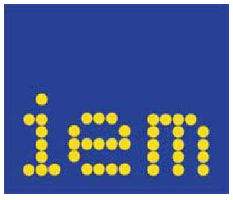
Voice Transformation

Diplomarbeit

durchgeführt und vorgelegt von
Jens Ahrens

Betreuung: o.Univ.-Prof. Mag. DI Dr. Robert Höldrich

durchgeführt am
Institut für Elektronische Musik und Akustik
Inffeldgasse 10/3
A-8010 Graz



Zusammenfassung

Die vorliegende Diplomarbeit befasst sich mit der Transformation der menschlichen Stimme. Der Schwerpunkt liegt dabei in der Manipulation von allgemeinen Charakteristika von Sprachsignalen (z.B. dem Geschlecht oder dem Alter eines Sprechers).
Anhand der bestehenden Literatur wurden die akustischen Unterschiede der Sprachsignale verschiedener Sprechergruppen wie z.B. Männern, Frauen, alten und jungen Erwachsenen ermittelt und zusammengefasst. Mit Hilfe zweier in der Software MATLAB implementierter Systeme wurden Transformationen von Sprachaufnahmen zwischen den Sprechergruppen und von stimmhafter Sprache in Flüstern durchgeführt. Das erste System beruht auf der *Spectral-Modeling-Synthese*, das zweite auf der Resynthese von einzelnen Glottispulsen.
Ein weiterer Teil der Arbeit gibt einen Überblick über die Literatur bezüglich der Verfahren zur Konvertierung einer Stimme in jene einer anderen (personifizierten) Person.

Abstract

The presented thesis deals with the transformation of the human voice. The focus is on the manipulation of general characteristics of speech signals (e.g. the sex or the age of a speaker).
The acoustic differences of speech signals of different groups of speakers such as men, women, young and old adults etc. have been determined and summarized on the basis of the available literature. With the help of two systems implemented in the software MATLAB, transformations of recordings of speech between the speaker groups and of voiced speech into whispering have been accomplished. The first system is based on *spectral modeling synthesis*, the second one is based on the resynthesis of single glottal pulses.
Furthermore, an overview over the literature about methods to map the voice of a speaker into the voice of a different person is given.

Danksagung

An dieser Stelle möchte ich all jenen danken, die durch ihre fachliche und persönliche Unterstützung zum Gelingen dieser Diplomarbeit beigetragen haben.

In erster Linie danke ich Herrn o.Univ.-Prof. Mag. DI Dr. Robert Höldrich für die kompetente Betreuung.

Darüberhinaus bedanke ich mich bei Herrn Thomas Musil, Herrn Univ. Ass. DI Dr. Alois Sontacchi und Herrn DI Markus Noisternig für die zahlreichen wissenschaftlichen Ratschläge, welche stets konstruktiv zu meiner Arbeit beigetragen haben.

Besonderer Dank gebührt natürlich auch meinen Eltern, die mir dieses Studium durch ihre vielfältige Unterstützung ermöglicht haben.

Inhaltsverzeichnis

Abkürzungen

DFT	Discrete Fourier Transformation
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transformation
FD-PSOLA	Frequency Domain Pitch Synchronous Overlap-Add
FFT	Fast Fourrier Transformation
IFFT	Inverse Fast Fourrier Transformation
IPSE	Improved Power Spectrum Envelope
LP	Linear Prediction
LPCs	Linear Prediction Coefficients
LSFs	Line Spectral Frequencies
MFPA	Maximally Flat Phase Alignment
OLA	Overlap Add
SD	Spectral Distortion
SMS	Spectral Modeling Synthesis
SPRM	Sinusoidal Plus Residual Model
STASC	Speaker Transformation Algorithm Using Segmental Codebooks
STFT	Short Time Fourier Transformation
TD-PSOLA	Time Domain Pitch Synchronous Overlap-Add
VC	Voice Conversion
VT	Voice Transformation

Verzeichnis der Hörbeispiele

Kapitel 1

Einleitung und Begriffsklärung

Begriffsklärung

In der bestehenden Literatur wird nicht konsequent zwischen den Begriffen *Voice Conversion* (VC) und *Voice Transformation* (VT) unterschieden. Jedoch wird der Begriff VC eher mit der Konvertierung eines Sprechers in einen anderen konkreten (personifizierten) Sprecher in Verbindung gebracht (z.B. Abe et al. 1988, Arslan 1999, Türk 2003). VT wird eher für die Transformation eines Merkmals wie z.B. das Geschlecht oder das Alter eines Sprechers verwendet (z.B. Traunmüller et al. 1989), obwohl für beide Fälle Ausnahmen zu finden sind.
In der vorliegenden Diplomarbeit wird jedoch genau zwischen den beiden oben genannten Fällen unterschieden, da sie beide zum Teil sehr unterschiedliche Ansätze der Manipulation darstellen. Um Verwechslungen zu vermeiden wird in dieser Arbeit die Konvertierung eines Sprechers in einen konkreten anderen Sprecher konsequent mit *Voice Conversion* bezeichnet; die Transformation eines Merkmals wird *Voice Transformation* genannt, auch wenn in der zugrunde liegenden Literatur zum Teil der jeweils andere Begriff verwendet wird.

Die vorliegende Diplomarbeit ist wie folgt gegliedert:

Das Kapitel 2 *Grundlagen der Sprachproduktion* beschreibt grob die anatomischen Grundlagen die zum Verständnis der Funktion des Sprechapparates nötig sind. In einem weiteren Abschnitt wird kurz auf die akustischen Grundlagen wie z.B. die je nach produziertem Laut unterschiedlichen Anregungssignale oder das Entstehen der Formanten eingegangen.
Darüber hinaus werden die akustischen Unterschiede zwischen verschiedenen Gruppen von Sprechern wie z.B. Männer, Frauen, Kinder etc. sowie die Veränderungen der Stimme mit zunehmendem Alter beschrieben.

Kapitel 3 gibt einen Überblick über die bestehenden Ansätze zur VC. Unter anderem werden die gängigsten Methoden zur Modellierung des Vokaltraktfilters und des Anregungssignals beschrieben. Es wird sowohl auf das Training von VC-Systemen als auch auf die verschiedenen Möglichkeiten bezüglich der Transformation der betreffenden

Parameter von Sprachsignalen eingegangen.

Die Kapitel 4 und 5 befassen sich mit der VT. Dort werden die im Rahmen dieser Diplomarbeit implementierten VT-Systeme vorgestellt und die Ergebnisse der Transformationen durch Hörbeispiele veranschaulicht.
Das System in Kapitel 4 verwendet die *Spectral-Modeling-Synthese*, das System in Kapitel 5 basiert auf der Resynthese von einzelnen Glottispulsen.

Auf der beiliegenden CD finden sich alle im Text erwähnten Hörbeispiele sowie sämtliche Matlab-Skripts die für die durchgeführten Manipulationen verwendet wurden.

Kapitel 2

Grundlagen der Sprachproduktion

2.1 Anatomische Grundlagen des Sprachapparates

Der Prozess der Sprachproduktion wird durch einen Luftstrom aus den Lungen eingeleitet. Dieser passiert den Larynx (Kehlkopf) welcher die Stimmlippen enthält. Der Bereich zwischen den Stimmlippen wird Glottis (Stimmritze) genannt. Die entstandene Anregung wird durch die Hohlräume des Vokaltraktes gefiltert, und das entstandene Signal wird abgestrahlt (Friedrich und Bigenzahn 1994). Abbildung 2.1 zeigt den Kehlkopf und den angrenzenden Vokaltrakt.

2.2 Akustische Grundlagen der Sprachproduktion

2.2.1 Modell der Sprache

Ein weit verbreitetes Modell der menschlichen Sprache ist das so genannte Quelle/Filter-Modell. Ein Signal wird als ein Anregungssignal angenommen, das ein Filter durchläuft. Bei Sprachsignalen repräsentiert das Filter den Einfluss des Vokaltrakts, der das Anregungssignal, z.B. der Glottis, zum abgestrahlten Signal formt (vgl. Abbildung 2.2).

2.2.2 Das Anregungssignal

Bei stimmhaften Lauten wie z.B. /a/ und /e/ vibrieren die Stimmlippen und erzeugen so ein quasi-periodisches Anregungssignal. Bei stimmlosen Lauten wie z.B. /s/ und /f/ sind die Stimmlippen geöffnet, und das Anregungssignal wird durch Luftturbulenzen hervorgerufen und hat eine rauschähnliche spektrale Energieverteilung (Shoup et al. 1988). Das Anregungssignal wird nun durch den Vokaltrakt gefiltert (siehe unten), wobei auch die Zunge und die Lippen Einfluss auf das Ausgangssignal haben.
Explosivlaute wie z.B. /p/ und /t/ werden erzeugt, indem im durch den Mund verschlossenen Vokaltrakt ein hoher Druck aufgebaut wird, welcher dann durch Öffnen des Mundes explosionsartig abgelassen wird. Darüber hinaus existieren noch Mischformen

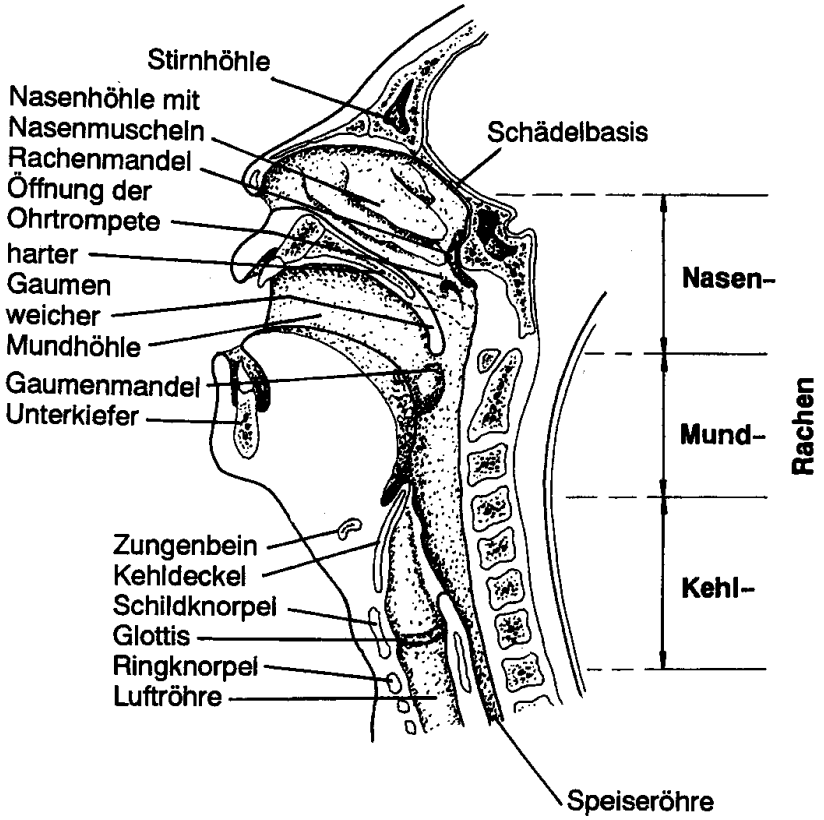


Abb. 2.1: Kehlkopf und Vokaltrakt (Friedrich und Bigenzahn 1994, Seite 37)

Abb. 2.2: Modell der Klangproduktion

der drei genannten Arten von Lauten, bei denen z.B. ein transienter Vorgang den stimmhaften Laut einleitet (z.B. /b/, /d/ oder /g/).

2.2.3 Die Vokaltrakttransferfunktion

Der Vokaltrakt kann als ein Rohr mit einem offenen Ende (geöffnete Lippen) und einem geschlossenen Ende (Glottis) beschrieben werden (Shoup et al. 1988, Atal und Hanauer 1971). Eine am geschlossenen Ende eingespeiste Welle pflanzt sich durch das Rohr fort und wird am offenen Ende teilweise reflektiert, wieder zurück in das Rohr geleitet und kann so stehende Wellen entstehen lassen (Resonanz). Diese stehenden Wellen sind durch eine bestimmte Verteilung von Schwingungsbäuchen und Schwingungsknoten charakterisiert und entstehen nur an bestimmten Frequenzen. Abbildung 2.3 zeigt die ersten vier Schwingungsmoden eines auf einer Seite offenen Rohres. A_i markiert die

Schwingungsbäuche.

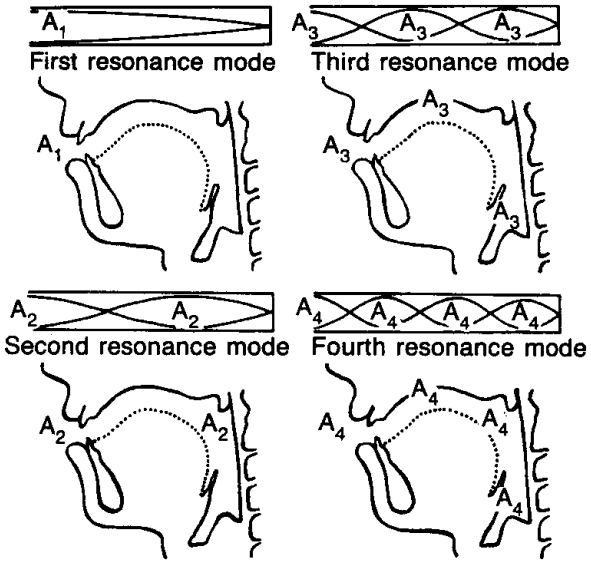


Abb. 2.3: Die ersten vier Schwingungsmoden eines auf einer Seite offenen Rohres (Shoup et al. 1988, Seite 179)

Die stehende Welle des ersten Schwingungsmodus weist die größte Wellenlänge – nämlich das Vierfache der Rohrlänge - und somit die tiefste Frequenz auf. Bei einer Rohrlänge von 17 cm, was etwa der Vokaltraktlänge eines erwachsenen Mannes entspricht, bedeutet dies eine Frequenz von ca. 500 Hz. Die nächst höheren Moden treten bei der 17-cm-Röhre bei 1500 Hz, 2500 Hz, usw. auf. Durch Veränderung der geometrischen Eigenschaften des Vokaltraktes (z.B. Schließen der Lippen) können die Resonanzfrequenzen verschoben werden.
Die Resonanzen werden Formanten genannt. Für die Wahrnehmung von Sprache sind meist nur die ersten zwei bis drei Formanten ausschlaggebend.
Die Verteilung der Formanten über den relevanten Frequenzbereich wird Transferfunktion des Vokaltraktes genannt. Sie spiegelt die Übertragungseigenschaften des Filters wieder und gibt das Verhältnis aus Ausgangs- und Eingangsignal des Filters an. Da sich die Filtereigenschaften des Vokaltrakts nur verhältnismäßig langsam mit der Zeit ändern, ist es möglich die Filterparameter eines Sprachsegmentes von 10ms-40ms zu schätzen (vgl. Abschnitt 3.2).

2.3 Beziehungen der charakteristischen Frequenzen von Vokalen zwischen verschiedenen Sprechern

2.3.1 Allgemeines

In den folgenden Abschnitten werden ausschließlich Unterschiede der charakteristischen Frequenzen (mittlere Grundfrequenz F_0, Formantmittenfrequenzen $F_1, F_2, \ldots$) von Vokalen zwischen verschiedenen Sprechergruppen beschrieben. Dies liegt in der Hauptsache daran, dass praktisch keine Messungen von Konsonanten in der Literatur verfügbar sind. Außerdem zeigte sich bei Versuchen der Analyse und Resynthese, nach erfolgter Transformation der charakteristischen Frequenzen zwischen verschiedenen Sprechergruppen, dass die angeführten Transformationsregeln nicht nur für Vokale sondern für alle Arten von Sprachsegmenten gelten (Traunmüller 1989).
Über die relativen Amplituden und die Bandbreiten der Formanten wird meist auch nicht berichtet. Nur Peterson und Barney (1952) merken an, dass sich bei ihrer Messung der Unterschiede zwischen den Vokalen von Männern und Frauen herausstellte, dass die relativen Amplituden der Formanten zwar starke individuelle Unterschiede zeigten, diese aber nicht auf das Geschlecht bezogen werden konnten.

2.3.2 Gesprochene Sprache

Die Abweichungen der charakteristischen Frequenzen von Vokalen zwischen verschiedenen Sprechern sind offensichtlich eine Konsequenz der Größenunterschiede der involvierten Organe. In erster Näherung können alters- und geschlechtsabhängige Unterschiede der Formantmittenfrequenzen mit einer entsprechenden Skalierung aller drei Dimensionen des Vokaltrakts beschrieben werden. Eine derartige Reskalierung würde aber die Verhältnisse der Formantfrequenzen unbeeinflusst lassen. Es zeigte sich aber (Fant 1975), dass die Verhältnisse der Formantfrequenzen zwischen verschiedenen Männern und Frauen gleichförmig abweichen. Diese Abweichungen führen zu dem Schluss, dass die Gestalt des Vokaltraktes bei Männern und Frauen bei der Produktion desselben Vokals nicht proportional ist. Traunmüller (1988) berichtet von verschiedenen Versuchen die geschlechtsspezifischen Unterschiede der Formantfrequenzen anhand anatomischer Daten zu berechnen. Die Unterschiede von F_2 und F_3 konnten zufrieden stellend vorhergesagt werden. Probleme ergaben sich allerdings bei F_1.
Bei der Betrachtung von ausschließlich erwachsenen Sprechern ist für einige Anwendungen die Annahme akzeptabel, dass die gleichen Vokale ungefähr gleiche Formantfrequenzen haben. Abbildung 2.4 zeigt, dass diese Annahme beim Vergleich von Erwachsenen und Kindern nicht mehr zulässig ist. Die von den mittleren Formantfrequenzen begrenzten Flächen sind nicht nur nicht übereinstimmend, sie überlappen nicht einmal.
Der Abbildung liegen Daten einer Untersuchung (Fujisaki et al. 1970) der fünf japanischen Vokale von sechs verschiedenen Sprechergruppen zugrunde. Jede der Gruppen (Männer, Frauen, Mädchen (12-14 Jahre), Jungen (12-14) vor dem Stimmbruch,

Jungen (12-14) nach dem Stimmbruch, Kinder (4-5)) bestand aus fünf Sprechern. Die abgebildeten Werte stellen Mittelwerte über jeweils alle fünf Sprecher einer Gruppe dar.

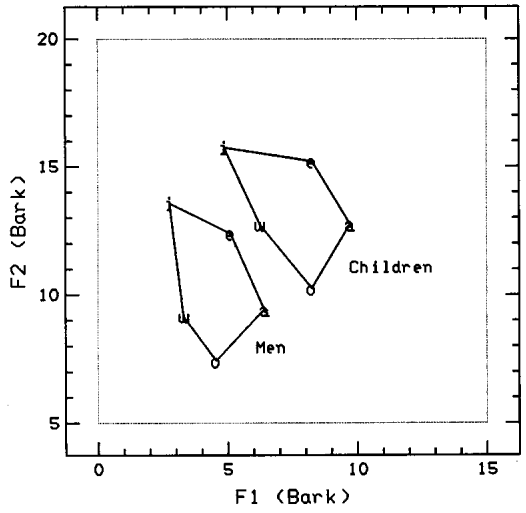


Abb. 2.4: Vergleich von F_1 (horizontal) und F_2 (vertikal) zwischen Männern und Kindern (4-5 Jahre), (Traunmüller 1988, Seite 11)

Die Beziehungen der Frequenzen der Formanten zwischen den Gruppen sind in den Abbildungen 2.5 (a)-(c) dargestellt.
Jede Abbildung zeigt die Abweichungen des Logarithmus der mittleren charakteristischen Formantfequenzen von einem willkürlichen Bezugswert für einige der verschiedenen Gruppen. Als Bezugswert wurde der Mittelwert über alle Sprecher gewählt, mit doppelter Gewichtung der ersten drei genannten Gruppen. Die Regressionsgeraden sind ebenfalls eingezeichnet.

Bemerkenswert ist der Umstand, dass die Beziehungen zwischen Männern und Jungen nach dem Stimmbruch mittels eines Skalierungsfaktors (1.14) beschrieben werden können: Die jeweiligen Regressionsgeraden verlaufen parallel. Die Unterschiede zwischen Männern und Frauen sowie zwischen Erwachsenen und Kindern lassen eine solche Beschreibung nicht zu. Die Regressionsgeraden verlaufen nicht parallel.
Nichtsdestotrotz lassen sich die Ergebnisse der Untersuchung mittels einer Regressionsgeraden beschreiben, deren Steigung sich mit der Maturation ändert (Traunmüller 1988). Die Gleichung 2.1 beschreibt die Beziehungen zwischen den Formatfrequenzen F_a und F_b verschiedener Sprechergruppen entsprechend der Regressionsgeraden:

$$F_b = k_1 \cdot F_a^p \tag{2.1}$$

Die Konstanten k_1 und p sind sprechergruppenspezifisch. k_1 repräsentiert den Skalierungsfaktor für $F_a = 1$ Frequenzeinheit. Tabelle 2.1 zeigt die Werte der Konstanten für einige ausgewählte Sprechergruppen. Skalierungsfaktoren für F_0 sowie angepasste Faktoren zur linearen Beschreibung der Beziehungen der Formantfrequenzen (k) sind ebenfalls aufgelistet. Um eine Referenz bezüglich der Unterschiede zweier Sprechergruppen zu haben sind die jeweiligen Skalierungsfaktoren bei $F_a = 300$ (k_{300}) und $F_a = 3000$ (k_{3000}) angegeben. Die Konturen der Grundfrequenz müssen generell

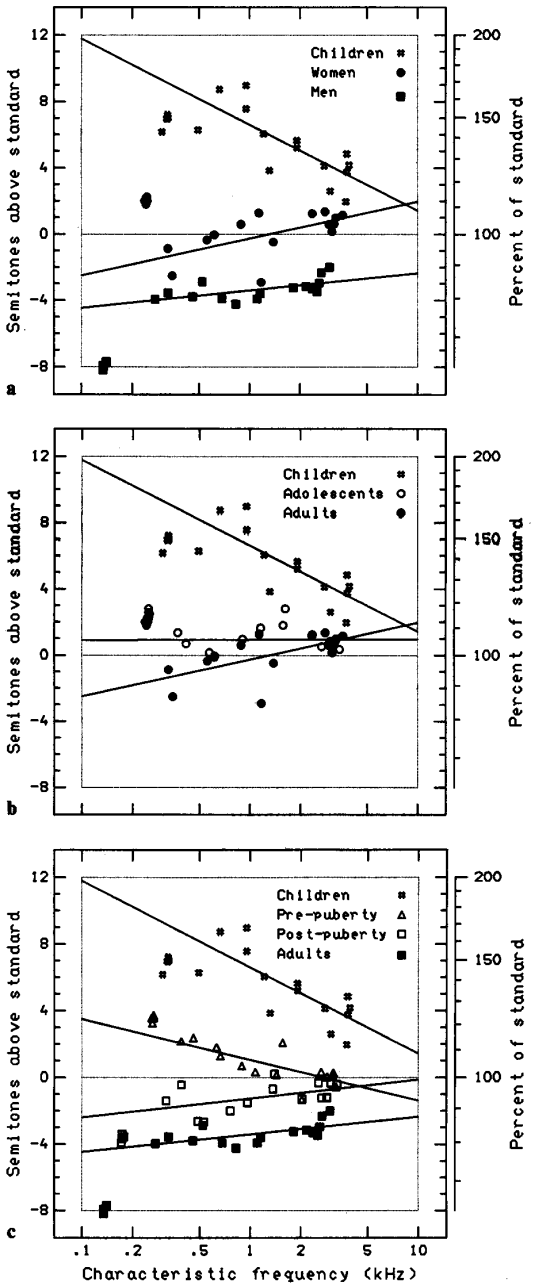


Abb. 2.5: Abweichung des Logarithmus von F_0, F_1, F_2 und F_3 verschiedener Sprechergruppen von einem Bezugswert (*standard*, siehe Text) gegenüber dem tatsächlichen Wert (Traunmüller 1988, Seite 12)

nicht korrigiert werden. Eine einfache Tonhöhenskalierung führt zu einem natürlich wirkenden Ergebnis (Eklund und Traunmüller 1997).

