Ronald Schulz

Stimmentransformationen

Diplomarbeit

vorgelegt am
Institut für Elektronische Musik und Akustik
an der Universität für Musik und darstellende Kunst in Graz
und der Technischen Universität Graz

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

November 2007

Kurzfassung

In dieser Diplomarbeit werden Algorithmen zur Transformation der menschlichen Stimme entwickelt. Dabei beschränkt sich die Betrachtung auf folgende Anwendungsgebiete: die Umwandlung des Stimmgeschlechts von weiblich zu männlich und umgekehrt, das künstliche Altern und Verjüngen einer beliebigen Stimme und die Veränderung des Stimmaufwandes („*vocal effort*“).

Anhand der wissenschaftlichen Literatur werden die wesentlichen Merkmale der einzelnen Transformationen herausgearbeitet. Bei diesen Transformationen ist keine Zielperson vorgegeben, nach der die modifizierte Stimme klingen soll. Das unterscheidet sie von der herkömmlichen *Voice Conversion*. Die Algorithmen arbeiten hauptsächlich im Zeitbereich und basieren auf der TD– PSOLA Methode. Sie werden in Matlab und PRAAT implementiert.

Zur Evaluierung der erarbeiteten Algorithmen wird ein Hörtest durchgeführt.

Abstract

In this diploma thesis algorithms for transformation of the human voice are developed. The following transformations are considered: changing the voice gender from male to female and conversely, artificial aging and rejuvenation of any voice and modification of the vocal effort. Based on the scientific literature, the essential characteristics of the transformations are identified. As opposed to „Voice Conversion“ there is no target voice, the transformed voice has to sound like. The algorithms work mainly in the time domain and are based on the TD– PSOLA method. The implementation is done in Matlab and PRAAT. To evaluate the developed algorithms a listening test is conducted.

Danksagung

Ich möchte mich bei all jenen bedanken, die durch ihre fachliche und persönliche Unterstützung zum Gelingen dieser Diplomarbeit beigetragen haben.

Insbesondere bedanke ich mich bei Herrn o.Univ.-Prof. Mag. DI Dr. Robert Höldrich und Herrn DI Markus Noisternig für die kompetente Betreuung. Darüberhinaus danke ich Herrn Univ. Ass. DI Dr. Alois Sontacchi für seine wertvollen Ratschläge meine Arbeit betreffend und den Kollegen aus dem ersten Stock für die zahlreichen Gespräche und gemeinsamen Kaffeepausen.

Außerdem möchte ich mich natürlich auch bei meinen Eltern für ihre Geduld und Unterstützung bedanken.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungen

AKF	Autokorrelationsfunktion
AMDF	Average Magnitude Difference Function
AR Prozeß	Autoregressiver Prozeß
CA	Chronologisches Alter
DFT	Discrete Fourier Transformation
FD– PSOLA	Frequency Domain– Pitch Synchronous Overlap and Add
FFT	Fast Fourier Transformation
FIR	Finite Impulse Response
GFB	Grundfrequenzbestimmung
HNM	Harmonic+ Noise Modell
IDFT	Inverse Discrete Fourier Transform
IFFT	Inverse Fast Fourier Transformation
LPC	Linear Predictive Coding
LP– PSOLA	Linear Predictive– Pitch Synchronous Overlap and Add
LSF	Line Spectral Frequency
LSP	Line Spectral Pair
PIF	Pol– Interaktions– Faktor
STFT	Short Time Fourier Transformation
SA	Stimmalter

SSD	Stimulus Sampling Discrimination
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum
TD– PSOLA	Time Domain– Pitch Synchronous Overlap and Add
WA	Wahrgenommenes Alter

Kapitel 1

Einleitung

Thema dieser Diplomarbeit ist die Transformation der menschlichen Stimme. Mit Stimmentransformationen können neue Stimmen kreiert werden, indem verschiedene Merkmale (*Features*) einer Stimme, wie Grundfrequenz, Formanten oder Sprechgeschwindigkeit modifiziert werden [16]. Es geht nicht, wie bei der *voice conversion*, um die Konvertierung der Stimme eines Sprechers in die eines anderen konkreten Sprechers [17, 18, 19].

Motivation. Transformationen der menschlichen Stimme sind für ein weites Feld von künstlerischen, wissenschaftlichen und industriellen Anwendungen nützlich. Eine Möglichkeit der Anwendung liegt im Bereich der verketteten Sprachsynthese (*concatenative speech synthesis*) [19]. Wenn zum Beispiel eine große Datenbank mit Sprachaufnahmen einer Sprecherin oder eines Sprechers erstellt wurde, dann ist es sehr aufwändig und teuer, für verschiedene Applikationen jeweils neue Sprachaufnahmen mit anderen SprecherInnen zu machen. Mit Stimmentransformationen kann die Datenbank an die gewünschte Anwendung angepasst werden. Weitere Anwendungsgebiete finden sich in der Film- und Spieleindustrie beim Editieren und Synchronisieren von Stimmen [20]. Die in dieser Arbeit entwickelten Algo-

rithmen arbeiten hauptsächlich im Zeitbereich und benötigen daher im Gegensatz zu Methoden im Frequenzbereich relativ wenig Rechenzeit. Dadurch lassen sie sich leichter in Echtzeitsysteme integrieren und sind somit auch für Live– Anwendungen im Bereich der Elektronischen Musik bzw. Neuen Musik interessant.

Gliederung der Diplomarbeit. Diese Arbeit lässt sich ziemlich genau in einen Theorie– und einen Praxisteil unterteilen. Auch wenn beides nicht eindeutig zu trennen ist, soll der theoretische Teil auf die wissenschaftlichen Grundlagen dieser Arbeit, die anatomischen und physiologischen Voraussetzungen und wichtige bereits entwickelte Methoden und Techniken der Sprachsignalverarbeitung verweisen.
Der Praxisteil beschreibt die Implementierung einiger der im ersten Teil vorgestellten Methoden und die Entwicklung entsprechender Algorithmen zur Stimmentransformation in Matlab und PRAAT[1]. Abgeschlossen wird dieser Teil mit der Erläuterung und Auswertung des durchgeführten Hörtests und einer Zusammenfassung der wichtigsten Resultate dieser Arbeit.

In Kapitel 2.1 werden zunächst anatomische und physiologische Grundlagen des menschlichen Sprechapparates erläutert. Danach folgt in Kapitel 2.2 ein kurzer Überblick über die akustische Phonetik mit der Beschreibung der wichtigsten Sprachlaute. Kapitel 2.3 stellt das weit verbreitete Quelle– Filter Modell der Spracherzeugung vor, welches auch Grundlage dieser Arbeit ist. Danach werden in Kapitel 2.4 zwei Modelle der Sprachsignalverarbeitung dargestellt.

Kapitel 3 beschreibt anhand der wissenschaftlichen Literatur Gemeinsamkeiten und Unterschiede von Frauen- und Männerstimmen, von Stimmen alter und jun-

[1]PRAAT ist ein bekanntes Sprachanalyse- und Syntheseprogramm [14].

ger Menschen und von Stimmen, die mit viel und wenig Stimmaufwand produziert werden.

Die im Kapitel 3 herausgearbeiteten akustischen Merkmale von Stimmen, werden in den Kapiteln 4 und 5 als für die Transformationen wichtige Parameter näher beleuchtet. Dabei stehen Methoden der Schätzung und Modifikation von Grundfrequenz und Formanten im Vordergrund.

Der Praxisteil beginnt in Kapitel 6 mit der Beschreibung des implementierten Systems zur Stimmentransformation. Danach folgt die Darstellung des durchgeführten Hörtests in Kapitel 7. Abschliessend gibt es eine kurze Zusammenfassung und einen Ausblick in Kapitel 8. Im Anhang A befinden sich einige Abbildungen und Tabellen, die im Zuge der Auswertung des Hörtests entstanden sind.

Teil I

Theorie

Kapitel 2

Grundlagen

Modelle des menschlichen Sprechapparates bilden die Grundlage für Transformationen der Stimme. Diese mathematischen Modelle beruhen auf den anatomischen und physiologischen Eigenschaften des Sprechapparates, welche am Anfang dieses Kapitels erläutert werden. Danach werden kurz die wichtigsten Typen von Phonemen vorgestellt, um einen Überblick über die Struktur der menschlichen Sprachlaute zu geben. Die akustische Theorie der Sprachproduktion gibt eine mathematische Beschreibung der Wellenausbreitung in den Sprechorganen ab. Der Vokaltrakt wird hierbei als akustisches Rohr modelliert. Für eine genaue Beschreibung dieser Theorie siehe [21, 1, 8]. Dieses sogenannte Röhrenmodell bildet die Grundlage für das klassische Quelle– Filter Modell der Spracherzeugung. Dabei wird das Sprachsignal als Ausgang eines zeitvarianten linearen Systems aufgefasst, welches entweder mit einem Rauschen oder mit einem quasiperiodischen Impulszug angeregt wird. Das Quelle– Filter Modell ist Thema in Abschnitt 2.3. Es dient vielen Analyse– Synthese Systemen der digitalen Sprachsignalverarbeitung als Grundlage, wie zum Beispiel STRAIGHT, welches in Abschnitt 2.4.2 vorgestellt wird. Abschnitt 2.4.1 beschreibt ein anderes populäres System der Sprachsignalverarbeitung, das *Harmonic + Noise* Modell.

2.1 Anatomische und physiologische Grundlagen des Sprechapparates

Die bei der Sprachproduktion verwendeten Organe, zu sehen in Abbildung 2.1, sind in erster Linie für die Atmung und Nahrungsaufnahme zuständig und erst in sekundärer Hinsicht für das Sprechen [1].

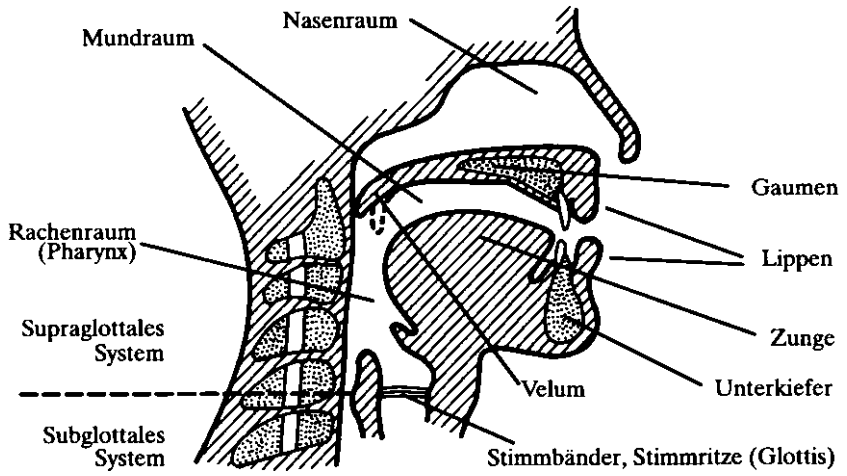

Abbildung 2.1: Bau des menschlichen Sprechapparates [1].

Das Sprachsignal entsteht in zwei Schritten: erstens der Anregung und zweitens der Filterung (vgl. Quelle– Filter Modell Abschnitt 2.3). An der Sprachproduktion sind hauptsächlich die Lungen, der Kehlkopf (*Larynx*) mit den Stimmbändern und der Stimmritze (*Glottis*) und der Mund- und Rachenraum (Vokaltrakt) beteiligt. Beim Sprechen strömt Luft aus der Lunge durch den Kehlkopf. Die Stimmerzeugung im Kehlkopf (siehe Abbildung 2.2), auch Phonation genannt, kann mit der myoelastisch- aerodynamischen Theorie der Stimmerzeugung erklärt werden. Danach schwingen die Stimmlippen bei der Produktion von Vokalen

> „durch das Wechselspiel zwischen subglottischem Druckanstieg und Einstellung der Kehlkopfmuskeln. Das periodische Wechselspiel dieser Kräf-

te erzeugt regelmäßige Stimmlippenschwingungen, die zur Erzeugung des Stimmklanges führen.“ [2]

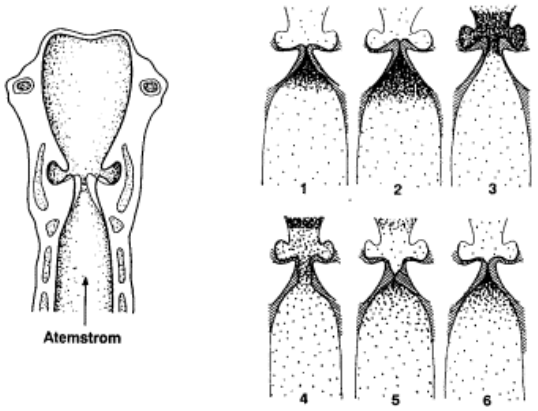

Abbildung 2.2: Stimmerzeugung im Kehlkopf. Der austretende Atemstrom drückt gegen die geschlossene Glottis (1, 2) und führt zu einem subglottischen Druckanstieg, durch welchen die Glottis geöffnet wird (3, 4). Demzufolge fällt der subglottische Druck ab und myoelastische und aerodynamische Kräfte leiten den Stimmlippenschluß ein (5, 6). Dieser Zyklus wiederholt sich und es kommt zu periodischen Verdichtungen und Verdünnungen des Ausatemstromes [2].

Aufgrund des abrupten Schließens der Glottis entsteht im Phonationssignal ein Knick, der ausschlaggebend für das Vorhandensein von Frequenzen bis hin zu mehreren Kilohertz bei stimmhaften Lauten ist. Das Anregungssignal durchläuft in weiterer Folge den Rachen- und Mundraum (Vokaltrakt), wo es entsprechend der aktuellen Form des Vokaltraktes gefiltert und dann vom Mund und/ oder der Nase abgestrahlt wird.

2.2 Grundlagen der akustischen Phonetik

Die akustische Phonetik, ein Zweig der Linguistik, beschäftigt sich mit den akustischen Eigenschaften der Sprachlaute (Phoneme). Jedes Phonem[1] kann auf unterschiedliche Weise artikuliert werden und die Artikulation liefert bei verschiedenen Vokaltrakten auch unterschiedliche klangliche Ergebnisse. Deshalb ist das Sprachsignal ein und desselben Phonems sehr variabel [8]. Dieser Abschnitt befasst sich mit den allgemeinen akustischen Eigenschaften verschiedener Phoneme.
Sprachlaute werden in Vokale und vokalähnliche Laute (Diphthonge) und in Konsonanten eingeteilt. Die Grenzen zwischen diesen Kategorien sind fließend. Laute, wie zum Beispiel „l“ oder „r“ sind je nach Sprache der einen, der anderen oder auch beiden Kategorien zuzuordnen [1].

Vokale. Vokale entstehen durch eine stationäre und stimmhafte Anregung des Vokaltrakts. Das Sprachsignal ist folglich quasi– periodisch. Die meiste Energie steckt dabei in den Frequenzen unterhalb von 1 kHz [8]. Es fehlen Verschlüsse und wesentliche Engstellen im Ansatzrohr. Die Schallabstrahlung erfolgt über die Mundöffnung (bei Nasalen auch zusätzlich über die Nase). Vokale sind die Phoneme mit der größten Intensität und dauern beim normalen Sprechen zwischen 50 und 400 ms. Sie lassen sich gut durch die ersten drei Formantfrequenzen unterscheiden. Eine klassische Art der Darstellung von Vokalen aufgrund akustischer Eigenschaften ist die Formantkarte. Dabei werden die beiden ersten Formantfrequenzen (F1 und F2) der Vokale gegeneinander aufgetragen. Abbildung 2.3 zeigt eine Formantkarte für deutsche Vokale.

[1] Das Phonem ist die kleinste bedeutungsunterscheidende Einheit einer Sprache, z. B. unterscheiden sich die beiden Wörter *Tisch* und *Fisch* nur durch ein Phonem. Im Deutschen gibt es in etwa 50 Phoneme [1, S. 40].

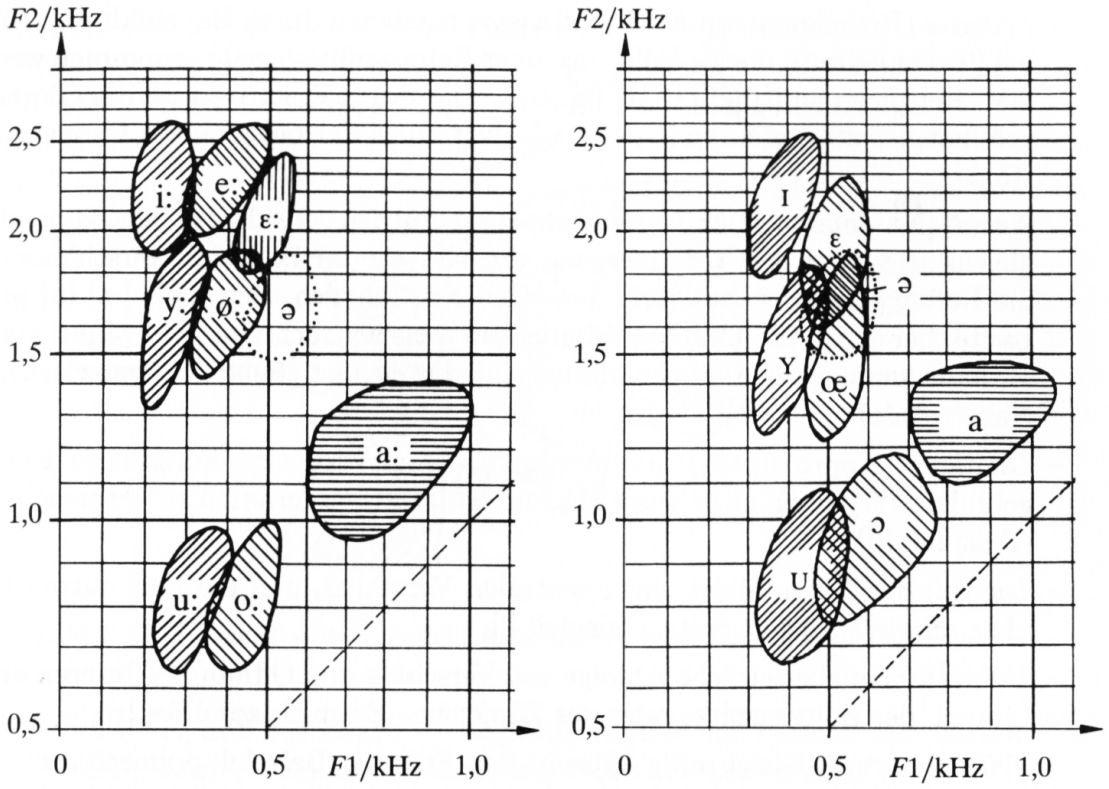

Abbildung 2.3: Formantkarte der deutschen Vokale. Langvokale (links), Kurzvokale (rechts); 16 Sprecherinnen und Sprecher [1, S. 47].

Diphthonge. Ein Diphthong gehört zu den vokalähnlichen Lauten. Es ist ein Doppellaut oder Zwielaut aus zwei Vokalen, wie zum Beispiel „ei“, „au“ oder „eu“ [22].

Konsonanten. Laut Vary et al. [1] erfolgt die Abstrahlung bei Konsonanten nicht unbedingt über den Mund. Man spricht dann von Nasalen. Die Anregung ist auch nicht (rein) stimmhaft oder es gibt wesentliche Engstellen oder Verschlüsse im Vokaltrakt [1]. Die rauschhafte Anregung regt nicht den ganzen Vokaltrakt

an, sondern nur den Bereich vor den Engstellen. Dadurch werden tiefe Frequenzen abgeschwächt [8]. Konsonanten können durch die Art ihrer Artikulation unterschieden werden [1]. Es gibt zum Beispiel Plosive (Verschlusslaute), Frikative (Reibelaute), Nasale, Gleitlaute, Laterale, Vibranten und Frikativvibranten. Eine genaue Beschreibung dieser Konsonanten findet sich in Vary et al. [1, S. 47].

2.3 Lineares Modell der Spracherzeugung

Grundsätzlich besteht die menschliche Sprache, wie in Abschnitt 2.2 beschrieben, aus einer Vielfalt von Lauten. Die statistischen Eigenschaften der Sprache ändern sich beim Sprechen ständig über die Zeit. Die Änderungen vollziehen sich jedoch meist relativ langsam. In Zeitabschnitten von 10 bis 30 ms kann man daher von einem quasi– stationären Signal ausgehen, welches annähernd als Ausgang eines linearen zeitinvarianten Systems dargestellt werden kann.
Das führt zum linearen Modell der Spracherzeugung, dem sogenannten Quelle– Filter Modell[2]. Es versucht, die menschliche Spracherzeugung nachzubilden. In [1, 8] oder [21] sind genaue Beschreibungen dieses Modells zu finden.
In der Praxis reicht es meist, das vereinfachte lineare Quelle– Filter Modell der Spracherzeugung zu betrachten. Es beruht auf einem rein rekursiven Filter, der also nur Polstellen besitzt. Man verwendet deshalb auch den Begriff Auto– Regressiver (AR) Prozeß. Für die Modellierung des Nasal– Traktes braucht man einen Pol– Nullstellen Filter. Zur Vereinfachung kann aber die Modellierung des Nasal– Trakts, der Glottis und der Lippen aus dem ursprünglichen Modell für Anwendungen in der Sprachsynthese vernachlässigt werden, ohne dass es zu nennenswerten Qualitätseinbußen kommt [1, S. 167]. Das lässt sich folgendermaßen begründen:

[2]Das Quelle– Filter Modell wird in der englischsprachigen Literatur auch als *terminal– analog model* bezeichnet [8].

Bei der Übertragungsfunktion $H(z)$ eines stabilen und kausalen Pol– Nullstellen Filters liegen sämtliche Polstellen innerhalb des Einheitskreises. Die Nullstellen können aber auch außerhalb des Einheitskreises liegen (vgl. Abbildung 2.4). In

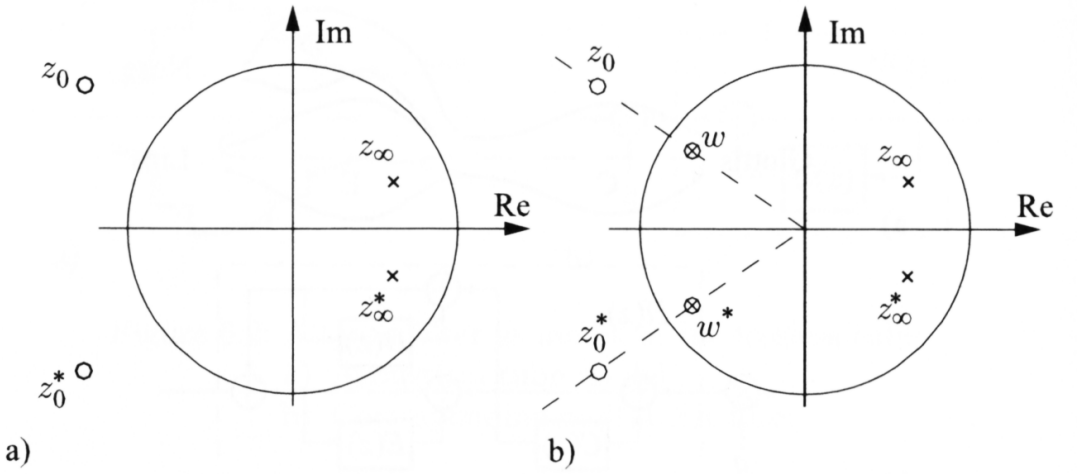

Abbildung 2.4: Pol– Nullstellen Diagramm. (a) originale Übertragungsfunktion $H(z)$. (b) Aufspaltung in minimalphasigen Filter und Allpass. o – Nullstellen, x – Polstellen [3].

diesem Fall lässt sich $H(z)$ in einen minimalphasigen Anteil $H_{min}(z)$ und eine Allpass- Übertragungsfunktion $H_{Ap}(z)$ laut

$$H(z) = H_{min}(z)H_{Ap}(z) \tag{2.1}$$

zerlegen. Zuerst werden die Nullstellen z_0 und z_0^*, die außerhalb des Einheitskreises liegen, nach innen gespiegelt, und durch Polstellen in gleicher Lage kompensiert. Zum minimalphasigen Anteil gehören dann die beiden Nullstellen

$$\begin{aligned} w &= \frac{1}{z_0^*} \\ w^* &= \frac{1}{z_0} \end{aligned} \tag{2.2}$$

entsprechend schreibt sich $H_{min}(z)$:

$$H_{min}(z) = \frac{(z-w)(z-w^*)}{(z-z_\infty)(z-z_0^*)}. \tag{2.3}$$

Die Übertragungsfunktion des Allpass setzt sich aus den außerhalb des Einheitskreises liegenden Nullstellen und den beiden Polen w und w^* zusammen:

$$H_{Ap}(z) = \frac{(z-z_0)(z-z_0^*)}{(z-w)(z-w^*)}. \tag{2.4}$$

Für die Sprachsynthese ist es ausreichend, den minimalphasigen Anteil zu realisieren, weil unser Ohr gegenüber den durch den Allpass modellierten Phasenänderungen weitgehend unempfindlich ist. Da nun sämtliche Pole und Nullstellen des minimalphasigen Filters innerhalb des Einheitskreises liegen, gibt es auch einen stabilen inversen Filter.

$$H_{min}^{-1}(z) = \frac{1}{H_{min}(z)} \tag{2.5}$$

Durch Anwendung dieses inversen Vokaltraktfilters $H_{min}^{-1}(z)$ auf das Sprachsignal lässt sich das Anregungssignal aus dem Gesamtsignal extrahieren. Darüber hinaus kann jedes minimalphasige Pol– Nullstellen Filter mit einem Allpol– Filter m– ten Grades angenähert werden. Damit erklärt sich die Verwendung eines Allpol– Filters bei der Sprachsynthese [1, S. 167 ff.]. Das vereinfachte lineare Quelle– Filter Modell ist in Abbildung 2.5 abgebildet.

