FOTODRUCK J.MAINZ GmbH
Neupforte 13, Nähe Markt, D - 5100 Aachen
Telefon: (0241) 27305 und (0241) 38606

Die Synthese der weiblichen Stimme unter besonderer Berücksichtigung der Phonation

Inaugural - Dissertation
zur
Erlangung des Doktorgrades
der Philosophischen Fakultät
der Universität zu Köln

vorgelegt von

Bernd - Joachim Kröger

aus Osnabrück

Köln 1989

Die Synthese der weiblichen Stimme unter besonderer Berücksichtigung der Phonation

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Philosophischen Fakultät
der Universität zu Köln

vorgelegt von
Bernd-Joachim Kröger
aus Osnabrück

Köln 1989

Inhaltsverzeichnis

Inhaltsverzeichnis III

Vorwort VIII

Einleitung und Überblick IX

I. THEORETISCHER TEIL:
ZUR PHONATION IN SYNTHETISCHER SPRACHE.
AUSWIRKUNGEN AUF STIMMQUALITÄT, GESCHLECHTSASSOZIATION UND VOKALPERZEPTION

1 *Differenzierte phonatorische Parameter als Grundlage für die Synthese weiblicher Stimmen* 1

1.1 Das Grundproblem der Synthese weiblicher Stimmen 1

1.2 Unterschiedliche Synthesemodelle mit differenzierten phonatorischen Parametern 2

1.2.1 Artikulatorische und phonatorische Parameter 2

1.2.2 Physiologie-orientierte und Wellenform-orientierte Parametersätze 5

1.2.3 Die Definition der Wellenform-orientierten Parameter 6

1.2.4 Modelldifferenzierungen aufgrund der Interaktion zwischen Glottis und Sprechtrakt: Die GEV-, GAI- und GMI-Modellgruppen 11

1.3 Systematische Unterschiede der phonatorischen Parameter bei der Synthese der weiblichen und der männlichen Stimme 15

1.4 Vor- und Nachteile der GAI-Modelle und der GMI-Modelle für die Synthese der weiblichen Stimme 17

2 *Stimmqualität und Geschlechtsidentifikation* 20

2.1 Zum Begriff Stimmqualität und Stimmklang 20

2.2 Perzeptiv relevante Stimmqualitätsmerkmale 23

2.3 Die perzeptiv relevanten Signalmerkmale zur Geschlechtsidentifikation 25

2.3.1 Die Grundfrequenz 27
2.3.2 Die mittlere Lage der Formanten 27
2.3.3 Der Stimmklang 29

3 *Vokalperzeption und Geschlechtsidentifikation* 33
3.1 Problemstellung 33
3.2 Das Modell der verschiebbaren Vokalschablone 33
3.2.1 Formanten und Vokalraum 33
3.2.2 Das Modell der verschiebbaren Schablone 36
3.2.3 Normierungsfaktoren und Klassifikationsfaktoren 38
3.2.4 Zur Dimension des Vokalraums 38
3.2.5 Die perzeptive Relevanz des Schablonenmodells 39
3.2.6 Extrinsische und intrinsische Faktoren 39
3.3 Der Zusammenhang zwischen Primärnormierung und Geschlechtsassoziation 40
3.3.1 Die These 40
3.3.2 Hinweise auf die Richtigkeit der These 42
3.3.2.1 Zum geringen Einfluß synthetischer Sprache auf die Verschiebung der Vokalschablone 43
3.3.2.2 Der Einfluß der Grundfrequenz auf die Vokalqualität 46
3.3.2.3 Interindividuelle Unterschiede in der Perzeption synthetischer Vokale 51
3.3.2.4 Das Versagen extrinsischer Normierungsfaktoren bei der Perzeption natürlicher Sprache 52
3.4 Ein Vergleich mit anderen Normierungstheorien 55
3.4.1 Zur perzeptiven Relevanz des Modells der Primärnormierung 55
3.4.2 Normierungsalgorithmen 59
3.4.3 Nichtproportionale Formantskalierung 61
3.4.4 Vokalperzeption anhand dynamischer Signalmerkmale 66

II. EXPERIMENTELLER TEIL:
ZUR SYNTHESE DER WEIBLICHEN STIMME UND
EXPERIMENTELLE UNTERSUCHUNGEN ZUR PERZEPTION
SYNTHETISCHER SPRACHE

4 *Die implementierten Synthesemodelle* 69
4.1 Die Verfahren der Sprachsynthese im Überblick 69
4.1.1 Signalorientierte und erzeugungsorientierte Syntheseverfahren 69
4.1.2 Zur Einordnung des KELLY-LOCHBAUM-Syntheseverfahrens: Formantensynthese, LPC-Synthese und KELLY-LOCHBAUM-Synthese 71
4.2 Das implementierte Sprechtraktmodell und seine Randbedingungen 73
4.2.1 Das KELLY-LOCHBAUM-Modell 73
4.2.1.1 Sprechtraktabschnitte und die Reflexion von Teilwellen 73
4.2.1.2 Simulation des Sprechtrakts in Raum und Zeit 75
4.2.1.3 Der Zusammenhang von Sprechtraktlänge und Samplingfrequenz 78
4.2.1.4 Die Berechnung von Druck und Volumenstrom im KELLY-LOCHBAUM-Modell 79
4.2.2 Verlustmechanismen im Sprechtrakt 80
4.2.3 Randbedingungen des KELLY-LOCHBAUM-Modells 81
4.2.3.1 Elektrische Analogiebildungen 81
4.2.3.2 Die Sprechtraktanregung 82
4.2.3.3 Die Abstrahlung 83
4.3 Die Modellierung der Glottis 85
4.3.1 Die Interaktionsarten zwischen Glottis und Sprechtrakt 85
4.3.2 Das GAI-Modell: Glottismodell mit akustischer Interaktion 87
4.3.2.1 Das Modell 87
4.3.2.2 Die Steuerfunktion 92
4.3.2.3 Einige Eigenschaften des GAI-Modells 94
4.3.2.4 Zur Definition der Analyseparameter 96

4.3.3 Das GEV-Modell: Glottismodell mit fest eingeprägtem Volumenstrom 101
4.3.3.1 Das Modell 101
4.3.3.2 Die Steuerfunktionen 102
4.4 Zur Synthese der Hörteststimuli 104

5 *Experimentelle Untersuchungen zum Einfluß der phonatorischen Parameter auf die Synthese der weiblichen Stimme* 106
5.1 Die Optimierung der phonatorischen Parameter nach HOLMBERG im GAI-Modell 106
5.2 Zur Durchführung der Hörexperimente 111
5.3 Der Einfluß der phonatorischen Parameter auf die Geschlechtsklassifikation und die Natürlichkeit: Experiment 1 112
5.3.1 Das Ziel des Experiments 112
5.3.2 Zur Synthese der Stimuli 113
5.3.3 Zur Durchführung des Experiments 115
5.3.4 Ergebnisse 115
5.4 Der Vergleich von GEV- und GAI-Modell: Experiment 2 118
5.4.1 Das Ziel des Experiments 118
5.4.2 Zur Synthese der Stimuli 118
5.4.3 Zur Durchführung des Experiments 123
5.4.4 Ergebnisse 123
5.5 Die Änderung der phonatorischen Parameter Verschluß-strom und Schließquotient: Experiment 3 und Experiment 4 125
5.5.1 Das Ziel der Experimente 125
5.5.2 Zur Synthese der Stimuli für Experiment 3: Die Änderung des Verschlußstroms 126
5.5.3 Zur Synthese der Stimuli für Experiment 4: Die Änderung des Schließquotienten 129
5.5.4 Zur Durchführung der Experimente 132
5.5.5 Ergebnisse 132

6 *Eine experimentelle Untersuchung zum Einfluß der Geschlechtsklassifikation auf die Vokalqualität: Experiment 5* 136
6.1 Das Ziel des Experiments 136
6.2 Zur Synthese der Stimuli 137
6.3 Zur Durchführung des Experiments 140
6.4 Ergebnisse 141

7 *Zusammenfassung der Ergebnisse und Diskussion* 148
7.1 Zum Einfluß der phonatorischen Parameter auf die Natürlichkeit der weiblichen Stimme 148
7.2 Die Unterschiede in den phonatorischen Einstellungen für die weibliche und die männliche Stimme 151
7.2.1 Die phonatorischne Unterschiede nach HOLMBERG et al. und ihre Verifikation anhand eines Syntheseexperiments dieser Arbeit 151
7.2.2 Ergebnisse aus anderen Untersuchungen mittels der Methode des inversen Filterns 152
7.2.3 Ergebnisse aus anderen Syntheseexperimenten 154
7.2.4 Der Einfluß von Verschlußstrom und Schließquotient 156
7.3 Zum Zusammenhang von Vokalperzeption und Geschlechtsassoziation 164

Verzeichnis der Abbildungen 167
Verzeichnis der Tabellen 171

Literaturverzeichnis 173

Vorwort

Für die Möglichkeit, die Arbeit in größtmöglicher Freiheit durchzuführen, bedanke ich mich bei Herrn Prof. Dr. G. Heike. Ebenso gilt mein Dank Herrn Prof. Dr. J.P. Fricke für die Betreuung der Arbeit.

Herrn Dr. S. Hilger danke ich für die vielen Stunden der Diskussion von Programmierproblemen und Herrn Dr. R. Greisbach für die Erstellung von Ansteuerdaten.

Allen Mitgliedern des IPKöln danke ich für die Unterstützung bei der Durchführung der Arbeit, und den Studenten, die an den Hörtests teilnahmen, danke ich für ihre Geduld und ihr Interesse.

Fürs Korrekturlesen danke ich Frau C. Riek und Ruth.

Einleitung und Überblick

In erster Linie wird Sprachsyntheseforschung mit dem Ziel betrieben, Systeme zu erstellen, die eine hohe Verständlichkeit bei schnellstmöglicher Generierung des Signals gewährleisten. Durch die zunehmende Qualitätssteigerung der mittels Synthese generierten Sprache ist es aber möglich geworden, anhand von Experimenten mit synthetisch generierten Stimuli die ektosemantische Sprachsphäre[1] zu untersuchen. So steht die Modellierung der weiblichen Stimme im Mittelpunkt dieser Arbeit, während in den Fällen, in denen vornehmlich phonetisch-phonologische Fragestellungen behandelt werden, zumeist Sprache in Anlehnung an das Vorbild der männlichen Stimme generiert wird. Die immer vorzunehmende Berücksichtigung geschlechtsspezifischer Signalmerkmale geht dort meist nicht über die Einstellung der Grundfrequenz und die Einstellung der Sprechtraktlänge bzw. der mittleren Lage der Formanten auf die durchschnittlichen Werte männlicher Stimmen hinaus.

Bei der Synthese der weiblichen Stimme tritt das Problem auf, daß die generierten Stimuli im Vergleich zur Synthese der männlichen Stimme für unnatürlich gehalten werden bzw. nicht als von einem weiblichen Sprecher erzeugt akzeptiert werden. Es ist somit notwendig, für die Synthese der weiblichen Stimme Modelle zu erstellen, in denen eine differenzierte und detaillierte Nachbildung vor allem der Phonation aber auch der Artikulation möglich ist.

Im theoretischen Teil dieser Arbeit wird in Kapitel 1 die Frage behandelt, welche Modellerweiterung bezüglich der Glottis vorzunehmen ist. Dazu werden die verschiedenen phonatorischen Parametersätze vorgestellt, die die Steuerung dieser erweiterten Synthesemodelle ermöglichen. Es wird anhand theoretischer Vorüberlegungen diskutiert, welches Synthesemodell bzw. welcher

---

[1] Siehe MEYER-EPPLER (1969, S. 3).

phonatorische Parametersatz zur Modellierung der geschlechtsspezifischen Unterschiede der Phonation am geeignetsten ist.

Da unterschiedliche Einstellungen der Phonation hauptsächlich zu Unterschieden in der Stimmqualität führen, werden in Kapitel 2 zunächst die Begriffe Stimmqualität und Stimmklang beleuchtet. Der Begriff Stimmklang wird eingeführt, um spezielle Signalaspekte der Stimmqualität zu beschreiben. Stimmklang umfaßt nur die Faktoren der Stimmqualität, die von der Phonation herrühren. Es wird gezeigt, daß die befriedigende Modellierung der wesentlichen geschlechtsspezifischen Charakteristika der Phonation eine wichtige Voraussetzung für die Erreichung eines hohen Natürlichkeitsgrades der synthetischen Sprache ist. Erst nach einer detaillierten Modellierung der Phonation assoziiert der Hörer mit den synthetisierten Stimuli weibliches Sprechergeschlecht. Interessanterweise bestehen auch enge Zusammenhänge zwischen der Klassifikation verschiedener Stimmqualitäten und der Geschlechtsidentifikation bzw. der mit dem Stimulus verbundenen Geschlechtsassoziation.

In Kapitel 3 wird die Normierungstheorie zur Vokalperzeption unter der Berücksichtigung aller im Sprachsignal vorhandenen geschlechtsspezifischen Merkmale betrachtet. Dabei wird die in der Literatur zu findende Bewertung der Formantfrequenzen als der wesentliche Faktor der Normierung kritisch diskutiert. Insbesondere findet die Fragestellung Beachtung, wie dem Hörer die schnelle Einstellung auf einen neuen, evtl. unbekannten Sprecher gelingt, da zwischen den Sprachsignalen verschiedener Sprecher große interindividuelle Unterschiede bestehen können. Es wird die Theorie der Primärnormierung entwickelt, die auf einem engen Zusammenhang zwischen der Geschlechtsassoziation und dem Verschieben der Vokalschablone[2] basiert.

---

[2]Zur Definition des Begriffs Vokalschablone siehe Abs. 3.2.2.

Im experimentellen Teil werden neben Fragen zur Funktionsweise und Implementierung der benutzten Synthesemodelle auch die perzeptiv relevanten Aspekte der generierten Sprachsignale untersucht.

In Kapitel 4 werden die benutzten Synthesemodelle vorgestellt. Es werden zwei Synthesemodelle beschrieben, die sich hauptsächlich durch Unterschiede in der Interaktionsart zwischen Glottis und Sprechtrakt unterscheiden. Insbesondere werden die mit der Phonation eng verbundenen Fragen zur Funktionsweise der Glottismodelle behandelt. Ferner wird auf Unterschiede bezüglich der phonatorischen Parametersätze in den beiden hier benutzten Modellen eingegangen.

In Kapitel 5 werden Experimente zur Messung des Einflusses der Phonation auf die Qualität der Synthese der weiblichen und der männlichen Stimme diskutiert. Neben der relativen Natürlichkeit wurde auch die Klassifikationsrate bezüglich des assoziierten Sprechergeschlechts experimentell gemessen. Ferner wurden die beiden benutzten Synthesemodelle und die Variation einzelner phonatorischer Parameter im Hinblick auf die Natürlichkeitssteigerung der Synthese der weiblichen Stimme untersucht.

In Kapitel 6 wird anhand der Ergebnisse einer experimentellen Untersuchung die in Kapitel 3 aufgestellte Theorie der engen Verbindung zwischen dem Verschieben der Vokalschablone und der Geschlechtsassoziation verifiziert. Dazu wurde in einem Hörtest die perzipierte Vokalqualität in Abhängigkeit von der Geschlechtsassoziation erfragt.

In Kapitel 7 werden die wichtigsten Ergebnisse dieser Arbeit zusammengestellt, mit Ergebnissen experimenteller Untersuchungen anderer Autoren verglichen und diskutiert. Zum einen werden die wichtigsten Tendenzen in der Einstellung der phonatorischen Parameter bei der weiblichen gegenüber der männlichen Stimme betrachtet, zum anderen werden die experimentellen Ergebnisse zur Verifikation der Theorie der Primärnormierung diskutiert.

I. THEORETISCHER TEIL:

ZUR PHONATION IN SYNTHETISCHER SPRACHE. AUSWIRKUNGEN AUF STIMMQUALITÄT, GESCHLECHTSASSOZIATION UND VOKALPERZEPTION

1 Differenzierte phonatorische Parameter als Grundlage für die Synthese weiblicher Stimmen

1.1 Das Grundproblem der Synthese weiblicher Stimmen

Die offensichtlichen Signalparameter, in denen sich der Unterschied zwischen den Sprachsignalen weiblicher und männlicher Sprecher manifestiert, sind die Grundfrequenz und die mittlere Lage der Formantfrequenzen. Während die Grundfrequenz weiblicher Sprecher um annähernd eine Oktave höher als die der männlichen Sprecher liegt (das entspricht einer Erhöhung um fast 100%), stellt man bei der mittleren Lage der Formantfrequenzen eine durchschnittliche Erhöhung der Frequenz um 15-20% fest.

In ersten Experimenten zur Synthese einer weiblichen Stimme wurden, ausgehend von der üblichen Synthese einer männlichen Stimme, F0 und die Formanten entsprechend der oben dargelegten Tendenzen geändert. Jedoch zeigte sich in Hörtests, daß diese so generierten Stimuli nicht als von einer weiblichen Person erzeugt akzeptiert werden. In der Testsituation mit eingeschränkter Wahlmöglichkeit ("forced-choice-situation"), in der nur gestattet ist, diese Stimuli als von weiblichen oder von männlichen Sprechern erzeugt zu klassifizieren, werden die aus der Synthese der weiblichen Stimme resultierenden Stimuli oft als männlich klassifiziert. Hingegen werden die aus der Synthese der männlichen Stimme resultierenden Stimuli nie als weiblich klassifiziert.[1] Die Testpersonen begründen ihre Ablehnung der weiblichen Stimuli mit dem Hinweis auf den ungewohnten und unnatürlichen

---

[1] Siehe Exp. 1 und Exp. 5 dieser Arbeit.

Klang der Stimme.

Damit wird die Stimmqualität zu einem für die Synthese entscheidenden Faktor. Wie auch andere Autoren[2] anmerken, muß bei der Synthese einer weiblichen Stimme besondere Sorgfalt auf die Modellierung der glottalen Schallerzeugung, der Phonation, gelegt werden. Während die Auswahl des Anregungssignals bei der Synthese der männlichen Stimme nicht kritisch ist und dort mit einer Vielzahl von Anregungsfunktionen Stimuli generiert werden können, die als von einem männlichen Sprecher erzeugt akzeptiert werden, so scheint dies bei der Synthese weiblicher Stimmen nicht der Fall zu sein.[3]

Im folgenden werden verschiedene Ansätze zur Modellierung der Phonation diskutiert und eine für die Synthese weiblicher Stimmen geeignete Erweiterung des am IPKöln[4] vorhandenen Synthesemodells vorgestellt. Einzelheiten zum Modell und zu seiner Realisation als Computersimulation werden in Kapitel 4 diskutiert.

1.2 Synthesemodelle mit differenzierten phonatorischen Parametern

1.2.1 Artikulatorische und phonatorische Parameter

Jedes Synthesemodell definiert einen zugehörigen Parametersatz, der die Steuerung dieses Modells ermöglicht. Ebenso definiert dieser Parametersatz das von dem Modell generierte Sprachsignal. Das Synthesemodell ist somit immer eng mit einem bestimmten Satz von Steuerparametern verbunden, wobei verschiedene Modelle evtl. sehr unterschiedliche Steuerparametersätze benöti-

[2] Siehe CHENG et al. (1987a und 1987b) und FANT et al. (1987).

[3] Eine Erklärung dieses experimentellen Befundes wird in Abs. 2.2 gegeben.

[4] Institut für Phonetik der Universität zu Köln.

gen. Ein Ziel der Sprachsyntheseforschung ist es, perzeptiv relevante Signalparameter und Steuerparameter einander anzunähern. Letztlich haben die Steuerparameter ihre phonetische Bedeutung erst dann bewiesen, wenn die aus diesen Parametern gewonnene Synthese des akustischen Signals auch wirklich die perzeptiv relevante Information enthält, die man mit den Parametern codieren wollte. Dann können die Steuerparameter auch als Signalparameter aufgefaßt werden, deren perzeptive Relevanz quasi durch Resynthese verifiziert worden ist. In diesem Sinne wird hier versucht, einen ersten Parametersatz zur Beschreibung von Stimmqualitätsunterschieden zwischen Frau und Mann zu entwickeln.

Die hier betrachteten Synthesemodelle lassen sich in die Klasse der erzeugungsorientierten Syntheseverfahren einordnen. Die Parametersätze beschreiben demnach artikulatorische Parameter, bzw. das Synthesemodell generiert das Sprachsignal nach dem Vorbild der menschlichen Spracherzeugung. Insofern sind auch die hier benutzten Synthesemodelle als "Quelle-Filter-Modelle" aufgebaut, wobei mit "Quelle" das Modell zur Anregung des Sprechtrakts entsprechend der Glottis und mit "Filter" die Modellierung des Sprechtrakts selbst gemeint ist (siehe Abb. 1.1). Jedoch unterscheiden sich die Modelle hinsichtlich der Modellierung der Glottis und des Sprechtraktes und bezüglich der Berücksichtigung von Interaktionsmechanismen zwischen Glottis und Sprechtrakt.

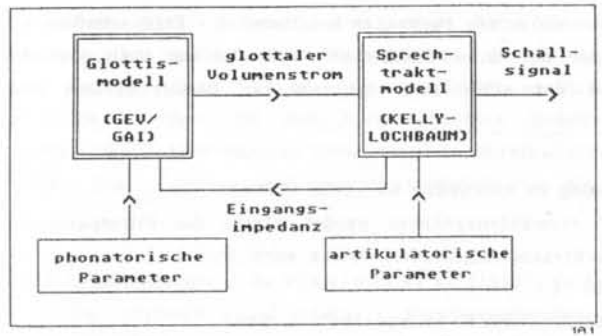


Abb. 1.1: Modell zur Synthese der weiblichen Stimme

Diesem Aufbau der Synthesemodelle entsprechend kann auch der zugehörige Parametersatz in die Gruppen der *phonatorischen Parameter* und der *artikulatorischen Parameter* unterteilt werden.[5] Der Begriff "artikulatorisch" ist hier im engen Sinne nur für den supralaryngalen Bereich zu verstehen.

Artikulatorische Parameter sind demnach die Parameter, die direkt die Einstellung des Sprechtrakts ermöglichen. Jedoch führt der Begriff artikulatorische Parameter sofort zu einer Beschränkung bei der Auswahl der Synthesemodelle. Gerade bezüglich der Realisation des Sprechtrakts gibt es sehr verschiedene Syntheseverfahren, die dementsprechend sehr unterschiedliche Parametersätze voraussetzen. So benötigt z.B. die Formantensynthese einen Satz von Formantfrequenzen, -amplituden und -bandbreiten zur Ansteuerung des Sprechtrakts.[6] Andererseits definiert die KELLY-LOCHBAUM-Synthese[7] einen Satz von Rohrquerschnitten, der aus der Sprechtraktform gewonnen werden kann. Die Steuerung der Sprechtraktform wird bei diesem Syntheseverfahren in der Regel von Parametersätzen übernommen, die fünf bis zehn Parameter umfassen, wie z.B. Kieferstellung, Zungenhöhe und Zungenposition.[8] Eine Mittelstellung nimmt die LPC-Synthese[9] ein, deren Koeffizienten einerseits wie bei der Formantensynthese leicht aus dem Sprachsignal gewonnen werden können, andererseits in der PARCOR-Darstellung artikulatorische Bedeutung besitzen und direkt zur Ansteuerung des KELLY-LOCHBAUM-Syntheseverfahrens verwendet werden können.

Die phonatorischen Parameter bestimmen die Eigenschaften der Quelle. Wegen der unterschiedlichen Verfahren zur Simulation der Quelle sind vor allem zwei Gruppen von Glottismodellen bzw.

[5] Bezeichnung in Anlehnung an TITZE (1984).

[6] Bei der Formantensynthese werden direkt die Filterparameter des Sprechtrakts angesteuert (Siehe auch Abs. 4.1).

[7] Siehe Abs. 4.1.

[8] Siehe HEIKE (1980, S. 11 und 1982, S. 93).

[9] Siehe Abs. 4.1.

phonatorischen Parametersätzen hervorzuheben, die im folgenden eingeführt werden.

1.2.2 Physiologie-orientierte und Wellenform-orientierte phonatorische Parameter

Physiologie-orientierte Parametersätze steuern Synthesemodelle, in denen die Öffnungsfläche zwischen den vibrierenden Stimmlippen als Funktion der Zeit nicht vorgegeben, sondern aus den physikalischen, d.h. aerodynamischen und biomechanischen Gesetzen berechnet wird. Je differenzierter diese Modelle werden,[10] d.h. je mehr physiologische und biomechanische Details und aerodynamische Effekte berücksichtigt werden, desto größer wird der notwendige Satz an Steuerparametern. All diesen Modellen gemeinsam sind aber die zwei Hauptparameter subglottaler Druck (p_sub) und Stimmlippenspannung (Q), während in vielen Modellen bereits die die Schwingungsart charakterisierende effektive Masse der Stimmlippen nur noch als Konstante eingeht. Allerdings ist die Schwingungsform der glottalen Welle, die aus der Schwingungsart der Stimmlippen resultiert, eine wichtige Größe zur Beschreibung der Unterschiede männlicher und weiblicher Stimmqualität.

Darum gehören die in dieser Arbeit betrachteten Synthesemodelle zur Gruppe der durch *Wellenform-orientierte Parametersätze*, auch *Wellenformparameter* genannt, angesteuerten Synthesemodelle. Bei diesen Modellen wird durch die Steuerparameter die glottale Wellenform direkt definiert. Die glottale Welle ist Ausgangssignal des Glottismodells bzw. Eingangssignal des Sprechtraktmodells und muß im Prinzip nur über die geforderten spektralen Charakteristika verfügen. In den hier benutzten Modellen soll die glottale Welle aber darüber hinausgehende physikalische Bedeutung haben. Das Ausgangssignal des Glottismodells ist hier immer der

[10] Einmassenmodell siehe FLANAGAN et al. (1968), Zweimassenmodell siehe ISHIZAKA et al. (1972), Mehrmassenmodell siehe TITZE (1973 und 1974).

glottale Volumenstrom (u_glott), d.h. die durch die Stimmlippen fließende Luftmenge pro Zeiteinheit.

1.2.3 Die Definition der Wellenform-orientierten Parameter

Bevor der hier benutzte phonatorische Parametersatz vorgestellt wird, sei noch angemerkt, daß natürlich auch das bisher am IPKöln benutzte Modell zur Synthese der männlichen Stimme über phonatorische Parameter verfügte. Allerdings beschränkte sich dieser Satz auf die Grundfrequenz (F0) und die Maximalamplitude, da die Wellenform fest vorgegeben war. Im jetzt vorgestellten Phonationsmodell wird der Parameter Grundfrequenz übernommen, allerdings wird die Amplitude nun durch die Wellenformparameter implizit definiert.

Die Wellenformparameter beziehen sich, wie oben gesagt wurde, direkt auf die Form des glottalen Volumenstroms, also auf die Schwingungsform des Signals, das in den Sprechtrakt eingespeist wird.[11] Der Wellenformparametersatz gliedert sich in die *Amplitudenparameter*, die die Amplitude der glottalen Welle kontrollieren, und in die *Zeitparameter*, die die Proportionen zwischen den einzelnen Zeitphasen innerhalb jeder Schwingungsperiode festlegen.

Unabhängig davon muß eine zweite Einteilung der phonatorischen Parameter in *primäre Parameter* und *sekundäre Parameter* vorgenommen werden. Aus der Definition des Synthesemodells ergibt sich direkt der primäre Parametersatz. Dabei handelt es sich um den minimalen Satz an Parametern, der direkt zur Steuerung des Modells benötigt wird und zur eindeutigen Steuerung des Modells

[11] Genaugenommen ist mit "glottaler Wellenform" also die Schwingungsform des Volumenstroms am Ort der Glottis gemeint. Dieser Begriff wurde aber in Anlehnung an den in der Literatur gebräuchlichen Terminus "glottal waveform" (siehe z.B. TITZE (1984, S. 570) oder HOLMBERG et al. (1988, S. 511)) gewählt.

auch hinreichend ist. Die sekundären Parameter sind eng mit den primären Parametern verbunden, aber nicht direkt zur Steuerung des Synthesemodells nötig. Sie können aus den primären Parametern oder dem Zeitverlauf der simulierten Signale berechnet werden. Bestimmte sekundäre Parameter können bestimmte primäre Parameter sogar gleichwertig ersetzen, ohne daß die Bedingung der eindeutigen Steuerung des Synthesemodells[12] verletzt wird. Die sekundären Parameter werden deswegen eingeführt, weil hierdurch die Menge der von den Parametern direkt beschriebenen Signalmerkmale erweitert wird. Beispielsweise beschreiben die durch die hier benutzten Synthesemodelle definierten primären Amplitudenparameter einige wesentliche Unterschiede zwischen der Phonation männlicher und weiblicher Stimmen nur mittelbar.

Die im folgenden definierten Wellenform-orientierten Parameter werden gemäß der Einteilung in Zeit- und Amplitudenparameter betrachtet. Die Unterscheidung von primären und sekundären Parametern ist hier lediglich in Tabelle 1.2 und Tabelle 1.3 aufgeführt. Sie wird erst bei der Beschreibung der Generierung der Stimuli größere Bedeutung erlangen.

Im Zeitbereich unterscheidet man innerhalb jeder Schwingungsperiode der glottalen Welle zunächst zwischen der *Verschlußphase* und der *Offenphase*. In der Offenphase bewegen sich die Stimmbänder, während sie in der Verschlußphase ruhen. Wie neuere Untersuchungen ergeben haben,[13] muß die Glottis während der Verschlußphase aber nicht völlig geschlossen sein. Im allgemeinen tritt ein kleiner Gleichstromwert des glottalen Volumenstroms auch während der Verschlußphase auf. Daneben kann die Offenphase noch in *Öffnungsphase* und *Schließphase* aufgespalten werden. In der Öffnungsphase entfernen sich die Stimmlippen voneinander (positive Stimmlippengeschwindigkeit), während sie sich in der Schließphase einander wieder nähern (negative Stimmlippengeschwindigkeit). Abbildung 1.2. stellt den glottalen Volu-

[12] Siehe Abs. 1.2.4.

[13] Siehe z.B. HOLMBERG et al. (1988).

menstrom (u_glott), dessen Ableitung (u_glott') und die vier oben beschriebenen Zeitphasen dar.

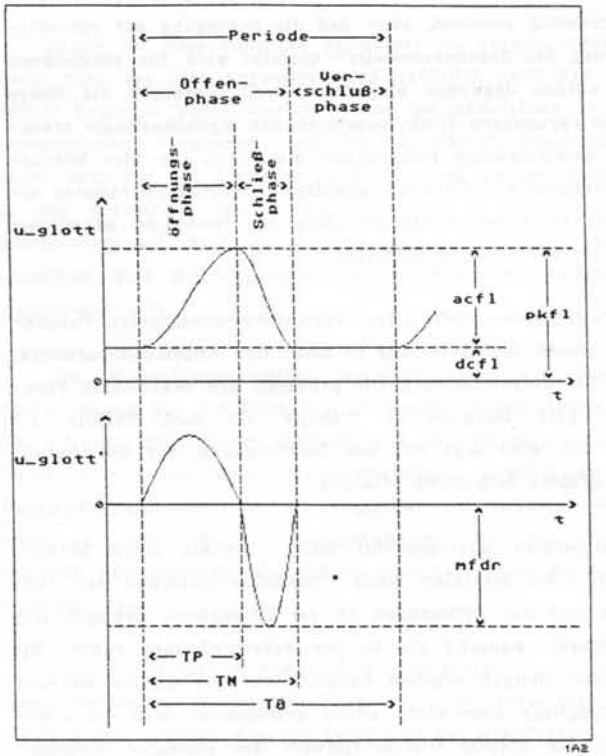


Abb. 1.2: Zur Definition der Wellenform-Parameter anhand der glottalen Welle

Setzt man den Nullpunkt der Zeit willkürlich an den Beginn einer Periode, so kann man TN, TP und T0 als die Zeitpunkte betrachten, die die Zeitphasen begrenzen. Der Zeitpunkt TN teilt somit Offenphase und Verschlußphase. TP ist der Zeitpunkt maximaler glottaler Öffnungsfläche und trennt die Öffnungsphase von der Schließphase.

0<t<TN	Offen-phase	TN	Offenzeit (open time)
TN<t<T0	Verschluß-phase	T0-TN	Verschlußzeit (close time)
0<t<TP	öffnungs-phase	TP	öffnungszeit (opening time)
TP<t<TN	Schließ-phase	TN-TP	Schließzeit (closing time)

1T1

Tab. 1.1: Die Zeitphasen einer Schwingungsperiode der glottalen Welle

Die hier definierten Zeitphasen einer Schwingungsperiode der glottalen Welle werden durch die phonatorischen Zeitparameter beschrieben, die günstigerweise relativ zur Periodendauer T0 definiert werden. Somit ergibt sich eine zu T0, TP und TN äquivalente Darstellung durch die *Grundfrequenz* (F0), den *Offenquotient* (opqu) und den *Schließquotienten* (clqu). Daneben tritt in der Literatur noch der *Geschwindigkeitsquotient* (spqu) auf, der in den hier benutzten Modellen zwar nicht primärer Steuerparameter ist, aber als sekundärer Parameter eingeführt wird, weil er die Symmetrie der glottalen Welle beschreibt. Eine glottale Welle mit maximalem Symmetriegrad hat einen Geschwindigkeitskoeffizienten von 1. Die Zeitspanne der Öffnungs- und Schließphase ist dann gleich bzw. der Offenquotient ist doppelt so groß wie der Schließquotient.

PRIM. PARAMETER	$F0 = \frac{1}{T0}$	Grundfrequenz (fundamental frequency)	
	$opqu = \frac{TN}{T0}$	Offenquotient (open quotient)	
	$clqu = \frac{TN-TP}{T0}$	Schließquotient (closing quotient)	
SEK. PARAMETER	$spqu = \frac{TP}{TN-TP}$	Geschwindigkeitsquotient (speed quotient)	$\frac{opqu - clqu}{clqu}$

1T2

Tab. 1.2: Zur Definition der Wellenform-Zeitparameter

Die Amplitudenparameter sind in Tabelle 1.3 zusammengestellt. Als primäre Parameter bestimmen die *Wechselstromamplitude* (acfl) und der *Verschlußstrom* (dcfl) die Schwingungsform des glottalen Volumenstroms (u_glott) (siehe auch Abb. 1.2). Ein äquivalenter primärer Parametersatz ist auch unter Austausch eines der beiden Parameter mit dem *Maximalstrom* (pkfl) möglich. Der *mittlere Strom* (avfl) ergibt sich aus der Integration des glottalen Volumenstroms über die Zeit. Neben den in Tabelle 1.3 definierten Amplitudenverhältnissen ist noch das *Stromänderungsmaximum* (mfdr), das als Maximum des Betrags der Ableitung des glottalen Volumenstroms nach der Zeit (u_glott') definiert ist, wichtig. Dieser maximale Wert der Ableitung ist meist negativ, liegt also meist in der Schließphase. Er gibt direkt die Anregungsstärke des Sprechtrakts an und bestimmt damit die Signalintensität.[14]

PRIM. PARAMETER	acfl	Wechselstromamplitude (ac flow)	
	dcfl	Verschlußstrom (dc flow)	
SEK. PARAMETER	pkfl	Maximalstrom (peak flow)	acfl + dcfl
	avfl	mittlerer Strom (average flow)	$\frac{1}{T0}\int_0^{T0} u_glott[t]\,dt$
	mfdr	Stromänderungsmaximum (maximum airflow declination rate)	$\max\{\lvert u_glott'[t]\rvert\}$
	dcfl / pkfl	relativer Verschlußstrom	$\frac{dcfl}{pkfl}$
	acfl / dcfl	Wechselstrom-Verschluß-strom-Verhältnis	$\frac{acfl}{dcfl}$

173

Tab. 1.3: Definition der Wellenform-Amplitudenparameter

Eine Zusammenstellung der die glottale Welle beschreibenden phonatorischen Parameter, also der zur Synthese benutzte Satz an Steuerparametern, ist in Tabelle 1.4 zusammen mit dem benutzten

[14] Zur Begründung siehe Abs. 4.3.2.4.

artikulatorischen Parametersatz[15] angegeben.

phonatorische Parameter	artikulatorische Parameter
Anzahl: 5	Anzahl: 10
Grundfrequenz Offenquotient Schließquotient Verschlußstrom Wechselstrom-amplitude	Stellung des Velums Lippenstülpung Lippenöffnung Lippenrundung Kieferstellung Zungenhöhe Zungenposition Zungenspitzenhöhe Zungenspitzenposition Höhe der Larynx

174

Tab. 1.4: Phonatorischer und artikulatorischer Steuerparametersatz

1.2.4 Modelldifferenzierungen aufgrund der Interaktion zwischen Glottis und Sprechtrakt: Die GEV-, GAI- und GM-Modellgruppen

Zwar definiert jedes Synthesemodell eindeutig den benötigten Satz an Steuerparametern, jedoch ist es möglich, ähnliche Synthesemodelle zu entwickeln, die vom gleichen Parametersatz angesteuert werden können. Dadurch kann aber bei gleichen Steuerparametern ein anderes Signal resultieren. So sind beispielsweise zwei Synthesemodelle denkbar, die vom gleichen Satz Physiologieorientierter Parameter angesteuert werden, d.h. denen das gleiche biomechanische Modell zugrunde liegt,[16] wobei ein Modell aber mehr aerodynamische Gesetzmäßigkeiten berücksichtigt als das andere.

Die zur Durchführung der in dieser Arbeit beschriebenen Experimente nötigen Differenzierungen der Synthesemodelle sind allerdings anderer Art. Sie betreffen die Interaktion zwischen Glottis und Sprechtrakt. Diese Interaktion kann in drei Hauptgruppen unterteilt werden: die *mechanische Interaktion*, die

[15] Siehe GREISBACH (1986, S. 48).

[16] Z.B. Repräsentation der Stimmlippen durch zwei Massen.

akustische Interaktion und die *Nichtberücksichtigung jeglicher Interaktion* zwischen Glottis und Sprechtrakt. Die mechanische Interaktion berücksichtigt die Rückwirkung des Sprechtraktes auf die Stimmlippenschwingung. Es wurde beispielsweise die Auswirkung der Sprechtrakteingangsimpedanz auf die Schwingungsfrequenz der Stimmlippen untersucht.[17] Die akustische Interaktion beschreibt die Rückwirkung des Sprechtraktes auf den glottalen Volumenstrom, aber ohne daß die Interaktion in der Lage wäre, die Schwingungsform der Stimmlippen, also den zeitlichen Verlauf der glottalen Öffnungsfläche, zu beeinflussen. Bei der Nichtberücksichtigung jeglicher Interaktion wird der fest vorgegebene glottale Volumenstrom einfach in den Sprechtrakt eingeprägt. Die *Glottismodelle mit mechanischer Interaktion*, auch kurz *GMI-Modelle* genannt, werden durch Physiologie-orientierte Parametersätze angesteuert. Es handelt sich also um Modelle mit frei schwingenden Stimmlippen. Die Glottismodelle mit Wellenform-orientierten Parametern können als *Glottismodelle mit akustischer Interaktion*, kurz *GAI-Modelle*, die akustische Interaktion berücksichtigen oder auch als *Glottismodelle mit eingeprägtem Volumenstrom*, kurz *GEV-Modelle*, ohne die Berücksichtigung jeglicher Wechselwirkungen zwischen Glottis und Sprechtrakt arbeiten. Beide Modelle, das GEV-Modell und das GAI-Modell, wurden im Rahmen dieser Arbeit implementiert; denn gerade bei hohem F0, bzw. wenn F0 in den Frequenzbereich des ersten Formanten gelangt oder sogar darüber hinausgeht, ist die Interaktion zwischen Glottis und Sprechtrakt wichtig.[18]

Insofern ist gerade bei der Synthese weiblicher Stimmen besonders auf die Interaktion zu achten, da die Formantfrequenzen im Mittel um 15-20% ansteigen, während die Grundfrequenz als typisches Geschlechtsmerkmal mit rund 100% viel stärker ansteigt. Wie im weiteren diskutiert wird, ist gerade die Interaktion zwischen Glottis und Sprechtrakt ein entscheidender qualitativer Unterschied zwischen weiblicher und männlicher Stimme. Es zeigte

---

[17] Siehe ISHIZAKA et al. (1972).

[18] Siehe FANT (1986).

sich, daß es nicht ausreicht, die weibliche Stimme im interaktionsfreien GEV-Modell durch ein einfaches "Heraufskalieren" der Grundfrequenz und der Formantfrequenzen zu synthetisieren.

Da des weiteren die Frage diskutiert werden muß, warum in bezug auf das Problem der Synthese weiblicher und männlicher Stimmqualität einem Wellenform-orientierten phonatorischen Parametersatz der Vorzug vor einem Physiologie-orientierten Parametersatz gegeben wurde, sollen hier einige Kriterien für die Güte eines Parametersatzes zur Untersuchung der in dieser Arbeit relevanten Fragestellungen aufgestellt werden.

1) Vollständigkeit des Parametersatzes
   Der Parametersatz muß die eindeutige Steuerung des Synthesemodells erlauben und damit zu durch den Parametersatz eindeutig definierten Signalen führen. Alle sich aus dem Synthesemodell ergebenden Steuerparameter müssen somit vom Parametersatz abgedeckt werden. Damit führt eine vollständige Anzahl von Parameterwerten zu einem eindeutigen akustischen Signal.

2) Perzeptive Relevanz der Parameter
   Die einen bestimmten perzeptiven Signalaspekt bestimmenden Parameter müssen das gewünschte perzeptive Resultat im synthetisierten Signal codieren. Beispielsweise ist ein in dieser Arbeit untersuchter perzeptiver Signalaspekt die Stimmqualität, die sich aus einer bestimmten Teilmenge der phonatorischen Parameter ergibt. Als perzeptives Resultat wird dazu die beim Hörer auftretende, mit dem synthetisierten Signal verbundene Geschlechtsassoziation gemessen. Insbesondere ist es nicht die Aufgabe eines Parametersatzes, ein natürlich gesprochenes akustische Signal so genau wie möglich zu rekonstruieren, sondern vor allem das gewünschte perzeptive Resultat befriedigend nachzubilden.

3) Akustische Relevanz der Parameter
   Es sollte prinzipiell möglich sein, eine Signalanalysemethode zu entwickeln, die es gestattet, die Steuerparameterwerte jedes Steuerparametersatzes aus dem akustischen Signal zu

extrahieren. Anders ausgedrückt: Man kann die Werte jedes Steuerparametersatzes auch als eine mögliche Codierung des Sprachsignals auffassen. Diese Codierung soll prinzipiell aus dem natürlichen Sprachsignal beliebiger Sprecher herleitbar sein.

4) Artikulatorische Relevanz der Parameter
   Jeder Parameter soll physiologisch sinnvoll sein, d.h. er soll artikulatorische bzw. erzeugungsorientierte Bedeutung haben.

5) Das Äquivalenzprinzip
   Zwischen zwei Parametersätzen, die den gleichen phonetischen Signalaspekt untersuchen (wie z.B. Untersuchung der Stimmqualitätsdifferenzen weiblicher und männlicher Sprecher durch Physiologie-orientierte oder Wellenform-orientierte Parametersätze), soll eine Abbildung existieren, die diese Parametersätze ineinander überführt.

6) Das Minimalprinzip
   Der Parametersatz ist so klein wie möglich zu halten und das Synthesemodell ist nur so kompliziert wie nötig zu gestalten, um das zu untersuchende Signalmerkmal perzeptiv befriedigend zu resynthetisieren. Insbesondere ist es nicht das Ziel, den menschlichen Sprechapparat möglichst genau nachzubilden, sondern anhand des benutzten Parametersatzes und Synthesemodells die für die entsprechende Fragestellung wesentlichen Signalmerkmale zu resynthetisieren.

7) Generierung der Parameterwerte eines Parametersatzes nach Regeln
   Es muß möglich sein, ein Regelsystem aufzustellen, das für einen bestimmten Parametersatz ohne Vorlage eines beispielhaften Sprachsignals die erforderlichen Parameterwerte generiert.

1.3 Systematische Unterschiede der phonatorischen Parameter bei der Synthese der weiblichen und der männlichen Stimme

Die Synthese weiblicher Stimmen ist ein noch relativ neues Gebiet, so daß es zur Zeit erst wenige Arbeiten zu diesem Thema gibt. CHENG et al. (1987a und 1987b) und FANT st al. (1987) benutzen verschiedene Synthesemodelle und auch verschiedene phonatorische Parametrisierungen. Gemeinsam ist diesen Arbeiten jedoch, daß die Synthesemodelle aus der GAI-Gruppe stammen.

Die in dieser Arbeit benutzte Einstellung der phonatorischen Parameter wurde durch eine Untersuchung von HOLMBERG et al. (1988) motiviert. In der Arbeit von HOLMBERG et al. wird die glottale Wellenform von weiblichen und männlichen Sprechern mit der Methode des inversen Filterns ermittelt, und erstmals wurden die gewonnenen Daten zur glottalen Wellenform detailliert veröffentlicht. Die Einstellung der phonatorischen Parameter in den im Rahmen dieser Arbeit benutzten Synthesemodellen wurde nach diesen Daten vorgenommen.

Es soll bereits hier vorweggenommen werden, daß mit den im Rahmen dieser Arbeit benutzten Synthesemodellen und dem hier entwickelten phonatorischen Parametersatz eine ausreichende Kontrolle der geschlechtsspezifischen Stimmqualität gelang. Die generierten Stimuli wurden als von einem weiblichen Sprecher geäußert anerkannt.[19]

Damit sind folgende Punkte erfüllt:

1) Die hier verwandten Synthesemodelle und die hier getroffene Auswahl der phonatorischen Parameter erlauben die Modellierung der Unterschiede in der Stimmqualität weiblicher und männlicher Sprecher.

---

[19] Siehe Exp. 1, Abs. 5.3.

2) Die von HOLMBERG et al. (1988) angegebenen Daten zur glottalen Wellenform weiblicher und männlicher Sprecher enthalten die wesentlichen Charakteristika der geschlechtsspezifischen Differenzen in der Stimmqualität.

3) Die wesentlichen Charakteristika der geschlechtsspezifischen Differenzen in der Stimmqualität bleiben trotz der hier durchgeführten Datenreduktion auf vier phonatorische Parameter (außer F0) erhalten. HOLMBERG et al. geben neben F0 immerhin zwölf Parameter an, die teilweise auch noch anders definiert sind als die hier benutzten Wellenformparameter.[20]

Ein weiteres Ziel dieser Arbeit war die Untersuchung der Frage, wie stark die Perzeption der weiblichen Stimmqualität durch die Änderung der einzelnen phonatorischen Parameter beeinflußt wird. Dadurch kann ermittelt werden, welche Parameter für die Stimmqualität von Frauen von übergeordneter Bedeutung sind. Weitgehende Einigkeit herrscht in der Literatur über die relative, d.h. auf T0 bezogene Vergrößerung der Offenphase bei weiblichen gegenüber männlichen Sprechern. Unterschiedliche Ergebnisse findet man aber bezüglich der Symmetrie der glottalen Wellenform. Während MONSEN et al. (1977), CARRELL (1981) und CHENG et al. (1987a und 1987b) die Symmetrisierung der glottalen Wellenform als das Hauptcharakteristikum des Stimmqualitätsunterschiedes zwischen Mann und Frau herausstellen, erkennen HOLMBERG et al. (1988) und FANT et al. (1987) keine starken Symmetrisierungstendenzen. Sie machen die Symmetrisierung nicht zum Hauptunterschied bei der Synthese der weiblichen und der männlichen Stimme. Dagegen diskutieren FANT et al. (1987) im Gegensatz zu CHENG et al. (1987a und 1987b) den Einfluß des Verschlußstroms und weisen diesem für die Synthese weiblicher und männlicher Stimmen große Bedeutung zu.