Abbildung 2.6 (a,b) zeigt, dass die Beziehungen der Formantfrequenzen desselben Sprechers bei unterschiedlicher Artikulation ebenfalls mit Gleichung 2.1 beschrieben werden können. Abbildung 2.6 (a) zeigt einen Vergleich von Rufen und Sprechen von zehn schwedischen Vokalen dreier Sprecher. Abbildung 2.6 (b) zeigt einen Vergleich von Flüstern und Sprechen von fünf Vokalen des amerikanischen Englisch von jeweils 15 Männern und Frauen.

F_a	F_b	k_{F0}	k	k_{300}	k_{3000}	k_1	p
Männer	Jungen (pubertiert, 12-14 J.)	1.29	1.14	1.14	1.13	1.14	0.999
Männer	Jungen (nicht-pubertiert, 12-14 J.)	1.93	1.26	1.41	1.16	2.29	0.915
Männer	Kinder, 4-5 Jahre	2.34	1.64	1.99	1.42	4.55	0.855
Frauen	Mädchen, 12-14 Jahre	1.03	1.05	1.15	1.01	1.62	0.940
Frauen	Kinder, 4-5 Jahre	1.32	1.34	1.75	1.18	4.72	0.827
Männer	Frauen	1.77	1.22	1.18	1.25	1.02	1.026
Männer	Frauen, aus (Fant 1975)	-	1.14	1.08	1.19	0.84	1.044
Gesprochen	gerufen (Männer)	2.12	1.14	1.36	0.99	2.99	0.862
Gesprochen	geflüstert (Männer)	-	1.14	1.43	0.97	3.73	0.831
Gesprochen	geflüstert (Frauen)	-	1.08	1.30	0.99	2.55	0.882

Tab. 2.1: Skalierungsfaktoren für Transformationen gemäß Gleichung 2.1; Daten aus (Traunmüller 1988)

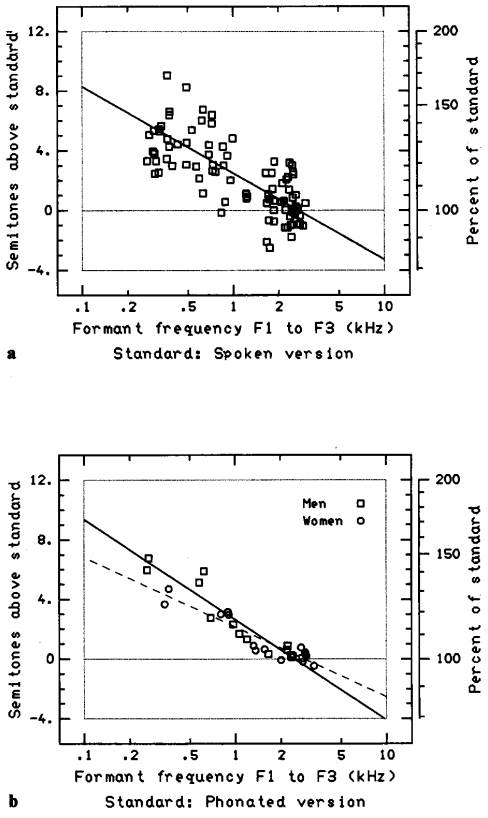


Abb. 2.6: Vergleich der Formantfrequenzen (F_1, F_2, F_3, logarithmisch) von Sprechern a: rufen und sprechen, b: flüstern und sprechen, Regressionsgerade für Frauen ist gestrichelt; (Traunmüller 1988, Seite 14)

2.3.3 Invariante Parameter

Traunmüller (1988) berichtet, dass Variationen der Größen wie Intonation und Phonation F_3 und die höheren Formantfrequenzen nur geringfügig beeinflussen. Diese Größen können nun herangezogen werden, um Aufschluss über die Körpergröße eines Sprechers zu erhalten, da die Lage dieser Formanten maßgeblich von der Größe des Vokaltraktes abhängig ist. Weitere Erläuterungen sind Traunmüller (1988) zu entnehmen. Allerdings liegen keine quantitativen Messungen vor.

2.3.4 Besonderheiten der Phonation von Frauen

Die oben angeführten Darstellungen beschreiben die Beziehungen der charakteristischen Frequenzen von Vokalen bei den meisten Kombinationen von Sprechergruppen sehr gut. Bei genauerer Analyse zeigen sich Diskrepanzen vor allem beim Vergleich von erwachsenen Männern und Frauen (Traunmüller 1988). Abbildung 2.7 (a) zeigt die Abweichungen der Frequenz von F_1, F_2 und F_3 in Halbtönen, Abbildung 2.7 (b) der Critical-Band-Rate in Bark. Die Regressionsgeraden sind für jeden Formanten separat (durchgezogen) sowie für alle Formanten zusammen (gestrichelt) eingezeichnet.

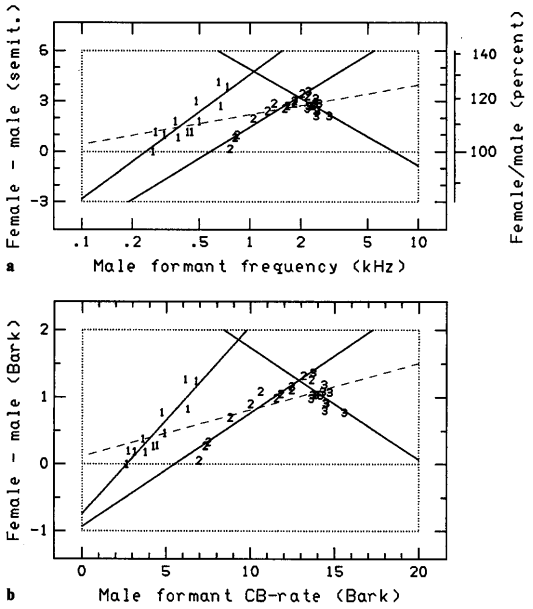


Abb. 2.7: Vergleich der Formantfrequenzen von F_1, F_2 und F_3 zwischen Männern und Frauen (Traunmüller 1988, Seite 20)

Die Abweichungen der Regressionsgeraden der einzelnen Formanten von der gemeinsamen Regressionsgeraden sind deutlich zu sehen. Die Formanten müssen also getrennt voneinander betrachtet werden.
Tabelle 2.2 listet für diesen Fall die entsprechenden Werte der Konstanten aus Gleichung 2.1 auf. Eine klangliche Verbesserung der Ergebnisse bei der Transformation von Männer- in Frauenstimmen bei separater Behandlung der Formanten wird von Traunmüller et al. (1989) berichtet.

	F_1	F_2	F_3
Uniform scale factor	1.116	1.140	1.168
Scale factor k_1 bei 1 Hz	0.362	0.373	-
Exponent p	1.186	1.155	-

Tab. 2.2: Faktoren und Exponenten der Gleichung (2.1) beim Vergleich Männer/Frauen für F_1, F_2 und F_3; Daten aus (Traunmüller 1988)

2.3.5 Flüstern

Unterschied zwischen Flüstern und gesprochener Sprache besteht hautsächlich in der unterschiedlichen Anregung des Vokaltraktes. An die Stelle der Anregung durch die Stimmlippen treten lediglich Luftturbulenzen im Rachenraum, der aufgrund der geöffneten Glottis von einer größeren Luftmenge durchströmt wird als beim Sprechen. Das Spektrum dieser Anregung ist ähnlich dem eines Rauschens (Eklund und Traunmüller 1997).
Die veröffentlichten Daten zeigen auch Erhöhungen der charakteristischen Frequenzen der ersten drei Formanten bei Flüstern (Peterson 1961, Kallail et al. 1984a,b). In den Daten von Kallail et al. (1984a,b) betragen die Erhöhungen 147 Hz für F_1, 70 Hz für F_2 sowie 46 Hz für F_3.
Die Autoren ordneten diese Erhöhungen den veränderten geometrischen Verhältnissen im Rachenraum aufgrund der offenen Glottis zu. Es erscheint allerdings unwahrscheinlich, dass vor allem die starke Erhöhung von F_1 so erklärt werden kann. Vielmehr zeigen die Formanten beim Flüstern ähnliche charakteristische Frequenzen wie bei Sprache, die zur besseren Verständlichkeit mit größerem Einsatz artikuliert wird. Eklund und Traunmüller (1997) halten es für angemessener, anzunehmen, dass Flüstern tendenziell mit mehr Einsatz erzeugt wird als gesprochene Sprache, um der geringeren Verständlichkeit aufgrund der geringeren Energie der Anregung entgegenzuwirken. Die Beziehungen zwischen den Formantfrequenzen von gesprochener Sprache und Flüstern sind bereits in Tabelle 2.1 aufgeführt.

Abbildung 2.8 zeigt den Unterschied des Schalldruckpegels von Flüstern und gesprochener Sprache (jeweils ausschließlich Vokale) im Abstand von ca. 5 cm von den Lippen. Die Frequenz ist logarithmisch (Basis 10) aufgetragen. Der dargestellte Frequenzbereich beträgt also 100 Hz bis 10 kHz.

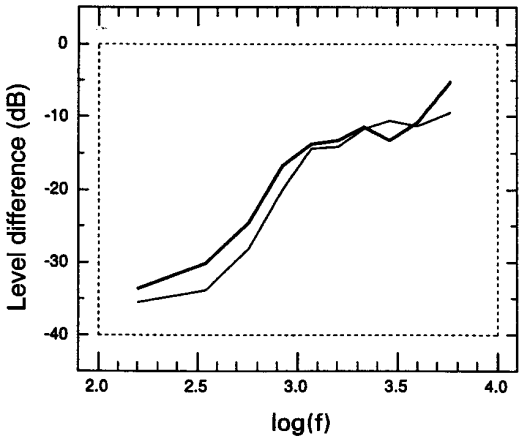


Abb. 2.8: Differenz des Schalldruckpegels zwischen gesprochener Sprache und Flüstern (Eklund und Traunmüller 1997, Seite 19)

Gesprochene Sprache kann nun in Flüstern transformiert werden, indem die ursprüngliche Erregung durch weißes Rauschen ausgetauscht wird, das entsprechend Abbildung 2.8 gefiltert wurde (Eklund und Traunmüller 1997, vgl. hierzu Hörbeispiele 1 und 2 in Abschnitt 3.2.1).

2.4 Veränderungen der Stimme mit zunehmendem Alter

2.4.1 Allgemeines

Die Fähigkeit von Hörern aufgrund von Sprachaufnahmen auf das Alter des jeweiligen Sprechers zu schließen wurde in etlichen Untersuchungen bestätigt (z.B. Ptacek et al. 1966; Ship und Hollien 1969). Ptacek et al. (1966) ließen das Alter (unter 35, über 65) von 72 männlichen und weiblichen Sprechern schätzen. Die zehn befragten Studenten erzielten bei der Schätzung von vorwärts abgespielten Aufnahmen einer gelesenen Passage eine Trefferquote von 99%, bei rückwärts abgespielten Aufnahmen 87% und bei ausgehaltenen Vokalen 78%. Shipp und Hollien (1969) ließen 25 untrainierte Hörer direkt das Alter von 125 männlichen Sprechern im Alter von 20 bis 89 Jahren anhand von Aufnahmen von gelesen Passagen schätzen. Das mittlere empfundene Alter korrelierte mit dem Faktor 0.88 mit dem tatsächlichen.
In weiterer Folge wurde der Einfluss des Alterns auf verschiedene akustische Eigenschaften der menschliche Stimme untersucht (z.B. Endres et al. 1971, Linville und Fisher 1985, Ramig und Ringel 1983). Als die Parameter mit dem größten Einfluss auf das empfundene Alter stellten sich die Stabilität von F_0, Jitter (Variationen in der Periodendauer aufeinander folgender Schwingungen der Grundfrequenz), Shimmer (Amplitudenschwankungen aufeinander folgender Schwingungen), Stabilität der Grundfrequenz sowie die Lage von F_1 (Linville und Fisher 1985) heraus.
Bei der Messung dieser Größen zeigten sich allerdings große individuelle Unterschiede innerhalb einer Altersgruppe (Ramig und Ringel 1983). Zwar wird von der Annahme ausgegangen, dass das physische Altern der beim Sprechen involvierten Organe die akustischen Veränderungen verursacht, Haberman (1972) stellt aber fest, dass eben bei diesem physischen Altern große Individuelle Unterschiede zu beobachten sind.

2.4.2 Formanten

Endres et al. (1971) nahmen über 29 Jahre hinweg regelmäßig Sprachproben von sieben verschiedenen Sprechern. Für jeden Sprecher S_i wurden zu jedem Zeitpunkt t_i einer Probe $N = 4$ Formant-Mittenfrequenzen bestimmt, welche zu einer bestimmten Gruppe K von Phonemen gehören. Dies wurde jeweils mit der größtmöglichen Anzahl M von Vertretern dieser Phoneme durchgeführt. Mittels der Gleichung 2.2

$$\langle \nu_{S_i,K}(t_i) \rangle = \frac{1}{M_{(S_i,K)}(t_i)} \frac{1}{N} \sum_{m=1}^{M} \sum_{n=1}^{N} \nu_{(S_i,K)mn}(t_i) \tag{2.2}$$

wurde ein Mittelwert $\langle \nu_{S_i,K}(t_i) \rangle$ berechnet welcher als *Point of Formant Concentration* (Endres et al. 1971) bezeichnet wird. Abbildung 2.9 zeigt eine graphische Darstellung der Ergebnisse für sechs der Sprecher. Bei allen zeigte sich ein Absinken der Points of Formant Concentration.

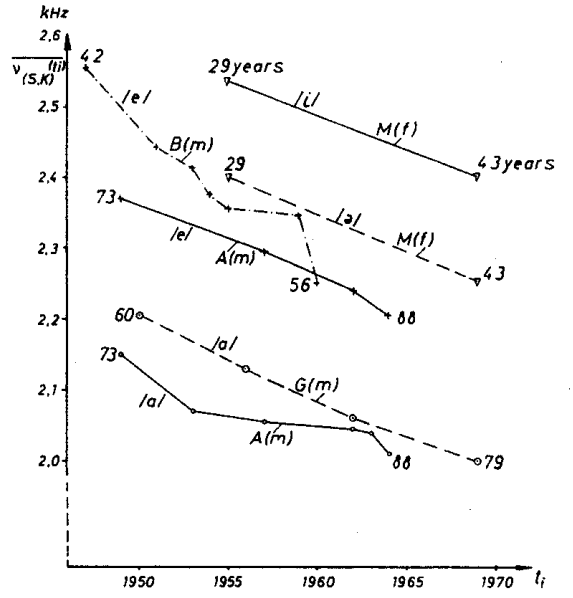


Abb. 2.9: Points of Formant Concentration in Abhängigkeit der Zeit; m steht für einen männlichen Sprecher, f für einen weiblichen; das Alter der Sprecher ist am Anfang und am Ende jeder Messreihe angegeben; (Endres et al. 1971, Seite 1844)

Linville und Fisher (1985) untersuchten 75 kaukasische Frauen der Altergruppen 25-35 Jahre, 45-55 Jahre und 70-80 Jahre unter anderem bezüglich ihrer ersten beiden Formanten bei ausgehaltenen Vokalen. Die Ergebnisse sind in Abbildung 2.10 dargestellt.

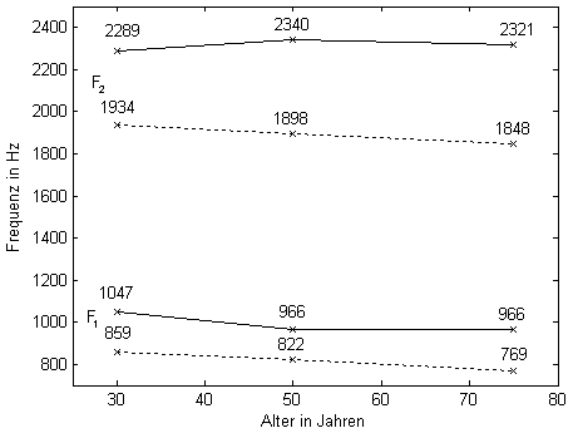


Abb. 2.10: Mittenfrequenzen von F_1 und F_2 gegenüber dem Alter der Sprecherinnen; Sprechen gestrichelt, Flüstern durchgezogen; Daten aus (Linville und Fisher 1985)

Laut der Hörtests von Linville und Fisher (1985) ist die Lage der Formanten vor allem bei der Beurteilung von Flüstern maßgebend, da ja dort keine Cues des Erregungssignals (Shimmer, Jitter etc.) vorhanden sind. Dabei wird vor allem ein tiefer erster Formant wird mit hohem Alter assoziiert.

2.4.3 Die Grundfrequenz

Bezüglich der Entwicklung der mittleren Grundfrequenz F_0 mit steigendem Alter finden sich widersprüchliche Angaben. Bei einigen Untersuchungen sinkt sie (z.B. Endres et al. 1971, Linville und Fisher 1985), bei anderen steigt sie (Hollien und Shipp 1972). Die

mittlere Grundfrequenz des Sprechers B der Untersuchung von Endres et al. (1971) (vgl. Abb. 2.9) z.B. sank im Zeitraum von 1947 bis 1960 monoton von 136 Hz auf 93 Hz.
Bei Hollien und Shipp (1972) hingegen zeigte sich eine Erhöhung von F_0, vor allem im hohen Alter. Sie untersuchten 172 Männer im Alter von 20 bis 89 Jahren. Abbildung Abb. 2.11 zeigt die Ergebnisse. Dargestellt sind die mittleren Grundfrequenzen der einzelnen Sprecher sowie die Mittlerwerte über jede Altersdekade (große Punkte).

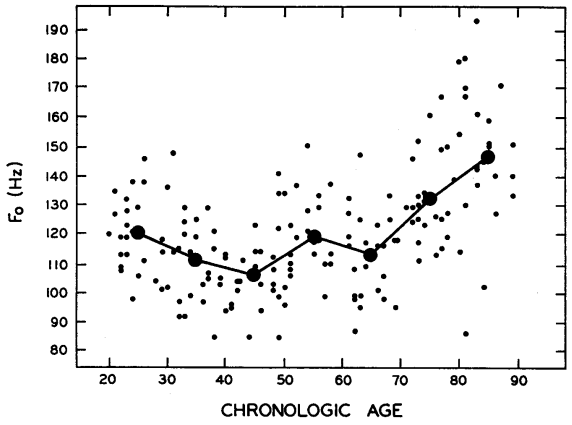


Abb. 2.11: Mittlere Grundfrequenzen verschiedener Sprecher im Vergleich zu deren Alter (Hollien und Shipp 1972, Seite 156)

Eine Erklärung für diese Diskrepanzen konnte nicht gefunden werden. Auffällig ist nur, dass die Autoren (z.B. Hollien und Shipp 1972, Endres et al. 1971, Linville und Fisher 1985) ihre Ergebnisse ausschließlich mit Messergebnissen anderer Untersuchungen vergleichen, die den gleichen Trend zeigen.

Einheitlich hingegen sind alle Beobachtungen bezüglich der Stabilität der Grundfrequenz (Endres et al. 1971, Linville und Fisher 1985, Ramig und Ringel 1983). Im zunehmenden Alter konzentrieren sich die momentanen Werte der Grundfrequenz zunehmend um deren Mittelwert. Abbildung 2.12 zeigt die Wahrscheinlichkeitsverteilungen der momentanen Grundfrequenz des bereits erwähnten Sprechers B (vgl. Abb. 2.9) aus der Untersuchung von Endres et al. (1971) für drei verschiedene Zeitpunkte (1947, 1955 und 1960). Die Maximalwerte der Verteilungen sind als die Mittelwerte der Grundfrequenz definiert.

2.4.4 Jitter und Shimmer

Tabelle 2.3 zeigt die Messergebnisse von Orlikoff (1990) bezüglich des Jitters und Shimmers bei der Artikulation von Vokalen durch sechs junge und sechs ältere Männer. Der mittlere Shimmer gibt den mittleren Amplitudenunterschied (in dB), der mittlere Jitter die mittlere Differenz der Periodendauer zweier aufeinander folgender Schwingungen des Erregungssignals an. Da der Jitter von der Grundfrequenz abhängig ist (Hollien et al. 1973; Orlikoff und Baken, 1990) (siehe unten), sind die Werte in % (bezogen auf die Periodendauer von F_0) angegeben. Informationen über ein derartiges Verhalten des

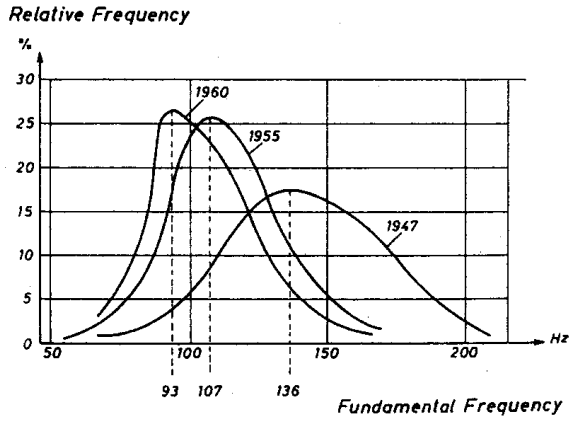


Abb. 2.12: Wahrscheinlichkeitsverteilungen der momentanen Grundfrequenz eines Sprechers zu drei Zeitpunkten (Endres et al. 1971, Seite 1844)

Shimmers sind nicht vorhanden. Der relative Shimmer in Tabelle 2.3 ist auf die mittlere Amplitude bezogen.

	Junge Männer	Ältere Männer
Mittlerer absoluter Jitter (ms)		
Mittelwert	0.042	0.053
SD	0.006	0.015
Bereich	0.034-0.051	0.038-0.081
Mittlerer relativer Jitter (%)		
Mittelwert	0.461	0.625
SD	0.067	0.102
Bereich	0.408-0.590	0.468-0.740
Mittlerer absoluter Shimmer (dB)		
Mittelwert	0.257	0.390
SD	0.088	0.113
Bereich	0.154-0.360	0.269-0.562
Mittlerer relativer Shimmer (%)		
Mittelwert	2.93	4.42
SD	1.00	1.24
Bereich	1.76-4.11	3.12-6.27

Tab. 2.3: Einfluss des Alterns auf ausgewählte akustische Eigenschaften der Stimme von Männern; Daten aus (Orlikoff 1990)

Ein Anstieg des Jitters sowie des Shimmers ist zu erkennen. Dies ist konsistent mit den Ergebnissen anderer Untersuchungen (Ramig und Ringel 1983, Linville und Fisher 1985). Es lassen sich jedoch keine quantitativen Aussagen machen.
In der Untersuchung von Linville und Fisher (1985) hat sich die Entwicklung des Jitters als das uneffektivste perzeptive Cue herausgestellt. Der Shimmer scheint für Wahrnehmung ausschlaggebender.
Orlikoff und Baken (1990) fanden heraus, dass eine signifikante Abhängigkeit des Jitters von der Lage der momentanen Grundfrequenz innerhalb des Stimmumfangs der Probanden besteht. Sie untersuchten sechs Männer und sechs Frauen im Alter

von 21-47 Jahren. Dabei wurde der Jitter (relativ und absolut) in Abhängigkeit des Abstandes der Grundfrequenz von einer selbst gewählten bequemen Tonhöhe gemessen. Abbildung 2.13 (a,b) zeigt die Ergebnisse. Die Frequenz ist relativ (in Halbtönen) zur selbst gewählten bequemen Tonhöhe angegeben. Der Jitter-Ratio ist definiert als mittlerer Jitter in Millisekunden dividiert durch die mittlere Periode und multipliziert mit 1000. Der Jitter-Ratio ist also das Zehnfache des relativen Jitters in %.
Die Kurve des Jitter-Ratio verläuft deutlich flacher als die des absoluten Jitters. Auffällig ist der starke Anstieg des Jitters bei sehr tiefen Frequenzen bei Männern (Abb. 2.13 (a)). Es scheint aber keine Geschlechtsabhängigkeit des Jitters gegeben zu sein.