Aus den vorhergehenden Betrachtungen ergeben sich die zwei primären funktionellen Verarbeitungsstufen des linearen Modells der Spracherzeugung: die Quelle und der Filter.
Die Quelle entspricht laut Abbildung 2.5 dem Anregungssignal v(k), welches der Anregung des menschlichen Vokaltraktes ähnelt. Bei stimmhaften Lauten erzeugen die Stimmbänder eine quasiperiodische Schwingung. Bei stimmlosen Lauten hingegen enstehen an Engstellen im Mund- oder Rachenraum bei geöffneten Stimm-

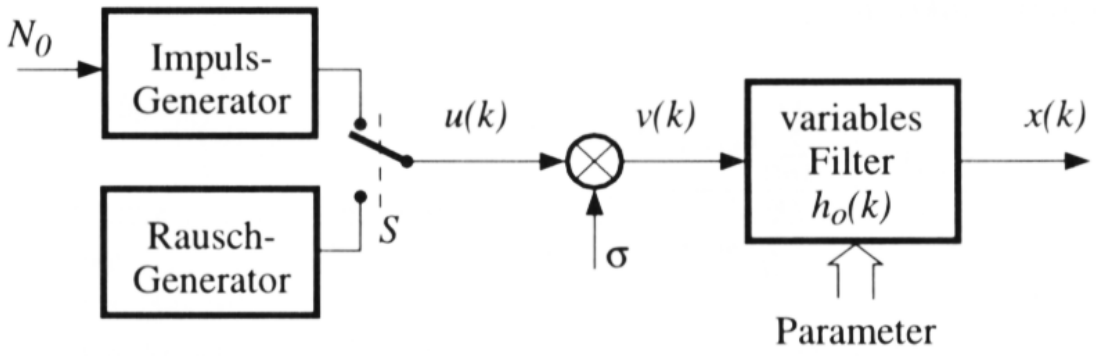

Abbildung 2.5: Vereinfachtes lineares Quelle– Filter Modell [1].

bändern turbulente Strömungen, ein Rauschsignal. Die quasiperiodische Schwingung und das Rauschen besitzen ein flaches Spektrum und dienen als Anregungssignale für die Spracherzeugung [1]. Folglich sind sie die Eingangssignale für den rein rekursiven, auto- regressiven (AR) digitalen Filter H(z), der die Resonanzen des Vokaltrakts modelliert. Sämtliche Modellparameter, wie Grundfrequenz, stimmhaft/stimmlos– Unterscheidung, Verstärkungsfaktor (*Gain*) und Resonanzfrequenzen und Bandbreiten sind zeitvariant, da sie sich beim Sprechen ständig ändern. Wie oben bereits erwähnt, sind jedoch beim Betrachten von kurzen Signalabschnitten (10 bis 30 ms) die Voraussetzungen für die Annahme von Linearität und Zeitinvarianz des Systems gegeben.

2.4 Modelle der digitalen Sprachsignalverarbeitung

Für die Repräsentation von Sprachsignalen in der digitalen Domäne gibt es verschiedene weit verbreitete Ansätze. In der Literatur unterscheidet man parametrische und nicht– parametrische Modelle, je nachdem ob Sprachparameter explizit ermittelt und für die Synthese verwendet werden oder nicht [11]. In diesem Kapitel werden zwei parametrische Sprachmodelle, das Harmonic+ Noise Modell und

STRAIGHT, vorgestellt.

2.4.1 Harmonic + Noise Modell

Das Harmonic + Noise Modell (HNM) wurde in den 1980er Jahren von Griffin und Lim [23] zunächst für Sprachcodierungsanwendungen entwickelt. In den nachfolgenden Jahren haben sich viele Forscher [11] mit diesem Modell beschäftigt und es auch in Hinblick auf Sprachmodifikationen erweitert. Es ist eine Weiterentwicklung des einfachen Sinus Modells, in dem die zeitvarianten spektralen Charakteristika des Sprachsignals als Summe von zeitvarianten Sinusschwingungen modelliert werden [10].

Beim HNM wird das Signal in eine deterministische Komponente und in eine stochastische Komponente zerlegt. Dabei besteht der deterministische Anteil aus Sinusschwingungen. Der stochastische Anteil modelliert einerseits die Schwankungen der Periode in stimmhaften Segmenten und andererseits die rauschhaften Anregungsgeräusche in stimmlosen oder stimmhaft/ stimmlos– gemischten Segmenten [11].

Das Spektrum wird in ein unteres und in ein oberes Frequenzband unterteilt. Die „*maximum voiced frequency*“ gibt an, wo diese Trennung passiert. Das untere Band wird nur durch Sinusschwingungen mit ihren Harmonischen und das obere Band durch ein mit einer Amplitudeneinhüllenden moduliertes Rauschen dargestellt [20]. Das Eingangssignal $s(t)$ wird als Summe $\hat{s}(t)$ von harmonischen Komponenten und einem stochastischen Anteil nachgebildet [24].

$$\hat{s}(t) = \sum_{k=-K(t)}^{K(t)} A_k(t) exp(jkt\omega_0(t)) + e(t) \tag{2.6}$$

$A_k(t)$ ist die komplexe harmonische Amplitude zum Zeitpunkt t, ω_0 die Grundfrequenz und e(t) die stochastische Komponente. Diese Parameter werden zu den Zeitpunkten t_i aktualisiert.

Die Grundfrequenz ω_0 wird in jedem Zeitfenster $[t_i - N/2, t_i + N/2]$ konstant gehalten, wobei N die Länge des Zeitfensters ist. Die komplexen Amplituden $A_k(t)$ sind Funktionen der Zeit.

$$A_k(t) = a_k(t_i) + (t - t_i)b_k(t_i) \tag{2.7}$$

$a_k(t_i)$ bezeichnet die ursprüngliche komplexe Amplitude, also die Amplitude und Phase der k'ten Harmonischen im Zeitfenster i. $b_k(t_i)$ ist die komplexe Neigung der Harmonischen und repräsentiert pseudolineare Variationen der Amplitude der Harmonischen und kleine Anpassungsfehler der Momentanfrequenz. $K(t)$ steht für die im deterministischen Anteil enthaltene zeitvariante Anzahl der Harmonischen der Grundfrequenz. Der stochastische Anteil $e(t)$ wird als ein durch Filterung von weißem Rauschem $n(t)$ mit einem zeitvarianten, normalisierten Allpol– Filter $A(t, Z)$ gewonnenes Signal angesehen. Das Ergebnis dieser Filterung wird mit einer Einhüllenden– Funktion $\omega(t)$ multipliziert, um die sich zeitlich ändernde Verteilung der Energie des stochastischen Anteils zu berücksichtigen [25].

$$e(t) = \omega(t)[A(t, Z) * e(t)] \tag{2.8}$$

Der Filter $A(t, Z)$ wird zu jedem Zeitpunkt t_i ausgewertet und auf einer Sample zu Sample Basis interpoliert [24].
Der Vorteil dieses parametrischen Sprachmodells liegt in der einfachen Handhabung von Grundfrequenz- und Zeitdauermodifikationen, da nur die einzelnen Parameter verändert werden müssen.

2.4.2 STRAIGHT

Ein anderes parametrisches Analyse– Synthese Verfahren für die Modellierung und Manipulation von Sprachsignalen ist das von H. Kawahara [26] entwickelte STRAIGHT (*Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum*). STRAIGHT wurde ursprünglich für die Erforschung

der Sprachwahrnehmung verwendet. Prinzipiell ist STRAIGHT ein Grundfrequenz–synchroner *Channel Vocoder* [27]. Als Grundlage dient das Quelle– Filter Modell der Spracherzeugung. Das Sprachsignal wird in Quelle und Filter zerlegt [28]. Ein großes Problem dabei sind Interaktionen zwischen Quellcharakteristika und zeitlich feinen Strukturen, wie Periodizität. Kawahara identifiziert drei Probleme, die es zu lösen gilt, um Sprachsignale mit hoher Qualität selbst bis in extreme Bereiche zu modifizieren. Ziel ist es danach, erstens Periodizitäteneffekte in der spektralen Analyse zu verhindern, zweitens eine zuverlässige Bestimmung der Grundfrequenz duchzuführen und drittens bei der Synthese eine genaue Kontrolle der Phaseninformation zu haben. Diese Ziele werden mit einer Kombination aus Fourier– und Wavelet Analyse Techniken erreicht.

Kapitel 3

Stimmentransformationen

Wie im ersten Kapitel ausgeführt, gibt es verschiedene Modelle der Spracherzeugung mit unterschiedlichen Vor- und Nachteilen. Mit Hilfe dieser Modelle ist es möglich, eine menschliche Stimme zu synthetisieren oder die Eigenschaften einer vorhandenen „echten“ Stimme zu verändern.

Um das Geschlecht und Alter einer menschlichen Stimme oder den Stimmaufwand zu modifizieren, müssen die dementsprechenden Merkmale von weiblichen und männlichen Stimmen extrahiert werden. Im folgenden Kapitel werden die spezifischen Charakteristika der zu verändernden Stimmgruppen erarbeitet. Zuerst geht es um den Unterschied von Frauen- und Männerstimmen, danach um die Eigenschaften der Stimme von jungen und alten Menschen und im letzten Teil dieses Kapitels um den Stimmaufwand.

3.1 Stimmgeschlecht

Die Unterschiede von Frauen- und Männerstimmen beruhen auf mehreren Faktoren. Laut [29] gibt es drei Typen von Parametern: Physiologische und akustische, welche objektiv gemessen werden können und perzeptive Parameter, die subjektiv

und somit nur mit Hilfe von psychoakustischen Versuchen bestimmbar sind.

Physiologische Unterschiede. Die physiologischen Unterschiede zwischen Frauen und Männern können einer Vielzahl dementsprechender Studien entnommen werden. Fant [30] zeigt, dass das Verhältnis von der Länge des weiblichen Vokaltraktes zum männlichen 0.87 beträgt. In [31] wird anhand von Messungen der Stimmlippen ein entsprechendes weiblich– männlich– Verhältnis von 0.8 festgestellt. Auch der Kehlkopf beider Geschlechter ist in vielerlei Hinsicht voneinander verschieden (genaueres siehe [32, 33]).
Boersma (zitiert in [34]) untersuchte das Wachstum von Mädchen und Knaben im Alter von 4 bis 14 Jahren und stellte fest, dass beide Geschlechter unterschiedlich schnell wachsen. Andere Untersuchungen bestätigen diese Ergebnisse. Merow und Broadbent (zitiert in [34]) fanden heraus, dass das Wachstum des männlichen Gesichts erst mit ungefähr 18 Jahren abgeschlossen ist. Dabei gibt es einen Entwicklungssprung in der Pubertät im Alter von 12 bis 14 Jahren. Bei den Mädchen ist die Entwicklung bereits im Alter von 14 Jahren abgeschlossen, mit einem Entwicklungssprung bei 10 bis 12 Jahren.

Akustische Korrelate. Aufgrund der physiologischen Unterschiede ergeben sich auch Unterschiede in der akustischen Domäne.
Die Grundfrequenz F0 ist eine sehr wichtige Größe zur Unterscheidung von Frauen- und Männerstimmen. Sie bewegt sich bei Frauen und Männern in signifikant unterschiedlichen Bereichen [35]. Peterson und Barney (zitiert in [36]) fanden heraus, dass die mittlere Grundfrequenz bei Frauen etwa 1.7 mal so groß ist wie die bei Männern. Der weibliche Vokaltrakt ist im Vergleich zum männlichen ungefähr 15 Prozent kürzer (Goldstein, zitiert in [36]). Die meisten Unterschiede finden sich aber beim *Pharynx* (Rachen). In [36] wurde gezeigt, dass die Unterschiede der For-

mantfrequenzen bei Männern und Frauen vom jeweiligen Vokal abhängig sind. Die Grundfrequenz von Frauen umfasst eine größere Spannweite als die von Männern. Nach Hollien und Shipp (zitiert in [29]) bewegten sich die gemessenen Grundfrequenzen von Männern im Bereich von 112 bis 146 Hz und die von Frauen laut Stoicheff [37] im Bereich von 170 bis 275 Hz. Andere wissenschaftliche Arbeiten zeigen einen Skalierungsfaktor von 1.6 entsprechend der unterschiedlichen Länge der weiblichen und männlichen Stimmlippen [33] oder stellen fest, dass die weibliche Grundfrequenz um etwa eine Oktave höher liegt als die männliche [37].
Neben der Grundfrequenz sind die Formanten wesentliche Parameter bei der Unterscheidung von Frauen- und Männerstimmen. Sie sind durch Frequenz, Amplitude und Bandbreite gekennzeichnet. Formantfrequenzen reflektieren die Größe und Form des menschlichen Vokaltrakts. So haben Untersuchungen gezeigt [29], dass die durchschnittlichen weiblichen und männlichen Formantfrequenzen über einen Skalierungsfaktor miteinander verbunden sind, der in etwa indirekt proportional zur Länge des Vokaltraktes ist. Frauen haben somit im Vergleich zu Männern etwa um 20 Prozent höhere Formantfrequenzen. Andere Literatur zeigt die Auswirkungen der unterschiedlichen Formen des Mund- und Rachenraumes auf die Bildung der Formantfrequenzen [34]. Huber [34] stellte auch fest, dass die untersuchten ersten drei Formanten bei weiblichen Kindern und Erwachsenen beiderlei Geschlechts ab 14 Jahren signifikant höher liegen als bei den männlichen Probanden. Sie untersuchte Kinder ab 4 Jahren und fand erst nach der Pubertät signifikante Unterschiede zwischen weiblichen und männlichen Probanden bei den Formanten.
Betreffend der Amplituden der Formanten existieren widersprüchliche Studien. Stathopoulos und Sapienza (zitiert in [34]) fanden bei Kindern und Frauen größere Öffnungen der Glottis sowie eine stärkere Kopplung zwischen Vokaltrakt und Anregung als bei Männern. Aufgrund der damit verbundenen stärkeren Dämpfung würden die Formantamplituden bei Kindern und Frauen geringer sein als die bei

Männern. Dagegen konnten Peterson und Barney (zitiert in [34]) keine Unterschiede der Formantamplituden zwischen Frauen und Männern feststellen. Sie hielten aber fest, dass zwischen verschiedenen SprecherInnen sehr große Variabilitäten vorhanden sind.

Vowel	F_1	F_2	F_3	B_1	B_2	B_3
			Male talker			
/a/	730	1090	2440	98	106	139
/i/	270	2290	3010	86	135	152
/u/	300	870	2240	87	101	134
			Female talker			
/a/	850	1220	2810	100	109	147
/i/	310	2790	3310	87	147	159
/u/	370	950	2670	89	103	144
			Children			
/a/	1030	1370	3170	105	113	156
/i/	370	3200	3730	89	157	170
/u/	430	1170	3260	90	108	158

Abbildung 3.1: Formantfrequenzen (Petersen und Barney, 1952) und Bandbreiten (Mannell, 1983) in Hertz von Männern, Frauen und Kindern von drei verschiedenen Vokalen [4].

Perzeptive Parameter. Betreffend der perzeptiven Parameter zur Stimmgeschlechtsunterscheidung, gibt es nur sehr wenige Angaben in der Literatur. Singh und Murry (1978) und Murry und Singh (1980) (zitiert in [29]) kamen zu dem Schluss, dass HörerInnen Informationen, die in der Grundfrequenz und in der Formantstruktur stecken, zum Erkennen von Frauen- und Männerstimmen nutzen. Die perzeptiven Parameter Stimmaufwand, Grundfrequenz und Nasalität werden für die Bestimmung von Frauenstimmen und Stimmaufwand, Grundfrequenz und Heiserkeit für die Bestimmung von Männerstimmen herangezogen. Singh und Murry gehen von unterschiedlichen perzeptiven Strategien der HörerInnen aus, je nachdem, ob sie weibliche oder männliche Stimmen klassifizieren.

Einige Forscher fanden heraus, dass melodische Eigenschaften, wie Intonation, Betonung und Koartikulation[1], Charakteristika der weiblichen Stimme sind [29]. Laut [39] basiert die Wahrnehmung des Geschlechts einer Stimme hauptsächlich auf der Grundfrequenz und zu einem geringeren Teil auf den höheren Formanten.
Andere Unterschiede im Stimmgeschlecht, wie eine hauchige Stimme (*breathy voice*), eine dynamischere Intonationskurve für Frauen oder unterschiedliche Dialekte bei Frauen und Männern scheinen angelernte Verhaltensweisen zu sein [36, S. 825].

	F0 Durchschnitt (Hz)	F0 Minimum (Hz)	F0 Maximum (Hz)
Männer	125	80	200
Frauen	225	150	350
Kinder	300	200	500

Tabelle 3.1: Mittlere, kleinste und größte Grundfrequenz für Männer, Frauen und Kinder [15].

3.2 Alterung der Stimme

Bereits in den 1960er Jahren wurden Untersuchungen zum Thema Alterung der Stimme unternommen. Damals wollte man herausfinden, ob es möglich sei, aus der Stimme und der Art des Sprechens auf das Alter der Person zu schließen. Verschiedene Experimente haben dies bestätigt. Unter anderem konnten Probanden in einer Untersuchung von Ptacek und Sander [40] mit großer Genauigkeit Sprachproben einer jungen Sprechergruppe (Durchschnittsalter 21 Jahre) bzw. ei-

[1] Als Koartikulation (lat. coarticulare 'zusammen artikulieren') wird die Beeinflussung eines Lautes durch den lautlichen Kontext bezeichnet. In der Phonetik ist Koartikulation die Bezeichnung für parallel verlaufende antizipierende (vorwegnehmende) Bewegungen bei der Artikulation. Diese Antizipation geschieht, in dem sich die Artikulatoren (zum Beispiel Zunge oder Lippen) während der Bildung eines Lautes bereits in die Stellung des folgenden Lautes begeben.[38]

ner alten Sprechergruppe (Durchschnittsalter 75 Jahre) zuordnen. Auch jüngere Studien, wie die von Traunmüller [41] stellen fest, dass Hörer das chronologische Alter eines Menschen aus dem Klang der Stimme ziemlich genau erkennen können.

Definitionen von Alter. In der wissenschaftlichen Literatur existieren unterschiedliche Begriffsbestimmungen von Alter, die hier kurz vorgestellt werden. Eine weitverbreitete Definition ist das chronologische Alter (CA) (*chronological age*) oder auch Kalenderalter (*calendar age*) genannt. CA ist die Zeit von der Geburt bis zum jetzigen Zeitpunkt. Für WissenschaftlerInnen, die sich mit dem Alter von SprecherInnen befassen (*speaker age*), sind zwei Definitionen wichtig, und zwar das wahrgenommene Alter (WA)(*perceived age*) und das Stimmalter (SA)(*vocal age*). Das WA ist das von HörerInnen subjektiv wahrgenommene, geschätzte Alter eines Sprechers. In Experimenten wird meist der mittlere Schätzwert von Testpersonen als Maß für das wahrgenommene Alter eines Sprechers verwendet [42, S. 8].
Das Stimmalter ist laut [43, S. 10] eine Zustandsbeschreibung des Prozesses der langfristigen Veränderung des Sprechapparates. Diese Änderungen sind im akustischen Sprachsignal zu beobachten und beruhen hauptsächlich auf physiologischen Veränderungen des Sprechapparates.
In dieser Arbeit werden beide Definitionen von Alter, das wahrgenommene Alter und das Stimmalter synonym verwendet. Wichtig ist die Abgrenzung beider Begriffe zum chronologischen Alter.

Anatomische und physiologische Veränderungen im Alter. Die Veränderungen der Stimme im Laufe der Jahre haben anatomische und physiologische Ursachen. So verringert sich mit zunehmendem Alter durch die schwindende Elastizität des Lungengewebes die Lungenkapazität und der Brustkorb wird steifer.

Dadurch kommt es zu erschwertem häufigerem Atmen. Der subglottale Druck ist nicht mehr so groß wie in jungen Jahren. Weiterhin kommt es mit den Jahren zu einer Verknöcherung (Ossifikation) der Kehlkopfknorpel, die Stimmlippen erschlaffen und verlieren ihre Elastizität. Das beeinträchtigt das Schwingungsverhalten der Stimmlippen und führt perzeptiv gesehen zu einem rauen Klangeindruck. Es ist auch möglich, dass sich die Stimmritze (*Glottis*) nicht mehr vollständig schliesst, was zu einer rauschhaften hauchigen Stimme (*breathy voice*) führt [5]. Durch das Absenken des Kehlkopfes (*Larynx*) im Alter vergrößert sich der Resonanzraum über der *Glottis*. Das führt zu einer Änderung der Formantstruktur.

Akustische Korrelate. Das neugeborene Baby beginnt gleich nach der Geburt, mit einer Frequenz um die 440 Hz (a1) zu schreien. Bis zum Alter von circa 9 Jahren sinkt die mittlere Grundfrequenz bei Mädchen und Knaben und liegt dann im Bereich zwischen 220 Hz (a) und 330 Hz (e1). Gleichzeitig erweitert sich der Stimmumfang auf etwa 2 Oktaven. Vom neunten Lebensjahr an entwickeln sich die Stimmen von Mädchen und Knaben unterschiedlich. In der Pubertät zwischen 12 und 16 Jahren wächst der Kehlkopf bei den Knaben durch eine erhöhte Produktion von Androgenen besonders stark. Auch die Stimmlippen werden um ungefähr 1 Zentimeter verlängert, was ein Absinken der mittleren Grundfrequenz um etwa eine Oktave zur Folge hat. Während dieser Entwicklungsphase ist die Stimme oft heiser und instabil. Man sagt auch, die Stimme befindet sich im Stimmbruch. Die Mädchen machen eine weniger starke Phase der Stimmenänderung bereits im Alter von 11 bis 14 Jahren durch. Dabei sinkt die Stimmlage um circa eine Terz ab [2]. Die Grundfrequenz (F0) ist für die Alterswahrnehmung wichtig [44]. Vergleicht man den Alterungsprozeß von Frauen und Männern fallen einige signifikante Unterschiede auf, wie folgend dargestellt. Linville [45] hat bei Frauen bis zur Menopause eine etwa gleichbleibende Grundfrequenz gemessen und danach ein Absinken die-

ser festgestellt. In der Arbeit von Hollien und Shipp (zitiert in [5]), die erwachsene Männer im Alter von 20 bis 89 Jahren untersucht haben, wurde gezeigt, dass die mittlere Grundfrequenz bis zum mittleren Alter sinkt und danach wieder langsam ansteigt. Die mittlere F0 betrug 119.5 Hz für die 20 bis 29 Jährigen und 146.3 Hz für die 80 bis 89 Jährigen. Bei Frauen und Männern nimmt der Stimmumfang mit zunehmendem Alter ab. Die untere Grenze steigt um etwa eine Quarte und die obere Grenze sinkt um ungefähr eine Terz [5].
Ein weiterer sehr wichtiger Parameter für die Wahrnehmung von Alter in der Stimme ist laut Linville (zitiert in [5]) die Stabilität der Grundfrequenz gemessen in der Standardabweichung, welche mit höherem Alter zunimmt.

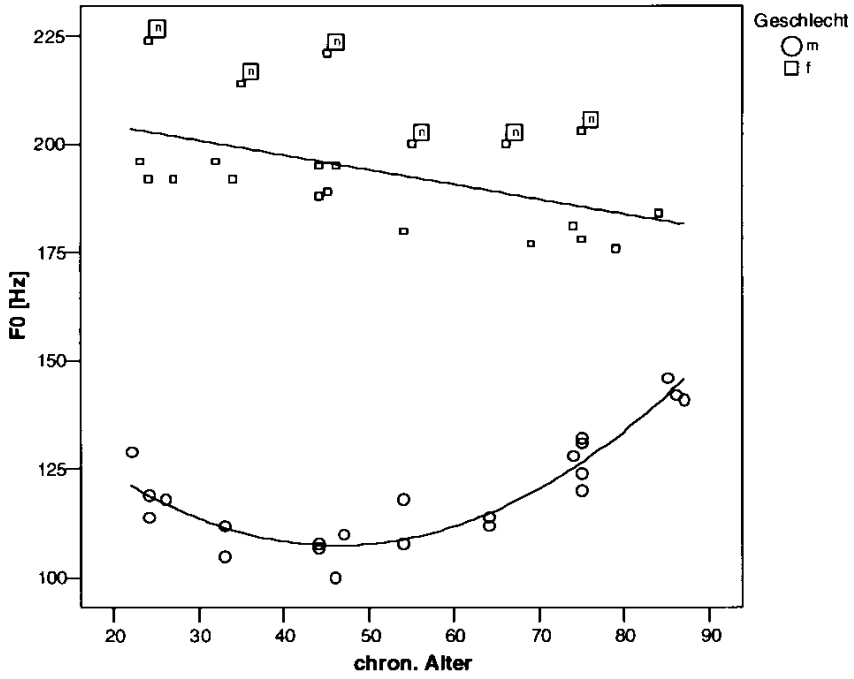

Abbildung 3.2: Grundfrequenz als Funktion des Alters. Das Label [n] kennzeichnet die weiblichen Sprecherinnen, die nie geraucht haben [5].

Die Perturbationsmaße Jitter und Shimmer haben weniger Einfluss auf das wahrgenommene Stimmalter als bisher angenommen. Shimmer bezeichnet Am-

plitudenstörungen und Jitter Schwankungen in der Periodendauer eines Signals. Einige Untersuchungen zeigten zwar einen Zusammenhang zwischen Alter und Jitter, andere wiederum nicht (zitiert in [5]).

Eine Korrelation zwischen Shimmer und Alter von SprecherInnen ist laut Berichten von Mori et al., Ramig und Ringel und Orlikoff und Baken gegeben (zitiert in [5, S. 140]).

Betrachtet man die Formantstruktur einer alten Stimme, so fällt auf, dass sich die ersten vier mittleren Formantfrequenzen in tiefere Frequenzbereiche verschieben. Das ist auf das Absenken des Kehlkopfes und die damit verbundene Verlängerung des Ansatzrohres zurückzuführen.