Daher wurden in dieser Arbeit einerseits der Schließquotient (clqu) und damit bei konstantem Offenquotient der Geschwindigkeitskoeffizient (spqu) bzw. die Symmetrie der Welle und andererseits der Verschlußstrom (dcfl) systematisch variiert und bezüglich ihrer Auswirkungen auf die Stimmqualität bei der Synthese der weiblichen Stimme getestet. Außer acht bleibt die Frage, wie sich verschiedene analytische Ansteuerfunktionen auf die Stimmqualität auswirken. Eine Betrachtung dieser Frage findet sich bei FANT et al. (1987).

[20] Mehr Details in Kap. 4 und in Kap. 5.

1.4 Vor- und Nachteile der GAI-Modelle und der GMI-Modelle für die Synthese der weiblichen Stimme

Für die Synthese der weiblichen Stimme wurden hier von vornherein Wellenform-orientierte und nicht Physiologie-orientierte Parametersätze gewählt. Der Hauptgrund ist, daß analytische Daten für die geschlechtsspezifischen Unterschiede in der laryngalen Aktivität bisher hauptsächlich aus Messungen mittels inverser Filtertechnik gewonnen werden konnten. Diese Daten sind weitgehend Wellenformparameter. Ein weiterer Grund ist die alleinige Wichtigkeit der glottalen Welle für die Auswirkung des Glottismodells auf das akustische Signal. Denn grundsätzlich ist es allein das Eingangssignal des Sprechtraktes, also der glottale Volumenstrom und dessen Spektrum, über das die phonatorischen Parameter das akustische Signal beeinflussen können. Auch in Modellen mit Physiologie-orientierten Parametersätzen bleibt es letztlich die glottale Welle, die allein das akustische Signal beeinflußt. Die Rekonstruktion der Wellenform nach vorgegebenen Daten ist aber in Modellen mit Wellenform-orientierten phonatorischen Parametern wesentlich einfacher als in Modellen mit Physiologie-orientierten phonatorischen Parametern. Es muß zunächst sogar in Frage gestellt werden, ob einfache selbstschwingende Glottismodelle überhaupt in der Lage sind, den glottalen Volumenstrom gemäß der wesentlichen geschlechtsspezifischen Tendenzen in der glottalen Wellenform zu modellieren. Es kann nicht unbedingt davon ausgegangen werden, daß ein selbstschwingendes Glottismodell in der Lage ist, die hier untersuchten Stimmqualitätsdifferenzen wiederzugeben.

Andererseits ist die durch die geringe Differenz von Grundfrequenz und Formantfrequenzen, wie es bei weiblichen Stimmen gegeben ist, auftretende mechanische Interaktion erheblich.[21] Wie auch in dieser Arbeit nachgewiesen wird, ist hingegen die akustische Interaktion nicht so entscheidend.[22] Die mechanische Interaktion führt insbesondere dazu, daß bei einer Änderung der Sprechtraktgeometrie und damit bei einer Änderung der Eingangseigenschaften des Sprechtraktes, auch eine Änderung in der Schwingungsform der Stimmlippen und damit in der Schwingungsform des glottalen Volumenstroms auftritt. CHENG et al. (1987a und 1987b) weisen darauf hin, daß bei der Synthese der weiblichen Stimme für verschiedene Sprechtrakteinstellungen verschiedene Offenquotienten angesetzt werden müssen. Bei männlichen Stimmen ist das wegen der geringen mechanischen Interaktion nicht nötig. Das bedeutet also, daß die artikulatorischen Parameter die phonatorischen Parameter beeinflussen, auch wenn die Stimmqualität konstant bleibt. Die mittels der Methode des inversen Filterns gewonnenen Daten beziehen sich aber meist nur auf einen Schwa-Laut. Damit können die phonatorischen Parameter nicht den Anspruch auf direkte Repräsentation der Stimmqualität erheben.[23] Jedoch ist noch nicht untersucht worden, wie stark die Stimmqualität bei konstanten Wellenform-orientierten Parametern unter der Änderung der artikulatorischen Parameter variiert, bzw. wie stark die Wellenform-orientierten phonatorischen Parameter bei Änderung der artikulatorischen Parameter auch geändert werden müßten, um die mechanische Interaktion zu berücksichtigen und die Stimmqualität konstant zu halten.

Der Vorteil der leichten Einstellung der Wellenform-orientierten Parameter wird also bezahlt mit der Erstellung eines komplizierten Systems an Steuerregeln für die Synthese eines Sprechvorganges, das zur Konstanthaltung der Stimmqualität bekannt sein muß. Ferner hat sich gezeigt, daß auch extreme

[21] Siehe TITZE (1986).

[22] Siehe auch NORD et al. (1984).

[23] Zur perzeptiven Relevanz der Parameter siehe Abs. 1.2.4.

Sprechtraktgeometrien, wie sie beispielsweise bei Verschlußlauten auftreten, in GMI-Modellen leicht modelliert werden können.[24] So bleibt also der Vorteil der Wellenform-orientierten Parametermodelle zunächst auf stationäre Vokale im Schwa-Bereich beschränkt, wie auch die mittels inversen Filterns gewonnenen Daten nur für stationäre Laute im Schwa-Bereich gelten. Andererseits eröffnet gerade die Aufstellung von Steuerregeln, wie es bei Wellenform-orientierten Parametermodellen wahrscheinlich nötig wird, einen Einblick in die perzeptiv wesentlichen Komponenten der laryngalen Aktivität.

Wie auch in dieser Arbeit gezeigt werden wird, ist die akustische Interaktion hingegen bei offenen Vokalen nicht erheblich. So reicht es im Fall der stationären offenen Vokale sogar aus, für die Synthese der weiblichen Stimme nur ein GEV-Modell zu verwenden. Jedoch wird dies auf den recht uninteressanten Fall der stationären Vokale beschränkt bleiben, da beispielsweise ein in natürlicher Sprache bei stärkeren F0-Transitionen auftretender Formantamplitudeneffekt nur von Modellen der GAI-Gruppe modelliert werden kann.[25]

Problematisch für die GMI-Modelle ist hingegen, daß gerade im Bereich der Glottis sehr komplizierte biomechanische und aerodynamische Verhältnisse vorliegen. Darum wurden neben Ein- und Zweimassenmodellen auch Mehrmassenmodelle vorgeschlagen,[26] um zu einer annähernd realen Beschreibung zu gelangen. Durch die Verengung der Luftpassage an den Stimmlippen ist aber auch die Aerodynamik hier um ein vielfaches komplizierter als bei der Wellenausbreitung im Sprechtrakt. So stellt sich immer die Frage nach der Realität der von GMI-Modellen errechneten Stimmlippenschwingungsverläufen.

[24] Zur Synthese der stimmhaften Plosive siehe z.B. FLANAGAN et al.(1975).

[25] Siehe FANT et al. (1982, S. 3).

[26] Siehe TITZE (1973 und 1974).

2 Stimmqualität und Geschlechtsidentifikation

Die Erzeugung einer natürlich klingenden weiblichen Stimme ist schwierig. Daher muß dem Faktor Stimmqualität große Bedeutung für die Synthese der weiblichen Stimme zugemessen werden. Im experimentellen Teil dieser Arbeit werden im Zusammenhang mit der in Kapitel 3 aufgestellten Theorie zur Primärnormierung Untersuchungen zur Geschlechtsklassifikation von synthetisch erzeugten Stimuli durchgeführt. Ziel dieses Kapitels ist es, den engen Zusammenhang zwischen Stimmqualität, Stimmklang und Geschlechtsidentifikation herauszuarbeiten. Damit begründet sich auch die in den experimentellen Untersuchungen gefundene Korrelation zwischen der Phonation bzw. den Einstellungen der phonatorischen Parameter und der Geschlechtsidentifikation. In Absatz 2.1 wird der Begriff Stimmqualität und Stimmklang definiert. In Absatz 2.2 werden die perzeptiven Faktoren der Stimmqualität untersucht und es wird gezeigt, daß die Stimmqualität einen wesentlichen Einfluß auf die Geschlechtsidentifikation hat. In Absatz 2.3 werden die für die Geschlechtsidentifikation bzw. für die mit dem Stimulus verbundene Assoziation des Sprechergeschlechts relevanten perzeptiven Signalmerkmale diskutiert.

2.1 Zum Begriff Stimmqualität

Nach PASCAL (1985) ist *Stimmqualität* ("voice quality") das perzeptive Resultat der anatomischen Charakteristika eines Sprechers und seiner gewohnheitsmäßigen "artikulatorischen Gesten". Dabei beschränkt PASCAL den Begriff artikulatorische Gesten nicht nur auf den supralaryngalen Bereich, sondern schließt auch alle laryngalen Bewegungsabläufe mit ein. So gesehen wird die Stimmqualität, wie eine enge Interpretation des Wortes vermuten ließe, nicht nur durch den Schwingungsmodus der Stimmlippen bzw. die Gesamtheit der laryngalen Aktivität, sondern durch die Gesamtheit der in einer Äußerung vorhandenen, den Sprecher charakterisierenden Signalmerkmale definiert. Die Stimmqualität umfaßt somit alle idiosynkratischen Merkmale eines Sprechers und damit die gesamte

Menge der ektosemantischen Merkmale[1]. Diese Signalmerkmale tragen beispielsweise zur Erkennung des Geschlechts, Alters, der Größe, der physischen und psychischen Verfassung und des sozialen Status bei. Daneben kann aber auch die Zugehörigkeit zu einer bestimmten Sprach- oder Dialektgemeinschaft zu Einflüssen auf die Stimmqualität führen. Prinzipiell sind die Merkmale der Stimmqualität immer segmentübergreifender Natur. Andererseits kann ein Sprecher gewisse Stimmqualitätsmerkmale aber auch während einer Äußerung ändern.

LAVER (1980) betrachtet die Stimmqualität vor allem vom Standpunkt der Spracherzeugung aus. Er führt unterschiedliche Stimmqualität auf Unterschiede in den artikulatorischen Einstellungen ("articulatory settings") zurück. Er unterteilt die artikulatorischen Einstellungen in supralaryngale und phonatorische Einstellungen. Beispiele für supralaryngale Einstellungen sind ständige Dentalisierung, Palatalisierung oder Nasalisierung. Die artikulatorische Einstellung beschreibt also die bei einem Sprecher vorliegende artikulatorische Grundtendenz. LAVER definiert die artikulatorische Einstellung aber nur in bezug auf eine neutrale Einstellung, die nach seiner Definition übergeordneter Natur ist und einen interindividuellen Mittelwert einer bestimmten Sprechergemeinschaft darstellt.

Entsprechende Beispiele für phonatorische Einstellungen sind die Tendenz zur Heiserkeit, Rauheit, Aspiration oder die Tendenz zu hoher bzw. tiefer Stimmlage. Darüber hinaus beschreibt LAVER in diesem Konzept der phonatorischen Einstellungen auch die verschiedenen Stimmregister. Diese bezeichnet er als phonatorische Modi, wie z.B. Modalmodus ("modal modus") oder Falsetto ("falsetto modus"). Dabei steht Modalmodus für den normalen Schwingungsmodus der Glottis bei der Erzeugung stimmhafter Sprachlaute.

---

[1] Siehe MEYER-EPPLER (1969, S. 3).

Allerdings wendet LAVER den Begriff der artikulatorischen Einstellung und damit den Begriff der Stimmqualität eher auf intraindividuelle als auf interindividuelle Unterschiede an. Auch die phonatorischen Einstellungen werden immer relativ zum Modalmodus beschrieben, während auf interindividuelle Differenzen der neutralen Einstellung selbst, wie sie beispielsweise bei dem Modalmodus weiblicher und männlicher Sprecher auftreten, nicht eingegangen wird.

Im Gegensatz dazu stehen die interindividuellen Unterschiede der Phonation im Mittelpunkt dieser Arbeit. Hier wird der Frage nachgegangen, wie sich der Modalmodus für weibliche und männliche Stimmen unterscheidet. Allerdings ist nicht unbedingt gewährleistet, daß der Modalmodus, der nach LAVER als periodische und effiziente Schwingung ohne hörbares Rauschen beschrieben wird, bei weiblichen Stimmen überhaupt in dieser Form zu finden ist. Beispielsweise vermuten mehrere Autoren[2], daß bedingt durch den steilen spektralen Abfall der glottalen Welle und durch den höheren spektralen Schwerpunkt des glottalen Rauschens bei weiblichen Stimmen genereil die Tendenz zur Aspiration vorliegt ("breathiness").

Neben dem hier eingeführten sehr umfassenden Begriff der Stimmqualität soll noch der Begriff *Stimmklang* definiert werden.[3] Während Stimmqualität, wie oben beschrieben, alle sprecherspezifischen Signalmerkmale, also auch solche der Artikulation, beinhaltet, wird der Begriff Stimmklang enger gefaßt. Der Stimmklang umfaßt die Menge der sprecherspezifischen Signalmerkmale, die sich aus der Phonation ergeben. Vom Standpunkt der Spracherzeugung aus gesehen ist der Stimmklang das perzeptive Korrelat zur Schwingungsform des glottalen Volumenstroms.

---

[2] Siehe HOLMBERG et al. (1988, S. 522), KLATT (1987) und KARLSSON (1988, S. 66).

[3] Siehe auch LINDNER (1981, S. 71).

2.2 Perzeptiv relevante Stimmqualitätsmerkmale

In diesem Absatz wird dargelegt, welche Signalmerkmale eng mit Stimmqualität verbunden sind, d.h. welche Signalmerkmale die Differenzierung zwischen verschiedenen Stimmen bewirken.

SINGH et al. (1978) untersuchen die Ähnlichkeit weiblicher und männlicher Stimmen. Dazu wurden die von Sprechern verschiedenen Geschlechts gesprochenen Stimuli in einem Hörtest paarweise dargeboten. Die Stimuli mußten von den Hörern auf einer siebenstufigen Rangskala von maximaler Ähnlichkeit bis zu maximaler Unähnlichkeit klassifiziert werden. Entsprechend der Definition von LAVER (1980) und PASCAL (1985) bezeichnen SINGH et al. (1978) dies als Klassifikation der Stimmqualität. Die Stimuli wurden ebenfalls auf akustische und perzeptive Merkmale analysiert. Mittels Faktorenanalyse wurde versucht, die Dimensionen des errechneten Raumes der Stimmqualität auf bestimmte akustische oder perzeptive Signalmerkmale zurückzuführen. Die Hauptachse wird danach als "Männlich-Weiblich Dimension"[4] bezeichnet, ohne daß diese Dimension aber genauer durch akustische Korrelate spezifiziert wird. Es wird lediglich die wichtige Rolle der Grundfrequenz für diese Dimension nachgewiesen.

Das Ergebnis bedeutet anders formuliert, daß die größten Stimmqualitätsdifferenzen gerade zwischen weiblichen und männlichen Stimmen auftreten. Bei der Erkennung und Klassifikation von Stimmen führt der Hörer also wahrscheinlich zunächst eine Geschlechtsidentifikation aufgrund der stärksten Stimmqualitätsmerkmale durch. Stimmqualität selbst ist also ein wesentlicher Faktor für die Geschlechtsidentifikation. Diese Ideen werden in Kapitel 3 bei der dort entwickelten Theorie zur Primärnormierung wieder aufgenommen.

Zur Unterscheidung von Stimmen innerhalb einer Geschlechtsgruppe verwendet der Hörer nach Ansicht mehrerer Autoren unter-

---

[4] "male-female factor", siehe SINGH et al. (1978, S. 84).

schiedliche Strategien.[5] Je nachdem, ob vom Hörer männliche oder weibliche Stimmen zu unterscheiden sind, werden also unterschiedliche Signalkomponenten zur perzeptiven Ermittlung der Stimmqualität herangezogen. Auch nach diesem Ergebnis muß davon ausgegangen werden, daß am Anfang der Einstellung des Hörers auf eine neue Stimme zunächst die Geschlechtsidentifikation steht. Bei der Klassifikation männlicher Stimmen scheint F0 die Hauptrolle zu spielen. Daneben konnte eine weitere, den "Rauhigkeitsgrad" der Stimme erfassende Dimension zur Stimmqualitätsklassifikation männlicher Stimmen nachgewiesen werden.[6] Bei den weiblichen Stimmen spielt die Grundfrequenz eine weniger dominierende Rolle. Die wesentliche Stimmqualitätsdimension weiblicher Stimmen wird nur relativ allgemein als "komplexe Stimmqualitätskomponente" beschrieben.[7]

Anhand der Ergebnisse obiger Untersuchungen kann auch erklärt werden, warum die Auswahl der Anregungsfunktion gerade bei der Synthese der weiblichen Stimme eine wichtige Rolle spielt: Bei tiefer Grundfrequenz assoziiert der Hörer männliches Sprechergeschlecht und auch eine weitergehende Stimmqualitätsklassifikation orientiert sich hauptsächlich an F0. Bei hoher Grundfrequenz hingegen wird zunächst weibliches Sprechergeschlecht assoziiert. Für eine weitergehende Stimmqualitätsklassifikation werden dann neben F0 auch andere perzeptive Signalfaktoren (wahrscheinlich auch der Stimmklang) benutzt.

Bei Untersuchungen zur Stimmqualität anhand längerer Phrasen spielen auch temporale suprasegmentale Faktoren eine wichtige Rolle, während bei Untersuchungen anhand stationärer Einzelvokale mit den Sprechtraktdimensionen verbundene Signalkomponenten wichtig werden.

---

[5] Siehe PASCAL (1985), ARONOWITCH (1976) und MURRY et al.(1980).

[6] Siehe MURRY et al. (1980, S. 1298).

[7] Siehe MURRY et al. (1980, S. 1299).

Auf dem hier nicht gesondert berücksichtigten Gebiet der Sprechererkennung werden apparative Hilfsmittel wie Langzeitspektren oder Algorithmen zur Bestimmung der mittleren Grundfrequenz benutzt.[8] Diese Techniken nutzen die Tatsache, daß Stimmqualität ein segmentübergreifendes, sprecherspezifisches Langzeitcharakteristikum ist.

2.3 Perzeptive Faktoren zur Geschlechtsidentifikation

Während im vorigen Absatz erörtert wurde, welche Faktoren die Wahrnehmung der Stimmqualität beeinflussen, soll in diesem Absatz die Frage behandelt werden, welche perzeptiven Faktoren die Geschlechtsidentifikation ermöglichen. Da jedoch die Hauptdimension der Stimmqualität in den geschlechtsspezifischen Unterschieden liegt, sind Stimmqualitätsperzeption und Geschlechtsidentifikation eng gekoppelt. Diese Kopplung vollzieht sich auch durch die Stimmklangkomponente der Stimmqualität, und es wird versucht, die Bedeutung des Stimmklangs für die Geschlechtsidentifikation hervorzuheben, obwohl dieser Faktor in der Literatur nur selten betrachtet wird.

Außer Frage steht, daß es in verschiedenen Dialektgruppen geschlechtsspezifische Unterschiede in den suprasegmentalen Signalmerkmalen gibt.[9] Beispielsweise wurden geschlechtsspezifische Unterschiede in der Intonation des Amerikanischen festgestellt.[10] Diese gelernten bzw. sozial bedingten Unterscheidungsmerkmale sind aber erst bei der Perzeption mehrsilbiger Äußerungen feststellbar.

In dieser Arbeit stehen die sich bereits anhand von stationären Vokalen oder Diphthongen manifestierenden perzeptiven Faktoren der Geschlechtsidentifikation im Vordergrund, da die hier

---

[8] Siehe z.B. SICKERT (1983, S. 258ff).

[9] Weiterführende Literatur siehe GROß (1987, S. 66-80).

[10] Siehe BREND (1971).

durchgeführten experimentellen Untersuchungen auch anhand stationärer Vokale und anhand von Diphthongen durchgeführt wurden. Einige dieser Faktoren können bereits aus signalanalytischen Arbeiten vermutet werden. Zum einen ist bekannt, daß die Formanten bei Frauen um durchschnittlich 15% bis 20% höher liegen als bei Männern.[11] Daneben wurde für Frauen generell eine geringere Signalintensität nachgewiesen.[12] Zum anderen liegt der wesentlichste Signalunterschied aber wohl in den gravierenden Grundfrequenzunterschieden zwischen männlichen und weiblichen Stimmen.[13] Physiologisch ist dies mit dem geschlechtsspezifisch unterschiedlichen Wachstum des Kehlkopfes während der Pubertät zu begründen. Während bei Männern das starke Wachstum des Kehlkopfes zum Absinken der Grundfrequenz um fast eine Oktave führt, bleibt F0 bei Frauen fast unverändert. Neben dem Anstieg der Formantmittelwerte um 15% bis 20%, der mit der entsprechenden Sprechtraktverkürzung begründet wird, ist diese F0-Differenz mit fast 100% gravierend.

Bei der Perzeption von Einzelvokalen und Diphthongen sind es also vor allen diese drei Komponenten der Stimmqualität, die die Geschlechtsidentifikation ermöglichen:

1) die durchschnittliche Höhe der Grundfrequenz
2) die durch die Sprechtraktlänge definierte mittlere Lage der Formantfrequenzen
3) der Stimmklang

In den nachfolgend referierten Untersuchungen wurde versucht, diese Faktoren zu isolieren und ihren Einfluß auf die Geschlechtsidentifikation nachzuweisen.

---

[11] Siehe z.B. PETERSON et al. (1952) oder FANT (1975).

[12] Siehe MARKEL et al. (1972).

[13] Siehe z.B. PETERSON et al. (1952).

2.3.1 Die Grundfrequenz

In allen Untersuchungen zur Geschlechtsidentifikation wird davon ausgegangen, daß der gravierende Einfluß der Grundfrequenz keines gesonderten Nachweises bedarf. Z.B. kann anhand der Studie von COLEMAN (1973 und 1976), in der die Stärke des Einflusses von F0 und der Formanten auf die Geschlechtsidentifikation untersucht wird, abgelesen werden, daß F0-Änderungen allein in der Lage sind, erhebliche Verschiebungen in der Geschlechtsidentifikation zu bewirken.

2.3.2 Die mittlere Lage der Formanten

COLEMAN (1971) läßt zehn Frauen und zehn Männer längere Phrasen spechen, wobei aber die Anregung des Sprechtraktes mittels eines Larynxvibrators vorgenommen wird. Somit wurden sowohl die Sprechtrakte der männlichen als auch die der weiblichen Sprecher mit einer Grundfrequenz von 85 Hz angeregt. Sinn dieser künstlichen Anregung sollte es sein, alle idiosynkratischen Merkmale der glottalen Welle, also die persönlichen Stimmqualitätsmerkmale der Tonhöhe und des Stimmklangs zu eliminieren. Insbesondere wurden auch der Sprechtrakt der Frauen nur mit 85 Hz angeregt. COLEMAN schließt aus der hohen Identifikationsrate, daß das Stimmqualitätsmerkmal der mittleren Formantlage allein schon zur Geschlechtsidentifikation ausreicht. Dabei ist tiefe mittlere Formantlage ein Attribut für männliche Stimmqualität und hohe mittlere Formantlage eins für weibliche. Unterstützt wird das Ergebnis noch durch den Befund, daß männliche Sprecher mit hohen Formantlagen und weibliche Sprecher mit tiefen Formantlagen bezüglich des Geschlechts weniger sicher klassifiziert wurden.

Dieses Experiment wurde von BROWN et al. (1977) in ähnlicher Form wiederholt, allerdings wurde der Larynxvibrator bei 120 Hz betrieben. Es ergaben sich gleichwertige Resultate. Allerdings muß angemerkt werden, daß in beiden Untersuchungen neben dem Merkmal der Formantlagen die dynamischen Stimmqualitätsmerkmale (z.B. artikulatorische Gesten) nicht eliminiert wurden, so daß

auch diese zu der hohen Erkennungsrate beitragen konnten.

LASS et al. (1976) und SCHWARTZ et al. (1968) wiesen nach, daß die Identifikation des Geschlechts allein anhand isolierter stationärer geflüsterter Vokale möglich ist. Allerdings ist anzumerken, daß neben den Signalmerkmalen der Sprechtraktlänge evtl. Signalmerkmale der Phonation perzeptive Relevanz für die Geschlechtsidentifikation haben. Es ist anzunehmen, daß z.B. das von der weiblichen und der männlichen Glottis erzeugte Rauschen verschiedene spektrale Energieverteilungen aufweist.

SCHWARTZ (1968) zeigte, daß die Geschlechtsidentifikation schon anhand der isolierten Frikative / s / und / ʃ / möglich ist. Hier wird als Grund der im Signal nachweisbare und für Frauen höher liegende Schwerpunkt der spektralen Energieverteilung des im Sprechtrakt erzeugten Rauschens angegeben.

Allerdings kann die mittlere Formantlage verschiedener Sprecher auch innerhalb einer Geschlechtsgruppe sehr unterschiedlich sein. Die Messungen von PETERSON et al. (1952) aber auch von COLEMAN (1973 und 1976) und von BROWN et al. (1977) zeigen, daß diese Unterschiede für einzelne Sprecher sogar so groß sind, daß Überlappungen bezüglich der Formantfrequenzen eines Vokals zwischen den Geschlechtsgruppen auftreten können. Ebenfalls können die Unterschiede in der mittleren Formantlage nicht ausschließlich mit Unterschieden in der Sprechtraktlänge begründet werden.

MATTINGLY (1966) stellte fest, daß die Separation zwischen den für beide Geschlechtsgruppen errechneten Mittelwerten der Formantlagen viel größer ist, als es aufgrund der durchschnittliche Variation der Sprechtraktlänge erwartet werden dürfte. Es wird daher vermutet, daß linguistische Konventionierungen eine Rolle spielen. Dazu wiesen WEINBERG et al. (1971) nach, daß perzeptive Geschlechtsidentifikation auch bei fünf- bis sechsjährigen Kindern möglich ist, obwohl hier keine geschlechtsspezifischen Formant- oder Grundfrequenzdifferenzen vorliegen.

In den Arbeiten von FANT (1966 und 1975) wurden eine Vielzahl von Formantfrequenzdaten weiblicher und männlicher Sprecher aus verschiedenen Sprachen verglichen. FANT stellte eine Nichtproportionalität in der Änderung der Formantfrequenzen zwischen weiblichen und männlichen Sprechern abhängig von der Vokalqualität als einheitliche Tendenz über mehrere Sprachen fest.[14] Nichtproportionalität heißt, daß einzelne Formantfrequenzen beim Vergleich verschiedener Vokale bei weiblichen Sprechern eine unterschiedliche prozentuale Frequenzerhöhung gegenüber den Formantfrequenzen männlicher Sprecher aufweisen. FANT führt diese ungleichmäßigen Frequenzänderungen darauf zurück, daß die physiologischen Sprechtraktdifferenzen nicht allein in einer proportionalen Verkleinerung der Sprechtraktgröße bestehen. Beispielsweise ist die Pharynx bei Frauen überproportional kürzer als bei Männern.

2.3.3 Der Stimmklang

CARRELL (1981) generierte für einige Vokale ein Stimuluskontinuum bezüglich der Formantfrequenzen ausgehend von männlichen bis zu weiblichen mittleren Formantfrequenzwerten. Aufgrund der Ergebnisse der Arbeit von MONSEN et al. (1977) wurden zwei Gruppen von Stimuli, zum einen mit symmetrischer und zum anderen mit antisymmetrischer glottaler Wellenform, generiert. In Abhängigkeit von der mittleren Formantlage ergibt sich eine Überlappung in der Geschlechtsidentifikation. Im Gegensatz zur antisymmetrischen Wellenform stellt sich bei der symmetrischen glottalen Wellenform schon bei tieferen Formantfrequenzen die "weiblich"-Identifikation ein.

In weiteren experimentellen Arbeiten untersucht COLEMAN (1973 und 1976), welcher der Faktoren Grundfrequenz oder mittlere Formantlage den stärkeren Einfluß auf die Geschlechtsidentifikation ausübt. Wegen der Erstellung der Stimuli mit dem Larynxvi-

---

[14] Siehe auch Abs. 3.4.3.

brator liefert diese Untersuchung aber, wie im folgenden dargelegt wird, eigentlich Ergebnisse bezüglich des Einflusses des Stimmklangs auf die Geschlechtsidentifikation. In Experiment 1 der Untersuchung von COLEMAN werden natürlich gesprochene Äußerungen männlicher und weiblicher Sprecher perzeptiv bezüglich des Sprechergeschlechts und akustisch bezüglich der Formant- und F0-Mittelwerte analysiert. Ergebnis ist, daß die Korrelation zwischen Geschlechtsassoziation und der Grundfrequenz wesentlich höher ist als die Korrelation zwischen Geschlechtsassoziation und den Formantmittelwerten.

In Experiment 2 der Untersuchung von COLEMAN hingegen werden die Äußerungen der weiblichen und der männlichen Sprecher mit dem Larynxvibrator erzeugt. Dabei müssen sowohl die weiblichen als auch die männlichen Sprecher den Larynxvibrator bei 120 Hz und bei 240 Hz betreiben. Das damit verbundene Signalmerkmal wird mit "tief" bzw. "hoch" bezeichnet. Es werden also Stimuli mit allen möglichen Kombinationen von Grundfrequenz und von mittlerer Formantlage, die hier je nach Geschlecht (d.h. in bezug auf die Sprechtraktgröße) mit "groß" bzw. "klein" bezeichnet werden, generiert. Neben den "natürlichen" Kombinationen S[groß/tief] für "männlich" und S[klein/hoch] für "weiblich" werden auch die "unnatürlichen" Kombinationen S[groß/hoch] und S[klein/tief] erzeugt, in denen demnach sich widersprechende Geschlechtsmerkmale in einem Stimulus vereinigt wurden. Die Hörer mußten die Stimuli ohne Alternative entweder als männlich oder als weiblich klassifizieren ("forced-choice-situation"). Die Gesamtergebnisse nach COLEMAN sind hier graphisch dargestellt (siehe Abb. 2.1).

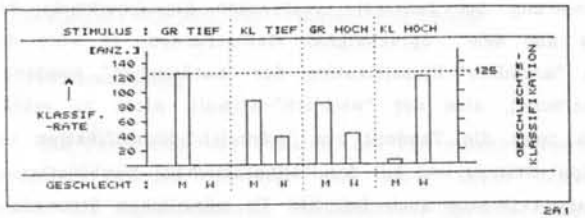


Abb. 2.1: Perzeptive Geschlechtsklassifikation von Stimuli, die mittels Larynxvibrator erzeugt wurden. TIEF: F0 = 120 Hz, HOCH: F0 = 240 Hz, GR: männl. Sprechtrakt, KL: weibl. Sprechtrakt. Daten nach COLEMAN (1976, S.176, Tab. 3 und Tab. 4) (vergl. auch mit Abb. 5.2 und mit Abb. 6.3 dieser Arbeit)

Die Ergebnisse der "unnatürlichen" Kombinationen zeigen annähernd gleiche Klassifikationstendenz für beide Kombinationen in Richtung "männlich"-Klassifikation. Nach Experiment 1 wäre aber ein stärkerer F0-Einfluß zu vermuten, so daß COLEMAN von einer "perzeptiven Schwäche der weiblichen Grundfrequenz" in Experiment 2 spricht.[15] Es ist aber nicht haltbar, einen perzeptiven Faktor in einem Experiment für stark und in einem anderen Experiment für schwach zu halten. Diese Diskrepanz ist allerdings unter Einbeziehung des perzeptive Faktors Stimmklang lösbar. Experiment 1 und 2 unterscheiden sich gerade durch die Verwendung des Larynxvibrators. Neben der Grundfrequenznormierung wird der Larynxvibrator nach Aussage von COLEMAN dazu verwendet, die idiosynkratischen Merkmale des Stimmklangs zu eliminieren. Die Eliminierung der persönlichen Merkmale des Stimmklangs darf aber nicht mit Eliminierung des perzeptiven Faktors Stimmklang gleichgesetzt werden. Vielmehr tritt an die Stelle des persönlichen Stimmklangs des Sprechers eine durch den Larynxvibrator selbst hervorgerufene Stimmklangkomponente, die vielleicht mit maschinellem Stimmklang bezeichnet werden könnte. Hätte tatsächlich

[15] Siehe COLEMAN (1976, S. 179).

eine Eliminierung der perzeptiv relevanten Signalmerkmale des Stimmklangs aus dem Sprachsignal stattgefunden, so wäre die auftretende "männlich"-Klassifikation der "natürlichen" Kombination S[klein/hoch], also der "weiblich"-Stimuli, nicht zu erklären. Jedoch zeigt die Tendenz zur "männlich"-Klassifikation bei den "weiblich"-Stimuli und bei den "unnatürlichen" Kombinationen, daß der Larynxvibrator auch bei 240 Hz männlichen Stimmklang impliziert. In diesem Experiment tritt also das gleiche Phänomen auf, das bereits in Kapitel 1 für die Synthese der weiblichen Stimme dargelegt wurde und das auch in den experimentellen Untersuchungen dieser Arbeit auftritt: Für die Synthese weiblicher Stimmen ist die künstliche Modellierung des Stimmklangs ein größeres Problem als für die Synthese männlicher Stimmen.

Festzuhalten bleibt, daß neben den Faktoren Grundfrequenz und mittlere Formantlage auch der Faktor Stimmklang wesentlich zur Geschlechtsidentifikation beiträgt. Entscheidend ist der Stimmklang zur Erreichung guter Stimmqualität bei der Synthese der weiblichen Stimme.

3 Vokalperzeption und Geschlechtsidentifikation

3.1 Problemstellung

Die Perzeption von Vokalen, die von verschiedenen Sprechern produziert wurden, bereitet dem Hörer keine Schwierigkeiten, obwohl die Signalparameter verschiedener Sprecher stark differieren können. Große interindividuelle Signalunterschiede sind somit für die Perzeption kein Problem. Insbesondere wurden bezüglich der Formantfrequenzen große und systematische Differenzen zwischen Männern, Frauen und Kindern festgestellt. Das bedeutet, daß die von verschiedenen Sprechern produzierten Vokale trotz unterschiedlicher Formantfrequenzen gleiche Vokalqualität haben. Wegen dieses Phänomens der Invarianz der Vokalqualität auch bei großen interindividuellen Signalunterschieden wurden Normierungstheorien für die Perzeption von Vokalen vorgeschlagen. Insbesondere wird durch den experimentellen Befund der Invarianz der Vokalqualität die Existenz einer perzeptiven Normierung der Formantfrequenzen nahegelegt.

In dieser Arbeit steht der Fall der Umstellung der Perzeption von einem weiblichen auf einen männlichen Sprecher und umgekehrt im Mittelpunkt. Es geht also um die Frage, wodurch die schnelle Einstellung auf einen neuen, evtl. auch unbekannten Sprecher geleistet wird. Dazu wird die Existenz eines sofort, d.h. in Echtzeit arbeitenden Normierungsprozesses angenommen. Dieser perzeptive Prozeß wird hier mit *Primärnormierung* bezeichnet, und es wird gezeigt, wie eng die Primärnormierung mit den perzeptiven Prozessen, die bei der Geschlechtsidentifikation ablaufen, zusammenhängt.

3.2 Das Modell der verschiebbaren Vokalschablone

3.2.1 Formanten und Vokalraum

Vokalqualität ist definiert als die Summe aller perzeptiv bzw. phonetisch relevanten Faktoren eines Vokals. Die Normierung

der Formantfrequenzen löst dann das Problem, in welcher Form die akustischen Signalparameter, insbesondere die Formantfrequenzen, auf die Vokalqualität abzubilden sind.

Der Formantbegriff ist seit dem 19. Jahrhundert bekannt. Es sind zwei Definitionen gängig. So ist es möglich, Formanten signalorientiert als Maxima der spektralen Einhüllenden oder erzeugungsorientiert als Resonanzfrequenzen des Sprechtraktes zu definieren.[1] Zunächst wurden die Formanten allein, genauer gesagt die absolute Lage der Maxima der Formanten, Formantfrequenzen genannt, als hinreichende perzeptiv relevante Signalfaktoren der Vokalqualität betrachtet. Nach dieser Theorie, auch Absoluttheorie genannt, ist die Klangfarbe eines Vokals also nur von der absoluten Lage der Formantfrequenzen abhängig. Die Absoluttheorie stellt somit eine erste Form der Abbildung der Signalparameter in den Raum der Vokalqualität dar. Es wird keine Normierung durchgeführt. Insbesondere wird Vokalqualität nach dieser Theorie auch als von anderen Signalparametern wie z.B. F0 und Stimmqualität unabhängig betrachtet.

Für die Betrachtung eines einzelnen Sprechers ist diese Theorie haltbar. Allerdings kann die Absoluttheorie nicht die interindividuellen Signalunterschiede insbesondere beim Vergleich von Männern, Frauen und Kindern erklären. JOOS (1948, S. 82-87) stellte fest, daß bei der Realisierung des gleichen Vokals aufgrund der unterschiedlichen Kopfgröße verschiedener Sprecher große Formantverschiebungen auftreten können. Er konstatiert, daß z.B. durch die unterschiedliche Kopfgröße von Mann und Kind eine Formantverschiebung von sieben Halbtönen auftreten kann. Das entspricht bei einem Sprecher einer Vokalqualitätsverschiebung von / æ / nach / u /. Diesen Befund starker idiosynkratischer Signalunterschiede belegt auch die von PETERSON et al. (1952) durchgeführte Großuntersuchung. Es wurden die Formantfrequenzen, die Intensität und die Grundfrequenz einer 76 Personen starken Sprechergruppe, bestehend sowohl aus Männern als auch aus Frauen

[1] Siehe STEVENS et al. (1961).

und Kindern, bei der Produktion der Vokale des Amerikanischen gemessen. So beträgt im Mittel die Erhöhung der Formantfrequenzen der Frauen gegenüber den Männern 17% und die Erhöhung der Formantfrequenzen der Kinder gegenüber den Männern 30%. In Abbildung 3.1 sind die mittleren Formantfrequenzen der Eckvokale[2], d.h. der Vokale, deren Formantfrequenzen für einen Sprecher die größt- bzw. kleinstmöglichen Werte annehmen, für männliche und weibliche Sprecher dargestellt. Abbildung 3.1 macht deutlich, daß bei einer genaueren Betrachtung der Eckvokale die Entwicklung von Perzeptionsmodellen, die die Normierung der Formantfrequenzen berücksichtigen, unumgänglich ist.

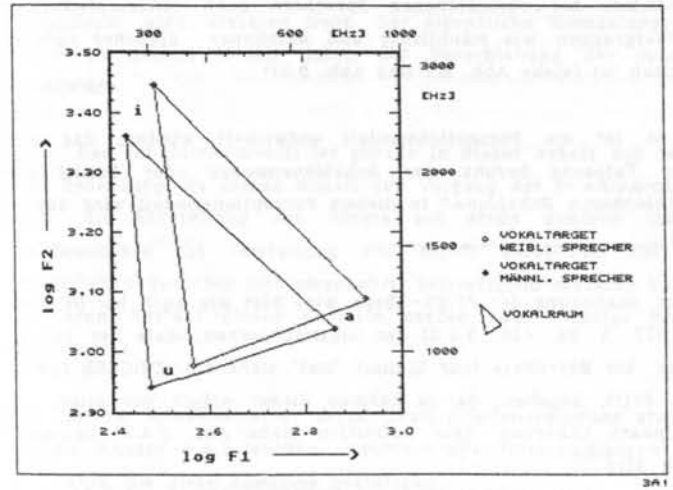


Abb. 3.1: Der Vokalraum weiblicher und männlicher Sprecher. Daten nach PETERSON et al. (1952)

Das innerhalb der die Eckvokale verbindenden Linien liegende F1/F2-Gebiet umfaßt den geschlechtsspezifischen *Vokalraum*[3]. Die

[2] Siehe auch JOOS (1948) und Abs. 3.4.1.

[3] "vowel space", siehe LIEBERMAN et al. (1988, S. 178).

Formantfrequenzen der vom zugehörigen Sprecher realisierten Vokale liegen in diesem Gebiet. Der Vokalraum ist für weibliche Sprecher gegenüber männlichen Sprechern insgesamt zu höheren Frequenzen hin verschoben.

3.2.2 Das Modell der verschiebbaren Schablone

In der akustisch orientierten Realisation der Vokalräume verschiedener Sprecher wird das Normierungsproblem offensichtlich: Die Vokalräume verschiedener Sprecher können gegeneinander verschoben liegen. Bemerkenswert ist jedoch die Tatsache, daß die Form des Vokalraumes, d.h. die relative Lage der Vokaltargets zueinander, bei verschiedenen Sprechern auch unterschiedlicher Sprechergruppen wie männlicher und weiblicher Sprecher nahezu identisch ist (siehe Abb. 3.1 und Abb. 3.3)[4].

So ist ein Perzeptionsmodell entwickelt worden, das auf dieser Tatsache beruht: das *Schablonenmodell* oder *Modell der verschiebbaren Schablone.*[5] In diesem Perzeptionsmodell wird davon

---

[4] Zur Skalierung der F1/F2-Ebene wird hier wie auch bei NEAREY (1977, S. 96, Abb. 3.1.2) der logarithmischen Skala der Vorzug vor der Mel-Skala (zur Einheit "mel" siehe z.B. ZWICKER (1982, S. 57ff)) gegeben, da im Rahmen dieser Arbeit vor allem die "lineare Näherung" (zur Definition siehe Abs. 3.4.3) betrachtet wird.

[5] "sliding template model", siehe NEAREY (1977, S. 93). Ähnliche Modellvorstellungen findet man auch bei JOOS (1948) und bei AINSWORTH (1975).
Gegenüber dem ebenfalls gebräuchlichen Terminus "Bezugssystem" hebt der hier Verwendung findende Begriff "Schablone" den perzeptiven Aspekt hervor. Insbesondere wird durch diesen Begriff die graphische Vorstellung von auf einer Schablone fixierten Targetpunkten, die jeweils die mittleren Formantfrequenzwerte der Vokale darstellen, nahegelegt (siehe auch Abb. 3.4 bis Abb. 3.6).

ausgegangen, daß die perzeptive Vokalidentifikation mittels einer für alle Sprecher identischen Vokalschablone durchgeführt wird, wobei diese Schablone im F1/F2-Raum verschiebbar ist und sich auf jeden Sprecher gesondert justiert. Bei der Umstellung von einem Sprecher auf einen anderen wird dann von Schablonenverschiebung gesprochen. Die Annahme einer solchen Schablonentheorie wird insbesondere dadurch motiviert, daß die Schablonenform in der doppeltlogarithmischen F1/F2-Ebene in erster Näherung invariant gegenüber den Vokalrealisationen verschiedener Sprecher ist, genau wie auch Vokalqualität sprecherunabhängig definiert werden kann. Hier liegt also ein Modell vor, das die Perzeption der von verschiedenen Sprechern realisierten Vokale über einen einheitlichen Verarbeitungsprozeß, der durch die Vokalschablone veranschaulicht wird, erklären kann. Der eigentliche Normierungsprozeß wird in diesem Modell durch die Verschiebung der Schablone realisiert.

Das Schablonenmodell ist gerade in dieser Arbeit von zentraler Bedeutung, da dieses Modell den Vorgang der Primärnormierung, d.h. die Einstellung des Hörers auf einen anderen Sprecher, insbesondere die Umstellung von einem weiblichen auf einen männlichen Sprecher und umgekehrt, befriedigend erklären kann. Im weiteren Verlauf dieses Kapitels werden zwei wichtige Probleme untersucht:

1) Ist die Annahme einer perzeptiven Primärnormierung statthaft? In Absatz 3.4.1 werden experimentelle Untersuchungen angeführt, die diese Annahme bestätigen.

2) Welche perzeptiv relevanten Signalmerkmale lösen die Primärnormierung aus? Diese Frage führt zu dem in Absatz 3.3 ausgeführten Kerngedanken dieses Kapitels: Primärnormierung und Geschlechtsidentifikation sind eng miteinander verbundene Vorgänge.

3.2.3 Normierungsfaktoren und Klassifikationsfaktoren

Im Modell der verschiebbaren Vokalschablone ist es wichtig, die perzeptiv relevanten Signalmerkmale, die zur Bestimmung der Vokalqualität führen, nach Normierungsfaktoren und nach Klassifikationsfaktoren einzuteilen. Als *perzeptive Normierungsfaktoren* werden diejenigen Signalmerkmale bezeichnet, die direkt zur Verschiebung der Schablone führen. Die Signalmerkmale, die nach vollzogener Einstellung der Schablone den Vergleich des zu klassifizierenden Vokals mit den Targetwerten der Schablone ermöglichen, heißen *perzeptive Klassifikationsfaktoren.*

3.2.4 Zur Dimension des Vokalraums

Bisher wurde stillschweigend vorausgesetzt, daß der Raum der Vokalqualität zweidimensional ist. Prinzipiell ist Vokalqualität aber eine multidimensionale Größe, da sie perzeptiv durch die spektrale Energieverteilung, d.h. die Form der spektralen Einhüllenden bestimmt ist. Anhand von Syntheseexperimenten wurden von DELATTRE et al. (1952) die Frequenzen der beiden ersten Formanten F1 und F2 als die dominierenden Parameter der Vokalqualität ausgewiesen. Es wurde gezeigt, daß mit zwei Formanten alle Vokale befriedigend synthetisiert werden können. Mit einem Formanten gelingt bereits eine befriedigende Synthese der Hinterzungenvokale. POLS et al. (1969) und PLOMP (1976) wiesen nach, daß die perzeptiven Abstände von Vokalen bezüglich ihrer Klangfarbe mittels multidimensionaler Skalierung bei einer Reduktion auf zwei bis drei Komponenten einen Großteil der Varianz erklären. Der Vergleich der akustischen und perzeptiven Faktoren der Vokale zeigte, daß F1 und F2 die akustischen Korrelate der beiden Hauptdimensionen der Vokalqualität darstellen.

3.2.5 Die perzeptive Relevanz des Schablonenmodells

Obwohl die Invarianz der Form der Vokalschablone für die Perzeption unterschiedlicher Sprecher die Schablonentheorie bereits motiviert, wurde durch die Untersuchungen von NEAREY (1977) die Vokalschablone als im menschlichen Perzeptionssystem existent nachgewiesen. In den von NEAREY (1977, S. 105–133) durchgeführten Hörtests wurde von den Testpersonen ein Netz von 200 synthetischen Stimuli, also ein Netz von 200 Punkten in der F1/F2-Ebene klassifiziert. Diese Vokale [V] wurden in zeitlichem Zusammenhang mit einer synthetischen [i]-Realisation als [iV–Vi]-Formen dargeboten. Der Rahmenvokal [i], durch den die perzeptive Normierung bewirkt werden sollte, wurde einerseits mit den Formantwerten eines Mannes, andererseits mit den Formantwerten eines Kindes synthetisiert. Die Klassifikation des Stimuluskontinuums ergab aber keine lokale Verschiebung der Phonemgrenzen im [i]-Bereich, sondern eine globale Verschiebung der Phonemgrenzen über den gesamten Vokalraum hinweg. Aus diesem Ergebnis kann die perzeptive Existenz der Vokalschablone gefolgert werden.