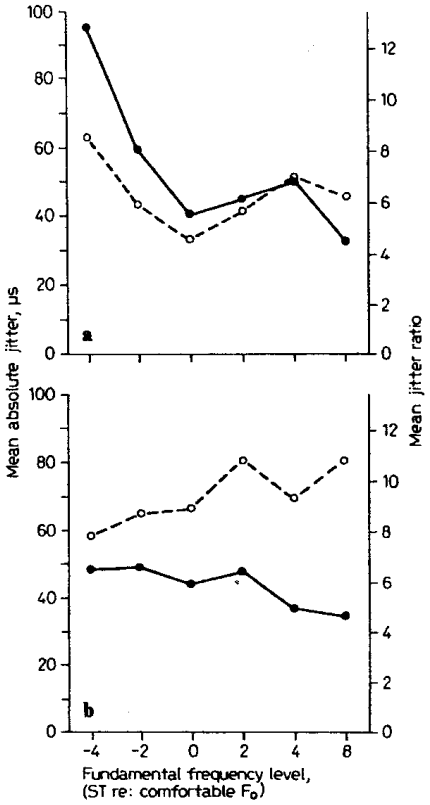


Abb. 2.13: Mittlerer absoluter Jitter (gefüllte Kreise) und mittlerer Jitter-Ratio (leere Kreise) in Abhängigkeit der Tonhöhe; a) Männer; b) Frauen; (Orlikoff und Baken 1990, Seite 37)

Die Ergebnisse unterschiedlicher Untersuchungen sind nun aufgrund der verschiedenen Messprozeduren nur bedingt vergleichbar. Orklikoff (1990) ließ Vokale in einer bequemen Tonhöhe artikulieren, Ramig und Ringel (1983) untersuchten gelesene Passagen bzw. freies Sprechen, und Linville und Fisher (1985) gaben bei ihren Messungen die Tonhöhe in einem gewissen Toleranzbereich vor. Darüber hinaus ist der Jitter meist nur relativ angegeben, was den Bezug zur Tonhöhe verschleiert.

Kapitel 3

Überblick über Voice-Conversion-Systeme

3.1 Einleitung

Voice Conversion (VC) bezeichnet die Manipulation der Stimme eines Sprechers, sodass sie wie die Stimme eines anderen Sprechers klingt. Der Sprecher dessen Stimme manipuliert wird wird Source-Sprecher genannt, der Sprecher dessen Stimme imitiert wird Target-Sprecher.

3.1.1 Aufbau eines Voice-Conversion-Systems

Es existiert eine Vielzahl von Ansätzen zur VC, wobei jeder Ansatz seine eigenen Vor- und natürlich auch Nachteile mit sich bringt. Die wichtigsten Vertreter werden in diesem Kapitel besprochen.
Allen in diesem Kapitel vorgestellten Verfahren liegt das in Abschnitt 2.2.1 beschriebene Quelle-/Filtermodell der Sprache zugrunde. Anregungssignal und Vokaltraktfilter werden getrennt voneinander modelliert und transformiert.
Alle resultierenden VC-Systeme haben zumindest die folgenden Komponenten gemeinsam:

- Eine Repräsentation der sprecherspezifischen Charakteristika eines Sprachsignals.
- Eine Methode die Charakteristika von Source- und Target-Sprecher miteinander in Beziehung zu setzen.
- Eine Methode die Charakteristika des Source-Sprechers so zu manipulieren, dass sie jenen des Target-Sprechers entsprechen.

Die VC gliedert sich in folgende Phasen (vgl. Abbildung 3.1):

Training: Beim Training des Systems mit geeigneten Sprachproben von Source- und Target-Sprecher werden die Parameter des Sprachmodells für beide Sprecher bestimmt und in Beziehung gesetzt. Diese Beziehungen werden in der *Transformationsfunktion* zusammengefasst.

Transformation: Bei der Transformation werden die Charakteristika eines Signals des Source-Sprechers mittels der Transformationsfunktion so manipuliert, dass sie jenen des Target-Sprechers entsprechen.

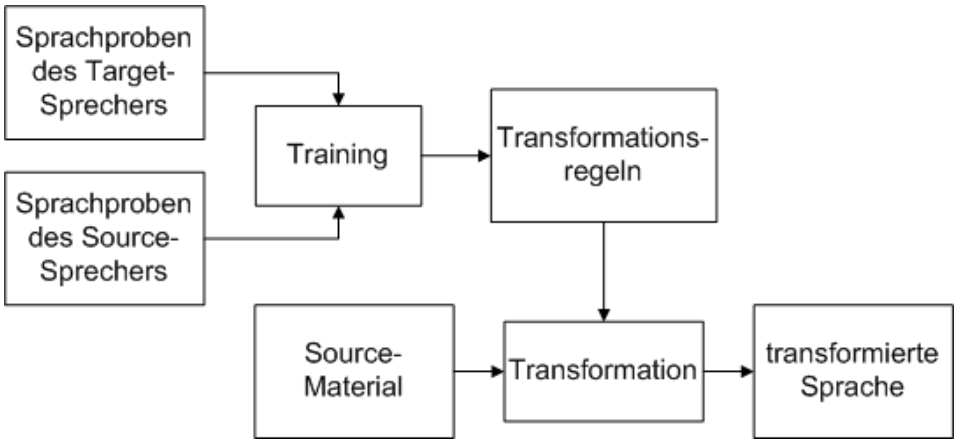


Abb. 3.1: Blockschaltbild der VC

Training

Um Transformationen zwischen zwei Sprechern durchführen zu können, muss das VC-System mit Sprachproben beider Sprecher trainiert werden. Bei diesem Training werden die akustischen und linguistischen Parameter von Source- und Target-Sprecher (z.B. Formantstrukur, Tonhöhe, Länge der Phoneme) bestimmt und in *Codebooks* gespeichert. Die Parameter können dann miteinander in Beziehung gesetzt und die Transformationsfunktion aufgestellt werden.

Die Mehrheit der VC-Systeme benötigt einen parallelen Sprachkorpus für das Training. Das bedeutet, von beiden Sprechern müssen die gleichen Sprachproben vorliegen. Die beiden Sprachproben werden vor dem Training zeitlich einander angeglichen.

Mouchtaris et al. (2005) schlagen ein Verfahren vor, das durch Adaption von Transformationsregeln eines Sprecherpaares die Konvertierung eines anderen Sprecherpaares ermöglicht, von dem kein paralleler Sprachkorpus verfügbar ist. Für die Aufstellung der Basistransformationsregeln ist allerdings ein paralleler Korpus notwendig.

Transformation

Sind die erforderlichen Modellparameter bestimmt, so kann die Transformation von Signalen durchgeführt werden.
Der klassische Ansatz ist die Verwendung von *Mapping Codebooks* (Abe et al. 1988). Mittels Vektorquantisierung werden die akustischen und linguistischen Eigenschaften der Sprecher partitioniert, und durch diskretes Mapping transformiert, was zu Diskontinuitäten im Ausgangssignal führte.
Eine Verbesserungsmöglichkeit ist die Verwendung entsprechender Interpolationsmethoden innerhalb einer diskreten Mapping-Umgebung. Ein Beispiel hierfür ist die *Speaker Transformation Algorithm using Segmental Codebooks* (STASC) von Arslan (1999).

Eine andere Möglichkeit besteht in der Verwendung von individuellen Transformationsfunktionen für jede Art bzw. Klasse von Segmenten von Sprachsignalen (lokale Transformationsfunktionen). Valbret et al. (1992) verwendeten sowohl lineare Regression (*Linear Multivariate Regression*) als auch *Dynamic Frequency Warping* als lokale Ansätze. Für jede Klasse berechnet ein Algorithmus während der Trainingsphase die optimalen Transformationen für beide Ansätze.
Mizuno und Abe (1995) berechneten ein von der Signalklasse abhängiges Set an linearen Transformationsregeln. Aufgrund der diskreten Natur der Transformationsfunktionen können aber auch hier Diskontinuitäten im Ausgangssignal entstehen.

Die Verwendung von *Artificial Neural Networks* oder statistischen Ansätzen wie *Gaussian Mixture Models* erlaubt das Aufstellen von kontinuierlichen Transformationsfunktionen. Ein Vergleich dieser Methoden ist in (Baudoin und Stylianou 1996) zu finden.

3.1.2 Anwendungen

VC bietet ein breites Spektrum an direkten und indirekten Anwendungen. Stellvertretend werden im Folgenden repräsentative Vertreter der verschiedenen Einsatzgebiete aufgezählt.

Nachbearbeitung von Sprachaufnahmen: In der Film- und Musikindustrie können unerwünschte Passagen in Aufnahmen ausgetauscht werden, wenn der eigentliche Sprecher nicht verfügbar ist. Ganze Produktionen können durch eine kleine Anzahl von Sprechern/Sängern verwirklicht werden.

Personifizierung von Dolmetschern: Dolmetscher können die Sprachcharakteristika der Person annehmen, die sie übersetzen. Filme können in andere Sprachen übersetzt werden, wobei die originale Stimme der Schauspieler beibehalten werden kann.

Erweiterung der Flexibilität von Text-to-Speech-Synthese-Systemen: Die Stimmenauswahl bei der Synthese von Sprache aus Text ist nicht mehr auf die dem System zu Grunde liegende Datenbasis beschränkt.

Signalkodierung: Sprachsignale können bei sehr niedrigen Bit-Raten ohne sprecherspezifische Merkmale übertragen werden. Beim Empfänger kann die Sprache dann wieder mit diesen Merkmalen resynthestisiert werden.

Unterhaltungsmedien: In Karaoke-Anwendungen kann jede beliebige Stimme in jene eines bekannten Künstlers transformiert werden.

3.2 Modellierung der Filterkomponente

3.2.1 Koeffizienten der linearen Prädiktion

Bei der linearen Prädiktion (*Linear Prediction* (LP)) wird das momentane Eingangssample $x[n]$ durch eine Linearkombination der vorhergehenden Samples des Eingangssignals geschätzt (Atal und Hanauer 1971):

$$\hat{x}[n] = \sum_{k=1}^{p} a_k x[n-k] \tag{3.1}$$

p bezeichnet die Ordnung der LP, a_k die Prädiktionskoeffizienten (*Linear Prediction Coeffitients* (LPCs)). Die Differenz aus dem tatsächlichen Eingangssignal $x[n]$ und seiner Prädiktion wird Prädiktionsfehler (*Prediction Error*) $e[n]$ genannt und gemäß

$$e[n] = x[n] - \hat{x}[n] = x[n] - \sum_{k=1}^{p} a_k x[n-k] \tag{3.2}$$

berechnet.

Mit der z-Transformation des Prädiktionsfilters

$$P(z) = \sum_{k=1}^{p} a_k z^{-k} \tag{3.3}$$

kann Gleichung (3.2) in der z-Domäne als

$$E(z) = X(z) - \hat{X}(z) = X(z) \cdot [1 - P(z)] \tag{3.4}$$

geschrieben werden.

Mit der Definition des Prädiktionsfehler-Filters $A(z)$ gemäß

$$A(z) = 1 - P(z) \tag{3.5}$$

lässt sich der Prädiktionsfehler nun über

$$E(z) = X(z) \cdot A(z) \tag{3.6}$$

berechnen. Mittels des Synthesefilters

$$V(z) = \frac{1}{A(z)} = \frac{1}{1 - P(z)} \tag{3.7}$$

kann das Eingangssignal wieder hergestellt werden:

$$X(z) = E(z) \cdot V(z) \tag{3.8}$$

$E(z)$ bzw. $e[n]$ beschreiben das Anregungssignal, $V(z)$ das Vokaltraktfilter. Das Vokaltraktfilter modelliert die Resonanzen und besitzt somit nur Pole.

Zur Berechnung der Filter- bzw. Prädiktorkoeffizienten a_k wurde eine Reihe von Verfahren entwickelt. Das bekannteste ist die auf der Levinson-Durbin-Rekursion basierende Autokorrelationsmethode. Sie und andere können z.B. (Atal und Hanauer 1971) entnommen werden.

Hörbeispiele

Hörbeispiel 1, (CD-Track 01): Frauenstimme, bandbegrenzt ($f_{Grenz} = 5.5kHz$, $f_{Sampling} = 22.05kHz$)

Hörbeispiel 2, (CD-Track 02): Hörbeispiel 1 durch Austauschen der Anregung mittels LP in Flüstern transformiert (vgl. Abschnitt 2.3.5)

Probleme mit den Prädiktorkoeffizienten:

Stabilität: Einige Verfahren zur Berechnung der Prädiktorkoeffizienten führen nicht zwingend zu einem stabilen Filter. Die Überprüfung ob ein bestimmtes Set an Koeffizienten ein stabiles Filter darstellt ist schwierig.

Quantisierung : Die Übertragungsfunktion eines Vokaltraktfilters reagiert sensibel auf Änderungen der Koeffizienten.

Interpolation: Die Interpolation zweier stabiler LPC-Sets führt nicht zu einer allmählichen Veränderung der Übertragungsfunktion. Die Stabilität ist auch nicht garantiert.

Aus den beschrieben Problemen heraus wurden andere Parametersets entwickelt. Am weitesten verbreitet sind die *Cepstral-Koeffizienten* und die *Line Spectral Frequencies*.

3.2.2 Cepstral-Koeffizienten

Das Cepstrum ist als die inverse Fourier-Transformation (IFT) des logarithmisierten Spektrums eines Signals $X(e^{j\omega})$ definiert (Oppenheim und Schafer 1999):

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \log(X(e^{j\omega}))e^{j\omega n} d\omega \tag{3.9}$$

Die Cepstral-Koefficienten c_n können direkt aus den LPCs a_k (Gleichung (3.1)) berechnet werden:

$$c_n = a_n + \frac{1}{n} \sum_{k=1}^{\min(P,n-1)} (n-k)c_{n-k}a_k \tag{3.10}$$

Die Koeffizienten werden komplexe Cepstral-Koefficienten genannt, obwohl sie reell sind (es wird aber ein komplexer Logarithmus verwendet). Die reellen Cepstral-Koeffizienten $\hat{c}_n$ werden ähnlich den komplexen berechnet, wobei jedoch nur der Betrag des Spektrums verwendet wird:

$$\hat{c}_n = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \log|X(e^{j\omega})|e^{j\omega n} d\omega \tag{3.11}$$

3.2.3 Line Spectral Frequencies

Die *Line Spectral Frequencies* (LSFs) werden aus dem einem symmetrischen und einem antisymmetrischen Polynom aus $A(z)$ (siehe Gleichung (3.5)) berechnet (Arslan 1999):

$$P(z) = A(z) + z^{-(P+1)}A(z^{-1}) \tag{3.12}$$

$$Q(z) = A(z) - z^{-(P+1)}A(z^{-1}) \tag{3.13}$$

Die Nullstellen dieser Polynome sind die LSFs.

Line Spectral Frequencies besitzen folgende vorteilhafte Eigenschaften:

- Die Prüfung der Stabilität ist einfach. Wenn alle LSFs steigend angeordnet und im Intervall [0,1] enthalten sind, ist das resultierende Filter stabil.
- Interpolation ist möglich.
- Da alle LSFs stark miteinander korreliert sind, können sie effizient quantisiert werden.
- Wenn zwei LSF-Werte nahe beieinander liegen, liegt meist ein spektraler Peak zwischen ihnen, was beim Tracking von Formanten und spektralen Peaks ausgenutzt werden kann.

Der Nachteil von LSFs ist die Notwendigkeit der Berechnung der Nullstellen von $P(z)$ und $Q(z)$. Bei einer hohen Abtastfrequenz muss die LP-Ordnung entsprechend hoch gewählt werden, was die Nullstellen-Berechnung aufwendig macht.

3.2.4 Improved-Power-Spectrum-Envelope-Analyse

Tanaka und Abe (1997) schlagen die *Improved-Power-Spectrum-Envelope*-Analyse (IPSE-Analyse) zur Extraktion der spektralen Einhüllenden eines Signals vor. Die spektrale Einhüllende wird pitch-synchron durch Interpolation der spektralen Peaks mittels einer Cosinus-Funktion berechnet. Der Algorithmus orientiert sich sowohl an den spektralen Peaks als auch an der momentanen Grundfrequenz F_0, da die spektralen Charakteristika von Sprachsignalen (vor allem im unteren Frequenzbereich) auch von der Tonhöhe abhängen.
Während des Trainings werden die Beziehungen zwischen der Grundfrequenz und der spektralen Einhüllenden eines Sprechers bestimmt und in einem Codebook gespeichert.

Die Vorgangsweise ist wie folgt:

- Fensterung eines Verarbeitungsblocks der Länge von 5 Perioden mit einem Hamming-Fenster und Berechung des logarithmisierten Leistungsspektrums
- Die lokalen Maxima des logarithmisierten Leistungsspektrums werden an den Stellen f_n resampelt ($nF_0 - F_0/2 < f_n < nF_0 + F_0/2$, wobei n eine ganze Zahl ist)
- Ist das Intervall zwischen f_n und f_{n+1} größer als das 1,5-fache von F_0 ist, so werden die lokalen Maxima des Leistungsspektrums innerhalb dieses Intervalls zur oben erhaltenen Sequenz hinzugefügt.
- Die Samples werden linear interpoliert und an F_0/n Intervallen abgetastet. n ist die ganze Zahl, die das Maximum für F_0/n ergibt, wobei $F_0/n < 50Hz$.
- Die resampelten Linien werden durch ein Cosinus-Modell approximiert:

$$Y(\lambda) = \sum_{i=0}^{M} A_i \cos(i\lambda) \qquad 0 \leq \lambda \leq \pi \tag{3.14}$$

  Der mittlere quadratische Fehler zwischen dem Modell und den resampelten Punkten wird minimiert.

3.2.5 True-Envelope-Schätzung

Die True-Envelope-Schätzung von Imai und Abe (1979) basiert auf der spektralen Glättung mittels des Cepstrum. Sie wird in den in den Kapiteln 4 und 5 beschriebenem Implementationen verwendet.

Das *reelle Cepstrum* eines diskreten Signals mit dem Spektrum $X[k]$ ist definiert als

$$\hat{X}[l] = \sum_{K}^{k=0} \log |X[k]| \cdot e^{j\frac{2\pi kl}{K}} \tag{3.15}$$

(vgl. Abschnitt 3.2.2). Multipliziert man $\hat{X}[l]$ mit einem Tiefpassfenster $w_{TP}[l]$ gemäß Gleichung (3.16) erhält man nach einer weiteren DFT die Einhüllende des Betragsspektrums (Oppenheim und Schafer 1999).

$$w_{TP}[l] = \begin{cases} 1 & l = 0, P_c \\ 2 & 1 \leq l < P_c \\ 0 & P_c < l \leq N-1 \end{cases} \tag{3.16}$$

P_c ist die Ordnung des Cepstrum (vgl. Abbildung 3.2).

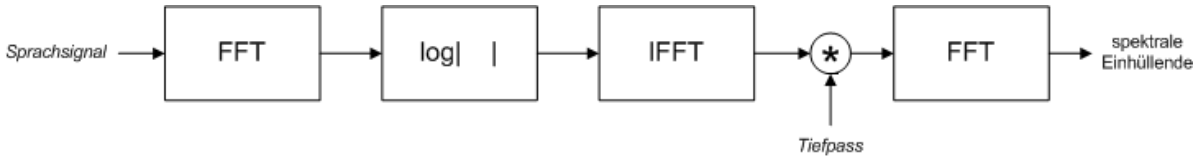


Abb. 3.2: Blockschaltbild zur Berechnung der spektralen Einhüllenden mittels des Cepstrums

Die iterative True-Envelope-Schätzung von Imai und Abe (1979) stellt eine Weiterentwicklung dieses Verfahrens dar, die zu einer Einhüllenden führt, die den Maxima eines Spektrums folgt und nicht dessen Mittelwert, so wie eine gemäß Abbildung 3.2 berechnete Einhüllende (vgl. auch Abbildung 3.3). Die Vorgangsweise ist wie folgt:

$V_i[k]$ sei die (cepstrum-basierende) spektrale Einhüllende bei der Iteration i. Mit den Initialwerten $A_0[k] = \log|X[k]|$ und $V_0[k] = -\infty$ ersetzt der Algorithmus iterativ das logarithmische Betragsspektrum des Eingangssignals gemäß

$$A_i[k] = \max(A_{i-1}[k], V_{i-1}[k]) \tag{3.17}$$

Bei jeder Iteration wird der Cepstralfilter auf das aktualisierte Betragsspektrum A_i angewendet.
Dadurch wächst die Einhüllende von Iteration zu Iteration, und die Täler zwischen den spektralen Peaks werden aufgefüllt. Als Abbruchkriterium wird der Parameter Δ benützt. Dieser definiert den Maximalwert, um den ein spektraler Peak die Einhüllende überragen darf. In den vorliegenden Anwendungen wird $\Delta = 2dB$ gesetzt. Abbildung 3.3 zeigt ein Betragsspektrum mit der cepstralen Einhüllenden und dem True Envelope.

Dieses Verfahren zeigt zwei wichtige Schwächen (Röbel und Rodet 2005):

- Der Rechenaufwand ist groß und nicht von der Ordnung des Ceptrum abhängig, sondern von der Länge der DFT.

- Das auf das Spektrum angewendete Fenster ist im Normalfall rechteckig. Dieses Rechteckfenster glättet das logarithmisierte Spektrum durch eine Faltung mit einer periodischen sinc-Funktion. Diese Glättung erzeugt aber Oszillationen an den Stellen, wo sich das logarithmisierte Spektrum schnell ändert (*Gibbs-Phänomen*).

Diese unerwünschten Oszillationen in der spektralen Einhüllenden lassen sich deutlich reduzieren, wenn ein Hanning-Fenster als Glättungsfilter in der Cepstral-Domäne verwendet wird. Wird die Länge des Hanning-Fensters ungefähr 1,66-mal so groß gewählt wie jene des Rechteckfensters, so wird ein Glättungsrumpf mit ungefähr der gleichen Hauptkeule aber beinahe ohne Nebenkeulen erzeugt.

Reduzierung des Rechenaufwandes

Da es kaum möglich ist, die Form einer spektralen Einhüllenden mit einer Präzision in der Größenordnung der Breite der spektralen Peaks wahrzunehmen, schlagen Röbel und Rodet (2005) vor, das logarithmisierte Spektrum so grob abzutasten, bis ein Bin nur noch die Breite eines spektralen Peaks hat. Die exakte Position der Amplitudensamples ist nicht ausschlaggebend und kann daher einfach quantisiert werden.
Jedoch müssen die Amplituden der spektralen Peaks vorsichtiger behandelt werden, da ja der Maximalwert eines Peaks die ausschlaggebende Information in Bezug auf die spektrale Einhüllende trägt. Deshalb wird das herunter getastete Spektrum $S[m]$ mittels eines Maximumfilters gemäß Gleichung (3.18) berechnet:

$$S[m] = \max_{k=r(m-0,5)}^{r(m+0,5)-1}(\log|X[k]|) \quad \text{mit} \quad r = \frac{K}{M} \tag{3.18}$$

K ist die Länge des Spektrums, M ist die Länge des heruntergetasteten Spektrums $S[m]$. M wird als größte 2-er-Potenz unter der Länge des Analysefensters gewählt.
Die berechnete spektrale Einhüllende wird dann durch lineare Interpolation zur Länge K hoch gesampelt.