Ein anderes Merkmal zur Unterscheidung von jungen und alten Stimmen ist die Sprechgeschwindigkeit (*speaking rate*). Obwohl einige Untersuchungen keinen Zusammenhang zwischen Sprechgeschwindigkeit und Sprecheralter bei freier Rede sehen konnten, gibt es auch andere Messungen von Walker et al. (zitiert in [5]), die bei älteren Frauen in freier Rede mehr Sprechunterbrechungen registrierten. Shipp et al. [46] ermittelten eine deutlich höhere Sprechgeschwindigkeit ohne Pausen (gemessen in Silben/ Sekunde) bei jungen Leuten im Vergleich zu älteren SprecherInnen. Zusätzlich atmeten die alten SprecherInnen während eines Satzes häufiger und machten dafür auch längere Pausen.

Die Experimente von Brückl und Sendlmeier [5] zeigen, dass die Stimmen alter Männer eine höhere und die alter Frauen eine tiefere Grundfrequenz aufweisen als die von jungen Männern und Frauen. Bei alten Männern sind die ersten beiden Formanten abgesenkt und der obere Frequenzbereich ist stark gedämpft. Die Stimme klingt durch verminderte Stimmstabilität rauer und hauchiger und die Sprechgeschwindigkeit beim lauten Vorlesen ist auch geringer im höheren Alter.

Frauen sprechen mit zunehmendem Alter zittriger und rauer, was auf eine größere Instabilität von F0 und Amplitudenperturbationen zurückzuführen ist. Außerdem

lesen sie langsamer vor.

Perzeptive Parameter. Wenn HörerInnen gefragt werden, nach welchen Kriterien sie das Alter einer Stimme bestimmen, werden meist Grundfrequenz, Sprechgeschwindigkeit, Lautheit, Rauhigkeit, Stimmqualität (*Vocal Tremor*, Hauchigkeit) und andere Dimensionen, wie Fehlen von Sanftheit, weniger präzise Artikulation und Lebendigkeit genannt [42, S. 15]. Insgesamt sind laut Linville [47] die Parameter Grundfrequenz und die Standardabweichung der F0 für die Wahrnehmung des Stimmalters entscheidend. Traunmüller [39] fand heraus, dass die Wahrnehmung des Alters einer Stimme hauptsächlich auf Grundfrequenz und höheren Formanten basiert.

3.3 Vom Flüstern zum Schreien (Stimmaufwand)

Der Stimmaufwand (*vocal effort*) ist eine Größe, die eine angestrengte oder angespannte Stimme beschreibt [6]. Geringer Stimmaufwand ist auf der Ebene der Wahrnehmung mit Flüstern und großer Stimmaufwand mit Schreien verbunden. Traunmüller und Eriksson [7] definieren *vocal effort* wie folgt:

> *„[...], vocal effort is the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance.“*

Ordinary bedeutet hierbei, dass die SprecherInnen keine relevanten körperlichen Störungen und keine besondere Stimmausbildung absolviert haben. In Abbildung 3.3 sieht man Frequenzspektren von einer hauchigen Stimme und einer Stimme mit viel Stimmaufwand. Das Spektrum der gehauchten Stimme hat einen geringeren höheren Frequenzanteil und einen stärkeren ersten Formanten. Die wichtigsten

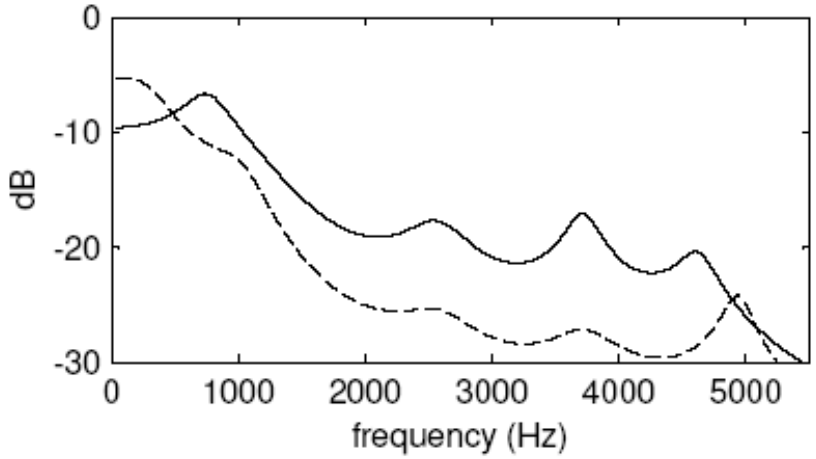

Abbildung 3.3: LPC– Spektrum einer hauchigen Stimme mit wenig Stimmaufwand (gestrichelte Kurve) und einer Stimme mit viel Stimmaufwand (durchgezogene Kurve). Die gleiche Stimme produziert den gleichen Vokal mit gleicher Grundfrequenz [6].

akustischen Charakteristika von Stimmaufwand sind laut Lienard und Benedetto [48] die Grundfrequenz und Amplitude der Grundfrequenz, die Frequenz und Amplitude des ersten Formanten, sowie die Amplituden des zweiten und dritten Formanten. Sie beschreiben einen linearen Zusammenhang zwischen Stimmaufwand und Grundfrequenz, in dem mit steigendem Stimmaufwand die Grundfrequenz um 5 Hz pro dB ansteigt. Darüber hinaus wird ein linearer Zusammenhang zwischen Stimmaufwand und Frequenz des ersten Formanten gezeigt, in dem mit steigendem Stimmaufwand die Frequenz von F1 um 3.5 Hz pro dB ansteigt. Dagegen stehen die Frequenzen des zweiten und dritten Formanten in keinem signifikanten Zusammenhang mit dem Stimmaufwand. Allgemein wurde ein Ansteigen der Amplituden der ersten drei Formanten mit steigendem Stimmaufwand beobachtet. Dazu kommt eine allgemeine Erhöhung der Amplituden im oberen Spektralbereich (Änderung des *spectral tilt*). Laut [6] besitzen Stimmen mit wenig Stimmaufwand eine hohe F1 Amplitude. Neben den oben genannten Kenngrößen wird auch eine etwas

langsamere und deutlichere Sprechweise mit erhöhtem Stimmaufwand verbunden, da man besonders gut verstanden werden will. Außerdem steigt laut Nordstrom und Driessen [6] mit erhöhtem Stimmaufwand das Verhältnis vom Harmonischen Anteil zum Rauschen, der Rauschanteil verringert sich.

Schreien. Bei der Gegenüberstellung von normaler Sprache und Schreien wird in vielen Untersuchungen eine starke Erhöhung der Grundfrequenz beim Schreien beobachtet. Zusätzlich steigt die erste Formantfrequenz F1 mit steigendem Stimmaufwand [7]. Beim Schreien wird der Mund weiter geöffnet als beim normalen Sprechen. Dadurch verändert sich der Vokaltrakt und folglich auch die Formanten, besonders F1. Der Lautstärkepegel steigt bei erhöhtem Stimmaufwand deutlich an [7]. Jedoch können Probanden in Hörtests auch bei normalisierter Lautstärke eine Stimme mit hohem Stimmaufwand von einer mit niedrigem unterscheiden [48].

Flüstern. Beim Flüstern entspannen sich die Stimmlippen und die Glottis schließt sich nicht mehr vollständig. Dadurch strömt die Luft aus der Lunge durch die Glottis und verursacht Turbulenzen, was zu einer rauschhaften Anregung führt. Das heißt, in geflüsterter Sprache gibt es keine Grundfrequenz. Abbildung 3.4 zeigt den Pegelunterschied über den Frequenzbereich zwischen Flüstern und normaler Lautbildung des gleichen Ausdrucks geäußert von Männern (schwarze Quadrate), Frauen (schwarze Kreise), Jungen (weiße Quadrate) und Mädchen (weiße Kreise).

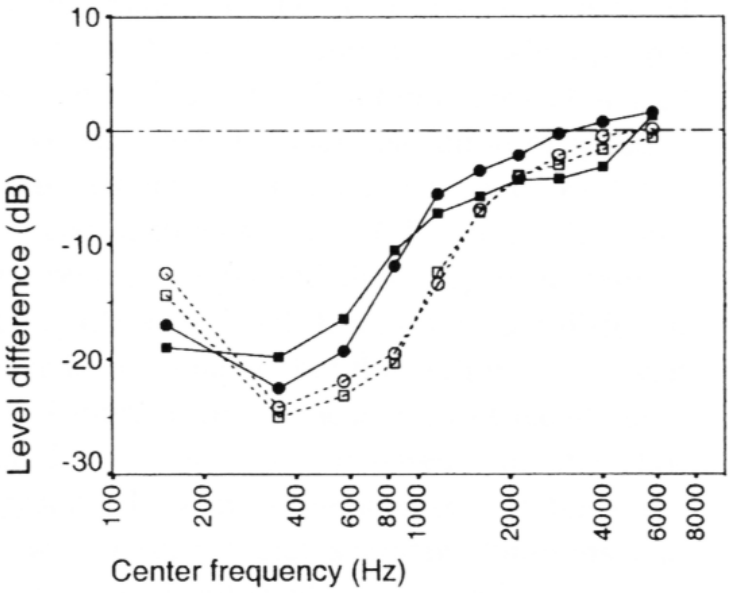

Abbildung 3.4: Pegelunterschied in Abhängigkeit von der Frequenz zwischen Flüstern und normaler Lautbildung des gleichen Ausdrucks geäußert von Männern (schwarze Quadrate), Frauen (schwarze Kreise), Jungen (weiße Quadrate) und Mädchen (weiße Kreise) [7].

Kapitel 4

Grundfrequenz

Für die in dieser Arbeit durchzuführenden Transformationen ist es notwendig, die Grundfrequenz zu schätzen. Das nächste Kapitel gibt zuerst einen kurzen Überblick über die gebräuchlichsten Methoden der Grundfrequenzbestimmung. Danach werden mehrere Algorithmen für die Modifikation der Grundfrequenz erläutert.

4.1 Schätzen der Grundfrequenz

Die Grundfrequenz gehört zu den wichtigsten Sprachsignalparametern. Mit ihr werden Informationen über die Prosodie[1] und damit zusammenhängend über das Geschlecht und das Alter übertragen. Es gibt in der Literatur mehrere Definitionen des Parameters Grundfrequenz [1, S. 198]. Sie unterscheiden sich, je nachdem welche Art von Grundfrequenzbestimmungs- Algorithmen (GFB– Algorithmen) verwendet werden.

Die in dieser Arbeit verwendete Definition für Algorithmen nach dem Prinzip der

[1] Prosodie (laut Bußmann [49, S. 618]):„Gesamtheit sprachlicher Eigenschaften wie Akzent, Intonation, Quantität, Sprechpausen. Sie beziehen sich im allgemeinen auf Einheiten, die größer sind als ein Phonem. Zur Prosodie zählt auch die Untersuchung von Sprechgeschwindigkeit, Rhythmus und Sprechpausen.“

Kurzzeitanalyse lautet:

> „T_0 ist definiert als die mittlere Dauer mehrerer aufeinanderfolgender Grundperioden. Auf welche Weise die Mittelung erfolgt und wieviele Perioden in die Messung involviert sind, bleibt dem einzelnen Algorithmus überlassen.“ [1, S. 198]

Wie bereits in Kapitel 2.3 in Zusammenhang mit dem Quelle– Filter Modell beschrieben, steckt die Information über F0 in der Anregung des Signals. Da nur stimmhafte Segmente eine Grundfrequenz haben, sollte vor jeder Grundfrequenzbestimmung noch die Bestimmung der Anregungsart erfolgen.
Jeder GFB– Algorithmus besteht laut [50] aus drei Verarbeitungsstufen; erstens die Vorverarbeitungsstufe, zweitens die Extraktionsstufe und drittens die Nachbearbeitungsstufe. Die Vorverarbeitungsstufe ist mit der Aufbereitung der Daten (z.B. Datenreduktion) für die nachfolgende Extraktionsstufe zuständig. Damit soll die Aufgabe der Extraktionsstufe erleichtert werden. In dieser zweiten Stufe findet dann die eigentliche Bestimmung der Grundfrequenz statt. Am Ausgang stehen der Schätzwert der Grundperiodendauer oder der Grundfrequenz (fundamentale Frequenz) zur Verfügung. Die Nachbearbeitungsstufe führt eine Fehlerkorrektur durch, glättet den Grundfrequenzverlauf oder stellt das Ergebnis graphisch dar [1].

Die GFB– Algorithmen werden in zwei Kategorien eingeteilt. Wenn das Eingangssignal der Extraktionsstufe die gleiche Zeitbasis wie das originale Signal besitzt, arbeitet der Algorithmus im Zeitbereich. Wird der Zeitbereich in der Vorverarbeitungsstufe verlassen, spricht man von Algorithmen mit Kurzzeittransformation.
Die meisten Probleme bei der F0 Bestimmung treten laut O´Shaughnessy [8] an den Grenzen zwischen stimmhaften und stimmlosen Lauten und bei plötzlichen Formant- oder Amplitudenänderungen auf.

Die Vorteile von GFB– Algorithmen im Zeitbereich liegen in der effizienten Berechnung und in der Bestimmung der Zeitpunkte der einzelnen Grundperioden. Dies ist wichtig für die Tonhöhen– synchron *(pitch– synchron)* arbeitende PSOLA Methode (vgl. 4.2.1) . Algorithmen mit Kurzzeittransformation geben keine Auskunft über die genauen Anfangs- und Endzeitpunkte der einzelnen Grundperioden, sind aber oft zuverlässiger in der F0 Schätzung [8, S. 219].

4.1.1 Autokorrelation

Die Autokorrelationsmethode gehört zu den zuverlässigen und einfach zu implementierenden Methoden zur Bestimmung der Grundfrequenz F0 und ist deshalb auch weit verbreitet [51, 52]. Die Autokorrelationsfunktion (AKF) $r(k)$ des Eingangssignals $s(n)$ lässt sich aus der inversen Fouriertransformation des Leistungsdichtespektrums $S(f)$ berechnen. Eine Möglichkeit der Berechnung der AKF im Zeitbereich ist in Gleichung 4.1 zu sehen. In der AKF sind Informationen über Amplituden der Harmonischen und Formanten und auch über Periodizitäten im Sprachsignal vorhanden. Deshalb kann man sie zur Bestimmung der Grundfrequenz F0 benutzen.
Die AKF ist eine gerade Funktion ($r(k) = r(-k)$), mit einem Maximum bei $k = 0$, wobei $r(0)$ bei stochastischen und periodischen Signalen der mittleren Leistung des Signals s(n) entspricht. Für periodische Signale $s(n)$ mit der Periode T_0 ist auch die AKF $r(k)$ periodisch mit T_0. Bei allen ganzzahligen Vielfachen von T_0 ergeben sich Maxima in der AKF [8, S. 183].

$$R(k) = \sum_{n=-\infty}^{\infty} s(n)s(n-k). \tag{4.1}$$

Die Autokorrelationsmethode ist robust gegenüber verrauschten Signalen, bekommt jedoch Schwierigkeiten bei Signalen mit dominanten Formanten . Es kann zur Verwechslung der Grundperiode mit der Periode der Schwingung des 1. For-

manten kommen [1].

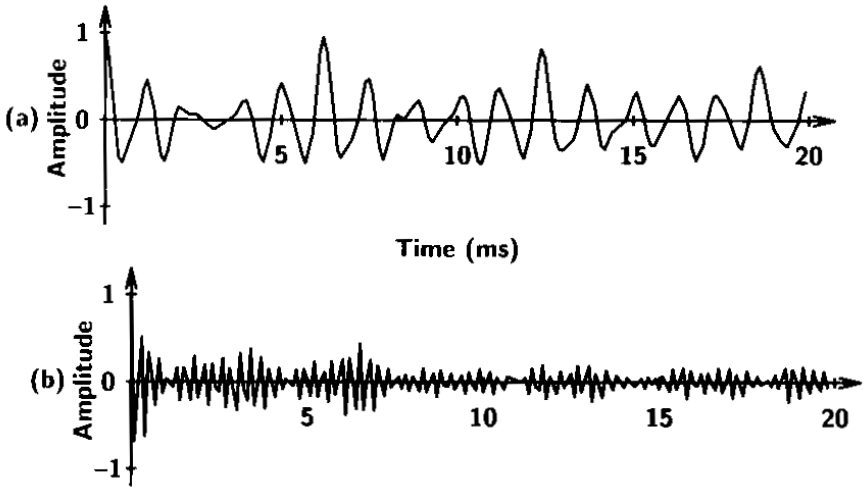

Abbildung 4.1: Autokorrelationsfunktion für (a) stimmhafte und (b) stimmlose Sprachlaute, unter Verwendung eines 20 ms Rechteckfensters (N= 201) [8].

Die Performance der Autokorrelationsmethode lässt sich durch eine Mittenbegrenzung (*center clipping*) des Signals verbessern (siehe Abbildung 4.2).

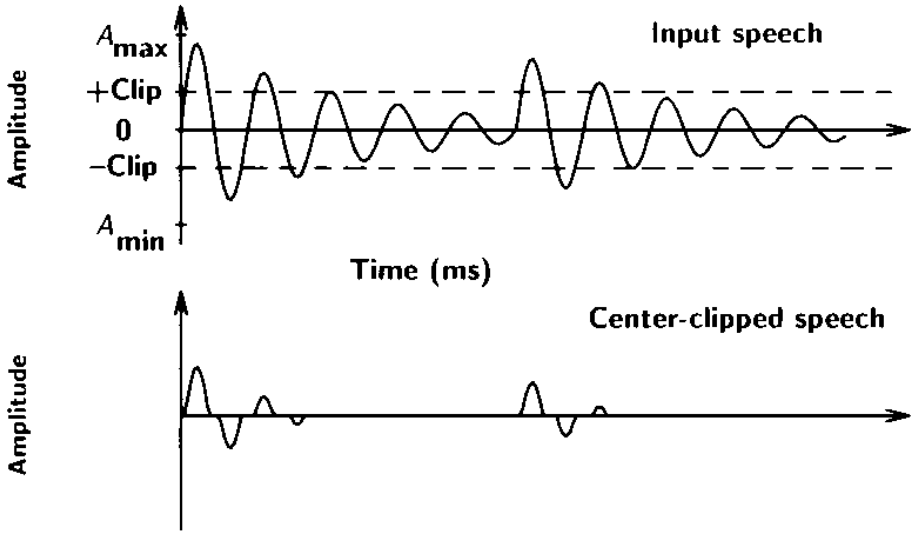

Abbildung 4.2: Beispiel von Mittenbegrenzung [8].

Dabei werden Signalanteile unterhalb eines bestimmten Amplitudenschwellwertes zu Null gesetzt. Übrig bleiben nur die stark ausgeprägten Spitzen am Beginn jeder Grundperiode. Dadurch erhöht sich die Robustheit gegenüber Verwechslungen der Grundfrequenz mit der Frequenz des 1. Formanten [1, S. 202].

Darüber hinaus ist die Autokorrelationsmethode unempfindlich gegenüber Phasenverzerrungen. Das ist vorteilhaft für Signale, die zum Beispiel über eine Telefonleitung übertragen werden oder deren Phase in irgendeiner anderen Weise verzerrt wird [51].

PRAAT. Das in dieser Arbeit genutzte Sprachanalyse– und Syntheseprogramm PRAAT [14] verwendet zur Schätzung der Grundfrequenz eine modifizierte Variante der Autokorrelationsmethode [9]. Es werden zwei Neuerungen eingeführt, die die Genauigkeit und Robustheit der F0– Schätzung erheblich vergrößern. Zur Bestimmung der Autokorrelationsfunktion $r_x(\tau)$ des originalen Signals x(t) wird die AKF $r_a(\tau)$ des gefensterten Signals a(t) durch die AKF $r_w(\tau)$ der Fensterfunktion dividiert.

$$r_x(\tau) \approx \frac{r_a(\tau)}{r_w(\tau)} \tag{4.2}$$

Durch die Normalisierung wird die Amplitude der AKF in die Nähe von 1 gebracht (vgl. Abbildung 4.3), was das Erkennen des ersten *Peaks*[2], der zur Bestimmung von F0 entscheidend ist, erleichtert.

Die zweite Verbesserung betrifft die durch die Samplingfrequenz beschränkte Genauigkeit der Schätzung. Die AKF eines abgetasteten Signals ist auch eine abgetastete Funktion. Durch Erhöhung der Abtastrate der AKF und einer $\sin x/x$–Interpolation erhöht sich die Genauigkeit der F0– Schätzung erheblich [9].

[2]Maximum

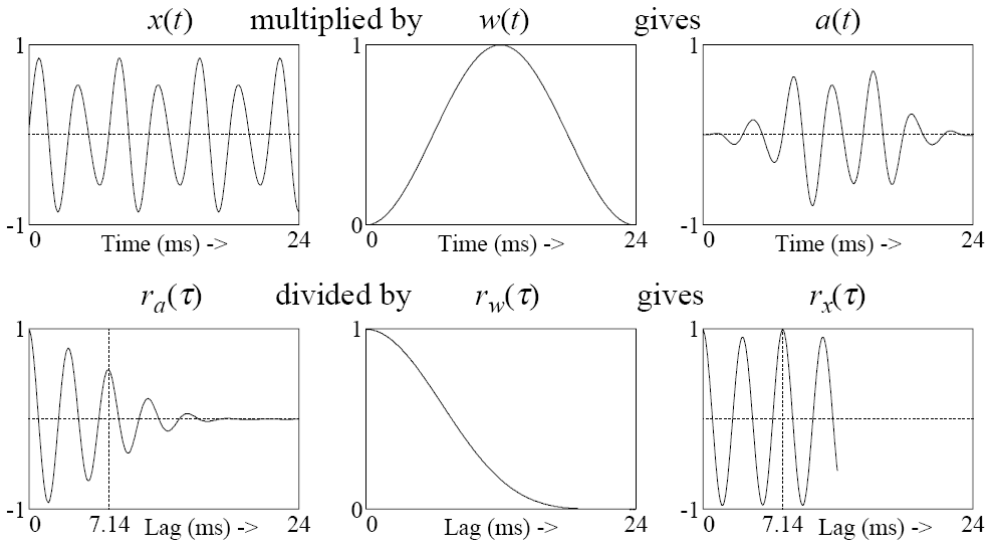

Abbildung 4.3: Grundfrequenzschätzung mittels AKF in PRAAT [9].

4.1.2 Betragsdifferenzfunktion

Eine andere weit verbreitete Methode zur Grundfrequenzbestimmung im Zeitbereich ist die Verwendung der Betragsdifferenzfunktion *(average magnitude difference function, AMDF)*. Hierbei wird das um k Samples verschobene Sprachsignal $s(n-k)$ vom nicht verschobenen Signal $s(n)$ abgezogen.

$$AMDF(k) = \sum_{n=-\infty}^{\infty} |s(n) - s(n-k)| \,. \tag{4.3}$$

Bei jeder Verzögerung k, die einem ganzzahligen Vielfachen der Grundfrequenzperiode entspricht, besitzt die AMDF ein Minimum. Die AMDF lässt sich auch mit sehr kurzen Messintervallen bestimmen. Die AMDF Methode ist schneller als die Autokorrelationsmethode, da nur simple Operationen, wie Subtraktion und Gleichrichtung und keine Multiplikationen zur Berechnung benötigt werden [8, S. 185].

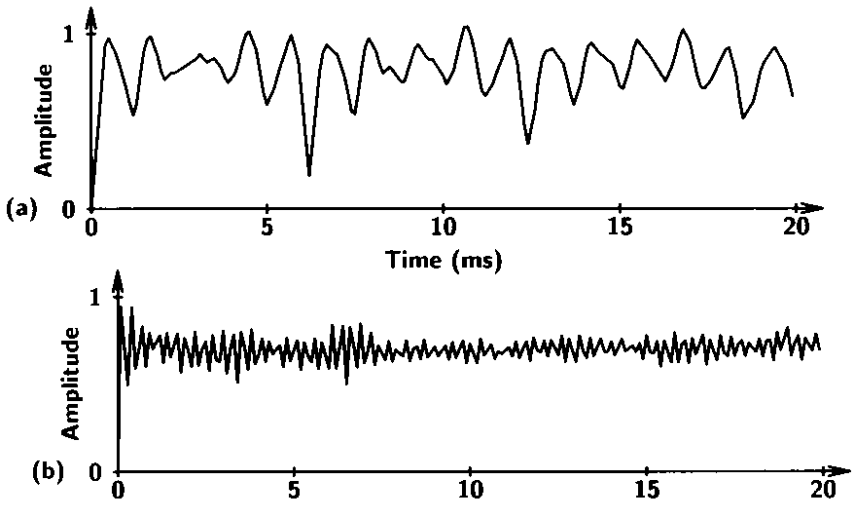

Abbildung 4.4: AMDF Funktion (normalisiert) für das gleiche Sprachsignal wie in Abbildung 4.1 [8].

4.1.3 Cepstrum

Mit dem Cepstrum $c(n)$ (siehe auch 5.1.2) ist es möglich, ein Sprachsignal

$$y(n) = x(n) * h(n) \tag{4.4}$$

in Quelle und Filter (vgl. Quelle– Filter Modell in 2.3) aufzutrennen. Speziell für die Grundfrequenzbestimmung wird die Quelle oder Anregung $x(n)$ benötigt. Gleichung 4.4 im Frequenzbereich ergibt

$$Y(e^{j\Omega}) = X(e^{j\Omega}) \cdot H(e^{j\Omega}). \tag{4.5}$$

Die Aufsplittung des Signals erfolgt durch Gewichtung des Cepstrums mit zwei Fensterfunktionen, einem Tiefpass– Fenster $w_{LP}(n)$ und einem Hochpass– Fenster $w_{HP}(n)$. Die niedrigen Cepstral– Werte (niedrige *quefrencies*) entsprechen der spektralen Einhüllenden in dB mit $log\left|H(e^{j\Omega})\right|$ und die hohen (hohe *quefrencies*) der Quelle mit $log\left|X(e^{j\Omega})\right|$.