Die Sonderstellung des Vokals [i] als Auslöser der Schablonenverschiebung resultiert nach NEAREY daraus, daß nur dieser Vokal für verschiedene Sprecher auch schon anhand von F1 und F2 eindeutig definiert ist. D.h. anhand von F1 und F2 kann die Lage der Schablone nur ausgehend vom [i] eindeutig bestimmt werden. Alle anderen Formantwerte sind bezüglich der Vokalqualität mehrdeutig, da der zugehörige Punkt in der F1/F2-Ebene innerhalb der Vokalräume verschiedener Sprecher liegt.[6]

3.2.6 Extrinsische und intrinsische Faktoren

Das von NEAREY (1977) durchgefürte Experiment macht deutlich, daß die perzeptiv relevanten Signalfaktoren, die zur

---

[6] Zur Begründung siehe NEAREY (1977, S. 99–100 und S. 96, Fig. 3.1.2).

Bestimmung der Vokalqualität führen, nicht unbedingt im Vokalsegment selbst liegen müssen. Nach AINSWORTH (1975, S. 105) unterscheidet man *extrinsische Faktoren* und *intrinsische Faktoren*. Während die intrinsischen Faktoren alle perzeptiv relevanten Signalmerkmale umfassen, die im Vokalsegment selbst liegen, zählen alle außerhalb des Segments liegenden Faktoren, die dessen Vokalqualität beeinflussen, zu den extrinsischen Merkmalen. Beispielsweise sind die Formantlagen des Vokals /i/ in dem Experiment von NEAREY extrinsische Faktoren, da sie die Lage der Vokalschablone beeinflussen und damit zur Änderung der Vokalqualität des Segments selber beitragen.

3.3 Der Zusammenhang zwischen Primärnormierung und Geschlechtsassoziation.

3.3.1 Die These

Die Frage nach den perzeptiven Faktoren, die die Verschiebung der Vokalschablone bewirken, also die Frage nach den perzeptiven Normierungsfaktoren führt zum zentralen Punkt dieses Kapitels. Wie in Absatz 3.4.2 näher ausgeführt wird, werden in der Literatur fast ausschließlich die Formanten als perzeptive Normierungsfaktoren betrachtet. Einige Autoren haben in experimentellen Untersuchungen zwar auch den Einfluß von F0 auf die Vokalqualität nachgewiesen, jedoch wird F0 in Arbeiten, die sich direkt mit dem Normierungsproblem befassen, nicht als Faktor für die Verschiebung der Vokalschablone in Betracht gezogen.

Zentraler Gedanke dieses Kapitels ist, daß bei der Einstellung des Perzeptionsmechanismus auf einen neuen, evtl. unbekannten Sprecher die *Primärnormierung* und die Klassifikation der Stimmqualität, d.h. in erster Linie die *Geschlechtsassoziation* bzw. Geschlechtsidentifikation, parallel ablaufende Prozesse sind. Prinzipiell ist also die Verschiebung der Vokalschablone nicht trennbar von der Geschlechtsidentifikation. Während einerseits die Lage der perzeptiven Vokalschablone zur Geschlechtsidentifikation beiträgt, beeinflußt andererseits die Geschlechtsidentifikation die Verschiebung und Lage der Vokalschablone.

Somit kann angenommen werden, daß neben den Formanten selbst auch F0 und alle weiteren perzeptiv relevanten Signalmerkmale, die zur Geschlechtsidentifikation beitragen, einen Einfluß auf die Primärnormierung haben. Es ist Ziel einer experimentellen Untersuchung dieser Arbeit[7] zu zeigen, daß allein die mit einem Stimulus verbundene Geschlechtsassoziation Einfluß auf die perzipierte Vokalqualität hat. Nach dieser Theorie tritt insbesondere die Größe Stimmklang als perzeptiv relevanter Faktor zur Primärnormierung hinzu. Somit kann prinzipiell die Stimmqualität selbst zu einem die Vokalqualität beeinflussenden Faktor werden.

Die Vorstellung der engen Verbindung der Schablonenverschiebung mit der Geschlechtsidentifikation kann auch als Idee für einen echtzeitfähigen Normierungsalgorithmus dienen. Grundlage des Algorithmus sind zwei vorab gelernte Schablonneneinstellungen mit den entsprechenden Targetwerten für jeden Vokal. Es ist lediglich eine Entscheidung bezüglich des Sprechergeschlechts zu treffen, die in der Hauptsache auf der Höhe der Grundfrequenz basieren sollte. Durch diese Entscheidung ist die Lage der Vokalschablone definiert, anhand deren die Bestimmung der Vokalqualität vorgenommen werden kann.

Jedoch benutzen die meisten bekannten Normierungsalgorithmen die Mittelwerte oder Extremwerte der Formantfrequenzen, die aus der Analyse vieler Sprachbeispiele eines Sprechers gewonnen wurden, als Normierungsfaktoren. Auch bei anderen Konzepten zur perzeptiven Normierung, die man anhand von Perzeptionsexperimenten zu verifizieren versuchte, stehen meist die Formantfrequenzen selbst im Vordergrund. Nur selten wird die F0-Information als wichtiger perzeptiver Normierungsfaktor anerkannt.

---

[7] Siehe Kap. 6, Exp. 5.

3.3.2 Hinweise auf die Richtigkeit der These

Im folgenden werden einige experimentelle Befunde aus der Literatur diskutiert, die die These der Kopplung von perzeptiver Primärnormierung und Geschlechtsidentifikation untermauern:

1) Der geringe Einfluß synthetischer Sprache auf die Verschiebung der Vokalschablone.
   Bei der Perzeption natürlicher und synthetischer Sprache wurde eine Differenz in der Verschiebungsweite der perzeptiven Vokalschablone bei der Umstellung von männlichen auf weibliche Sprecher festgestellt. Die Verschiebungsweite fällt bei synthetischen Stimuli geringer aus. Diese Differenz kann durch die bei den synthetischen Stimuli fehlende weibliche Stimmqualität erklärt werden und beweist somit die Wichtigkeit der Geschlechtsassoziation für die perzeptive Vokalnormierung.

2) Der Einfluß der Grundfrequenz auf die Vokalqualität.
   Obwohl die meisten Normierungstheorien von einer Normierung nur anhand der Formanten selbst ausgehen, gibt es experimentelle Arbeiten, die allein den Einfluß von F0 auf die Vokalqualität bei Konstanz aller anderen Signalparameter nachweisen. Da F0 der wichtigste perzeptive Faktor für die Geschlechtsidentifikation ist,[8] untermauert auch dies den Zusammenhang von Primärnormierung und Geschlechtsidentifikation.

3) Interindividuelle Unterschiede in der Perzeption synthetischer Vokale.
   Es wurde festgestellt, daß große Unterschiede in Abhängigkeit vom Hörer bei der Perzeption desselben synthetischen Vokals auftreten können. Dieser Befund könnte in Verbindung mit einer unterschiedlichen Geschlechtsassoziation stehen, die verschiedene Hörer mit demselben Stimulus verbinden. Gerade bei synthetisch generierten Stimuli kann eine große Offenheit

[8] Siehe Abs. 2.3.

bezüglich der Geschlechtsassoziation vorliegen. Diese Idee führt direkt zu einer im Rahmen dieser Arbeit durchgeführten experimentellen Untersuchung.[9]

4) Das Versagen extrinsischer Normierungsfaktoren bei der Perzeption natürlicher Sprache.
   Es konnte nachgewiesen werden, daß sich die Vokalqualität desselben synthetischen Stimulus ändert, wenn die mittlere Lage der Formanten des Rahmensatzes (extrinsische Normierungsfaktoren[10]), in den der Stimulus eingebettet ist, variiert wird. Diese Änderung der Vokalqualität wird dadurch erklärt, daß die Lage der perzeptiven Vokalschablone durch den Rahmensatz beeinflußt wird. Allerdings konnten diese Vokalqualitätsänderungen nicht anhand natürlich gesprochener Sprache und auch nicht anhand synthetischer Sprache hoher Qualität nachgewiesen werden. Da qualitativ weniger hochwertige Synthese frei von vielen Stimmqualitätsmerkmalen ist, ist dies ein Indiz dafür, daß zumindest eine Stimmklangänderung erfolgen muß, damit eine Schablonenverschiebung entsteht. Der Hörer nimmt erst bei erfolgter Stimmklangänderung an, daß ein Sprecherwechsel vorliegt. Dies unterstützt die These, daß die Echtzeitnormierung vor allem von dem intrinsischen Faktor Stimmqualität abhängt.

Die hier angesprochenen Punkte werden im folgenden detailliert diskutiert.

3.3.2.1 Zum geringen Einfluß synthetischer Sprache auf die Verschiebung der Vokalschablone

Es ist noch nicht geklärt, warum bei der Klassifikation synthetischer Vokale durch Erhöhung der Formanten und F0 von männlichen auf weibliche Durchschnittswerte nicht die entspre-

[9] Siehe Kap. 6, Exp. 5.

[10] Definition in Abs. 3.2.6.

chend zu erwartende Verschiebung der perzeptiven Vokalschablone um rund 15–20% erreicht wird.

SLAWSON (1968) benutzt Formantensynthese zur Generierung der Stimuli. Die Anregung des Sprechtraktmoduls geschah mit einer trapezförmigen Anregungsfunktion. Es wurden stationäre Vokale mit vier Formanten generiert. Die Grundfrequenzen betrugen 135 Hz und 270 Hz. Es wurde festgestellt, daß sich durch die Erhöhung von F0 um eine Oktave die Vokalqualität trotz Konstanz der Formantfrequenzen ändert. Die so entstehende Vokalqualitätsdifferenz bei der Erhöhung von F0 um eine Oktave kann aber auf ein Minimum reduziert werden, wenn die Formanten um 10% erhöht, d.h. wenn die Formantfrequenzen mit dem Faktor 1.1 multipliziert werden.

Auch AINSWORTH (1975) benutzte Formantensynthese zur Generierung der Stimuli. Die Anregung des Sprechtraktmoduls erfolgte hier über einen Impulsgenerator.[11] Es wurden zweiformantige Stimuli mit Vokalen im [hVd]-Kontext generiert. Die Grundfrequenz betrug 120 Hz, 240 Hz und 360 Hz. Zusätzlich wurde versucht, eine Verschiebung des perzeptiven Vokalraumes anhand von Rahmenvokalen[12] zu erreichen. Dazu wurden die Formantfrequenzen der Rahmenvokale, die den durchschnittlichen Formantwerten eines Mannes entsprachen, auch um 15% und um 30% erhöht. Die Hörer mußten ein Stimuluskontinuum bezüglich F1 und F2 klassifizieren. Da AINSWORTH auch das Modell eines in der F1/F2-Ebene verschiebbaren Vokalraumes benutzt, berechnete er die mittlere Verschiebung der perzeptiven Vokalschablone durch F0 und durch die Formanten der Rahmenvokale anhand der Verschiebung der zu einem Vokal klassifizierten Stimuli. Das in diesem Zusammenhang wichtige Teilergebnis ist die nur sehr geringe Verschiebung der perzeptiven Vokalschablone um 6% bis 7%, trotz einer Erhöhung von F0 um eine Oktave

[11] Siehe HOLMES et al. (1964).

[12] Im folgenden werden die zeitlich kurz vor dem zu klassifizierenden Stimulus dargebotenen Signale, bzw. die Signale, in denen der zu klassifizierende Stimulus eingebettet ist, als Rahmensignale bzw. Rahmenvokale oder Rahmensätze bezeichnet.

und einer gleichzeitigen Erhöhung der Formantfrequenzen um 30%. Allerdings konstatiert AINSWORTH, daß dies auch durch einige Faktoren des Untersuchungsverfahrens selbst bedingt sein kann.

CARLSON, GRANSTRÖM und FANT (1970) stellen bei einer ähnlichen Untersuchung fest, daß trotz der bei der Synthese vorgenommenen Einstellung der Grundfrequenz und der Formantfrequenzen auf aus akustischen Messungen bekannte Durchschnittswerte für männliche und weibliche Sprecher nur eine um den Faktor zwei bis vier kleinere Verschiebung der Vokalschablone nachgewiesen werden kann.

Zunächst besagen diese Resultate, daß allein F0 die Vokalqualität zu ändern vermag. Des weiteren ist den hier referierten experimentellen Daten gemeinsam, daß die bei der Umstellung von einer männlichen auf eine weibliche Stimme zu beachtende Verschiebung der Vokalschablone um 15% bis 20% bei der Perzeption *synthetischer* Sprache nicht erfolgt. Die experimentell ermittelten Werte liegen bei 10% (SLAWSON (1968)), 6% bis 7% (AINSWORTH (1975)) und 5% bis 10% (CARLSON et al. (1970)). Obwohl auch andere Gründe für diese Differenz verantwortlich sein könnten, wird hier die These vertreten, daß die Hörer mit den dargebotenen Stimuli wegen fehlender Stimmqualitätsmerkmale wie z.B. des typisch weiblichen Stimmklangs keine weibliche Stimme assoziierten. Es ist nicht anzunehmen, daß ein einfacher Impulsgenerator oder eine trapezförmige Anregungsfunktion in der Lage ist, weiblichen Stimmklang zu modulieren. Es könnte also genau das Fehlen der Geschlechtsidentifikation der Grund für die nur schwache Verschiebung der perzeptiven Vokalschablone sein.

Andererseits ist aber auch die Tatsache, daß überhaupt durch die F0-Erhöhung auch eine Verschiebung der perzeptiven Vokalschablone zu höheren Frequenzen erfolgt, bemerkenswert. Eine F0-Anhebung bewirkt immer eine Verschiebung der Vokalschablone in die gleiche Richtung, nämlich zu höheren Frequenzen. Diese Korrelation muß mit den gelernten Einstellungen der Schablone für weibliche und männliche Sprecher zusammenhängen. Durch eine hohe Grundfrequenz wird die Erwartung eines weiblichen Sprechers

geweckt, wie es die tendenzielle Verschiebung der Schablone belegt. Auch diese Überlegungen basieren auf der Annahme eines engen Zusammenhangs zwischen Geschlechtsassoziation und Schablonenverschiebung. Womit sonst wäre die regelmäßige Verschiebung der Vokalqualität an beliebigen Orten in der F1/F2-Ebene und bei beliebigen Grundfrequenzen zu begründen?

3.3.2.2 Der Einfluß der Grundfrequenz auf die Vokalqualität

Es wird versucht, entgegen der Haupttendenz in der Literatur, nach der nur die Formanten selbst als Grundlage für die Normierung berücksichtigt werden, hier die Bedeutung von F0 gerade für die Primärnormierung herauszustellen. Der Einfluß von F0 ist ein Hinweis darauf, daß Geschlechtsassoziation und Vokalnormierung eng verwandte perzeptive Prozesse sind.

In Absatz 3.4. werden die wichtigsten Normierungstheorien und Vokalklassifikationsalgorithmen vorgestellt. Sie beruhen durchweg auf der ausschließlichen Betrachtung der Formantfrequenzen und ihrer Mittel- oder Extremwerte als Normierungsfaktoren. Man spricht auch von Vokaltraktnormierung ("vowel tract normalization"), da hinter diesen Theorien die Idee steht, daß der Hörer durch die Lage der Formanten Information über die Größe und insbesondere die Länge des Vokaltrakts eines Sprechers erhält und anhand dieser Information die Normierung durchführt. Da zur Ermittlung dieser Werte eine längere zeitliche Mittelwertbildung nötig ist, kann die mittlere Lage der Formanten nicht der wesentliche perzeptive Faktor für die sofortige Einstellung der Vokalschablone sein. Allenfalls können die höheren Formanten F3, F4 und evtl. noch F5, deren Frequenzlagen recht konstant sind, da diese Formanten relativ unabhängig von der Sprechtrakteinstellung sind, als perzeptive Faktoren zur Primärnormierung beitragen.[13]

[13] Siehe z.B. MEYER-EPPLER (1969, S. 280ff) zur Einbeziehung des dritten Formanten.

Obwohl F0 von den Normierungstheorien als perzeptiver Faktor unberücksichtigt bleibt, gab es schon früh experimentelle Befunde, die Vokalqualitätsänderungen nur durch Änderung der Grundfrequenz nachwiesen. MILLER (1953) zeigte mittels Formantensynthese, daß allein durch F0 eine Verschiebung der Phonemgrenzen zwischen den Vokalen in der F1/F2-Ebene zu höheren Frequenzen hin stattfindet. Er benutzte die Grundfrequenzen 140 Hz und 288 Hz. Neben dieser Untersuchung wies auch SLAWSON (1968) nur durch F0 bedingte Vokalqualitätsänderungen nach. Beide Studien zeigen aber, daß die F0-Änderung im Gegensatz zu einer Formantfrequenzänderung quantitativ viel gravierender ausfallen muß, um eine experimentell nachweisbare Vokalqualitätsänderung zu bewirken. SLAWSON zeigte, daß eine Anhebung der Grundfrequenz um 100%, also eine Änderung auf den doppelten Wert, einer Verschiebung der Vokalschablone um nur 10% entspricht.

SUOMI (1984) untersuchte die Güte eines maschinellen Klassifikationsverfahrens für das Geschlecht des Sprechers und die Vokale des Finnischen. Die Vokalklassifikation geschah dabei aufgrund der Bestimmung des kürzesten Abstandes zwischen dem Prototyp jeder Klasse - das entspricht einem Targetpunkt der Vokalschablone - und dem zu klassifizierenden Stimulus. Das Kriterium für den Abstand zweier Vokale ist hier die Ähnlichkeit der Form der spektralen Energieverteilung der Vokale. Durch die Trennung der zu klassifizierenden Vokale nach weiblichen und nach männlichen Sprechern ergab sich eine wesentliche Verbesserung der Klassifikationsrate als bei der gemeinsamen Klassifikation, bei der für jeden Vokal nur ein Prototyp existiert. Eine gute Identifikation des Sprechergeschlechts allein ist ebenfalls möglich. Schlechte Ergebnisse bringt aber die gleichzeitige Klassifikation von Vokalen und Sprechergeschlecht, obwohl dazu zwei separate Prototypen (nach weiblicher und männlicher Sprechergruppe) für jeden Vokal existierten. Das schlechte Klassifikationsergebnis kann auf die Nichtbenutzung der im Stimulus vorhandenen F0-Information zurückgeführt werden. Würde diese Information hinzugezogen, so wäre allein hierdurch eine annähernd fehlerfreie Geschlechtsidentifikation entsprechend dem hier vorgeschlagenen Algorithmus möglich.

Ein anderes Ziel der oben beschriebenen Untersuchung von AINSWORTH (1975) war es, abzuschätzen, ob F0 oder die Formanten selbst für die Verschiebung der Schablone von größerer Bedeutung seien. Zwar wurde angenommen, daß die Formantlagen einen stärkeren Einfluß auf die Normierung haben als F0, jedoch ist dies aus den experimentellen Daten nicht ersichtlich.

Wie bei AINSWORTH stand auch bei der experimentellen Untersuchung von FUJISAKI et al. (1968) die Frage im Mittelpunkt, ob F0 oder die Lage der Formanten den wichtigeren perzeptiven Faktor für den Normierungsprozess darstellt. FUJISAKI et al. betrachteten den Einfluß der höheren Formanten F3 und F4 auf die perzeptive Normierung, während AINSWORTH die Einstellung der Schablone anhand von F1 und F2 durch die Rahmenvokale /i/, /a/ und /u/ betrachtete. Neben F0 sind auch die höheren Formanten zu den intrinsischen Faktoren zu rechnen. Während AINSWORTH die Formanten als extrinsische Faktoren untersuchte, betrachten FUJISAKI et al. die Formanten als intrinsische Faktoren. FUJISAKI et al. generierten synthetische, vierformantige Vokale im Bereich von /o/ bis /a/ und im Bereich von /u/ bis /e/ mit F0 und Formantwerten japanischer Sprecher.[14] Im Perzeptionstest, der mit japanischen Muttersprachlern durchgeführt wurde, wurden die Phonemgrenzen zwischen /o/ und /a/ und zwischen /u/ und /e/ ermittelt und insbesondere ihre Abhängigkeit von F0 und den höheren Formanten untersucht. Obwohl FUJISAKI et al. dies nicht so interpretierten, zeigen die Versuchsergebnisse deutlich, daß F0 allein schon zu starken Verschiebungen der Phonemgrenzen in der Lage ist, während die höheren Formanten keinen so großen Einfluß auf die Phonemgrenzen haben. Dort werden Faktoren definiert, die die relative Verschiebung der Phonemgrenzen dF1b/F1b bezogen auf die relative Verschiebung der höheren Formanten dF3/F3 bzw. auf die relative Verschiebung der Grundfrequenz

[14] FUJISAKI et al. (1968, S. 74) benutzten feste, lineare Zusammenhänge zwischen F1 und F2 jeweils für den Übergang zwischen den Vokalen /e/ und /u/ und zwischen den Vokalen /o/ und /a/.

dF0/F0 angeben.[15] Eine vollständige Verschiebung der perzeptiven Vokalschablone im Sinne der Umstellung von weiblichen auf männliche Sprecher liegt bei (dF1b/F1b)/(dF3/F3) = 1 vor. Diesem Wert entspricht der Fall, daß sich höhere und tiefere Formanten relativ gesehen um die gleichen Beträge verschieben. Jedoch ist eine vollständige Verschiebung der Vokalschablone bei der Betrachtung des auf die Grundfrequenz bezogenen Faktors (dF1b/F1b)/(dF0/F0) schon bei Werten von 0.15 bis 0.2 gegeben, weil beim Vergleich weiblicher und männlicher Sprecher die Grundfrequenz wesentlich stärker ansteigt als die Formantfrequenzen. Während F0 um 100% ansteigt, erhöhen sich die Formanten um 15% bis 20%. FUJISAKI et al. ermittelten für die auf F0 bezogenen Anstiegsfaktoren den Wert von 0.22 für die Verschiebung der /u/-/e/-Grenze und einen Wert von 0.29 für die Verschiebung der /o/-/a/-Grenze anhand der Änderung von F0. Diese Werte sind also so zu interpretieren, daß hier eine viel stärkere Verschiebung der perzeptiven Vokalschablone vorliegt als z.B. bei der Verschiebung der /u/-/e/-Grenze anhand der höheren Formanten mit dem auf F3 bezogenen Faktor von 0.63. Die Verschiebung der /o/-/a/-Grenze anhand der höheren Formanten ergab sogar nur einen Faktor von 0.13.

Diese Resultate sind auch direkt aus den Meßergebnisse ersichtlich.[16] Man kann für die F0-Erhöhung von 150 Hz auf 300 Hz eine Verschiebung der Phonemgrenzen /u/-/e/ von 560 Hz nach 630 Hz und für /o/-/a/ von 630 Hz nach 800 Hz ablesen.[17] Dies entspricht einer Verschiebung der Grenzen um 13% bis 27% und deutet an, daß entgegen der Schlußfolgerung der Autoren sehr wohl F0 allein in der Lage ist, die perzeptive Verschiebung der Phonemgrenzen in der durch die akustischen Signalunterschiede weiblicher und männlicher Sprecher auch perzeptiv zu fordernden

[15] Siehe FUJISAKI et al. (1968, S. 76, Tab. 1).

[16] Siehe FUJISAKI et al. (1968, S. 76, Fig. 6 und Fig. 7).

[17] Grundfrequenzen im Bereich von 300 Hz sind auch für weibliche Sprecher sehr hoch. Dies erklärt die hohen Frequenzwerte, die bei der Verschiebung der Phonemgrenzen für F1 auftreten.

Größenordnung zu bewirken. Die Verschiebung der Phonemgrenzen des ersten Formanten durch die Grundfrequenz nach den Experimenten von FUJISAKI und KAWASHIMA sind in Abbildung 3.2 als Funktion der Grundfrequenz wiedergegeben. Zum Vergleich wurde die Verschiebung des ersten Formanten des Schwa-Lautes - und die Parallelen dazu im Abstand von 100 Hz - beim Übergang von männlichen zu weiblichen Sprechern nach Durchschnittswerten aus der signalanalytischen Arbeit von PETERSON et al. (1952) eingetragen. Dabei wurde eine mittlere Grundfrequenz von F0(männl.) = 130 Hz und F0(weibl.) = 220 Hz und eine Formantfrequenzerhöhung um rund 17% von 500 Hz auf 585 Hz angesetzt. Abbildung 3.2 zeigt, daß nach den Messungen von FUJISAKI et al. im Fall der /o/-/a/-Phonemgrenze der Signalfaktor F0 allein zu der erforderlichen Verschiebung der Vokalschablone in der Lage ist. Diese Steigung fällt sogar größer aus als die nach den signalanalytischen Durchschnittswerten geforderte.

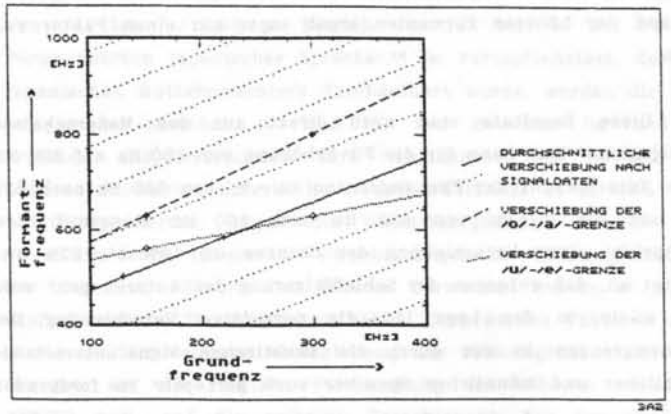


Abb. 3.2: Kopplung zwischen Grundfrequenz und Lage des ersten Formanten des Schwa-Lautes nach Signaldaten (PETERSON et al. (1952)) und gemessene Werte für die Verschiebung der Phonemgrenzen nach FUJISAKI et al. (1968)

Der Grund für die Fehlinterpretation der Versuchsergebnisse von FUJISAKI et al. durch die Autoren selbst liegt also in der fehlenden Berücksichtigung der Tatsache, daß der natürliche geschlechtsspezifische Signalunterschied bezüglich der Grundfrequenz viel größer als der Unterschied bezüglich der Mittelwerte der Formantfrequenzen ist. Genau deswegen treten auch nur geringe Vokalqualitätsänderungen bei starken Grundfrequenzänderungen auf.[18] Es ist aber falsch, aus diesen nur geringen Vokalqualitätsänderungen zu schließen, daß die Grundfrequenz nur untergeordnete Bedeutung als perzeptiv relevantes Signalmerkmal für die Vokalqualität hat, denn in natürlichen Sprachsignalen treten ja entsprechend große Variationen der Grundfrequenz über einen Bereich von 80-400 Hz auf.

3.3.2.3 Interindividuelle Unterschiede in der Perzeption synthetischer Vokale

Der folgende experimentelle Befund kann durch die Offenheit synthetischer Stimuli bezüglich der Geschlechtsassoziation erklärt werden. AINSWORTH (1976, S. 72) berichtet von erheblichen individuellen Variationen in der perzeptiven Verschiebung der Vokalschablone, die anhand der Klassifikation synthetischer Vokale festgestellt wurden. Für einige Hörer ist die perzeptive Verschiebung vernachlässigbar, während sie für andere Hörer größer ausfällt als die sich aus den akustischen Signalwerten ergebende Verschiebung der Schablone bei der Perzeption weiblicher und männlicher Sprecher. Nach der in diesem Kapitel entwickelten Theorie deutet das darauf hin, daß auch hier die Stimuli trotz ihrer typisch weiblichen F0- und Formantwerte wegen der Außerachtlassung der Stimmqualität nicht überzeugend weiblich klangen. So kann es sein, daß ein Teil der Hörer mit demselben Stimulus weibliches, der andere Teil männliches Sprechergeschlecht assoziierte.

---

[18] Siehe SLAWSON (1968) und AINSWORTH (1975).

3.3.2.4 Das Versagen extrinsischer Normierungsfaktoren bei der Perzeption natürlicher Sprache

Gerade in Hinblick auf die perzeptive Normierung zeigen sich in der Literatur systematische Unterschiede zwischen natürlicher und synthetischer Sprache. Nur bei synthetischer Sprache wurde gezeigt, daß ein Verschieben der Vokalschablone durch extrinsische Faktoren möglich ist, während ein Verschieben anhand natürlicher Sprache nicht nachgewiesen werden konnte.

LADEFOGED et al. (1957)[19] ließen die Vokalrealisationen der Silbe [bVt] klassifizieren. Dabei wurden zwei Testvarianten durchgeführt. In der ersten Variante wurde der Silbe mit dem zu klassifizierenden Vokal der Satz "Please say what this word is:" vorangestellt, während in der zweiten Variante die Vokale ohne vorlaufenden Satz zu klassifizieren waren. Die Stimuli, d.h. sowohl der Satz als auch die Silbe, wurden mittels Sprachsynthese erzeugt. Es wurden vier Formanten realisiert und die Anregung des Sprechtraktmoduls geschah mittels eines Impulsgenerators. Es wird konstatiert, daß bei weitem kein natürlicher Sprachklang erreicht wurde. Das Ergebnis der Untersuchung war, daß signalidentische Vokalrealisationen der Silbe [bVt] je nach Variation der Formantfrequenzlagen des vorlaufenden Satzes, die in der Größenordnung der Formantdifferenzen weiblicher und männlicher Sprecher lagen, unterschiedlich perzipiert wurden. Es wurde bei der Auswahl des vorlaufenden Satzes darauf geachtet, daß dieser bezüglich F1 und F2 den gesamten Vokalraum überstreicht. F1 und F2 variierten von ihren tiefsten bis zu ihren höchsten Frequenzen (realisiert im /i/, /a/ und /w/ des Amerikanischen). Dies Ergebnis konnte dadurch erklärt werden, daß sich durch den vorlaufenden Satz der perzeptive Vokalraum des Hörers nach dem dargebotenen akustischen Vokalraum des Sprechers justiert. Unter der Voraussetzung, daß der perzeptive Vokalraum einer gewissen Trägheit unterworfen ist, wird der zu klassifizierende Stimulus in der durch den vorlaufenden Satz gegebenen Vokalraumeinstellung

[19] Siehe auch LADEFOGED (1967).

perzipiert. Die Unterschiede in der Vokalqualität eines identischen Stimulus ergeben sich also aus der unterschiedlichen Einstellung der Vokalschablone des Hörers, bedingt durch den vorlaufenden Satz.

Ähnliche Experimente mit synthetischen Stimuli wurden von THOMPSON et al. (1970) durchgeführt. Sie reduzierten allerdings den vorlaufenden Satz auf die Darbietung eines Vokals. Die Hörer mußten von zwei in kurzem zeitlichen Abstand aufeinander folgenden Vokalen jeweils den zweiten klassifizieren. THOPSON et al. sprechen hierbei von beeinflussendem Vokal ("affecting vowel") und beeinflußtem Vokal ("affected vowel"), womit der erste und zweite Vokal gemeint sind. Auch hier wurde eine Beeinflussung der Perzeption durch extrinsische Faktoren, also durch außerhalb des Vokals liegende Signalfaktoren festgestellt. Als extrinsische Faktoren wurden auch hier die Formanten des vorlaufenden Vokals angenommen.

Im Gegensatz dazu konnte dieser perzeptive Effekt anhand natürlicher Sprache nicht nachgewiesen werden. DECHOVITZ (1977) wiederholte das oben beschriebene Experiment von LADEFOGED et al. (1957) mit natürlicher Sprache. Dazu wurde der vorlaufende Satz und die Silbe [bVt] mit dem zu klassifizierenden Vokal [V] von zwei männlichen Sprechern mit stark unterschiedlichen Vokaltraktdimensionen gesprochen. Die Teststimuli wurden aber in jeder Kombination zusammengeschnitten, d.h. es wurde auch der vorlaufende Satz von Sprecher 1 zusammen mit der zu klassifizierenden Silbe [bVt] von Sprecher 2 dargeboten. Da bei der Aufnahme der Sätze darauf geachtet wurde, daß beide Sprecher mit gleicher Tonhöhe und gleichem Sprechtempo sprachen, schließt DECHOVITZ, daß die Hörer nicht merken, daß ein Stimulus evtl. aus von verschiedenen Sprechern gesprochenen Teilen besteht. Jedoch ergab sich keine perzeptive Verschiebung aufgrund der Formantlageänderung im vorlaufenden Satz. Das gleiche Experiment mit größeren Formantvariationen, wozu ein Kind und ein Mann als Sprecher herangezogen wurden, fiel ebenfalls negativ aus. Nach ähnlichen,

von LADEFOGED[20] durchgeführten Experimenten mit synthetischer Sprache hoher Qualität konnte ebenfalls keine Beeinflussung der Vokalqualität durch extrinsische Faktoren nachgewiesen werden.

Entgegen der Auffassung von DECHOVITZ ist aber anzunehmen, daß aufgrund der Vielschichtigkeit der Stimmqualität (siehe Kap. 2) sehr wohl eine Sprecheridentifikation auch anhand nur einer Silbe möglich ist. Stimmqualität wird ja neben Tonhöhe und Sprechtempo noch von vielen anderen Faktoren, beispielsweise dem Stimmklang, beeinflußt. So ist bei natürlich gesprochenen Stimuli die Änderung der Stimmqualität unvermeidbar, nicht jedoch in synthetisch erzeugten. Die oben aufgeführten experimentellen Ergebnisse können dahingehend interpretiert werden, daß synthetisch erzeugte Sprache eben nicht die starken perzeptiven Faktoren enthält, die zur Primärnormierung, d.h. zur sofortigen Umstellung der Vokalschablone führen. Insbesondere geben hier die bestehenden Grenzen der Sprachsynthese, also das Problem der Erzeugung eines natürlichen Stimmklangs und die Tatsache der noch nicht möglichen Differenzierung zwischen verschiedenen idiosynkratischen Stimmqualitätsmerkmalen, einen Hinweis darauf, daß gerade die mit diesen, im Signal nicht vorhandenen Komponenten gekoppelten perzeptiven Faktoren, zur Schablonenverschiebung führen. Dies belegt die These, daß eine Änderung der Schablonenlage um große Abstände immer an eine Wahrnehmung von Stimmqualitätsänderung gebunden ist. Da synthetische Sprache weitgehend frei von idiosynkratischen Stimmqualitätsmerkmalen ist, kann die Verschiebung der Vokalschablone durch den Rahmensatz, also durch extrinsische Faktoren nachgewiesen werden, da die daraus resultierenden Vokalqualitätsänderungen gerade durch ein Nichtverschieben der perzeptiven Vokalschablone beim zu klassifizierenden Stimulus zu erklären sind.

Auf der anderen Seite soll hier ein Einfluß extrinsischer Faktoren auf den gesamten perzeptiven Normierungsprozeß gar nicht bestritten werden, jedoch spielen diese Faktoren beim Prozeß der

[20] Private Mitteilung an PAPÇUN, siehe PAPÇUN (1980, S. 7).

Primärnormierung keine Rolle. Es wurde nachgewiesen, daß Realisierungen von Vokalen durch nur einen Sprecher mit geringerer Fehlerrate klassifiziert werden als von verschiedenen Sprechern gesprochene, in kurzen zeitlichen Abständen aufeinander folgende Vokale.[21] Auch das Sprechtempo des Rahmensatzes hat einen systematischen Einfluß auf die Perzeption von Vokalen.[22] Jedoch treten diese Effekte vor allem bei einem kontinuierlichen und längeren Wahrnehmen eines Sprechers auf. Hier geht es um die Untersuchung einer anderen Kommunikationssituation. Es wurde ein Modell für die Perzeption der ersten Silbe eines neuen, evtl. unbekannten Sprechers entwickelt. Insbesondere wird der Fall der Umstellung von einem weiblichen auf einen männlichen Sprecher untersucht. Dies ist ein Fall, in dem normalerweise große interindividuelle Signalunterschiede vorliegen, die vom Hörer aber sofort und ohne Schwierigkeiten verkraftet werden. Also kann – auch durch die Definition der Aufgabenstellung, aufgrund deren auch der Begriff Primärnormierung als ein grobes erstes Einstellen auf einen neuen Sprecher anderen Geschlechts zu verstehen ist – die Primärnormierung nur an intrinsische Signalparameter gebunden sein.

3.4 Der Vergleich mit anderen Normierungstheorien

3.4.1 Zur perzeptiven Relevanz des Modells der Primärnormierung

Bezüglich der perzeptiven Normierung sind verschiedene Modellvorstellungen entwickelt worden. Im folgenden werden einige Befunde aus Perzeptionsexperimenten referiert, die die Hypothese der Existenz der Primärnormierung verifizieren. Es wird gezeigt, daß die in Echtzeit ablaufende Primärnormierung eine realistische Beschreibung der Umstellung der Perzeption von weiblichen auf männliche Sprecher und umgekehrt ist.

Auf JOOS (1948) geht die Vorstellung zurück, daß zur

[21] Siehe z.B. VERBRUGGE et al. (1976).

[22] Siehe AINSWORTH (1974).

sicheren Identifikation der Vokale ein Einhörvorgang notwendig ist. Dabei wird die besondere Bedeutung der den Vokalraum begrenzenden Eckvokale ("point vowels") /i/, /a/ und /u/ für den Einhörprozeß unterstrichen. LIEBERMAN (1973) betont, daß diese besondere Rolle der Eckvokale mehrere Gründe hat. Neben ihrer Bedeutung als Begrenzer des Vokalraumes in akustischer Hinsicht (minimale und maximale Formantwerte) und in artikulatorischer Hinsicht (extreme Artikulatorpositionen ohne Friktionsbildung) stellt LIEBERMAN heraus, daß die Eckvokale die einzigen Vokale sind, deren akustische Daten eine eindeutige Rücktransformation in eine Areafunktion des Sprechtraktes zulassen, während die akustische Struktur anderer Vokale durch verschiedene Gruppen von Areafunktionen erreicht werden kann.[23] Also kann der Hörer nur anhand dieser Vokale eindeutige Rückschlüsse auf die Vokaltraktgröße ziehen, womit die Sonderstellung dieser Vokale für den perzeptiven Normierungsprozeß gegeben wäre. Allerdings beweisen VERBRUGGE et al. (1976) in einem Klassifikationsexperiment anhand natürlich gesprochener Einsilber, daß als Rahmenvokale[24] dargebotene Eckvokale desselben Sprechers die Identifikationsrate nicht erhöhen. Es wird geschlossen, daß die extrinsischen Parameter der Eckvokale nicht die zur Normierung des Vokalraumes führenden perzeptiven Faktoren sind.

Eine andere Gruppe von Forschern betont, daß die Einbettung des Vokals in eine Konsonantenumgebung eine wesentliche Voraussetzung für das Funktionieren der perzeptiven Normierung darstellt. So ermittelten STRANGE et al. (1976) ein sehr schlechtes Klassifikationsergebnis für isoliert gesprochene Vokale, während die Erkennung durch die Darbietung einer [CVC]-Silbe schon wesentlich gesteigert wird. Das Stimulusmaterial zu dieser Untersuchung bestand aus natürlich gesprochenen Vokalen von fünf Frauen, fünf Männern und von fünf Kindern. In einer späteren Untersuchung von STRANGE et al. (1983) wird sogar versucht, eine Vokalerkennung allein anhand der dynamischen Merkmale, der Formanttransitionen, nachzuweisen.

[23] Siehe auch STEVENS (1969).

[24] Siehe Abs. 3.3.2.1.

VERBRUGGE et al. (1976) lassen von Männern, Frauen und Kindern gesprochene [hVd]- und [pVd]-Silben klassifizieren, indem alle Stimuli aller Sprecher in einem Perzeptionstest gemischt nacheinander dargeboten werden ("randomized-talker-condition"). Bereits hier ergibt sich ein gutes Klassifikationsergebnis, das allerdings noch etwas gesteigert werden kann, wenn nur die Aufgabe besteht, die Vokale eines Sprechers zu klassifizieren ("isolated-talker-condition"). Auch VERBRUGGE et al. lassen also keine isolierten Vokale, sondern nur [CVC]-Silben klassifizieren, und es werden hohe Erkennungsraten dieser Einsilber auch dann erreicht, wenn die Hörer Vokale von ständig wechselnden Sprechern klassifizieren müssen.

Ein ähnliches Experiment mit natürlich gesprochenen Stimuli von fünf Frauen und von fünf Männern wurde von ASSMANN et al. (1982) durchgeführt und führte zum gleichen Resultat. Damit ist erwiesen, daß eine einzige Silbe eines beliebigen Sprechers genug Information enthält, um die perzeptive Normierung zu bewirken. Eine erste, ungefähre Normierung geschieht also ohne Einhörprozeß, d.h. in Echtzeit. Damit ist die Existenz einer Primärnormierung auf perzeptiver Ebene bewiesen.

Da bei oberflächlicher Betrachtung die Steigerung der Erkennungsrate anhand dynamischer Signalmerkmale im Widerspruch zur Schablonentheorie steht, wird von einigen Verfechtern der Schablonentheorie auch bestritten, daß die dynamischen Parameter wie z.B. die Formanttransitionen wesentliche perzeptive Faktoren für den Normierungsprozeß sind. Hierzu existieren Arbeiten, die vor allem das schlechte Klassifikationsergebnis von STRANGE et al. (1976) für die isoliert gesprochenen Vokale in Frage stellen. In einer von KAHN (1978) durchgeführten Untersuchung müssen die im [hV]-Kontext dargebotenen Vokale zweier weiblicher und zweier männlicher Sprecher klassifiziert werden. Da der glottale Frikativ /h/ ein durch Koartikulation vom nachfolgenden Vokal stark beeinflußte Sprechtrakteinstellung aufweisen kann, liegen hier nur geringe Formanttransitionen vor. Trotzdem ergab sich eine

Erkennungsrate von 97% bei gemischter Darbietung von Vokalrealisationen aller Sprecher. In einem Experiment von MACCHI (1980) wurde neben dem Klassifikationsexperiment mit [CVC]-Silben auch ein Experiment mit isoliert gesprochenen Vokalen durchgeführt. Die Aufnahme der Stimuli wurde unter Hochqualitätsbedingungen gemacht, und im Gegensatz zu STRANGE et al. (1976) wird hier eine hohe Erkennungsrate auch für isolierte Vokale erreicht. In einem Klassifikationsexperiment von DIEHL et al. (1981) mit synthetisch erzeugten und natürlich gesprochenen Vokalen in Isolation und im [CVC]-Kontext wurde kein Unterschied zwischen der Identifikation isolierter Vokale und im Kontext eingebundener Vokale festgestellt. DIEHL et al. stellen in Frage, daß speziell die Formanttransitionen eine wesentliche Rolle für den Normierungsprozeß spielen.

Letztlich spricht aber auch die Wichtigkeit dynamischer Signalparameter nicht gegen die Primärnormierung oder gegen die Schablonentheorie und auch nicht gegen die Verbindung von Primärnormierung und Geschlechtsassoziation.[25] Es ist einerseits möglich, die Formanttransitionen eines Vokals in die benachbarten Konsonanten als perzeptive Normierungsfaktoren[26] zu betrachten. Da Formanttransitionen wesentliche akustische Korrelate der artikulatorischen Gesten eines Sprechers sind, ist es denkbar, daß sie somit ebenfalls wesentliche Stimmqualitätsinformationen enthalten und dadurch zum Verschieben der Vokalschablone beitragen. Andererseits können die Formanttransitionen auch als perzeptive Klassifikationsfaktoren[27] betrachtet werden. Sie geben dann Aufschluß über die Targetwerte des zu klassifizierenden Vokals. So betrachtet stellen sie ebenfalls keinen Widerspruch zur Schablonentheorie dar, da diese Theorie völlig offen läßt, wie der Perzeptionsapparat zur Extraktion der Targetpositionen des perzipierten Vokals gelangt.[28]

---

[25] Siehe Abs. 3.4.4.

[26] Zur Definition siehe Abs 3.2.3.

[27] Zur Definition siehe Abs. 3.2.3.

[28] Siehe Abs 3.4.4.

Die im Rahmen dieser Arbeit durchgeführten Experimente zur Vokalqualität werden auch schon deshalb mit Diphthongen durchgeführt, weil nachgewiesen wurde, daß die zeitliche Variation des Signals gerade bei synthetischen Stimuli wesentlich dafür ist, daß das Signal auch wirklich als Sprachsignal erkannt und entsprechend verarbeitet wird.

3.4.2 Normierungsalgorithmen

Nach der hier entwickelten Vorstellung der perzeptiven Primärnormierung aufgrund der Geschlechtsassoziation läßt sich, wie oben ausgeführt, ein einfacher Normierungsalgorithmus entwickeln. Anhand der Grundfrequenz wird eine Geschlechtsklassifikation durchgeführt und damit die Einstellung der Vokalschablone als Grundlage für die danach folgende Vokalklassifikation vorgenommen. Es wird dabei die Existenz einer bereits vorab gelernten Schablone vorausgesetzt.

Dieser Ansatz wird durch zwei wesentliche Merkmale charakterisiert.

1) Es werden neben den Formantfrequenzen auch andere Signalmerkmale als perzeptive Normierungsfaktoren mit einbezogen. Insbesondere vollzieht sich hier die Normierung selbst anhand von F0.

2) Es ist keine Einlernphase für die Einstellung auf einen Sprecher nötig. Die Targetpunkte der Vokale der Schablone sind bereits vorab bekannt. Die Verschiebung der Schablone bei der Einstellung auf einen bestimmten Sprecher geschieht nicht anhand einiger vorab dargebotener Normierungsvokale. Ein solcher Algorithmus kann prinzipiell auch die Einstellung auf einen Sprecher in Echtzeit vornehmen.

Jedoch erfüllen die bekannten Normierungsalgorithmen diese beiden Bedingungen nicht gleichzeitig, obwohl sie hohe Erkennungsraten erzielen. Jedoch erfolgt aus den guten Klassifika-

tionsergebnissen, die ein Normierungsalgorithmus erbringt, nicht unbedingt, daß die perzeptive Normierung ähnlich ablaufen müßte.

Eine Einlernphase zur Einstellung auf einen Sprecher benötigen die Algorithmen von GERSTMAN (1968) und von LOBANOV (1971), denen trotz unterschiedlicher Berechnung der normierten Formantfrequenzen die gleiche Idee zugrunde liegt. Zunächst müssen viele Vokalrealisationen eines Sprechers vorliegen, anhand derer die maximalen und minimalen Frequenzwerte des ersten und zweiten Formanten dieses Sprechers ermittelt werden. Diese Werte bilden die Grundlage für die Transformation der Formantfrequenzen des Sprechers in einen normierten Vokalraum. Es wird hier die Idee der Normierung anhand von Eckvokalen nach JOOS (1948) erkennbar.