Röbel und Rodet (2005) schlagen noch weitere Möglichkeiten zur Reduzierung des Rechenaufwandes vor, welche aber in der vorliegenden Arbeit nicht berücksichtigt werden, da die verwendeten Implementationen nicht echtzeitfähig sein brauchen.

3.2.6 Subband-Verarbeitung mit der DWT

Türk und Arslan (2002) schlagen eine Modifikation des STASC-Systems (vgl. Abschnitt 3.1.1) durch Subband-Verarbeitung vor, die vor allem bei hohen Abtastraten Verbesserungen bringt.
Subband-Verarbeitung hat gegenüber der Breitbandverarbeitung dann folgende Vorteile:

- Um die Target- und Source-Spektren bei hohen Abtastraten im Detail zu modellieren muss die Anzahl der LSFs entsprechend erhöht werden. Die Genauigkeit der Algorithmen zur Bestimmung der Nullstellen lässt aber bei Polynomen höherer Ordnung nach.

- Die Interpolation der LSFs bei Breitbandverarbeitung kann zu Formantverschiebungen führen. LSFs sind reine Frequenzgrößen. Je größer der Bereich ist, über

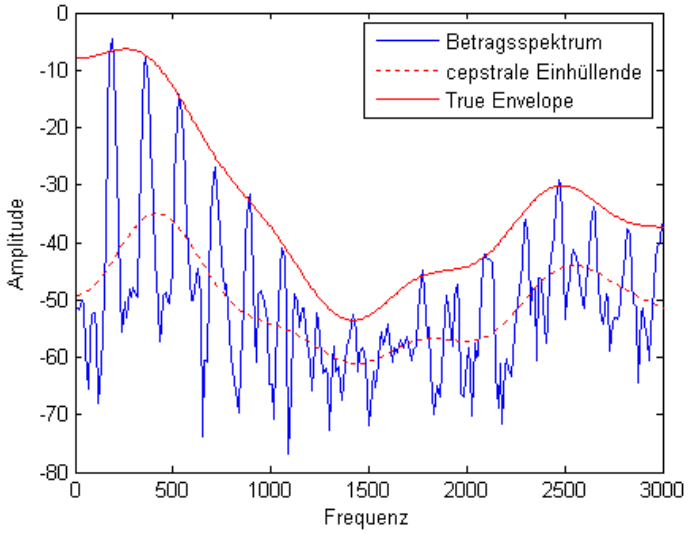


Abb. 3.3: Betragsspektrum mit cepstraler Einhüllender und True Envelope

den sie reichen, desto mehr resultiert die lineare Interpolation in Verschiebungen von LSF-Paaren.

- In höheren Frequenzbereichen finden sich Signalkomponenten die nicht unmittelbar zur Sprache gehören. Werden diese - wie bei der Breitbandverarbeitung - wie Sprache modelliert und modifiziert kann dies zu Artefakten führen.

- Das Training ist in Bezug auf den Rechenaufwand der aufwendigste Teil der Voice Conversion, vor allem bei Verwendung von umfangreichen Trainingsdaten. Dieser Rechenaufwand ist bei Subband-Verarbeitung geringer, da geringere Prädiktionsordungen verwendet werden können. Außerdem kann die Abtastrate verringert werden (siehe unten)

Die Subband-Dekomposition kann effizient mittels der *diskreten Wavelet-Transformation* (DWT) implementiert werden. Die folgenden Eigenschaften machen die DWT zu einem attraktiven Werkzeug zum Design von Filterbänken:

- Bei der Verwendung geeigneter Filter ist eine perfekte Rekonstruktion des Signals gewährleistet.

- Es können FIR-Filter benutzt werden. Diese sind garantiert stabil und haben einen linearen Phasengang.

- Die Subband-Dekomposition/-Rekonstruktion kann vollständig im Zeitbereich durchgeführt werden.

Diskrete Wavelet-Transformation

Die kontinuierliche Wavelet-Transformation $T_f(b,a)$ eines Signals $f(t)$ ist als

$$T_f(b,a) = \frac{1}{a}\int f(t)\varphi(\frac{t-b}{a})dt \tag{3.19}$$

definiert (Torrésani 1999). $\varphi(\frac{t-b}{a})$ ist eine skalierte und zeitlich verschobene Kopie des Mutter-Wavelet $\varphi(t)$. Die diskrete Wavelet-Transformation (DWT) erhält man durch digitales Filtern und herunterabtasten des Signals $x[n]$:

$$y_{high}[k] = \sum_n x[n]g[2k-n] \tag{3.20}$$

$$y_{low}[k] = \sum_n x[n]h[2k-n] \tag{3.21}$$

$g[n]$ ist ein Hochpass-, $h[n]$ ein Tiefpassfilter. Die Subband-Dekomposition erhält man durch eine entsprechende Kaskade dieser Operationen.
Das Signal kann mittels inverser diskreter Wavelet-Transformation wieder aus seinen Subbandkomponenten $y_{low}[k]$ und $y_{high}[k]$ gemäß Gleichung (3.22) rekonstruiert werden:

$$\hat{x}[n] = \sum_{k=-\infty}^{\infty} (y_{high}[k]g[-n+2k]) + (y_{low}[k]g[-n+2k]) \tag{3.22}$$

Training

Während der Trainingsphase werden für jedes Frequenzband separate Codebooks erstellt. Türk und Arslan (2002) benutzten vier Subbänder mit einer Bandbreite von jeweils 5,5 kHz bei einer Abtastrate von 44,1 kHz.
Aufgrund der physiologischen Eigenschaften des Sprachapparates sind annähernd alle für die Erstellung der Codebooks notwendigen Sprachkomponenten im untersten Subband enthalten. Dadurch kann die Abtastrate in der Trainingsphase auf 11,025 kHz reduziert und so der Prozess erheblich beschleunigt werden. Obwohl die Subband-Dekomposition zusätzlichen Rechenaufwand bedeutet ist das Training in Subbändern schneller, da die Dekomposition sehr effektiv im Zeitbereich mittels FIR-Filtern implementiert werden kann.

Transformation

Die tieferen Subbänder werden mittels der Codebook-Einträge transformiert. Türk und Arslan (2002) stellten fest, dass die Transformation höherer Frequenzbänder wenig zur Qualität der Transformation beiträgt, dafür aber zusätzliche Artefakte verursacht.

Die transformierten Subbänder werden mittels der Rekonstruktionsfilterbank generiert, die nicht transformierten Bänder werden dem Ausgangssignal hinzuaddiert.

Prosodische Modifikationen inklusive Tonhöhenskalierung etc. werden dann auf das Breitbandausgangssignal angewendet.

Jedoch bringt das DWT-basierende System folgende Nachteile mit sich:

- Die Rekonstruktionsfilterbank liefert keine perfekte Rekonstruktion, da jedes Subband eine modifizierte Version des Originalsignals ist.
- Spektrale Peaks die an den Grenzen der Subbänder auftreten werden nicht angemessen modelliert und transformiert.

3.2.7 Selektive Vorverstärkung

Um die oben genannten Nachteile des DWT-basierenden Systems zu umgehen schlägt Türk (2003) die selektive Vorverstärkung vor. Diese erlaubt es, verschiedene Frequenzbereiche mit unterschiedlicher Genauigkeit zu modellieren.
Wie bereits in Abschnitt 3.2.3 erwähnt, kumulieren sich LSFs um spektrale Peaks herum. Das führt dazu, dass der Durchlassbereich eines Bandpassfilters bei gleicher LP-Ordnung genauer modelliert wird als der betreffende spektrale Bereich im Breitbandsignal.
Eine einfache Erhöhung der LP-Ordnung führt zwar auch zur Erhöhnung der Genauigkeit der Modellierung eines Breitbandsignals, jedoch ist die LP-Ordnung durch die Nullstellenberechnung begrenzt (vgl. Abschnitt 3.2.3).
Wird nun ein Signal in einzelne Frequenzbänder aufgeteilt, so kann die Genauigkeit der Modellierung des Spektrums auch bei niedriger LP-Ordnung gesteuert werden.

Durch entsprechende Überlappung der Frequenzbänder kann das erwähnte Problem der Modellierung von spektralen Peaks an den Rändern der Subbänder bei der Transformation mittels der DWT verhindert werden.

3.3 Modellierung der Anregung

3.3.1 Impuls-/Rausch-Modell

Das einfachste Modell für das Anregungssignal unterscheidet nur zwischen stimmhaften und stimmlosen Segmenten (Oppenheim und Schafer 1999). In stimmhaften Segmenten wird das Anregungssignal durch eine Folge von Impulsen modelliert. Der Abstand der Impulse entspricht einer Periodendauer der momentanen Grundfrequenz.
In stimmlosen Segmenten kann die Anregung durch Rauschen approximiert werden.

3.3.2 Multiband-Anregungs-Modell

Beim Multiband-Anregungs-Modell wird die Entscheidung stimmhaft/stimmlos in verschiedenen Subbändern getroffen (Griffin und Lim 1988).

Die Subbänder sind um die ganzzahligen Vielfachen der Grundfrequenz zentriert. Diese Approximation ist besser als das einfache Impuls-/Rausch-Modell, da die höheren Frequenzbereiche natürlicher Sprachsignale auch in stimmhaften Segmenten rauschähnliche Komponenten besitzen, welche durch das Impuls-/Rausch-Modell nicht gut beschrieben werden.

Der Vorteil des Multiband-Anregungs-Modells liegt in der detaillierten Repräsentation des Anregungssignals. Der Nachteil ist die Erfordernis einer robusten Pitch-Detektierung und robusten stimmhaft/stimmlos Entscheidungen.

3.3.3 Sinus-Modell

Beim Sinus-Modell wird das Anregungssignal als eine Linearkombination von Sinusschwingungen beliebiger Frequenz ω_l, Amplitude A_l und Phase modelliert (McAuley und Quatieri 1986):

$$e(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \phi_l) \tag{3.23}$$

L ist die Anzahl der Sinuskomponenten, ϕ_l ist ein Phasen-Offset, der die Phasenbeziehungen zwischen den einzelnen Komponenten beschreibt.

3.4 Modifikation der Anregung

Im Normalfall müssen sowohl die Tonhöhe als auch die prosodischen Charakteristika (z.B. die Dauer einzelner Phoneme) des Eingangssignals manipuliert werden. Die zwei am häufigsten dafür eingesetzten Verfahren werden im Folgenden beschrieben.

3.4.1 Time Domain Pitch Synchronous Overlap-Add (TD-PSOLA)

Beim *Time-Domain-Pitch-Synchronous-Overlap-Add*(TD-PSOLA)-Verfahren wird das Eingangssignal pitch-synchron in Segmente aufgeteilt (Moulines und Charpentier 1990). Diese Segmente werden in einer geeigneten Weise zusammengesetzt, um die gewünschten Modifikationen zu erhalten. Die Vorgangsweise ist wie folgt:

- Mittels eines geeigneten Verfahrens (z.B. Gold und Rabiner 1969) müssen die Perioden der Grundfrequenz exakt markiert werden.
- Pitch-synchron werden Segmente der Länge von zwei bis vier Perioden des Eingangssignals extrahiert.

- Zeit- bzw. Tonhöhenmodifikationen werden wie unten beschrieben durchgeführt und das Ausgangssignal wird mittels Overlap-Add-Synthese (siehe unten) konstruiert.

Modifikation der Zeitachse

Um Modifikationen der Zeitachse zu erhalten werden beim TD-PSOLA ganzzahlige Anzahlen von Segmenten wiederholt oder weggelassen. Die Wiederholung von Segmenten streckt, das Weglassen von Segmenten staucht das Signal im Zeitbereich.

Modifikation der Tonhöhe

Zur Tonhöhenmodifikation verändert TD-PSOLA das Ausmaß des Überlappens zweier aufeinander folgender Segmente. Um die Tonhöhe zu erhöhen werden die Segmente näher aneinandergerückt; um die Tonhöhe zu vermindern werden sie auseinandergezogen. Die dabei entstehende zeitliche Stauchung oder Streckung des Signals muss mittels einer geeigneten Zeitmodifikation (siehe oben) kompensiert werden.

Overlap-Add-Synthese

Nachdem Position und Überlappung der Segmente des Ausgangssignals bestimmt wurden werden die einzelnen Segmente Hamming-gefenstert (zur Vermeidung von Diskontinuitäten) und dann mit der entsprechenden Überlappung addiert.
Vorteil der Overlap-Add-Synthese (OLA-Synthese) ist ihre Einfachheit und Effizienz. Aber je größer die vorgenommene Skalierung ist, desto schlechter wird das Ergebnis. Darüber hinaus sind folgende Nachteile zu erwähnen:

- Zeitskalierungen können nur quantisiert mit einer Auflösung von einer Periodendauer durchgeführt werden.
- Bei Zeitexpansion kann das Wiederholen von Segmenten (vor allem von nichtperiodischen Segmenten) hörbare Artefakte verursachen.
- Bei Sprache besteht ein Zusammenhang zwischen der spektralen Einhüllenden und der Tonhöhe (vgl. Abschnitt 3.2.4). Für sehr große und sehr kleine Tonhöhenskalierungsfaktoren wir dieser Umstand hörbar.

3.4.2 Frequency Domain Pitch Synchronous Overlap-Add (FD-PSOLA)

Das *Frequency Domain Pitch Synchronous Overlap-Add*(FD-PSOLA)-Verfahren ist das Äquivalent zum TD-PSOLA im Frequenzbereich (Moulines und Charpentier 1990). Es kann auch ohne Markierungen der Perioden des Signals verwendet werden. Allerdings erfordert es eine robuste Tonhöhendetektion.

- Das Kurzzeitspektrum eines Eingangssignals wird mit einer Fensterlänge von 2-5 Perioden der Grundfrequenz berechnet.
- Die spektrale Einhüllende des Verarbeitungsblocks wird geschätzt. Mittels inverser Filterung wird das Spektrum des Anregungssignals berechnet.
- Die Tonhöhe des Anregungssignals wird verändert. Dazu können verschiedene Verfahren verwendet werden. Stellvertretend werden unten zwei Methoden beschrieben.
- Das modifizierte Anregungssignal wird mit der spektralen Einhüllenden des Eingangssignals gefiltert und ergibt so das Ausgangssignal.

Spektrale Kompression/Expansion

Bei der spektralen Kompression/Expansion wird die Frequenzachse gemäß eines Skalierungsfaktors β linear gestaucht oder gestreckt. Die DFT-Koeffizienten werden gemäß Gleichungen (3.24) und (3.24) linear interpoliert.

$$Y(k_s) = (1 - \alpha)X(k_v) + \alpha X(k_v + 1) \tag{3.24}$$

k_v erhält man durch abschneiden der Nachkommastellen von k/β. Das Gewicht α erhält man aus Gleichung (3.25).

$$\alpha = k_s - \frac{k}{\beta} \tag{3.25}$$

Hier muss die spektrale Einhüllende exakt bestimmt werden, da die Komponenten des Anregungsspektrums in andere Frequenzbereiche transformiert werden. Sind Unregelmäßigkeiten im Anregungsspektrum vorhanden werden diese ebenfalls verschoben und verursachen Artefakte.

Wird die Frequenzachse gestaucht, so entsteht eine leere Region im oberen Bereich des Spektrums. In diesen muss entweder ein tiefer Teil des Spektrums kopiert oder ein hoher Teil umgeklappt werden.

Wiederholung/Eliminierung von Teiltönen

Die Tonhöhe kann auch durch Einsetzen bzw. Eliminieren von Teiltönen verändert werden. Um die Tonhöhe zu verringern werden die Teiltöne näher zusammengerückt, und durch Wiederholung von vorhandenen Teiltönen werden neue hinzugefügt. Um die Tönhöhe zu erhöhen wird der Abstand der Teiltöne vergrößert, und Teiltöne werden gelöscht.
Dieser Ansatz erfordert eine präzise Bestimmung der Grundfrequenz.

Synthese

Die Synthese erfolgt wie beim TD-PSOLA als OLA-Synthese (Abschitt 3.4.1), wobei aber die Überlappung und die Position der Syntheseblöcke durch die Analyse festgelegt sind.

3.5 Transformation der Anregung

Neben der oben besprochenen Manipulation von prosodischen Charakteristika und der Grundfrequenz können weitere Merkmale des Anregungssignals transformiert werden. Im Folgenden werden zwei Ansätze dazu beschrieben.

3.5.1 Transformation der Anregung beim STASC

Beim *Speaker Transformation Algorithm Using Segmental Codebooks* (STASC, Arslan 1999) werden die LP-Residualsignale während des Trainings gesammelt und entsprechend der Segmente (z.B. Phoneme) in denen sie auftreten kategorisiert. Sowohl für den Source- als auch für den Target-Sprecher werden die mittleren Kurzzeit-Betragsspektren für jedes Segment bestimmt. Anhand dieser Daten kann dann für jedes Segment eine Transformationsfunktion aufgestellt werden. Dadurch werden nicht nur die allgemeinen Charakteristika des Anregungssignals transformiert, sondern auch die Nullstellen des Vokaltraktfilters, die durch die LP-Analyse nicht optimal beschrieben werden.

3.5.2 High Resolution Voice Conversion

Kain (2001) geht noch einen Schritt weiter und transformiert zusätzlich zum Betragsspektrum des Anregungssignals auch das Phasenspektrum. Das vorgeschlagene System modelliert das Anregungssignal sorgfältiger als die herkömmliche LP-Konvertierung da es ihm spektrale Details hinzufügt. Kain (2001) nennt deshalb seinen Ansatz *High Resolution Voice Transformation*.
Die hier zu Grunde liegende Annahme ist wie in Abschnitt 3.5.1, dass die Residualsignale für einen Sprecher innerhalb einer phonetischen Klasse ähnlich sind, und so also ein Zusammenhang zwischen der spektralen Einhüllenden (die ja auch innerhalb einer phonetischen Klasse ähnlich ist) und dem Residualsignal besteht.

Training

Mittels eines Klassifikators, der die Zugehörigkeit eines LP-Parameter-Sets zu einer phonetischen Klasse beschreibt, wird ein Codebook für das Residualsignal angelegt, wobei Betrags- und Phasenspektrum getrennt gespeichert werden.
Beim Betragsspektrum wird der Codebook-Eintrag als normalisierte und gewichtete

Summe aller Residual-Betragsspektren berechnet, wobei die Gewichte den Grad der Zugehörigkeit des Segmentes zu einer Klasse beschreiben.
Für die Phase muss ein anderer Ansatz gewählt werden, da sie in Werten modulo 2 π vorliegt, und so die Summation zu einem unbrauchbaren Ergebnis führen würde. Selbst die Verwendung der entwickelten (*unwrapped*) Phasen liefert nur für tiefe Frequenzen verlässliche Ergebnisse, da sich die Phasen nur dort langsam mit der Frequenz verändern.
Deshalb wählt Kain (2001) den Phasenvektor des Zentroides jeder Klasse als Codebook-Eintrag. Der Nachteil dabei ist, dass ein Residual-Phasenspektrum ausgewählt werden kann, dass nicht repräsentativ für die Trends in einer Klasse ist.

Synthese

Das Betragsspektrum des Residualsignals wird aus einer gewichteten Summe der Codebook-Einträge berechnet. Die Gewichtung orientiert sich wiederum am Grad der Zugehörigkeit des momentanen Phonems zu einer Klasse. Als Phasenspektrum wird der wahrscheinlichste Codebook-Eintrag ausgewählt.

3.6 Modellierung und Transformation der F_0-Kontur

Die Kontur der Grundfrequenz enthält weitere Charakteristika eines Sprechers. Auch hier existiert eine Vielzahl an unterschiedlichen Ansätzen zur Modellierung und Transformation. Einige Beispiele werden stellvertretend im Folgenden beschrieben.

3.6.1 Mittelwert-/Varianz-Modell

Das Mittelwert-/Varianz-Modell beruht auf der Annahme, dass die Werte der momentanen Grundfrequenz eines Sprechers aus einer einzigen Gaußverteilung stammen. Die stimmhaften Segmente eines Signals werden benutzt, um die Parameter (Mittelwert und Standardabweichung) der Verteilung zu schätzen.

Aus diesen Parametern wird dann ein variabler Tonhöhenskalierungsfaktor $\beta(t)$ bestimmt:

$$\beta(t) = \frac{a f_0^s(t) + b}{f_0^s(t)}, \qquad a = \frac{\sigma_t}{\sigma_s}, \qquad b = \mu_t - \mu_s \frac{\sigma_t}{\sigma_s} \tag{3.26}$$

$f_0^s(t)$ ist die momentane Source-F_0, μ_s und μ_t sind die Source- und Target-F_0-Mittelwerte, σ_s und σ_t sind die Standardabweichungen von Source- und Target-F_0.

3.6.2 Satz-Codebooks

Wie für die anderen Parameter ist es auch möglich Codebooks für die F_0-Kontur zu generieren (Chappel und Hansen 1998). Der Vorteil dabei ist, dass reale F_0-Konturen bei der Synthese benutzt werden. Der Nachteil ist, dass die Anzahl der Codebook-Einträge die Anzahl der möglichen Synthesekonturen einschränkt.
Dieser Ansatz liefert gute Ergebnisse jedoch nur bei einem beschränkten Vokabular oder speziellen Anwendungen, wo die Variabilität der Konturen eingeschränkt ist.

3.6.3 Fujisakis Modell

Fujisakis Modell (Fujisaki und Kawai 1982) beruht auf der Annahme, dass eine F_0-(Intonations-)Kontur aus zwei Komponenten besteht: der Phrase und dem Akzent. Der Prozess der diese Konturen generiert wird durch einen glottalen Oszillationsmechanismus repräsentiert, der Phrasen- und Akzentinformationen als Eingangsignal benutzt und daraus kontinuierliche F_0-Konturen generiert.
Das Eingangssignal des Mechanismus sind entweder Impulse - zur Modellierung der Phrasen - und Stufenfunktionen - zur Modellierung der Akzente.
Der Mechanismus besteht aus zwei kritisch abgestimmten Filtern zweiter Ordnung. Das erste Filter modelliert die Phrasen, das zweite Filter die Akzente.
Phrasen und Akzente können beliebig aneinandergehängt werden, um eine komplette Pitch-Kontur zu erhalten.
Mit den entsprechenden Parametern wird dann eine passende Kontur für das Ausgangssignal der VC-Systems generiert.

3.6.4 Segmentales F_0-Kontour-Modell

Türk (2003) schlägt ein segmentales Modell der Kontour der Grundfrequenz F_0 vor, um die Charakteristika der Intonation eines Sprechers detaillierter zu modellieren und zu konvertieren als es der statistische Ansatz aus Abschnitt 3.6.1 erlaubt. Dabei kategorisiert er die Segmente der F_0-Kontur des Source- und des Target-Sprechers und verwendet diese Information bei der Transformation.

Dies geschieht wie folgt:

Das Trainingsmaterial von Source- und Target-Sprecher wird phonetisch angeglichen und jedem stimmhaften Segment des Source-Sprechers wird das korrespondierende Segment des Target-Sprechers zugeordnet.
Bei der Transformation wird mittels eines Distanzmaßes der Unterschied jedes Eingangssegmentes zu jedem der Codebook-Einträge berechnet. Daraus werden dann entsprechende Gewichtungen berechnet und die Synthese-F_0-Kontur als gewichteter Mittelwert einer durch einen Parameter einstellbaren Anzahl an Codebook-Einträgen

zusammengesetzt.

Diese Methode beschreibt die generelle F_0-Kontur eines Sprechers sehr gut. Plötzliche Sprünge der Tonhöhe werden jedoch nicht adäquat modelliert.

3.7 Evaluierung der Voice Conversion

In der Literatur ist eine große Anzahl an objektiven und subjektiven Methoden zur Evaluierung der Performance eines VC-Algorithmus zu finden. Da der Mensch letztendlich der Empfänger eines CV-Signals ist, ist die perzeptive Evaluierung aber das Maß der Dinge.