Zur Bestimmung der Grundfrequenz wird der Bereich mit den hohen *quefrencies* herangezogen. Wenn das Signal periodisch ist, ist auch das Cepstrum entspre-

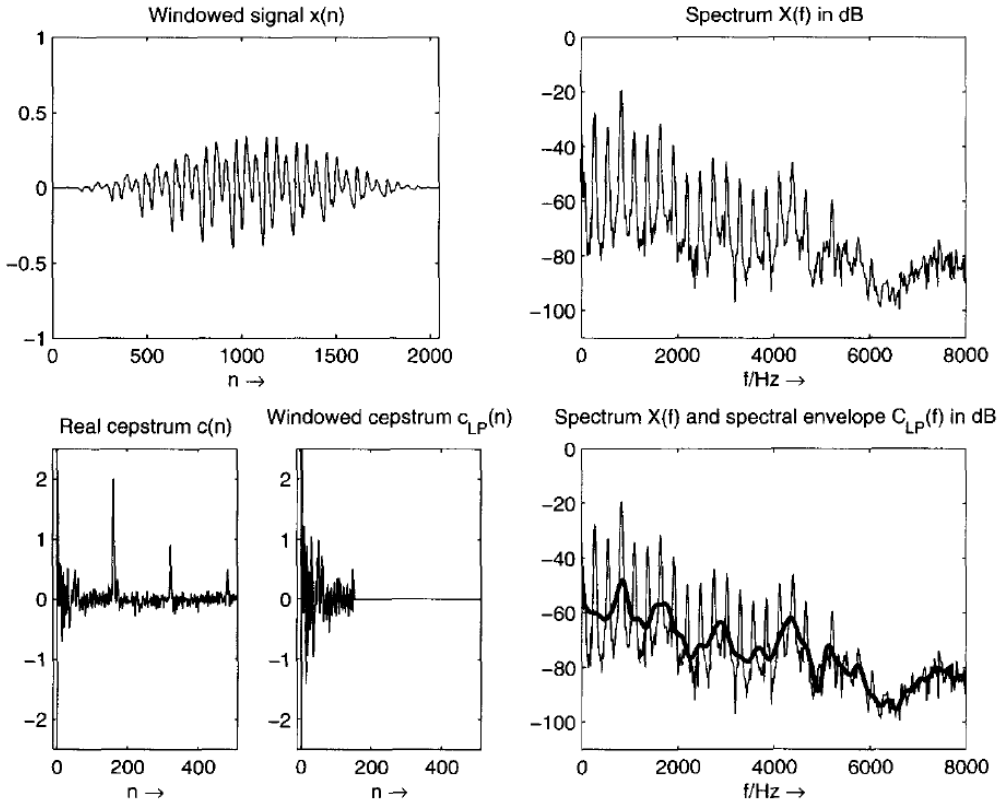

Abbildung 4.5: Gefenstertes Signalsegment, Spektrum (FFT Länge $N = 2048$), Cepstrum, Gefenstertes Cepstrum ($N_1 = 150$) und spektrale Einhüllende [10].

chend der Grundfrequenz periodisch. Der erste signifikante *Peak* im Cepstrum, nachdem der Tiefpass Anteil weggeschnitten wurde, entspricht der Grundfrequenz in Samples (siehe Abbildung 4.5). Das Cepstrum Verfahren für die Schätzung der Grundfrequenz ist im Gegensatz zur Autokorrelationsmethode unempfindlich gegenüber stark ausgeprägten Formanten. Dafür kann es bei verrauschten Signalen Probleme geben [1, S. 205].

4.1.4 Harmonische Analyse

Die Harmonische Analyse arbeitet im Frequenzbereich. Bei der direkten Bestimmung der Grundfrequenz F0 aus dem ersten Maximum des Leistungsdichtespektrums kann es zu Ungenauigkeiten kommen. Das Signal kann bandpassgefiltert sein (Telefon) oder die Grundfrequenz besitzt wenig Energie im Vergleich zum ersten Formanten [8]. Ein weiterer Schwachpunkt dieser Methode ist die begrenzte Auflösung der Frequenzen. Eine N– Punkte *Fast Fourier Transformation* (FFT) besitzt eine Frequenzauflösung von

$$\Delta f = \frac{f_s}{N} \tag{4.6}$$

Hierbei steht f_s für die Samplingfrequenz. Wird im Betragsfrequenzgang des Signals beim FFT– Bin k_0 ein lokales Maximum detektiert, dann berechnet sich der Schätzwert für die Grundfrequenz folgendermaßen:

$$\tilde{f} = k_0 \cdot \Delta f = k_0 \frac{f_s}{N}. \tag{4.7}$$

Bei einer Samplingfrequenz von 44100 Hz und einer FFT– Länge von $N = 1024$ beträgt die Frequenzauflösung demzufolge $\Delta f \approx 43.07$ Hz, was für einige Anwendungen eventuell zu wenig sein kann. Die Frequenzauflösung lässt sich mit verschiedenen Methoden vergrößern [10, S. 337].

Verbesserungen anderer Art können erzielt werden, indem die harmonische Struktur des Signals in die F0– Bestimmung mit einbezogen wird. Zuerst müssen die Frequenzen der Harmonischen detektiert werden. Diese befinden sich bei ganzzahligen Vielfachen von F0. Der größte gemeinsame Teiler dieser Vielfachen ist ein guter Schätzwert für die Grundfrequenz. Eine andere Möglichkeit der F0 Schätzung ist die Stauchung der Frequenzachse mit den ganzzahligen Faktoren (2, 3, 4, ... usw.). Die Summe dieser Spektren besitzt ein signifikantes Maximum bei der Grundfrequenz, da die Harmonischen an die Stelle der Grundfrequenz verschoben werden [8].

4.2 Modifikation der Grundfrequenz

Nachdem die Grundfrequenz mit den oben beschriebenen Methoden erfolgreich geschätzt wurde, ist es nun möglich, diese auch zu modifizieren. Dabei wird grundsätzlich in parametrische und nicht– parametrische Methoden unterschieden [11]. Der PSOLA– Algorithmus wird den nicht– parametrischen, der Phasenvocoder hingegen den parametrischen Methoden zugeordnet. Diese beiden Verfahren werden in den folgenden Kapiteln genauer beschrieben.

4.2.1 PSOLA

Eine seit vielen Jahren weit verbreitete Methode, die Grundfrequenz eines Sprachsignals zu verändern, ist der von Moulines *et. al* [53, 54] vorgeschlagene *Pitch–Synchronous Overlap and Add* (PSOLA)– Algorithmus. Eine Zeitbereichs– Implementierung der PSOLA Methode, das *Time Domain* (TD)– PSOLA Verfahren, basiert auf der Annahme, dass das Eingangssignal durch eine Grundfrequenz (*pitch*) charakterisiert werden kann. Dies trifft unter anderem für stimmhafte Laute der Sprache zu. Neben der Grundfrequenz lassen sich mit TD– PSOLA auch die Dauer und Amplitude eines Sprachsignals modifizieren. Der PSOLA Algorithmus besteht aus drei Phasen: der Analyse–, der Modifikations– und der Synthese– Phase.

Analysephase. In der Analysephase wird zuerst die Grundfrequenz des Eingangssignals bestimmt. Danach wird jede Grundperiode[3] mit Markern versehen. Bei stimmhaften Lauten werden die Marker *pitch*– synchron an die Stelle der größten Amplitude oder der glottalen Pulse gesetzt. Bei stimmlosen Lauten geschieht dies in konstanten Abständen von meist 10 bis 15 ms. Anschließend wird das Signal $x(n)$ in Blöcke aufgeteilt, indem zeitverschobene Fensterfunktionen $w(n)$ mit

[3] auch Pitchperiode genannt

$x(n)$ multipliziert werden. Die Länge der Fenster L_i ist proportional zur lokalen Grundperiode T0, das heisst $L_i = kT0$, wobei der Proportionalitätsfaktor k bei TD– PSOLA meist 2 beträgt.

Modifikationsphase. In der Modifikationsphase werden die Kurzzeit– Analysesignale in eine modifizierte Kette von Kurzzeit– Synthesesignalen umgewandelt, die mit neuen Synthese– Pitchmarkern synchronisiert werden. Je nachdem, welche Tonhöhen– (Parameter β) oder Zeitdauermodifikationen (Parameter α) vorgenommen werden, müssen die Anzahl, der Abstand zwischen den Segmenten und möglicherweise die jeweiligen Kurzzeitsegmente[4] selbst modifiziert werden.

Synthesephase. In der Synthesephase werden die einzelnen Segmente anhand der neuen Synthese– Pitchmarker gemäß dem *overlap– add* Verfahren wieder zu einem vollständigen Signal zusammengefügt. Das Synthesefenster ist gleich dem Analysefenster.
Bei der Tonhöhenmodifikation betrachtet man nur die stimmhaften Laute, da diese den periodischen Anteilen der Sprache entsprechen. Das heißt, ein Impulszug $g(n)$ mit einer bestimmten Frequenz durchläuft den Resonanzfilter v(n), wird also gefaltet und ergibt das Ausgangssignal $y(n)$. Nun werden jeweils um die glottalen Impulse herum kurze Signalabschnitte aus $y(n)$ extrahiert und in anderen Abständen wieder zurückkopiert oder zusammengesetzt. Dabei wird jedes Segment mit einer Fensterfunktion gewichtet. Um zeitliche Modifikationen zu kompensieren, werden Segmente wiederholt oder weggelassen. Der Name „*pitch– synchronous*“ verweist somit auf die *pitch–* synchrone Extrahierung der Segmente. Jedes kurze Segment kann als Faltung eines Impulses mit der Impulsantwort des Resonanzfilters aufge-

[4] siehe dazu Kapitel 5.2.1

fasst werden. Da keine Abtastratenmodifikation (*resampling*) durchgeführt wird, sondern die Segmente lediglich umkopiert werden, bleibt die Formantstruktur erhalten, die Impulsantwort des Filters also unverändert. Die veränderte Periode der Segmente geht mit einer Tonhöhenänderung einher. Es besteht ohne Weiteres die Möglichkeit auch die Formanten zu ändern, indem vor dem Umkopieren die kurzen Signalsegmente *resampled* werden (vgl. Abschnitt 5.2.1).

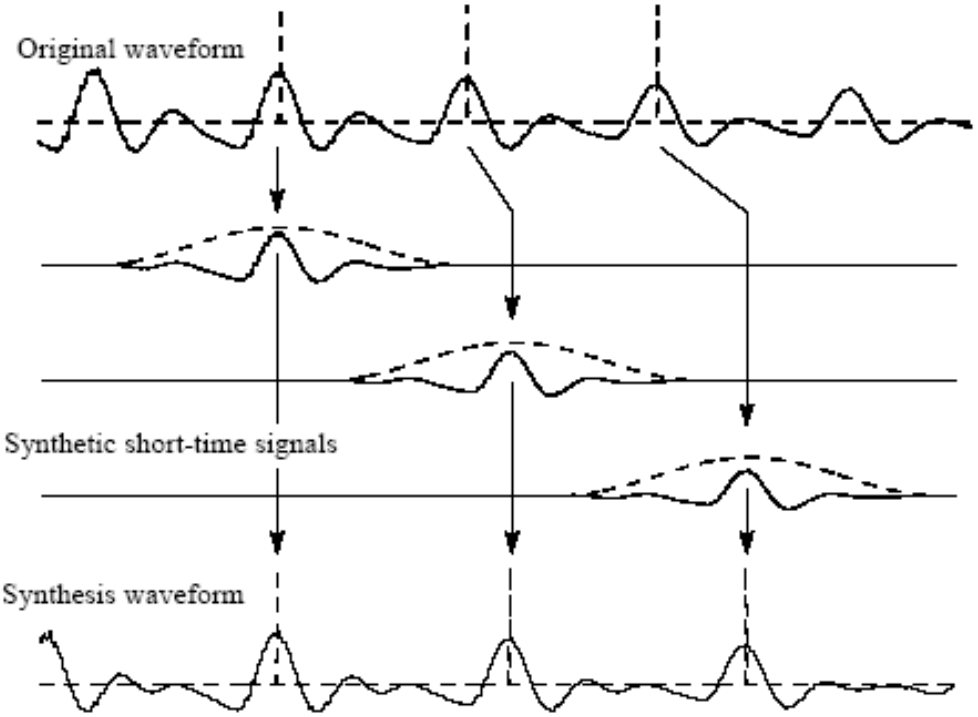

Abbildung 4.6: Grundfrequenz– Modifikation mit TD– PSOLA [11].

Bei der TD– PSOLA wird die zeitliche Entwicklung der Tonhöhe (Grundfrequenz) als bekannt vorausgesetzt. Dazu wird eine Analyse– Stufe vor der eigentlichen Tonhöhenänderungsstufe benötigt. Aufgrund dessen kann es zu Schwierigkeiten bei einer Echtzeit– Implementierung dieses Verfahrens kommen.

Die PSOLA Methode ist relativ robust gegenüber Hall- und Chorusproblemen, welche typische Kennzeichen von Methoden im Frequenzbereich sind (vgl.

Phasenvocoder Abschnitt 4.2.2) [11]. Das gilt zumindest für moderate Parameteränderungen und für quasi– periodische Signale. Bei nichtperiodischen Signalen können hingegen starke Artefakte auftreten. Ein weiteres Problem dieses Verfahrens stellt das sogenannte *transient doubling* dar. Wird das Signal genau inmitten eines Transienten zerschnitten, kommt es zu ungewollten Verdopplungen. Neuere Zeitbereichsmethoden erkennen Transienten und verhindern eine Wiederholung dieser Segmente. Alles in allem sind Zeitbereichsmethoden für Zeitskalierungs- und Tonhöhenänderungen wegen ihrer Einfachheit und guten Performance für kleine Modifikationen gut geeignet.

Bei Zeitdehnungen von Werten $\alpha > 2$ kommt es bei stimmlosen Phonemen zu einem hörbaren Brummen, da einzelne Segmente des Signals regelmässig wiederholt werden. Wird hingegen jedes stimmlose Segment zeitlich gespiegelt, kann aufgrund der verringerten Kurzzeit– Korrelation aufeinanderfolgender Segmente dieser Effekt vermindert werden.
Das PSOLA Verfahren kann im Zeitbereich als *Time Domain* (TD)– PSOLA oder als *Linear Predictive* (LP)– PSOLA oder im Frequenzbereich als *Frequency Domain* (FD)– PSOLA implementiert werden.

LP– PSOLA. *Linear Predictive*– PSOLA funktioniert ähnlich der TD– PSOLA, wobei nicht das Zeitsignal direkt, sondern das aus der LPC Analyse gewonnene Anregungssignal des Quelle– Filter Modells verwendet wird. Durch Trennung von Anregung und Filter besteht die Möglichkeit, die Grundfrequenz und die spektrale Einhüllende unabhängig voneinander zu verändern.

FD– PSOLA. Die *Frequency Domain*– PSOLA Technik limitiert die Anwendbarkeit auf die Modifikation der Grundfrequenz. Das *pitch*– synchrone Kurzzeit– Analyse Signal wird vor der Synthese in den Frequenzbereich transformiert, dort

modifiziert und rücktransformiert. So kann der Abstand zwischen Harmonischen um den Faktor β neu skaliert werden. Bei einer Erhöhung von F0 muß ein Teil des oberen Spektrums abgeschnitten werden. Bei einer F0 Verringerung dagegen entsteht durch die Stauchung der Frequenzachse eine leere „Stelle“ im oberen Frequenzbereich, die wieder aufgefüllt werden muß. Eine nähere Beschreibung dieser Methode ist in [54, 11] zu finden.

4.2.2 Phasen– Vocoder

Ein weiteres bekanntes und weit verbreitetes Analyse- Synthese System ist der Phasen– Vocoder. Da es sehr viel Literatur über dieses Thema gibt, siehe unter anderem [10, 8], werden die Grundlagen nur kurz umrissen.
Das Eingangssignal $x(n)$ wird in Blöcke der Länge N unterteilt und mit einer N–Punkte langen Fensterfunktion gewichtet. Die gefensterten Segmente werden nun mittels Fouriertransformation in den Frequenzbereich transformiert. Die zeitvarianten Kurzzeitspektren (*short time fourier transform- STFT*) des Zeitsignals $x(n)$ schreiben sich wie folgt:

$$\begin{aligned} X(n,k) &= \sum_{m=-\infty}^{\infty} x(m)h(n-m)W_N^{mk},\ k = 0,1,...,N-1 \\ W_N &= e^{-j2\pi/N} \\ &= X_R(n,k) + jX_I(n,k) = |X(n,k)|\, e^{j\varphi(n,k)} \end{aligned} \tag{4.8}$$

$X(n,k)$ ist komplex und vereint den Betragsfrequenzgang $|X(n,k)|$ und den Phasengang $\varphi(n,k)$ mit den Frequenzbins $0 \leq k \leq N-1$ und dem Zeitindex n. Der Betrag und die Phase jedes Blocks können modifiziert und dann mit Hilfe der inversen Fouriertransformationen (IFFTs) wieder in den Zeitbereich rücktransformiert werden. In dieser Synthesestufe werden die Segmente wiederum gefenstert

und überlappend addiert (*overlap add*). In Abbildung 4.7 sind die drei verschiedenen Stufen (Analyse, Transformation, Synthese) eines Phasenvocoders abgebildet.

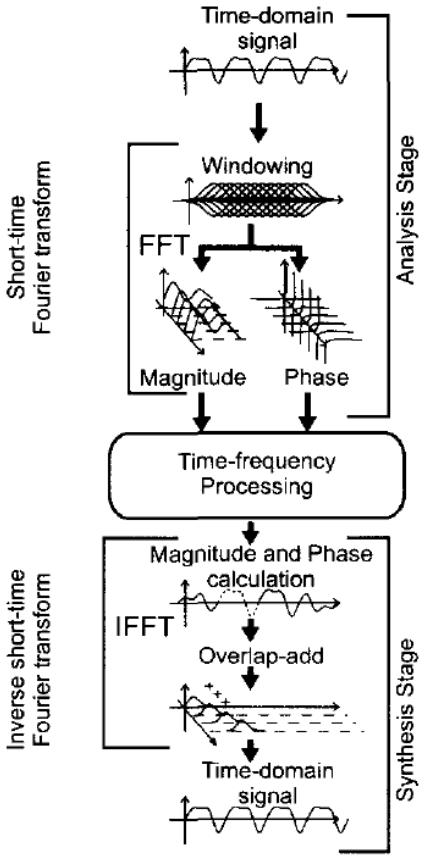

Abbildung 4.7: Zeit– Frequenz Bearbeitung mit dem Phasenvocoder [10].

Bei Zeitdauer– und Tonhöhenmodifikationen mittels der Standardausführung des Phasenvocoders treten oft Artefakte, wie Halligkeit, *„Phasiness“* oder der Verlust von Präsenz, auf [55]. Durch verschiedene neue Techniken lässt sich das Auftreten dieser Artefakte jedoch um ein Vielfaches verringern [56, 55, 57, 58].

Kapitel 5

Formanten

Formanten sind Verstärkungen von bestimmten Frequenzbereichen (Resonanzen), die durch Frequenz, Amplitude und Bandbreite gekennzeichnet sind. Die Formantfrequenzen hängen von der Größe und Form des Vokaltrakts ab. Die Amplituden der Formanten geben neben der Intensität der Anregung auch Aufschluss über die Kopplung von Vokaltrakt und Anregung bzw. Quelle. Je größer die Kopplung, desto geringer sind die Formantamplituden. Aufgrund der stärkeren Kopplung vergrößern sich die Bandbreiten der Formanten. Gleichzeitig verringert sich die Intensität [34].

Informationen über Formanten stecken also hauptsächlich in der spektralen Einhüllenden des Sprachsignals. Für die Schätzung des Spektrums des Vokaltrakts gibt es viele verschiedene Methoden, wovon einige wichtige im folgenden Kapitel vorgestellt werden.

5.1 Schätzen der spektralen Einhüllenden

5.1.1 Lineare Prädiktion (LPC)

Geht man von dem in Kapitel 2.3 genannten Quelle– Filter Modell der Spracherzeugung aus, ist es wünschenswert, die Koeffizienten des diesem Modell zugrundeliegenden Allpol– Filters zu bestimmen. Dies ist mit Hilfe der linearen Prädiktion im Sinne einer Systemidentifikation möglich. Das vorhandene Sprachsignal lässt sich somit in die Bestandteile Quelle und Filter, bzw. Anregung und spektrale Einhüllende zerlegen. Da die „wahren“ Modellkoeffizienten c_k unbekannt sind, kann bei der linearen Prädiktion (LPC) das jeweils aktuelle Sample des Ausgangssignals $s(n)$ aus einer Linearkombination vergangener Samples $s(n-i)(i=1,2,...k)$ und der Verwendung eines FIR Filters mit Koeffizienten a_k angenähert werden.

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{5.1}$$

p bezeichnet die LPC Ordnung und a_k die zu bestimmenden Prädiktionskoeffizienten. Die Differenz zwischen Originalsignal $s(n)$ und geschätztem Signal $\hat{s}(n)$ lautet unter der Voraussetzung $a_0 = 1$:

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) = \sum_{k=0}^{p} a_k s(n-k). \tag{5.2}$$

Das Differenzsignal $e(n)$ wird als Residuum oder Prädiktionsfehler bezeichnet, der minimiert werden soll. Bestimmt man die Prädiktorkoeffizienten a_k so, dass sie den „wahren“ Modellkoeffizienten c_k entsprechen, ist der Prädiktionsfehler gleich der gesuchten Anregung $u(n)$. Im Frequenzbereich ergibt sich die Übertragungsfunktion des FIR Prädiktionsfehlerfilters, oder auch LP– Analyse Filter genannt, folgendermaßen:

$$A(z) = \sum_{k=0}^{p} a_k z^{-k}. \tag{5.3}$$

Damit schreibt sich Gleichung 5.2 im Frequenzbereich

$$E(z) = S(z)A(z). \tag{5.4}$$

Mit der Definition des Allpol– oder auch LP– Synthese– Filters $H(z)$

$$H(z) = \frac{1}{A(z)} \tag{5.5}$$

erhält man das Eingangssignal S(z) durch Filterung der Anregung U(z) mit dem Synthese Filter H(z).

$$S(z) = U(z) \cdot H(z). \tag{5.6}$$

$H(z)$ modelliert die Resonanzen der Sprache, entspricht also der spektralen Einhüllenden des Signals. Der Synthese Filter ist minimalphasig, weshalb auch immer eine stabile Inverse $A(z)$ existiert. Führt man nun eine LPC Analyse des Eingangssprachsignals durch, erhält man die spektrale Einhüllende des Signals in Form der Auto– Regressiven (AR)– Koeffizienten a_k. Die AR Koeffizienten lassen sich mit der Autokorrelationsmethode und dem Levinson- Durbin Algorithmus bestimmen. Da die LPC Analyse nur mit einer endlichen Ordnung p durchgeführt wird, ergibt sich ein geglättetes Spektrum $H(z)$. Mit der Inverse von $H(z)$, dem Analyse Filter $A(z)$ lässt sich nun eine Schätzung des Anregungssignals $u(n)$ ermitteln. Die linearen Prädiktionskoeffizienten a_k bestimmen die Pole des LP– Analyse Filters. Sie werden durch Minimierung des quadratischen Fehlers $e(n)$ ermittelt. Die Minimierung des quadratischen Fehlers $e(n)$ ergibt:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} d\omega = 1. \tag{5.7}$$

$P(\omega)$ ist das Leistungsdichtespektrum des Signals und $\hat{P}(\omega)$ das durch den LP–Filter geschätzte Leistungsdichtespektrum. Bei der Fehlerminimierung löschen sich Fehler, die durch $P(\omega) > \hat{P}(\omega)$ und durch $\hat{P}(\omega) > P(\omega)$ entstehen, gegenseitig aus. Dadurch ergibt sich nur eine ungefähre spektrale Einhüllende, die nicht durch alle spektralen Punkte läuft. Ein weiterer Nachteil der Linearen Prädiktion ist, dass sich die geschätzten Pole meist in Richtung der Harmonischen also der ganzzahligen Vielfachen der Grundfrequenz bewegen. Grund dafür ist das Aliasing, welches während der Schätzung der Prädiktionskoeffizienten in der Autokorrelationssequenz auftritt [59]. Die LPC Koeffizienten können nicht direkt für die weitere Verarbeitung verwendet werden, da sie zu empfindlich gegenüber Quantisierungsfehlern sind und somit die Gefahr von Instabilität besteht. Die LPC Koeffizienten können in *Line Spectral Frequencies* (LSF) oder cepstrale Koeffizienten transformiert werden, die wesentlich robuster gegenüber Quantisierungsfehlern sind.

5.1.2 Cepstrale Koeffizienten

Eine andere Möglichkeit, die spektrale Einhüllende eines Signals darzustellen, bietet sich mit den cepstralen Koeffizienten an. Zum einen können diese mit einer iterativen Formel direkt aus den LPC Koeffizienten berechnet werden [60]. Zum anderen kann der Übergang vom Zeitsignal zum Cepstrum mittels einer logarithmierten diskreten Fouriertransformation (DFT) mit nachfolgender IDFT erfolgen [10]. Von den jeweiligen komplexen DFT Koeffizienten

$$X(k) = \sum_{n=0}^{N-1} x(n) W_{kn}^{N} = |X(k)|\, e^{j\varphi_x(k)}, k = 0, 1, ..., N-1 \tag{5.8}$$

wird der Logarithmus berechnet

$$\hat{X}(k) = logX(k) = log\,|X(k)| + j\varphi_x(k). \tag{5.9}$$

Da in der Praxis das reelle Cepstrum meistens ausreichend ist [8, S. 212], nimmt

man nur den Realteil von $\hat{X}(k)$

$$\hat{X_R}(k) = log\,|X(k)| \tag{5.10}$$

und führt eine inverse DFT durch

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \hat{X_R}(k) W_{-kn}^{N}. \tag{5.11}$$

Tiefpass– Fensterung der Cepstralkoeffizienten $c(n)$ mit nachfolgender Fouriertransformation ergibt eine geglättete Version der spektralen Einhüllenden $C_{LP}(k)$ des Eingangssignals $x(n)$ [10, S. 311]. Die einzelnen Schritte für die Berechnung der spektralen Einhüllenden mittels des Cepstrums sind in Abbildung 5.1 zu sehen.