Die im weiteren vorgestellten Algorithmen arbeiten zwar ohne Einlernphase nur anhand intrinsischer Normierungsfaktoren, allerdings benutzen auch diese Algorithmen nicht die in der Grundfrequenz enthaltenen Information. LADEFOGED et al. (1978) bestimmen artikulatorische Daten anhand einer festen, für alle Sprecher geltenden Formel, in die nur die Formantfrequenzverhältnisse von F1, F2 und F3 eingehen. Die resultierenden Daten bilden einen normierten Vokalraum, allerdings in artikulatorisch orientierter Realisation. Im Vordergrund der Arbeit von LADEFOGED et al. steht allerdings das Problem der Bestimmung von Sprechtraktgeometrien aus akustischen Signalparametern. Die Idee der Extraktion artikulatorischer Parameter aus dem Sprachsignal benutzt auch WAKITA (1977), der Vokalnormierung aufgrund der Bestimmung der individuellen Vokaltraktlänge aus den höheren Formanten vornimmt. Bei dem von SUOMI (1984) vorgestellten Algorithmus wird die Vokalklassifikation zusammen mit der Geschlechtsklassifikation durchgeführt. Dieser Algorithmus kommt den in diesem Kapitel entwickelten Ideen am nächsten. Jedoch fallen die Klassifikationsergebnisse dürftig aus, da hier nur die Signalinformation der spektralen Energieverteilung benutzt wird. Wie in Absatz 3.3 angedeutet, ist ein besseres Funktionieren zu erwarten, wenn auch F0-Information hinzugezogen wird.

3.4.3 Nichtproportionale Formantskalierung

FANT analysierte die Formantdaten von Vokalen, die von Männern, Frauen und Kindern acht verschiedener Sprachgruppen realisiert wurden.[29] Er stellte universelle Tendenzen in der Abweichung der Formantfrequenzen von einer simplen Proportionalverschiebung besonders beim Vergleich weiblicher und männlicher Sprecher fest.[30] Dazu definierte Fant den k-Faktor, der die relative Frequenzänderung eines Formanten i beim Vergleich weiblicher und männlicher Sprecher für jeden Vokal angibt.

(3.1) $$k_i = dF_i/F_i = (F_i(\text{weibl.}) - F_i(\text{männl.}))/F_i(\text{männl.})$$

mit i=1,2,3 Nr. des Formanten

Im Mittel beträgt der k-Faktor für die drei ersten Formanten aller Vokale 0.17, aber Fant stellt generelle sprachübergreifende Tendenzen in der Abweichung des k-Faktors von diesem Mittelwert in Abhängigkeit vom Vokal und von den Formanten selbst fest. FANT versucht diese Tendenzen artikulatorisch zu begründen. Teils kann die akustische Nichtproportionalität darauf zurückgeführt werden, daß Frauen im Vergleich zu Männern eine überproportional kürzere Pharynx aufweisen.[31] Ein anderer Teil dieser akustischen Nichtproportionalität kann nur durch die Annahme von geschlechtsspezifischen Unterschieden in der Artikulation erklärt werden.

Allerdings konnte NORDSTRÖM (1975) in Syntheseexperimenten mit nichtproportionaler artikulatorischer Skalierung die zu erwartenden nichtproportionalen akustischen Formantfrequenz-

---

[29] Siehe FANT (1975) und FANT (1973, S. 84-93).

[30] Geometrisch interpretiert ist diese Feststellung von FANT äquivalent mit der Aussage, daß in der doppeltlogarithmischen F1/F2-Ebene geschlechtsspezifische Unterschiede in der *Form* der Vokalschablone vorliegen.

[31] Neuere Untersuchungen zur Pharynxgeometrie wurden von SUNDBERG et al. (1987) vorgenommen.

variationen nicht modellieren. Ferner ist der experimentelle Nachweis dieser quantitativen Modifikationen in Perzeptionstests nicht möglich, da diese Modifikationen eine genaue quantitative Ermittlung der perzeptiven Verschiebungsweite erfordern. Im Experiment dieser Arbeit konnte nur die Richtung der Verschiebung der Schablone, nicht aber der Betrag der Verschiebung ermittelt werden. Allerdings ist dies zur Überprüfung der Hypothese, daß Geschlechtsassoziation ein Verschieben der Schablone in bestimmte Richtungen bewirkt, auch nicht nötig.

Entsprechend wird hier auch nur eine einfache quantitative Realisierung der Schablonentheorie als Modell benutzt. Dieses Modell, von NEAREY (1977, S. 89) mit CRH ("Constant Ratio Hypothesis") und hier im folgenden mit *linearer Näherung* bezeichnet, geht von der Vorstellung aus, daß eine Verschiebung der Schablone nur entlang konstanter Frequenzverhältnisse (F1/F2 = const.) erfolgt, und daß die Frequenzverhältnisse der Schablonentargets ebenfalls konstant bleiben. In der F1/F2-Ebene mit logarithmischer Frequenzskalierung entspricht dies dem Bild der Verschiebung der Vokalschablone entlang von Parallelen zur logF1/log F2=1-Geraden ohne Verzerrung der Form der Schablone (siehe Abb. 3.3).[32] Artikulatorisch entspricht dies einer proportionalen oder linearen Längenänderung des Sprechtraktes bei Invarianz der Sprechtraktform für alle Vokale. Abbildung 3.3, in der eine streng proportionale Verschiebung in der Form des Vokaldreiecks (lineare Näherung) und die nach PETERSON et al. (1952) gemessenen Formantwerte der Vokale /u/, /a/ und /i/ (reale Werte) abgetragen wurden, macht deutlich, daß die lineare Näherung die realen Werte bereits gut wiedergibt.

[32] Die Forminvarianz der Vokalschablone rührt gerade daher, daß ein Frequenzverhältnis oder auch Tonhöhenintervall in der logarithmischen Skalierung durch eine Distanz konstanter Länge definiert ist.

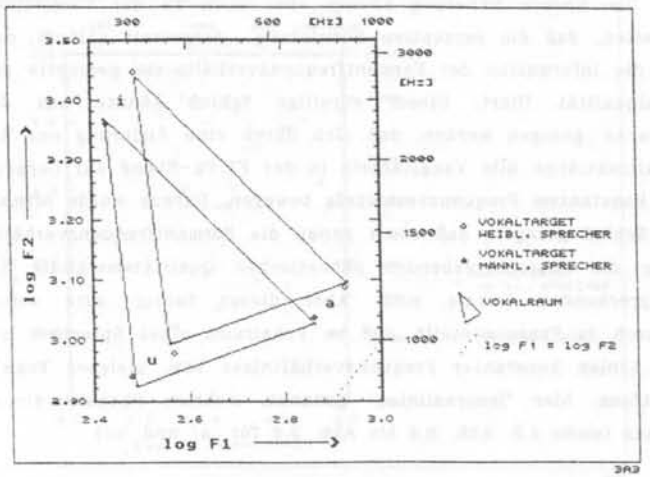


Abb. 3.3: Der Vokalraum weiblicher und männlicher Sprecher in linearer Näherung. Daten nach PETERSON et al.(1952)

Da in den in dieser Arbeit durchgeführten Experimenten nur die Richtung der Verschiebung der Schablone nachgewiesen werden soll, ist es ausreichend, die Testergebnisse in dieser linearen Näherung zu interpretieren. Entsprechend wurde auch zur Synthese der weiblichen Stimme nur eine proportionale Verkürzung des Sprechtraktes um 17% ausgehend von den Werten der männlichen Stimme vorgenommen. Dabei blieben die relativen Sprechtraktgeometrien erhalten.

Wie oben ausgeführt, kann die Form der Vokalschablone in der doppelt-logarithmischen F1/F2-Ebene in linearer Nährung als gleichbleibend angesehen werden. Hieraus leitet sich ein großer Vorteil des Modells der verschiebbaren Schablone ab. Nach diesem Modell reicht es für die Perzeption unterschiedlicher Sprecher einer Dialektgruppe aus, daß das menschliche Perzeptionssystem zur Klassifikation der Vokale über nur eine gelernte Vokalschablone verfügt.

Die lineare Näherung könnte aber auch zu der Vorstellung verleiten, daß die perzeptive Normierung dergestalt abläuft, daß nur die Information der Formantfrequenzverhältnisse perzeptiv zur Vokalqualität führt. Dieser voreilige Schluß könnte aus der Tatsache gezogen werden, daß sich durch eine Änderung nur der Vokaltraktlänge alle Vokaltargets in der F1/F2-Ebene auf Geraden mit konstantem Frequenzverhältnis bewegen. Daraus wurde oftmals der Schluß gezogen, daß somit genau die Formantfrequenzverhältnisse die ausschlaggebenden phonetischen Qualitätsmerkmale der entsprechenden Vokale sind. Aber dieser Schluß wird sofort dadurch in Frage gestellt, daß im Vokalraum eines Sprechers auf den Linien konstanter Frequenzverhältnisse bzw. gleicher Vokaltraktform, hier "Isoarealinien" genannt, mehrere Phoneme liegen können (siehe z.B. Abb. 3.4 bis Abb. 3.6 für /a/ und /o/).

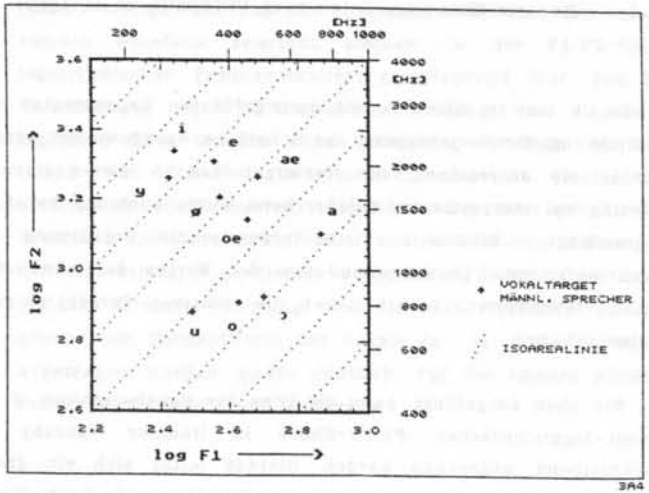


Abb. 3.4: Die Vokalschablone männlicher Sprecher mit Isoarealinien. Daten gemittelt über mehrere Sprecher mehrerer Sprachen nach FANT (1975)

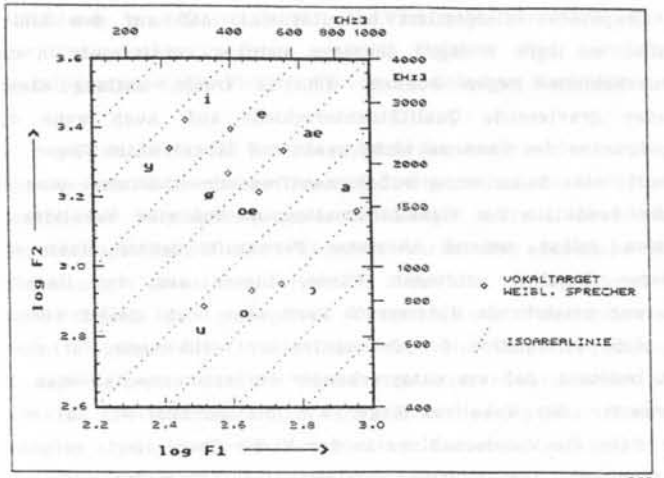


Abb. 3.5: Die Vokalschablone weiblicher Sprecher mit Isoarealinien. Daten gemittelt über mehrere Sprecher mehrerer Sprachen nach FANT (1975)

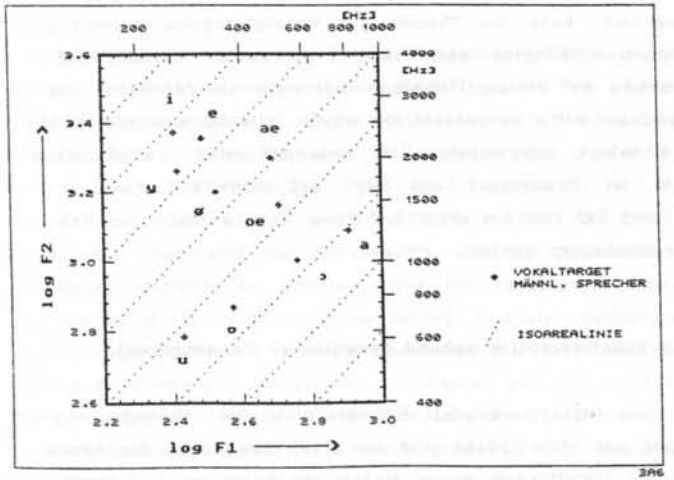


Abb. 3.6: Die Vokalschablone eines männlichen Sprechers mit Isoarealinien. Daten eines Sprechers des Hochdeutschen nach JOERGENSEN (1969)

Graphisch interpretiert bedeutet das, daß auf den Linien parallel zu logF1 = logF2 durchaus mehrere Vokaltargets in der Vokalschablone liegen können, d.h. es treten entlang dieser Geraden gravierende Qualitätsunterschiede auf. Auch wenn die Targetpunkte der Phoneme nicht exakt auf Isoarealinien liegen, so bedeutet eine Reduzierung auf Formantfrequenzverhältnisse eine zu starke Reduktion der Signalinformation, so daß eine Vokalklassifikation selbst anhand absoluter Formantfrequenzen dann ein besseres Resultat erbringen würde. Dieser aus der linearen Näherung entstehende Widerspruch kann aber leicht gelöst werden. Die bloße Information der Formantfrequenzverhältnisse für einen Laut bedeutet, daß ein entsprechender Perzeptionsmechanismus die Information der Vokaltraktlänge, d.h. die Information, auf welcher Höhe die Vokalschablone in der F1/F2-Ebene liegt, vollständig ignoriert. Die richtige Vorstellung des Perzeptionsprozesses auch in der linearen Näherung ist aber die, daß die Lage der Schablone, d.h. die mittlere Sprechtraktlänge auch auf der Verarbeitungsebene der Signalparameter zur Vokalqualität bekannt ist und unter Berücksichtigung dieser Information die absoluten Formantfrequenzen normiert werden. Aus einem anderen Blickwinkel betrachtet, kann die Theorie der Vokalperzeption anhand bloßer Frequenzverhältnisse auch dadurch verworfen werden, daß eine Reduktion auf Formantfrequenzverhältnisse die Dimension des Vokalraumes auf 1 heruntersetzen würde. Allerdings wurde in Absatz 3.2.4 belegt, daß Vokalqualität eine mindestens zweidimensionale Größe ist. Dementsprechend folgt, daß diese Reduktion zu stark ist, und daß sich bei einer Reduktion in eine Dimension Vokaltargets überlagern werden.

3.4.4 Vokalperzeption anhand dynamischer Signalmerkmale

Das Perzeptionsmodell der verschiebbaren Vokalschablone geht zurück auf JOOS (1948). Auf der einen Seite zeigt die Arbeit von NEAREY (1977), daß dieses Modell bis heute an Aktualität nicht verloren hat. Insbesondere weist NEAREY die Existenz einer perzeptiven Schablone nach. Auf der anderen Seite werden stationäre Formantfrequenzen, durch die die Targetpunkte der Vokalschablone ja definiert sind, als wichtigste perzeptive Faktoren der Vokale aber in Frage gestellt. Vielmehr wird auch für die Vokalperzeption die Wichtigkeit von dynamischen Merkmalen betont.

LINDBLOM (1963) konnte anhand der Analyse des akustischen Signals den Nachweis erbringen, daß eine um so stärkere Zentralisation der Formantfrequenzen der Vokale auftritt, je höher das Sprechtempo ist. Dieses Phänomen wurde als akustische Unterschreitung ("acoustic undershoot") bezeichnet. Entsprechend wurde eine Theorie der perzeptiven Überschreitung ("perceptive overshoot") aufgestellt, die besagt, daß die Targetfrequenzen der Vokale perzeptiv aus den Formanttransitionen extrapoliert werden können. Ein experimenteller Nachweis der perzeptiven Überschreitung in Abhängigkeit von verschiedenen Konsonantumgebungen wurde von WILLIAMS (1987) erbracht. AINSWORTH (1974) untersuchte den Einfluß der Vokaldauer auf die Klassifikation der Vokale. Es ergab sich, daß die relative Vokaldauer, d.h. die Länge eines Vokals in Abhängigkeit vom Sprechtempo, ein wichtiger Faktor der Vokalperzeption ist. So gelang es AINSWORTH, den Parameter Sprechtempo als extrinsischen Normierungsfaktor nachzuweisen.

Allerdings widersprechen diese Versuchsergebnisse in keinem Punkt der Schablonentheorie. Der Begriff perzeptive Überschreitung macht eher deutlich, daß auch bei den dynamischen Theorien die Vorstellung einer Schablone mit Vokaltargets im Vordergrund steht. Denn schließlich sind es die Targetwerte selbst, die unter- bzw. überschritten werden. LINDBLOM (1963, S. 1773) hebt sogar die Bedeutung der Targetwerte als die eigentlichen invarianten Attribute der Vokale hervor. Der Targetwert selbst wird bei LINDBLOM als unabhängig vom Kontext und als unabhängig vom Sprechtempo angesehen. Widersprochen wird nur der Annahme, daß konstante Formantfrequenzen die wichtigsten perzeptiven Klassifikationsfaktoren[33] sind, also daß konstante Formantfrequenzen die entscheidenden perzeptiven Merkmale sind, die zur Identifikation des Vokals mit dem entsprechenden Targetpunkt führen.

---

[33] Zur Definition siehe Abs. 3.2.3.

Sowohl die perzeptive Überschreitung als auch der Einfluß extrinsischer Normierungsparameter beschreiben aber auch eher kontextabhängige Perzeptionsphänomene. Es geht dabei nicht um den Prozeß der Primärnormierung, sondern eher um intraindividuelle als um interindividuelle Variationen. Aber auch diese Phänomene stellen, wie oben bereits ausgeführt wurde, keinen Widerspruch zur Schablonentheorie dar, sondern sind als Ergänzungen dieser Theorie aufzufassen.

II. EXPERIMENTELLER TEIL:

ZUR SYNTHESE DER WEIBLICHEN STIMME UND EXPERIMENTELLE UNTERSUCHUNGEN ZUR PERZEPTION SYNTHETISCHER SPRACHE

4 Die implementierten Synthesemodelle

4.1 Die Verfahren der Sprachsynthese

4.1.1 Signalorientierte und erzeugungsorientierte Syntheseverfahren

Die Sprachsynthese kann in die Hauptgruppen der erzeugungsorientierten und der signalorientierten Syntheseverfahren eingeteilt werden. Bei der Gruppe der signalorientierten Verfahren steht die Frage nach der Codierung des Sprachsignals mit möglichst wenig Speicherplatz im Vordergrund. Hierzu wurden eine Vielzahl von Techniken entwickelt.[1] Unterschiedlich ist auch die Länge der abgespeicherten Spracheinheiten, die von Satzteilen über Wörter und Silben bishin zu Diphonen führt. Je kleiner diese Spracheinheiten sind, desto weniger Speicherplatz ist zur Speicherung des gesamten Inventars nötig. Allerdings verkompliziert sich damit auch die Resynthese.

Die in dieser Arbeit benutzten Synthesemodelle gehören zur Gruppe der erzeugungsorientierten Syntheseverfahren. Diese Verfahren arbeiten nach dem Vorbild der menschlichen Erzeugung von Sprachlauten. Ziel dieser Verfahren ist es, die in Hinblick auf das akustische Signal wesentlichen Mechanismen der Spracherzeugung nachzubilden. Deshalb haben die Modelle dieser Gruppe auch prinzipiell den gleichen dreiteiligen Aufbau: sie bestehen aus Glottismodell, Sprechtraktmodell und Abstrahlmodell. Das Ausgangssignal des Glottismodells, die glottale Welle, ist das

---

[1] Siehe SICKERT (1983, S. 156–175).

Eingangssignal oder auch Anregungssignal des Sprechtraktes. Das Sprechtraktmodell ist ein Filter, dessen Durchlassfrequenzen den Frequenzen der Sprechtraktresonanzen entsprechen. Die so vom Sprechtraktfilter überformte glottale Welle wird vom Abstrahlmodell in eine Schallwelle umgeformt. Letztlich entspricht auch das Abstrahlmodell einem Filter und wird gelegentlich mit dem Sprechtraktfilter zusammengefaßt.

Während die signalorientierte Synthese, auch "Signalcodierung"[2] genannt, erst mit der stürmischen Entwicklung der Digitaltechnik an Bedeutung gewann, hat die erzeugungsorientierte Synthese eine lange Tradition. KRATZENSTEIN (1782) entwickelte anläßlich eines Wettbewerbs 1779 ein mechanisches Modell zur Vokalsynthese. Dabei wurde jeder Vokal durch einen entsprechend geformten starren Hohlraumresonator generiert. Es soll damit bereits akzeptable Vokalqualität erreicht worden sein. Weitere mechanische Synthesemodelle wurden von VON KEMPELEN (1791) entwickelt. Bei diesen Modellen war die Resonatorform bereits flexibel, so daß zeitveränderliche akustische Sprachsignale synthetisiert werden konnten. DUDLEY (1940) stellte 1939 den mit analoger Filtertechnik arbeitenden VODER ("Voice operation demonstrator") vor, dessen Steuerparameter wegen ihrer Vielzahl von einem Menschen mit Händen und Füßen kontrolliert und variiert werden mußten.

Aber auch heute haben erzeugungsorientierte Syntheseverfahren nicht nur im Bereich der Forschung große Bedeutung. Ein Grund dafür ist, daß in der Nachbildung der Spracherzeugung der Schlüssel zu einer wenig Speicherplatz verbrauchenden Codierung liegt. SICKERT (1983, S. 175) spricht auch von Quellencodierung, wobei mit Quelle die Quelle des Sprachsignals, also die menschliche Spracherzeugung gemeint ist. Letztlich ist also eine Codierung durch diejenigen Parameter, die artikulatoriche und perzeptive Relevanz haben, am effektivsten. Während Signalcodierung immer Abspeicherung des Signals selbst bedeutet, werden bei

[2]Siehe Sickert (1983, S. 158).

der Quellencodierung die artikulatorischen und damit auch die perzeptiv wichtigsten Signalparameter wie z.B. Formantfrequenzverläufe oder LPC-Koeffizientensätze abgespeichert. Den geringsten Speicherplatz benötigt die Vollsynthese oder Synthese nach Regeln, bei der es nur noch nötig ist, die Phonemkette der zu synthetisierenden Äußerung anzugeben.

4.1.2 Zur Einordnung des KELLY-LOCHBAUM-Syntheseverfahrens: Formantensynthese, LPC-Synthese und KELLY-LOCHBAUM-Synthese

Die Gruppe der erzeugungsorientierten Syntheseverfahren kann andererseits nach der hinter dem Sprechtraktmodell stehenden Theorie in zwei entgegengesetzte Modellgruppen unterteilt werden. Beispiele dieser Gruppen bilden die Formantensynthese und die im Rahmen dieser Arbeit benutzte KELLY-LOCHBAUM-Synthese.

Die Formantensynthese berücksichtigt nur die akustischen Eigenschaften des Sprechtraktes. Es wird nur das Eingangssignal und das durch die Übertragungsfunktion modulierte Ausgangssignal betrachtet, also berücksichtigt das Modell nur die Gesamtübertragungseigenschaften des Sprechtrakts.[3] Dagegen wird bei der KELLY-LOCHBAUM-Synthese von der Sprechtraktgeometrie ausgegangen und das detaillierte Übertragungsverhalten an jedem Ort des Sprechtraktes betrachtet. So kann prinzipiell an jedem Ort des Sprechtrakts der Volumenstrom und der Druck der sich im Trakt ausbreitenden Welle berechnet werden.[4] Entsprechend verschieden sind auch die Sätze der Steuerparameter. Während die Formantensynthesemodelle durch Formantfrequenzen und Formantenamplituden angesteuert werden, benötigt die KELLY-LOCHBAUM-Synthese Sätze von Sprechtraktkonfigurationen.

[3]FLANAGAN (1972, S. 214) bezeichnet diese Modellgruppe als "terminal analog synthesizer".

[4]FLANAGAN (1972, S. 229) bezeichnet diese Modellgruppe als "transmission-line-synthesizer".

Die Formantensynthesemodelle werden durch einen Filter pro Formant realisiert. Die Filter können parallel oder seriell geschaltet sein. Beides ist gebräuchlich, weil sich daraus verschiedene Vor- und Nachteile ergeben und letztere sich nicht ausschließen lassen. Die Filterrealisation kann analog oder digital sein.

In dieser Arbeit wird das KELLY-LOCHBAUM-Syntheseverfahren benutzt, bei dem der Sprechtrakt digital simuliert wird.[5] Dazu muß der Sprechtrakt quantisiert werden, d.h. er wird in diskrete, gleichlange Rohrabschnitte mit unterschiedlichen Querschnitten zerlegt. Nach der Wellentheorie treten zwischen Rohrabschnitten verschiedenen Querschnitts Impedanzsprünge auf, wodurch Reflexionen der sich im Sprechtrakt ausbreitenden Teilwellen entstehen.

Neben der Formantensynthese und der KELLY-LOCHBAUM-Synthese ist die LPC-Synthese ("linear predictive coding"[6]) verbreitet. Sie ähnelt der Formantensynthese insofern, als daß die zur Ansteuerung des Modells notwendigen LPC-Koeffizieneten wie die Formantfrequenzen aus dem Sprachsignal gewonnen werden können. Allerdings ist die Gewinnung der LPC-Koeffizienten unproblematischer, da nicht wie bei der Bestimmung der Formantfrequenzen Maximum-Such-Algorithmen ("peak picking algorithms") angewendet werden müssen. Daraus begründet sich auch die weite Verbreitung der LPC-Synthese, die vor allem dann benutzt wird, wenn es um die Resynthese eines konkreten akustischen Sprachsignals geht. Aus den LPC-Koeffizienten ist leicht die Berechnung der Übertragungsfunktion des Sprechtrakts möglich, bzw. die Koeffizienten dienen als Filterkoeffizienten eines rekursiven digitalen Filters, dessen Übertragungsfunktion der Übertragungsfunktion des Sprechtrakts entspricht. Andererseits gibt es eine alternative Darstellung der LPC-Koeffizienten, PARCOR-Koeffizienten genannt ("partial correlation"[7]). Diese Koeffizienten sind genau die

[5] Siehe KELLY et al. (1962).

[6] Siehe z.B. MARKEL et al. (1976).

[7] Siehe MARKEL et al. (1976, S. 32ff).

Reflexionskoeffizienten zwischen den Sprechtraktabschnitten nach KELLY und LOCHBAUM. Somit steht die LPC-Synthese mit ihrer sowohl akustisch orientierten als auch artikulatorisch orientierten Realisationsmöglichkeit in der Mitte zwischen Formantensynthese und KELLY-LOCHBAUM-Synthese. Wegen des engen Zusammenhangs der Reflexionskoeffizienten mit den Querschnittsflächen der Traktabschnitte,[8] erlaubt das PARCOR-Verfahren auch die Extraktion von Sprechtraktgeometrien aus dem Sprachsignal. Allerdings sind die so gewonnenen Geometrien stark von der glottalen Welle und von den Randbedingungen, also der Abstrahlung und der Interaktion der Glottis mit dem Sprechtrakt, abhängig, wodurch diese Anwendung des LPC-Verfahrens starken Einschränkungen unterliegt.

4.2 Das implementierte Sprechtraktmodell und seine Randbedingungen

4.2.1 Das KELLY-LOCHBAUM-Modell

4.2.1.1 Sprechtraktabschnitte und die Definition von Teilwellen

Das KELLY-LOCHBAUM-Modell[9] ist eine einfache, aber sehr effektive Art, die Wellenausbreitung im Sprechtrakt zu simulieren. Zur Berechnung des Volumenstroms u zu einem Zeitpunkt t an einem bestimmten Ort des Sprechtraktes verlangt das KELLY-LOCHBAUM-Modell eine Quantisierung des Sprechtraktes in N gleichlange zylindrische Sprechtraktabschnitte mit konstanter Querschnittsfläche A[i]. A[1] bezeichnet die Querschnittsfläche des ersten supraglottalen Traktabschnitts oberhalb der Glottis und A[N] die des letzten Traktabschnitts an den Lippen vor der Abstrahlung der Schallwelle. Rohrabschnitte konstanten Querschnitts führen für die sich ausbreitende Welle im Sprechtrakt zu Abschnitten konstanten Wellenwiderstandes, d.h. zu konstanter

[8] Siehe Gl. 4.2 in Abs. 4.2.1.

[9] Siehe KELLY et al. (1962).

Impedanz Z[i].

(4.1) $Z[i] = rho * c / A[i]$ mit $i = 1,2,...,N$

Die Dichte der Luft wird mit rho, die Schallgeschwindigkeit mit c bezeichnet. Eine Zusammenstellung der benutzten Konstanten befindet sich in Tabelle 4.1[10]. Zur Berechnung der sich im Trakt ausbreitenden Welle werden die Orte betrachtet, an denen zwei Rohrabschnitte aneinanderstoßen (siehe Abb. 4.1).

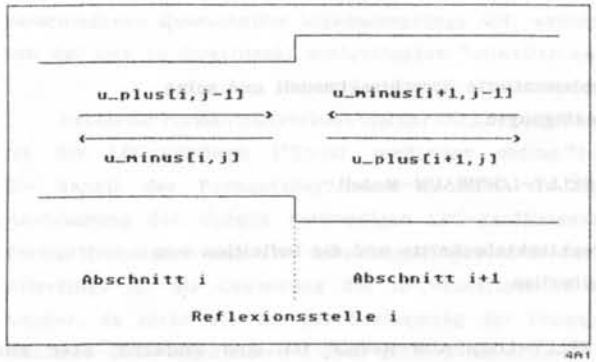


Abb. 4.1: Einfallende und reflektierte Teilwellen zum Zeitpunkt t[j] an der Reflexionsstelle i

Diese Orte, im folgenden Reflexionsstellen genannt, sind diejenigen Stellen des quantisierten Sprechtraktes, an denen die vor- und rücklaufenden Teilwellen im Falle eines Impedanzsprunges, d.h. wenn Z[i] und Z[i+1] bzw. A[i] und A[i+1] unterschiedlich sind, reflektiert werden. Innerhalb der Traktabschnitte laufen die Teilwellen störungsfrei. Der Impedanzsprung an einer Reflexionsstelle i kann durch den Reflexionskoeffizienten k[i]

[10] Siehe Abs. 4.3.2.1.

beschrieben werden.[11]

(4.2a) $k[i] = (A[i] - A[i+1]) / (A[i] + A[i+1])$
mit $i = 1,2,...,N-1$

(4.2b) $k[i] = (Z[i+1] - Z[i]) / (Z[i+1] + Z[i])$
mit $i = 1,2,...,N-1$

Die Reflexionskoeffizienten geben in anschaulicher Weise an, wie groß der Anteil der an einer Reflexionsstelle ankommenden Teilwelle ist, der reflektiert bzw. transmittiert werden muß.

4.2.1.2 Simulation des Sprechtrakts in Raum und Zeit

Betrachtet man nun die Reflexionsstelle i am Zeitpunkt t[j], so ergibt sich folgende Situation für die Teilwellen (siehe Abb.4.2).

[11] Siehe MARKEL et al. (1976, S. 63ff).

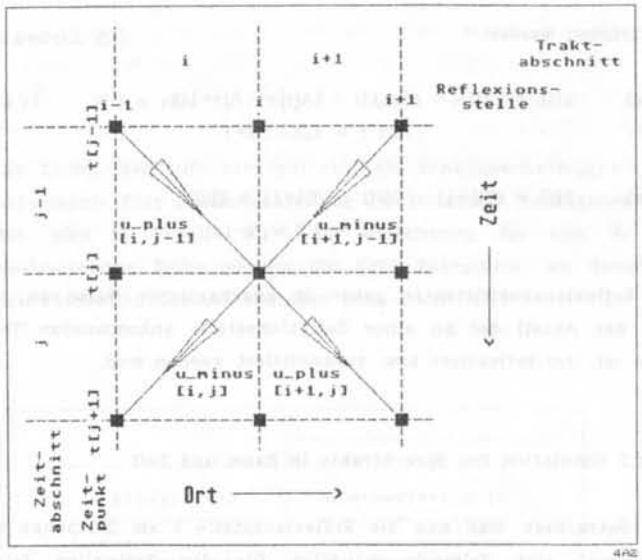


Abb. 4.2: Raum-Zeit-Darstellung für die Teilwellen an dem Reflexionspunkt [i,j]

Die Teilwellen u_plus[i,j-1] und u_minus[i+1,j-1] treffen vor der Reflexion (Zeitpunkt j) an der Stelle i zusammen und spalten sich nach den Streugleichungen in die reflektierten Wellen u_plus[i+1,j] und u_minus[i,j] auf. Die Streugleichungen ergeben sich nach KELLY et al. (1962) aus den Kontinuitätskriterien für Druck und Volumenstrom. Sie werden hier in der eleganten Formulierung von ITAKURA und SAITO vorgestellt.[12]

(4.3a) u_plus[i+1,j] = u_plus[i,j-1] - u_refl[i,j]

(4.3b) u_minus[i,j] = u_minus[i+1,j-1] + u_refl[i,j]

(4.3c) u_refl[i,j] = k[i] * (u_plus[i,j-1] + u_minus[i+1,j-1])

[12] Siehe MARKEL et al. (1976, S. 123).

Dieses Modell mit einer Multiplikation ("one multiplier model") von ITAKURA und SAITO bringt durch die Ausführung nur einer Multiplikation bei der Berechnung einer Reflexionsstelle Geschwindigkeitsvorteile gegenüber der ursprünglichen Formulierung von KELLY und LOCHBAUM, bei dem vier Multiplikationen auszuführen waren. Ferner hat der in dieser Formulierung auftretende Term u_refl als der eigentliche durch die Reflexion zwischen den Teilwellen ausgetauschte Anteil des Volumenstroms physikalisch anschauliche Bedeutung. Obwohl die oben aufgestellte Streugleichung nur für den Zeitpunkt t[j] gilt, bzw. aus den für jeden Zeitpunkt geltenden Kontinuitätsbedingungen hergeleitet worden ist, wird durch die Definition von zwei Teilwellenpaaren, je für Zeiträume kurz vor und kurz nach t[j] (u[*,j-1] und u[*,j]), die Zeitabhängigkeit deutlich, die den Streugleichungen neben der auf der Hand liegenden Ortsabhängigkeit innewohnt. Die Dimension des Raumes und der Zeit werden bei den verschiedenen Implementierungsmöglichkeiten wichtig. So wurde das Modell von ITAKURA und SAITO im Rahmen dieser Arbeit dergestalt implementiert, daß die Refkexionspunkte entlang der Diagonalen in der Raum-Zeit-Ebene berechnet werden (siehe Abb. 4.3).

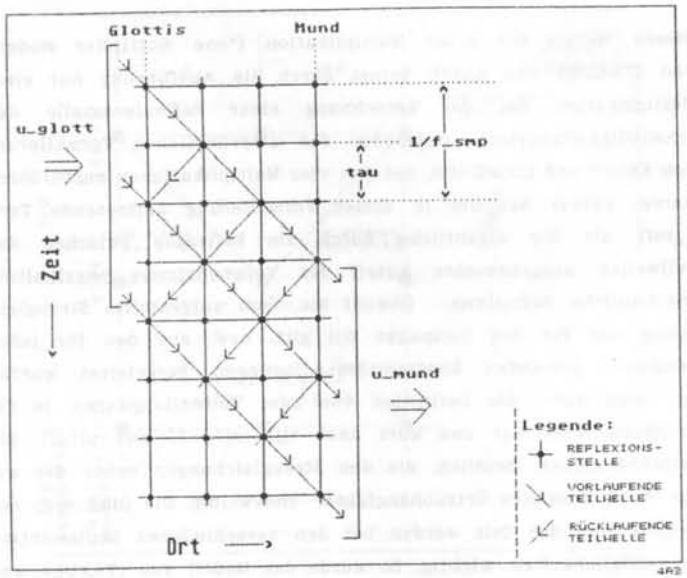


Abb. 4.3: Zum Prinzip der Berechnung der Reflexionsstellen in der Darstellung der Raum-Zeit-Ebene

Dann müssen nur die rücklaufenden Teilwellenwerte des Volumenstroms u_minus abgespeichert werden, während die positiven Teilwellenwerte direkt weiterverrechnet werden können. Abbildung 4.3 zeigt ferner, daß bei der Verwendung dieses eleganten Rechenverfahrens überhaupt nur jede zweite Reflexionsstelle in der Raum-Zeit-Ebene berechnet werden muß.

4.2.1.3 Der Zusammenhang von Sprechtraktlänge und Samplingfrequenz

Die Länge der Traktabschnitte l wird von der Samlingfrequenz f_smp - also der Frequenz, mit der die berechneten abzustrahlenden Druckwerte mittels D/A-Wandlung in analoge Größen umgesetzt werden - bzw. von dem Zeittakt, mit dem die Amplitudenwerte des glottalen Volumenstroms u_glott in den Sprechtrakt eingespeist

werden, diktiert. Für die Laufzeit tau der sich mit Schallgeschwindigkeit c in den Traktabschnitten der Länge l ausbreitenden Teilwellen gilt

(4.4a) $\quad tau = l / c .$

Abb.4.3. zeigt, daß, bezogen auf die Zeitachse, nur zu jedem zweiten Reflexionspunkt ein Wert des glottalen Volumenstroms eingeprägt wird. Das bedeutet, daß nur an jedem zweiten Reflexionspunkt ein Druckwert der resultierenden Schallwelle berechnet wird. Danach ergibt sich für den Zusammenhang zwischen der Länge der Traktabschnitte l und der Samplingfrequenz f_smp

(4.4b) $\quad f_smp = 0.5 \cdot c / l .$

Somit gibt es zwei Möglichkeiten der Variation der Sprechtraktlänge. Die einfachste Möglichkeit besteht darin, die Gesamtzahl der Traktabschnitte zwischen Glottis und Lippen zu variieren. Von dieser Methode wird in den Experimenten 1 und 2 Gebrauch gemacht. Gemäß dem hier aufgezeigten Zusammenhang zwischen der Samplingfrequenz und der Länge der Sprechtraktabschnitte ist es aber auch möglich, die Sprechtraktlänge bei konstanter Gesamtzahl der Traktabschnitte durch die Variation der Samplingfrequenz zu ändern. Bei dieser in Experiment 5 angewendeten Methode ist allerdings zu beachten, daß eine Änderung der Samplingfrequenz auch die Einspeisung der glottale Welle betrifft.

4.2.1.4 Die Berechnung von Druck und Volumenstrom im KELLY-LOCHBAUM-Synthesemodell

Trotz ihrer unmittelbaren Anschaulichkeit ist die Teilwellendarstellung ein nur beim KELLY-LOCHBAUM-Verfahren benutztes Hilfsmittel. Physikalisch relevant sind hingegen der an einem Ort i und zu einer Zeit t[j] herrschende Druck p[i,j] und Volumenstrom u[i,j] im Sprechtrakt. Diese Größen ergeben sich aus der Superposition der Teilwellenwerte.

(4.5) p[i,j] = (u_plus[i,j] + u_minus[i,j]) * Z[i]

(4.6) u[i,j] = u_plus[i,j] - u_minus[i,j]

Das Minuszeichen bei der Berechnung des Volumenstroms resultiert aus der Definition des Richtungssinns für vor- und rücklaufende Teilwellen.

4.2.2 Verlustmechanismen im Sprechtrakt

Die durch den Sprechtrakt strömende Luft ist den verschiedensten Mechanismen ausgesetzt, die zum Energieverlust in den im Sprechtrakt angeregten Schwingungsmoden führt. Verluste entstehen beispielsweise durch die viskose Reibung der Luftschichten in der Nähe der Sprechtraktwände, durch die Wärmeleitung während der durch die Luftdruckschwankungen der Schallwelle entstehenden Kompression und durch die Interaktion der Schallwelle mit den Sprechtraktwänden, wie z.B. Wandvibration und Wandabsorbtion. Bei einem quantitativen Vergleich dieser Verluste mit denen, die durch die Glottis selbst und durch die Abstrahlung der Druckwelle am Mund entstehen, treten erstere in den Hintergrund.[13] Alle Verlustmechanismen führen zur Dämpfung, also zum Abklingen der Schwingungsmoden im Sprechtrakt. Die Dämpfung bestimmt ferner die Bandbreite der Formanten. Die Bandbreite kann geradezu als Maß der Dämpfung und damit als Maß der im Sprechtrakt auftretenden Verluste angesehen werden.

Da die Verluste durch Abstrahlung der Schallwelle am Mund erst oberhalb von 1.5 kHz bis 2 kHz wesentlich werden, ist es eine gute erste Näherung, die Verluste des Sprechtraktes als frequenzunabhängig und als am Ort der Glottis lokalisiert zu betrachten. Im KELLY-LOCHBAUM-Modell kann dies durch einfaches Anfügen eines nullten Traktabschnittes realisiert werden, so daß die (nach dem Zeitpunkt t[j]) in Abschnitt 0 reflektierte rücklaufende Teilwelle u_minus[0,j] dem Vokaltrakt verloren geht

[13] Siehe LILJENCRANTS (1985, S. 2/13).

und somit für eine Dämpfung sorgt (siehe Abb.4.4). Der zur Reflexionsstelle 0 gehörige Reflexionskoeffizient k[0] wurde empirisch auf eine realistische Halbwertsbreite von rund 50 Hz für F1 und F2 eingestellt.

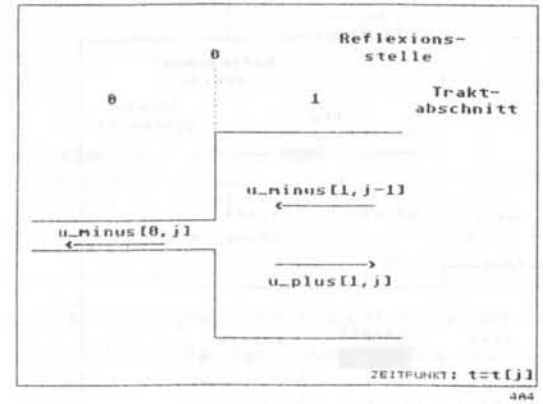


Abb. 4.4: Modellierung der Sprechtraktverluste

4.2.3 Randbedingungen des KELLY-LOCHBAUM-Modells

4.2.3.1 Elektrische Analogiebildungen

Der Anschluß von Glottis und die Abstrahlung am Mund werden oft mithilfe eines elektrisches Schaltbildes verdeutlicht. Diese Darstellung ist möglich, weil das KELLY-LOCHBAUM-Modell des Sprechtraktes einem passiven digitalen Filter entspricht und daher ebenfalls als analoge elektrische Schaltung realisiert werden kann.[14] Bei der hier gewählten Analogiebildung entspricht der elektrische Strom I[i] des Traktabschnitts i dem Volumenstrom u[i] und die elektrische Spannung dieses Traktabschnitts E[i] dem in diesem Abschnitt auftretenden Druck p[i] (siehe Abb. 4.5). Die

[14] Siehe FLANAGAN (1972, S. 229) oder LINGGARD (1985, S. 41).

physikalischen Eigenschaften der in einem Traktabschnitt befindlichen Luftsäule wie Kompressibilität und Trägheit der Luftmasse werden durch die elektrischen Größen Kapazität C[i] und Induktivität L[i] vertreten. Im elektrischen Ersatzschaltbild gibt es allerdings keine Analogie zu den im KELLY-LOCHBAUM-Modell berechneten Teilwellen des Volumenstroms.

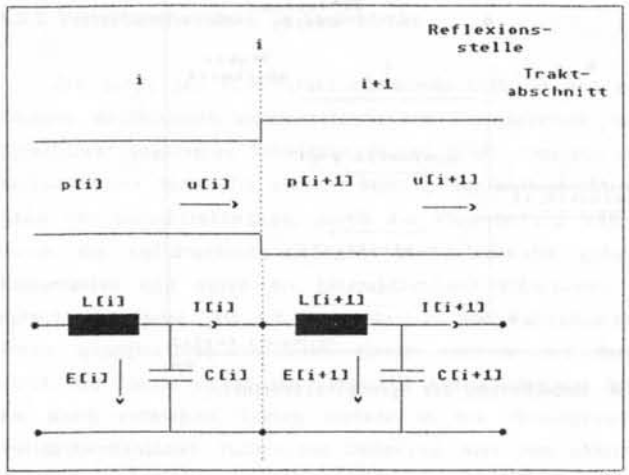


Abb. 4.5: Analogien zwischen akustischem und elektrischem Sprechtraktmodell

4.2.3.2 Die Sprechtraktanregung

Die Anregung des Sprechtrakts wird durch das Einprägen des glottalen Volumenstroms u_glott (im elektrischen Ersatzschaltbild: i_glott) als Zusatzstrom zwischen Abschnitt 0 und 1 realisiert (siehe Abb. 4.6).

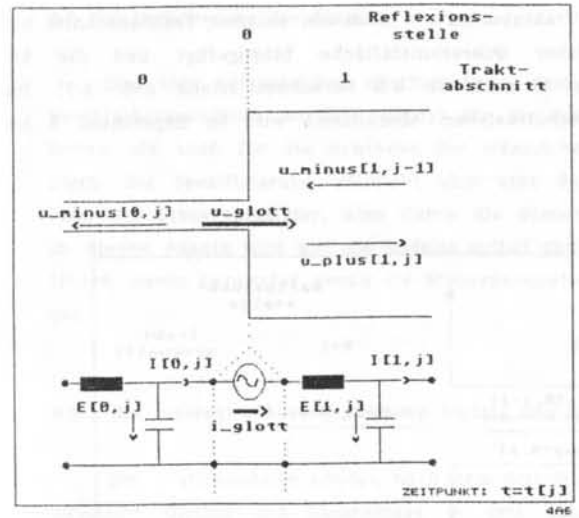


Abb. 4.6: Das Einprägen des glottalen Volumenstroms in den Sprechtrakt zum Zeitpunkt t[j]

Das Einprägen des glottalen Volumenstroms entspricht im analogen elektrischen Schaltbild dem Zwischenschalten einer Stromquelle. Da an der Reflexionsstelle 0 nahezu Totalreflexion auftritt, wird fast die gesamte rücklaufende Teilwelle u_minus[1,j-1] als u_plus[1,j] reflektiert, während nur ein geringer Teil der Welle als u_minus[0,j] in Rohr 0 gestreut wird. Damit entspricht der glottale Volumenstrom dem Gesamtstrom in Abschnitt 1, d.h. der Volumenstrom wurde in Abschnitt 1 eingeprägt. Diese glottale Randbedingung gilt für beide im Rahmen dieser Arbeit benutzten Glottismodelle.

4.2.3.3 Die Abstrahlung

Die Randbedingung am Sprechtraktende betrifft die Rückstreuung und Abstrahlung der Schallwelle im Bereich der Lippen. Im Rahmen dieser Arbeit wurden zwei Abstrahlungsmodelle benutzt. Im Modell des schallweichen Abschlusses des Sprechtraktes wird

zum letzten Traktabschnitt N noch ein weiterer Traktabschnitt N+1 mit unendlicher Querschnittsfläche hinzugefügt und die Reflexionsstelle N nach Gl. 4.3 berechnet (siehe Abb. 4.7). Das Modell des schallweichen Abschlusses wird in Experiment 5 benutzt.