3.7.1 Objektive Evaluierung

Ein weit verbreitetes Fehlermaß in der Sprachsignalverarbeitung ist die *Spectral Distortion* (SD) zwischen zwei Signalen. Diese ist definiert als

$$\mathrm{SD}(X_1, X_2) = \frac{1}{M} \sum_{m=1}^{M} \ldots \sqrt{\frac{1}{K} \sum_{k=0}^{K-1} \left(20 \cdot \log_{10} |X_1^m(e^{j2\pi \frac{k}{K}})| - 20 \cdot \log_{10} |X_2^m(e^{j2\pi \frac{k}{K}})| \right)^2} \quad . \quad (3.27)$$

X_1 und X_2 sind die Spektren den beiden zu untersuchenden Signale. M ist die Anzahl der betrachteten Verarbeitungsblocks.
Zur Beurteilung der Performance von VC-Systemen wird die SD zwischen Source- und Target-Signal sowie zwischen transformiertem Signal und dem Target-Signal berechnet (z.B. Arslan 1999, Abe et al. 1988). Ist die SD zwischen transformiertem Signal und Target geringer als zwischen Source und Target, so ist dies ein Indiz für eine erfolgreiche Konvertierung.
Eine andere Möglichkeit besteht darin, das transformierte Signal als Eingangsignal zu einem Sprecheridentifikationssystem zu benutzen und die Wahrscheinlichkeiten der Identifikation der involvierten Sprachsignale zu bestimmen (z.B. Arslan 1999).

3.7.2 Subjektive Evaluierung

Bei der subjektiven Evaluierung sind vor allem drei Größen von Interesse: Sprecheridentifikation, Natürlichkeit und Verständlichkeit.

Ein weit verbreiteter Ansatz zum Testen der Sprecheridentität ist der ABX-Test (z.B. Arslan 1999, Abe et al. 1988). Die Probanden hören dabei die Stimuli A, B und

X und müssen sich entscheiden, ob entweder A oder B näher an X liegt. X ist dabei meist das transformierte Signal, A und B sind der Target- und der Source-Sprecher.
Allerdings ist hier anzumerken, dass eine Trefferquote von 100% für das Target nur bedeutet, dass das transformierte Signal näher am Target- als am Source-Sprecher liegt. Ob das transformierte Signal vom Target-Sprecher ununterscheidbar ist kann mit einem ABX-Test nicht bestimmt werden.
Eine Verbesserung des ABX-Test stellt der Paarvergleich dar. Die Probanden hören ein Stimuluspaar und beurteilen die Ähnlichkeit der Sprecher auf einer Skala (z.B. Abe et al. 1988).

Zum Testen der Natürlichkeit und der Verständlichkeit existiert eine Reihe genormter Verfahren, die vor allem im Telekommunikationsbereich weit verbreitet sind (z.B. ITU 1996).

Kapitel 4

Voice Transformation mittels der Spectral-Modeling-Synthese

4.1 Einleitung

Die Abschnitte 2.3 und 2.4 erläutern die Parameter von Sprachsignalen, die manipuliert werden müssen, um Transformationen zwischen verschiedenen Gruppen von Sprechern wie z.B. Frauen, Männern etc. durchführen zu können. Dieses Kapitel beschreibt ein im Rahmen dieser Diplomarbeit in der Software MATLAB implementiertes *Spectral-Modeling-Synthesis*(SMS)-System (Serra 1997, Amatrain et al. 2002), das die notwendigen Manipulationen ermöglicht. Die SMS beruht nicht auf dem in Abschnitt 2.2.1 vorgestellten Quelle-/Filter-Modell des Sprachsignals, sondern auf einem *Sinusoidal Plus Residual Model* (SPRM).
Das SPRM ist eine Erweiterung des in Abschnitt 3.3.3 vorgestellten Sinus-Modells. Es bietet die Möglichkeit zu entscheiden, welcher Teil des Spektrums eines Signals durch Sinusschwingungen modelliert und welcher dem übrigen Fourier-Spektrum zugeordnet wird. Das Sinusmodell wird dabei nur für die Teiltöne, also den deterministischen Teil eines Klangs verwendet. Der Residualteil beschreibt alles Übrige, was im Idealfall ein stochastisches Signal darstellt.

Ein Signal $s[n]$ wird durch

$$s[n] = \sum_{r=1}^{R} A_r[n] \cos(\phi_r[n]) + e[n] \tag{4.1}$$

beschrieben. $A_r[n]$ bezeichnet die momentane Amplitude und $\phi_r[n]$ die momentane Phase der r-ten Sinuskomponente und $e[n]$ die Residualkomponente zum Zeitpunkt [n]. R ist die Gesamtzahl der beschriebenen Teiltöne.

4.2 Analyse

Die generelle Analyse-/Synthese-Struktur ist in Abbildung 4.1 als Blockschaltbild dargestellt.

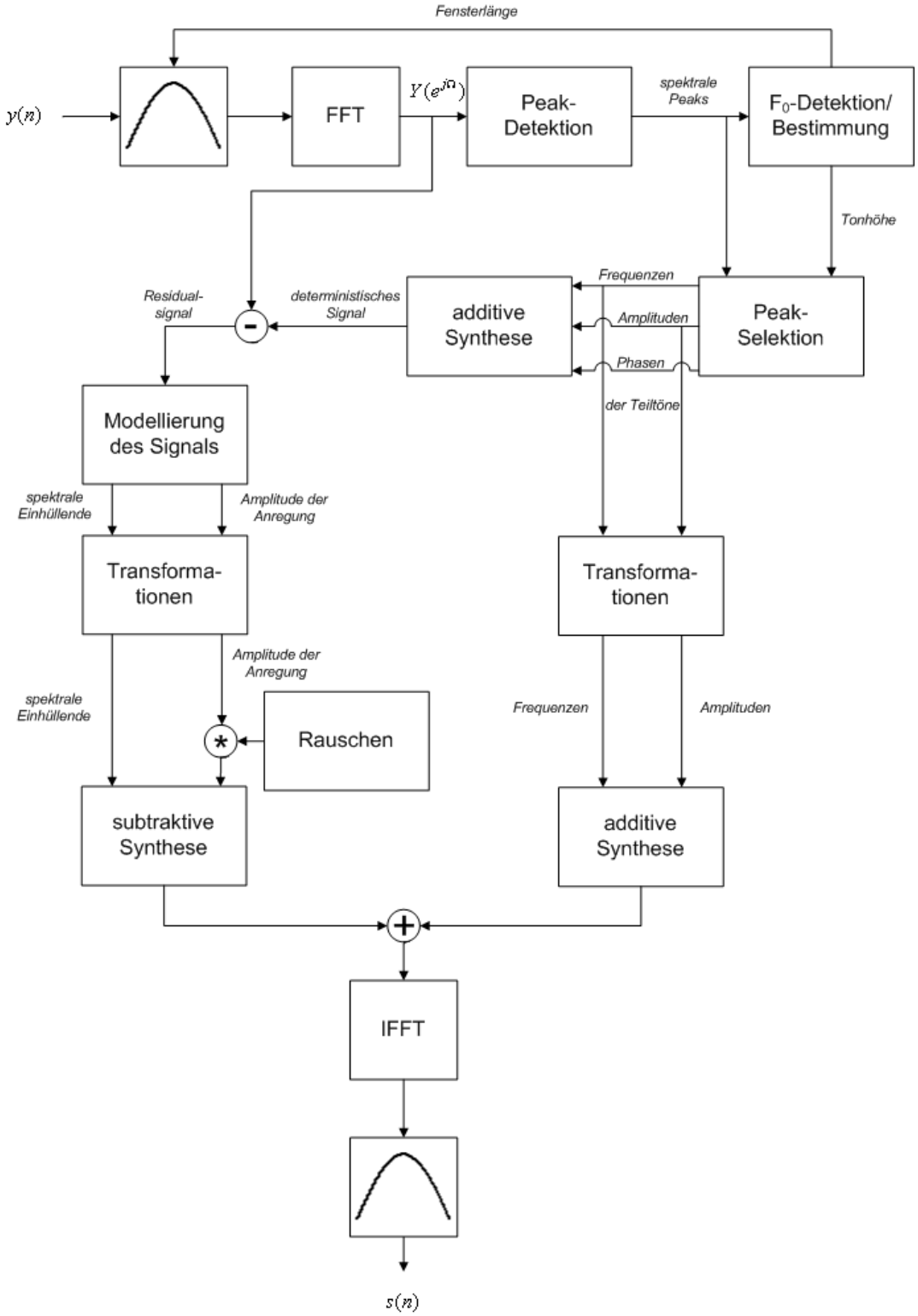


Abb. 4.1: Blockschaltbild der Analyse-/Synthesestruktur des verwendeten Systems

Zur Analyse des Sprachsignals wird in der vorliegenden Anwendung ein Blackman-Harris-92dB-Fenster benutzt (Gleichung (4.2)). Dieses besitzt zwar eine verhältnismäßig breite Hauptkeule von 9 Bins, aber dafür ein sehr hohes Verhältnis der Höhen der Hauptkeule zu den Nebenkeulen von 92dB (Harris 1978). Dieses Verhältnis entspricht beinahe dem Dynamikumfang des verwendeten 16-bit Systems (96dB), sodass nur der Einfluss der Hauptkeule auf das Signal berücksichtigt zu werden braucht, was vor allem die Resynthese vereinfacht (siehe Abschnitt 4.3.1, bzw. Abb. 4.2).

$$\begin{aligned} w_{BH92}(n) &= 0.35875 - 0.48829 \cdot \cos\frac{2\pi n}{N} + \\ &\quad +0.14128 \cdot \cos\frac{4\pi n}{N} - 0.01168 \cdot \cos\frac{6\pi n}{N} \end{aligned} \tag{4.2}$$

N ist die Länge des Fensters, n der Sample-Index.

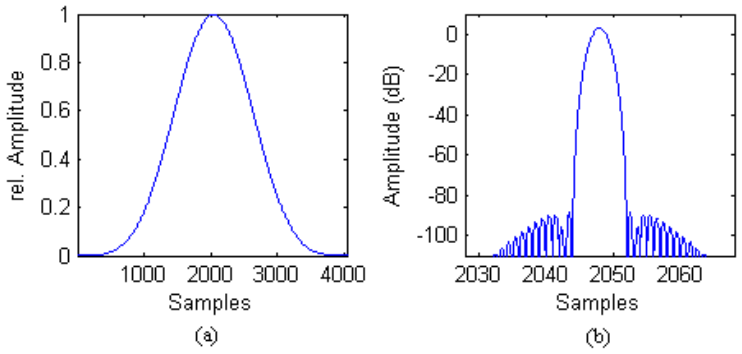


Abb. 4.2: Blackman-Harris-92dB-Fenster im Zeitbereich (a) und dessen Hauptkeule im Detail (b)

Ein weiterer wichtiger Schritt vor der FFT-Analyse ist das *Buffer Centering*. Normalerweise liegt der Zeitursprung am linken Rand eines Fensters. Durch das Buffer Centering wird er in die Mitte verschoben, was einen in der Mitte des Analysefensters liegenden Puls null-phasig macht. Somit vermeidet man den linearen Term, der sonst der Phase des Signals überlagert wäre (*"Zero-phase-windowing"*).
Das Buffer Centering wird durch eine zirkulare Verschiebung des Signals im Zeitbereich erreicht, das dann eine Kommutation der ersten und zweiten Hälfte des Analyse-Buffers ist (vgl. Abbildung 4.3).

4.2.1 Sinuskomponente

Um nun die Sinuskomponente vom Signal trennen zu können müssen die stabilen Teiltöne im Spektrum detektiert werden.
Eine praktikable Vorgehensweise zur Detektierung der Teiltöne eines Signals ist die Detektierung aller Maxima seines momentanen Betragsspektrums. Mittels des *Two-Way-Mismatch*(TWM)-Verfahrens von Maher und Beauchamp (1994) kann dann

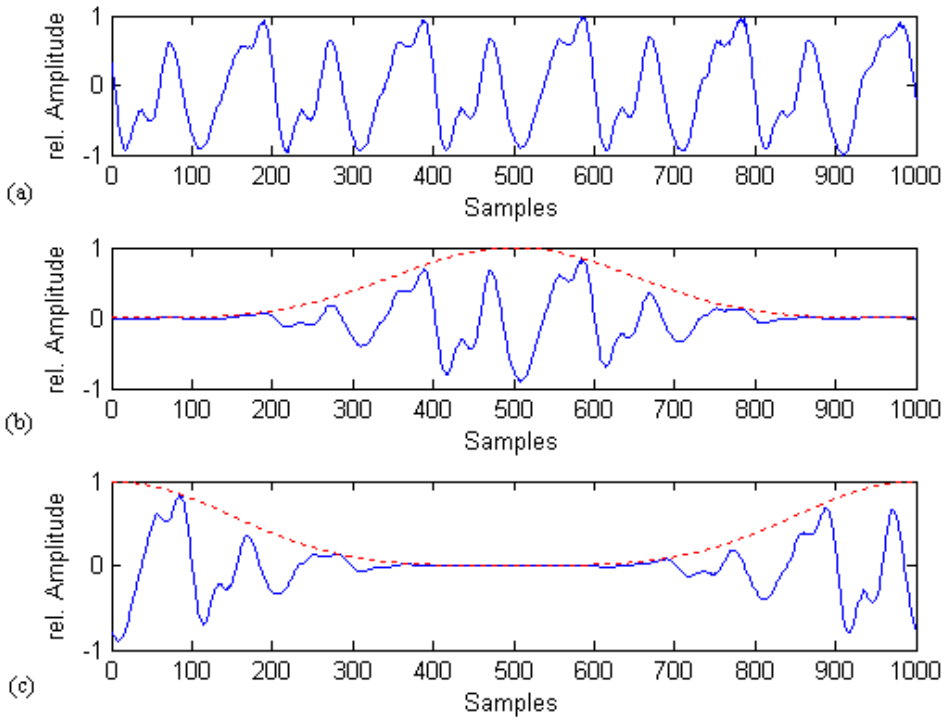


Abb. 4.3: Fensterung eines Analyseblocks und Buffer Centering

die momentane Grundfrequenz des Signals bestimmt werden, indem aus einem Pool potenzieller Kandidaten derjenige ausgewählt wird, dessen generierte Teiltöne am ehesten mit den detektierten Maxima im Spektrum des Signals übereinstimmen. Mit der Grundfrequenz können nun die Maxima des Spektrums ausgewählt werden, die zu Teiltönen gehören.
Der Algorithmus gibt darüber hinaus auch an, ob der momentane Verarbeitungsblock überhaupt ein periodisches Signal enthält oder nicht. Wird ein Block als nicht-harmonisch eingestuft, so wird ihm keine Grundfrequenz zugeordnet und das gesamte Signal als Residualsignal behandelt.

Peak-Detektion

Zur Detektion der Peaks des Betragsspektrums wird dieses zunächst differenziert. An den Stellen, wo im differenzierten Betragsspektrum ein Vorzeichenwechsel von "+" nach "−" stattfindet befindet sich ein Peak.
Die Frequenzauflösung der FFT ist auf Grund der Analyseblocklänge begrenzt. Bei einer Blocklänge von beispielsweise 512 Samples und der Samplingfrequenz von 22,05 kHz liegt die Auflösung bei $\frac{22050}{512} \approx 43Hz$. Mittels *Zero Padding* kann sie noch verringert werden, sodass die Genauigkeit der Peak-Detektion höher wird. Da jedoch das Zero Padding einen entsprechend hohen Rechenaufwand bedeutet wird es in der vorliegenden Anwendung nur soweit angewendet, bis eine quadratische Interpolation des Spektrums möglich ist, die nur Samples aus der unmittelbaren Nachbarschaft des Maximums

benutzt (Abbildung 4.4). Damit können die Frequenzen der Peaks mit für die weitere Analyse/Synthese ausreichender Genauigkeit bestimmt werden.

Bei der quadratischen Interpolation wird das Betragsspektrum um einen Peak herum durch eine Parabel approximiert (Smith und Serra 1987):

$$y(x) = a(x-p)^2 + b \tag{4.3}$$

p ist die Abszisse und b die Ordinate am Scheitelpunkt der Parabel, a ist ein Maß für deren Konkavität. Das Bezugssystem ist um den Punkt $(i, 0)$ zentriert, wobei i der Bin ist, wo der Peak im Betragsspektrum auftritt. Es hat sich gezeigt, dass der Fehler der Approximation im logarithmischen Maßstab geringer ist als im linearen (Smith und Serra 1987).
Die drei in Betracht gezogenen Punkte des Spektrums sind:

$$\begin{aligned}\alpha &= 20 \cdot \log_{10} |X(i-1)| \\ \beta &= 20 \cdot \log_{10} |X(i)| \\ \gamma &= 20 \cdot \log_{10} |X(i+1)| \quad ,\end{aligned}$$

wobei gilt:

$$\begin{aligned}y(-1) &= \alpha \\ y(0) &= \beta \\ y(1) &= \gamma\end{aligned}$$

Die Frequenz f_{max} in Bins am Scheitelpunkt der Parabel ist

$$f_{max} = i + p \quad \text{mit} \quad p = \frac{1}{2}\frac{\alpha - \gamma}{\alpha - 2\beta + \gamma} \quad . \tag{4.4}$$

Die Amplitude am Scheitelpunkt ist

$$y(p) = \beta - \frac{1}{4}(\alpha - \gamma) \cdot p \quad . \tag{4.5}$$

Die Phase des Peaks erhält man durch Interpolation des Phasenspektrums.

F_0-Detektion und Bestimmung

Die Grundfrequenz sei hier als der gemeinsame Teiler der harmonischen Reihe definiert, die die Teiltöne des momentanen Spektrums am besten beschreibt.
Um eben diesen Teiler für ein gegebenes Spektrum zu finden werden potentielle Kandidaten für die Grundfrequenz definiert. Dazu werden von jedem der drei höchsten Peaks im momentanen Spektrum die ersten zehn durch ganzzahlige Division der Frequenz des Peaks entstehenden Teiler ausgewählt.
Für jeden dieser 30 Kandidaten wird dann bestimmt, wie gut eine auf ihm aufgebaute

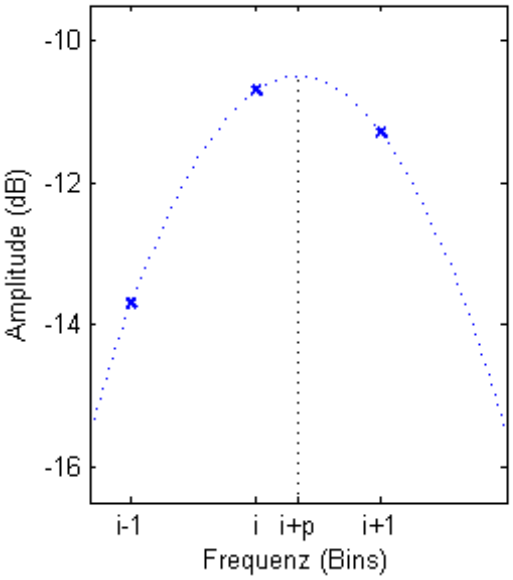


Abb. 4.4: Interpolation der Bins des Amplitudenspektrums zur Erhöhung der Genauigkeit der Peak-Detektion

harmonische Reihe die tatsächlichen spektralen Peaks beschreibt. Das von Maher und Beauchamp (1994) dafür vorgeschlagene Maß basiert auf den gewichteten Differenzen zwischen den Frequenzen der detektierten Peaks und jenen der idealen harmonischen Reihe. Es wird sowohl die Abweichung der detektierten (*"measured"*) Peaks von den idealen (*"predicted"*) Teiltönen als auch die Abweichung in umgekehrter Richtung in jeweils einem Fehlermaß ausgedrückt. Aus beiden Fehlermaßen wird dann der Gesamtfehler berechnet.

Der predicted-zu-measured Fehler ist als

$$\begin{aligned} Err_{p\rightarrow m} &= \sum_{n=1}^{N} Err(\Delta f_n, f_n, a_n, A_{max}) = \\ &= \sum_{n=1}^{N} \Delta f_n \cdot (f_n)^{-p} + \frac{a_n}{A_{max}} \times [q\Delta f_n \cdot (f_n)^{-p} - r] \end{aligned} \tag{4.6}$$

definiert. N ist die Anzahl der idealen Teiltöne, Δf_n die Differenz zwischen einem idealen Teilton und dem naheliegendsten detektierten Peak. f_n und a_n sind die Frequenz und die Amplitude der idealen Teiltöne; A_{max} ist die maximale im momentanen Spektrum vorkommende Amplitude. Der Index p steht für *predicted*; der Index m für *measured*.

Der measured-zu-predicted Fehler wird analog zu Gleichung (4.6) als

$$\begin{aligned} Err_{m\to p} &= \sum_{k=1}^{K} Err(\Delta f_k, f_k, a_k, A_{max}) = \\ &= \sum_{k=1}^{K} \Delta f_k \cdot (f_k)^{-p} + \frac{a_k}{A_{max}} \times [q\Delta f_k \cdot (f_k)^{-p} - r] \end{aligned} \tag{4.7}$$

definiert. K ist die Anzahl der detektierten spektralen Peaks.

Der Gesamtfehler Err_{ges} ist dann

$$Err_{ges} = \frac{Err_{p\to m}}{N} + \rho\frac{Err_{m\to p}}{K} \quad . \tag{4.8}$$

Für die Konstanten schlagen Maher und Beauchamp (1994) $p = 0.5$, $q = 1.4$, $r = 0.5$ und $\rho = 0.33$ vor. Der Kandidat der den kleinsten Gesamtfehler produziert wird als Grundfrequenz des momentanen Verarbeitungsblocks gewählt. Ist dieser minimale Gesamtfehler aber größer als 1.5 wird der aktuelle Block als nicht-periodisch eingestuft.

Bei Sprachsignalen kann es aber vorkommen, dass das TWM-Verfahren Segmente als periodisch einstuft, die nicht als periodisch empfunden werden (vor allem der Laut /ss/). Eine Tonhöhentransformation dieser Segmente ist nicht sinnvoll, da sie den Klang der enthaltenen Laute verändert.
Um diese Artefakte zu umgehen wird in der vorliegenden Anwendung zusätzlich zum TWM eine Betrachtung der spektralen Energieverteilung zur Beurteilung der Periodizität verwendet. Bei den oben genannten Lauten, die fälschlicherweise als periodisch eingestuft werden ist das Verhältnis aus der Energie in den höheren Frequenzbereichen und der Energie in den niedrigeren Frequenzbereichen deutlich höher als bei den tatsächlich periodischen Lauten.
Um ein einfaches Maß zur Einschätzung der spektralen Energieverteilung zu erhalten werden die Bins des Betragsspektrums im Bereich zwischen 4 und 5 kHz sowie zwischen 50 Hz und 2 kHz aufsummiert. Ist der Quotient der beiden Summen größer als 0,3, so wird der betreffende Verarbeitungsblock sofort als nicht-periodisch eingestuft, ohne, dass das TWM-Verfahren angewendet wird (vgl. Abbildung 4.5).

Korrektur der F_0-Kontur

Da der oben beschriebene F_0-Detektions- und Bestimmungs-Algorithmus nicht absolut robust arbeitet, kann die Performance des gesamten Systems noch durch eine Nachbearbeitung der berechneten F_0-Kontur verbessert werden. Dabei wird wie folgt vorgegangen (vgl. Abbildung 4.6):

- Nicht-periodische Segmente werden über bis zu 20 ms hinweg interpoliert.
- Periodische Segmente die kürzer als 20 ms sind werden als nicht-periodisch eingestuft.