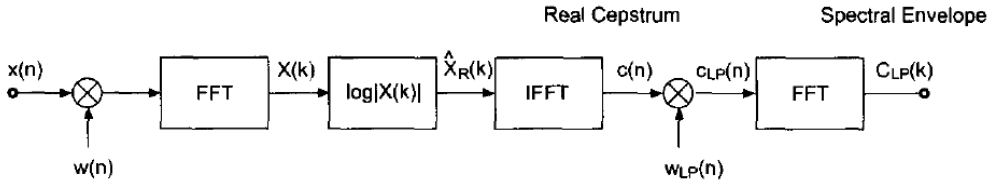

Abbildung 5.1: Berechnung der spektralen Einhüllenden mit Hilfe des Cepstrums [10, S. 311].

5.1.3 Line Spectral Frequencies

Line Spectral Frequencies (LSF) bieten eine alternative Darstellungsform der LPC Koeffizienten, die aber einige wichtige Vorteile bringt. Sie sind robust gegenüber Quantisierungsfehlern, Interpolation ist möglich und unter den unten genannten Bedingungen bilden sie ein stabiles System. LSF lassen sich durch Zerlegung des LP– Analyse Filters $A(z)$ (siehe Gleichung (5.3)) der Ordnung p in ein Spiegelpolynom $P(z)$ und in ein Antispiegelpolynom $Q(z)$ berechnen.

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A(z^{-1}) \\ Q(z) &= A(z) - z^{-(p+1)} A(z^{-1}) \end{aligned} \tag{5.12}$$

Die Nullstellen dieser sogenannten *Line Spectral Pair* (LSP)– Polynome bilden die LSF. Das Polynom $A(z)$ lässt sich daraus wieder eindeutig rekonstruieren.

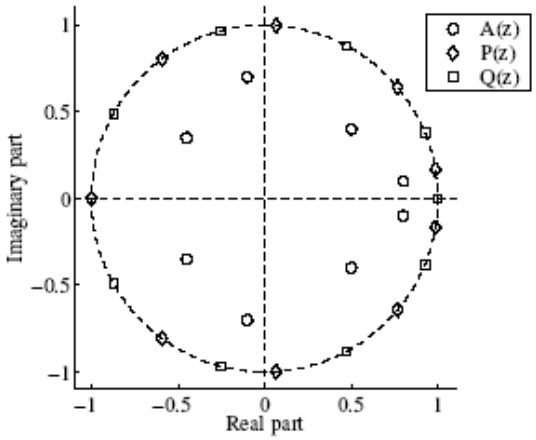

Abbildung 5.2: Nullstellen der *Line Spectral Pair*– Polynome $P(z)$ und $Q(z)$ berechnet aus dem LP– Polynom $A(z)$ [12].

$$A(z) = \frac{1}{2}[P(z) + Q(z)] \tag{5.13}$$

Betrachten wir nun die Nullstellen α_i und β_i von $P(z)$ und $Q(z)$. Diese Nullstellen befinden sich auf dem Einheitskreis der z– Ebene, $|\alpha_i| = |\beta_i| = 1$ und lassen sich wie folgt darstellen:

$$\alpha_i = e^{i\pi\lambda_i} \text{ und } \beta_i = e^{i\pi\gamma_i}. \tag{5.14}$$

Die LSF Parameter λ_i und γ_i bezeichnen die Winkel der Nullstellen auf dem Einheitskreis. Sie liegen in aufsteigender Reihenfolge im Bereich zwischen 0 und 1. Mit ihnen lässt sich $A(z)$ wieder herstellen. Es muss darauf geachtet werden, dass die auf dem Einheitskreis befindlichen Nullstellen der Polynome $P(z)$ und $Q(z)$ ineinander verschachtelt alternieren. Ist das der Fall, entspricht die Summe beider Polynome einem minimalphasigen System und die Stabilität der Allpol Filter $P^{-1}(z)$ und $Q^{-1}(z)$ ist gesichert. An den Polstellen von $P^{-1}(z)$ und $Q^{-1}(z)$ ergeben

sich im Frequenzbereich unendlich hohe Werte, wie in Abbildung 5.3 dargestellt, die jeweils den Winkeln der Nullstellen entsprechen. Diese sind im Spektrum als vertikale Linien erkennbar. Diese Linien werden als *Line Spectral Frequencies* bezeichnet [12].

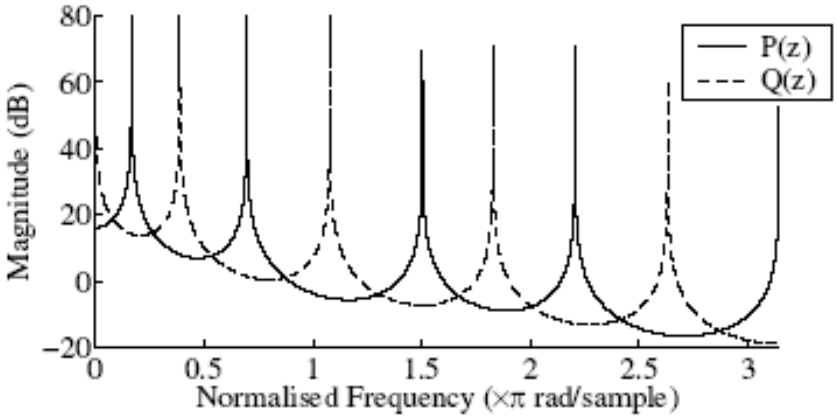

Abbildung 5.3: *Line Spectral Frequencies* in den Spektren von P(z) und Q(z) [12].

5.2 Modifikation der Formanten

Für die Modifikation von Formanten gibt es verschiedene Methoden, die im folgenden Kapitel vorgestellt werden. Die spektrale Einhüllende, als Träger der Formantinformation, lässt sich durch Abtastratenmodifikation (*resampling*), *Frequency Warping*) und durch direkte Manipulation der LPC Pole modifizieren.

5.2.1 Abtastratenmodifikation

In Verbindung mit dem PSOLA System lassen sich die Formantfrequenzen linear verschieben, indem die jeweiligen Segmente des Eingangssignals vor der Synthese zeitlich gedehnt oder gestaucht werden. Eine zeitliche Dehnung entspricht einer

Stauchung im Frequenzbereich und eine Stauchung im Zeitbereich korrespondiert mit einer Expansion im Frequenzbereich. Zur Erhöhung der Formanten um einen Faktor γ muss jedes Segment um den Faktor $1/\gamma$ mit dementsprechendem *Resampling* verkürzt werden [10, S. 224]. Abbildung 5.4 zeigt die Formantskalierung mittels Abtastratenreduktion.

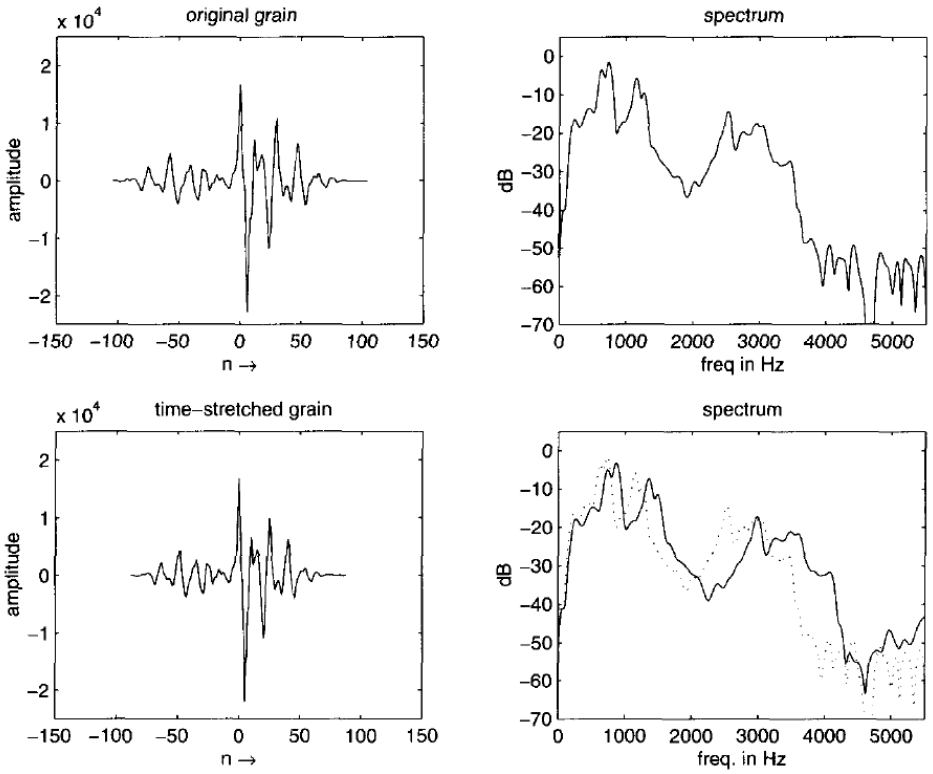

Abbildung 5.4: PSOLA Formant Skalierung mittels Abtastratenreduktion [10, S. 227].

5.2.2 Frequency Warping

Mit dem *Frequency Warping* können Formanten verschoben werden, indem man die spektrale Einhüllende mit einer *Warping* Funktion umformt. Bei einer Formantverschiebung wird zuerst der zu verändernde Frequenzbereich mit unterer

und oberer Grenzfrequenz f_u und f_o festgelegt. Zusätzlich bestimmt man die originale Formant– Mittenfrequenz f_1 und danach die neue Ziel– Mittenfrequenz f_2.

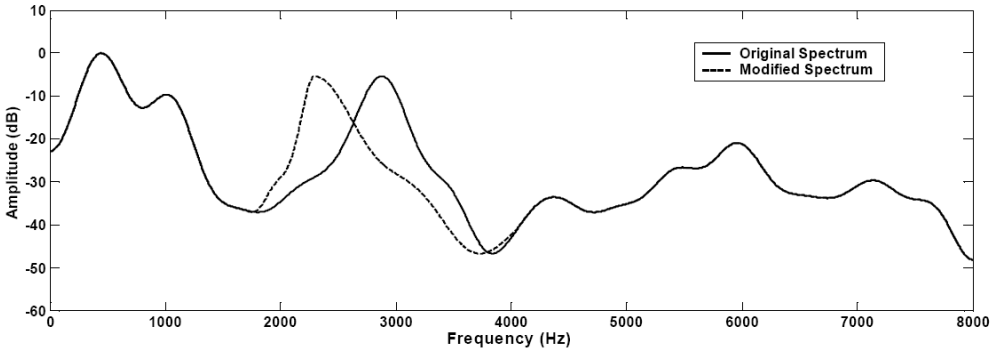

Abbildung 5.5: *Frequency Warping*: Beispiel einer Formantverschiebung [13].

Frequency Warping ermöglicht eine gute Kontrolle über die Formanten, solange die nicht zu modifizierenden und die zu modifizierenden Formanten ausreichenden Abstand voneinander aufweisen. Dann können sie als voneinander unabhängig angesehen werden. Abbildung 5.5 zeigt einen solchen Fall. Sind die Formanten zu dicht beieinander, beeinflussen sie sich gegenseitig. Damit ist dann die Veränderung des einen Formanten mit der unerwünschten Veränderung des nicht zu modifizierenden Formanten verbunden. Auch ist es mit *Frequency Warping* nicht möglich, Formanten zu verschmelzen oder zu trennen.

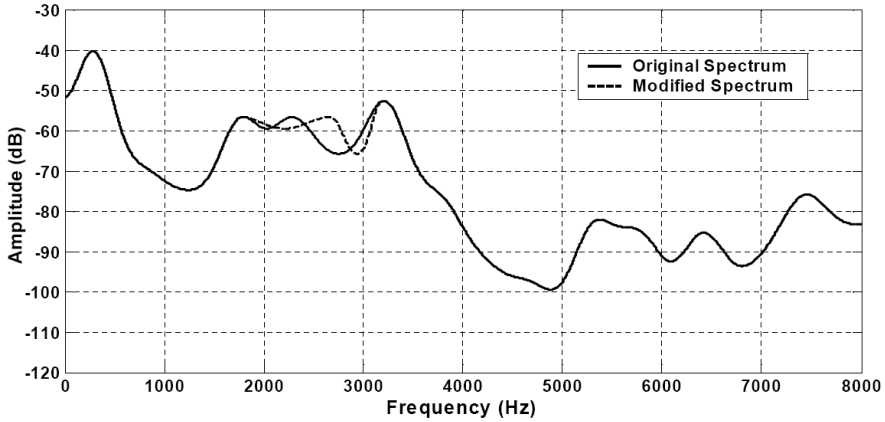

Abbildung 5.6: Probleme beim *Frequency Warping*: Verschiebung von F3 von 2300 Hz nach 2700 Hz, F3 löst sich nicht von F2 (1800 Hz) und verschmilzt nicht mit F4 (3200 Hz) [13].

In Abbildung 5.6 wurde der dritte Formant F3 von 2300 Hz nach 2700 Hz mittels *Warping* verschoben. Dabei sieht man, dass sich F3 nicht von F2 (Mittenfrequenz 1800 Hz) lösen kann und auch nicht mit F4 (Mittenfrequenz 3200 Hz) verschmilzt [13, S. 28].

5.2.3 LPC– Pole

Formanten können, wie bereits in Kapitel 5.1.1 ausgeführt, mit Hilfe der linearen Prädiktion durch Polstellen modelliert werden. Dazu wird das LP– Polynom faktorisiert, die Nullstellen ermittelt und den Resonanzen im Vokaltrakt können die entsprechenden LP– Polynom Nullstellen zugeordnet werden. Durch Modifikation dieser komplexen Nullstellen ist ein direkter Zugriff auf die spektrale Einhüllende des Signals möglich. Nach inverser Filterung mit dem neuen LP– Synthesefilter ergibt sich eine neue Formantstruktur. Die Nullstellen des LP– Polynoms $A(z)$ sind konjugiert komplexe Polpaare $z_i = r_i \cdot e^{j\phi_i}$, aus denen sich die Formantfrequenzen,

$$F_i = \frac{\phi_i}{2\pi} \tag{5.15}$$

und die Bandbreiten der Formanten,

$$B_i = -\frac{ln(r_i)}{\pi} \tag{5.16}$$

schätzen lassen. Ist die LP– Filterordnung p doppelt so groß wie die Anzahl der Formanten im Frequenzbereich von 0 bis $f_s/2$ Hz, entsprechen die Nullstellen z_i in etwa den Formanten. Ist p größer als die Anzahl der Formanten, können einige Nullstellen z_i neben den Formanten auch sogenannte falsche Pole repräsentieren, die kleineren „unwichtigen“ *Peaks* im Spektrum und reellen Polen bei 0 und $f_s/2$ Hz entsprechen. Die reellen Pole stellen die Steigung der spektralen Einhüllenden dar. Falsche Pole besitzen eine große Bandbreite. Im Falle, dass p kleiner als die Anzahl der Formanten ist, werden einige nebeneinanderliegende Formanten im höheren Frequenzbereich zusammengefasst und durch einen Pol repräsentiert. Beim Bestimmen der Formanten ist es wichtig die falschen Formanten auszusortieren [15, S. 123]. Dies kann durch das Ausschließen von Polen geschehen, die ein großes Bandbreiten/ Frequenz– Verhältnis besitzen.

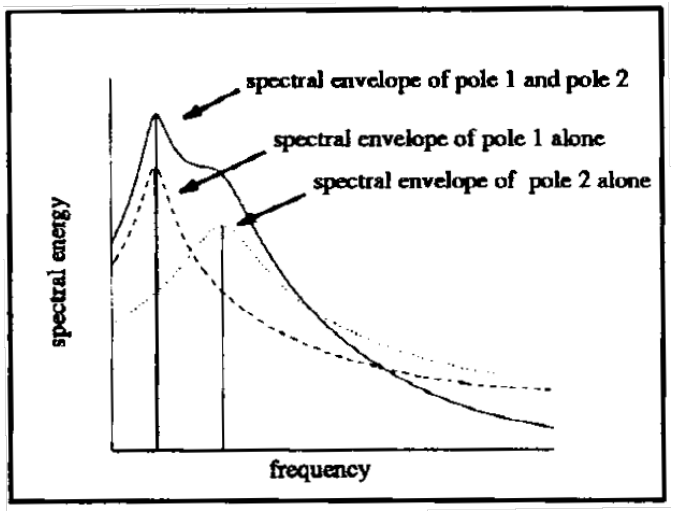

Abbildung 5.7: Beispiel für Polinteraktion.

Liegen Resonanzen bzw. Pole in zu geringem Abstand nebeneinander können

diese nicht mehr aufgelöst werden und erscheinen als nur ein *Peak* im Spektrum [61]. Dies wird in der Literatur auch als Polinteraktions– Problem bezeichnet (vgl. Abbildung 5.7).
Es gibt mehrere Ansätze, dieses Problem zu lösen. Hsiao und Childers [62] modifizieren die Bandbreiten der Formantpole, ohne die allgemeine Formantstruktur zu verändern. Angenommen, es gibt zwei Pole z_i und z_j mit den entsprechenden Winkeln ϕ_i und ϕ_j und den Radien r_i und r_j, dann schreibt sich das Leistungsdichtespektrum für den Winkel ϕ_i:

$$\left|H(e^{j\phi_i})\right|^2 = \frac{1}{(1-r_i)^2}\Delta\left|H\right|_j . \tag{5.17}$$

Dabei ist $\Delta\left|H\right|_j$ der sogenannte Pol- Interaktions Faktor (PIF),

$$\Delta\left|H\right|_j = \frac{1}{1-2r_j cos(\phi_i - \phi j) + r_j^2} \tag{5.18}$$

der den Interaktionseffekt zwischen den Polen z_i und z_j messen soll. Der Radius r_i des Pols z_i wird nun iterativ so lange verändert, bis das Fehlermaß D,

$$D = \sum_{i=1}^{N} \left|H(e^{j\phi_i})\right|^2 - \left|H'(e^{j\phi_i})\right|^2 , \tag{5.19}$$

welches die Differenz zwischen dem Ziel– Leistungsdichtespektrum $\left|H(e^{j\phi_i})\right|^2$ und dem modifizierten Leistungsdichtespektrum $\left|H'(e^{j\phi_i})\right|^2$ ist, einen bestimmten Schwellwert unterschreitet.

Mizuno, Abe und Hirokawa [63] schlagen eine andere Methode vor, das Polinteraktions–Problem zu lösen. Mit den Methoden von Hsiao und Childers und Mizuno *et al.* ist es nicht möglich, die Amplituden und Bandbreiten der Formanten unabhängig voneinander zu modifizieren.
Morris und Clements [64] verändern die *Line Spectral Frequencies* und können damit Formantfrequenzen und Bandbreiten unabhängig voneinander modifizieren.

Teil II

Praxis

Kapitel 6

Implementierung der Stimmentransformationen

Das in dieser Arbeit implementierte System zur Stimmentransformation wurde in Matlab und dem Sprachanalyse– Programm PRAAT [14] erstellt. Die implementierten Systeme wurden durch Hörtests evaluiert (siehe Kapitel 7).

Sprachmaterial. Die Sprachaufnahmen für sämtliche Experimente stammen aus dem Projekt „Varietäten des Österreichischen Deutsch: Standardaussprache und Varianten der Standardaussprache“, welches am Institut für elektronische Musik und Akustik (IEM) an der Universität für Musik und darstellende Kunst Graz erstellt wurde [65]. Dieses Österreichische Aussprachewörterbuch beinhaltet unter anderem Sprachaufnahmen von neun Frauen und neun Männern aus verschiedenen Regionen Österreichs. Für die vorliegende Arbeit wurden pro SprecherIn jeweils etwa 10 Sekunden dieser mit einer Samplingrate von 22050 Hz aufgenommenen Sprachbeispiele verwendet. Die aufgenommenen Personen befinden sich alle in einem Alter von durchschnittlich etwa 39 Jahren und arbeiten beim Fernsehen oder Hörfunk als professionelle SprecherInnen. In Tabelle 6.1 sind die mittleren Grund-

frequenzen der SprecherInnen in den verwendeten Sprachaufnahmen aufgelistet. Die Herkunftsbezeichnungen bedeuten 1– Burgenland, 2– Kärnten, 3– Niederösterreich, 4– Oberösterreich, 5– Salzburg, 6– Steiermark, 7– Tirol, 8– Vorarlberg und 9– Wien.

Herkunft	1	2	3	4	5	6	7	8	9
Frauen	130.4	170.1	155	141.3	176.9	174.9	200.9	147.7	152.6
Männer	106.4	116.5	100.1	89.5	115.5	117.9	91.6	102.1	94.2

Tabelle 6.1: Mittlere Grundfrequenzen der verwendeten Stimmen in Hz.

6.1 Änderung des Stimmgeschlechts

Wie bereits in Kapitel 3.1 ausgeführt, sind die Grundfrequenz F0 und die Formantstruktur die wichtigsten Charakteristika zum Erkennen des Stimmgeschlechts. Um diese beiden Parameter zu modifizieren, wird ein Algorithmus entwickelt, der TD–PSOLA mit einer Abtastratenmodifikation (*Resampling*) verbindet. Zum einen sind beide Methoden in der Sprachsignalverarbeitung sehr etabliert[1] und zum anderen liefern sie Sprachsyntheseergebnisse mit hoher Qualität [11]. In Abbildung 6.1 ist das Blockschaltbild des implementierten Systems zur Änderung des Stimmgeschlechts dargestellt. Das Sprachsignal wird zuerst in PRAAT analysiert und die Grundfrequenz bestimmt (vgl. Abschnitt 4.1.1). Im Zuge der Bestimmung von F0 werden die einzelnen Grundperioden T_0 mit Pitchmarkern versehen. Da dies nur für stimmhafte Segmente geschieht, müssen die Marker für die stimmlosen und stillen Abschnitte des Signals manuell gesetzt werden. Dazu wird ein fixer Abstand von 15 ms gewählt. Die Sprachdatei wird nun zusammen mit der Textdatei, die

[1] Diese Methoden werden auch in PRAAT [14] zur Umwandlung des Stimmgeschlechts verwendet.

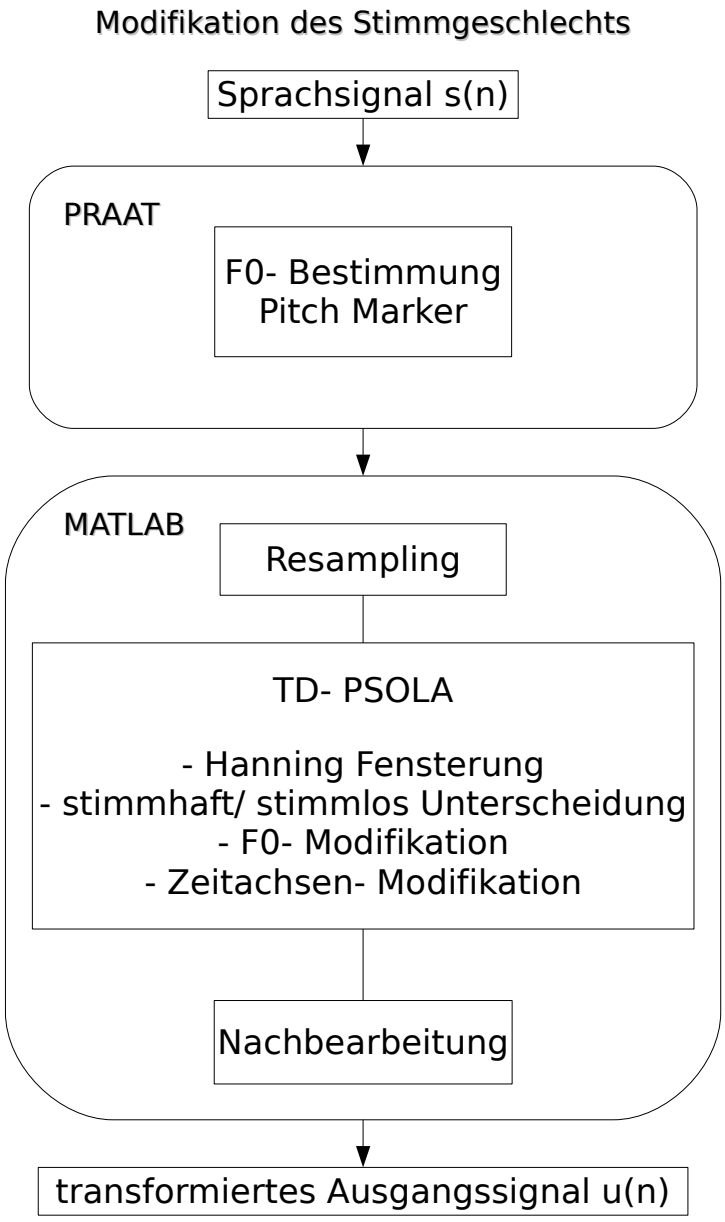

Abbildung 6.1: Blockschaltbild– System für die Änderung des Stimmgeschlechts.

die Positionen der Pitchmarker enthält, nach Matlab exportiert.
Für die Änderung des Stimmgeschlechts werden die Formanten durch Resampling verschoben (siehe Kapitel 5.2.1) und die Grundfrequenz mit TD– PSOLA modifiziert (siehe Kapitel 4.2.1).
Nach Experimenten mit den Sprachaufnahmen der 9 verschiedenen männlichen Sprechern wird für die Mann– Frau Transformation ein Resamplingfaktor von $res_{MF} = 8/9$ gewählt. Wenn man das Signal mit der originalen Samplingfrequenz von $f_s = 22050$ Hz abspielt, erhöhen sich die Formantfrequenzen und gleichzeitig auch die Grundfrequenz um den Faktor $formant_{MF} = 1,125$ bzw. um $+12,5$ %. In weiterer Folge muss die Grundfrequenz noch weiter erhöht werden. Dies geschieht mittels TD– PSOLA. Das Signal wird entsprechend der gesetzten Pitchmarker in Hanning– gefensterte Segmente unterteilt, deren Länge jeweils zwei Pitchperioden entspricht. Um die durch *Resampling* entstandene zeitliche Verkürzung auszugleichen, muss die Zeitachse um den Faktor $\alpha = 1/res_{MF} = 9/8$ gedehnt werden. Für die endgültige Erhöhung der Grundfrequenz wird ein Faktor $\beta_{MF} = 1,7$ gewählt. Damit ergibt sich insgesamt ein F0– Skalierungsfaktor von $f0_{MF} = 1,91$ bzw. in Prozent eine Steigerung von +91 %. Die Modifikation der Grundfrequenz wird beim TD– PSOLA nur für die stimmhaften Segmente durchgeführt.
Die Transformation von Frauen- zu Männerstimmen wird durch dementsprechende

	Skalierungsfaktoren			
Transformationen	Frau– Mann		Mann– Frau	
Grundfrequenz	0.6	−40 %	1.91	+91.0 %
Formanten	0.86	−14 %	1.125	+12.5 %

Tabelle 6.2: Transformation des Stimmgeschlechts: Skalierungsfaktoren für Grundfrequenz und Formanten.