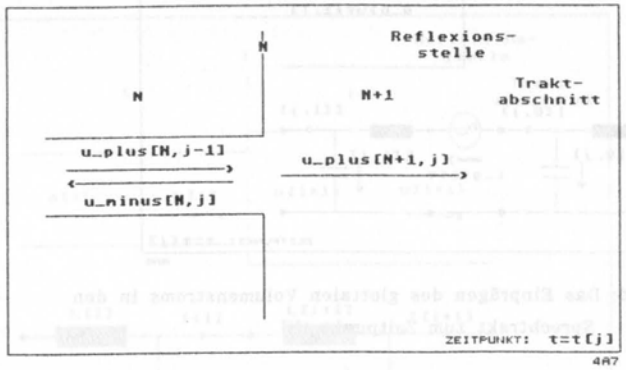


Abb. 4.7: Der schallweiche Abschluß

Ein komplexeres Modell stellt der Kolben-Kugel-Abstrahlfilter dar. Hierbei werden die durch die Abstrahlung entstehenden frequenzabhängigen Verluste berücksichtigt. Am Ende des Sprechtrakts ist hier ein digitales Filter erster Ordnung angeschlossen, dessen Pol und Nullstelle so liegt, daß die entstehende Übertragungsfunktion das Frequenzverhalten der Strahlungsimpedanz eines in einer Kugeloberfläche schwingenden Kolbens wiedergibt.[15] Dieses Modell findet in Experiment 1 bis 4 Anwendung.

Durch Differenzierung des aus dem Sprechtrakt austretenden Volumenstroms ergibt sich in beiden Abstrahlmodellen die Schalldruckwelle des Sprachsignals.

[15] Siehe LILJENCRANTS (1985, S. 3/3).

4.3 Die Modellierung der Glottis

Die hier vorgestellten Glottismodelle sind prinzipiell nicht geschlechtsspezifisch, sondern sowohl für die Synthese der weiblichen als auch für die Synthese der männlichen Stimme konzipiert. Die Spezifizierung vollzieht sich erst durch die Einstellung der Steuerparameter, also durch die Steuerung des Modells. In diesem Absatz wird auf die Modelle selbst und ihre Steuerfunktionen, deren Parameter genau die Steuerparameter sind, eingegangen.

4.3.1 Die Interaktionsarten zwischen Glottis und Sprechtrakt

Die Glottismodelle können nach dem Kriterium der Interaktion zwischen Glottis und Sprechtrakt in drei Kategorien eingeteilt werden:[16]

1) Das *GEV-Modell.* Glottismodell mit fest eingeprägtem Volumenstrom.
   Dieses Modell prägt ungeachtet aller physikalischen Gesetze einen willkürlich definierten glottalen Volumenstrom in den Sprechtrakt ein. Da in diesem Modell keine Kopplung zwischen Glottis und Sprechtrakt vorliegt, kann es auch als "lineares Modell" der Spracherzeugung bezeichnet werden. Erst durch die Rückwirkung des Sprechtraktes auf die Glottis wird das Synthesemodell nichtlinear.

2) Das *GAI-Modell.* Glottismodell mit akustischer Interaktion.
   In diesem Modell ist die Bewegungsfunktion der Stimmlippen, d.h. die Zeitfunktion für die glottale Öffnungsfläche willkürlich und ungeachtet der physikalischen Realität vorgegeben. Allerdings wird der daraus resultierende glottale Volumenstrom, wie er sich durch das Zusammenwirken der sub- und supraglottalen Druckdifferenz mit den geometrischen Verhält-

[16] Siehe auch Abs. 1.2.4.

nissen an der glottalen Engstelle ergibt, nach physikalischen Gesetzen berechnet. Durch das Einbeziehen des supraglottalen Drucks wird die Rückwirkung des Sprechtraktes auf die Glottis und damit die Interaktion zwischen Glottis und Sprechtrakt realisiert. Dieses Modell ist nichtlinear und von der Vorgeschichte abhängig, d.h. die bereits vorliegende Anregung der Schwingungsmoden des Sprechtrakts wirkt auch auf die weitere zeitliche Entwicklung der glottalen Welle ein.

3) Das *GMI-Modell.* Glottismodell mit mechanischer Interaktion.
Hier wird auch die Schwingung der Stimmlippen nicht mehr vorgegeben, sondern nach grundlegenden physikalischen Gesetzmäßigkeiten berechnet.

Zur Lösung der in dieser Arbeit vorgegebenen Fragestellungen ist ein Glottismodell mit mechanischer Interaktion nicht sinnvoll.[17] Die direkte und kontrollierbare Einstellung der Schwingungsform der glottalen Welle ist eine Anforderung an das Modell. Im GEV-Modell wird die Schwingungsform direkt eingestellt. Wegen der bei der weiblichen Stimme wichtigen Interaktion zwischen Glottis und Sprechtrakt wurde auch das GAI-Modell benutzt. Allerdings stellen die Steuerparameter dort die Öffnungsfläche als Funktion der Zeit und nicht direkt die Form der glottalen Welle ein. Gerade bei starker Interaktion können auch qualitative Unterschiede zwischen der Form der glottalen Welle und der Schwingungsform der Stimmlippen auftreten.[18]

Bezüglich der Durchführung von Perzeptionsexperimenten besteht ein Nachteil des GMI-Modells darin, daß es in diesem Modell nicht möglich ist, die Grundfrequenz exakt zu kontrollieren. Beispielsweise wird aber in Experiment 5 zur Bestimmung der Vokalqualitätsdifferenzen eine solche exakte Kontrolle verlangt. Im GMI-Modell kann die Grundfrequenz nur mittelbar über die Stimmlippenspannung und den subglottalen Druck beeinflußt werden,

[17] Siehe auch Abs. 1.4.

[18] Siehe Abs. 4.3.2.3 und Abs. 4.3.2.4.

ist aber auch dann noch abhängig von der Sprechtraktgeometrie.[19] Somit war neben der schlechten Kontrolle der Form der glottalen Welle auch die schlechte Kontrolle der Grundfrequenz selbst ein Grund, zunächst von der Implementierung des GMI-Modells abzusehen.

4.3.2 Das GAI-Modell: Glottismodell mit akustischer Interaktion

4.3.2.1 Das Modell

Das GAI-Modell betrachtet die Glottis als eine Engstelle in einem Rohr. Durch diese Engstelle wird der subglottale Bereich (Lunge, Bronchien und Luftröhre) vom supraglottalen Bereich (Sprechtrakt) getrennt. Der glottale Volumenstrom u_glott an der durch die Stimmlippen gebildeten Engstelle wird aus der Differenz delta_p von supra- und subglottalem Druck p_supra und p_sub berechnet.

(4.7) delta_p = p_sub - p_supra

Die Engstelle selbst stellt für den Luftstrom einen Widerstand dar. Im elektrischen Ersatzschaltbild (siehe Abb. 4.8) wird sie folglich als Impedanz z_glott[t] betrachtet, die entsprechend der glottalen Öffnungsfläche zeitabhängig ist.

[19] Das GMI-Modell entspricht mehr der natürlichen Sprachproduktion, in der ebenfalls keine absolut exakte Kontrolle der Grundfrequenz gegeben ist.

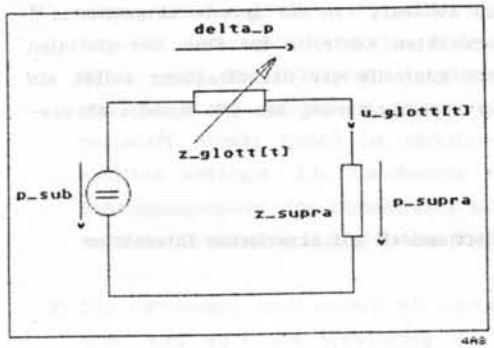


Abb. 4.8: Elektrisches Ersatzschaltbild der Glottis im GAI_Modell

Die Kopplung von Glottis und Sprechtrakt besteht in der Abhängigkeit des glottalen Volumenstroms auch von der Eingangsimpedanz des Sprechtrakts z_supra (siehe auch Abb. 1.1). Interaktion zwischen Glottis und Sprechtrakt tritt dann auf, wenn die Eingangsimpedanz des Sprechtrakts nicht mehr gegenüber der Glottisimpedanz vernachlässigbar ist.

Zur Berechnung des glottalen Volumenstroms u_glott muß die glottale Impedanz z_glott als Funktion der momentanen Querschnittsfläche a_glott bekannt sein. Der Volumenstrom ergibt sich dann aus dem BERNOULLI-Gesetz (Gleichung 4.8).

(4.8) $\quad delta_p = 0.5 \cdot rho \cdot (u_glott/a_glott)^2$

Gemäß dem OHMschen Gesetz kann danach ein aus Gl. 4.8 folgender Impedanzterm R_bern abgeleitet werden. Ein Faktor k wird zur Korrektur der bezüglich der Geometrie idealisierten Annahmen eingeführt.[20] Die Werte der benutzten Konstanten werden in Tabelle 4.1 zusammengestellt.

[20] Siehe auch FLANAGAN (1972, S. 44) und ANANTHAPADMANABHA et al. (1982, S. 2-3).

(4.9) $\quad R_bern = 0.5 \cdot k \cdot rho \cdot u_glott / a_glott^2$

Dieser Widerstand ist nichtlinear, da er von dem ihn durchfließenden Strom u_glott abhängt.

rho	= 1.14 Ns^2/m^4	Dichte der Luft
c	= 350 m/s	Schallgeschwindigkeit
mue	= $1.86 \cdot 10^{-5}$ Ns/m^2	Viskositätskonstante der Luft
k	= 1.1	Geometrie-korrekturfaktor
l_glott	= 18 mm	Länge der Stimmritze
h_glott	= 3 mm	Höhe der glottalen Enge

Tab. 4.1: Zusammenstellung der physikalischen Konstanten und der Geometriefaktoren

Neben dem Hauptterm der Glottisimpedanz R_bern, der sich aus dem BERNOULLI-Effekt ergibt, treten noch zwei weitere Impedanzfaktoren auf, die aber gegenüber dem Hauptfaktor nur von untergeordneter Bedeutung sind. Daher ist es auch nicht erforderlich, die in Tabelle 4.1 aufgelisteten Geometriefaktoren für die Synthese der weiblichen und der männliche Stimme getrennt einzustellen.

Erstens: Durch die Enge der glottalen Querschnittsfläche sollte die dort durch die viskose Reibung entstehende Zusatzimpedanz R_visc mitberücksichtigt werden.

(4.10) $\quad R_visc = 12 \cdot mue \cdot l_glott^2 \cdot h_glott / a_glott^2$

Dabei bezeichnet mue die Viskositätskonstante von Luft (siehe Tab. 4.1), l_glott die Länge der Stimmritze (siehe Abb. 4.9 und Tab. 4.1) und h_glott die Höhe der glottalen Enge (siehe Abb. 4.9 und Tab. 4.1).

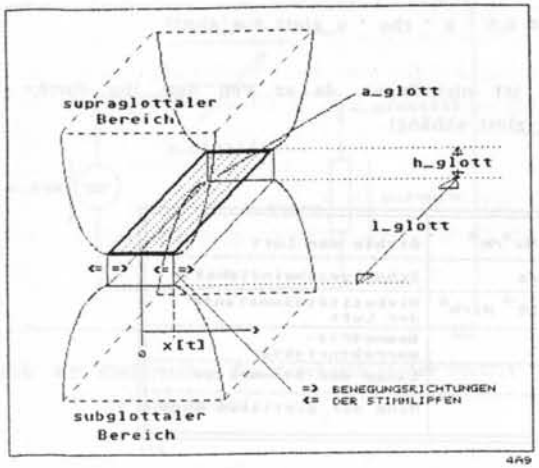


Abb. 4.9: Die stark vereinfachte Geometrie der glottalen Engstelle im GAI-Modell. x[t] symbolisiert den Abstand der Stimmlippen voneinander

Zweitens: Die Trägheit der schwingenden Luftsäule am Ort der glottalen Verengung liefert einen induktiven Impedanzanteil L_glott.

(4.11) L_glott = rho * h_glott / a_glott

Wegen des geringen Beitrags, den R_visc und L_glott neben R_bern zu z_glott liefern, ist es nicht nötig, die in R_visc und L_glott auftretenden Faktoren zu optimieren. Insbesondere ist es unwichtig, die Länge und Höhe der glottalen Verengung auf die geschlechtsspezifischen Unterschiede hin einzustellen. Der im Bernoulliterm auftretende Geometriefaktor wird in der Literatur ebenfalls mit leicht differierenden Werten angegeben.[21] Damit kann gemäß der Schaltung der Impedanzen (siehe Abb. 4.8) der

[21] Siehe FLANAGAN (1972, S. 45) und ANANTHAPADMANABHA et al. (1982, S. 4).

glottale Volumenstrom u_glott berechnet werden.

(4.12) u_glott * (R_bern + R_visc) + d/dt [u_glott * L_glott] = delta_p

Die Bestimmung des glottalen Druckabfalls delta_p erfolgt, indem der supraglottale Druck aus den Teilwellenwerten des KELLY-LOCHBAUM-Modells berechnet wird (siehe Gl. 4.5). Der subglottale Druck ist vorgegeben. Durch das Einbeziehen der Eingangsimpedanz des Sprechtrakts p_supra in die Berechnung des glottalen Volumenstroms vollzieht sich die Rückkopplung des Sprechtrakts auf die Glottis. Der so berechnete glottale Volumenstrom wird in der in Absatz 4.2 beschriebenen Weise in den Sprechtrakt eingeprägt. Dadurch wird der supraglottale Druck beeinflußt, der seinerseits wieder auf die Berechnung von u_glott zurückwirkt.

Gleichung 4.12 ist eine nichtlineare Differentialgleichung erster Ordnung in der Variablen u_glott. Diese Differentialgleichung kann nicht leicht integriert werden. Das Problem ist aber lösbar, indem die Ableitung u_glott' des glottalen Volumenstroms gemäß Gleichung 4.13 durch den glottalen Volumenstrom u_glott selbst ersetzt wird.[22]

(4.13) u_glott' = (u_glott - u_glott_delay) / delta_t

Die Größe u_glott_delay bezeichnet das jeweils zuletzt berechnete Sample des glottalen Volumenstroms, und delta_t bezeichnet das Samplingintervall (delta_t = 1 / f_smp). Nach Einsetzen von Gl. 4.13 geht die zu lösende nichtlineare Differentialgleichung erster Ordnung (Gl. 4.12) in eine einfach lösbare quadratische Gleichung über.

[22] Siehe z.B. STIEFEL (1976, S. 122ff).

4.3.2.2 Die Steuerfunktion

Im GAI-Modell übernimmt die *Steuerfunktion* die Aufgabe der Umsetzung der Wellenformparameter[23] in eine Funktion der Zeit, die die glottale Öffnungsfläche a_glott[t] zu jedem Zeitpunkt t definiert. Die hier benutzte Funktion beschreibt die Offenphase der Glottis durch zwei Sinus- oder Cosinus-Halbwellen. Sie wird also *trigonometrische Halbwelle-Halbwelle-Funktion* oder im folgenden kurz als *THH-Funktion* bezeichnet (zur Definition siehe Tab. 4.2 und Abb. 4.10). In der Literatur werden neben den trigonometrischen Funktionen auch eine Vielzahl anderer Steuerfunktionen angegeben, wie z.B. die Exponentialfunktion oder rationale Funktionen.[24]

Zeitraum	a_glott(THH)[t]
0<t<TP	$dc_a_glott + \frac{ac_a_glott}{2}\left(1-\cos\left(\frac{t}{TP}\pi\right)\right)$
TP<t<TN	$dc_a_glott + \frac{ac_a_glott}{2}\left(1+\cos\left(\frac{t-TP}{TN-TP}\pi\right)\right)$
TN<t<T0	dc_a_glott

4T2

Tab. 4.2: Definition der THH-Steuerfunktion im GAI-Modell

[23] Zur Definition siehe Abs. 1.2.3.

[24] Siehe ROSENBERG (1971) oder FANT et al. (1985).

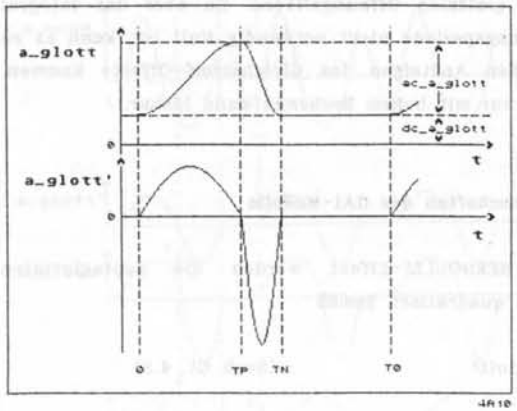


Abb. 4.10: Die THH-Steuerfunktion im GAI-Modell

Im GAI-Modell können allerdings nur Steuerfunktionen mit stetiger erster Ableitung eingesetzt werden, da sonst ein starker Rückkopplungseffekt auftritt. Unstetigkeiten in der ersten Ableitung führen zur Anregung hochfrequenter Schwingungsmoden im Sprechtrakt, die auf den glottalen Volumenstrom selbst zurückwirken und somit ein Aufschaukeln der hohen Frequenzen beim glottalen Volumenstrom, im Sprechtrakt und in der abgestrahlten Druckwelle bewirken. Das GAI-Modell arbeitet im allgemeinen erst bei stetiger Ableitung der glottalen Öffnungsfläche. Dies bedingt einen Abfall des Quellenspektrums mit -18dB/Okt. , während für reale Quellenspektren im allgemeinen ein spektraler Abfall um -12 dB/Okt. festgestellt wurde.[25] Um schwächer abfallende Quellenspektren zuzulassen, müßte der Sprechtrakt im Frequenzbereich ab 3 kHz übermäßig stark gedämpft werden. Abhilfe könnte evtl. auch eine analytische Steuerfunktion wie z.B. das LF-Modell schaffen. Beim LF-Modell[26] kann die Steilheit des spektralen Abfalls des Quellenspektrums zwischen -18dB/Okt. und -12dB/Okt. frei eingestellt werden. Jedoch bezieht sich diese analytische Funktion auf

[25] Siehe MONSEN et al. (1977).

[26] LILJENCRANTS-FANT-Modell, siehe FANT et al. (1985).

die Ableitung der glottalen Öffnungsfläche. Da aber das Integral über eine Schwingungsperiode nicht notwendig Null ist, kann es zu einem unrealistischen Ansteigen des Gleichstrom-Offsets kommen. Dieses Problem ist nur mit hohem Rechenaufwand lösbar.

4.3.2.3 Einige Eigenschaften des GAI-Modells

Durch den BERNOULLI-Effekt werden die supraglottalen Druckschwankungen quadratisch gemäß

$$delta_p \sim u_glott^2 \qquad \text{(nach Gl. 4.8)}$$

auf den glottalen Volumenstrom u_glott übertragen und wieder in den Sprechtrakt eingespeist. Der quadratische Zusammenhang zeigt deutlich die Nichtlinearität des Modells. Ferner ist der glottale Volumenstrom u_glott über den supraglottalen Druck p_supra von der "Vorgeschichte" des Modells abhängig. Ist beispielsweise ein Formant in der Verschlußphase angeregt - und das ist die Regel -, so wird durch diese Anregung, auch bei konstanter glottaler Öffnungsfläche während der Verschlußphase, eine Oszillation des glottalen Volumenstroms in der Verschlußphase mit der Frequenz des Formanten bewirkt. Dieser Effekt wird hier auch in den Teststimuli zur Synthese der weiblichen Stimme besonders in der Ableitung des glottalen Volumenstroms sichtbar (siehe Abb. 4.11, Pfeil 1). Dieser *Überlagerungseffekt*[27], im glottalen Volumenstrom u_glott durch Frequenzkomponenten höher als F0, wurde ursprünglich in der Offenphase erkannt, da ältere Modelle noch ohne Gleichstrom-Offset in der Verschlußphase arbeiteten.

[27] "ripple", siehe FANT et al. (1987, S. 15).

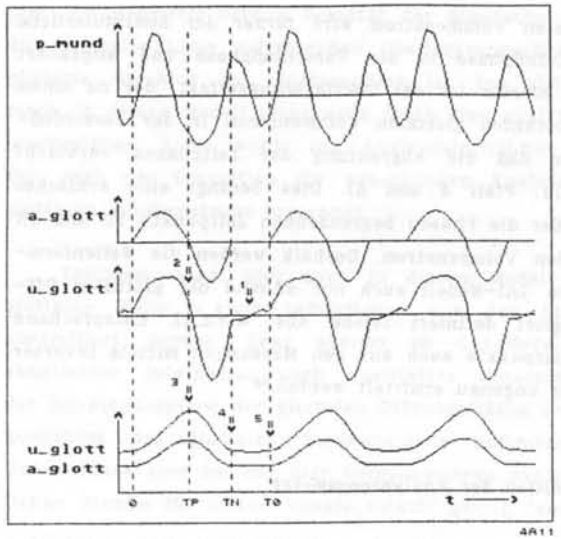


Abb. 4.11: Zum Überlagerungseffekt (Pfeil 1), zur Impulsmaximumverschiebung (Pfeil 2 und 3) und zur Phasenbegrenzung von u_glott und a_glott (Pfeil 4 und 5) im GAI-Modell

Ein weiterer im GAI-Modell und auch in der Realität auftretender Effekt ist die Verschiebung des Zeitpunktes TP, kurz *Impulsmaximumverschiebung*[28] genannt. Hier ist das Maximum des glottalen Volumenstroms gegenüber dem Maximum der glottalen Öffnungsfläche zu einem späteren Zeitpunkt hin verschoben. Dieser Effekt ist bei geringer Kopplung aber kaum sichtbar. Er beruht auf der an der glottalen Engstelle schwingenden Luftsäule, wird also vom induktiven Term der Glottisimpedanz L_glott hervorgerufen.[29] Der Effekt tritt aber bei den hier generierten Schwa-Lauten der weiblichen Stimme nur in geringem Maße auf (siehe Abb. 4.11, Pfeil 2 und 3).

[28] "pulseskewing", siehe FANT et al. (1987, S. 15).

[29] Siehe ROTHENBERG (1981).

Beim glottalen Volumenstrom wird ferner der kontinuierliche Übergang der Offenphase in die Verschlußphase und umgekehrt sichtbar. Grund hierfür ist der Überlagerungseffekt, der zu einem nicht exakt konstanten glottalen Volumenstrom in der Verschlußphase führt, so daß die Abgrenzung der Zeitphasen verwischt (siehe Abb. 4.11, Pfeil 4 und 5). Dies bedingt eine schlechte Definierbarkeit der die Phasen begrenzenden Zeitpunkte T0 und TN aus dem glottalen Volumenstrom. Deshalb werden die Wellenform-Zeitparameter im GAI-Modell auch nur anhand der glottalen Öffnungsfläche a_glott definiert (siehe Abs. 4.3.2.4). Entsprechend können diese Zeitpunkte auch aus den Messungen mittels inverser Filtertechnik nur ungenau ermittelt werden.[30]

4.3.2.4 Zur Definition der Analyseparameter

Durch die Funktionsweise des GAI-Modells bedingt, wird hier nicht der glottale Volumenstrom u_glott direkt, sondern die glottale Öffnungsfläche a_glott angesteuert. Dieses Charakteristikum des GAI-Modells ist der Grund für die Einführung von *Analyseparametern*. Die im GAI-Modell simulierte glottale Welle wird numerisch analysiert, um anhand dieser Prozedur zu Parametern zu gelangen, die die Form der glottalen Welle charakterisieren. Darum ist die Definition der Analyseparameter auch eng an die Definition der Steuerparameter angelehnt.[31]

Zwar ist im GAI-Modell zwischen den auf die glottale Öffnungsfläche a_glott wirkenden Steuerparametern und den aus dem glottalen Volumenstrom u_glott berechneten Analyseparametern zu unterscheiden, aber trotzdem gelingt schon durch die Steuerparameter die genaue Einstellung des glottalen Volumenstroms. Im GAI-Modell werden die Parameter des glottalen Volumenstroms lediglich erst nach Ablauf der Analyseprozedur zugänglich, weil der glottale Volumenstrom bereits Resultat der Simulation ist und durch die in erster Linie auftretenden Überlagerungseffekte eine komplizierte Struktur hat. Insbesondere ist der glottale Volumenstrom in diesem Modell nicht mehr durch eine analytische Funktion beschreibbar. Somit wurde ein Analysealgorithmus implementiert, der nach der Simulation die wesentlichen Analyseparameter des glottalen Volumenstroms berechnet.

---

[30] Siehe HOLMBERG et al. (1988).

[31] Vgl. Tab. 4.3 und Abb. 4.12 mit Tab. 1.3 und Abb. 1.2 in Abs. 1.2.3.

Trotzdem kann aber auch in diesem Modell die Form der glottalen Welle u_glott befriedigend von den Steuerparametern kontrolliert werden. Zwar können im GAI-Modell bei starker akustischer Interaktion auch qualitative Unterschiede zwischen der Schwingungsform der glottalen Öffnungsfläche und der Schwingungsform des glottalen Volumenstroms auftreten, aber diese Unterschiede sind bei der hier vorgenommenen Synthese der weiblichen Stimme für offene Vokale relativ gering (siehe Abb. 4.11).

Die Analyse der beiden Schwingungsformen zeigt, daß die die Öffnungs-, Schließ- und Verschlußphasen trennenden Zeitpunkte T0, TP und TN ungefähr gleich bleiben. Allerdings ist die Amplitude des glottalen Volumenstroms neben der glottalen Öffnungsfläche auch vom subglottalen Druck p_sub abhängig. So werden im GAI-Modell hauptsächlich die Amplitudenparameter des glottalen Volumenstroms mittels dieser Analyse bestimmt. Diese Parameter erlangen Bedeutung bei der exakten Abgleichung der simulierten Volumenstromwerte mit den aus Messungen mittels der Methode des inversen Filterns bekannten Daten.

Der Analysealgorithmus kann aber auch auf analytische Funktionen angewendet werden. So liefert er beispielsweise auch die Parameter des glottalen Volumenstroms im GEV-Modell.[32] Die Parameter werden in Tabelle 4.3 definiert (siehe auch Abb. 4.12). Dabei werden die zur Berechnung des mittleren Stroms und des Verschlußstroms notwendigen Integrationen numerisch als Amplitudensummationen ausgeführt.

---

[32] Siehe Abs. 4.3.3.

$pkfl = max \{ u_glott[t] \}$	Maximalstrom
$dcfl = \frac{1}{T0-TN} \int_{TN}^{T0} u_glott[t]\ dt$	Verschlußstron
$acfl = pkfl - dcfl$	Stronschwingungs-differenz
$avfl = \frac{1}{T0} \int_{0}^{T0} u_glott[t]\ dt$	mittlerer Stron
$mfdr = max \{ \lvert u_glott'[t] \rvert \}$	Stromänderungs-maximum

4T3

Tab. 4.3: Definition der Analyseparameter

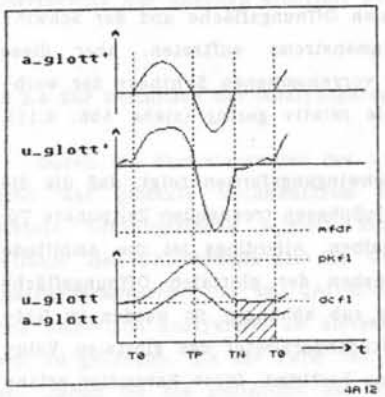


Abb. 4.12: Zur Definition der Analyseparameter

Wie in Kapitel 1 bereits ausgeführt wurde, ist es ein Ziel dieser Untersuchung, die synthetisierte glottale Wellenform nach den Analyseparametern von HOLMBERG et al. (1988), die die glottale Wellenform mit der Technik des inversen Filterns ausgemessen haben, zu optimieren. Darum wurden die hier definierten Analyseparameter eng an die Definitionen der Analyseparameter von HOLMBERG et al. angelehnt.

Jedoch sind auch einige Unterschiede zu vermerken. Während bei HOLMBERG et al. der "minimale Strom" eingeführt wird, wird hier die Integralgröße "Verschlußstrom" benutzt. Der Verschluß-

strom ist als der in der Verschlußphase auftretende mittlere Strom definiert (siehe Tab. 4.2), während der minimale Strom das in der Verschlußphase liegende Minimum des glottalen Volumenstroms angibt. Da beabsichtigt ist, über die Steuerparameter und damit auch über die Analyseparameter suprasegmentale Eigenschaften wie die Stimmqualität vorzugeben, scheint es nicht sinnvoll zu sein, eine Größe wie das Minimum zu benutzen, die vom Überlagerungseffekt durch die Sprechtraktkopplung, also von den segmentalen Artikulatorpositionen des Sprechtrakts, beeinflußt ist.

Neben den sich direkt auf den glottalen Volumenstrom beziehenden Amplitudenparametern tritt auch der wichtige Parameter des Stromänderungsmaximums (mfdr) auf, der das Maximum der Ableitung des glottalen Volumenstroms u_glott_ab angibt. Diese Größe definiert die Stärke der Anregung des Sprechtrakts und damit die Energie bzw. die Lautstärke des abgestrahlten Signals.[33] Da die Stimmlippen kurz vor dem Verschluß, also in der Schließphase, ihren größten Geschwindigkeitsbetrag erreichen, tritt das Stromänderungsmaximum immer kurz vor ihrem Verschluß, also dem Zusammenstoßen der Stimmlippen, auf. Die große Änderung des Volumenstroms in der Schließphase, d.h. die steile negative Flanke des glottalen Volumenstroms sorgt also im wesentlichen für die Anregung des Sprechtrakts, wie auch anhand der im KELLY-LOCHBAUM-Modell simulierten Signalverläufe für die abgestrahlte Druckwelle und den glottalen Volumenstrom gezeigt werden kann (siehe Abb. 4.13).

[33] Siehe FANT et al. (1987, S. 21).

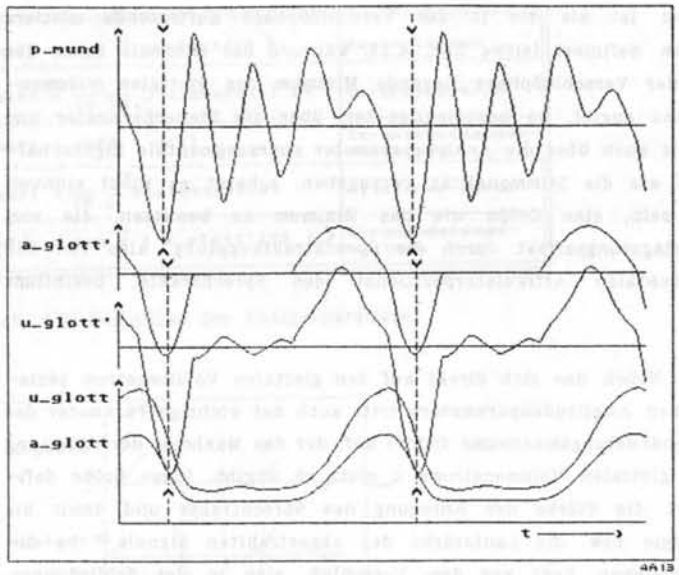


Abb. 4.13: Sprechtraktanregung und Stromänderungsmaximum

Abbildung 4.13 stellt die simulierten Werte für die Ableitung des glottalen Volumenstroms u_glott' und für die abgestrahlte Druckwelle p_mund, also die Ableitung des am Mund auftretenden Volumenstroms, dar. Durch die spezielle Implementierung der Sprechtraktsimulation, bei der die Reflexionsstellen entlang der Diagonalen in der Raum-Zeit-Ebene berechnet werden, wird die Abstrahlung p_mund ohne Zeitverzögerung zur Anregung durch den glottalen Volumenstrom u_glott bzw. dessen Ableitung u_glott' dargestellt. Die durch die Schallgeschwindigkeit gegebene Verzögerung der Ausbreitung des Glottisimpulses im Sprechtrakt ist eben in der hier benutzten Sprechtraktsimulation enthalten. Abbildung 4.13 zeigt anschaulich, daß der in den Sprechtrakt eingespeiste Glottisimpuls (siehe Pfeil und gestrichelte Linie) genau vom Stromänderungsmaximum bestimmt wird. Durch den Glottisimpuls werden die Schwingungsmoden im Sprechtrakt (Formanten) angeregt, die dann in der Verschlußphase zu abklingenden Oszillationen gemäß den Formantfrequenzen führen,

bis der Trakt vom nächsten Glottisimpuls erneut angeregt wird.[34]

Während die Amplitudenparameter, wie hier ausgeführt wurde, anhand einer Analyseprozedur aus dem glottalen Volumenstrom bestimmt werden, bleibt andererseits die Definition der Wellenform-Zeitparameter auf die glottale Öffnungsfläche a_glott bezogen. Zwar sind die Zeitphasen auch anhand der Stimmlippenbewegung definiert worden[35], aber entscheidend für das synthetisierte Signal ist immer der in den Sprechtrakt eingeprägte Volumenstrom. Darum sind die Zeitparameter auch schon in Kapitel 1 graphisch am glottalen Volumenstrom veranschaulicht worden (siehe Abb. 1.2). Diese schwache Trennung zwischen den Zeitparametern des glottalen Volumenstroms und der glottalen Öffnungsfläche ist gerechtfertigt durch die in diesem Modell und wohl auch in der Realität vorliegenden geringen Unterschiede zwischen der Schwingungsform beider Wellen. Erst bei starker akustischer Interaktion können stärkere Unterschiede bezüglich beider Wellen auch in den Zeitparametern auftreten (Impulsmaximumverschiebung), jedoch ergaben sich für die hier synthetisierten Vokale keine Probleme.

4.3.3 Das GEV-Modell: Glottismodell mit fest eingeprägtem Volumenstrom

4.3.3.1 Das Modell

Die Realisation des GEV-Modells ist im Gegensatz zum GAI-Modell sehr einfach, da nur die sich aus der Steuerfunktion ergebenden Werte in der in Absatz 4.2 beschriebenen Weise in den Sprechtrakt eingeprägt werden müssen. Demzufolge tritt die Größe

---

[34] Eine geringfügige Verschiebung gegenüber der abgestrahlten Druckwelle p_mund ergibt sich aus der Überlagerung des glottalen Impulses mit der im Sprechtrakt zu diesem Zeitpunkt vorliegenden abklingenden Druckschwankung.

[35] Siehe Abs. 1.2.3.

der glottalen Öffnungsfläche in diesem Modell nicht auf. Die Steuerfunktion beschreibt direkt den glottalen Volumenstrom, während sie im GAI-Modell die glottale Öffnungsfläche beschreibt.

4.3.3.2 Die Steuerfunktionen

Im Gegensatz zum GAI-Modell muß im GEV-Modell nicht die Bedingung der Stetigkeit der ersten Ableitung der Steuerfunktion eingehalten werden. Darum kann neben der *THH-Funktion* hier auch die *trigonometrische Halbwelle-Viertelwelle-Funktion*, kurz *THV-Funktion* genannt, verwendet werden (zur Definition siehe Gl. 4.15, Gl. 4.16 und Abb. 4.14 , 4.15). Die THV-Funktion führt nach der Untersuchung von ROSENBERG (1971) unter einer Vielzahl von Steuerfunktionen bei der Synthese der männlichen Stimme zur natürlichsten Stimmqualität.

Zeitraum	u_glott(THV)[t]
0<t<TP	$dc_u_glott + \frac{ac_u_glott}{2}\left(1-\cos\left(\frac{t}{TP}\pi\right)\right)$
TP<t<TN	$dc_u_glott + ac_u_glott * \cos\left(\frac{t-TP}{TN-TP} * \frac{\pi}{2}\right)$
TN<t<T0	dc_u_glott

4T4

Tab. 4.4: Definition der THV-Steuerfunktion im GEV-Modell

Zeitraum	u_glott(THH)[t]
0<t<TP	$dc_u_glott + \frac{ac_u_glott}{2}\left(1-\cos\left(\frac{t}{TP}\pi\right)\right)$
TP<t<TN	$dc_u_glott + \frac{ac_u_glott}{2}\left(1+\cos\left(\frac{t-TP}{TN-TP}\pi\right)\right)$
TN<t<T0	dc_u_glott

4T5

Tab. 4.5: Definition der THH-Steuerfunktion im GEV-Modell

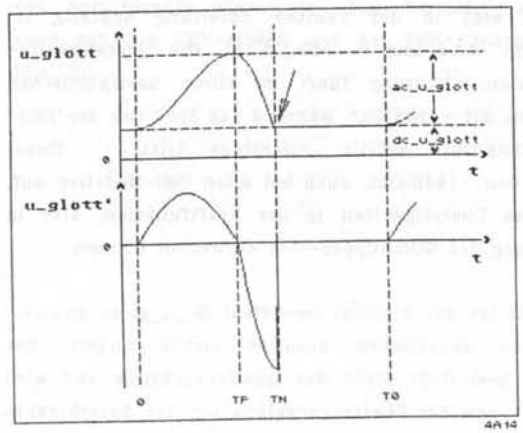


Abb. 4.14: Die THV-Steuerfunktion im GEV-Modell

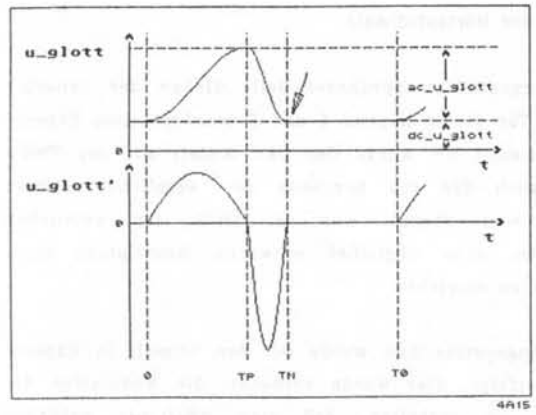


Abb. 4.15: Die THH-Steuerfunktion im GEV-Modell

Der wesentliche Unterschied dieser beiden trigonometrischen Steuerfunktionen besteht darin, daß die Schließphase bei der THV-Funktion einen Knick beim Übergang zur Verschlußphase, d.h. eine

Unstetigkeitsstelle in der ersten Ableitung, aufweist, während die THH-Funktion erst in der zweiten Ableitung unstetig ist (siehe Pfeil in Abb. 4.14 und in Abb. 4.15). Die Unstetigkeitsstelle in der ersten Ableitung führt zu einem asymptotischen Abfall des Spektrums mit -12dB/Okt. während das Spektrum der THH-Funktion mit -18dB/Okt. abfällt. Allerdings tritt dieser starke Abfall von -18dB/Okt. auch bei allen GMI-Modellen auf, da diese frühestens Unstetigkeiten in der Kraftfunktion, also in der zweiten Ableitung des Stimmlippenortes aufweisen können.

Im GEV-Modell ist der Gleichstrom-Offset dc_u_glott unwichtig, da dieser am akustischen Resultat nichts ändert. Der Gleichstrom-Offset beeinflußt nicht das Quellenspektrum und wird insofern auch nicht von der Übertragungsfunktion des Sprechtrakts erfaßt. Wichtig ist dieser Offset aber im GAI-Modell, so daß er der einfacheren Vergleichbarkeit halber auch im GEV-Modell beibehalten wurde.

4.4 Zur Synthese der Hörteststimuli

Die hier vorgestellten Synthesemodelle dienen der Generierung der Stimuli für die in Kapitel 5 und 6 beschriebenen Experimente. Für Experiment 5[36] wurde das GEV-Modell mit der THV-Steuerfunktion auch für die Synthese der weiblichen Stimme benutzt. Ziel dieser Stimuli war es, trotz der weiblichen Grundfrequenzwerte, eine möglichst schwache Assoziation einer weiblichen Stimme zu erreichen.

Das entgegengesetzte Ziel wurde mit den Stimuli in Experiment 1 bis 4 verfolgt. Hier wurde versucht, die Modulation der Stimmqualität so zu gestalten, daß eine möglichst natürlich klingende weibliche Stimme resultiert. Dazu wurde das GAI-Modell mit der THH-Funktion für die Ansteuerung benutzt. Ferner wurde hier der Kugel-Kolben-Abstrahlfilter gewählt, während zur Generierung der Stimuli für Experiment 5 der schallweiche Abschluß des Sprechtrakts ausreichte. Zum Vergleich wurden die Stimuli auch mit dem GEV-Modell und der THH-Funktion generiert (siehe Exp. 2).

[36] Siehe Kap. 6.

5 Experimentelle Untersuchungen zum Einfluß der phonatorischen Parameter auf die Synthese der weiblichen Stimme

Wie in Experiment 1 und 5 gezeigt wird, reicht es nicht, F0 und die Formantfrequenzen auf die durchschnittlichen Werte weiblicher Sprecher einzustellen. Die Stimuli werden danach noch nicht unbedingt als von Frauen erzeugt akzeptiert. Hier soll nachgewiesen werden, daß die Eindeutigkeit der Geschlechtsassoziation entscheidend von der Einstellung der phonatorischen Parameter abhängt.

In Absatz 5.1 wird die Optimierung der glottalen Parameter im GAI-Modell nach physiologischen Daten von HOLMBERG et al. (1988) für die Synthese der weiblichen und der männlichen Stimme vorgestellt. Experiment 1 (siehe Abs. 5.2) untersucht den Einfluß der Optimierung der glottalen Parameter auf die Eindeutigkeit der Geschlechtsassoziation. In Absatz 5.3 wird der Einfluß der unterschiedlichen Synthesemodelle auf den Stimmklang bei der Synthese der weiblichen Stimme untersucht (siehe Exp. 2). Die Experimente 3 und 4 (siehe Abs. 5.4) beschäftigen sich mit der Stärke der Veränderung des Signals durch die verschiedenen phonatorischen Parameter und deren Auswirkung auf den Stimmklang bei der Synthese der weiblichen Stimme.

5.1 Die Optimierung der phonatorischen Parameter nach HOLMBERG im GAI-Modell

In der von HOLMBERG et al. (1988) durchgeführten Studie wurde bei mehreren weiblichen und männlichen Sprechern der subglottale Druck und der Zeitverlauf des glottalen Volumenstroms mithilfe der Technik des inversen Filterns ermittelt. Dabei wurden eine Vielzahl von Parametern, im folgenden HOLMBERG-Parameter genannt, definiert, die den Verlauf der ermittelten Wellenform des glottalen Volumenstroms charakterisieren. Außerdem wurden dort die Mittelwerte der HOLMBERG-Parameter für die Sprechergruppen männlich und weiblich und für die Lautstärkegrade leise, normal und laut angegeben. Die in Kapitel 1 eingeführten

Wellenform-Parameter und die in Kapitel 4 eingeführten Analyseparameter sind in Anlehnung an die HOLMBERG-Parameter definiert worden. Zur Erzielung akzeptabler weiblicher und männlicher Stimmqualität wurden die Steuerparameter im GAI-Modell nach den Parameterwerten der Untersuchung von HOLMBERG et al. für weibliche und männliche Sprecher eingestellt.

Zunächst wurde der subglottale Druck nach den HOLMBERG-Daten eingestellt. Ferner wurden HOLMBERG-Zeitparameter direkt für die Einstellung der THH-Steuerfunktion benutzt und nicht mehr anhand des daraus resultierenden glottalen Volumenstroms nachgemessen (siehe Diskussion in Abs. 4.3.2). Die Einstellung der Amplitudenparameter geschieht im GAI-Modell anhand der Vorgabe der Steuerparameter dc_a_glott und ac_a_glott für die Einstellung der Steuerfunktion a_glott[t] (siehe Abs. 4.3.2). Allerdings wurde hier der äquivalente Amplitudensteuerparametersatz bestehend aus der Maximalöffnungsfläche und dem Verhältnis Verschlußfläche zu Maximalöffnungsfläche benutzt (siehe Tab. 5.1). Die Maximalöffnungsfläche a_glott_pk und das Verhältnis Verschlußfläche zu Maximalfläche a_glott_dc/pk führen wie folgt auf die nach Kapitel 4 definierten Steuerparameter:

(5.1) $\quad dc_a_glott = a_glott_pk * a_glott_dc/pk$

(5.2) $\quad ac_a_glott = a_glott_pk - dc_a_glott$

Die Einstellung dieser beiden Amplitudensteuerparameter wurde so vorgenommen, daß sich daraus bezüglich der HOLMBERG-Parameter eine Optimierung für vier Analyseparameter des glottalen Volumenstroms ergeben. Diese vier wesentlichen Amplitudenparameter sind der Maximalstrom (pkfl), der mittlere Strom (avfl), der Verschlußstrom (dcfl) und das Stromänderungsmaximum (mfdr) (siehe Tab. 5.1).

	Variable	Einheit	HOLM Wert	weiblich EINST. 1 Wert	weiblich EINST. 1 Fehler [%]	weiblich EINST. 2 Wert	weiblich EINST. 2 Fehler [%]	weiblich EINST. 3 Wert	weiblich EINST. 3 Fehler [%]	männlich HOLM Wert	männlich EINST. 1 Wert	männlich EINST. 1 Fehler [%]
VOREINSTELLUNG DER STEUERPARAMETER	a_glott_pk	[mm²]	/	7	/	7	/	7	/	/	12	/
	a_glott_dc/pk	[/]	/	0.4	/	0.38	/	0.5	/	/	0.25	/
	p_sub	[cm H_2O]	6.09	6.09	0	6.09	0	6.09	0	5.91	5.91	0
OPTIMIERTE PARAMETER - ANALYSE	F0	[Hz]	208	208	0	208	0	208	0	115	115	0
	clou	[/]	0.279	0.279	0	0.307	+10.0	0.307	+10.0	0.218	0.218	0
	qopu	[/]	0.735	0.735	0	0.735	0	0.735	0	0.609	0.609	0
	pkvk	[l/sec³]	0.226	0.225	- 0.4	0.225	- 0.4	0.222	- 1.8	0.354	0.365	+ 3.1
	avfl	[l/sec³]	0.139	0.136	- 2.2	0.137	- 1.4	0.151	+ 8.6	0.185	0.177	- 4.3
	dcfl	[l/sec³]	0.076	0.090	+18.4	0.090	+18.4	0.113	+48.7	0.101	0.093	- 7.9
	mfdr	[l/sec²]	117	142	+21.4	132	+12.8	110	- 6.0	239	230	- 3.8

Tab. 5.1: Einstellung der HOLMBERG-Parameter im GAI-Modell für die Synthese der weiblichen und der männlichen Stimme

Bei der Optimierung wurde versucht, die Abweichung dieser Analyseparameter von den HOLMBERG-Parametern in 10%-Schranken zu halten. Lediglich der Analyseparameter Verschlußstrom durfte vom äquivalenten HOLMBERG-Parameter des minimalen Stroms abweichen.[1]

[1] Grund hierfür ist die geringe perzeptive Relevanz des Verschlußstroms (siehe Exp. 3, Abs. 5.5.5).