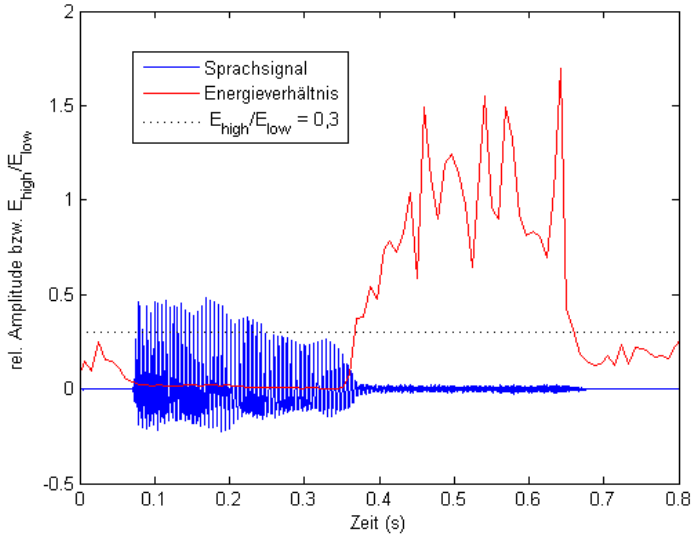


Abb. 4.5: Wellenform des Wortes "eins" (blau) und das Verhältnis der Energiemaße (rot, siehe Text)

- Abschließend wird die Kontur durch einen Medianfilter 10. Ordnung geglättet um durch Oktavierung der eigentlichen Grundfrequenz entstandene Ausreißer zu korrigieren.

Mittels der korrigierten F_0-Kontur werden dann die stabilen Teiltöne aus dem Spektrum des Analysesignals bestimmt und Trajektorien zugeordnet.

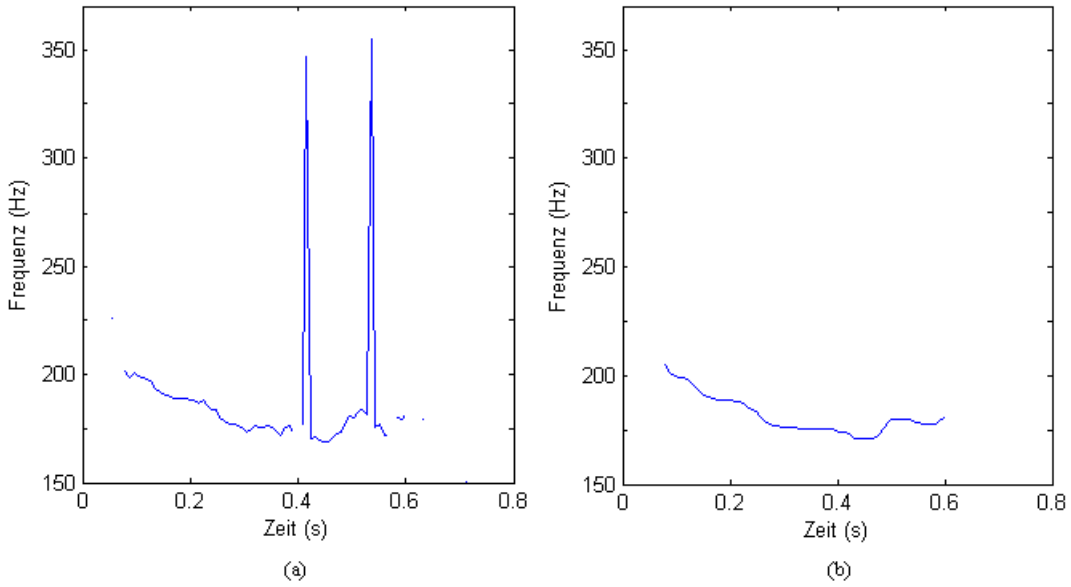


Abb. 4.6: Ausschnitt einer F_0-Kontur ohne Korrekturen (a) und mit Korrekturen (b)

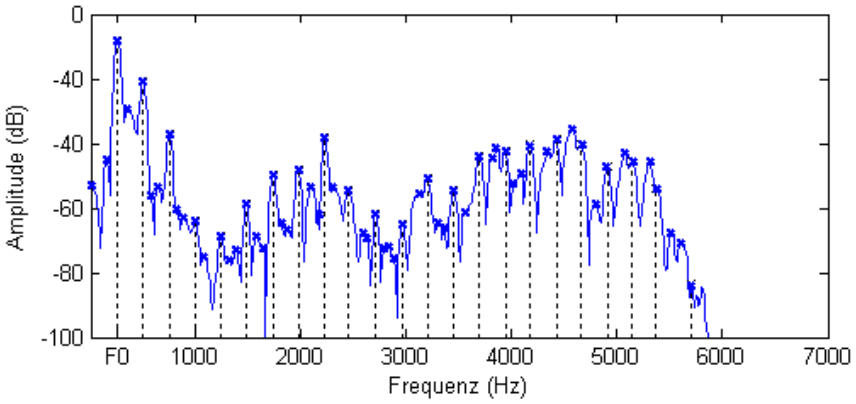


Abb. 4.7: Spektrum eines ausgewählten Blocks des Eingangssignals; Kreuze: detektierte Maxima; gestrichelte Linien: identifizierte Teiltöne

Peak-Trajektorien

Um einem Teilton bei der Resynthese eine momentane Phase zuordnen zu können muss dieser Teilton bis zu seinem Ursprung zurückverfolgt werden können (Gleichung (4.15)). Zu diesem Zweck werden die Teiltöne Trajektorien zugeordnet.
Dabei darf aber eine Trajektorie nicht einfach von einem Teilton zu dem naheliegendsten Teilton im folgenden Block geführt werden. Ändert sich die Grundfrequenz stark in zwei aufeinander folgenden Blöcken werden die Trajektorien so nicht sauber geführt (Abbildung 4.8 (a)).
Die Information über die momentane Grundfrequenz kann benutzt werden, um für jeden virtuellen Teilton dieser Grundfrequenz einen *Guide* zu definieren. Dieser Guide sucht sich dann von Block zu Block den stärksten Peak in dem ihm zugewiesenen Frequenzbereich und führt so eine Trajektorie weiter, beendet sie oder startet eine neue (Abbildung 4.8 (b)).

Einzelne Trajektorien werden dabei über bis zu 20 ms hinweg interpoliert. Die Amplitude des entsprechenden Teiltons wird linear interpoliert, die Frequenz f im i-ten Block gemäß

$$f_i = f_{i-1} \cdot \frac{F_{0,i}}{F_{0,i-1}} \quad . \tag{4.9}$$

F_0 ist die Grundfrequenz im jeweiligen Block. Die Phase wird gemäß

$$\phi_i = \text{princarg}(\phi_{i-1} + 2\pi f_i \Delta t) \tag{4.10}$$

berechnet. Δt ist der zeitliche Abstand zweier Blöcke. *princarg* steht für *Principle Argument*, eine Funktion, die einen beliebigen Phasenwert durch geeignete Addition ganzer Vielfacher einer Periode (2π) in das Intervall $]-\pi, \pi]$ transferiert.

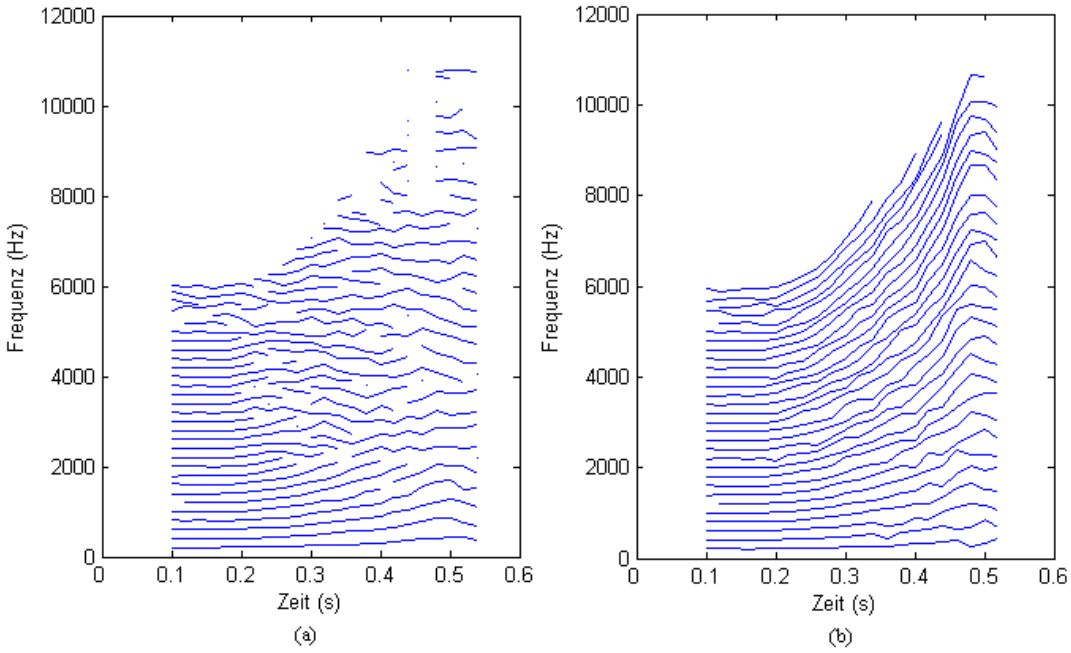


Abb. 4.8: Peak-Tracking ohne (a) und mit Guides (b)

4.2.2 Residualkomponente

Sobald die stabilen Teiltöne des Signals identifiziert sind, kann das Spektrum des deterministischen Signalanteils unter Berücksichtigung des Einflusses der Analyse-Fensterung approximiert werden (siehe Synthese, Abschnitt 4.3.1). Die Subtraktion des Spektrums der Sinuskomponente vom Signalspektrum ergibt das Spektrum der Residualkomponente.

4.3 Synthese

4.3.1 Sinuskomponente

Eine Sinusschwingung im Frequenzbereich stellt eine *sinc*-ähnliche Funktion, nämlich die Transformation des benutzten Fensters, dar. Wegen des bereits erwähnten hohen Verhältnisses der Höhen der Hauptkeule zu den Nebenkeulen brauchen im vorliegenden Fall nur die Samples der Hauptkeule mit entsprechender Frequenz, Phase und Amplitude berechnet zu werden. Durch entsprechende Addition solcher Keulen können beliebig viele Sinusschwingungen in ein Spektrum eingefügt werden (additive Synthese, vgl. Abb. 4.10 (b)). Eine inverse Fourier-Transformation überführt das Signal dann wieder in den Zeitbereich (vgl. hierzu Abschnitt 4.3.3).

Real- und Imaginärteil des Spektrums eines Teiltons erhält man gemäß

$$\Re_r^k = A_{HK,r}^k \cdot A_r \cos\phi_r \tag{4.11}$$
$$\Im_r^k = A_{HK,r}^k \cdot A_r \sin\phi_r \quad . \tag{4.12}$$

$A_{HK,r}^k$ ist die Amplitude der Hauptkeule des r-ten Teiltons am Bin k, A_r ist seine Amplitude und ϕ_r seine Phase. Das Zeigerdiagramm in Abbildung 4.9 veranschaulicht die Gleichungen (4.11) und (4.12).

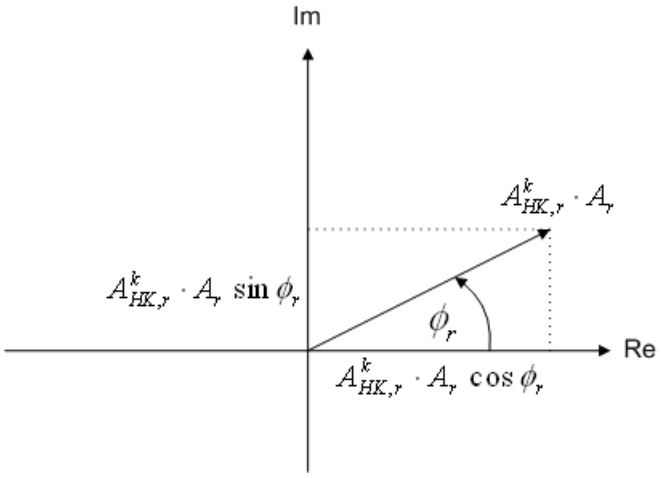


Abb. 4.9: Veranschaulichung der Berechnung des komplexen Spektrums des r-ten Teiltons am Bin k in der komplexen Zahlenebene; $A_{HK,r}^k$ ist die Amplitude der Hauptkeule am Bin k, A_r ist die Amplitude und ϕ_r die Phase des r-ten Teiltons

Das verwendete Blackman-Harris-Fenster kann im Frequenzbereich als

$$W_{BH92}(\Omega) = \sum_{m=1}^{4}(-1)^m \frac{a_m}{2}\left[D\left(\Omega - \frac{2\pi}{N}m\right) + D\left(\Omega + \frac{2\pi}{N}m\right)\right] \tag{4.13}$$

dargestellt werden (Loscos 2003). N ist die Länge des Fensters, die Parameter a_m sind $a_1 = 0,35875$, $a_2 = 0,48829$, $a_3 = 0,14128$ und $a_4 = 0,01168$. $D(\Omega)$ ist der Dirichlet-Kernel

$$D(\Omega) = e^{j\frac{\Omega}{2}} \frac{\sin(\frac{N}{2})\Omega}{\sin(\frac{1}{2})\Omega} \quad . \tag{4.14}$$

Für jeden zu resynthetisierenden Teilton wird nun eine Hauptkeule des Fensters mit der entsprechenden Amplitude, Mittenfrequenz und Phase in das Spektrum addiert. Da die Hauptkeule des Fensters als null-phasig angenommen werden kann (Loscos 2003), erhalten alle Bins der Hauptkeule die Phase des zugehörigen Teiltons (vgl. Gleichung (4.15) bezüglich der Phase).

Da die Frequenzen der Teiltöne auch Werte zwischen den Bins annehmen können müssten eigentlich viele diskrete Hauptkeulenspektren bestimmt werden. Um diesen

hohen Rechenaufwand zu umgehen wird in der vorliegenden Implementation eine detaillierte Vorlage des Spektrums der Hauptkeule mit 4096 Abtastpunkten berechnet, aus der dann die Amplitudenwerte ausgelesen werden.

Da das menschliche Gehör nicht in der Lage ist, die genauen Phasenbeziehungen zwischen den Teiltönen eines Sprachsignals wahrzunehmen, können die Phasen der synthetisierten Teiltöne durch

$$\phi_r(t) = \int_0^t \omega_r(\tau) d\tau \tag{4.15}$$

approximiert werden (Serra 1997). $\omega_r(t)$ ist die Kreisfrequenz des r-ten Teiltons. Das Alter t eines Teiltons erhält man aus den Peak-Trajektorien.

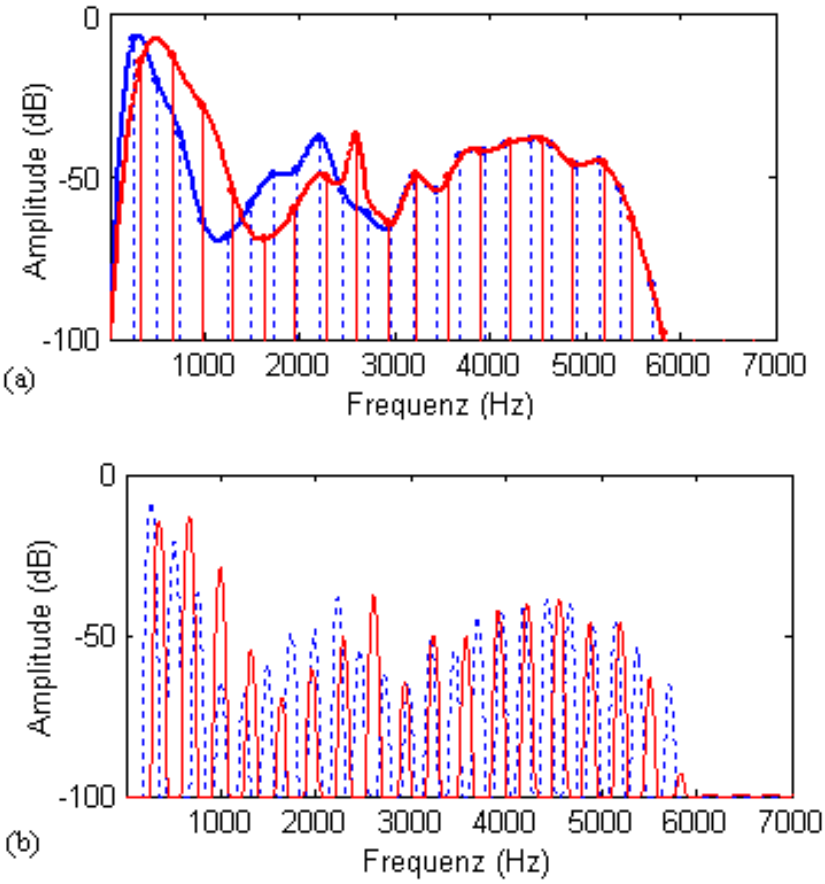


Abb. 4.10: Sinuskomponente des gleichen Blocks wie aus Abbildung 4.7; (a) spektrale Einhüllenden mit Teiltönen; (b) Betragsspektren

4.3.2 Residualkomponente

Dem hier verwendeten Modell liegt die Annahme zugrunde, dass das Residualsignal $e(t)$ ein stochastisches Signal ist. Es kann also vollständig durch seine Amplitude und seine

allgemeinen spektralen Eigenschaften beschrieben werden. Es ist nicht nötig, weder die Phasenbeziehungen noch die Details des Betragsspektrums zu beschreiben.
Das momentane Residualsignal kann also vollständig durch ein Filter, das die allgemeinen spektralen Eigenschaften des Signals beschreibt (also z.B. der Einhüllenden des Betragsspektrums), beschrieben werden:

$$e(t) = \int_0^t h(t,\tau)u(\tau)d\tau \tag{4.16}$$

$u(\tau)$ ist weißes Rauschen, $h(t,\tau)$ ist die Impulsantwort eines zeitvarianten Filters zum Zeitpunkt t.

Die naheliegendste Vorgangsweise um das Residualsignal zu resynthetisieren wäre nun, weißes Rauschen blockweise mit den entsprechenden spektralen Einhüllenden zu filtern. Eine einfachere Lösung ist aber, aus den spektralen Einhüllenden mit Hilfe der IFFT ein stochastisches Signal zu generieren. Das heißt, es muss das (komplexe) Spektrum eines stochastischen Signals generiert werden.
Als Betragsspektrum kann die Einhüllende direkt verwendet werden. Die Phase wird aus einer Abfolge von zufälligen Werten im Intervall $]-\pi,\pi]$ gebildet.
Allerdings sind in diesem Spektrum im Gegensatz zum Spektrum des deterministischen Signals weder die Analyse-Fensterung noch das bei der Analyse verwendete Zero-Padding berücksichtigt. Um das Zero-Padding zu simulieren werden nicht für jedes Bin des Spektrums Phasenwerte generiert, sondern für jedes ganzzahlige Vielfache des Zero-Padding-Faktors (also z.B. für jedes 4. Bin bei 4-fachem Zero-Padding) und zwischen diesen Werten interpoliert. Bei der Interpolation muss das Phasen-Wrapping berücksichtigt werden. (vgl. Abschnitt 5.4.2).
Um dem Spektrum den Einfluss der Analysefensterung aufzuprägen wird das generierte Residualspektrum mit der Hauptkeule des Analysefensters gefaltet. Nun muss nur noch die Energie dieses generierten Spektrums der Energie des durch Subtraktion des deterministischen Signals vom Eingangsignal erhaltenen Residualspektrums angeglichen werden.
Die Energie eines Spektrums $X(k)$ der Länge K ist

$$NRG_X = \sum_{k=0}^{K-1} |X(k)|^2 \quad . \tag{4.17}$$

Das generierte Residualspektrum $E'(k)$ wird gemäß

$$E''(k) = E'(k) \cdot \sqrt{\frac{NRG_E}{NRG_{E'}}} \tag{4.18}$$

skaliert. E ist das Referenzspektrum.

Danach können Residual- und Teiltonspektrum addiert und mit einer einzigen IFFT in den Zeitbereich transformiert werden.

4.3.3 Synthesefensterung

Eigentlich müssten die Länge N des Fensters w und Schrittweite (*Hop Size*) H so gewählt werden, dass sich die resultierende Einhüllende zu einer Konstanten addiert:

$$A_w(n) = \sum_{m=-\infty}^{\infty} w(n - mH) \approx const. \tag{4.19}$$

Ein Maß für die Abweichung von A_w von einer Konstanten ist die Differenz des maximalen und des minimalen Wertes der Einhüllenden relativ zum Maximalwert:

$$d_w = \frac{\max_n(A_w(n)) - \min_n(A_w(n))}{\max(A_w(n))} \tag{4.20}$$

d_w wird *Amplitude Deviation* des Overlap-Faktors genannt und sollte unter 1% liegen. Beim Blackman-Harris-Fenster liegt d_w aber nur bei sehr hohen Overlap-Raten niedrig genug. Um den Rechenaufwand bei diesen hohen Oberlap-Raten zu umgehen wird in der vorliegenden Implementation ein zweites Fenster verwendet, das die gewünschten Überlappungseigenschaften im Zeitbereich aufweist (Rodet und Depalle 1992):

Das Ergebnis der IFFT des Spektrums eines Verarbeitungsblocks ist das Signal $w_{BH92} \cdot s[n]$, wobei w_{BH92} das Blackman-Harris-Fenster und $s[n]$ das gesuchte Signal ist. Eine Division durch w_{BH92} ergibt $s[n]$. Gewichtet man dieses Signal nun mit einem Dreieckfenster w_Δ, so prägt man dem Signal die für den Overlap-Add-Prozess notwendigen Eigenschaften auf.
Die Länge des Dreieckfensters wird doppelt so lange wie die Hop-Size gewählt. Dadurch addieren sich die Fenster bei der Overlap-Add-Prozedur exakt zu 1. Außerdem ist so die Synthese von der Analyse unabhängig. Die Länge des Analysefensters kann den momentanen Anforderungen (z.B. der Grundfrequenz) angepasst werden.
Ein weiterer Vorteil dieser Vorgangsweise ist, dass die Amplituden und Frequenzen der Teiltöne zwischen aufeinander folgenden Blöcken linear interpoliert werden, und somit Diskontinuitäten vermieden werden.

4.4 Transformationen

Für alle durchgeführten Transformationen bezüglich der Formanten gilt, dass das Potenzgesetz aus Abschnitt 2.3 nur auf Signalkomponenten unter 2000 Hz angewendet wird. Über 3000 Hz sind das Original- und das transformierte Signal identisch, dazwischen befindet sich ein Übergangsbereich zwischen beiden Bereichen (vgl. Abbildung 4.10).

4.4.1 Sinuskomponente

Formanten und Grundfrequenz

Zur Verschiebung der Formanten wird die spektrale Einhüllende des deterministischen Signalanteils bestimmt und skaliert. Die Amplituden der Teiltöne werden dann entsprechend der skalierten Einhüllenden gewichtet.
Die Einhüllende wird in der vorliegenden Implementation mittels der optimierten True-Envelope-Schätzung berechnet, da sich die Einhüllende an den Maxima des Betragsspektrums orientieren soll (vgl. Abschnitt 3.2.5). Abbildung 4.10 (a) illustriert die Skalierung einer Einhüllenden.
Zur Änderung der Grundfrequenz müssen die Frequenzen der einzelnen Teiltöne lediglich mit dem entsprechenden Faktor multipliziert und gemäß der spektralen Einhüllenden gewichtet werden (vgl. ebenfalls Abbildung 4.10 (a)). Wird die Grundfrequenz nach unten verschoben, so wird ein Teil des Spektrums an die im oberen Frequenzbereich entstehende leere Stelle kopiert.

Wird auch die Kontur der Grundfrequenz manipuliert, so muss zuerst die mittlere Grundfrequenz des gesamten zu transformierenden Sprachsignals bestimmt werden, indem der Mittelwert der Grundfrequenz aller als periodisch eingestufter Segmente berechnet wird. Nach einem Vergleich der momentanen Grundfrequenz mit der mittleren können die entsprechenden Skalierungen vorgenommen werden. Diese Skalierungen müssen dann auf die Frequenzen aller Teiltöne angewendet werden (vgl. Abbildung 4.11).