Wahl der Skalierungsfaktoren durchgeführt, das Verfahren bleibt dabei gleich. Der

Resamplingfaktor für die Frau– Mann- Transformation beträgt $res_{FM} = 7/6$, was einer Formantverschiebung um den Faktor $formant_{FM} = 0.86$ bzw. einem Prozentwert von -14 % entspricht. Mit TD– PSOLA wird die Skalierung der Zeitachse um $\alpha = 6/7$ korrigiert. Weiters wird ein F0– Skalierungswert von $\beta_{FM} = 0,7$ gewählt, was zusammengenommen einer Grundfrequenzverringerung um den Faktor $f0_{FM} = 0,6$ entspricht, bzw. in Prozent einer Abnahme um -40 %. In der Tabelle 6.2 sind die oben genannten Skalierungsfaktoren zusammengefasst. Im Zusammen-

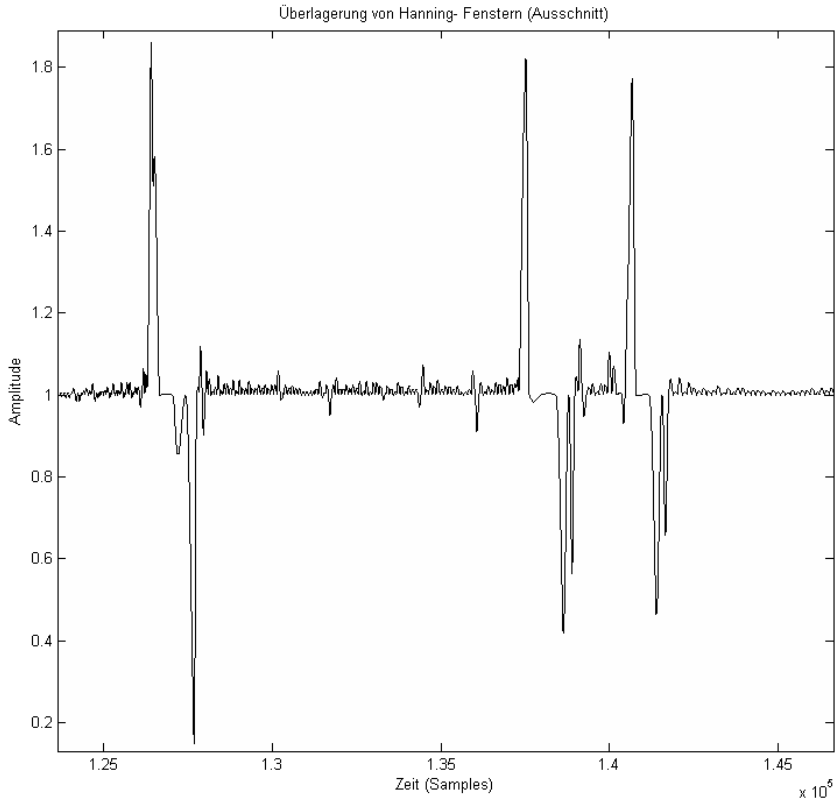

Abbildung 6.2: Überlagerung von Hanning– Fenstern bei Frau– Mann Transformation (Ausschnitt).

hang mit der Änderung der Grundfrequenz entstehen bei der PSOLA Synthese durch die „verschobene“ Überlagerung der gefensterten Segmente Amplitudenper-

turbationen, die sich als Rauigkeit im Klang bemerkbar machen. Die Entstehung dieser Artefakte kann verhindert werden, indem das gesamte modifizierte Ausgangssignal mit der Inversen der überlagerten Hanning– Fensterfunktionen multipliziert wird. Dies geschieht im Nachbearbeitungsblock. In Abbildung 6.2 sieht man einen Ausschnitt der überlagerten Hanning- Fenster am Ausgang der PSOLA Synthese bei der Umwandlung von einer Frauen- in eine Männerstimme. Durch die Verringerung der F0 werden die gefensterten Blöcke im Zeitbereich auseinandergezogen und die Überlagerung der Fensterfunktionen ergibt keine konstante Amplitude mit $gain = 1$. Es kommt zu Verstärkungen und Abschwächungen des Pegels.

Sprecherinnen	F0	modif. F0	Sprecher	F0	modif. F0
f1	130.4	92.0	m1	106.4	186.0
f3	155.0	93.7	m2	116.5	200.8
f4	141.3	87.0	m3	100.1	179.0
f5	176.9	100.9	m4	89.5	159.4
f6	174.9	107.6	m6	117.9	207.6
f7	200.9	115.8	m7	91.6	169.4
f8	147.7	91.1	m8	102.1	189.7
f9	152.6	102.2	m9	94.2	174.8
mittlere F0	159.9	98.8	mittlere F0	102.3	183.3

Tabelle 6.3: Transformation des Stimmgeschlechts: Grundfrequenz– Modifikation (analysiert mit PRAAT); mittlere F0 in Hertz vor und nach der Transformation.

6.2 Alterung der Stimme

Für die Wahrnehmung des Stimmalters sind, wie bereits in Kapitel 3.2 ausgeführt, hauptsächlich die Parameter Grundfrequenz, Grundfrequenzstabilität, Formantstruktur und die Sprechgeschwindigkeit wichtig. Zunächst wurden aber einige Experimente mit der Amplitudenmodulation (AM) durchgeführt, die den Shimmer in einer alten Stimme simulieren sollte. Die Ergebnisse waren jedoch nicht zufriedenstellend, was die Wahrnehmung des Alters und die Natürlichkeit der modifizierten Stimme betraf. Deshalb wurde auf diese Art der Modifikation verzichtet.
Dagegen lieferten Versuche mit Änderungen der Stabilität der Grundfrequenz sehr gute Ergebnisse. Insgesamt werden für die Transformation des Stimmalters die Parameter Grundfrequenz, F0 Stabilität, Formanten und Sprechgeschwindigkeit modifiziert. Das Blockschaltbild des implementierten Systems zur Änderung des Stimmalters ist in Abbildung 6.3 dargestellt. In Verbindung mit TD– PSOLA wird F0 bei der künstlichen Alterung von Frauen um den Faktor $f0_F = 0.9$ verringert. Dazu kommt eine Verringerung der Stabilität der Grundfrequenz mit Hilfe von tiefpassgefiltertem Rauschen $z(n)$. Dies soll die brüchige und zittrige Stimme eines alten Menschen nachbilden. Das gefilterte Rauschen stellt eine Folge von Zufallszahlen dar. F0 schwankt über einen Bereich frei wählbarer Länge und innerhalb frei wählbarer Grenzen. Konkret bedeutet das für die Frauenstimmen folgendes: durch eigene Versuche wurde ein Bereich der Länge von jeweils 5 Segmenten festgelegt, in dem die Grundfrequenzmodifikation konstant gehalten wird. Das heißt, alle 5 Segmente ändert sich der PSOLA Grundfrequenzparameter β um den Wert $z(n)$, was zum Beispiel für die ersten 5 Segmente bedeutet: $\beta = 0.9 + z(1)$. $z(n)$ kann sich in einem wiederum frei wählbaren Bereich von -1 bis $+1$ bewegen. Stellt man nun für $z(n)$ die Ober- und Untergrenzen bei $+0.4$ bis -0.4 ein, dann schwankt die Grundfrequenz in einem Rahmen von $0.5 \leq \beta \leq 1.3$. Nimmt man in einem Signalabschnitt beispielsweise eine mittlere Grundfrequenz von 160 Hz an, dann

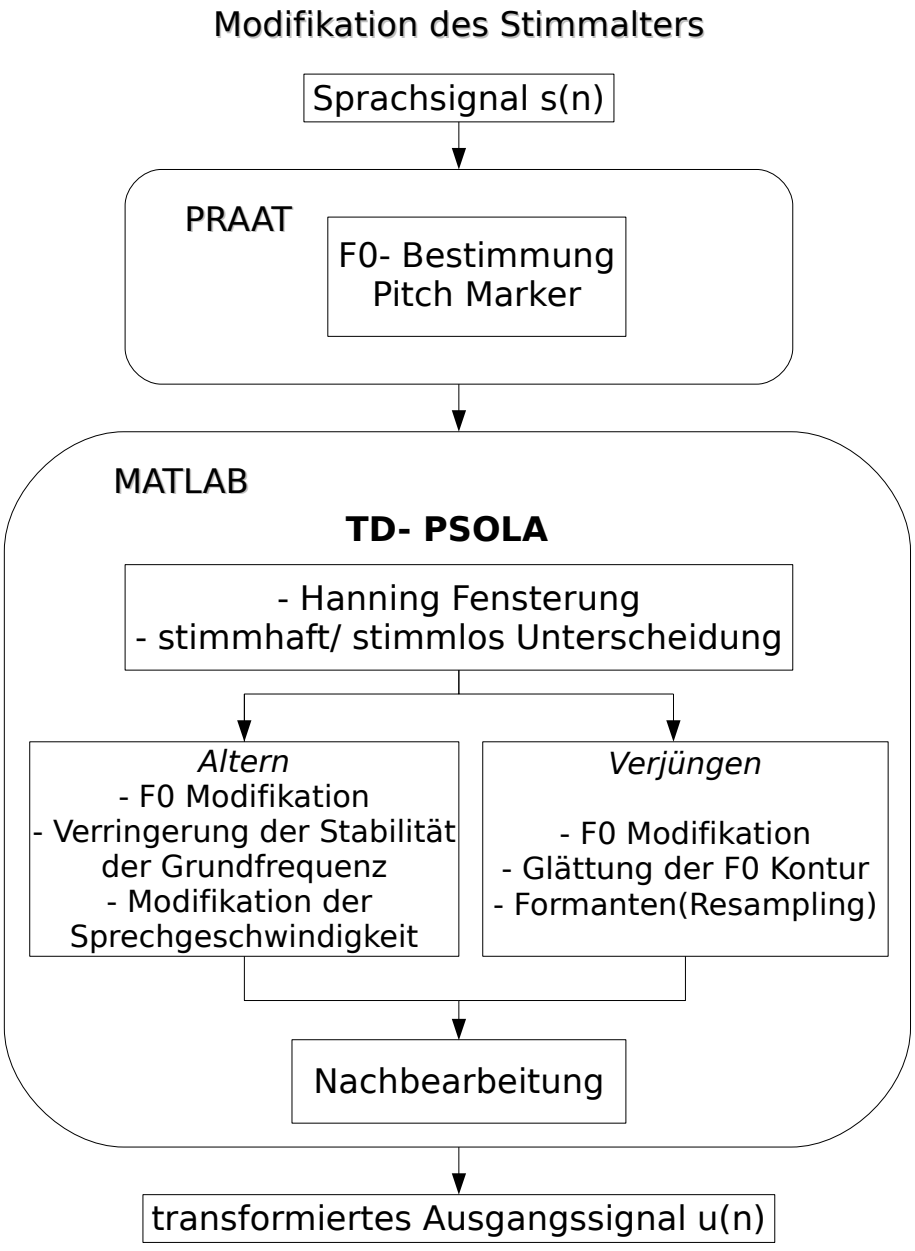

Abbildung 6.3: Blockschaltbild– System für die Änderung des Stimmalters.

kann diese in einem Bereich von 80 Hz bis 203 Hz schwanken. Für die Transformation des Alters von Männerstimmen wird die Grundfrequenz um den Faktor $f0_M = 1.2$ erhöht. Die Anzahl der Segmente in denen die F0– Stabilität konstant bleibt, liegt bei 10 und die Grenzen für $z(n)$ sind mit $+0.35$ und -0.35 festgelegt. Die Grundfrequenz schwankt also in einem Bereich von $0.85 \leq \beta \leq 1.55$, was bei einer mittleren Grundfrequenz von beispielsweise 120 Hz einem Schwankungsbereich von 102 Hz bis 186 Hz entspricht[2].

Zusätzlich wird bei Frauen– und Männerstimmen die Zeitachse um den Faktor $\alpha = 1.2$ gedehnt und damit die langsamere Sprechweise von alten Menschen berücksichtigt. In den Tabellen 6.4 und 6.5 sind die verwendeten Parameter für die künstliche Erhöhung des Stimmalters bei Frauen und Männern zusammengefasst.

Stufen	F0	Range	Anzahl der Segmente	Dehnung der Zeitachse
jung	1	–	5	1
Original	unverändert			
alt	0.9	0.4	5	1.2

Tabelle 6.4: Parameter für die Modifikation des Alters bei Frauen.

Stufen	F0	Range	Anzahl der Segmente	Dehnung der Zeitachse
jung	1	–	10	1
Original	unverändert			
alt	1.2	0.35	10	1.2

Tabelle 6.5: Parameter für die Modifikation des Alters bei Männern.

[2]Im Nachhinein hat sich herausgestellt, dass der Schwankungsbereich für eine natürlicher klingende Stimme etwas kleiner gewählt werden sollte. Außerdem ist es von Vorteil, den Schwankungsbereich von der Höhe der Grundfrequenz abhängig zu machen.

Neben der künstlichen Alterung wurden auch Versuche unternommen, die Stimmen zu verjüngen. Das geschieht bei den Frauen und Männern durch eine Erhöhung der Grundfrequenz und eine leichte Verschiebung der Formanten in den oberen Frequenzbereich. Bei den Frauenstimmen wurde ein Resamplingfaktor von $res_{Fjung} = 16/17$ und bei den Männerstimmen ein Faktor von $res_{Mjung} = 19/20$ gewählt. Das entspricht einer Formantverschiebung und einer Grundfrequenzänderung um den Faktor 1.0625 bzw. 1.0526. Zusätzlich wurde die Grundfrequenz– Kontur (*pitch contour*) mit einem Tiefpassfilter mit einer Grenzfrequenz von $f_g = 330.75$ Hz geglättet und ein Hochpass mit $f_g = 220.5$ Hz verstärkt noch die hohen Frequenzen, um den Eindruck von Jugendlichkeit, Frische und Lebendigkeit zu simulieren.

6.3 Vom Flüstern zum Schreien (Stimmaufwand)

Bei der Modifikation des Stimmaufwandes kommen zu den bereits bei den Transformationen des Stimmgeschlechts und des Stimmalters eingesetzten Verfahren, wie *Resampling* und TD– PSOLA, noch das *Frequency Warping* hinzu. Mit dem bereits in Kapitel 5.2.2 erwähnten *Frequency Warping* ist es möglich, die spektrale Einhüllende direkt zu modifizieren. Die Implementierung des Stimmaufwandes (*vocal effort*) lässt sich in zwei Teile zerlegen. Der erste Teil realisiert 2 Arten von Flüstern und der zweite eine hauchige Stimme (*breathy*) und eine Stimme mit erhöhtem Stimmaufwand (*stressed*). Die Auftrennung erfolgt hinsichtlich der verwendeten Anregung. In Zusammenhang mit dem Quelle Filter Modell und der Linearen Prädiktion wird die geschätzte Anregung beim Modellieren von Flüstern durch gefiltertes Rauschen ersetzt. Bei der hauchigen und gestressten Stimme bleibt die Originalanregung erhalten. Abbildung 6.4 zeigt das implementierte System zur Transformation des Stimmaufwandes. Bei der Veränderung des Stimmaufwandes ging es vor allen Dingen um die Änderung der Grundfrequenz, der

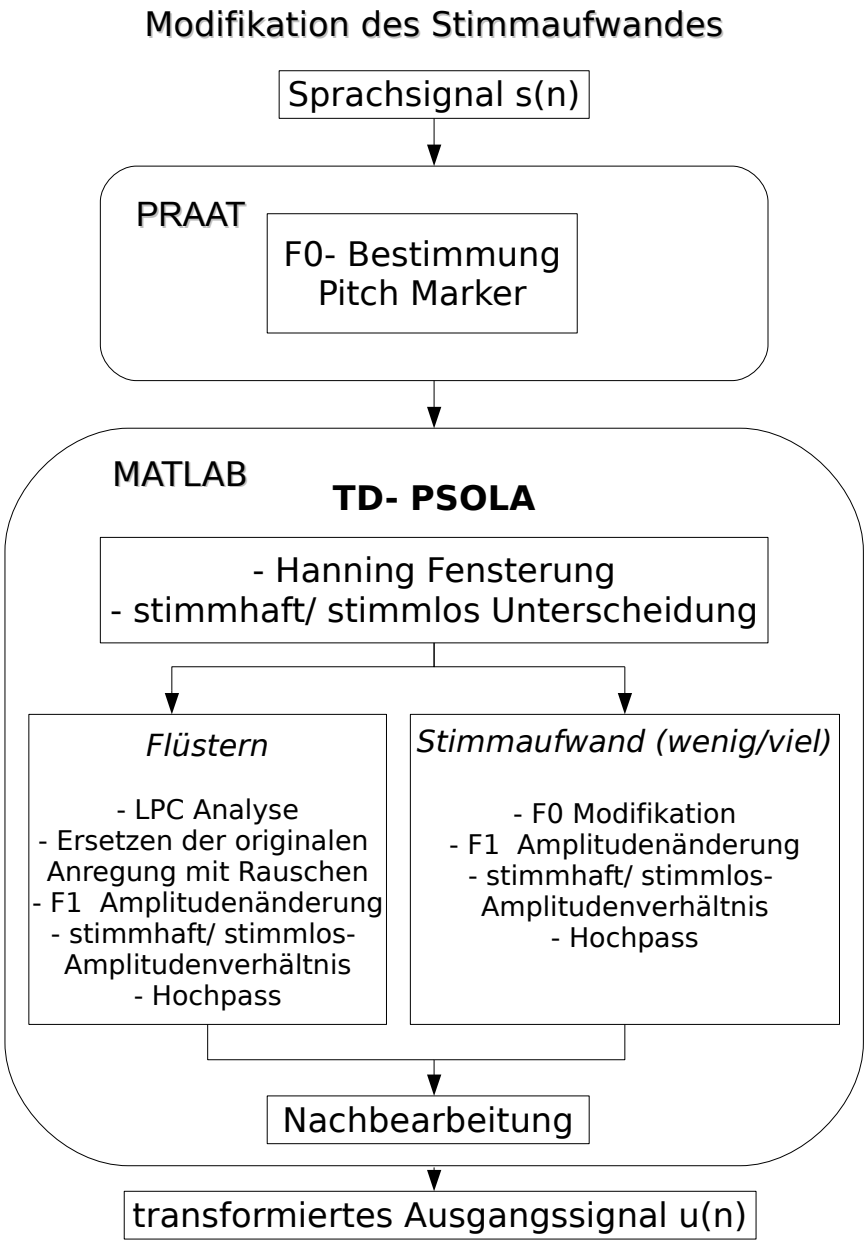

Abbildung 6.4: Blockschaltbild– System für die Änderung des Stimmaufwandes.

spektralen Neigung und um die Veränderung der Amplitude des ersten Formanten. Eine Stimme mit wenig Stimmaufwand besitzt eine hohe F1 Amplitude. Zur Erhöhung dieser wird zuerst die Frequenz von F1 mithilfe des LPC– Spektrums geschätzt. Danach erfolgt eine Erhöhung der Amplitude des ersten Formanten direkt im Betragsspektrum. Um eine minimalphasige Phase zu erhalten, wird der modifizierte Betragsfrequenzgang in den Cepstralbereich überführt, gefenstert und wieder in den Zeitbereich rücktransformiert. Die verwendeten Parameter sind in den Tabellen 6.6 und 6.7 aufgeführt.

Flüstern. Für die Implementierung von Flüstern wird das durch eine LPC–Analyse der Ordnung $p = 27$ gewonnene Anregungssignal durch bandpassgefiltertes Rauschen ersetzt. Der Bandpass lässt Frequenzen von 300 Hz bis 3400 Hz passieren. Zusätzlich durchläuft das Ausgangssignal einen Hochpass mit der Grenzfrequenz $f_{gH} = 275$ Hz. Für die erste Flüster– Variante wird die Amplitude des ersten Formanten erhöht, um den Stimmaufwand noch weiter zu verringern. Beim zweiten Flüstern bleibt die Amplitude des ersten Formanten unverändert. Jedoch ändert sich das Verhältnis der Amplituden von stimmhaften zu stimmlosen Segmenten. Wie in den Tabellen 6.6 und 6.7 zu sehen, erhöht sich die Lautstärke des stimmlosen Anteils. Damit bekommt das Flüstern eine etwas schärfere Klangfarbe und der Stimmaufwand steigt etwas.

Stufen	F0	Amplitudenfaktor von F1	Spektrale Neigung	Stimmhaft/-los Amplitude
Flüstern 1	–	2	Hochpass gain= 1	1/ 1
Flüstern 2	–	–	Hochpass gain= 1	1/ 3
Wenig	0.95	3.3	–	–
Original	–	–	–	–
Viel	1.05	0.33	Hochpass gain= 2	1/ 1.5

Tabelle 6.6: Parameter für die Änderung des Stimmaufwandes bei Frauen.

Stufen	F0	Amplitudenfaktor von F1	Spektrale Neigung	Stimmhaft/-los Amplitude
Flüstern 1	–	2	Hochpass gain= 1	1/ 1
Flüstern 2	–	–	Hochpass gain= 1	1/ 3
Wenig	–	3.3	Tiefpass	–
Original	–	–	–	–
Viel	1.05	0.33	Hochpass gain= 2	1/ 1.5

Tabelle 6.7: Parameter für die Änderung des Stimmaufwandes bei Männern.

Kapitel 7

Hörtest

Im Zuge dieser Arbeit wurde auch ein Hörtest zur Evaluierung der Ergebnisse der durchgeführten Stimmentransformationen durchgeführt. Dieser Test soll eine Einschätzung der Modifikationsalgorithmen in Bezug auf das Erkennen der Transformationen bzw. auf die Plausibilität der gemachten Modifikationen liefern. Beim Testen der Algorithmen zur Änderung des Stimmgeschlechts wird versucht, herauszufinden, inwieweit die Testpersonen die transformierten Stimmproben *(Samples)* aus einer Reihe von originalen Stimmen erkennen können. Dies geschieht mit der *Stimulus Sampling Discrimination*– Methode, die in Abschnitt 7.1.1.1 erläutert wird.

Die Algorithmen für die Modifikation des Stimmalters und des Stimmaufwandes werden in Hinblick auf Plausibilität getestet. Es geht darum, ob bestimmte Parameteränderungen wie Grundfrequenz oder Formantfrequenzen auch Änderungen in der Wahrnehmung von Stimmalter und/ oder Stimmaufwand nach sich ziehen. Beim Stimmalter– Test stellt sich die Frage, ob die Alterstransformationen auch von den Testpersonen als solche erkannt werden. Das Gleiche gilt für den Stimmaufwand– Test. Wird das Erhöhen oder Verringern des Stimmaufwandes auch von den Probanden als Erhöhung oder Verringerung des Stimmaufwandes

wahrgenommen?
In diesem Kapitel werden zuerst die Testdesigns vorgestellt, danach der Ablauf der Hörtests beschrieben, und zum Schluß die Ergebnisse präsentiert.

7.1 Testdesign

Probanden. Für den Hörtest standen 12 Experten (*„expert listeners“*) zur Verfügung, die alle seit mehreren Jahren Elektrotechnik– Toningenieur an der TU Graz sowie an der Universität für Musik und Darstellende Kunst, Graz studieren oder bereits ihr Studium abgeschlossen haben. Zwei Frauen und zehn Männer, die alle zwischen 23 und 29 Jahre alt sind, haben den Test jeweils nacheinander mit Kopfhörer durchgeführt. Die Testdauer betrug im Durchschnitt etwa 40 Minuten.

7.1.1 Stimmgeschlecht– Test

Den Probanden wurden 40 Male zwei Sets *à* 3 Sprachsamples vorgespielt. Ein Set von beiden enthielt in jedem Fall ein modifiziertes Signal und zwei Originale und das andere Set nur Originale. Die HörerInnen hatten die Aufgabe, das Set zu bestimmen, in welchem sich das transformierte Signal ihrer Meinung nach befand. Weiterhin sollten sie die Position des künstlichen Signals 1,2 oder 3 kennzeichnen. Dieser Test wurde mit PRAAT durchgeführt. Abbildung 7.1 zeigt die PC–Testoberfläche.