Die Einstellung der Parameter für die männliche Stimme bietet keine Schwierigkeiten. Bei der Parametereinstellung der weiblichen Stimme ergibt sich bei einer exakten Voreinstellung der Zeitparameter genau nach den HOLMBERG-Daten das Problem, daß trotz eines Verschlußstroms, der bereits auf die obere 10%-Schranke eingestellt wurde, ein noch um 20% zu hohes Stromänderungsmaximum auftrat (siehe Tab. 5.1, Einstellung 1). Um aber die Unterschiede in der glottalen Wellenform zwischen weiblichen und männlichen Sprechern besser zu berücksichtigen, wurde für die weibliche Stimme in einer zweiten Einstellung der Schließquotient um 10% vergrößert (siehe Tab. 5.2, Einstellung 2). Dadurch wird eine Verlängerung der Schließphase und somit eine Symmetrisierung der glottalen Wellenform bewirkt. Dies führt zu einer weiteren Absenkung des Stromänderungsmaximums. Einstellung 2 ist bezüglich der Fehlerminimierung die optimale. Um jedoch die geschlechtsspezifischen Unterschiede bezüglich der Sprechlautstärke im Sinne der in den Daten von HOLMBERG et al. streckenden Tendenzen richtig zu berücksichtigen, wurde noch eine dritte Einstellung vorgenommen. Dabei weicht der Verschlußstrom stark von den HOLMBERG-Werten ab. Dies läßt sich aber, wie in Experiment 3 nachgewiesen wird, damit rechtfertigen, daß die Änderung des Verschlußstroms im GAI-Modell nur recht geringe perzeptive Änderungen nach sich zieht. Einstellung 3 gibt somit die in den HOLMBERG-Parametern abzulesenden Tendenzen in den geschlechtsspezifischen Unterschieden der glottalen Wellenform am besten wieder.

Diese Hauptunterschiede zwischen der weiblichen und der männlichen glottalen Wellenform sind nach HOLMBERG et al. folgende: Der Maximalstrom liegt bei Männern rund doppelt so hoch wie bei Frauen. Während der Verschlußstrom bei der weiblichen Stimme im Verhältnis zum Maximalstrom bei ca. 50% liegt, beträgt er bei der männlichen Stimme nur ca. 30%; d.h. im Verhältnis zum Maximalstrom liegt bei weiblichen Stimmen im Vergleich zu männlichen Stimmen ein annähernd doppelt so hoher Verschlußstrom vor. Die Offenzeit ist bezogen auf die Periodendauer beim Mann etwas kürzer als bei der Frau. Wesentlicher ist die Schließzeit, die beim Mann nur ein Drittel der Offenzeit, bei der Frau aber schon

fast die Hälfte der Offenzeit beträgt. Dieser Unterschied in der Schließzeit wird auch mit Symmetrisierung der glottalen Welle bei weiblichen Stimmen bezeichnet.

Aus den hier auftretenden Differenzen zwischen HOLMBERG-Parametern und den Analyseparametern des GAI-Modells darf aber nicht der Schluß gezogen werden, daß das GAI-Modell mit der THH-Steuerfunktion keine gute physikalische oder phonetische Beschreibung der weiblichen Stimme zuläßt. Vielmehr muß bedacht werden, daß die HOLMBERG-Parameter aus der Mittelung von Daten vieler verschiedener Sprecher eines Geschlechts berechnet wurden. Interindividuelle Mittelwerte von Wellenformparametern sind aber problematisch, da die glottalen Wellenformen auch innerhalb einer Geschlechtsgruppe von Sprecher zu Sprecher stark variieren können. Darum kann nicht gefolgert werden, daß die Mittelwerte der phonatorischen Parameter mit den phonatorischen Parametern eines durchschnittlichen Sprechers einer Geschlechtsgruppe identisch sind. Deshalb wird in Experiment 1 auch die Einstellung 3 für die weibliche Stimme benutzt, da es sinnvoller ist, die in den Daten vorhandenen Tendenzen zu modellieren, anstatt die HOLMBERG-Daten in dem vorliegenden Synthesemodell zu optimieren.

Die Einstellung der Synthesemodelle erfolgte anhand der Sprechtraktgeometrie des Schwa-Lautes, da die HOLMBERG-Parameter auch bei einer ähnlichen Sprechtraktgeometrie ermittelt wurden. Um einen ersten Eindruck vom Einfluß der Sprechtraktgeometrie auf die glottale Welle im GAI-Modell zu bekommen, wurde in einem hier nicht näher dokumentierten Test untersucht, wie stark die Analyseparameter, also die Form des glottalen Volumenstroms, bei fest vorgegebener Einstellung der Steuerfunktion a_glott[t] von der Änderung der Sprechtraktgeometrie beeinflußt werden. Es wurde die Steuerfunktion der weiblichen Stimme nach Einstellung 3 benutzt, wobei sich zeigt, daß auch bei extremen Sprechtraktformen, wie sie z.B. bei den Vokalen /u/ und /i/ vorliegen, Änderungen der Analyseparameter nur in einem Rahmen von 5% auftreten. Dieses Ergebnis besagt aber nicht, daß die Form der glottalen Welle in der Realität bei anderen Sprechtrakteinstellungen der hier vorgegebenen Einstellung nach dem Schwa-Laut entsprechen muß, da hier

die Steuerfunktion, die die glottale Öffnungsfläche vorgibt, ohne Berücksichtigung der mechanischen Interaktion konstant gehalten wurde. Der Test belegt nur die geringe Auswirkung der akustischen Interaktion auf die glottale Welle.

In Experiment 1 (siehe Abs. 5.2) wird nun überprüft, ob durch die hier vorgenommene Einstellung der phonatorischen Parameter für die weibliche und die männliche Stimme die Stimmqualität dergestalt beeinflußt wird, daß mit den Stimuli eine eindeutige Geschlechtsassoziation verbunden wird.

5.2 Zur Durchführung der Hörexperimente

In allen hier durchgeführten Experimenten (Exp. 1 bis 5) wurden die Signaldaten der Stimuli mittels der Algorithmen der Synthesemodelle per Computer errechnet.[2] Die generierten Signaldatenfiles können danach in ein analoges Signal gewandelt werden (Samplingfrequenz f_smp = 20 kHz für Exp. 1 bis Exp. 4).[3] Die Signale wurden dabei tiefpassgefiltert[4] und dann auf Tonband (REVOX A 77) zum Hörtest zusammengestellt und per Lautsprecher (ITT Synton S3-50) dargeboten. Die Tests wurden im SS 1988 und im WS 1988/1989 mit bezüglich der Sprachsynthese unerfahrenen Phonetikstudenten der Universität zu Köln durchgeführt, die an diesen Tests als Einzelpersonen oder in Gruppen von maximal fünf Personen teilnahmen.

---

[2] Die Algorithmen wurden vom Autor dieser Arbeit in BASIC nach den in Abs. 4.2 und Abs. 4.3 beschriebenen Ideen auf dem Computer ATARI ST entwickelt.

[3] Dazu dient das Programm ATARI-SONATA (ATARI-SONATA ist eine Weiterentwicklung des PDP-SONATA (siehe HILGER (1988, S. 126)) auf dem Computer ATARI ST). Das Format der in den Syntheseprogrammen generierten Signaldatenfiles entspricht dem Format der Signaldatenfiles des Programms ATARI-SONATA.

[4] Es wurde Tiefpassfilterung bei einer Grenzfrequenz von 3 kHz vorgenommen (siehe auch Abb. 6.1).

Die Stimuli wurden immer paarweise dargeboten, wobei alle Kombinationen einer Stimulusgruppe auftraten. Jedes Paar wurde dreimal hintereinander wiederholt, und die erfragte perzeptive Größe sollte in entsprechend vorbereitete Formulare eingetragen werden.

Während für Experiment 5 auch Diphthongfolgen synthetisiert wurden, war es für Experiment 1 bis Experiment 4, die sich mit den perzeptiven Auswirkungen der Änderung der phonatorischen Parameter beschäftigen, nur erforderlich, kurze Schwa-Laute konstanter Grundfrequenz zu erzeugen. Für die Synthese der männlichen Stimme wurde ein Sprechtrakt von 17.5 cm Länge (das entspricht 20 Traktabschnitten bei f_smp = 20 kHz), für die der weiblichen Stimme ein Sprechtrakt von 14.9 cm Länge (das entspricht 17 Traktabschnitten bei f_smp=20 kHz) angekoppelt. Für die Stimuli von Experiment 1 bis Experiment 4 wurde eine Stimuluslänge von nur 150 msec gewählt, damit der durch diese Synthese erzeugte stationäre Klang als natürlich gesprochener Sprachlaut aufgefaßt wird.

5.3 Der Einfluß der phonatorischen Parameter auf die Geschlechtsklassifikation und die Natürlichkeit: Experiment 1

5.3.1 Das Ziel des Experiments

Anhand von Experiment 1 soll gezeigt werden, daß durch systematische Änderungen des Stimmklangs allein, also nur durch die Änderung der phonatorischen Parameter, signifikante Änderungen bei der perzeptiven Geschlechtsklassifikation auftreten. Es werden zwei Stimulusgruppen generiert und im Test verglichen. In der GEV-ROS-Gruppe wird ohne Änderung der glottalen Wellenformparameter nur F0 und die Sprechtraktlänge für die Synthese der männlichen und der weiblichen Stimme variiert. In der GAI-HOLM-Gruppe werden für die Grundfrequenz und die Sprechtraktlänge die gleichen Werte gewählt, allerdings wird zusätzlich auf eine sorgfältige geschlechtsspezifische Einstellung der phonatorischen

Parameter geachtet.

5.3.2 Zur Synthese der Stimuli

Ausgangspunkt ist die Optimierung der phonatorischen Parameter für die Synthese der weiblichen und der männlichen Stimme nach HOLMBERG et al. (1988) (siehe Abs. 5.1). Zunächst wurde die weibliche Stimme (Einstellung 3) und die männliche Stimme (Einstellung 1) mit zugehöriger Sprechtraktlänge (siehe Abs. 5.2) synthetisiert. Diese Stimuli bilden den Stamm der GAI-HOLM-Stimulusgruppe (Synthese im GAI-Modell mit phonatorischen Parametern nach HOLMBERG et al.(1988)).

In der GEV-ROS-Stimulusgruppe (Synthese im GEV-Modell mit phonatorischen Parametern nach ROSENBERG[5]) wurden Stimuli mit den GAI-HOLM-Stimuli entsprechenden Grundfrequenz-, Formantfrequenz- und Intensitätswerten generiert, jedoch unter Nichtberücksichtigung bzw. Nichtänderung weiterer phonatorischer Parameter. Es wurde eine glottale Wellenform verwendet, die auch im bisher am IPKöln benutzten Synthesemodell Anwendung fand. Diese Wellenform basiert auf einer Untersuchung von ROSENBERG (1971). Resultat der Untersuchung von ROSENBERG ist ein Satz optimaler phonatorischer Zeitparameter für die Synthese in GEV-Modellen: Der Offenquotient beträgt danach 0.56 (opqu=0.56) und der Schließquotient 0.16 (clqu=0.16). Da in dem Modell von ROSENBERG kein Verschlußstrom auftritt (dcfl=0), wird durch den frei wählbaren Maximalstrom pkfl allein die Signalintensität bestimmt. Der Intensitätsabgleich auf die GAI-HOLM-Stimuli erfolgte durch den Abgleich des Stromänderungsmaximums mfdr. Insbesondere wurde diese glottale Wellenform ungeändert für die Synthese der männlichen und der

[5] Die phonatorischen Parameter nach ROSENBERG (1971) werden im folgenden beschrieben.

weiblichen Stimme der GEV-ROS-Gruppe benutzt (siehe Abb. 5.1[6]).

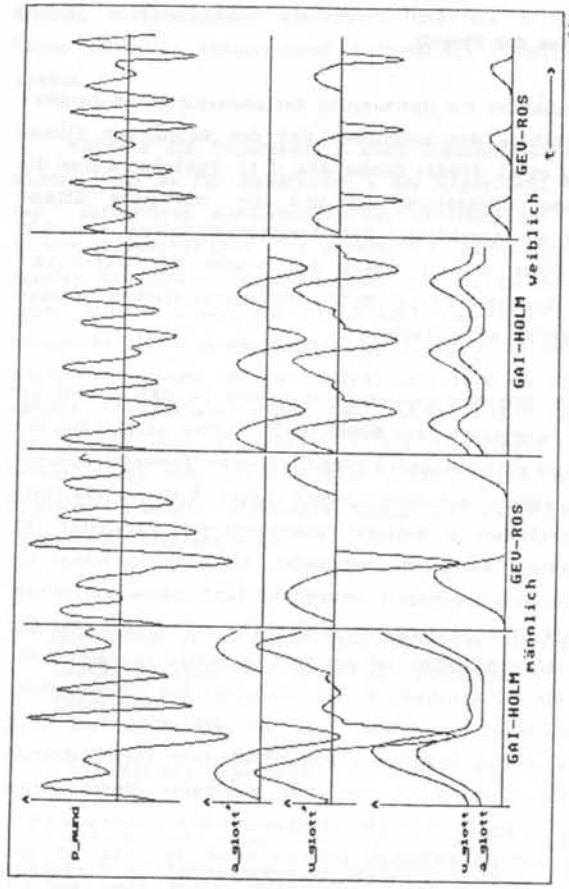


Abb. 5.1: Signaldarstellung der Stimuli zu Exp. 1

[6]Zu den im folgenden dargestellten Signalverläufen (siehe Abb. 5.1, Abb. 5.3, Abb. 5.5 und Abb. 5.6) ist anzumerken, daß die Amplituden- und Zeitskalierung für alle dort dargestellten Signalausschnitte identisch ist, so daß eine gute Vergleichbarkeit der Signalverläufe gegeben ist.

Aus den oben beschriebenen Stammstimuli ("männlich", "weiblich") jeder Stimulusgruppe wurden unter Wahrung der mit "hoch" und "tief" bezeichneten geschlechtsspezifischen Einstellungen der phonatorischen Parameter weitere Stimuli unter "Austausch" der mit "groß" und "klein" bezeichneten geschlechtsspezifischen Sprechtrakte generiert. So besteht also jede Stimulusgruppe aus den vier Stimuli S[kl/hoch], S[gr/hoch], S[kl/tief] und S[gr/tief]. Neben den "natürlichen" Kombinationen (oder auch Stammstimuli) S[kl/hoch] für die weibliche Stimme und S[gr/tief] für die männliche treten also noch die "unnatürlichen" Kombinationen S[kl/tief] und S[gr/hoch]. Abbildung 5.1 stellt allerdings nur die Signalverläufe der natürlichen Kombinationen dar.

5.3.3 Zur Durchführung des Experiments

Experiment 1 wurde mit 15 Testpersonen durchgeführt. Gefragt wurde nach dem Geschlecht des den Schwa-Laut produzierenden Sprechers. Es mußte für jeden Stimulus eine Entscheidung zwischen "männlich" oder "weiblich" getroffen werden ("forced-choice-situation"). Allerdings hatten die Testpersonen die Möglichkeit, jedem Stimulus zusätzlich das Attribut "unnatürlich" zu verleihen. Obwohl die Stimuli paarweise dargeboten wurden, konnten auch beide Stimuli als weiblich oder männlich klassifiziert werden. Paarweise verglichen wurden allerdings nur Stimuli einer Stimulusgruppe. Jedoch wurden in wechselnder Reihenfolge Stimuluspaare beider Gruppen dargeboten.

5.3.4 Ergebnisse

Da jeder Stimulus dreimal innerhalb des Hörtests auftritt, ist die Klassifikationsrate 45 das Maximum (siehe Abb. 5.2). In dem Fall wäre der entsprechende Stimulus von allen Testpersonen in allen Paaren immer dem gleichen Geschlecht zugeordnet worden. Hingegen ist die Unnatürlichkeitsrate ein Relativwert und wird in Prozent angegeben. Diese Prozentwerte geben an, wie oft ein Stimulus bezogen auf die Klassifikation zu einem bestimmten

Geschlecht das Attribut "unnatürlich" bekam. Die Werte sind in Tabelle 5.2 zusammengestellt und zusätzlich als Blockdiagramm dargestellt worden (siehe Abb. 5.2).

			STIMULUS			
			GR TIEF	KL TIEF	GR HOCH	KL HOCH
KLASSIFIKATIONS-RATE [ANZ.]	GEV-ROS	MÄNNLICH	45	43	32	8
		WEIBLICH	0	2	13	37
	GAI-HOLM	MÄNNLICH	45	45	23	3
		WEIBLICH	0	0	22	42
UNNATÜRLICHKEITS-RATE [%]	GEV-ROS	MÄNNLICH	16	60	72	88
		WEIBLICH	/	100	46	49
	GAI-HOLM	MÄNNLICH	7	49	57	67
		WEIBLICH	/	/	18	10

ST2

Tab. 5.2: Testergebnisse zu Exp. 1: Geschlechtsklassifikationsrate und Unnatürlichkeitsrate

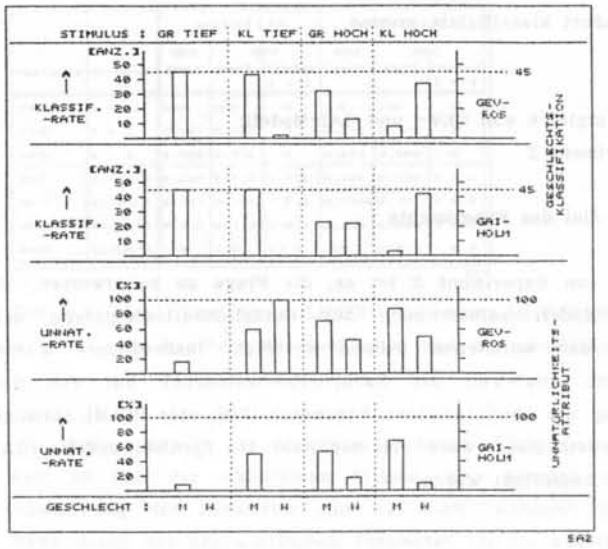


Abb. 5.2: Balkendiagramm zur Geschlechtsklassifikationsrate und zur Unnatürlichkeitsrate nach Exp. 1

Wichtigstes Ergebnis von Experiment 1 ist die Tendenz der stärkeren "weiblich"-Klassifikation der GAI-HOLM-Stimuli im Vergleich zu den GEV-ROS-Stimuli bei hoher Grundfrequenz, also bei den Stimuli S[gr/hoch] und S[kl/hoch]. Diese Verschiebung in der Geschlechtsklassifikation ist überzufällig. Für S[gr/hoch] ergibt sich ein Signifikanzniveau von 5%, für S[kl/hoch] von knapp 10%.[7] Diese Verschiebung kann auf die Einstellung der phonatorischen Parameter zurückgeführt werden, da sich die Stimuli beider Gruppen nur in der Form der glottalen Welle unterscheiden.

Interessant ist auch die wesentlich geringere Unnatürlichkeitsrate der GAI-HOLM-Stimuli gegenüber den GEV-ROS-Stimuli. Dies ist vor allem dann der Fall, wenn die Stimuli als von einer

[7] Die Berechnung der Signifikanzniveaus erfolgte nach dem 2I-Dach-Test von KULLBACK (siehe HELLER et al. (1974, S. 222)).

Frau geäußert klassifiziert wurden.

5.4 Der Vergleich von GEV- und GAI-Modell: Experiment 2

5.4.1 Das Ziel des Experiments

Ziel von Experiment 2 ist es, die Frage zu beantworten, ob das GAI-Modell selbst noch zur Natürlichkeitssteigerung der Synthese der weiblichen Stimme beiträgt. Insbesondere wurde abgeschätzt, inwieweit der Natürlichkeitseindruck nur von der Einstellung der phonatorischen Parameter (ROS oder HOLM) abhängt und inwieweit dieser auch von der Wahl des Synthesemodells (GAI oder GEV) beeinflußt wurde.

5.4.2 Zur Synthese der Stimuli

Neben den bereits aus Experiment 1 vorliegenden Stimuli zur Einstellung der HOLMBERG-Parameter im GAI-Modell und der ROSENBERG-Parameter im GEV-Modell ist es nun noch erforderlich, die Einstellung der phonatorischen Parameter nach HOLMBERG im GEV-Modell und die der Parameter nach ROSENBERG im GAI-Modell zu synthetisieren.

Bei der Einstellung der HOLMBERG-Parameter im GEV-Modell wurde nicht die Optimierung auf die HOLMBERG-Parameter selbst angestrebt, sondern eine Optimierung auf die im hier benutzten GAI-Modell eingestellten Parameter (siehe Abs. 5.1) vorgenommen; Denn in Experiment 2 soll der Einfluß der Modelle bei gleicher glottaler Welle untersucht werden. Die im GEV-Modell optimierten phonatorischen Parameter sind in Tabelle 5.3 zusammen mit den Werten im GAI-Modell angegeben.

Variable	Einheit	weiblich GAI Wert	weiblich GEV Wert	weiblich GEV Fehler [%]	männlich GAI Wert	männlich GEV Wert	männlich GEV Fehler [%]
F0	[Hz]	208	208	0	115	115	0
clqu	[/]	0.307	0.307	0	0.218	0.218	0
opqu	[/]	0.735	0.735	0	0.609	0.609	0
pkfl	[l/sec]	0.222	0.215	- 3.2	0.365	0.370	+ 1.2
avfl	[l/sec]	0.151	0.148	- 2.0	0.1766	0.177	+ 0.2
dcfl	[l/sec]	0.113	0.112	- 0.9	0.093	0.094	+ 0.6
mfdr	[l/sec^2]	110	112	+1.8	230	229	- 0.4

ST2

Tab. 5.3: Einstellung der HOLMBERG-Parameter im GEV-Modell für die Synthese der weiblichen und der männlichen Stimme

Bemerkenswert ist die generell bessere Optimierungsmöglichkeit im Fall der männlichen Stimme. Hier beträgt die mittlere Abweichung der Parameter nur ca. 0.5%, während die mittlere Abweichung der phonatorischen Parameter für die weibliche Stimme auf ca. 2% ansteigt. Da das Optimierungsmaximum hauptsächlich von den Differenzen in der Kurvenform zwischen Steuerfunktion und glottalem Volumenstrom bestimmt wird, ist diese Vergrößerung in der Abweichung eine Folge des Anstieges der akustischen Interaktion zwischen Glottis und Sprechtrakt aufgrund der Anhebung von F0. Trotzdem ist diese Abweichung in den Wellenformparametern von ca. 2% ebenfalls gering, so daß schon hier vermutet werden kann, daß der Wechsel zwischen den beiden Modellen keinen entscheidenden perzeptiven Unterschied bewirken wird.

Zur Übertragung der glottalen Wellenform nach ROSENBERG (1971) auf das GAI-Modell wurden die Zeitparameter direkt anhand der Steuerfunktion der glottalen Öffnungsfläche a_glott[t] eingestellt und die Glottis wurde während der Verschlußphase fast ganz geschlossen (siehe Abb. 5.3).[8] Das Maximum der glottalen Öffnungsfläche a_glott_pk wurde so eingestellt, daß alle Stimuli

---

[8] Da der Algorithmus des GAI-Modells eine Glottisfläche von 0 nicht zuläßt, wurde hier eine Restöffnungsfläche von unter 1 ‰ des Maximalstroms eingestellt.

den gleichen Wert des Stromänderungsmaximums (mfdr) und damit gleiche Intensität aufwiesen. Ein Randproblem stellt die durch die unterschiedliche mittlere Öffnungsweite der Glottis bedingte, unterschiedliche glottale Dämpfung im GAI-Modell dar. Glottale Dämpfung tritt im GEV-Modell durch die Glottis selbst nicht auf und wird durch die Teilwellenreflexion in einen angekoppelten nullten Traktabschnitt modelliert.[9] Mithilfe der Variation der Querschnittsfläche dieses Traktabschnitts A[0] muß demnach die Traktdämpfung im GAI-Modell immer auf den in allen Modellen gleichen Gesamtwert nachgeregelt werden, um in allen Modellen zu vergleichbarer Traktdämpfung zu führen.

So wurden insgesamt acht Stimuli für beide Parametersätze (HOLM und ROS) in beiden Modellen (GAI und GEV) für beide Geschlechtsgruppen synthetisiert (siehe Abb. 5.3). Da im Rahmen dieser Arbeit die Synthese der weiblichen Stimme von vorrangiger Bedeutung ist, wurde Experiment 2 mit den Stimuli der Synthese der weiblichen Stimme S[kl/hoch] durchgeführt.

[9] Siehe Abs. 4.2.2.

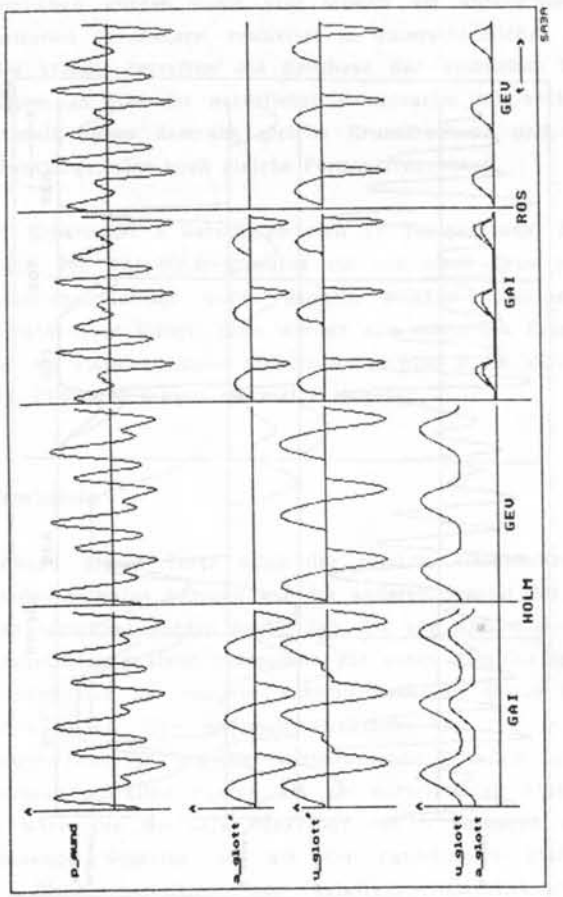


Abb. 5.3A: Signaldarstellung der Stimuli zu Exp. 2. Synthese der weiblichen Stimme

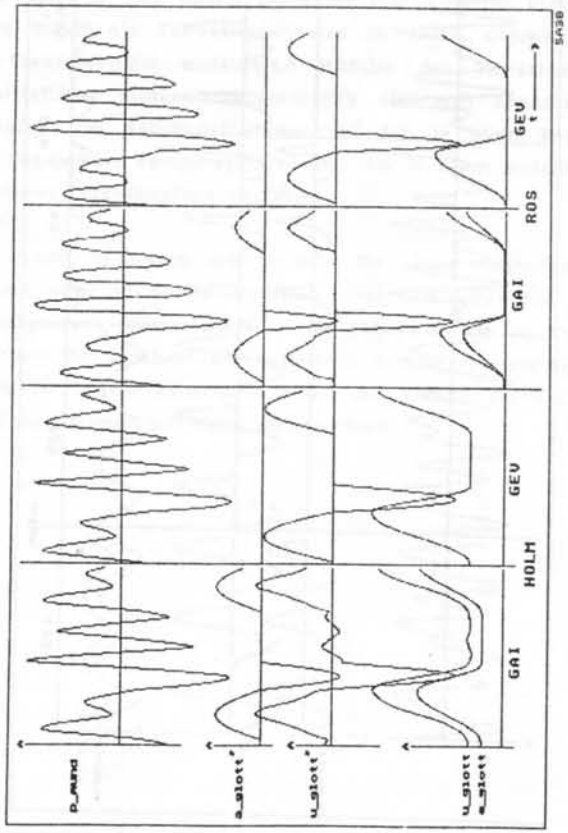


Abb. 5.3B: Signaldarstellung der Stimuli zu Exp. 2. Synthese der männlichen Stimme

5.4.3 Zur Durchführung des Experiments

Verglichen wurden somit vier Stimuli mit unterschiedlichen phonatorischen Parametern, realisiert in unterschiedlichen Modellen. Alle Stimuli betreffen die Synthese der weiblichen Stimme, sie entsprechen also der natürlichen Kombination S[kl/hoch]. Alle vier Stimuli haben demnach gleiche Grundfrequenz und gleiche Sprechtraktlänge, also auch gleiche Formantfrequenzen.

An Experiment 2 beteiligten sich 17 Testpersonen, die von vornherein den GAI-HOLM-Stimulus als von einer Frau geäußert klassifizierten. Gefragt wurde danach, welcher Stimulus eines Paares natürlicher klingt. Dazu wurden alle möglichen Paarkombinationen der vierelementigen Stimulusmenge {GAI-HOLM, GEV-HOLM, GAI-ROS, GEV-ROS} jeweils zweimal dargeboten.

5.4.4 Ergebnisse

Anhand dieses Tests kann der relative Natürlichkeitsgrad eines jeden Stimulus bezogen auf die anderen Stimuli der Stimulusmenge berechnet werden (siehe Tab. 5.4 und Abb. 5.4). Im Fall einer Entscheidung einer Testperson für einen Stimulus in einem Paar erhöht sich der relative Natürlichkeitsgrad dieses Stimulus um einen Punkt. Eine maximale Punktzahl von 102 in diesem Experiment bedeutet, daß der entsprechende Stimulus von allen Testpersonen in allen Paaren als am natürlichsten klassifiziert worden wäre. Die minimale Punktzahl von 0 bedeutet, daß der entsprechende Stimulus nie als der natürlichere klassifiziert wurde. Diese Definition des relativen Natürlichkeitsgrades bezieht sich aber nur auf die Natürlichkeit eines Stimulus in Relation zur Natürlichkeit aller anderen Stimuli der getesteten Stimulusgruppe. Mithilfe der Größe der kritischen Differenz[10] (d_krit) kann in bezug auf ein vorgegebenes Signifikanzniveau

[10] "critical difference"; Zur Berechnung siehe DAVID (1969, S.37 und S. 111).

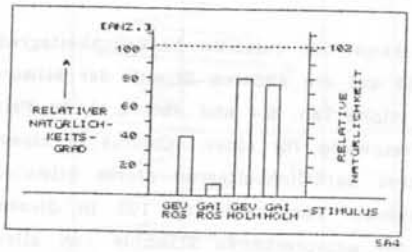
ermittelt werden, innerhalb welcher Intervalle des relativen Natürlichkeitsgrades zwei Stimuli als perzeptiv unterscheidbar abgelehnt werden müssen.

In Experiment 2 ergibt sich d_krit=15 bei einem Signifikanzniveau von 1%, so daß die Stimuli GEV-HOLM und GAI-HOLM als perzeptiv unterscheidbar abgelehnt werden müssen.

	STIMULUS			
	GEV-ROS	GAI-ROS	GEV-HOLM	GAI-HOLM
RELATIVER NATÜRLICHKEITSGRAD [ANZ.]	40	8	80	76

ST4

Tab. 5.4: Ergebnisse zu Exp. 2: Der relative Natürlichkeitsgrad



SA4

Abb. 5.4: Balkendiagramm zum relativen Natürlichkeitsgrad nach Exp. 2

Hauptergebnis von Experiment 2 ist demnach, daß die Einstellung der phonatorischen Parameter entscheidender als die Auswahl zwischen einem Modell mit oder ohne akustische Interaktion ist. Es ergeben sich kaum Unterschiede zwischen den gemessenen relativen Natürlichkeitsgraden der Stimuli beider Modelle bei der Einstellung der phonatorischen Parameter nach HOLMBERG. Einen wesentlich geringeren relativen Natürlichkeitsgrad hingegen weisen die Stimuli mit der Einstellung der phonatorischen Parameter

nach ROSENBERG auf. Allerdings wurde diese Einstellung im GEV-Modell noch als natürlicher empfunden als die Verwirklichung der entsprechenden Einstellung im GAI-Modell. Der Grund hierfür ist wahrscheinlich die Notwendigkeit einer verbleibenden Restöffnungsfläche der Glottis im GAI-Modell, wie es auch in der Realität der Fall ist.

5.5 Die Änderung der phonatorischen Parameter Verschlußstrom und Schließquotient: Experiment 3 und Experiment 4

5.5.1 Das Ziel der Experimente

Verschlußstrom und Schließquotient sind die entscheidenden phonatorischen Wellenformparameter, die zur Differenzierung der Stimmqualität zwischen Mann und Frau führen.[11] Darum ist es von grundsätzlichem Interesse, die perzeptiven Auswirkungen der systematischen Änderung dieser beiden Parameter in den hier benutzten Modellen zu studieren. In Experiment 3 wird die Änderung des Verschlußstroms (dcfl) und in Experiment 4 die Änderung des Schließquotienten (clqu) betrachtet.

Trotz der geringen perzeptiven Unterschiede zwischen GEV- und GAI-Modell nach Experiment 2, werden diese Parameteränderungen am GAI-Modell durchgeführt, da dieses Modell das zur Generierung zeitveränderlicher Signale relevante ist.[12] Die Verwendung des GAI-Modells ist auch deshalb sinnvoll, weil die Verschlußstromänderung im GEV-Modell keine akustischen Auswirkungen nach sich zieht.

Ausgangspunkt für die systematischen Parameteränderungen ist die Einstellung der phonatorischen Parameter nach HOLMBERG. Von dieser Einstellung ausgehend wird der zu untersuchende Parameter

---

[11] Siehe Abs. 1.3.

[12] Siehe auch Abs. 7.1.

variiert. Die anderen Parameter sollen dabei möglichst unverändert bleiben. Jedoch zeigt sich, daß bei der Änderung des Offenquotienten weitere Parameter geändert werden müssen, um die Stimuli zumindest bezüglich der Intensität vergleichbar zu machen.

5.5.2 Zur Synthese der Stimuli für Experiment 3: Die Änderung des Verschlußstroms

Ausgehend von der Einstellung der phonatorischen Parameter nach HOLMBERG wird hier nur die Höhe des Verschlußstroms geändert. Dabei bleibt die Wechselstromkomponente der Steuerfunktion a_glott[t] völlig unberührt; Es wird bei sonst gleichbleibenden Steuerparametern nur der Verschlußwert der Steuerfunktion dc_a_glott geändert. Betrachtet man den Verschlußwert der Steuerfunktion als auf den nach HOLMBERG optimierten Wert dc_a_glott(HOLM) normiert, so wurden Stimuli mit dem relativen Verschlußwert

dc_a_glott / dc_a_glott(HOLM) = 0, 1/4, 1/2, 1 und 3/2

generiert. Diese Stimuli werden mit ansteigendem Verschlußwert mit DC01 bis DC05 bezeichnet (siehe Tab. 5.5). Der Stimulus DC04 mit relativem Verschlußwert 1 entspricht der Einstellung nach HOLMBERG.

Variable	Einheit	DC01	DC02	DC03	DC04 HOLM	DC05
dc_a_glott / dc_a_glott(HOLM)	[/]	0	1/4	1/2	1	3/2
AE03	[cm²]	50	40	30	20	0
pkf1	[1/sec]	0.111	0.139	0.167	0.222	0.277
avf1	[1/sec]	0.0396	0.0674	0.0953	0.151	0.206
dcf1	[1/sec]	$3.3 \cdot 10^{-8}$	0.0277	0.0569	0.113	0.170
Afdr	[1/sec²]	116	114	113	110	107
acf1	[1/sec]	0.111	0.111	0.110	0.109	0.107
acf1/dcf1	[/]	$3.36 \cdot 10^{+6}$	4.01	1.957	0.959	0.633
dcf1/pkf1	[/]	$2.98 \cdot 10^{-7}$	0.200	0.338	0.511	0.613

STS

Tab. 5.5: Einstellungen der phonatorischen Parameter der Stimuli aus Exp. 3: Variation des Verschlußstroms

Weitere Steuerparameter mußten nicht geändert werden. Wegen der Invarianz des Wechselstromanteils bleibt insbesondere das Stromänderungsmaximum und damit die Intensität der generierten Stimuli konstant.

Durch die drastisch geänderte mittlere glottale Öffnungsfläche ändert sich aber die glottale Dämpfung und damit der Gesamtbetrag der Sprechtraktdämpfung stark. Der Steuerparameter a_glott_dc hat also starken Einfluß auf die Bandbreite der Formanten. Allerdings muß die Sprechtraktdämpfung für alle Stimuli gleich sein, um eine gute Vergleichbarkeit zu gewährleisten. Ziel des Experiments ist es ja, nur Unterschiede bezüglich der Phonation zu untersuchen. Im GEV-Modell hingegen trägt das Glottismodell selbst nicht zur Dämpfung bei. Hier wird die gesamte Sprechtraktdämpfung von der Rückstreuung in den Traktabschnitt 0 übernommen. Das Problem wird hier dadurch gelöst, daß die Querschnittsfläche von Traktabschnitt 0 im GAI-Modell je nach der Stärke der Dämpfung durch die Glottis selbst so eingestellt wird, daß beide Dämpfungsmechanismen zusammen zu einer für alle Stimuli gleichbleibenden Gesamtdämpfung des Sprechtrakts führen.[13] Diese Einstellung der Querschnittsfläche von Traktabschnitt 0 (A[0]) wurde empirisch anhand der Betrachtung des

[13] Siehe auch Abs. 4.2.2.

Ausschwingverhaltens des Modells eingestellt. Dies führt jedoch zu einer Begrenzung des relativen Verschlußwertes auf 1.5, da in diesem Fall die gesamte Sprechtraktdämpfung allein von der Glottis des GAI-Modells bewirkt wird.

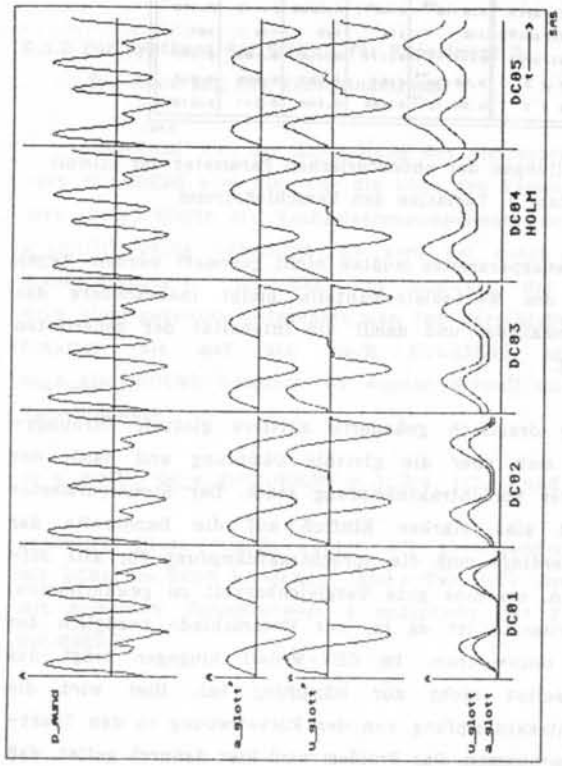


Abb. 5.5: Signaldarstellung der Stimuli zu Exp. 3:
Variation des Verschlußstroms

Abbildung 5.5 zeigt, daß in der Zeitfunktion der Ableitung des glottalen Volumenstroms u_glott' nur geringe Unterschiede auszumachen sind. Allerdings erkennt man, wie bei zunehmendem Verschlußstrom mehr und mehr Überlagerungsschwingungen ("ripple")

aus der Offenphase in die Verschlußphase übergehen. Da außer diesen durch die akustische Interaktion bewirkten Signalunterschieden keine weiteren Unterschiede auftreten, kann vermutet werden, daß sich wie in Experiment 2 auch in Experiment 3 kaum Unterschiede bezüglich der Perzeption der Stimuli feststellen lassen.

5.5.3 Zur Synthese der Stimuli für Experiment 4: Die Änderung des Schließquotienten

Der Schließquotient (clqu) wird nur bei konstantem Offenquotient (opqu) variiert. Auch hier wird von der Einstellung der Steuerparameter nach HOLMBERG ausgegangen. Parallel zum relativen Verschlußwert wird hier ein relativer Schließquotient clqu/clqu(HOLM) definiert, bei dem der Wert auf den nach HOLMBERG optimierten Schließquotienten normiert ist. Es wurden die Stimuli CL01 BIS CL06 synthetisiert, wobei der relative Schließquotient von 1/3 ausgehend in äquidistanter Schrittfolge von +1/3 bis zum Wert 2 erhöht wurde (siehe Tab. 5.6).

clqu / clqu(HOLM) = 1/3, 2/3, 1, 4/3, 5/3 und 2

Der Stimulus CL03 mit dem relativen Schließquotienten von 1 entspricht der Synthese mit den nach HOLMBERG optimierten phonatorischen Steuerparametern.

Variable	Einheit	CL01	CL02	CL03 HOLM	CL04	CL05	CL06
clqu / clqu(HOLM)	E / 3	1/3	2/3	1	4/3	5/3	2
a_glott_pk	[mm²]	5	6	7	7.25	6	5
a_glott_dc/pk	E / 3	0.75	0.575	0.5	0.45	0.575	0.7
opqu	E / 3	0.732	0.732	0.732	0.732	0.732	0.732
clqu	E / 3	0.103	0.206	0.309	0.412	0.515	0.609
spqu	E / 3	6.1	2.55	1.37	0.775	0.42	0.283
mfdr	El/sec²]	-111	-113	-109	-109	+105	+113
dcfl	El/sec]	0.123	0.112	.0113	0.106	0.112	0.114
avfl	El/sec]	0.133	0.139	0.151	0.149	0.139	0.127
pkfl	El/sec]	0.156	0.189	0.222	0.226	0.183	0.153
acfl	El/sec]	0.0331	0.0766	0.109	0.200	0.0708	0.0591
acfl/dcfl	E / 3	0.274	0.685	0.959	1.13	0.631	0.345
dcfl/pkfl	E / 3	0.787	0.594	0.511	0.469	0.613	0.744

ST6

Tab. 5.6: Einstellungen der phonatorische Parameter der Stimuli aus Exp. 4: Variation des Schließquotienten

Wird der relative Schließquotient größer als 4/3, so tritt der physikalisch unrealistische Fall ein, daß sich die Stimmlippen mit höherer Geschwindigkeit öffnen, als sie sich schließen. Insbesondere liegt jetzt das Stromänderungsmaximum in der Öffnungsphase. Darum wurde entgegen der üblichen Definition[14] das Stromänderungsmaximum mfdr in Tabelle 5.6 mit Vorzeichen angegeben.

Werden neben dem Schließquotienten alle weiteren Steuerparameter konstant gehalten, so treten beim Vergleich der Stimuli große Intensitätsunterschiede auf. Diese Unterschiede dürfen aber nicht durch ein willkürliches Auspegeln der abgestrahlten Druckwelle p_mund ausgeglichen werden, da die im GAI-Modell berechneten Absolutwerte des Volumenstroms physikalisch relevant sind. Die Wiedergabelautstärke der berechneten abgestrahlten Druckwelle p_mund muß daher fest eingestellt bleiben. Sie ist anhand des HOLMBERG-Parametersatzes für die normal-laute Stimme festgelegt worden. Die Änderung der Lautstärke darf im GAI-Modell wie bei

[14] Siehe Abs. 1.2.3.

der natürlichen Spracherzeugung nur über die Variation der Intensität der glottalen Welle, bzw. über die Variation des Stromänderungsmaximums mfdr erfolgen. Darum muß der Intensitätsabgleich mithilfe der Steuerparameter selbst durchgeführt werden. Hauptsächlich wurde dies durch die Variation des Maximalwerts a_glott_pk erreicht. Es wurde darauf geachtet, daß der mittlere Strom annähernd konstant gehalten wurde, um ein Gleichbleiben der glottalen Dämpfung zu gewährleisten.

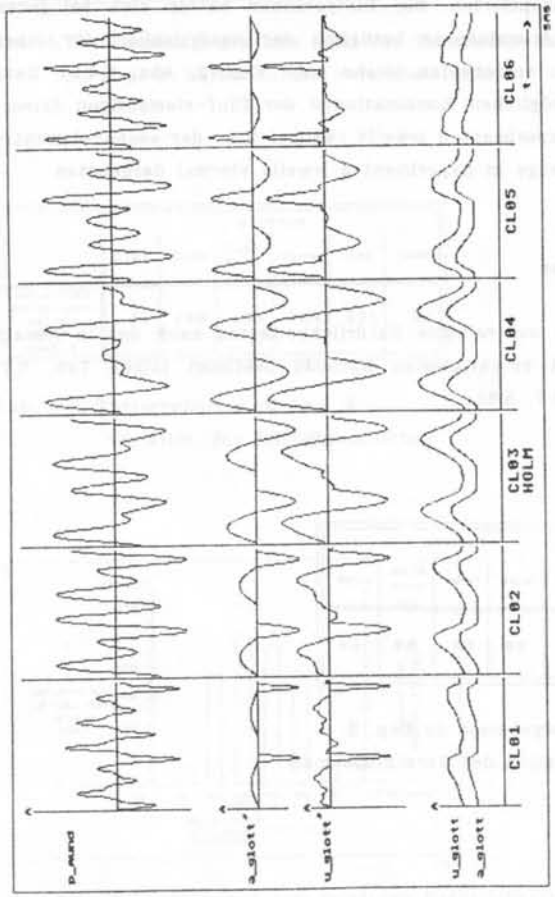


Abb. 5.6: Signaldarstellung der Stimuli zu Exp. 4: Variation des Schließquotienten

5.5.4 Zur Durchführung der Experimente

Die Stimuli zur Variation des Verschlußstroms DC01 bis DC05 und zur Variation des Schließquotienten CL01 bis CL06 wurden für die weibliche Stimme, d.h. für die natürliche Kombination S[kl/hoch] synthetisiert. An beiden Tests nahmen 12 Testpersonen teil, die alle den Stimulus mit der nach HOLMBERG optimierten Einstellung der phonatorischen Parameter als von einer Frau gesprochen akzeptierten. Die Testpersonen hatten sich bei jedem dargebotenen Stimuluspaar bezüglich der Natürlichkeit für einen der Stimuli zu entscheiden (siehe auch Exp. 2, Abs. 5.4.2). Dazu wurden alle möglichen Kombinationen der fünf-elementigen Stimulusmenge in Experiment 3 jeweils zweimal bzw. der sechs-elementigen Stimulusmenge in Experiment 4 jeweils viermal dargeboten.

5.5.5 Ergebnisse

Es wurde der relative Natürlichkeitsgrad nach der in Absatz 5.4.3 (Exp. 2) beschriebenen Methode bestimmt (siehe Tab. 5.7, 5.8 und Abb. 5.7, 5.8).