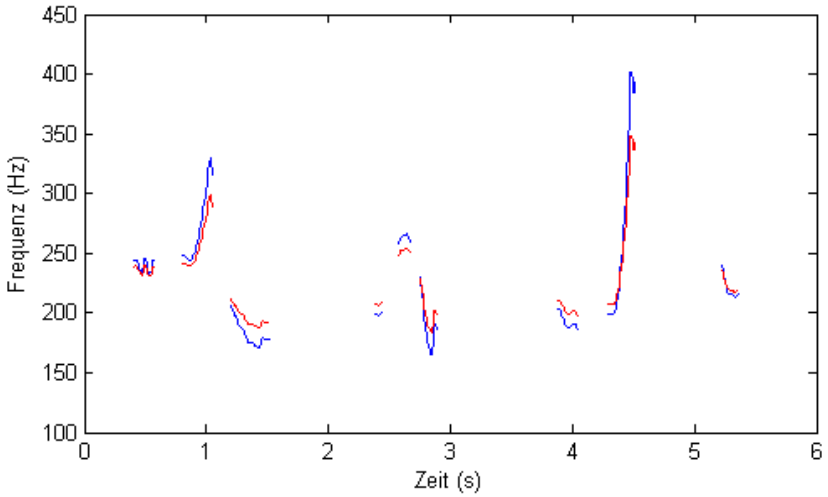


Abb. 4.11: Kontur der Grundfrequenz während eines kurzen Satzes; ursprüngliche Kontur (blau); Kontur um den Faktor 0.7 gestaucht (rot)

Bestimmung der Ordnung des Cepstrum

Die *Cut Quefrency* (Ordnung) der Cepstral-Analyse ist ein kritischer Parameter bei der Bestimmung der spektralen Einhüllenden. Die Einhüllende sollte nicht so detailliert sein,

dass einzelne Teitöne des Spektrums abgebildet werden, sondern es sollen die Lücken zwischen den Teiltönen aufgefüllt werden.
Da sich der Abstand zweier aufeinander folgender Teiltöne unter anderem durch Tonhöhenschwankungen des Signals ständig verändert muss die Cut Quefrency auf die jeweiligen momentanen Anforderungen adaptiert werden.
Bei einem Segment mit tiefer Grundfrequenz liegen die Teiltöne näher beieinander als bei einem Segment mit hoher Grundfrequenz, somit müssen im ersteren Fall mehr Details beschrieben werden. Wird die Einhüllende des Spektrums auch bei hohen Grundfrequenzen genauso detailliert beschrieben werden die Lücken zwischen den Teiltönen nicht aufgefüllt, was aber erwünscht ist.
Als sinnvolle Faustregel für die Ordnung P_c des Ceptrums hat sich

$$P_c \leq \frac{F_s}{2\delta_F} \tag{4.21}$$

erwiesen. F_s ist die Samplingfrequenz und δ_F die größte zu überbrückende Lücke im Betragsspektrum.

Manipulation von Jitter und Shimmer

Um die Stimme eines erwachsenen Sprechers künstlich zu altern muss dem Signal Jitter und Shimmer, also Schwankungen der Dauer und Amplitude zweier aufeinander folgender Perioden, hinzugefügt werden (vgl. Anhang 2.4). Die Idee hinter dem hier angewendeten Verfahren ist, das Signal um eine Oktave nach unten zu transponieren, das transponierte Signal um $T_0 + T_{jit}$ zu verschieben und mit dem Faktor $10^{A_{shim}/20}$ zu skalieren (Loscos und Bonada 2004). T_0 ist die momentane Periodendauer des Signals, T_{jit} und A_{shim} sind mittelwertfreie Zufallsvariablen.
Durch Addition der beiden Signale erhält man das Ausgangssignal. Abbildung 4.12 veranschaulicht die Vorgangsweise im Zeitbereich, Abbildung 4.13 zeigt das entsprechende Blockschaltbild.

In der vorliegenden Anwendung ist es vorteilhafter den Algorithmus im Frequenzbereich zu implementieren, da die Position der einzelnen Glottispulse nicht bekannt ist.
Die Transponierung des Eingangssignals erfolgt durch Hinzufügen von Subharmonischen (Schwingungen zwischen den eigentlichen Teiltönen) an den Stellen $1,5 \cdot F_0$, $2,5 \cdot F_0$, $3,5 \cdot F_0$, etc., mit der Grundfrequenz F_0 des Signals (vgl. Abschnitt 3.4.2). Die Subharmonische bei $0,5 \cdot F_0$ kann als von der Grundschwingung verdeckt angenommen werden.
Loscos und Bonada (2004) schlagen vor, als Phase einer Subharmonischen die Phase des nächstliegenden Teiltons zuzüglich des Offsets

$$\Delta\phi_{sh} = 2\pi \cdot f_r \cdot \left(\frac{f_{sh}}{f_r} - 1\right) \cdot \Delta t \tag{4.22}$$

zu verwenden und die Subharmonischen somit mit den nächstliegenden Teiltönen zu synchronisieren. f_{sh} ist die Frequenz der Subharmonischen, f_r die Frequenz des

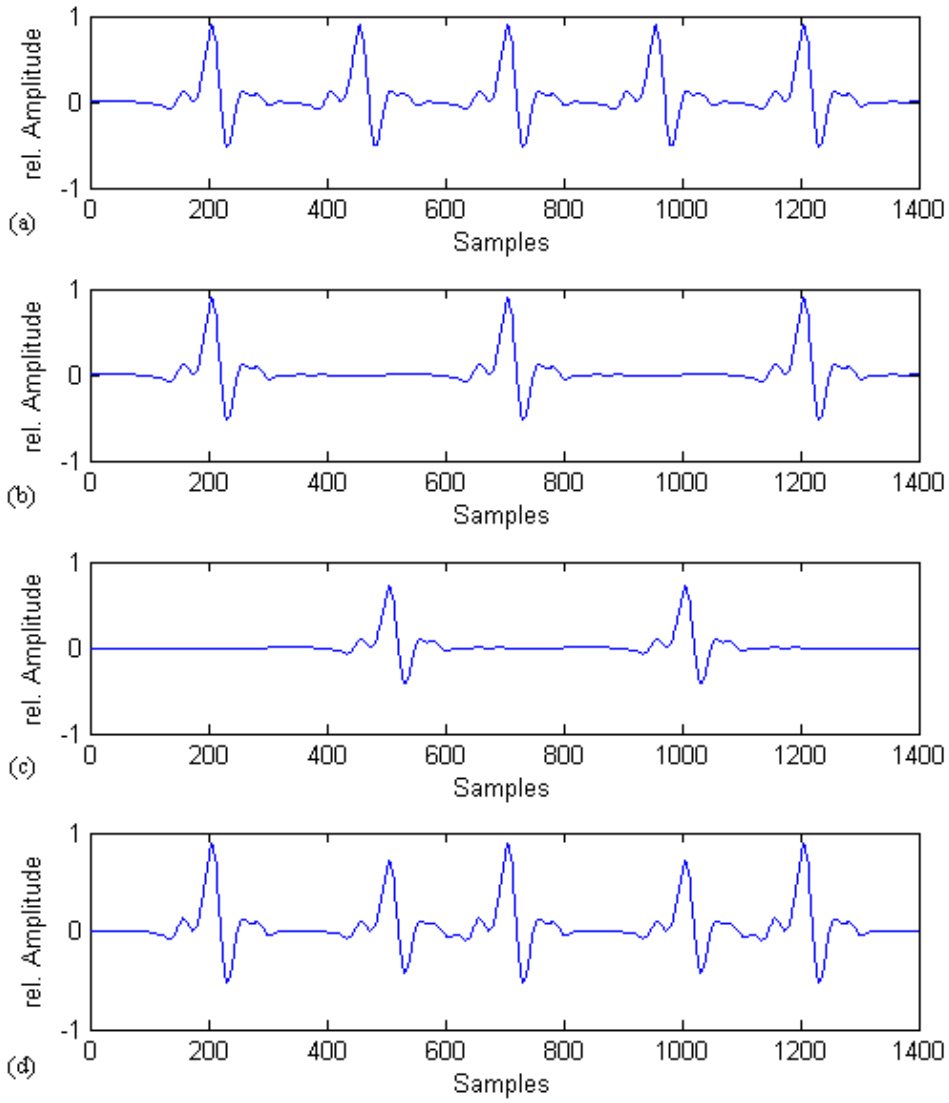


Abb. 4.12: (a) Eingangssignal; (b) um eine Oktave nach unten transponiertes Signal; (c) transponiertes Signal um T_0+T_{jit} verschoben mit $10^{A_{shim}/20}$ skaliert; (d) Ausgangssignal

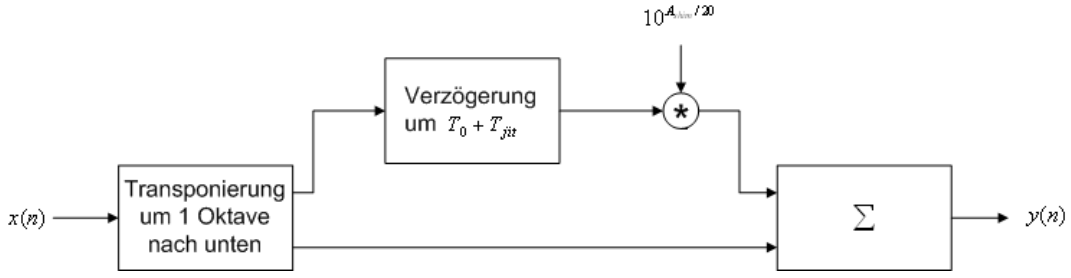


Abb. 4.13: Blockschaltbild im Zeitbereich

nächstliegenden Teiltons und Δt ist die Dauer eines Frames (vgl. Abbildung 4.14).

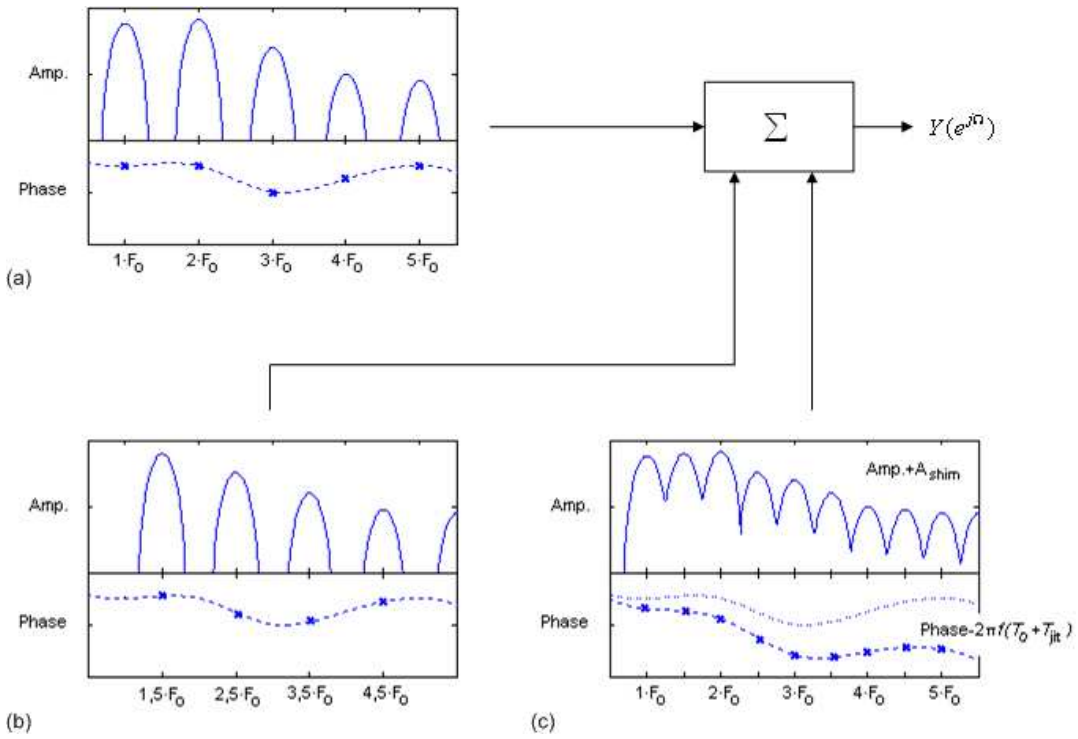


Abb. 4.14: (a) Spektrum des Signals; (b) Subharmonische; (c) transponiertes Signal um $T_0 + T_{jit}$ verschoben und mit $10^{A_{shim}/20}$ skaliert

In der vorliegenden Implementation hat es sich aber als zuverlässiger erwiesen, auch die Subharmonischen entsprechenden Trajektorien zuzuordnen und die Phase gemäß Gleichung (4.15) zu approximieren.
Die Zufallsbariablen T_{jit} und A_{shim} haben Varianzen von 5% bzw. 4dB.

Die zeitliche Verschiebung des transponierten Signals wird durch Addition des entsprechenden Phasenterms zum Spektrum erreicht:

$$\Delta\phi_r = -2\pi \cdot f_r \cdot (T_0 + T_{jit}) \tag{4.23}$$

Da dem Signal mit der Addition von Hauptkeulen auch Energie hinzugefügt wird, muss das resultierende Spektrum entsprechend skaliert werden, sodass seine Energie der Energie des ursprünglichen Spektrums entspricht (Gleichung (4.18)).

4.4.2 Residualkomponente

Da das Residualsignal keine periodischen Komponenten mehr besitzt, müssen nur die Formanten entsprechend skaliert werden. Dies geschieht ähnlich der Formantverschiebung bei der Sinuskomponente mittels einer geeigneten Skalierung der spektralen Einhüllenden.
Bei der Berechnung der Einhüllenden des Residualspektrums wird aber nicht die True-

Envelope-Methode verwendet, da diese sich hauptsächlich an den spektralen Peaks orientiert. Dies ist beim Residualsignal nicht erwünscht, da ein einzelner freistehender Peak die Einhüllende in einem größeren Bereich um den Peak herum anhebt. Deshalb wird hier die spektrale Glättung mittels des Cepstrum verwendet (vgl. Abbildung 3.3).
Aus der skalierten Einhüllenden wird dann ein stochastisches Signal generiert (vgl. Abschnitt 4.3.2)

4.5 Hörbeispiele

Hörbeispiel 3 (CD-Track 03):	Frauenstimme, bandbegrenzt ($f_{Grenz} = 5.5kHz$, $f_{Sampling} = 22.05kHz$)
Hörbeispiel 4 (CD-Track 04):	Hörbeispiel 3 mittels SMS resynthetisiert
Hörbeispiel 5 (CD-Track 05):	Resynthetisierte Sinuskomponente des Signals aus Hörbeispiel 3
Hörbeispiel 6 (CD-Track 06):	Residualkomponente des Signals aus Hörbeispiel 3, durch ein stochastisches Signal approximiert
Hörbeispiel 7 (CD-Track 07):	Hörbeispiel 3 in einen Mann transformiert
Hörbeispiel 8 (CD-Track 08):	Hörbeispiel 3 in ein Kind transformiert
Hörbeispiel 9 (CD-Track 09):	Hörbeispiel 3 künstlich gealtert

Kapitel 5

Voice Transformation mittels der Resynthese von Glottispulsen

5.1 Einleitung

In diesem Kapitel wird ein weiteres im Rahmen dieser Diplomarbeit in der Software MATLAB implementiertes System vorgestellt, dass die Manipulation der in den Abschnitten 2.3 und 2.4 beschriebenen Parameter von Sprachsignalen ermöglicht.
Dazu werden einzelne Glottispulse im Frequenzbereich modelliert (Bonada 2004). Das System erlaubt gleichzeitig eine unabhängige Kontrolle über die Glottispulse im Zeitbereich sowie flexible Klangfarben- und Phasenmodifikationen im Frequenzbereich.

5.2 Modellierung

Die Kurzzeit-Fourier-Transformation (*Short-Time Fourier Transformation*, STFT) eines gefensterten Verarbeitungsblocks $x[n]$ kann als

$$\begin{aligned} x[n] &= s[n] \cdot w[n] \\ X(e^{j\Omega}) &= \sum_{n=0}^{N-1} x[n] e^{-j\Omega n} \end{aligned} \tag{5.1}$$

beschrieben werden. $s[n]$ bezeichnet das Eingangssignal, $w[n]$ das Fenster.

Im Folgenden wird ein Rechteckfenster der Länge N und Höhe 1 angenommen, sowie ein endliches Eingangssignal, das kürzer als N und vollständig vom Fenster abgedeckt ist.

Wird ein Signal um Δn Samples verzögert, so ergibt sich für seine STFT:

$$\begin{aligned} S_{delayed\Delta n}(e^{j\Omega}) &= \sum_{n=0}^{N-1} s[n-\Delta n]e^{-j\Omega n} = \\ &= \sum_{n=0}^{N-1} x[n-\Delta n]e^{-j\Omega n} = \\ &\overset{m=n-\Delta n}{=} \sum_{m=-\Delta n}^{N-1-\Delta n} x[m]e^{-j\Omega(m+\Delta n)} \approx \\ &\approx X(e^{j\Omega})e^{-j\Omega\Delta n} \end{aligned} \tag{5.2}$$

Ist das verzögerte Signal ebenfalls vollständig vom Fenster abgedeckt, so ist die Näherung in der letzten Zeile eine Identität.

Betrachten wir nun ein Signal $y[n]$ das aus der Summe von R identischen Signalen $s[n]$ besteht, die um Δn Samples verschoben sind. Überlappungen sind dabei möglich.

$$y[n] = s[n] + s[n-\Delta n] + s[n-2\Delta n] + \ldots + s[n-(R-1)\Delta n] \tag{5.3}$$

Die STFT von $y[n]$ ist dann:

$$\begin{aligned} Y(e^{j\Omega}) &= \sum_{n=0}^{N-1} y[n]e^{-j\Omega n} \approx \\ &\approx X(e^{j\Omega}) \cdot [1 + e^{-j\Omega\Delta n} + e^{-2j\Omega\Delta n} + \ldots + e^{-(R-1)j\Omega\Delta n}] = \\ &= X(e^{j\Omega}) \cdot \sum_{n=0}^{R-1} e^{-j\Omega\Delta nr} = \\ &= X(e^{j\Omega}) \cdot \frac{1-e^{-j\Omega\Delta nR}}{1-e^{-j\Omega\Delta n}} = \\ &= X(e^{j\Omega}) \cdot e^{-j\Omega\Delta n\frac{R-1}{2}} \cdot \frac{\sin(0.5\Omega\Delta nR)}{\sin(0.5\Omega\Delta n)} \triangleq \\ &\triangleq X(e^{j\Omega}) \cdot \mathrm{sinc}_R(\Omega\Delta n) \end{aligned} \tag{5.4}$$

D.h., das Spektrum des Signals $y[n]$ entspricht also ungefähr dem des Signals $x[n]$ mit der Gewichtung $\mathrm{sinc}_R(\Omega\Delta n)$. Abbildung 5.1 illustriert den sinc_R-Term.

Der Effekt der Gewichtung mit $\mathrm{sinc}_R(\Omega\Delta n)$ entspricht ungefähr dem Abtasten des Spektrums von $x[n]$. Der Term $\mathrm{sinc}_R(\Omega\Delta n)$ kann als Folge von Pulsen im Abstand $\frac{2\pi}{\Delta n}$ mit der Höhe R gesehen werden (Beweis siehe Bonada 2004).

Wenn davon ausgegangen wird, dass sich $|X(e^{j\Omega})|$ und $\angle X(e^{j\Omega})$ verhältnismäßig langsam mit der Zeit ändern, kann nun das Spektrum $X(e^{j\Omega})$ durch Interpolation der Werte von $Y(e^{j\Omega})$ an den Frequenzen Ω_R (d.h. an den Frequenzen seiner Teiltöne) geschätzt

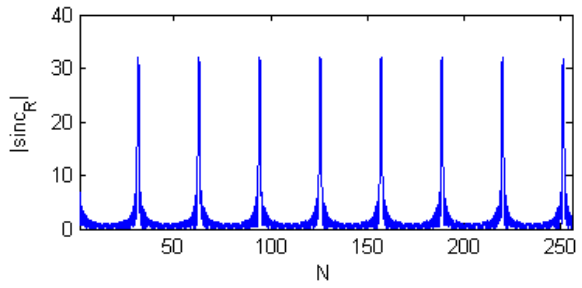


Abb. 5.1: $|\mathrm{sinc}_R(\Omega\Delta n)|$ für $R = 32$, $\Delta n = 8$, $N = 256$;

werden.
Das rekonstruierte Signal $x'[n]$ kann dann mittels inverser STFT aus dem geschätzten Spektrum $X'(e^{j\Omega})$ berechnet werden. Schließlich kann nun das Eingangssignal $s[n]$ durch

$$s'[n] = x'[n] \tag{5.5}$$

angenähert werden.

Bei stimmhaften Segmenten von Sprachsignalen entspricht $s[n]$ einem durch den Vokaltrakt gefilterten Glottispuls. Allerdings wird man es meist mit vielen überlappenden Pulsen zu tun haben, die nicht vollständig vom Fenster abgedeckt werden, wie in Abbildung 5.2 dargestellt. In diesem Fall hängt Δn direkt über $\Delta n = \frac{f_S}{Grundfrequenz}$ mit der Grundfrequenz des Segmentes zusammen. f_S bezeichnet die Samplingfrequenz.
Es muss beachtet werden, dass die Pulse dann aber nicht mehr identisch sind, da sich die Charakteristika des sprachproduzierenden Systems ständig verändern. Wenn das Fenster aber kurz genug gewählt wird, sodass das Signal quasi-stationär ist, können die einzelnen Pulse geschätzt werden. Als guten Kompromiss zwischen Frequenz- und Zeitauflösung hat sich eine Fensterlänge von drei Perioden der Grundfrequenz bewährt.

5.3 Analyse

Zur Analyse des Sprachsignals wird das bereits in Kapitel 4 verwendete System benutzt, um die Maxima des momentanen Spektrums detektieren und - bei harmonischen Segmenten - mit Hilfe der Grundfrequenz die Teiltöne heraussortieren zu können.
Werden die Amplituden der detektierten Maxima mit $\frac{1}{R}$ gewichtet, so ergeben sich bei der Synthese mittels inverser STFT automatisch die korrekten Amplituden der Glottispulse (vgl. oben). R kann über

$$R = \frac{N \cdot Grundfrequenz}{f_S} \tag{5.6}$$

bestimmt werden.