7.1.1.1 Stimulus Sampling Discrimination (SSD)

Für das Testen der Änderung des Stimmgeschlechts wurde die psychoakustische *Stimulus Sampling Discrimination (SSD)*– Methode gewählt. Die SSD wurde 1987 von Sorkin et al. (zitiert in [66]) für die Quantifizierung von Informationen in Zusammenhang mit Auditiven und Visuellen Displays vorgestellt. Mehrere Forscher,

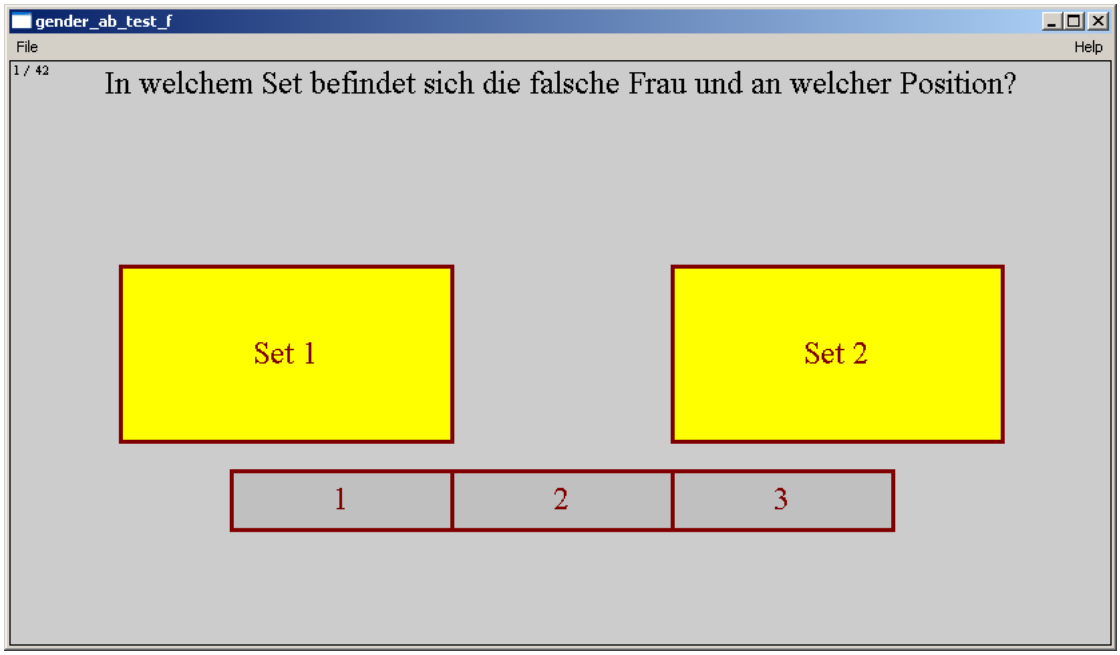

Abbildung 7.1: Testoberfläche für den Stimmgeschlecht– Test (durchgeführt mit PRAAT [14]).

wie Lutfi, M. Mellody und G. H. Wakefield [66, 67] entwickelten diese Methode weiter, indem Konzepte aus der Informations– und Signaldetektionstheorie integriert wurden. Die SSD lässt sich deshalb auch gut für die Evaluierung von Klangsynthesealgorithmen verwenden.

Bei herkömmlichen perzeptiven Diskriminierungstests wird meist die Frage gestellt, ob das synthetische Signal $r \in R$ vom originalen Signal $o \in O$ zu unterscheiden ist. Die Synthesequalität wird dabei über den AB– Vergleich von einzelnen *Samples* o und r aus einem Ensemble von *Samples* O und R bewertet. Dabei kann es passieren, dass die TesthörerInnen sich auf für den Test irrelevante Aufnahmeartefakte im Original konzentrieren. Das kann zum Beispiel Mikrofonrauschen oder Hall sein. Das lässt sich eventuell verhindern, indem man die Testpersonen dementsprechend instruiert, diese Artefakte nicht zu beachten oder indem man beiden *Samples* o und r die entsprechenden Artefakte beifügt. Trotz allem ist bei einem direkten AB– Vergleich die Gefahr groß, dass Unterscheidungen zwischen Original und synthetisiertem Signal aufgrund von sekundären Merkmalen getroffen werden.

Darum ist für Mellody und Wakefield [67] das ideale psychophysikalische Experiment jenes, bei welchem die HörerInnen entscheiden können, ob zwei Klänge von der gleichen Quelle stammen, zum Beispiel zwei Samples von einem Sänger mit dem gleichen Vokal und Grundfrequenz, ohne dass die *Samples* in allen perzeptiven Charakteristika vollkommen übereinstimmen müssen. Folglich wird die *Stimulus Sampling Discrimination* Prozedur für das Messen der Synthesequalität vorgeschlagen.

In welchem Ausmaß zwei verschiedene Samples zu unterscheiden sind, hängt auch vom Kontext ab, in dem beide Proben präsentiert werden. Man spricht hier auch vom Auftreten einer Maskierung der Information „*informational masking*“. Beispielsweise kann die Unterscheidungsschwelle beim Frequenzvergleich zweier Sinustöne unterschiedlich hoch sein, je nachdem in welchem Zusammenhang den Hö-

rerInnen die Töne angeboten werden. Es ist für die Probanden schwieriger einen Unterschied festzustellen, wenn jeder Sinus pro Beobachtungsintervall immer als dritter Ton innerhalb einer Sequenz von 5 Sinustönen präsentiert wird und nicht nur jeweils ein Sinus pro Intervall. Die TesthörerInnen können die unwichtigen Informationen (Sinustöne 1,2,4 und 5) nicht einfach ignorieren und werden vom eigentlich zu bewertenden Sinus 3 abgelenkt. Es kommt zu einer Maskierung von Information.

Die SSD ist eine Methode, bei der Samples von verschiedenen Signalklassen (Original, Synthetisch) verglichen werden. Jedoch findet dieser Vergleich zwischen den Ensembles beider Klassen statt und nicht zwischen einzelnen Ausprägungen dieser Ensembles. Der Proband hört also Samples vom Original– Ensemble und vom Vergleichs– Ensemble und muss entscheiden, in welchem Intervall sich eine Probe vom Vergleichs– Ensemble befindet. Gibt es keine Variation innerhalb der Ensembles, reduziert sich das Experiment zu einem einfachen AB– Vergleich. Gibt es jedoch Variationen, misst man die Fähigkeit der HörerInnen zwischen den statistischen Verteilungen der *Samples* innerhalb jedes Ensembles zu unterscheiden [66].

7.1.2 Stimmaufwand– Test

Um den Algorithmus zur Modifikation des Stimmaufwandes zu bewerten, wurde ein Hörtest in Form eines *Ranking*– Experiments durchgeführt. Dabei bekamen die Probanden fünf verschiedene *Samples*, das Original und 4 unterschiedliche Modifikationen zur Auswahl, welche sie nach der Stärke des Stimmaufwandes von 1 (wenig Stimmaufwand) bis 5 (viel Stimmaufwand) ordnen, d.h. in eine Rangordnung bringen sollten. Insgesamt wurde dieser Test mit 6 verschiedenen Proben von SprecherInnen (3 Frauen, 3 Männer) durchgeführt. Die Testsoftware für den Stimmaufwand– Test (GUI siehe Abbildung 7.2) und den Stimmalter– Test wurde

in modifizierter Form von einer anderen Arbeit [68] übernommen.

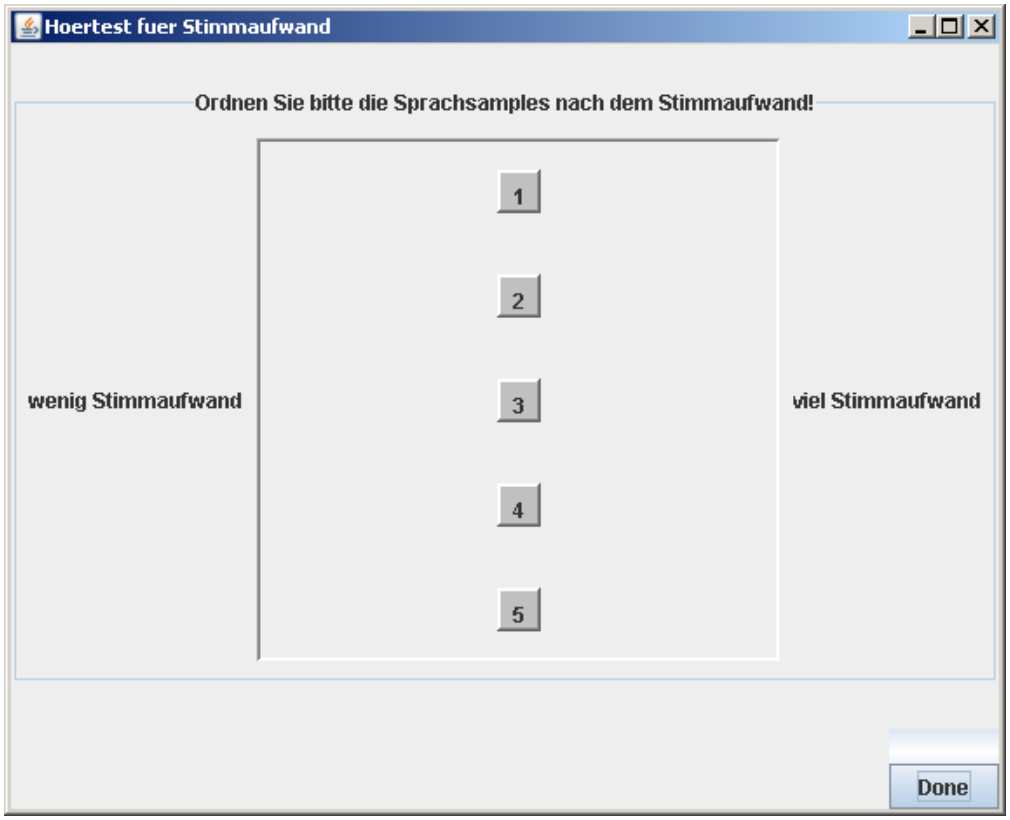

Abbildung 7.2: Testoberfläche für den Stimmaufwand– Test [68].

7.1.3 Stimmalter– Test

Der Stimmalter– Test ist dem Stimmaufwand– Test im Design sehr ähnlich. Er ist auch ein *Ranking*– Hörversuch. Nur dass dieses Mal die Probanden nur 3 verschiedene Samples nach dem wahrgenommenen Stimmalter zu ordnen hatten. Das Original wurde einmal künstlich verjüngt und einmal wurde das Stimmalter erhöht. Dieser Versuch wurde mit den Proben von 10 verschiedenen SprecherInnen (5 Frauen, 5 Männer) durchgeführt.

7.2 Ergebnisse

7.2.1 Stimmgeschlecht

Die Ergebnisse des Tests über die Transformation des Stimmgeschlechts sind in Tabelle 7.1 dargestellt. Die Wahrscheinlichkeit, aus den beiden Sets das richtige mit dem modifizierten Sprachsample auszuwählen, beträgt $p = 0.5$. Es handelt sich hierbei um Ereignisse, die in zwei Alternativen mit gleicher Wahrscheinlichkeit auftreten, entweder ist die Antwort „richtig“ ($p = 0.5$) oder „falsch“ ($p = 0.5$). Der Test lässt sich beliebig oft wiederholen und die Wahrscheinlichkeiten bleiben bei jeder Wiederholung gleich. Deshalb spricht man auch von einem Bernoulli–Experiment. Die Ergebniswahrscheinlichkeiten sind binomialverteilt.
Die Wahrscheinlichkeit, dass bei n Wiederholungen k mal das Ereignis „richtig“ eintrifft, ist

$$P(X = k) = \binom{n}{k} \cdot p^k q^{n-k} k = 0, 1, 2, ..., n. \tag{7.1}$$

$\binom{n}{k}$ wird auch Binomialkoeffizient genannt.

Für die Auswertung der Ergebnisse ist es wichtig zu wissen, wie signifikant die Aussagen sind. Wie groß ist die Wahrscheinlichkeit, dass das Ergebnis durch Raten zustande gekommen ist? Zunächst wird Proband Nummer 3 aufgrund zu großer Abweichungen von den anderen Testpersonen aus dieser Testreihe ausgeschlossen. Die geringe Fehlerrate dieser Testperson kam unter anderem dadurch zustande, dass sie bereits bei einem Vortest zur Verfügung stand. Damit kannte die Testperson bereits die Sprachbeispiele.
Beim Testen der Transformation Frauen– zu Männerstimme haben die Probanden bei 40 *Trials* durchschnittlich etwa 13 mal das Set mit dem modifizierten Signal nicht erkannt. Mit Gleichung (7.1) ergibt sich eine Ratewahrscheinlichkeit von $p = 0.019239$. Das heißt, die modifizierten Stimmsamples konnten von den originalen Stimmen signifikant bei einem Signifikanzniveau von 95 % unter-

	Transformation	
Testpersonen	Frau zu Mann	Mann zu Frau
1	11	4
2	15	6
3	*2*	*1*
4	11	2
5	12	5
6	13	6
7	17	3
8	17	8
9	16	2
10	9	4
11	9	9
12	12	7

Tabelle 7.1: Transformation des Stimmgeschlechts: gemachte Fehler pro Testperson.

schieden werden, nicht jedoch sehr signifikant (99 %). In Zusammenhang mit der Transformation von Männer– zu Frauenstimmen haben die Testpersonen bei 40 *Trials* durchschnittlich 5 Fehler gemacht, was einer Ratewahrscheinlichkeit von $p = 6.9131e - 007$ entspricht. Die transformierten Stimmen wurden von den Probanden bei einem Signifikanzniveau von 99 % mit Sicherheit erkannt.
Abbildung 7.3 zeigt den Boxplot [69] der Fehlerrate[1]. Man erkennt eine wesentlich höhere Fehlerrate bei der Transformation von Männer– zu Frauenstimme als bei der Wandlung von Frauen– zu Männerstimme. Die meisten Probanden konnten demnach die transformierten Männerstimmen schlechter als die transformierten Frauenstimmen erkennen.

Die Abbildung 7.4 gibt einen Überblick über die Abhängigkeit der gemachten Fehler vom jeweiligen Sprecher. Man sieht, dass einige SprecherInnen leichter als „modifiziert“ zu erkennen sind. Die Abkürzungen f und m stehen für weibliche und männliche SprecherInnen. Die Frau– Mann Transformation der Sprecherin f3 wurde zum Beispiel im Vergleich zu den anderen Sprecherinnen am besten erkannt. Bei der Mann– Frau Transformation konnte Sprecher m8 am besten von den Originalen unterschieden werden. Vergleicht man die gemachten Fehlerraten mit den Grundfrequenzen der jeweiligen originalen Sprachaufnahmen, lässt sich kein signifikanter Zusammenhang herstellen.

7.2.2 Stimmaufwand und Alter

Für die statistische Auswertung der beiden anderen Hörtests wird der eindimensionale Chi Quadrat Test gewählt. Damit können die Unterschiede zwischen den Häufigkeiten der einzelnen Merkmalsabstufungen verglichen werden. Ziel ist es, zu erkennen, ob die Probanden die unterschiedlich modifizierten Sprachproben auch

[1] Fehler bedeutet, die Testperson hat das Set, in dem sich das transformierte Sample befand, nicht erkannt.

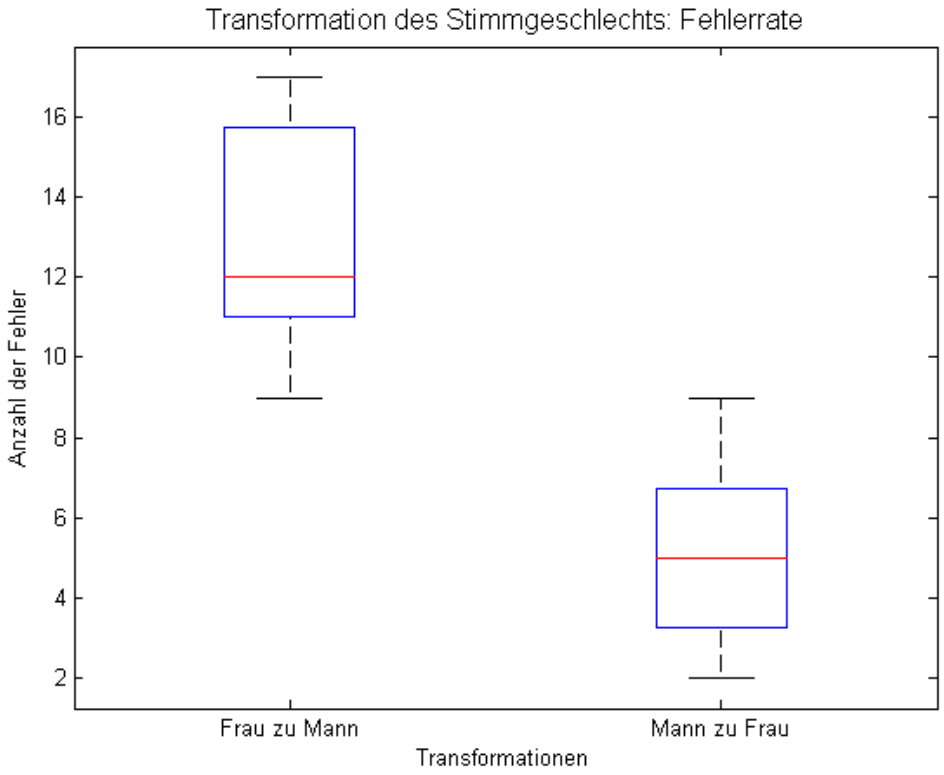

Abbildung 7.3: Boxplot der Fehlerrate.

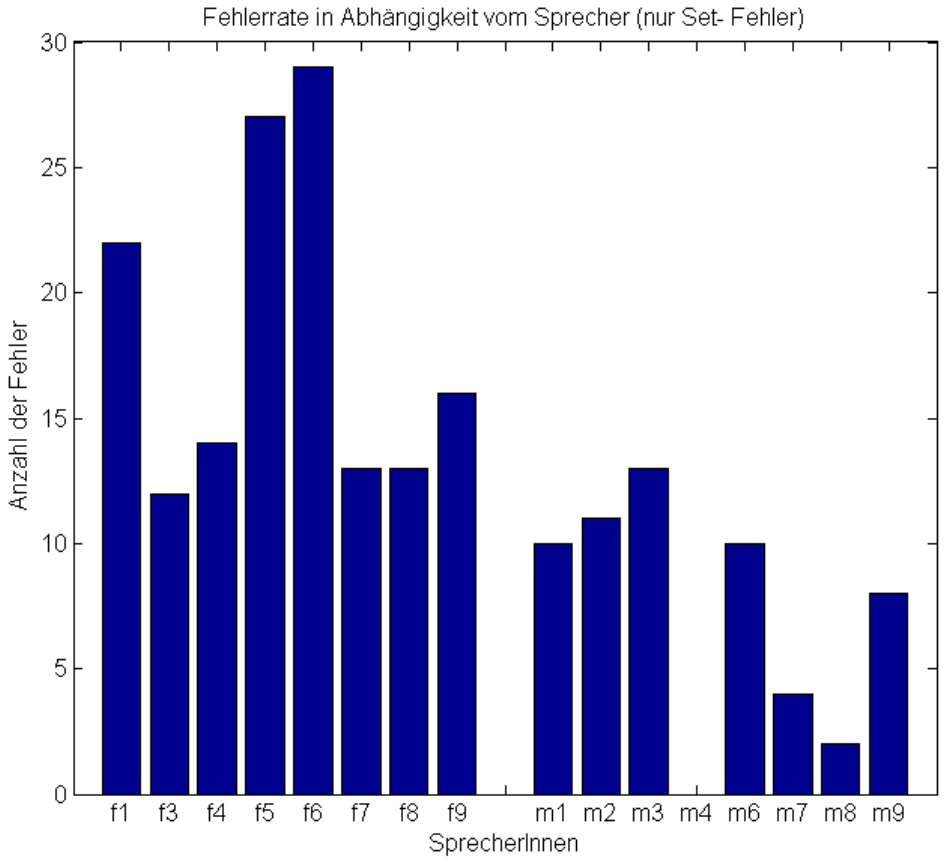

Abbildung 7.4: Transformation des Stimmgeschlechts: Fehlerrate pro SprecherInnen.

den jeweiligen Eigenschaften von wenig Stimmaufwand zu viel Stimmaufwand oder von jung zu alt signifikant zuordnen konnten. Zählt man nun die Häufigkeiten für alle Probanden bezogen auf die Zuordnung von Stimmaufwand 1 bis 5 zusammen, bekommt man die Anzahl der Proben die an erster Stelle (wenig Stimmaufwand) auch die Probe mit dem geringsten Stimmaufwand gewählt haben usw. . Wenn man nachweisen kann, dass, zum Beispiel bei Stimmaufwand 1, wo 3 Personen die Probe 2 an erster Stelle und 10 Personen die Probe 1 an erster Stelle gesetzt haben, beide Häufigkeiten gleichverteilt (Nullhypothese H0) oder nicht gleichverteilt

	Fehler bei Transformation		Fehler bei Transformation
Sprecherinnen	Frau zu Mann	Sprecher	Mann zu Frau
f1	22	m1	10
f3	12	m2	11
f4	14	m3	13
f5	27	m4	0
f6	29	m6	10
f7	13	m7	4
f8	13	m8	2
f9	16	m9	8

Tabelle 7.2: Transformation des Stimmgeschlechts: Fehlerrate pro SprecherInnen.

(Gegenhypothese H1) sind, dann kann man die Nullhypothese entweder bestätigen oder verwerfen.

Mit der Chi Quadrat Methode vergleicht man die erwarteten mit den beobachteten Häufigkeiten [70]. Dabei repräsentieren die erwarteten Häufigkeiten die Nullhypothese H0. Man spricht hier von einem Vergleich der Häufigkeiten eines k– fach gestuften Merkmals. Der Test auf Gleichverteilung geht von H0 aus, dass jeder Stimmaufwandsstufe gleichviele Sprachsamples zugeordnet werden. Das würde bedeuten, die gemessenen Häufigkeitsunterschiede sind zufällig zustandegekommen. Die Frage ist, ob die für Stimmaufwand 1 bis 5 gewählten Sprachsamples pro Merkmal auch aus unterschiedlichen Stichproben stammen. Zuerst benötigt man

die erwarteten Häufigkeiten. Diese berechnen sich folgendermaßen:

$$\begin{aligned} f_{e(1)} &= 1/5 \cdot 72 = 16.5, \\ f_{e(2)} &= 1/5 \cdot 72 = 16.5, \\ f_{e(3)} &= 1/5 \cdot 72 = 16.5, \\ f_{e(4)} &= 1/5 \cdot 72 = 16.5, \\ f_{e(5)} &= 1/5 \cdot 72 = 16.5. \end{aligned} \tag{7.2}$$

Den χ^2- Wert erhält man durch die Formel

$$\chi^2 = \sum_{j=1}^{k} \frac{(f_{b(j)} - f_{e(j)})^2}{f_{e(j)}}. \tag{7.3}$$

Setzt man die erwarteten und beobachteten Häufigkeiten (siehe Tabelle 7.3) ein, dann erhält man einen χ^2- Wert von $\chi_1^2 = 147.78$ für die erste Stufe des Stimmaufwandes (wenig Stimmaufwand). Für die anderen Stufen 2 bis 5 des Stimmaufwandes ergeben sich entsprechende Werte $\chi_2^2 = 103.92$, $\chi_3^2 = 48.694$, $\chi_4^2 = 79.25$ und $\chi_5^2 = 87.306$. Nun vergleicht man diese Ergebnisse mit dem entsprechenden Wert aus der Tabelle der χ^2- Verteilungen [70, S. 817]. Mit einem Freiheitsgrad von $df = 5 - 1 = 4$ (5 verschiedene Merkmalsausprägungen) und einem gewählten Signifikanzniveau von 99 % ergibt sich aus der Tabelle der χ^2- Wert von $\chi^2_{4,99\%} = 13.2767$. Da alle berechneten χ_i^2– Werte größer als der ausgelesene $\chi^2_{4,99\%}$– Wert sind, wird die Nullhypothese der Gleichverteilung verworfen. Man kann davon ausgehen, dass die Unterschiede in den Zuordnungen der Sprachsamples zu den Stimmaufwandstufen statistisch sehr signifikant sind.

Nach dieser Gesamtsignifikanz kann man noch testen, ob sich die Häufigkeiten innerhalb einer Stimmaufwands– oder Stimmaltersstufe signifikant unterscheiden. Dazu wird die beobachtete Häufigkeit $f_{b(1)}$ mit dem Durchschnitt der restlichen beobachteten Häufigkeiten verglichen [70, S. 163]. Hier ergeben sich auch signifikante Unterschiede in den Häufigkeiten.

Die Probanden haben die transformierten Sprachsamples den verschiedenen Abstufungen des Stimmaufwandes richtig zuordnen können. Das Gleiche gilt für das Stimmalter (siehe Abbildung 7.6). Das heißt, die Modifikationen der Stimmmerkmale werden monoton auf die Wahrnehmung von Stimmalter und Stimmaufwand abgebildet.

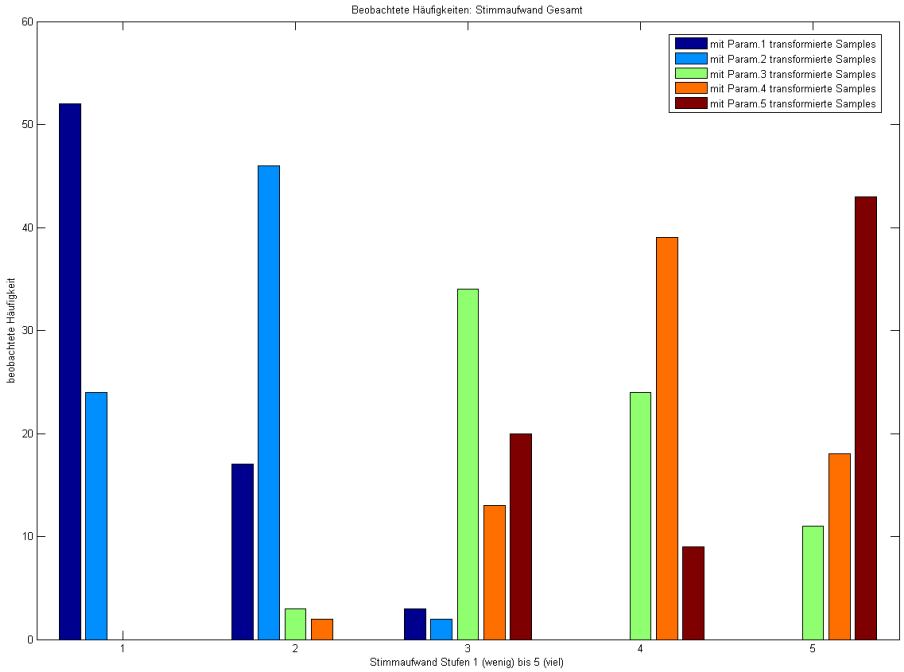

Abbildung 7.5: Stimmaufwand: Beobachtete Häufigkeiten Gesamt.