	STIMULUS				
	DC01	DC02	DC03	DC04 HOLM	DC05
RELATIVER NATÜRLICH-KEITS-GRAD EANZ.3	0	58	59	59	64

ST7

Tab. 5.7: Testergebnisse zu Exp. 3:
Variation des Verschlußstroms

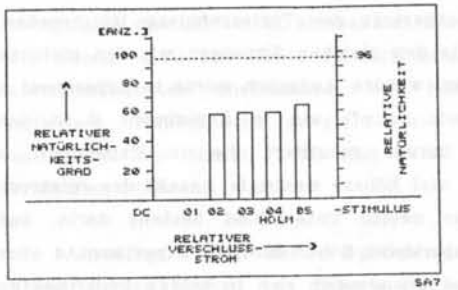


Abb. 5.7: Balkendiagramm zum relativen Natürlichkeitsgrad nach Exp. 3: Variation des Verschlußstroms

	STIMULUS					
	CL01	CL02	CL03 HOLM	CL04	CL05	CL06
RELATIVER NATÜRLICH-KEITS-GRAD EANZ.3	21	148	205	184	131	31

ST8

Tab. 5.8: Testergebnisse zu Exp. 4:
Variation des Schließquotienten

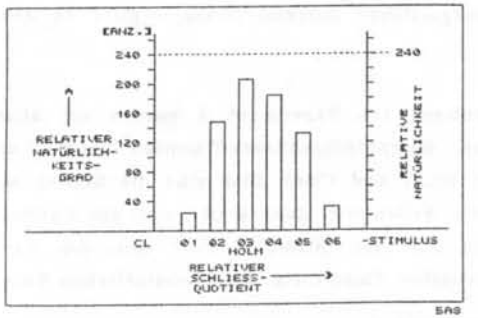


Abb. 5.8: Balkendiagramm zum relativen Natürlichkeitsgrad nach Exp. 4: Variation des Schließquotienten

Eine gute Vergleichbarkeit der Testergebnisse ist gegeben, da beide Tests in jeweils den gleichen Sitzungen mit den gleichen Testpersonen durchgeführt wurden. Lediglich wurde in Experiment 4 jede Kombination doppelt so oft wie in Experiment 3 in den Hörtest aufgenommen. Daraus resultiert die in Experiment 4 gegenüber Experiment 3 viel höhere maximale Anzahl des relativen Natürlichkeitsgrades. Der zweite Unterschied besteht darin, daß die Stimulusmenge in Experiment 3 nur fünf, in Experiment 4 aber sechs Stimuli umfaßt. Deshalb ergeben sich in beiden Experimenten auch unterschiedliche kritische Differenzen[15] (d_krit):

d_krit = 14 in Exp. 3 (Signifikanzniveau = 1%),
d_krit = 21 in Exp. 4 (Signifikanzniveau = 1%).

Somit sind nach Experiment 3 das Stimulustripel DC02 bis DC04 und nach Experiment 4 die Stimuluspaare CL03 und CL04, CL02 und CL05 sowie das Paar CL01 und CL06 als perzeptiv unterscheidbar auf einem Signifikanzniveau von 1% abzulehnen.

Das Hauptergebnis von Experiment 3 ist somit, daß der Stimulus DC01 (kein Verschlußstrom) als sehr unnatürlich gegenüber den anderen Stimuli empfunden wird, während die Größe des Verschlußstroms - so lange dieser größer als Null ist - keine perzeptiven Auswirkungen hat. Die Stimuli DC02 bis DC04 werden von vielen Testpersonen als nicht unterscheidbar empfunden, obwohl der Überlagerungseffekt[16] zunimmt (siehe u_glott' in Abb. 5.5).

Nach den Ergebnissen von Experiment 4 werden vor allem Stimuli mit extremen Geschwindigkeitskoeffizienten (spqu) als unnatürlich abgelehnt (CL01 und Cl06). Dies sind die Stimuli mit stark antisymmetrischer Wellenform. Zwar ergibt sich ein maximaler Natürlichkeitsgrad für den Stimulus CL03 mit der nach HOLMBERG et al. optimierten Einstellung der phonatorischen Parameter, aber das darf nicht überschätzt werden, da neben CL03 auch die Stimuli CL02 und CL04, die sich durch ähnliche Geschwindigkeitskoeffizienten auszeichnen, ein hohes Natürlichkeitsniveau aufweisen. Insbesondere ist die Distanz der relativen Natürlichkeitsgrade der Stimuli DC03 und DC04 gerade noch gleich der kritischen Differenz.

[15] Siehe Abs. 5.4.4.

[16] Siehe Abs. 4.3.2.3.

Allerdings zeigt der Vergleich der als recht natürlich klassifizierten Stimuli von Experiment 3 und von Experiment 4, daß Änderungen im Schließquotienten viel empfindlicher wahrgenommen werden als Änderungen in der Verschlußstromhöhe.

6 Eine experimentelle Untersuchung zum Einfluß der Geschlechtsklassifikation auf die Vokalqualität: Experiment 5

6.1 Das Ziel des Experiments

In Experiment 5 soll nachgewiesen werden, daß die mit einem Stimulus verbundene Geschlechtsassoziation Einfluß auf die Vokalqualitätsklassifikation hat. Zwischen zwei Hörergruppen, die mit demselben Stimulus einen weiblichen bzw. einen männlichen Sprecher assoziieren, werden in Experiment 5 signifikante Differenzen in der Vokalqualitätsklassifikation festgestellt. Ziel dieses Experiments ist es somit, die in Kapitel 3 aufgestellte Theorie zur Primärnormierung zu bekräftigen. Dazu sind drei Bedingungen an die Teststimuli zu stellen.

1) Es muß gewährleistet sein, daß die Formanten bzw. die Form der spektralen Einhüllenden und die Grundfrequenz exakt kontrollierbar sind. Eine solche exakte Kontrolle ist bei synthetischer Spracherzeugung in der Regel gegeben.

2) Insbesondere müssen die Formantfrequenzen bzw. die Form der spektralen Einhüllenden unabhängig von der Grundfrequenz sein. Die Konstanz der Formantfrequenzen unter Variation der Grundfrequenz ist wesentliche Voraussetzung für dieses Experiment. Diese Voraussetzung ist insbesondere im GEV-Modell erfüllt. Da in diesem Modell keine Kopplung zwischen Glottis und Sprechtrakt vorliegt, wird die Übertragungsfunktion des Sprechtrakts von Grundfrequenzänderungen nicht beeinflußt. Ebenso tritt keine Beeinflussung der glottalen Wellenform und insbesondere keine Beeinflussung von F0 durch Änderungen der Sprechtraktgeometrie auf.

3) Die Stimuli sollten offen bezüglich der möglichen Geschlechtsassoziation sein. Da es das Ziel des Experiments ist, anhand *eines* Stimulus Differenzen in der Vokalqualität nachzuweisen, sollten sich zwei möglichst gleichstarke Hörergruppen bezüglich der mit diesem Stimulus verbundenen Assoziation des

Sprechergeschlechts als "männlich" oder als "weiblich" ergeben. Dies wird in Experiment 5 dadurch erreicht, daß bei Erhöhung der Grundfrequenz die nach ROSENBERG (1971) eingestellten phonatorischen Parameter ungeändert bleiben. Somit führt die Nichtberücksichtigung der parallel zur Grundfrequenzänderung notwendigen Stimmklangmodulation zu der gewünschten Offenheit der Stimuli bezüglich der Geschlechtsassoziation.

6.2 Zur Synthese der Stimuli

Es wurden vier Stimuli des Logatoms [ɛɪɛ] im GEV-Modell synthetisiert. Die Einstellung der phonatorischen Parameter, die nach den Werten von ROSENBERG (1971) vorgenommen wurde, ist für alle Stimuli bis auf die Grundfrequenz gleich. Es wurden hier Grundfrequenzwerte von 130 Hz und von 220 Hz gewählt. Im folgenden werden diese beiden Einstellungen der phonatorischen Parameter, die sich aber nur bezüglich der Grundfrequenz unterscheiden, mit "tief" bzw. "hoch" bezeichnet. Die zur Steuerung der artikulatorischen Parameter nötigen, aus der Folge von Sprechtraktgeometrien resultierenden Sätze an Querschnittsflächen für die Traktabschnitte wurde mithilfe des Programms M4[1] auf dem Großrechner PDP11/73 generiert und mittels Filetransfer auf den Computer ATARI ST übertragen und dort von dem zur Berechnung der Stimuli nach dem GEV-Modell geschriebenen Simulationsprogramm[2] als Steuerdaten benutzt. Alle vier Stimuli wurden mittels desselben artikulatorischen Steuerparametersatzes erstellt. Somit liegen allen Stimuli dieselben Sprechtraktformen als Funktion der

---

[1] Siehe GREISBACH (1986).

[2] Wegen des großen Umfangs dieses Syntheseprogramms, das die Generierung zeitveränderlicher Signale erlaubt, wurde die Software vom Autor dieser Arbeit (in Anlehnung an das in FORTRAN geschriebene Syntheseprogramm auf PDP11/73 (siehe HILGER 1988, S. 123)) in der Programmiersprache C auf dem Computer ATARI ST entwickelt.

Zeit zugrunde. Geändert wurde allerdings die mittlere Sprechtraktlänge, um so die Sprechtrakte eines weiblichen und eines männlichen Sprechers in der linearen Näherung[3] zu modellieren.

Während in Experiment 1 zur Variation der Sprechtraktlänge die Anzahl der den gesamten Sprechtrakt bildenden Traktabschnitte variiert wurde, wird hier ein anderer Weg eingeschlagen. Da die Länge eines Traktabschnitts von der Samplingfrequenz f_smp abhängt, kann durch das Heraufsetzen der Samplingfrequenz die Länge eines Traktabschnitts und damit bei gleichbleibender Gesamtzahl der Traktabschnitte die Sprechtraktlänge herabgesetzt werden.[4] Da die Verringerung der Gesamtzahl der Traktabschnitte mit einem ungenaueren "Abtasten" der Sprechtraktgeometrie verbunden ist, bleibt die Invarianz der Sprechtraktform unter Sprechtraktlängenänderung am besten durch die Längenänderung der Traktabschnitte gewahrt. Es werden Stimuli für die Samplingfrequenzen f_smp = 20.000 kHz und f_smp = 23.255 kHz generiert. Das entspricht Traktabschnittlängen von 0.875 cm bzw. von 0.753 cm (siehe Gl. 4.4) oder einer Sprechtraktlänge von 17.5 cm bzw. 15.05 cm bei 20 Traktabschnitten. Im folgenden werden diese beiden artikulatorischen Einstellungen mit "groß" für den im Mittel 17.5 cm langen und mit "klein" für den um 14% kürzeren Sprechtrakt bezeichnet.

Aus dem Sonagramm (siehe Abb. 6.1) kann der Verlauf der ersten drei Formanten als Funktion der Zeit abgelesen werden.[5] Die Formantfrequenzwerte variieren vom [ɑ] mit F1 = 750 Hz und F2 = 1400 Hz bis zum [ɪ] mit F1 = 300 Hz und F2 = 2300 Hz. Im Hörtest wurde allerdings nur die Klangfarbe des letzten [ɑ] des Stimulus [ʃɑrɑ] erfragt. Die Hauptfunktion der mit der Zeit variierenden Formanten lag vor allem darin, sicherzustellen, daß die Stimuli von den Testpersonen auch als Sprachsignale akzep-

[3] Siehe Abs. 3.4.3.

[4] Siehe Abs. 4.2.1.3.

[5] Um die analoge Tiefpassfilterung (siehe Abs. 5.2) zu berücksichtigen, wird das Sonagramm vom analogen Signal erstellt.

tiert wurden.

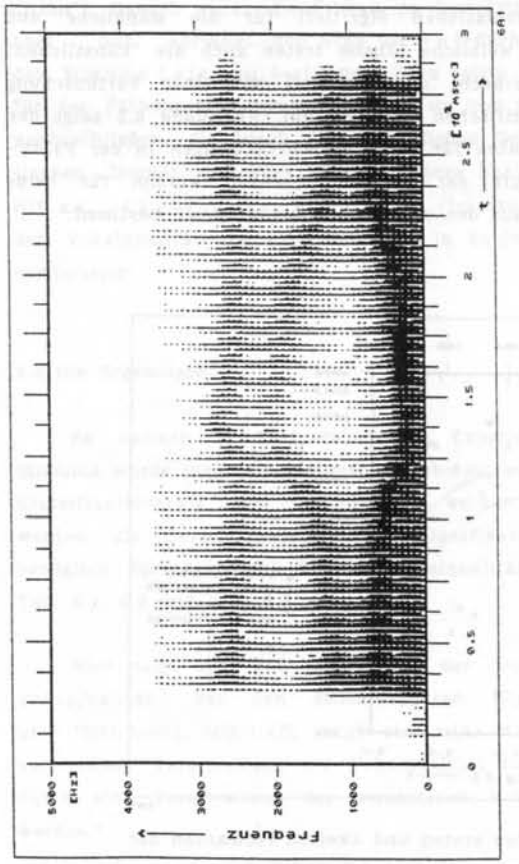


Abb. 6.1: Sonagramm des Logatoms [ʃɪd] für den Stimulus S[gr/tief] aus Exp. 5. Digitales Sonagramm[6] des mit 10 kHz zurückgesampelten analogen Signals

[6] auf der Basis von 128-Punkte-FFTs (Fast Fourier Transform) mit einer Auflösung von 70 dB.

Für den Hörtest zu Experiment 5 wurde eine vierelementige Stimulusmenge mit allen möglichen Kombinationen der phonatorischen und artikulatorischen Parameter generiert. Neben den "natürlichen" Kombinationen S[gr/tief] für die männliche und S[kl/hoch] für die weibliche Stimme treten auch die "künstlichen" Kombinationen S[gr/hoch] und S[kl/tief] mit einer Vertauschung der geschlechtsspezifischen Merkmale auf. Abbildung 6.2 zeigt den Verlauf der Formanten für beide Sprechtraktlängen in der F1/F2-Ebene. Die Verläufe der Formantfrequenzen wurden für beide Sprechtraktlängen aus den zugehörigen Sonagrammen bestimmt.

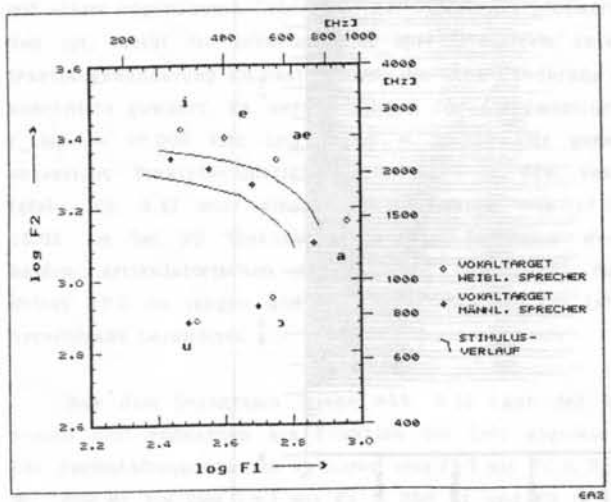


Abb. 6.2: Verlauf des ersten und zweiten Formanten der Stimuli aus Exp. 5 in der F1/F2-Ebene

6.3 Zur Durchführung des Experiments

Die Stimuli wurden paarweise dargeboten. Dabei treten alle möglichen Kombinationen der in der Stimulusmenge enthaltenen Elemente als Stimuluspaare auf. Bei der Darbietung jedes Paares

war zunächst eine Geschlechtsklassifikation vorzunehmen, wobei der Hörer jeden Stimulus ohne Alternative dem Sprechergeschlecht "weiblich" oder "männlich" zuordnen mußte ("forced-choice-condition"). Danach waren die Stimuli zu transkribieren. Dabei wurde aber nur verlangt, den Grad der [ɔ]-Färbung im letzten [ɐ] des Stimulus [ɐɾɐ] zu bestimmen. Dazu sollte eine einfache Skala für den Färbungsgrad benutzt werden, um von vornherein Probleme auszuschließen, die durch unterschiedlichen Gebrauch von diakritischen Zeichen auftreten könnten. Diese Skala besteht aus den mit ɐɐ , ɐɔ , ɔɐ und ɔɔ bezeichneten vier Stufen, die die Strecke der Vokalqualität von [ɐ] bis [ɔ] in äquidistanten Abständen symbolisiert.

6.4 Die Ergebnisse

Es nahmen 33 Testpersonen an Experiment 5 teil. Jeder Stimulus wurde insgesamt dreimal dargeboten, so daß eine maximale Klassifikationsrate von 99 erreicht werden konnte. Zunächst werden die Gesamtergebnisse der Klassfikation jedes Stimulus bezüglich Sprechergeschlecht und Vokalqualität dargestellt (siehe Tab. 6.1, 6.2 und Abb. 6.3, 6.4).

Hier zeigt sich gut der Einfluß der Grundfrequenz auf die Vokalqualität.[7] Bei den Stimuluspaaren {S[gr/hoch], S[gr/tief]} und {S[kl/hoch], S[lk/tief]} zeigt sich eine mit der F0-Erhöhung verbundene Verschiebung der Vokalqualität zum [ɔ]. Dies kann durch eine Verschiebung der perzeptiven Vokalschablone erklärt werden.[8]

---

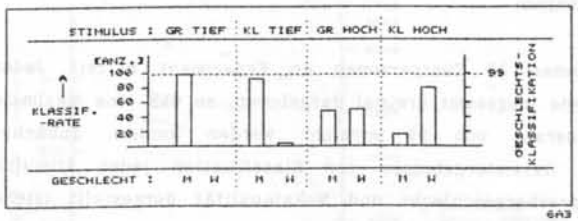
[7] Zur Theorie siehe auch Kap. 3 bzw. MILLER (1953) oder SLAWSON (1968).

[8] Zur Theorie siehe Kap. 3.

KLASSIFIKATIONS-RATE ENZ.%	STIMULUS			
	GR TIEF	KL TIEF	GR HOCH	KL HOCH
MÄNNLICH	99	94	48	18
WEIBLICH	0	5	51	81

6T1

Tab. 6.1: Geschlechtsklassifikation der Stimuli aus Exp. 5. Gesamtklassifikation



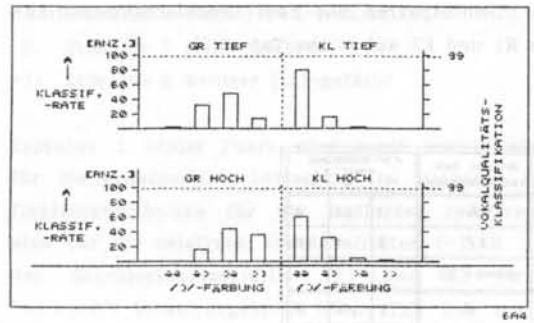
6A3

Abb. 6.3: Balkendiagramm zur Geschlechtsklassifikation. Gesamtklassifikation in Exp. 5

KLASSIFIKATIONS-RATE ENZ.%	STIMULUS			
	GR TIEF	KL TIEF	GR HOCH	KL HOCH
aa	2	81	8	60
a›	33	17	17	33
›a	49	1	45	5
››	15	8	37	1

6T2

Tab. 6.2: Vokalqualitätsklassifikation der Stimuli aus Exp. 5 Gesamtklassifikation



6A4

Abb. 6.4: Balkendiagramm zur Vokalqualitätsklassifikation. Gesamtklassifikation in Exp. 5

Desweiteren werden die Teilergebnisse für die Klassifikation der Vokalqualität bei bestimmten Stimuluspaaren betrachtet. Es soll nachgewiesen werden, daß bei demselben Stimulus, bedingt durch unterschiedliche Geschlechtsassoziation, auch Unterschiede in der Vokalqualitätsklassifikation auftreten. Somit ist für den entsprechenden Stimulus die *geschlechtsassoziationsbedingte Vokalqualität* oder einfacher die *bedingte Vokalqualität* zu ermitteln. Die Klassifikationsergebnisse der Hörer bezüglich der Vokalqualität werden somit, je nachdem, ob die Hörer mit dem Stimulus gleichzeitig weibliches oder männliches Sprechergeschlecht assoziieren, getrennt ausgewertet. Dazu sind zwei Stimuluspaare zu betrachten, deren Stimuli jeweils die gleiche artikulatorische Einstellung aufweisen, deren Formantfrequenzen also jeweils gleich sind. Bei diesen im folgenden mit P1 und P2 bezeichneten Paaren wechselt die Einstellung der phonatorischen Parameter, also die Grundfrequenz, von "tief" nach "hoch".

	Stimulus 1		Stimulus 2
P1:	S[gr/tief]	->	S[gr/hoch]
P2:	S[kl/tief]	->	S[kl/hoch]

Obwohl der Stimulus S[kl/tief] von einigen Hörern in anderen Stimuluspaaren auch als "weiblich" klassifiziert wurde, treten

bei diesen beiden Stimuluspaaren nur zwei Geschlechtsklassifikationskombinationen K1 und K2 auf (siehe Tab. 6.3).

	GESCHLECHTS-KLASSIFIKATION	ANZAHL DER TESTPERSONEN	/ɔ/-FÄRBUNG ANZ. 3		
			-1	0	+1
P1	K1: M->M	16	1	11	4
	K2: M->W	17	3	4	10
		SUMME: 33			
P2	K1: M->M	6	1	5	0
	K2: M->W	27	0	19	8
		SUMME: 33			

6T3

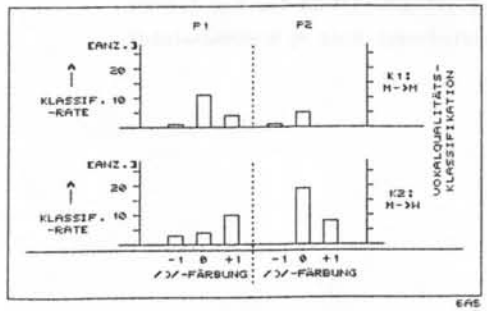
Tab. 6.3: Kombinationen der Geschlechtsassoziation und die bedingten relativen Vokalqualitäten für Stimuluspaar 1 und 2 aus Exp. 5

Interessant ist vor allem der Stimulus S[gr/hoch] in Paar 1, der von gleich vielen Testpersonen als "männlich" oder "weiblich" klassifiziert wird. Auch die Gesamtklassifikationsergebnisse zeigen, daß besonders dieser Stimulus sehr offen bezüglich der Geschlechtsassoziation ist (siehe Tab. 6.1). Dadurch eignet sich dieser Stimulus in hohem Maße für den hier angestrebten Nachweis einer Vokalqualitätsdifferenz bei unterschiedlicher Geschlechtsassoziation.

Zur Analyse der Vokalqualitätsdifferenzen wird wie folgt vorgegangen. Es wird nicht wie bei der Erstellung der Gesamtergebnisse der Klassifikation die absolute Transkription innerhalb der oben angegebenen vierstufigen Skala betrachtet, sondern bezogen auf das zu untersuchende Stimuluspaar wird nur ausgewertet, ob eine Testperson den zweiten Stimulus des Paares *relativ* zum ersten Stimulus als stärker [ɔ]-gefärbt, gleich gefärbt oder weniger [ɔ]-gefärbt transkribiert hat. Zur Benennung dieser *relativen Vokalqualitäten* werden folgende Symbole eingeführt:

+1: Stimulus 2 stärker [ɔ]-gefärbt
0: Stimulus 2 gleich gefärbt
-1: Stimulus 2 weniger [ɔ]-gefärbt

Stimulus 1 beider Paare wird somit jeweils als Referenzstimulus für die Vokalqualität betrachtet. Im folgenden werden die Klassifikationsergebnisse für die *bedingten relativen Vokalqualitäten*, also für die relativen Vokalqualitäten (-1, 0, +1), getrennt nach der Geschlechtsassoziation (K1 und K2 für "männlich" bzw. "weiblich") zusammengestellt (siehe Tab. 6.4 und Abb. 6.5).



6A5

Abb. 6.5: Balkendiagramm der bedingten relativen Vokalqualitäten für Stimuluspaar 1 und 2 nach Exp. 5

Bei beiden Paaren zeigen sich Unterschiede in der relativen Vokalqualität, bedingt durch verschiedene Geschlechtsassoziation. Diese auftretenden Vokalqualitätsdifferenzen sind überzufällig.

Für Paar 1 und Paar 2 ergibt sich ein Signifikanzniveau besser als 5%.[9]

Auch die Richtung, in die sich die Vokalqualität beim Vergleich der beiden, bei jedem Paar auftretenden Kombinationen verschiebt, steht in Einklang mit der in Kapitel 3 aufgestellten Theorie zur Primärnormierung. Nach dieser Theorie verschiebt sich die Vokalschablone bei weiblicher Geschlechtsassoziation hin zu höheren Formantfrequenzen (dargestellt durch die perzeptive Vokalschablone in Abb. 6.6). Für ein und denselben Stimulus, also für konstant bleibende Formantfrequenzen, bedeutet das - vom Bezugssystem der Vokalschablone aus gesehen - ein Verschieben der Vokalqualität hin zu tieferen Frequenzen. Im hier untersuchten Fall der Formantlage im [ɑ]-[ɔ]-Bereich ergibt dies eine Verschiebung der Vokalqualität hin zum [ɔ]. Diese von der Theorie der Primärnormierung geforderte stärkere [ɔ]-Färbung im Fall der weiblichen Geschlechtsassoziation spiegelt sich in den Testdaten eindeutig wider (siehe Tab. 6.3 und Abb. 6.5).

[9] Die Berechnung des Signifikanzniveaus erfolgte nach dem 2I-Dach-Test von KULLBACK (siehe HELLER et al. (1974, S. 222)). Der einfache Kontingenztest nach PEARSON auf der Grundlage der Chi-Quadrat-Verteilung (siehe z.B. BAMBERG et al. (1982, S. 202)) kann nicht durchgeführt werden, weil die Häufigkeitsverteilung auch Werte kleiner als fünf annimmt.

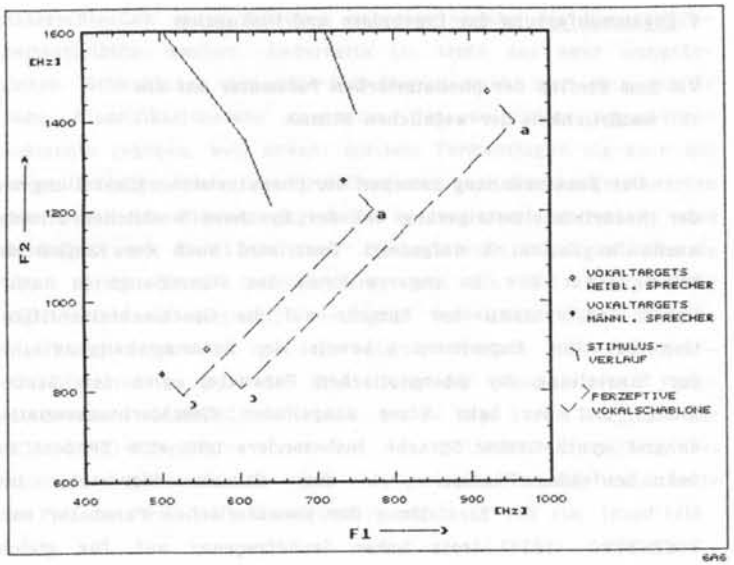


Abb. 6.6: [ɑ]-[ɔ]-Ausschnitt der F1/F2-Ebene mit Stimulusverlauf und der Einstellung der perzeptiven Vokalschablone je nach Sprechergeschlecht

7 Zusammenfassung der Ergebnisse und Diskussion

7.1 Zum Einfluß der phonatorischen Parameter auf die Natürlichkeit der weiblichen Stimme

Der Zusammenhang zwischen der phonatorischen Einstellung und der Natürlichkeitssteigerung bei der Synthese weiblicher Stimmen wurde in Kapitel 2 aufgezeigt. Dort wird auch der Einfluß der Stimmqualität bzw. im engeren Sinne des Stimmklangs in natürlicher und synthetischer Sprache auf die Geschlechtsidentifikation diskutiert. Experiment 1 beweist den Zusammenhang zwischen der Einstellung der phonatorischen Parameter, also dem Stimmklang, und der beim Hörer ausgelösten Geschlechtsassoziation anhand synthetischer Sprache. Insbesondere tritt eine Tendenz zur "männlich"-Klassifikation bei den Stimuli S[gr/hoch] und S[kl/hoch] mit der Einstellung der phonatorischen Parameter nach ROSENBERG (1971) trotz hoher Grundfrequenz auf. Die gleiche Tendenz zeigen auch die Ergebnisse der Untersuchung von COLEMAN (1976) (siehe auch Abs. 2.3.3 und vgl. Abb. 2.1 mit Abb. 5.2).

Die vom Hörer durchzuführende Klassifikation nach weiblichem oder männlichem Sprechergeschlecht stellt ein indirektes Maß für den absoluten Natürlichkeitsgrad der synthetischen Stimme dar. Beispielsweise könnte man eine synthetische Stimme dann als "natürlich" bezeichnen, wenn in einem dem Experiment 1 entsprechenden Klassifikationsexperiment ein gewisser Prozentsatz an Fehlklassifikationen unterschritten wird. Für die HOLMBERG-Einstellung beträgt die Fehlklassifikationsrate nach Experiment 1 nur 6.7% gegenüber einer Fehlklassifikationsrate von 18% bei der ROSENBERG-Einstellung. Andererseits kann die Natürlichkeit einer synthetischen Stimme auch direkt erfragt werden, wie es in Experiment 1 durch das Unnatürlichkeitsattribut, das der Hörer dem Stimulus optional verleihen kann, geschieht. Diese direkte Erfragung ergibt für die ROSENBERG-Einstellung einen Unnatürlichkeitsgrad von 49%, während dieser für die HOLMBERG-Einstellung auf 10% absinkt. Jedoch ist diese Unnatürlichkeitsskala starken subjektiven Einflüssen ausgesetzt. Je nachdem wie kritisch der Hörer den synthetischen Stimuli gegenübersteht, wird dieser

unterschiedlich stark Gebrauch von der Angabe des Unnatürlichkeitsattributs machen. Anderseits ist trotz des sehr unnatürlichen Stimmklangs der GEV-ROS-Stimuli auch hier eine relativ hohe Klassifikationsrate zugunsten des weiblichen Sprechergeschlechts gegeben, weil sowohl mittlere Formantlagen als auch der enorm wichtige Faktor der Grundfrequenz nach Durchschnittswerten der weiblichen Stimme eingestellt worden sind. Interessant sind deswegen auch die "künstlichen" Kombinationen der Signalmerkmale von verschiedenen Geschlechtern, wie sie bei den Stimuli S[gr/hoch] und S[kl/tief] vorliegen. Vor allem beim Stimulus S[gr/hoch] zeigt sich gut die starke Tendenz zur "weiblich"-Klassifikation im Fall der Einstellung der phonatorischen Parameter nach HOLMBERG (siehe Abb. 5.2 zu Exp. 1). Im Vergleich zur ROSENBERG-Einstellung tritt bei der HOLMBERG-Einstellung bei allen vier Kombinationen auch eine geringere Unnatürlichkeitsrate im Fall der "weiblich"-Klassifikation auf.

Im Gegensatz dazu wird in Experiment 2, 3 und 4 nur die relative Natürlichkeit eines Stimulus in bezug auf die im entsprechenden Experiment insgesamt untersuchte Stimulusmenge ermittelt. Durch Experiment 2 wird bewiesen, daß eine gute Einstellung der phonatorischen Parameter stärkeren Einfluß auf die Natürlichkeit der Stimuli hat als die Berücksichtigung bzw. Außerachtlassung der akustischen Interaktion. Es ist bei stationären offenen Vokalen also nicht entscheidend, ob die Synthese der weiblichen Stimme im GEV-Modell oder im GAI-Modell erfolgt. NORD et al. (1984) kommen zu ähnlichen Ergebnissen. Sie synthetisieren Vokale mit fünf Formanten in einem Modell der GEV-Gruppe und in einem Modell der GAI-Gruppe. Allerdings wird der Sprechtrakt mittels Formantensynthese realisiert und eine andere phonatorische Steuerfunktion benutzt. Es wird festgestellt, daß sich die in beiden Modellen generierten stationären Vokale perzeptiv nur unwesentlich unterscheiden. Allerdings treten nach NORD et al. (1984) zwischen den Vokalen der weiblichen Stimme etwas größere perzeptive Unterschiede auf. Trotzdem wird in neueren

Arbeiten[1] immer wieder ein Modell der GAI-Gruppe benutzt. Hauptgrund für diese Bevorzugung von GAI-Modellen ist das Versagen der GEV-Modelle im Fall stärkerer Interaktion (z.B. bei geschlossenen Vokalen) und im Fall der Generierung zeitveränderlicher Signale. So kann ein GEV-Modell beispielsweise den bei F0-Variationen auftretenden Amplitudeneffekt[2] nicht wiedergeben.

Andererseits wird darauf hingewiesen, daß die mechanische Kopplung für die Synthese der weiblichen Stimme sehr wichtig ist.[3] Das bedeutet, daß sowohl die Schwingungsform der Stimmlippen als auch ihre Frequenz und Amplitude stark von der Eingangsimpedanz des Sprechtrakts und damit von der Sprechtraktform abhängen. Daher stellt sich die Frage, ob die für den Schwa-Laut mittels der Technik des inversen Filterns gewonnenen Daten zur glottalen Wellenform bei anderen Vokalen bzw. Sprechtraktkonfigurationen noch relevant sind. Es müßte zunächst überprüft werden, inwieweit die Natürlichkeit der hier benutzten Einstellungen der phonatorischen Parameter bei anderen Vokalen erhalten bleibt. Falls dies nicht der Fall ist, müßte die glottale Wellenform dann entweder anhand perzeptiver Tests iterativ eingestellt werden, oder es müßte versucht werden, z.B. anhand eines Modells der GMI-Gruppe die Haupttendenzen der phonatorischen Parameter in Abhängigkeit von der Sprechtraktform zu ermitteln. Allerdings liegen heute noch keine Ergebnisse zur Simulation der weiblichen Stimme in einem Modell der GMI-Gruppe vor, obwohl diese Modelle zur Natürlichkeitssteigerung synthetischer Sprache benutzt werden.[4] Es ist auch fraglich, ob ein Ein- oder Zweimassenmodell überhaupt die nötige Flexibilität in der Änderung der Form der glottalen Welle besitzt, um die erforderlichen Stimmqualitätsänderungen zu modellieren.

---

[1] Siehe z.B. FANT (1986) oder FANT et al. (1987).

[2] Siehe FANT et al. (1982, S. 3).

[3] Siehe TITZE (1986) und CHENG et al. (1987a und 1987b).

[4] Siehe KOIZUMI et al. (1987).

7.2 Die Unterschiede in den phonatorischen Einstellungen für die weibliche und männliche Stimme

7.2.1 Die phonatorischen Unterschiede nach HOLMBERG et al. und ihre Verifikation anhand eines Syntheseexperimentes dieser Arbeit

Experiment 1 beweist nicht nur den Einfluß der phonatorischen Parameter auf die Geschlechtsidentifikation, sondern verifiziert auch, daß die phonatorischen Einstellungen nach HOLMBERG et al. (1988) gegenüber den Einstellungen nach ROSENBERG (1971) gerade für die Synthese der weiblichen Stimme auf größere Akzeptanz stoßen. Dies belegen sowohl die Ergebnisse der Geschlechtsklassifikation als auch die Werte der Unnatürlichkeitsrate. Daher kann angenommen werden, daß die in Experiment 1 benutzten phonatorischen Einstellungen nach HOLMBERG et al. viele wesentliche Tendenzen der geschlechtsspezifischen Unterschiede der glottalen Wellenform erfassen. In der Hauptsache sind die Tendenzen in der weiblichen gegenüber der männlichen Phonation nach HOLMBERG et al. (1988) die folgenden:

a) höhere Grundfrequenz (F0)
b) höherer Offenquotient (opqu)
c) geringeres Stromänderungsmaximum (mfdr)
d) höherer relativer Verschlußstrom (dcfl/pkfl)
e) geringerer Maximalstrom (pkfl)
f) geringere Wechselstromamplitude (acfl)
g) geringerer mittlerer Strom (avfl)

Keine signifikanten Tendenzen wurden für die folgenden Größen festgestellt:

a) subglottaler Druck (p_sub)
b) Geschwindigkeitsquotient (spqu)

Damit gibt es nach HOLMBERG et al. auch keine signifikanten Unterschiede für die Symmetrie der Wellenform. Besonderer Wert wurde bei der Einstellung der phonatorischen Parameter auf das

Herausarbeiten der oben angegebenen Tendenzen gelegt.[5] Diese Tendenzen werden im folgenden anhand anderer Messungen der glottalen Wellenform und anhand anderer Untersuchungen zur Synthese der weiblichen und der männlichen Stimme diskutiert.

7.2.2 Ergebnisse aus anderen Untersuchungen mittels der Methode des inversen Filterns

Zur Untersuchung der Wellenform des glottalen Volumenstroms der weiblichen und der männlichen Stimme mittels der Methode des inversen Filterns liegen neben der Studie von HOLMBERG et al. zwei weitere Untersuchungen vor (siehe KARLSSON (1985) und MONSEN et al. (1977)). Allen Untersuchungen gemeinsam ist das Ergebnis, daß es bezüglich der Wellenform des glottalen Volumenstroms große interindividuelle Unterschiede gibt. Es gibt keine typische geschlechtsspezifische Schwingungsform der glottalen Welle. Ferner wurden in allen zu besprechenden Untersuchungen nur Vokale im Schwa-Bereich untersucht. Daher liefern diese experimentell bestimmten Daten keinen Aufschluß über die Variation der Schwingungsform bei verschiedenen Sprechtrakteinstellungen.

MONSEN et al. (1977) untersuchten zehn weibliche und zehn männliche Sprecher bei unterschiedlicher Sprechlautstärke und in unterschiedlichen Phonationsmodi (normal, falsetto, creak). Es wurden die folgenden Tendenzen bezüglich der weiblichen gegenüber der männlichen glottalen Welle festgestellt:

a) stärkere Symmetrie der Wellenform
b) um rund 6 dB geringere Gesamtintensität der glottalen Welle
c) steilerer Abfall des Spektrums der glottalen Welle

Die Durchschnittswerte des spektralen Abfalls liegen für weib-

[5] Siehe insbesondere Einstellung 3 für die weibliche Stimme und Einstellung 1 für die männliche Stimme nach Tab. 5.1 und die Diskussion dazu in Abs. 5.1.

liche Stimmen bei 15 dB/Okt. und bei männlichen Stimmen bei 12 dB/Okt. Die Symmetrisierung wird nach MONSEN et al. als ein allgemeines Korrelat der F0-Erhöhung und der Abnahme der Sprechintensität auch bei männlichen Stimmen ausgewiesen.

Die Untersuchung von KARLSSON (1985) ist mit sechs weiblichen und einer männlichen Testperson neben den Untersuchungen von MONSEN et al. (1977) und von HOLMBERG et al. (1988) weniger umfangreich. Während die Aufnahme des Sprachsignals bei der Untersuchung von MONSEN et al. (1977) mit einem reflexionslosen Metallrohr[6] durchgeführt wurde, wird hier und auch bei der Untersuchung von HOLMBERG et al. (1988) das Sprachsignal mithilfe der Rothenberg-Maske[7] aufgenommen. Wie HOLMBERG et al. stellt auch KARLSSON fest, daß bei der glottalen Welle fast aller Sprecherinnen ein Verschlußstrom ("leakage") auftritt. Die Größe des Verschlußstroms steigt sogar noch mit dem Übergang von normal-lauter zu leiser Stimme.

Die Analyse des Datenmaterials aller drei Untersuchungen ergibt, daß der Offenquotient bei der weiblichen Stimme größer als bei der männlichen ist. Allerdings wird die Symmetrisierungstendenz der glottalen Wellenform kontrovers diskutiert. Zwar nimmt der Schließquotient der weiblichen Stimme nach HOLMBERG et al. auch einen recht großen Wert an, jedoch liegt nach HOLMBERG et al. in der Symmetrisierung der glottalen Wellenform kein entscheidender Unterschied zwischen der Wellenform der weiblichen und der männlichen Stimme. Der Geschwindigkeitskoeffizient drückt die Eigenschaft der Symmetrie der glottalen Welle am eindeutigsten aus. Nimmt er den Wert 1 an, so sind die Öffnungs- und die Schließphase gleich lang und es liegt eine symmetrische Wellenform vor. Bei den Daten nach HOLMBERG et al. wurde kein signifikanter geschlechtsspezifischer Unterschied für den Geschwindigkeitskoeffizienten festgestellt. Vielmehr heben HOLMBERG et al. und KARLSSON die Rolle des Verschlußstroms hervor.

[6] Siehe SONDHI (1975).
[7] Siehe ROTHENBERG (1973).

Ein starker Verschlußstrom bei weiblichen Sprechern kann auch insofern von Bedeutung sein, als durch diesen Gleichstromanteil glottales Rauschen verursacht werden kann.[8] Wegen des steileren Abfalls des Wechselstromanteils des Anregungsspektrums des Sprechtrakts wird dieses glottale Rauschen bei der weiblichen Stimme nicht so stark überdeckt wie ein gleichstarkes Rauschen bei einer männlichen Stimme. Somit könnte zumindest für eine bestimmte Klasse der weiblichen Sprecher das Stimmqualitätsattribut Behauchung ("Breathiness") zum normalen Phonationsmodus gehören.

7.2.3 Ergebnisse aus anderen Syntheseexperimenten

Zur Einstellung der phonatorischen Parameter für die weibliche und die männliche Stimme werden im folgenden die Ergebnisse von fünf Untersuchungen diskutiert.

Ausgehend von den Ergebnissen der Untersuchung von MONSEN et al. (1977) synthetisiert CARRELL (1981) die weibliche und die männliche Stimme. In dieser Untersuchung wird der geschlechtsspezifische Unterschied des Stimmklangs, der neben F0 hier vor allem in der Symmetrie der glottalen Wellenform liegt, modelliert. Es wird gezeigt, daß eine symmetrische Wellenform eher die weibliche Stimme charakterisiert.

CHENG et al. (1987a und 1987b) benutzen ebenfalls eine symmetrische Wellenform. Als weiteres wesentliches Merkmal der glottalen Wellenform der weiblichen Stimme wird hier der Offenquotient angesehen. Im speziellen werden die perzeptiven Auswirkungen der Variation des Offenquotienten bei unterschiedlichen Sprechtraktgeometrien untersucht. Es ergeben sich anhand von Hörtests optimierte Werte je nach Vokal. Diese Werte für den Offenquotienten liegen zwischen 0.65 und 0.8 .

---

[8] Siehe HOLMBERG et al. (1988, S. 522), KARLSSON (1988, S. 66) und KLATT (1987).

In einer Studie von NORD et al. (1984) wird ebenfalls die Tendenz der Vergrößerung des Offenquotienten beim Vergleich der weiblichen mit der männlichen Stimme hervorgehoben. Es werden allerdings sehr niedrige Werte angesetzt (opqu(männl.) = 0.5 und opqu(weibl.) = 0.6). Ferner wird angenommen, daß bei weiblichen Sprechern höherer subglottaler Druck vorliegt als bei männlichen. Dies wird aber durch die neueren Messungen mittels der Technik des inversen Filterns deutlich widerlegt.[9]

In der Arbeit von FANT et al. (1987) wird eine symmetrische Wellenform sowohl bei der Synthese der weiblichen als auch bei der Synthese der männlichen Stimme verwendet. Neben der Tendenz eines größeren Offenquotienten bei der weiblichen gegenüber der männlichen Stimme wird hier auch der Verschlußstrom als wesentliches Merkmal weiblicher Stimmen ausgewiesen. Allerdings werden hier für die Synthese der weiblichen und der männlichen Stimme auch unterschiedliche Steuerfunktionen benutzt. Die Benutzung des LF-Modells[10] für die Synthese der weiblichen Stimme erlaubt eine Variation des spektralen Abfalls von 12 dB/Okt. bis hin zu 18 dB/Okt. .

Obwohl die Arbeit von MONSEN et al. (1978) nicht zum Ziel hat, die weibliche Stimme zu synthetisieren, sind einzelne Ergebnisse dieser Studie auch hier von Interesse, da die Autoren mit einem Modell der GMI-Gruppe die Änderung der glottalen Wellenform bei F0-Erhöhung bis über 200 Hz studieren. Zum einen wird F0 durch Anheben des subglottalen Drucks erhöht, wobei sich neben einer allgemeinen Amplitudenerhöhung keine Änderung der Form der glottalen Welle ergibt. Zum anderen tritt jedoch bei der F0-Erhöhung durch die Änderung der Stimmlippenspannung eine Tendenz zur Verlängerung der Offenphase und zur Symmetrisierung der Wellenform auf. Dies führt auch zu einem steileren Abfall des Quellenspektrums.

---

[9] Siehe HOLMBERG et al. (1988) und KARLSSON (1988).

[10] LILJENCRANTS-FANT-Modell, siehe FANT et al. (1987).

7.2.4 Der Einfluß von Verschlußstrom und Schließquotient

Die Ergebnisse der hier referierten fünf Untersuchungen zur Synthese der weiblichen und der männlichen Stimme spiegeln ebenfalls die Tendenzen aus den Messungen mittels der Technik des inversen Filterns wider. Während in allen Arbeiten die Tendenz der Verlängerung der Offenpase bei der weiblichen gegenüber der männlichen Stimme auftritt, wird einerseits in der Symmetrisierung, andererseits in der Einführung eines Verschlußstroms der wichtigste geschlechtsspezifische Unterschied bezüglich der glottalen Wellenform gesehen.

Um die perzeptive Auswirkung der Änderung der Symmetrie oder des Verschlußstroms bei der weiblichen Stimme im hier benutzten Synthesemodell zu studieren, wurden Experiment 3 und Experiment 4 durchgeführt. Tatsächlich wird eine Absenkung des Verschlußstroms auf annähernd Null nach den Ergebnissen aus Experiment 3 von den Hörern abgelehnt. Allerdings sind die perzeptiven Unterschiede durch die Änderung der Stärke des Verschlußstroms nur sehr gering. Das steht aber im Einklang mit der Tatsache, daß die akustische Interaktion bei stationären Vokalen nicht wichtig ist (siehe Exp. 2). Die Änderung der Verschlußstromhöhe führt nur zu Änderungen in den durch die akustische Interaktion hervorgerufenen Überlagerungsschwingungen. Je höher die glottale Öffnungsfläche und damit der Verschlußstrom, desto mehr Anteile der Überlagerungsschwingung verlagern sich aus der Offenphase in die Verschlußphase (siehe Abb. 5.5). Grund für die perzeptive Ablehnung des glottalen Totalverschlusses kann in der dann abrupt auftretenden Konstanz des glottalen Volumenstroms liegen. Dies führt zu Unstetigkeiten in den höheren Ableitungen des glottalen Volumenstroms und damit zum weniger starken Abfall des Spektrums des Volumenstroms. Abbildung 7.1a und Abbildung 7.1b zeigen, daß die Spektren des glottalen Volumenstroms der Stimuli DC02 bis DC05 nahezu identisch sind (vgl. auch mit den zugehörigen perzeptiven Resultaten von Experiment 3 (Abb. 5.7)). Im Gegensatz dazu weisen die Spektren für Stimulus DC01 ab ca. 1000 Hz höhere Intensität auf.

Eine andere Auswirkung der Anhebung des Verschlußstroms liegt im Auftreten von glottalem Rauschen. Insofern ist es von Interesse, das im Bereich der Glottis durch Luftturbulenzen entstehende Rauschen bei der Synthese der weiblichen Stimme zu berücksichtigen.

Der Vergleich der Ergebnisse von Experiment 3 und Experiment 4 zeigt, daß die Variation des Schließquotienten und damit die Variation des Symmetriegrades der glottalen Welle empfindlicher als die Variation des Verschlußstroms wahrgenommen wird. So zeigt sich nach Experiment 4 ein Maximum im relativen Natürlichkeitsgrad bei der glottalen Wellenform mit einem Geschwindigkeitskoeffizienten von annähernd 1. Somit bestätigen die Ergebnisse von Experiment 4 die Symmetrisierungstendenz der glottalen Welle für die weibliche Stimme. Allerdings ergibt sich nach HOLMBERG et al. auch für die männliche glottale Welle ein stärkerer Symmetrisierungsgrad als nach ROSENBERG.