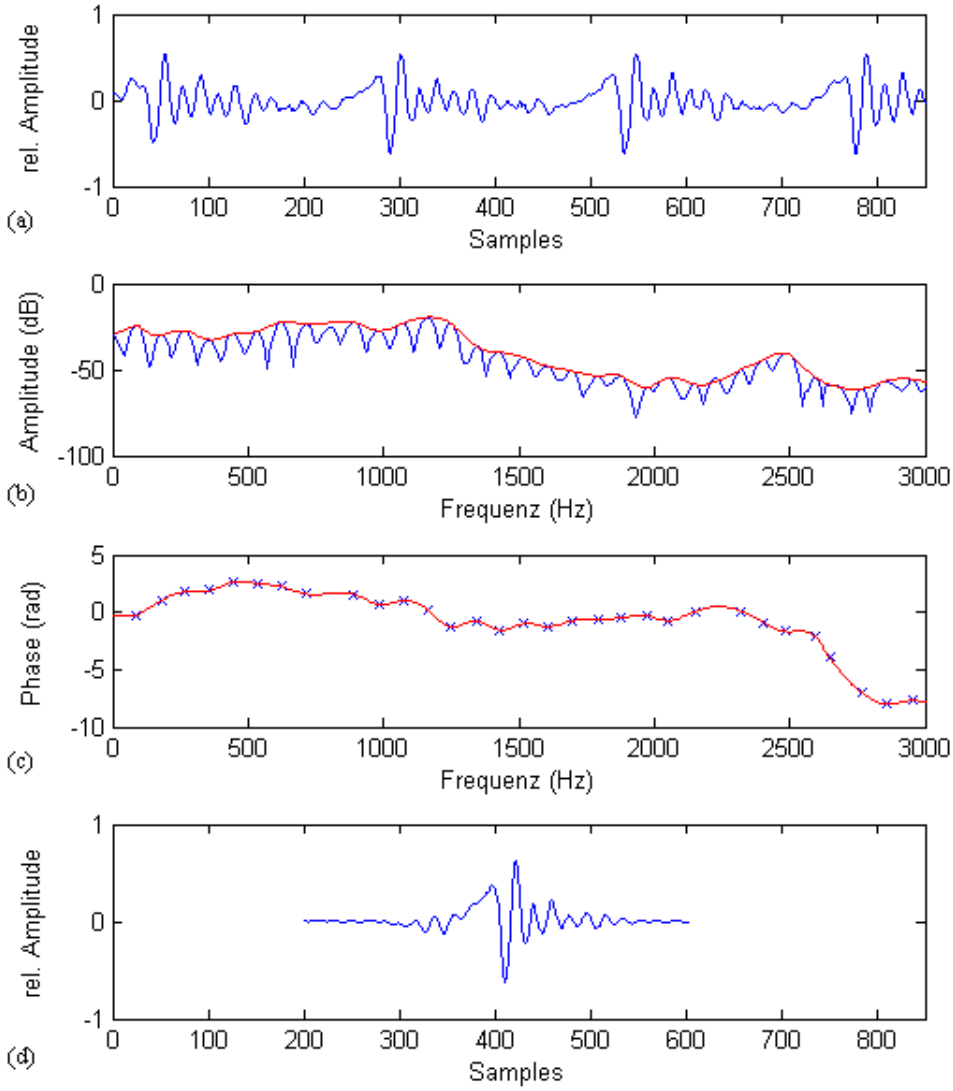


Abb. 5.2: Resynthese eines Glottispulses: (a) Eingangssignal; (b) Betragsspektrum des Eingangssignals (blau) mit Einhüllender des Teiltonspektrums (rot); (c) Phasen der Teiltöne (blaue Kreuze) mit Interpolation (rot); (d) resynthetisierter Glottispuls; $f_S = 20.05kHz$

5.4 Synthese

In der vorliegenden Anwendung wird das Betragsspektrum nicht zwischen den detektierten Teiltönen interpoliert, sondern es wird die mittels des True-Envelope-Verfahrens geschätzte spektrale Einhüllende als Pulsspektrum benutzt (vgl. Abschnitt 3.2.5). Die Einhüllende wird entsprechend der gewünschten Formantverschiebungen skaliert. Die Phasen der Teiltöne werden interpoliert.
Mit Hilfe der bei der Analyse bestimmten - und gegebenenfalls transformierten - momentanen Grundfrequenz können nun die Pulse entsprechend im Ausgangssignal platziert werden (vgl. Abschnitt 5.4.1). Dabei können gewünschter *Jitter* und *Shimmer* berücksichtigt werden. Analog zu der Vorgangsweise in Abschnitt 4.4.1 wird nur jeder

zweite Puls mit $10^{A_{shim}/20}$ skaliert bzw. um T_{jit} verschoben.

Es gelten die in Abschnitt 4.3.3 angeführten Überlegungen bezüglich der Fensterungen, allerdings braucht hier der resynthetisierte Puls nicht mehr durch das Analysefenster dividiert werden, da das Spektrum des Pulses nicht mit dem des Fensters gefaltet wurde.

5.4.1 Maximally Flat Phase Alignment

Um bei der Synthese die Kontrolle über die exakte zeitliche Position des Pulses zu haben ist es notwendig, die Pulse im Synthesefenster zu zentrieren.
Ist ein Puls im Fenster zentriert, so ist sein Phasenspektrum annähernd flach mit kleinen Verschiebungen unter den Formanten (vgl. Abbildung 5.2c). Auf diesem Umstand beruht das von Bonada (2004) vorgeschlagene *Maximally Flat Phase Alignment* (MFPA). Dabei wird diejenige Zeitverschiebung Δt_c gefunden, bei der das Phasenspektrum am flachsten - und somit der Puls zentriert - ist:

1. Zuerst müssen einige potenzielle Kandidaten ϕ'_{c0} für die Phase der Grundfrequenz im Intervall $]-\pi, \pi]$ definiert werden.

2. Für jeden Kandidaten muss dann die entsprechende Zeitverschiebung Δt_c auf alle anderen Teiltöne angewendet werden. Das führt zu einer Phasenverschiebung von $\Delta\phi_r = 2\pi f_r \Delta t_c$. f_r ist dabei die Frequenz des r-ten Teiltons.

3. Nun müssen die Phasendifferenzen aufeinander folgender Teiltöne gemäß $\phi_{diff} = \sum_r |\text{princarg}(\phi'_{r+1} + \phi'_r)|$ aufsummiert werden. *princarg* bezeichnet das *Principle Argument*, eine Funktion, die einen beliebigen Phasenwert durch geeignete Addition ganzer Vielfacher einer Periode (2π) in das Intervall $]-\pi, \pi]$ transferiert.

Ist nun diese Aufsummierung für alle Kandidaten durchgeführt, so beschreiben die verschiedenen Werte von ϕ_{diff} eine Funktion ähnlich eines Sinus. Ihr Minimum bestimmt dann den Kandidaten für die Phase der Grundfrequenz ϕ_{min}, bei dem der Puls im Fenster zentriert ist (Abbildung 5.3).

Dem MFPA liegt die Annahme zugrunde, dass der periodische Teil eines Sprachsignals durch eine Folge von Impulsen beschrieben werden kann. Dies trifft jedoch nicht zwingend zu. Bei einigen der stimmhaften Laute wird die Energie nicht impulshaft abgegeben, sondern ist über die ganze Periode des Signals verteilt (vgl. Abbildung 5.4). Somit ist die für das MFPA notwenige Phasensychronizität nicht gegeben, und die Pulse werden nicht zuverlässig im Fenster zentriert. Dies resultiert in einem eher rauen

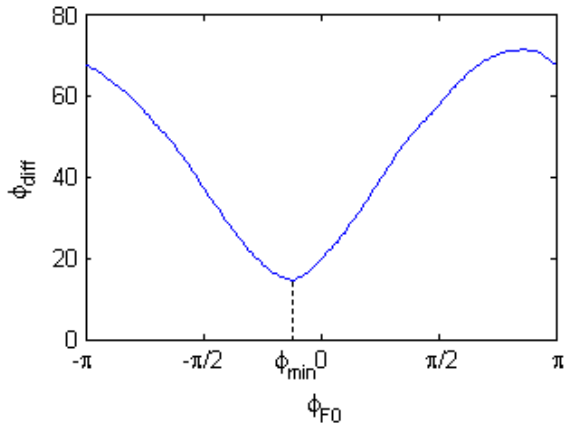


Abb. 5.3: Fehlerfunktion ϕ_{diff} beim MFPA

Charakter der resynthetisierten Stimme.

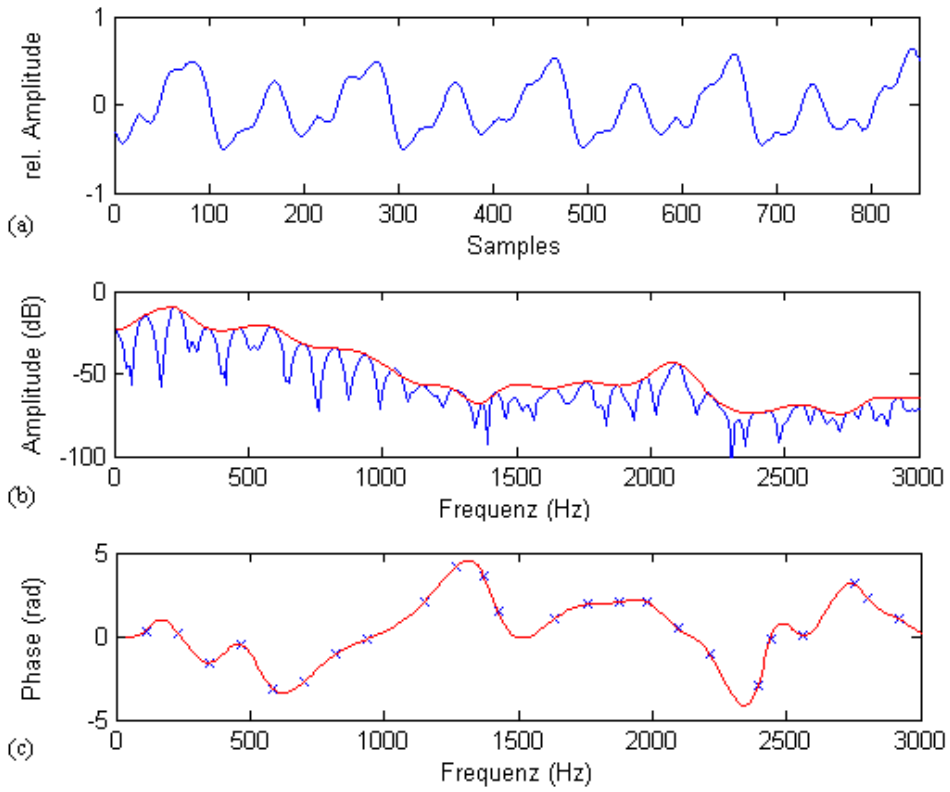


Abb. 5.4: (a) Wellenform des Lautes /u/; (b) Betragsspektrum (blau) mit Einhüllender (rot); (c) Phasen der Teiltöne (blaue Kreuze) mit Interpolation (rot)

Die Analyse-/Synthesestruktur des gesamten Systems ist in Abbildung 5.5 in einem Blockschaltbild zusammengefasst.

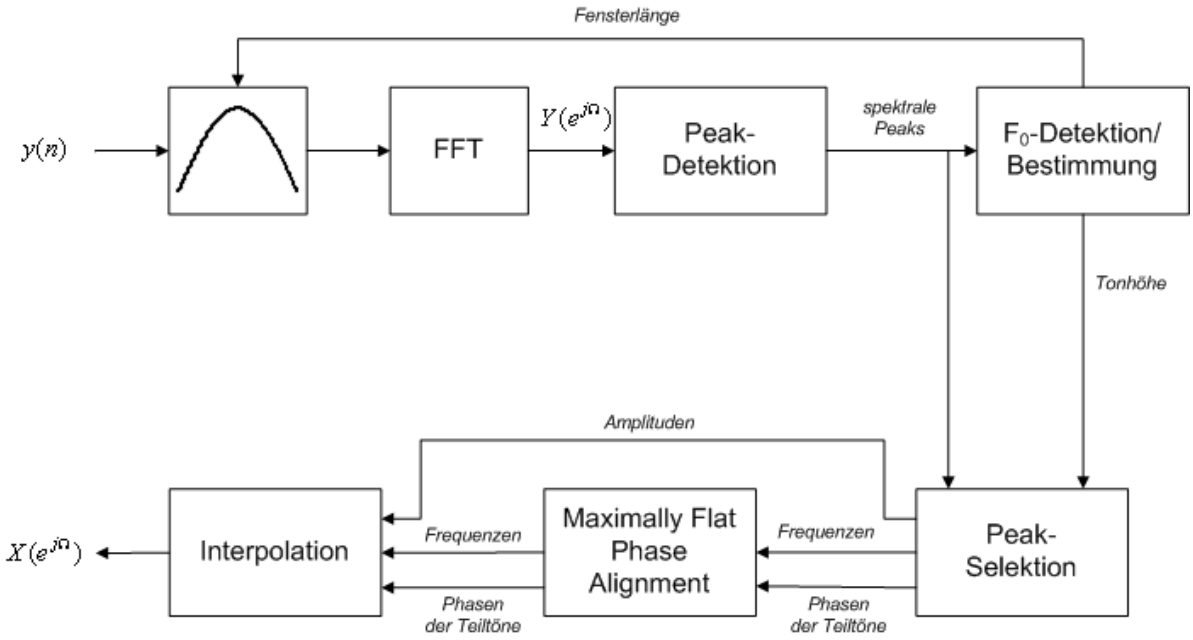


Abb. 5.5: Blockschaltbild des Algorithmus

5.4.2 Unwrapping der Phase

Vor der Interpolation der Phase muss diese entwickelt (*unwrapped*) werden, da die Phasen der Teiltöne modulo 2π vorliegen. Dabei ist die Phase ϕ_1 des ersten Teiltons der Initialwert:

$$\varphi_1 = \phi_1 \tag{5.7}$$

Die entwickelten Phasen der anderen Teiltöne werden iterativ gemäß

$$\varphi_r = \varphi_{r-1} + princarg(\phi_r - \phi_{r-1}) \tag{5.8}$$

berechnet. Die Unwrapping-Prozedur garantiert, dass bei der Interpolation der kürzeste Weg zwischen zwei aufeinander folgenden Phasen gewählt wird.

5.5 Residualkomponente

Bei der Schätzung von $X'(e^{j\Omega})$ wurden sämtliche Informationen des Spektrums des Eingangssignals zwischen den Teiltönen außer Acht gelassen. Diese Teile des Spektrums beschreiben allerdings inhärente Unregelmäßigkeiten sowie verschiedene andere Charakteristika der Stimme, wie z.B. Luftgeräusche etc. .
Bei den durchgeführten Transformationen wurde das Residualsignal zum resynthetisierten Ausgangssignal hinzuaddiert, um die Natürlichkeit des Ergebnisses zu verbessern.
Der beschriebene Algorithmus wird nur in stimmhaften Segmenten benutzt, da er von einem periodischen Eingangssignal ausgeht. Die stimmlosen Segmente werden vollständig durch das Residualsignal beschrieben.

5.6 Hörbeispiele

Hörbeispiel 10 (CD-Track 10): Männerstimme, bandbegrenzt ($f_{Grenz} = 5.5kHz$,$f_{Sampling} = 20.05kHz$)

Hörbeispiel 11 (CD-Track 11): Hörbeispiel 10 aus Glottispulsen und Residualsignal resynthetisiert

Hörbeispiel 12 (CD-Track 12): Aus Glottispulsen resynthetisierte Sinuskomponente des Signals aus Hörbeispiel 10

Hörbeispiel 13 (CD-Track 13): Residualsignal aus Hörbeispiel 10

Hörbeispiel 14 (CD-Track 14): Sprecher aus Hörbeispiel 10 künstlich gealtert

Literaturverzeichnis

Abe, M., Nakamura, S. Shikano, K. und Kuwabara, H. (1988): *Voice Conversion through vector quantisation*. Proc. of ICASSP 1988, New York, pp. 655-658

Amatriain, X., Bonada, J., Loscos, A, Serra, X.: *Spectral Processing*. In Zölzer, U. (Hrsg.): *DAFX: Digital Audio Effects*. John Wiley & Sons, Ltd. (2002)

Arslan, L.M. (1999): *Speaker Transformation Algorithm Using Segmental Codebooks*. Speech Communication 28, pp. 211-226

Atal, B.S.; Hanauer, S.L.: *Speech analysis and synthesis by linear prediction of the speech wave*. Journal of the Acoustical Society of America (JASA) 50(2, Part 2): 637-655 (1971)

Baudoin, G. und Stylianou, Y. (1996): *On the transformation of the speech spectrum for voive conversion*. Proc. of ICSLP 1996, Philadelphia, PA

Bonada, J.: *High quality voice transformations based on modelling radiated voice pulses in frequency domain*. Proc. of the 7th Int. Conf. on Digital Audio Effects (DAFX-04), Naples, Italy, October 5-8, 2004

Chappel, D.T. und J.H.L. Hansen (1998): *Speaker-Specific Pitch Contour Modeling and Modification*. Proceedings of the IEEE ICASSP, Seattle, Washington, May 1998, Vol. II, pp. 885-888

Eklund, I.; Traunmüller, H.: *Comparative Study of Male and Female Whispered and Phonated Versions of the Long Vowels of Swedish*. Phonetica 54: 1-21 (1997)

Endres, W.; Bambach, W.; Flösser, G.: *Voice spectograms as a function of age, voice disguise, and voice imitation*. JASA 49, No. 6(2): 1842-1849 (1971)

Fant, G.: *Non-uniform vowel normalisation*. Q.Prog.Status Rep., Speech Transm.Lab., R. Inst. Technol., Stockholm, No. 2/3, 1-19 (1975)

Friedrich, G.; Bigenzahn, W.: *Phoniatrie*. Verlag Hans Huber, Bern, Göttingen, Toronto Seattle (1994)

Fujisaki, H. und H. Kawai (1982): *Modeling the Dynamic Characteristics of Voice Fundamental Frequency with Applications to Analysis and Synthesis of Intonation*. Working Group on Intonation, 13th International Congress of Linguists, Tokyo

Fujisaki, H.; Yoshimune; Nakamura: *Formant frequencies of sustained vowels in Japanese obtained by analysis-by-synthesis of spectral envelopes*. University of Tokyo, unveröffentlicht (1970)

Gold, B. und L. R. Rabiner (1969): *Parallel Processing Techniques for Estimating Pitch periods of Speech in the Time Domain*. JASA, Vol. 46(2) Pt. 2, pp. 442-448

Griffin, D.W. und Lim, J.S. (1988): *Multi-Band Excitation Vocoder*. IEEE Transactions on acoustics, speech, and signal processing 36(8), pp. 1223-1235

Haberman, G.: *Functional aspects of the aging larynx*. HNO (Berlin) 20: 121-124 (1972)

Harris, F.J. (1978): *On the use of windows for harmonic analysis with the discrete Fourier transform*. Proceedings IEEE, vol. 66, pp. 51-83.

Hollien, H.; Michel, J.; Doherty, E.T.: *A method for analyzing vocal jitter in sustained phonation*. Journal of Phonetics 1: 85-91 (1973)

Hollien, H.; Shipp, T.: *Speaking fundamental frequency and chronologic age in males*. Journal of Speech and Hearing Research (JSHR) 15: 155-159 (1972)

Imai, S. und Y. Abe (1979): *Spectral envelope extraction by improved cepstral method*. Electron. and Commun. in Japan, vol. 62-A, no. 4, pp. 10–17

ITU-T Recommendation P.800 (1996): *Methods for subjective determination of transmission quality*

Kain, A.B. (2001): *High Resolution Voice Transformation*. PhD Dissertation, OGI School of Science and Engineering at Oregon Health and Science University

Kallail, K.J.; Emanuel, F.W.: *Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects*. JSHR 27, 245-251 (1984a)

Kallail, K.J.; Emanuel, F.W.: *An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects*. J. Phonet. 12, 175-186 (1984b)

Linville, S.E.; Fisher, H.B.: *Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females*. JASA 78: 40-48 (1985)

Loscos, A. und Bonada, J. (2004): *Emulating rough and growl voice in spectral domain*. Proc. of the 7th Int. Conf. on Digital Audio Effects, Naples, Italy, October 5-8

Loscos, A. (2003): *Issues on Modeling the Singing Voice*. Doctoral report, Universitat Pompeu Fabra, Barcelona

Maher, R.C., Beauchamp, J.W.: *Fundamental frequency estimation of musical signals using a two-way mismatch procedure*. JASA, 95(4), pp.2254-2263, (1994)

McAuley, R.J. und Quatieri, T.F.: *Speech analysis/synthesis based on a sinusoidal representation*. IEEE Transactions on Acoustics, Speech and Signal Processing, 34(4), p. 744-754 (1986)

Mizuno, H. und Abe, M. (1995): *Voice Conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt*. Speech Communication Vol. 16(2), pp. 153-164

Mouchtaris, A, Van der Spiegel, J und Mueller, P (2005): *Nonparallel Training for Voice Conversion Based on a Parameter Adaptation Approach*. IEEE Trans. on Speech and Audio Processing

Moulines, E. und F. Charpentier (1990): *Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones*. Speech Communication 9, pp. 453-467

Moulines, E. und Laroche, J. (1995): *Non-parametric techniques for pitch-scale modification of speech*. Speech Communication 16, pp. 175-205

Oppenheim, A.V. und Schafer, R.W.: *Zeitdiskrete Signalverarbeitung*. R. Oldenburg Verlag München Wien (1999)

Orlikoff, R.F.: *The relationship of age and cardiovascular health to certain acoustic characteristics of male voices*. JSHR 33: 450-457 (1990)

Orlikoff, R.F.; Baken, R.J.: *Consideration of the Relationship between the Fundamental Frequency of Phonation and Vocal Jitter*. Folia Phoniatrica 42: 31-40 (1990)

Peterson, G. H.; Barney, H.L.: *Control Methods Used in a Study of Vowels*. In Ronald W. Schafer, John D. Markel (Hrsg.): *Speech Analysis*, 1952, IEEE Press

Peterson, G.: *Parameters of Vowel Quality*. JSHR 4: 10-29 (1961)

Ptacek, P.H.; Sander, E.K.; Maloney, W.H.; Jackson, C.R.: *Phonatory and rela-*

ted changes with advance age. JSHR 9: 353-360 (1966)

Ramig, L.A.; Ringel, R.L.: *Effects of physiological aging on selected acoustic characteristics of voice*. JSHR 26: 22-30 (1983)

Röbel, A. und X. Rodet (2005): *Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation*. Proc. of the 8th Int. Conference on Digital Audio Effects (DAFx'05), Madrid, Spain, September 20-22

Rodet, X. und Depalle, Ph. (1992): *Spectral envelopes and inverse FFT synthesis*. Proc. 93. AES Convention, San Francisco, AES Preprint No. 3393

Serra, H. (1997): *Musical Sound Modeling With Sinusoids Plus Noise*. In C. Roads, S. Pope, A. Picialli, G. De Poli (Hrsg.): *Musical Signal Processing*, Swets & Zeitlinger

Shipp, T.; Hollien, H.: *Perception of the aging male voice*. JSHR 12: 703-710 (1969)

Shoup, J.E.; Lass, N.J.; Kuehn, D.P.: *Acoustics of Speech*. In Lass, Mc. Reynolds, Northern Yoder (Hrsg.): *Handbook of SPEECH-LANGUAGE PATHOLOGY AND AUDIOLOGY*, B.C. Becker Inc., Toronto, Philadelphia (1988)

Smith, J.O. III und Serra, X. (1987): *PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation*. Proc. of the Int. Computer Music Conf. (ICMC-87, Tokyo), Computer Music Association

Tanaka, K. und Abe, M. (1997): *A New Fundamental Frequency Modification Algorithm with Transformation of Spectrum Envelope According to f0*. ICASSP 1997, Vol. 2, pp. 951-954

Torrésani, B. (1999): *An Overview of Wavelet Analysis and Time-Frequency Analysis (A Minicourse)*. In V.B. Priezzhev and V.P. Spiridonov (eds.), *Self-Similar Systems*, Proceedings of the International Workshop 1998, JINR, E5-99-38, Dubna, 1999, pp. 9-34

Traunmüller, H.: *Paralinguistic Variation and Invariance in the Characteristic Frequencies of Vowels*. Phonetica 45: 1-29 (1988)

Traunmüller, H.; Branderud, P.; Bigestans, A.: *Paralinguistic speech signal transformations*. PERILUS No. X, Inst. Linguist., Univ. Stockh., 47-64 (1989)

Türk, O. (2003): *New Methods for Voice Conversion*. Master of Science thesis, Boğaziçi University

Türk, O. und L.M. Arslan (2002) *Subband Based Voice Conversion*. Proceedings

of the ICSLP 2002, Vol. 1, pp. 289-292, September 2002, Denver, Colorado, USA

Valbret, H., Moulines, E. und Tubach, J.P. (1992): *Voice transformation using PSOLA technique*. Speech Communication Vol. 11(2), pp. 175-187