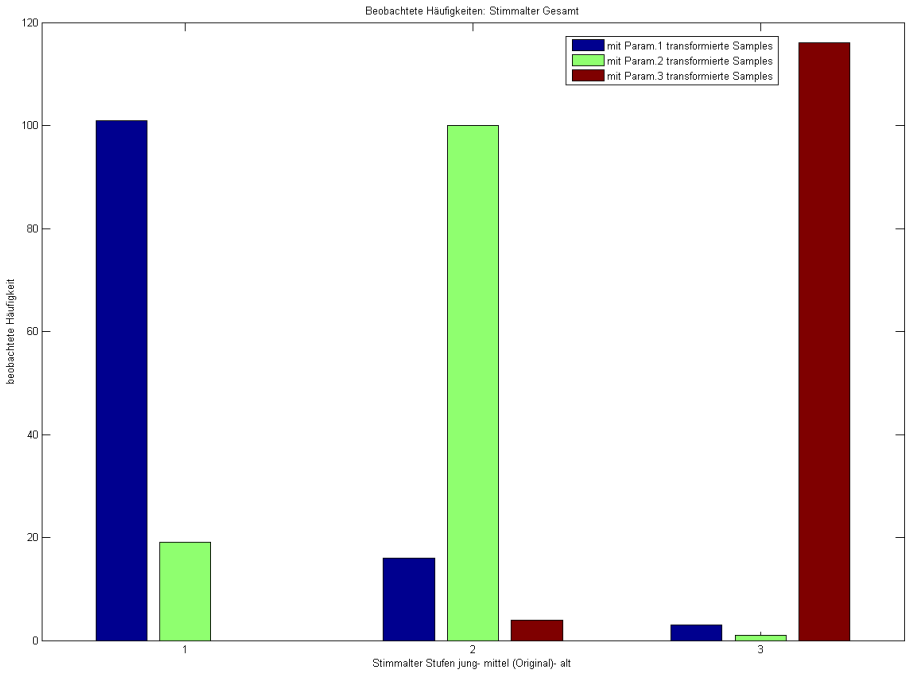

Abbildung 7.6: Stimmalter: Beobachtete Häufigkeiten Gesamt.

	Stufen Stimmaufwand				
transformierte Samples	**1 (wenig)**	**2**	**3**	**4**	**5 (viel)**
Param. 1	52	17	3	0	0
Param. 2	24	46	2	0	0
Param. 3	0	3	34	24	11
Param. 4	0	2	13	39	18
Param. 5	0	0	20	9	43

Tabelle 7.3: Beobachtete Häufigkeiten Stimmaufwand.

	Stufen Alter		
transformierte Samples	**jung**	**mittel**	**alt**
Param. 1	101	16	3
Param. 2	19	100	1
Param. 3	0	4	116

Tabelle 7.4: Beobachtete Häufigkeiten Alter Gesamt.

Kapitel 8

Zusammenfassung und Ausblick

Diese Diplomarbeit beschäftigt sich mit Transformationen der menschlichen Stimme. Auf Grundlage der wissenschaftlichen Literatur werden Algorithmen zur Wandlung des Stimmgeschlechts, zur Transformation des Stimmalters und des Stimmaufwandes entwickelt. Die Algorithmen arbeiten im Zeitbereich und basieren in erster Linie auf der TD– PSOLA Methode.

Für die Transformation des Stimmgeschlechts werden die Grundfrequenz F0 und die Formanten der verwendeten Sprachaufnahmen modifiziert. F0 wird mit TD– PSOLA und die Formanten werden mit einer Abtastratenmodifikation (*resampling*) verändert. Die verwendeten Skalierungsfaktoren sind mit Hilfe der wissenschaftlichen Literatur und aus eigenen Experimenten gefunden worden.

Die Veränderung des wahrgenommenen Stimmalters wird durch die Modifikation der Merkmale Grundfrequenz, Stabilität der Grundfrequenz, Grundfrequenz–Kontur und Zeitdauer vollzogen. Für die Verjüngung der Stimme wird F0 bei beiden Geschlechtern mit TD– PSOLA erhöht und die Grundfrequenz– Kontur geglättet. Beim künstlichen Altern verringert sich die Grundfrequenz bei Frauen

und erhöht sich bei Männern. Die Stabilität der Grundfrequenz und die Sprechgeschwindigkeit nehmen bei Frauen und Männern gleichermaßen ab.

Der Stimmaufwand wird mit der Modifikation der Parameter Grundfrequenz, Amplitude des ersten Formanten F1, Spektrale Neigung *(Spectral Tilt)* und Stimmhaft/ Stimmlos– Amplitude (beim Flüstern) transformiert. Mit erhöhtem Stimmaufwand erhöht sich auch die Grundfrequenz und die höheren Anteile des Spektrums werden verstärkt. Um eine Stimme mit geringerem Stimmaufwand zu erhalten, wird die Amplitude des ersten Formanten erhöht und bei Frauen die Grundfrequenz geringfügig verringert. Beim Flüstern wird die ursprüngliche Anregung durch bandpassgefiltertes Rauschen ersetzt.

Die Algorithmen werden in einem Hörversuch getestet. Die Ergebnisse des Stimmaufwand– Tests und des Stimmalter– Tests zeigen eine große Übereinstimmung der Wahrnehmung des Stimmaufwandes bzw. des Stimmalters mit dementsprechenden Merkmalsänderungen. Die Parameteränderungen werden demnach monoton auf die Wahrnehmung von Stimmalter und Stimmaufwand abgebildet. Bei der Umwandlung des Stimmgeschlechts sind jedoch Verbesserungen in Hinblick auf Sprachqualität der transformierten Stimmproben notwendig. Die Testpersonen konnten die transformierten Proben von den originalen Proben signifikant unterscheiden. Dabei zeigte sich, dass die Transformation von Männer– zu Frauenstimme leichter zu erkennen war, als die Transformation von Frauen– zu Männerstimme.

Die Algorithmen wurden für alle Transformationen der Aufnahmen der 9 Frauen– bzw. Männerstimmen mit den jeweils gleichen Parametern verwendet. Würde man das Parameterset auf eine konkrete Stimmprobe hin optimieren, ließe sich die Qualität der Transformation steigern. Es gibt weitere Möglichkeiten der Verbesserung.

Die Performance von TD– PSOLA hängt sehr stark von der Genauigkeit der Bestimmung der Grundfrequenz ab. In dieser Arbeit wurde F0 mit Hilfe von PRAAT bestimmt. Vielleicht sind andere Grundfrequenzbestimmungsalgorithmen besser geeignet [71]. Für eine unabhängige Modifikation von spektraler Einhüllenden und Anregungssignal ist die Verwendung von LP– PSOLA von Vorteil.
Eventuell lässt sich mit parametrischen Methoden, wie zum Beispiel dem Harmonic+ Noise Modell, STRAIGHT oder dem Auto– Regressive eXogenous Liljencrant–Fant (ARX–LF) Modell [72] eine bessere Synthesequalität erreichen.

Anhang A

Weitere Ergebnisse des Hörtests

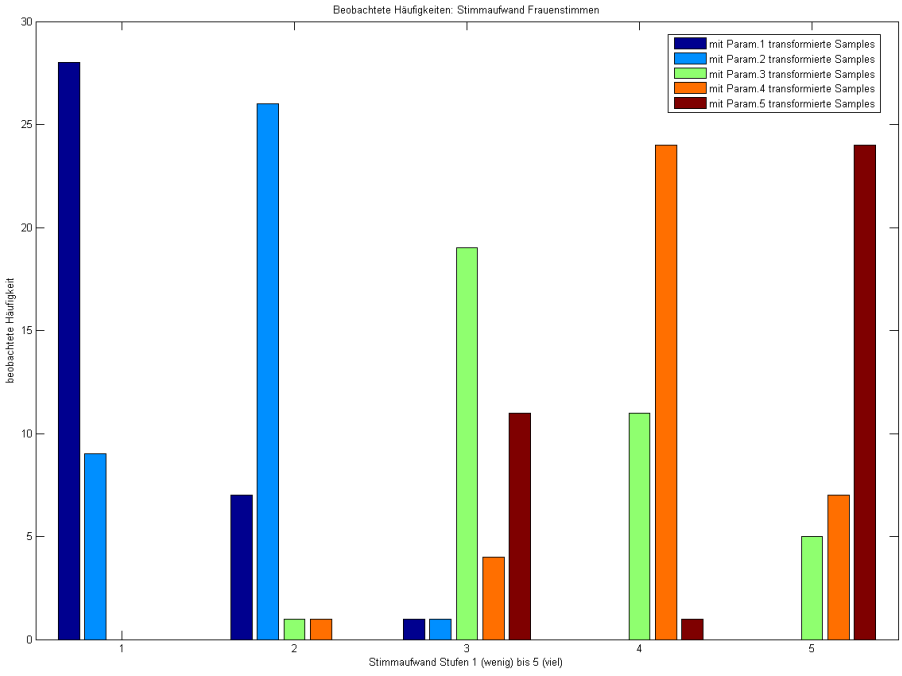

Abbildung A.1: Beobachtete Häufigkeiten: Stimmaufwand Frauenstimmen.

	Stufen Stimmaufwand				
transformierte Samples	**1 (wenig)**	**2**	**3**	**4**	**5 (viel)**
Param. 1	28	7	1	0	0
Param. 2	9	26	1	0	0
Param. 3	0	1	19	11	5
Param. 4	0	1	4	24	7
Param. 5	0	0	11	1	24

Tabelle A.1: Beobachtete Häufigkeiten: Stimmaufwand Frauenstimmen.

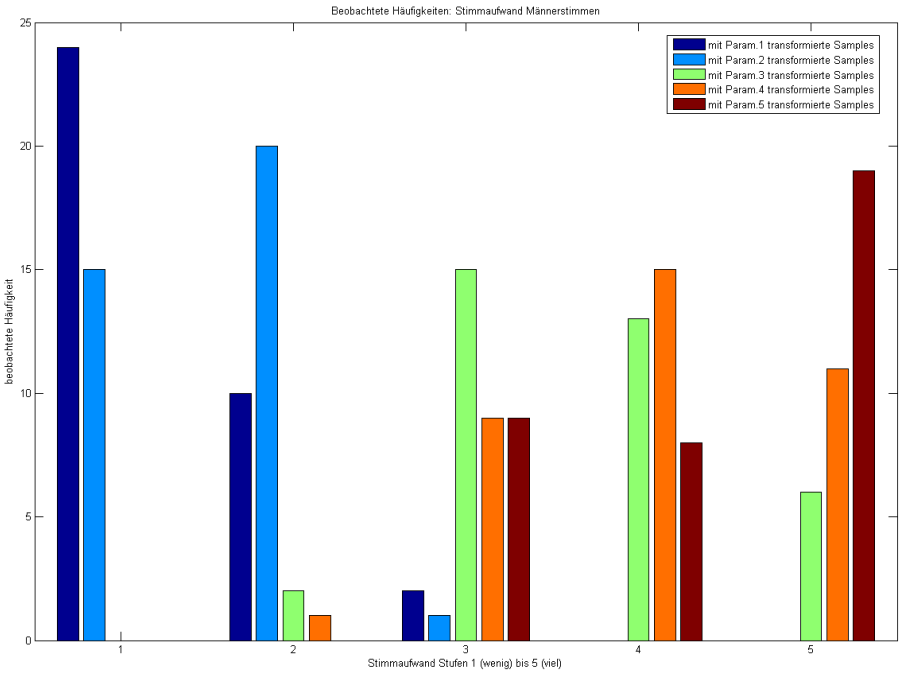

Abbildung A.2: Beobachtete Häufigkeiten: Stimmaufwand Männerstimmen.

	Stufen Stimmaufwand				
transformierte Samples	**1 (wenig)**	**2**	**3**	**4**	**5 (viel)**
Param. 1	24	10	2	0	0
Param. 2	15	20	1	0	0
Param. 3	0	2	15	13	6
Param. 4	0	1	9	15	11
Param. 5	0	0	9	8	19

Tabelle A.2: Beobachtete Häufigkeiten: Stimmaufwand Männerstimmen.

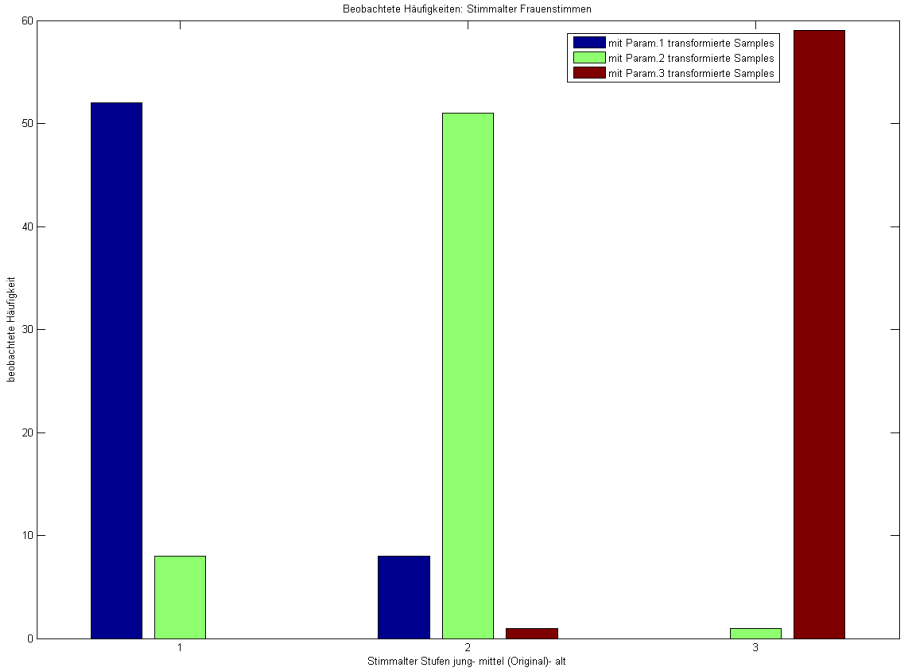

Abbildung A.3: Beobachtete Häufigkeiten: Alter Frauenstimmen.

	Stufen Alter		
transformierte Samples	**jung**	**mittel**	**alt**
Param. 1	52	8	0
Param. 2	8	51	1
Param. 3	0	1	59

Tabelle A.3: Beobachtete Häufigkeiten: Alter Frauenstimmen.

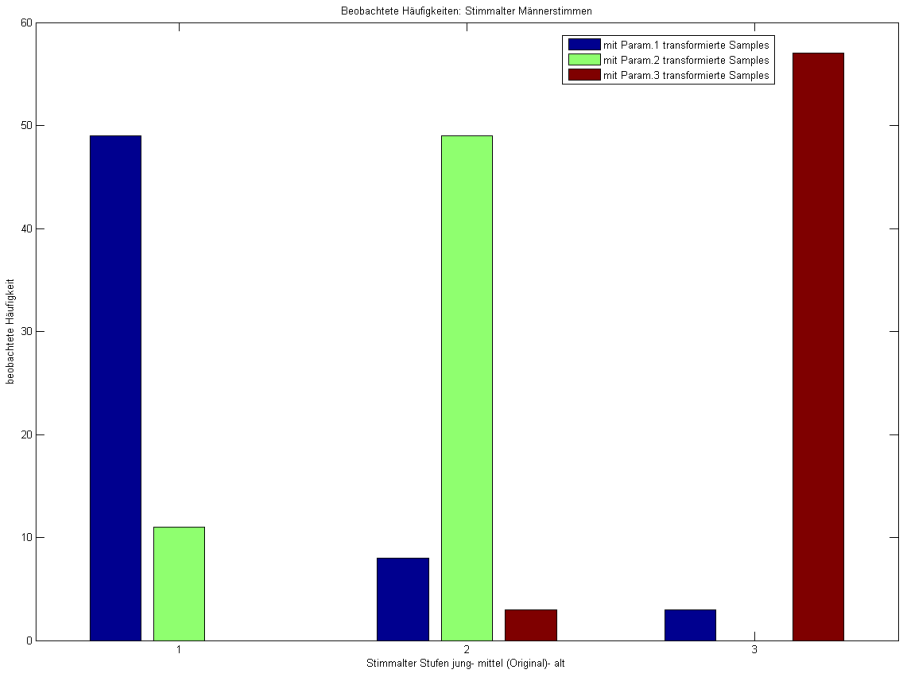

Abbildung A.4: Beobachtete Häufigkeiten: Alter Männerstimmen.

	Stufen Alter		
transformierte Samples	**jung**	**mittel**	**alt**
Param. 1	49	8	3
Param. 2	11	49	0
Param. 3	0	3	57

Tabelle A.4: Beobachtete Häufigkeiten: Alter Männerstimmen.

Literaturverzeichnis

[1] P. Vary, U. Heute, and W. Hess, *Digitale Sprachsignalverarbeitung*, P. Vary, Ed. B.G. Teubner Stuttgart, 1998.

[2] K. Kiesler, "Stimmstörung und heiserkeit," http://www.meduni-graz.at/phoniatrie/diagnostik10.htm, Oktober 2007. [Online]. Available: http://www.meduni-graz.at/phoniatrie/diagnostikkeit/10.htm

[3] P. Vary and R. Martin, *Digital Speech Transmission. Enhancement, Coding and Error Concealment.* Wiley, 2006.

[4] M. Iseli, Y.-L. Shue, and A. Alwan, "Age, sex and vowel dependencies of acoustic measures related to the voice source," *J. Acoust. Soc. Am.*, vol. April 121(2283- 2295), 2007.

[5] M. Brückl and W. Sendlmeier, *Stimmlicher Ausdruck in der Alltagskommunikation, Mündliche Kommunikation.* Logos Verlag Berlin, 2005, ch. Junge und alte Stimmen.

[6] K. I. Nordstrom and P. F. Driessen, "Variable pre-emphasis lpc for modeling vocal effort in the singing voice," *DAFx-06*, 2006.

[7] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *J. Acoust. Soc. Am.*, vol. 107, 2000.

[8] D. O'Shaughnessy, *Speech Communication- Human and Machine.* IEEE Press, 2000.

[9] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *IFA Proceedings 17*, 1993.

[10] U. Zölzer, Amatriain, and Arfib..., *DAFX- Digital Audio Effects*, U. Zölzer, Ed. John Wiley & Sons, LTD, 2002.

[11] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, 1995.

[12] T. Bäckström, "Linear predictive modelling of speech - constraints and line spectrum pair decomposition," Ph.D. dissertation, Helsinki University of Technology, 2004.

[13] M. E. Lee, "Acoustic models for the analysis and synthesis of the singing voice," Ph.D. dissertation, Georgia Institute of Technology, 2005.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 4.5.13) [computer program]." http://www.praat.org/, 2007.

[15] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," Ph.D. dissertation, Institute of Phonetics, Saarland University, 2004.

[16] D. G. Childers, *Speech Processing and Synthesis Toolboxes.* John Wiley & Sons, Inc., 2000.

[17] M. Tang, C. Wang, and S. Seneff, "Voice transformations: From speech synthesis to mammalian vocalizations," *Presented at the 7th European Conference on Speech Communication and Technology, Aalborg, Denmark,*, 2001.

[18] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, vol. March, VOL. 6, NO. 2, 1998.

[19] M. Abe, "A segment-based approach to voice conversion," *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pp. 765–768 vol.2, 1991.

[20] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum," *IEEE*, 2001.

[21] Y. Stylianou, "Voice transformation." Interspeech, 2007.

[22] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, A. V. Oppenheim, Ed. Prentice Hall, 1978.

[23] http://de.wikipedia.org/wiki/Diphthong.

[24] D. W. Griffin and J. S. Lim, "Multiband- excitation vocoder," *IEEE*, 1988.

[25] J. Laroche, Y. Stylianou, and E. Moulines, "Hns: Speech modification based on a harmonic + noise model," *IEEE*, 1993.

[26] Y. Stylianou, *Nonlinear Speech Modeling*. Springer Verlag, 2005, ch. Modeling Speech Based on Harmonic Plus Noise Models.

[27] H. KAWAHARA, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," *IEEE*, 1997.

[28] H. Kawahara, "Exemplar-based voice quality analysis and control using a high quality auditory morphing procedure based on straight."

[29] http://www.wakayama-u.ac.jp/ kawahara/PSSws/, October 2007.

[30] K. Wu and D. G. Childers, "Gender recognition from speech. part i: Coarse analysis," *J. Acoust. Soc. Am.*, vol. 90, no. 4, pp. 1828–1839, 1991.

[31] G. Fant, "Vocal tract energy functions and non- uniform scaling," *J. Acoust. Soc. Jpn.*, vol. 11, 1976.

[32] M. Hirano, J. Kurita, and T. Nakahima, "Growth, development and aging of human vocal folds," *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, 1983.

[33] I. R. Titze, "Physiology of the female larynx," *J. Acoust. Soc. Am.*, vol. 82, 1987.

[34] ——, "Physiologic and acoustic differences between male and female voices," *J. Acoust. Soc. Am.*, vol. 85, 1989.

[35] J. E. Huber, E. T. Stathopoulos, Curione, Ash, and Johnson, "Formants of children, women, and men: The effects of vocal intensity variation," *J. Acoust. Soc. Am.*, vol. 106 , No. 3, 1532-1542, 1999.

[36] S. Lee, A. Potamianos, and S. Narayan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of Acoust. Soc. Amer.*, vol. 105(3), 1999.

[37] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, 1990.

[38] M. Stoicheff, "Speaking fundamental frequency characteristics of non- smoking female adults," *J. Speech Hear. Res.*, vol. 24, 1981.

[39] http://de.wikipedia.org/wiki/Koartikulation. [Online]. Available: http://de.wikipedia.org/wiki/Koartikulation

[40] H. Traunmüller, A. Eriksson, and L. Menard, "Perception of speaker age, sex and vowel quality investigated using stimuli produced with an articulatory model," *Speech Commun.*, 1984.

[41] P. Ptacek and E. Sander, "Age recognition from voice," *Journal of Speech and Hearing Research*, vol. 9, 1966.

[42] H. Traunmüller, "Perception of speaker sex, age, and vocal effort."

[43] S. Schötz, "Perception, analysis and synthesis of speaker age," Ph.D. dissertation, Lund University, 2006.

[44] M. Brückl, "Altersbedingte veränderungen von frauenstimmen. eine akustische und perzeptive analyse," Master's thesis, Institut für Kommunikationswissenschaft, Technische Universität, Berlin, 2002.

[45] S. Linville and H. Fisher, "Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females," *J. Acoust. Soc. Am.*, vol. 78, 1985.

[46] S. E. Linville, "The sound of senescence," *Journal of Voice*, vol. 10, 1996.

[47] T. Shipp, Y. Qi, R. Huntley, , and H. Hollien, "Acoustic and temporal correlates of perceived age," *Journal of Voice*, 1992.

[48] S. E. Linville, "Acoustic- perceptual studies of aging voice in women," *Journal of Voice*, vol. 1, 1987.

[49] J.-S. Lienard and M.-G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *J. Acoust. Soc. Am.*, vol. 106(411-422), 1999.

[50] Bußmann and Hadumod, *Lexikon der Sprachwissenschaft*, 1990.

[51] W. J. Hess, *Pitch determination of Speech Signals.* Springer Verlag, 1983.

[52] L. R. RABINER, "On the use of autocorrelation analysis for pitch detection," *IEEE*, 1977.

[53] X.-D. MEI, J. Pan, and S.-H. SUN, "Efficient algorithms for speech pitch estimation," *Proceedings of 2001 lnternational Symposium on Intelligent Multimedia, video and Speech Processing May 24 2001 Hong Kong*, 2001.

[54] E. Moulines, F. Charpentier, and C. Hamon, "A diphone synthesis system based on time-domain prosodic modifications of speech," *Proceedings ICASSP*, 1989.

[55] E. Moulines and F. Charpentier, "Pitch- synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, 1990.

[56] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE*, 1999.

[57] ——, "Phase-vocoder: About this phasiness business," *IEEE*.

[58] ——, "New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects," *IEEE*, 1999.

[59] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," *DAFx(05)*, 2005.

[60] J. Ajmera, "Effect of age and gender on lp smoothed spectral envelope," *IEEE*, 2006.

[61] J. Makhoul, "Linear prediction: A tutorial review," *IEEE*, 1975.

[62] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE*, 1974.

[63] Y.-S. Hsiao and D. G. Childers, "A new approach to formant estimation and modification based on pole interaction," *IEEE*, 1997.

[64] H. Mizuno, M. Abe, and T. Hirokawa, "Waveform-based speech synthesis approach with a formant frequency modification," *IEEE*, 1993.

[65] R. W. Morris and M. A. Clements, "Modification of formants in the line spectrum domain," *IEEE*, 2002.

[66] R. Muhr, R. Höldrich, and G. Nierhaus, "Varietäten des Österreichischen deutsch: Standardaussprache und varianten der standardaussprache." [Online]. Available: http://iem.at/projekte/dsp/varietaeten/

[67] G. H. Wakefield, "A mathematical/ psychometric framework for comparing the percedptual response to different analysis- synthesis techniques: Ground rules for a synthesis bake-off," *ICMC*, 2000.

[68] M. Mellody and G. H. Wakefield, "A tutorial example of stimulus sample discrimination in perceptual evaluation of synthesized sounds: Discriminationbetween original and re- synthesized singing," *Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland*, 2001.

[69] S. Bech and N. Zacharov, *Perceptual Audio Evaluation- Theory, Method and Application.* John Wiley & Sons, Ltd, 2006.

[70] J. Bortz, *Statistik.* Springer, 2005.

[71] A. de Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111 (4), 2002.

[72] D. Vincent, O. Rosec, and T. Chonavel, "A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and hnm modeling," *IEEE ICASSP*, 2007.