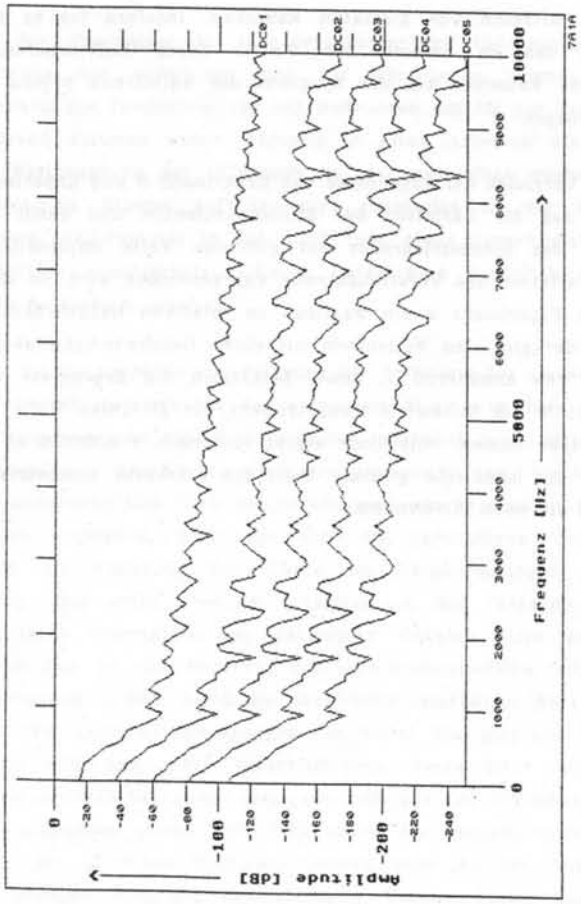


Abb. 7.1a: 256 Punkte-FFTs[11] des glottalen Volumenstroms (u_glott) aller Stimuli aus Exp. 3. Von oben: DC01 bis DC05 im Abstand von 26 dB. Alle FFT-Fenster bei gleicher Phasenlage

[11] Fast Fourier Transform

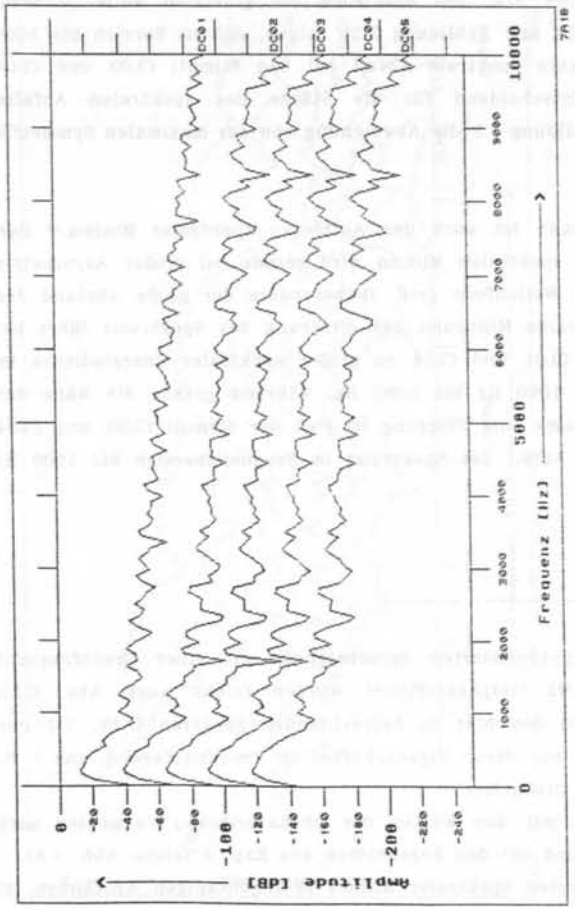


Abb. 7.1b: 256 Punkte-FFTs der Ableitung des glottalen Volumenstroms (u_glott') aller Stimuli aus Exp. 3. Von oben: DC01 bis DC05 im Abstand von 26 dB. Alle FFT-Fenster bei gleicher Phasenlage

Wesentliches Merkmal für die Natürlichkeit der weiblichen Stimme scheint auch in Experiment 4 die Stärke des Abfalls des Signalspektrums bzw. des Spektrums der glottalen Welle zu sein. Abbildung 7.2a und Abbildung 7.2b zeigen, daß im Bereich bis 1000 Hz der stärkste spektrale Abfall bei den Stimuli CL03 und CL04 vorliegt.[12] Entscheidend für die Stärke des spektralen Abfalls ist nach Abbildung 7.2 die Abweichung von der maximalen Symmetrie (spqu=1).[13]

Interessant ist auch das Auftreten spektraler Minima.[14] Der Abstand der spektralen Minima wird gerade bei großer Asymmetrie der glottalen Wellenform groß. Insbesondere der große Abstand des ersten spektralen Minimums zum Ursprung des Spektrums führt bei den Stimuli CL01 und CL06 zu großer spektraler Energiedichte im Bereich von 1000 Hz bis 2000 Hz, während gerade die Nähe des ersten Minimums zum Ursprung im Fall der Stimuli CL03 und CL04 zum starken Abfall des Spektrums im Frequenzbereich bis 1000 Hz führt.

---

[12] Da die synthetisierten Sprachsignale mit einer Grenzfrequenz von 3 kHz tiefpassgefiltert wurden (siehe auch Abs. 5.2), werden bei den hier zu betrachtenden Spektren (Abb. 7.1 und Abb. 7.2) nur deren Eigenschaften im Frequenzbereich von 0 Hz bis 3 kHz diskutiert.

[13] Vgl. auch mit den Werten der phonatorischen Parameter nach Tab. 5.6 und mit den Ergebnissen aus Exp. 4 (siehe Abb. 5.8).

[14] Das Auftreten spektraler Minima in regelmäßigen Abständen ist insbesondere ein Charakteristikum von Rechteckimpulsfolgen (siehe z.B. MEYER et al. (1974, S. 40)). Dabei ist der Abstand der spektralen Minima umgekehrt proportional zur Impulsdauer. Gerade bei den sehr unsymmetrischen glottalen Wellenformen (kurze Schließzeit) kann der Zeitverlauf der Ableitung des glottalen Volumenstroms mit einer Impulsfolge verglichen werden (siehe Abb. 5.6). So kann die Schließzeit in erster Näherung als Impulsdauer angesehen werden.

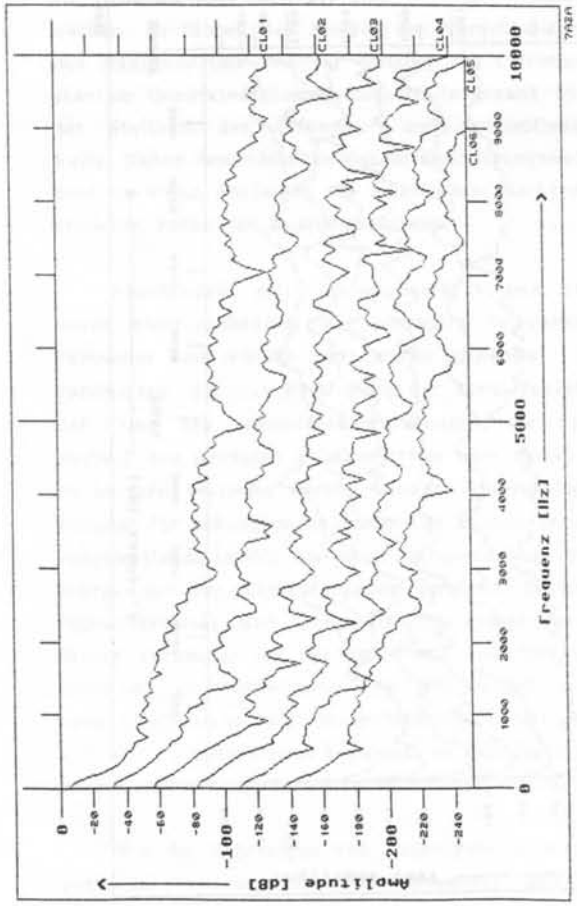


Abb. 7.2a: 256 Punkte-FFTs des glottalen Volumenstroms (u_glott) aller Stimuli aus Exp. 4. Von oben: CL01 bis CL06 im Abstand von 26 dB. Alle FFT-Fenster bei gleicher Phasenlage

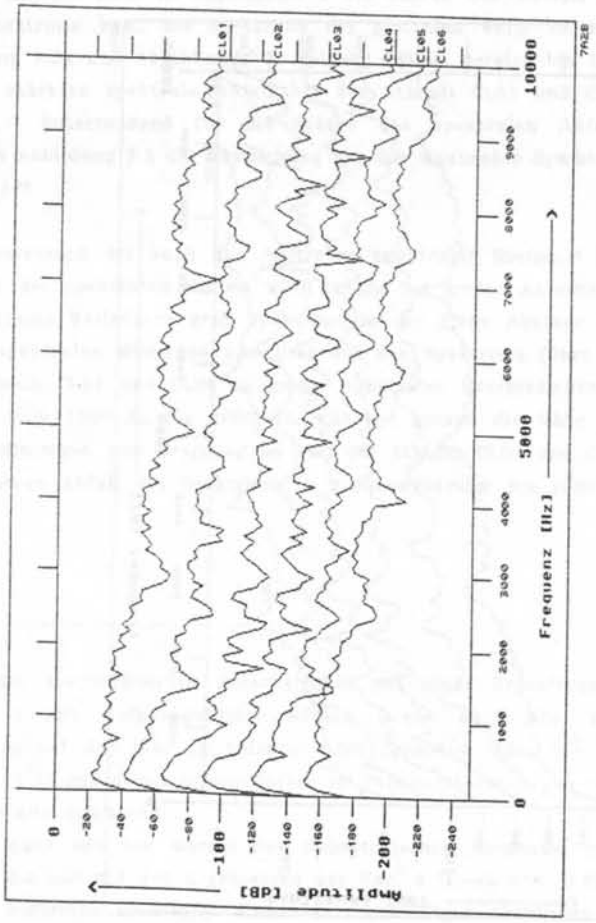


Abb. 7.2b: 256 Punkte-FFTs der Ableitung des glottalen Volumenstroms (u_glott') aller Stimuli aus Exp. 4. Von oben: CL01 bis CL05 im Abstand von 26 dB. Alle FFT-Fenster bei gleicher Phasenlage

Die Tendenz des steileren spektralen Abfalls des Quellenspektrums bei weiblicher Phonation kann neben dem Symmetrisierungsgrad aber auch aus anderen Wellenformparametern abgelesen werden. So führen der Anstieg des Verschlußstroms, das Absinken des Maximalstroms und der Anstieg des Offenquotienten bei konstantem Geschwindigkeitsquotienten insgesamt zu einem Abnehmen der Steilheit der Öffnungs- und Schließflanke der glottalen Welle. Neben dem Absinken des Stromänderungsmaximums führt auch dies zu einem Ansteigen der Neigung des Spektrums, also zu einem steileren Abfall des Quellenspektrums.

Abschließend soll noch angemerkt werden, daß die Steuerparameter nicht unbedingt die perzeptiv relevanten phonatorischen Parameter sein müssen. Dazu wurde in Kapitel 1 zwischen primären Parametern, die zur Festlegung der Steuerfunktion benötigt werden, und den sekundären Parametern, die aus dem zeitlichen Verlauf des glottalen Volumenstroms oder direkt aus den primären Parametern bestimmt werden können, unterschieden. Ein wichtiges Beispiel für sekundäre phonatorische Parameter ist das Stromänderungsmaximum (mfdr). Da dieser Parameter ein direktes Maß für die Stärke der Sprechtraktanregung darstellt, bestimmt er direkt die Signalintensität und ist damit von großer perzeptiver Relevanz. Dieser Parameter ist bei der THH- und THV-Steuerfunktion nur mittelbar zu kontrollieren. Es gibt jedoch andere Synthesemodelle, bei denen die Steuerfunktion direkt die Ableitung des glottalen Volumenstroms bestimmt, so daß das Stromänderungsmaximum direkt als primärer Steuerparameter eingehen kann.[15]

Wie die Ergebnisse von Experiment 1 bestätigen, ist jedoch auch im hier benutzten Synthesemodell gewährleistet, daß die primären Steuerparameter genügend variabel sind, um die geschlechtsspezifischen phonatorischen Unterschiede zu modellieren. Die drei Zeitparameter Grundfrequenz, Offenquotient und Schließquotient sind, wie diese Diskussion gezeigt hat, wichtige phonatorische Parameter, wobei der Schließquotient gleichwertig durch

---

[15] Siehe FANT et al. (1985).

den Geschwindigkeitskoeffizienten ersetzt werden kann. Allerdings scheinen die Absolutwerte von Gleich- und Wechselstromamplitude perzeptiv wenig relevante Größen zu sein. Wesentlichere Amplitudenparameter sind wohl eher der relative Verschlußstrom dcfl/pkfl und das Stromänderungsmaximum für die Phonation. Dementsprechend wurde bei der Generierung der Stimuli die Einstellung dieser Größen iterativ vorgenommen.

7.3 Zum Zusammenhang von Vokalperzeption und Geschlechtsassoziation

Die Theorie zur Primärnormierung (siehe Abs. 3.1) wurde bereits anhand von Untersuchungsergebnissen verschiedener Autoren in Kapitel 3 diskutiert und motiviert. Zwar ist das Phänomen der Formantnormierung wegen der engen Verbindung mit der Vokalperzeption häufig untersucht worden, jedoch bringt gerade die Betrachtung der Normierungsproblematik unter dem Gesichtspunkt der geschlechtsspezifischen Signalunterschiede neue Ideen und Aspekte. Ziel ist es, gerade die Bedeutung der Formantfrequenzen als Faktoren für die Verschiebung der Vokalschablone zu relativieren. Durch die Aufdeckung des engen Zusammenhanges zwischen der Geschlechtsidentifikation und dem Verschieben der Vokalschablone werden neben den Formanten vor allem die Grundfrequenz, aber auch der Stimmklang und letztlich alle weiteren Signalmerkmale der Stimmqualität als wichtige perzeptive Normierungsfaktoren ausgewiesen. Der Grund für die festgestellte starke Unterbewertung anderer Signalmerkmale als der Formanten liegt in der Redundanz des Sprachsignals. Eine Normierung anhand der Formanten ist zwar mathematisch möglich, jedoch liegen im Signal eben weitere Merkmale wie z.B. die Grundfrequenz vor, die ein Verschieben der Vokalschablone einleiten können. Betrachtet man die Sprechergruppen der Männer, Frauen und Kinder, so ist die Erhöhung der Formantfrequenzen immer mit einer Erhöhung von F0 und weiteren systematischen Tendenzen z.B. bei der phonatorischen Einstellung gekoppelt.

Die hier durchgeführte experimentelle Untersuchung zum Zusammenhang von Geschlechtsassoziation und Vokalperzeption (Exp. 5) ist punktuell, d.h. die Theorie der Primärnormierung wurde im Experiment nur anhand einiger Punkte in der F1/F2-Ebene bestätigt. Hingegen findet man in der Literatur häufig experimentelle Untersuchungen zu der Thematik der Verschiebung der Vokalschablone, in denen der gesamte Formantraum "ausgemessen" wird.[16] Ziel der hier durchgeführten Untersuchung ist es aber nicht, die Existenz einer perzeptiven Vokalschablone nachzuweisen oder die Stärke der Verschiebung der Schablone quantitativ zu bestimmen, sondern den Nachweis zu erbringen, daß der Faktor der Geschlechtsassoziation eine Schablonenverschiebung in der nach der Theorie vorhergesagten Richtung überhaupt bewirkt. Durch den bereits von anderen Autoren[17] erbrachten Nachweis der Existenz einer starren, aber verschiebbaren Vokalschablone im Perzeptionssystem ist es auch völlig ausreichend, nur punktuelle Untersuchungen durchzuführen. Es ist deshalb möglich, von den hier im [a]-[ɔ]-Bereich ermittelten Ergebnissen auf eine Gesamtverschiebung der Vokalqualität zu schließen.

Die nur sehr geringen perzeptiven Unterschiede, die von diesem Experiment aufgedeckt werden müssen, stellen hohe Anforderungen an die Meßgenauigkeit, die von Experiment 5 erreicht werden mußte. Schließlich sollte ein Unterschied in der Vokalqualität *eines* Stimulus, also *eines identischen* Signals bei verschiedenen Hörern festgestellt werden. Ein solcher Effekt wird nur schwer durch die Ermittlung absoluter Vokalqualitäten oder durch die Verschiebung von Phonemgrenzen, die bei Vokalen ohnehin unschärfer als bei Konsonanten sind, nachzuweisen sein. Diese Differenzen können nur, wie in Experiment 5 aufgezeigt, durch das Erfragen von Vokalqualitätsänderungen gegenüber einem Referenzstimulus aufgedeckt werden. Die dazu nötige Darbietung zweier Stimuli in enger zeitlicher Abfolge entspricht auch der nach der Theorie der Primärnormierung geforderten Hörsituation des Ein-

---

[16] Siehe AINSWORTH (1975) oder NEAREY (1977).

[17] Siehe NEAREY (1977).

hörens auf einen neuen Sprecher. Die Erfragung der Vokalqualitätsdifferenz zweier Stimuli bezieht sich damit genau auf die Verschiebung der Vokalschablone im Perzeptionssystem.

Trotzdem ist es von Interesse, den gleichen Effekt auch an anderen Punkten der Formantebene nachzuweisen. Dazu ist es immer nötig, einen Bereich zu finden, in dem eine Verschiebung der Formantwerte durch die Variation der Sprechtraktlänge (Isoarealinien) mit einer starken Qualitätsänderung innerhalb der festliegenden Vokalschablone verbunden ist. Um zu meßbaren Vokalqualitätsdifferenzen zu gelangen, müssen also die Bereiche in der Formantebene ausgesucht werden, in denen zwei Targetpunkte auf einer Parallelen zu der durch die Längenänderung des Sprechtrakts definierten Verschiebungsrichtung der Vokalschablone liegen. Diese Bedingung wird am besten von der [ɑ]-[ɔ]-Strecke erfüllt (siehe Abb. 3.4 bis Abb. 3.6). Eventuell bietet sich auch der [i]-[y]-Bereich oder der [e]-[ø]-Bereich[18] an.

Ein anderes bzw. alternatives Experiment zur Verifizierung der Theorie der Primärnormierung wäre ein Nachweis der Verschiebung der Vokalschablone durch die Änderung des Stimmklangs bei konstanter Grundfrequenz und gleichbleibender Geometrie des Ansatzrohres.[19] Jedoch hätte das positive Ergebnis eines solchen Experiments nicht die Aussagekraft, wie sie im Ergebnis des hier durchgeführten Experiments liegt, da Änderungen im Stimmklang immer mit Änderungen in der Einhüllenden des Quellenspektrums verbunden sind und somit hierdurch die Formantfrequenzen direkt beeinflußt werden.

---

[18] Siehe FANT et al. (1974).

[19] Siehe z.B. CARRELL (1981).

Verzeichnis der Abbildungen

1.1 Modell zur Synthese der weiblichen Stimme 3

1.2 Zur Definition der Wellenform-Parameter anhand der glottalen Welle 8

2.1 Perzeptive Geschlechtsklassifikation von Stimuli, die mittels Larynxvibrator erzeugt wurden. Daten nach COLEMAN (1976) 31

3.1 Der Vokalraum weiblicher und männlicher Sprecher. Daten nach PETERSON et al. (1952) 35

3.2 Kopplung zwischen Grundfrequenz und Lage der ersten Formanten des Schwa-Lautes nach Signaldaten (PETERSON et al. (1952)) und gemessene Werte für die perzeptive Verschiebung der Phonemgrenzen nach FUJISAKI et al. (1968) 50

3.3 Der Vokalraum weiblicher und männlicher Sprecher in linearer Näherung. Daten nach PETERSON et al. (1952) 63

3.4 Die Vokalschablone männlicher Sprecher mit Isoarealinien. Daten gemittelt über mehrere Sprecher mehrerer Sprachen nach FANT (1975) 64

3.5 Die Vokalschablone weiblicher Sprecher mit Isoarealinien. Daten gemittelt über mehrere Sprecher mehrerer Sprachen nach FANT (1975) 65

3.6 Die Vokalschablone eines männlichen Sprechers mit Isoarealinien. Daten eines männlichen Sprechers des Hochdeutschen nach JOERGENSEN (1969) 65

4.1 Einfallende und reflektierte Teilwellen zum Zeitpunkt t[j] an der Reflexionsstelle i 74

4.2 Raum-Zeit-Darstellung für die Teilwellen an dem Reflexionspunkt [i,j] 76

4.3 Zum Prinzip der Berechnung der Reflexionsstellen in der Darstellung der Raum-Zeit-Ebene 78

4.4 Modellierung der Sprechtraktverluste 81

4.5 Analogien zwischen akustischem und elektrischem Sprechtraktmodell 82

4.6 Das Einprägen des glottalen Volumenstroms in den Sprechtrakt zum Zeitpunkt t[j] 83

4.7 Der schallweiche Abschluß 84

4.8 Elektrisches Ersatzschaltbild der Glottis im GAI-Modell 88

4.9 Die stark vereinfachte Geometrie der glottalen Engstelle im GAI-Modell 90

4.10 Die THH-Steuerfunktion im GAI-Modell 93

4.11 Zum Überlagerungseffekt, zur Impulsmaximumverschiebung und zur Phasenbegrenzung von u_glott und a_glott im GAI-Modell 95

4.12 Zur Definition der Analyseparameter 98

4.13 Sprechtraktanregung und Stromänderungsmaximum 100

4.14 Die THV-Steuerfunktion im GEV-Modell 103

4.15 Die THH-Steuerfunktion im GEV-Modell 103

5.1 Signaldarstellung der Stimuli zu Exp. 1 114

5.2 Balkendiagramm zur Geschlechtsklassifikationsrate und zur Unnatürlichkeitsrate nach Exp. 1 117

5.3a Signaldarstellung der Stimuli zu Exp. 2: Synthese der weiblichen Stimme 121

5.3b Signaldarstellung der Stimuli zu Exp. 2: Synthese der männlichen Stimme 122

5.4 Balkendiagramm zum relativen Natürlichkeitsgrad nach Exp. 2 124

5.5 Signaldarstellung der Stimuli zu Exp. 3: Variation des Verschlußstroms 128

5.6 Signaldarstellung der Stimuli zu Exp. 4: Variation des Schließquotienten 131

5.7 Balkendiagramm zum relativen Natürlichkeitsgrad nach Exp. 3: Variation des Verschlußstroms 133

5.8 Balkendiagramm zum relativen Natürlichkeitsgrad nach Exp. 4: Variation des Schließquotienten 133

6.1 Sonagramm des Logatoms [ɛɪɤ] für den Stimulus S[gr/tief] aus Exp. 5 139

6.2 Verlauf von F1 und F2 der Stimuli aus Exp. 5 in der F1/F2-Ebene 140

6.3 Balkendiagramm zur Geschlechtsklassifikation. Gesamtklassifikation in Exp. 5 142

6.4 Balkendiagramm zur Vokalqualitätsklassifikation. Gesamtklassifikation in Exp. 5 143

6.5 Balkendiagramm zu den bedingten relativen Vokalqualitäten für Stimuluspaar 1 und 2 nach Exp. 5 145

6.6 [ɛ]-[ɔ]-Ausschnitt der F1/F2-Ebene mit Stimulusverlauf und der Einstellung der perzeptiven Vokalschablone je nach assoziiertem Sprechergeschlecht 147

7.1a 256-Punkte-FFTs des glottalen Volumenstroms aller Stimuli aus Exp. 3: Variation des Verschlußstroms 158

7.1b 256-Punkte-FFTs der Ableitung des glottalen Volumenstroms aller Stimuli aus Exp. 3 159

7.2a 256-Punkte-FFTs des glottalen Volumenstroms aller Stimuli aus Exp. 4: Variation des Schließquotienten 161

7.2b 256-Punkte-FFTs der Ableitung des glottalen Volumenstroms aller Stimuli aus Exp. 4 162

Verzeichnis der Tabellen

1.1 Die Zeitphasen einer Schwingungsperiode der glottalen Welle 9

1.2 Definition der Wellenform-Zeitparameter 9

1.3 Definition der Wellenform-Amplitudenparameter 10

1.4 Phonatorischer und artikulatorischer Steuerparametersatz 11

4.1 Zusammenstellung der physikalischen Konstanten und der Geometriefaktoren 89

4.2 Definition der THH-Steuerfunktion im GAI-Modell 92

4.3 Definition der Analyseparameter 98

4.4 Definition der THV-Steuerfunktion im GEV-Modell 102

4.5 Definition der THH-Steuerfunktion im GEV-Modell 102

5.1 Einstellung der HOLMBERG-Parameter im GAI-Modell für die Synthese der weiblichen und der männlichen Stimme 108

5.2 Testergebnisse zu Exp. 1: Geschlechtsklassifikationsrate und Unnatürlichkeitsrate 116

5.3 Einstellung der HOLMBERG-Parameter im GEV-Modell für die Synthese der weiblichen und der männlichen Stimme 119

5.4 Ergebnisse zu Exp. 2: Der relative Natürlichkeitsgrad 124

5.5 Einstellungen der phonatorischen Parameter der Stimuli aus Exp. 3: Variation des Verschlußstroms 127

5.6 Einstellungen der phonatorischen Parameter der Stimuli aus Exp. 4: Variation des Schließquotienten 130

5.7 Testergebnisse zu Exp. 3: Variation des Verschlußstroms 132

5.8 Testergebnisse zu Exp. 4: Variation des Schließquotienten 133

6.1 Geschlechtsklassifikation der Stimuli aus Exp. 5. Gesamtklassifikation 142

6.2 Vokalqualitätsklassifikation der Stimuli aus Exp. 5. Gesamtklassifikation 142

6.3 Kombinationen der Geschlechtsassoziation und die bedingten relativen Vokalqualitäten für Stimuluspaar 1 und 2 aus Exp. 5 144

Literaturverzeichnis

AINSWORTH, W. A. (1974): "The influence of precursive sequences on the perception of synthesized vowels", *Language and Speech* 17, 103-109.

AINSWORTH, W. A. (1975): "Intrinsic and extrinsic factors in vowel judgements", in: FANT und TATHAM (1975), 103-113.

AINSWORTH, W. A. (1976): *Mechanisms of speech perception.* Pergamon Press, Oxford.

ANANTHAPADMANABHA, T. V., FANT, G. (1982): "Calculation of true glottal flow and its components", *Speech Transm. Lab.-Quart. Prog. Stat. Rep.* 1/1982, 1-30. Royal Inst. Techn., Stockholm.

ARONOVITCH, D. S. (1976): "The voice of personality: Stereotyped judgements and their relation to voice quality and sex of speaker", *Journ. Soc. Psychol.* 99, 207-220.

ASSMANN, P. F., NEAREY, T. M., HOGAN, J. T. (1982): "Vovel identification: orthographic, perceptual, and acoustic aspects", *Journ. Acoust. Soc. Am.* 71, 975-989.

ATAL, B. S., CHANG, J. J., MATHEWS, M. V., TUKEY, J. W. (1978): "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", *Journ. Acoust. Soc. Am.* 63, 1535-1555.

BAMBERG, G., BAUR, F. (1982): *Statistik.* Oldenbourg Verlag, München (2. überarb. Aufl.).

BREND, R. M. (1971): "Male-female intonation patterns in American English", *Proc. 7th Int. Cong. Phon. Sci.*, 866-870. Mouton, The Hague.

BROWN, W. S., FEINSTEIN, S. H. (1977): "Speaker sex identification utilizing a constant laryngeal source", *Folia Phoniatrica* **29**, 240–248.

CARLSON, R., GRANSTRÖM, B., FANT, G. (1970): "Some studies concerning perception of isolated vowels", *Speech Transm. Lab.–Quart. Prog. Stat. Rep.* 2–3/1970, 19–35. Royal Inst. Techn., Stockholm.

CARRELL, T. D. (1981): "Effects of glottal waveform on the perception of talker sex", *Journ. Acoust. Soc. Am.* **70** Suppl., S97.

CHENG, Y. M., GUERIN, B. (1987a): "Control parameters in male and female glottal sources", in: T. Bear, C. Sasaki, K. S. Harris (eds.), *Laryngeal function in phonation and respiration*, 219–238. Little, Brown and Comp., Boston.

CHENG, Y. M., GUERIN, B. (1987b): "Study of the source–filter interactive concept and its application to male and female speech synthesis", *Bulletin du Laboratoire de la Communication Parlée* 1A, 29–66. Institut de la Communication Parlée, Grenoble.

COLEMAN, R. O. (1971): "Male and female voice quality and its relationship to vowel formant frequencies", *Journ. Sp. Hearing Res.* **14**, 565–577.

COLEMAN, R. O. (1973): "A comparison of the contributions of two vocal characteristics to the perception of maleness and femaleness in the voice", *Speech Transm. Lab.–Quart. Prog. Stat. Rep.* 2–3/1973, 13–22. Royal Inst. Techn., Stockholm.

COLEMAN, R. O. (1976): "A comparison of the contributions of two vocal characteristics to the perception of maleness and femaleness in the voice", *Journ. Sp. Hearing Res.* **19**, 168–180.

DAVID, H. A. (1969): *The method of paired comparisons.* Charles Griffin & Company. London (2nd edition).

DECHOVITZ, D. R. (1977): "Information conveyed by vowels: a negative finding", *Journ. Acoust. Soc. Am.* **61**, Suppl.1, S39.

DELATTRE, P., LIBERMAN, A. M., COOPER, F. S., GERSTMAN, L. J. (1952): "An experimental study of the acoustic determinants of vowel color; Observations on one– and two– formant vowels synthesized from spectrographic patterns", *Word* **8**, 195–210.

DIEHL, R. L., BUCHWALD McCUSKER, S., CHAPMAN, L. S. (1981): "Perceiving vowels in isolation and in consonantal context", *Journ. Acoust. Soc. Am.* **69**, 239–248.

DUDLEY, H. (1940): "The carrier nature of speech", *Bell System Techn. Journ.* 19, 495–515.

FANT, G. (1966): "A note on vocal tract size factors and non–uniform f–pattern scaling", *Speech Transm. Lab.–Quart. Prog. Stat. Rep.* 4/1966, 22–30. Royal Inst. Techn., Stockholm.

FANT, G. (1973): *Speech sounds and features.* MIT–Press, Cambridge.

FANT, G. (1975): "Non uniform vowel normalization", *Speech Transm. Lab.–Quart. Prog. Stat. Rep.* 2–3/1975, 1–19. Royal Inst. Techn., Stockholm.

FANT, G. (1986): "Glottal flow: models and interaction", *Journal of Phonetics* **14**, 393–399.

FANT, G., ANANTHAPADMANABHA, T. V. (1982): "Truncation and superposition", *Speech Transm. Lab.–Quart. Prog. Stat. Rep.* 2–3/1982, 1–17. Royal Inst. Techn., Stockholm.

FANT, G., CARLSON, R., GRANSTRÖM, B. (1974): "The [e]-[ø] ambiguity", *Spech Communication Seminar* **3**, 117-121. Almqvist & Wiksell, Stockholm.

FANT, G., LILJENCRANTS, J., LIN, Q. (1985): "A four-parameter model for glottal flow", *Speech Transm. Lab.-Quart. Prog. Stat. Rep.* 4/1985, 1-13. Royal Inst. Techn., Stockholm.

FANT, G., LIN, Q. (1987): "Glottal source-vocal tract acoustic interaction", *Speech Transm. Lab.-Quart. Prog. Stat. Rep.* 1/1987, 13-27. Royal Inst. Techn., Stockholm.

FANT, G., TATHAM, M. A. A. (eds.) (1975): *Auditory analysis and perception of speech.* Academic Press, London.

FLANAGAN, J. L. (1972): *Speech analysis, synthesis and perception.* Springer Verlag, Berlin, Heidelberg, New York (2nd Edition).

FLANAGAN, J. L., ISHIZAKA, K., SHIPLEY, K. L. (1975): "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", *Bell System Techn. Journ.* **54**, 485-506.

FLANAGAN, J. L., LANDGRAF, L. (1968): "Self-oscillating source from vocal tract synthesizers", *IEEE Transactions on Audio and Electroacoustics* **AU-16**, 57-64. Nachdruck in: FLANAGAN und RABINER (1973), 140-147.

FLANAGAN, J. L., RABINER, L. R. (eds.) (1973): *Speech synthesis.* Dowden, Hutchinson & Ross, Stoudsburg, Pennsilvania.

FUJISAKI, H., KAWASHIMA, T. (1968): "The roles of pitch and higher formants in the perception of vowels", *IEEE Transactions on Audio and Electroacoustics* **AU-16**, 73-77.

GERSTMAN, L. J. (1968): "Classification of self-normalized vowels", *IEEE Transactions on Audio and Electroacoustics* **AU-16**, 78-80.

GREISBACH, R. (1986): "M4: Ein Produktionsmodell mit einer allgemein-phonetischen Steuerung", *IPKöln-Berichte* **13**, 41-58. Inst. f. Phonetik, Köln.

GROβ, U. (1987): "Phonetische Aspekte des sprachlichen Verhaltens von Frauen", *IPKöln-Berichte* **14**, 47-104. Inst. f. Phonetik, Köln.

HEIKE, G. (1980): "Artikulatorische Synthese: Beschreibung des Programmpakets ARSDEM", *IPKöln-Berichte* **10**, 8-31. Inst. f. Phonetik, Köln.

HEIKE, G. (1982): "Quasiphonemische Einheiten und phonetische Regeln in einem Artikulationsmodell des Deutschen", in: T. Vennemann (ed.): *Silben, Segmente, Akzente*, 87-97. Max Niemeyer Verlag, Tübingen.

HELLER, K., ROSEMANN, B. (1974): *Planung und Auswertung empirischer Untersuchungen.* Ernst Klett Verlag, Stuttgart (1. Aufl.).

HELMHOLTZ, H. von (1896): *Die Lehre von den Tonempfindungen.* Leipzig (5. Aufl.).

HILGER, S. (1988): *Entwicklung eines Systems zur interaktiven Analyse und Synthese von Sprache unter besonderer Berücksichtigung der Glottissimulation.* Dissertation, Köln.

HOLMBERG, E. B., HILLMAN, R. E., PERKELL, J. S. (1988): "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voice", *Journ. Acoust. Soc. Am.* **84**, 511-529.

HOLMES, J. N., MATTINGLY, I. G., SHEARME, J. N. (1964): "Speech synthesis by rule", *Language and Speech* **7**, 127-140.

ISHIZAKA, K., FLANAGAN, J. L. (1972): "Synthesis of voiced sounds from a two-mass model of the vocal cords", *Bell System Techn. Journ.* **51**, 1233-1268. Nachdruck in: FLANAGAN und RABINER (1973), 148-183.

JOERGENSEN, H. P. (1969): "Die gespannten und ungespannten Vokale in der norddeutschen Hochsprache mit einer spezifischen Untersuchung der Struktur ihrer Formantfrequenzen", *Phonetica* **19**, 217-245.

JOOS, M. (1948): "Acoustic phonetics", *Language* **24**, Suppl. 1-136. Waverly Press, Baltimore.

KAHN, D. (1978): "On the identifiability of isolated vowels", *UCLA Working Papers in Phonetics* **41**, 26-31. University of California, Los Angeles.

KARLSSON, I. (1985): "Glottal waveforms for normal female speakers", *Speech Transm. Lab.-Quart. Prog. Stat. Rep.* 1/1985, 31-36. Royal Inst. Techn., Stockholm.

KARLSSON, I. (1988): "Glottal waveform parameters for different speaker types", *Speech Transm. Lab.-Quart. Prog. Stat. Rep.* 2-3/1988, 61-67. Royal Inst. Techn., Stockholm.

KELLY, J. L., LOCHBAUM, C. C. (1962): "Speech synthesis", *Proc. 4th Intern. Congr. Acoust.*, Paper G42, 1-4. Nachdruck in: FLANAGAN und RABINER (1973), 127-131.

KEMPELEN, W. von (1791): *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine.* Faksimile-Neudruck der Ausgabe Wien 1791. Friedrich Frommermann Verlag, Stuttgart.

KLATT, D. H. (1987): "Acoustic correlates of breathiness: first harmonic amplitude, turbulence noise, and tracheal coupling", *Journ. Acoust. Soc. Am.* **82**, Suppl. 1, S91.

KOIZUMI, T., TANIGUCHI, S., HIROMITSU, S. (1987): "Two-mass models of the vocal cords for natural sounding voice synthesis", *Journ. Acoust. Soc. Am.* **82**, 1179-1192.

KRATZENSTEIN, C. G. (1782): "Sur la naissance de la formation des voyelles", *Journal de Physique* **21**, 358-388.

LADEFOGED, P. (1967): *Three areas of experimental phonetics.* Oxford University Press, London.

LADEFOGED, P., BROADBENT, D. E. (1957): "Information conveyed by vowels", *Journ. Acoust. Soc. Am.* **29**, 98-104.

LADEFOGED, P., HARSHMAN, R., GOLDSTEIN, L., RICE, L. (1978): "Generating vocal tract shapes from formant frequencies", *Journ. Acoust. Soc. Am.* **64**, 1027-1035.

LASS, N. J., HUGHES, K. R., BOWYER, M. D., WATERS, L. T., BOURNE, V. T. (1976): "Speaker sex identification from voiced, whispered, and filtered vowels", *Journ. Acoust. Soc. Am.* **59**, 675-678.

LAVER, J. (1980): *The phonetic description of voice quality.* Cambridge University Press, Cambridge.

LIEBERMAN, P. (1973): "On the evolution of human language", in: Anderson, S., Kiparsky, P. (eds.): *A Festschrift for Morris Halle*, 107-120. Holt, Rinehart and Winston Inc., New York.

LIEBERMAN, P., BLUMSTEIN, S. E. (1988): *Speech physiology, speech perception, and acoustic phonetics.* Cambridge University Press. Cambridge.

LILJENCRANTS, J. (1985): *Speech synthesis with a reflection-type line analog.* Dissertation. Royal Institute of Technology, Stockholm.

LINDBLOM, B. (1963): "Spectrographic study of vowel reduction", *Journ. Acoust. Soc. Am.* **35**, 1773-1781.

LINDNER, G. (1981): *Grundlagen und Anwendungen der Phonetik.* Akademie Verlag, Berlin.

LINGGARD, R. (1985): *Electronic synthesis of speech.* Cambridge University Press, Cambridge.

LOBANOV, B. M. (1971): "Classification of Russian vowels spoken by different speakers", *Journ. Acoust. Soc. Am.* **49**, 606-608.

MACCHI, M. J. (1980): "Identification of vowels spoken in isolation versus vowels spoken in consonantal context", *Journ. Acoust. Soc. Am.* **68**, 1636-1642.

MARKEL, J. D., GRAY, A. H. (1976): *Linear prediction of speech.* Springer Verlag, Berlin, Heidelberg, New York.

MARKEL, N. N., PREBOR, L. D., BRAND, J. F. (1972): "Biosocial factors in dyadic communication: sex and speaking intensity", *Journal of Personality and Social Psychology* **23**, 11-13.

MATTINGLY, G. (1966): "Speaker variation and vocal-tract size", *Journ. Acoust. Soc. Am.* **39**, 1219.

MEYER, E., GUICKING, D. (1974): *Schwingungslehre.* Friedr. Vieweg & Sohn, Braunschweig.

MEYER-EPPLER, W. (1969): *Grundlagen und Anwendungen der Informationstheorie.* Springer Verlag, Berlin, Heidelberg, New York (2. neubearbeitete Aufl.).

MILLER, R. L. (1953): "Auditory tests with synthetic vowels", *Journ. Acoust. Soc. Am.* **25**, 114-121.

MONSEN, R. B., ENGEBRETSON, A. M. (1977): "Study of variations in the male and female glottal wave", *Journ. Acoust. Soc. Am.* **62**, 981-993.

MONSEN, R. B., ENGEBRETSON, A. M., VEMULA, N. R. (1978): "Indirect assessment of the contribution of subglottal air pressure and vocal-fold tension to changes of fundamental frequency in English", *Journ. Acoust. Soc. Am.* **64**, 65-80.

MURRY, T., SINGH, S. (1980): "Multidimensional analysis of male and female voices", *Journ. Acoust. Soc. Am.* **68**, 1294-1300.

NEAREY, T. M. (1977): *Phonetic feature systems for vowels.* Ph. D. thesis, Univ. Conneticut. Reproduced 1978 by the Indiana Univ. Linguistic Club, Bloomington.

NORD, L., ANANTHAPADMANABHA, T. V., FANT, G. (1984): "Signal analysis and perceptual tests of vowel responses with an interactive source filter model", *Speech Transm. Lab.-Quart. Prog. Stat. Rep.* 2-3/1984, 25-52. Royal Inst. Techn., Stockholm.

NORDSTRÖM, P. E. (1975): "Attempts to simulate female and infant vocal tracts from male area functions", *Speech Transm. Lab.-Quart. Prog. Stat. Rep.* 2-3/1975, 20-33. Royal Inst. Techn., Stockholm.

PAPÇUN, G. (1980): "How do different speakers say the same vowels? Discriminant analyses of four imitation dialects", *UCLA Working Papers in Phonetics* **48**, 1-140. University of California, Los Angeles.

PASCAL, D. (1985): "Qualité de la voix et caractéristiques individuelles", *Bulletin de l' Institut de Phonétique de Grenoble* **14**, 73-127. Grenoble.

PETERSON, G. E., BARNEY, H. L. (1952): "Control methods used in a study of the vowels", *Journ. Acoust. Soc. Am.* **24**, 175–184.

PLOMP, R. (1976): *Aspects of tone sensation.* Academic Press, London.

POLS, L. C. W., VAN DER KAMP, L. J. T., PLOMP, R. (1969): "Perceptual and physical space of vowel sounds", *Journ. Acoust. Soc. Am.* **46**, 458–467.

ROSENBERG, A. E. (1971): "Effect of glottal pulse shape on the quality of natural vowels", *Journ. Acoust. Soc. Am.* **49**, 583–590.

ROTHENBERG, M. (1973): "A new inverse filtering technique for deriving the glottal air flow waveform during voicing", *Journ. Acoust. Soc. Am.* **53**, 1632–1645.

ROTHENBERG, M. (1981): "An interactive model for the voice source", *Speech Transm. Lab.-Quart. Prog. Stat. Rep.* 4/1985, 1–17. Royal Inst. Techn., Stockholm.

SCHWARTZ, M. F. (1968): "Identification of speaker sex from isolated, voiceless fricatives", *Journ. Acoust. Soc. Am.* **43**, 1178–1179.

SCHWARTZ, M. F., RINE, H. E. (1968): "Identification of speaker sex from isolated, whispered vowels", *Journ. Acoust. Soc. Am.* **44**, 1736–1737.

SICKERT, K. (ed.) (1983): *Automatische Spracheingabe und Sprachausgabe.* Verlag Markt und Technik, Haar bei München.

SINGH, S., MURRY, T. (1978): "Multidimensional classification of normal voice qualities", *Journ. Acoust. Soc. Am.* **64**, 81–87.

SLAWSON, A. W. (1968): "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency", *Journ. Acoust. Soc. Am.* **43**, 87–101.

STEVENS, K. N. (1969): "The quantal nature of speech: Evidence from articulatory-acoustic data", in: David, E. E., Denes, P. B. (eds.): *Human communication: a unified view*, 51–66. McGraw-Hill, New York.

STEVENS, K. N., HOUSE, A. S. (1961): "An acoustical theory of vowel production and some of its implications", *Journ. Sp. Hearing Res.* **4**, 303–320.

STIEFEL, E. (1976): *Einführung in die numerische Mathematik.* Teubner Verlag, Stuttgart (5. Aufl.).

STRANGE, W., JENKINS, J. J., JOHNSON, T. L. (1983): "Dynamic specification of coarticulated vowels", *Journ. Acoust. Soc. Am.* **74**, 695–705.

STRANGE, W., VERBRUGGE, R. R., SHANKWEILER, D. P., EDMAN, T. R. (1976): "Consonant environment specifies vowel identity", *Journ. Acoust. Soc. Am.* **60**, 213–224.

SUNDBERG, J., JOHANSSON, C., WILBRAND, H., YTTERBERGH, C. (1987): "From sagittal distance to area", *Phonetica* **44**, 76–90.

SUOMI, K. (1984): "On talker and phoneme information conveyed by vowels: a whole spectrum approach to the normalization problem", *Speech Communication* **3**, 199–209.

THOMPSON, C. L., HOLLIEN, H. (1970): "Some contextual effects on the perception of synthetic vowels", *Language and Speech* **13**, 1–13.

TITZE, I. (1973): "The human vocal cords: a mathematical model, Part I", *Phonetica* **28**, 129–170.

TITZE, I. (1974): "The human vocal cords: a mathematical model, Part II", *Phonetica* 29, 1-21.

TITZE, I. (1984): "Parameterization of the glottal area, glottal flow, and vocal fold contact area", *Journ. Acoust. Soc. Am.* 75, 570-580.

TITZE, I. (1986): "Glottal flow models", *Journal of Phonetics* 14, 405-406.

VERBRUGGE, R. R., STRANGE, W., SHANKWEILER, D. P., EDMAN, T. R. (1976): "What information enables a listener to map a talker's vowel space?", *Journ. Acoust. Soc. Am.* 60, 198-212.

WAKITA, H. (1977): "Normalization of vowels by vocal-tract length and its application to vowel identificaton", *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-25, 183-192.

WEINBERG, B., BENNETT, S. (1971): "Speaker sex recognition of 5- and 6-year-old children's voices", *Journ. Acoust. Soc. Am.* 50, 1210-1213.

WILLIAMS, D. R. (1987): *The role of dynamic information in the perception of coarticulated vowels.* Dissertation. Univ. Microfilms Internat., Ann Arbor.

ZWICKER, E. (1982): *Psychoakustik.* Springer Verlag, Berlin, Heidelberg, New York.

L E B E N S L A U F

10. 4. 1959	Geburt in Osnabrück als erster Sohn des Dipl.-Ing. Bernhard Kröger und seiner Frau Marie-Luise
1965 - 1968	Marienschule Georgsmarienhütte
1968 - 1977	Gymnasium Oesede
Mai 1977	Abitur
1977 - 1978	Wehrdienst in Rheine und Delmenhorst
1979 - 1985	Physikstudium an der Universität Münster
Okt. 1983	Diplomhauptprüfung
1983 - 1985	Anfertigung der Diplomarbeit am Institut für Metallforschung der Universität Münster
April 1985	Erlangung des akademischen Grades des Diplom-Physikers
1985 - 1989	Studium der Phonetik, Musikwissenschaft und Physik an der Universität zu Köln
seit 1987	Wissenschaftlicher Mitarbeiter am Institut für Phonetik der Universität zu Köln
Aug. 1989	Promotionsprüfung