

3D-Artikulatorische Sprachsynthese

Dissertation

zur

Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

der Fakultät für Informatik und Elektrotechnik

der Universität Rostock

vorgelegt von

Peter Birkholz

aus Rostock

Gutachter: Prof. Dr.-Ing. habil. Dietmar Jackel
Universität Rostock, Institut für Informatik

Prof. Dr.-Ing. habil. Erika Müller
Universität Rostock, Institut für Nachrichtentechnik

Prof. Dr. Walter F. Sendlmeier
*Technische Universität Berlin,
Institut für Sprache und Kommunikation*

Tag der Verteidigung: 20. Oktober 2005

Zusammenfassung

Diese Arbeit beschreibt ein artikulatorisches Sprachsynthesesystem, das in der Lage ist, synthetische Sprachäußerungen in hoher Qualität zu generieren. Das System umfasst ein Modell des Sprechapparats, ein aerodynamisch-akustisches Simulationsverfahren und ein Steuermodell für die Generierung der artikulatorischen Bewegungsabläufe. Der Sprechapparat wird als ein verzweigtes Rohrsystem modelliert, das aus Abschnitten für die Lunge und Luftröhre, die Stimmritze und den Rachen-, Mund- und Nasenraum zusammengesetzt ist. Dieses Rohrmodell bildet die Grundlage für das aerodynamisch-akustische Simulationsverfahren, das gegenüber den meisten konventionellen Verfahren eine kontinuierliche Änderung der Rohrlänge erlaubt und ein theoretisch basiertes Verfahren für die Simulation von Friktionsrauschen einschließt. Die Rohrgeometrie des Rachen- und Mundraums wird mit Hilfe eines dreidimensionalen geometrischen Modells des Vokaltrakts berechnet, das im Rahmen dieser Arbeit neu entwickelt wurde. Das Steuermodell basiert auf dem Konzept der artikulatorischen Geste und ermöglicht in einfacher Weise die Simulation koartikulatorischer Phänomene. Jede Äußerung wird als eine Abfolge von Gesten definiert und vom Steuermodell in die Zeitfunktionen der artikulatorischen Parameter überführt.

Abstract

This work describes an articulatory speech synthesizer that is capable of producing speech utterances in high quality. The system contains a model of the speech apparatus, an aerodynamic-acoustic simulation and a control model for the generation of articulatory movements. The speech apparatus is modelled as a branched tube system, that comprises sections for the lung and trachea, the glottis, and the pharynx, mouth and nasal cavities. This tube model is the basis for the aerodynamic-acoustic simulation method, that – compared to most of the conventional approaches – allows for continual length changes of the tube, and includes a theoretically based method for the simulation of friction and aspiration. The tube geometry of the pharynx and mouth cavities are calculated by means of a three-dimensional vocal tract model, that was especially developed for this purpose. The control model is based on the concept of articulatory gestures and facilitates the simulation of coarticulatory phenomena. Each utterance is defined by means of a gestural score and is transformed into the trajectories of the articulatory parameters by the control model.

Danksagung

An dieser Stelle möchte ich allen Personen danken, die direkt oder indirekt zum Gelingen dieser Arbeit beigetragen haben. Prof. Dietmar Jackel danke ich für die Betreuung der Arbeit und für die große Unterstützung bei der Veröffentlichung der Ergebnisse. Ich möchte mich auch besonders bei Ralf Winkler, Sascha Fagel und Karl Schnell für die zahlreichen Gespräche bedanken, aus denen nicht selten Anregungen für die vorliegende Arbeit hervorgegangen sind. Astrid Päschke und Prof. Walter Sendlmeier danke ich dafür, dass sie mir die Emotionssprachdatenbank des Instituts für Sprache und Kommunikation der TU Berlin für eigene Untersuchungen zur Verfügung gestellt haben. Diese Daten haben zur Verbesserung einiger Analysewerkzeuge beigetragen, die ich im Rahmen der Promotion entwickelt und eingesetzt habe. Frau Prof. Ursula Kantorczyk und Frau Dr. Mannewitz vom Institut für Slavistik an der philosophischen Fakultät haben das polnische Begleitbuch zu einer alten Sammlung von Röntgenaufnahmen des Vokaltrakts durchgearbeitet und dazu viele meiner Fragen beantwortet. Vielen Dank dafür! Prof. Bernd Kröger danke ich ganz herzlich für die vielen Hinweise, die zur Verbesserung dieses Manuskripts beigetragen haben. Meiner lieben Freundin Mandy-Nicole Gensow gilt meine Dankbarkeit besonders für ihre Bereitschaft zu zahlreichen Sprachaufnahmen und für ihre kritische Beurteilung der Syntheseergebnisse. Auch Katja Ihsberner danke ich für ihre Teilnahme bei Sprachaufnahmen. Meinem restlichen Freundeskreis und meiner Familie möchte ich für die gemeinsame vergnügliche Zeit zwischendurch, die Gelegenheit zur Zerstreuung bot, danken. Ebenfalls gilt mein Dank Prof. Erika Müller und Prof. Walter Sendlmeier für ihre Bereitschaft, die Arbeit zu begutachten.

Inhaltsverzeichnis

1	Einleitung	1
1.1	Motivation	1
1.2	Mechanismus der Spracherzeugung	2
1.3	Methoden der Sprachsynthese	5
1.4	Artikulatorische Sprachsynthese	7
1.5	Beitrag und Aufbau der Arbeit	9
2	Modellierung des Sprechapparats	11
2.1	Einleitung	11
2.2	Literaturüberblick	12
2.3	Geometrische Modellierung des Vokaltrakts	17
2.3.1	Grundkonzepte	18
2.3.2	Geometrie und Parameter des Vokaltrakts	19
2.3.3	Berechnung der Mittellinie und der Flächenfunktion	34
2.4	Nasenraum	38
2.5	Stimmlippen	39
2.6	Subglottale Luftwege	43
2.7	Zusammenfassung und Diskussion	44
3	Aerodynamisch-akustische Simulation	47
3.1	Grundlagen	47
3.2	Literaturüberblick	48
3.3	Grundgleichungen	49
3.3.1	Impulssatz	49
3.3.2	Kontinuitätsgleichung	51
3.3.3	Zustandsgleichung	53
3.3.4	Schwingungsgleichung für die Rohrwand	53
3.4	Verluste im Sprechtrakt	54
3.4.1	Verluste durch Reibung an der Rohrwand	54
3.4.2	Verluste durch Wirbelbildung	57
3.4.3	Verluste durch Schallabstrahlung	61
3.4.4	Verluste durch elastische Rohrwände	63
3.5	Der Sprechapparat als akustisches Netzwerk	64
3.5.1	Örtliche Diskretisierung der Grundgleichungen	64
3.5.2	Verluste und Energiequellen	67

3.5.3	Netzwerkstruktur des gesamten Sprechapparats	69
3.6	Simulation im Frequenzbereich	72
3.6.1	Theoretische Grundlagen	73
3.6.2	Vierpolanalyse des Netzwerks für den Sprechapparat	76
3.7	Simulation im Zeitbereich	80
3.7.1	Methoden der zeitdiskreten Approximation	80
3.7.2	Diskretisierung der Netzwerkgleichungen	82
3.7.3	Simulation	86
3.7.4	Vergleich mit der Simulation im Frequenzbereich	88
3.8	Modellierung sekundärer Schallquellen	93
3.8.1	Mechanismus der Sekundärschallerzeugung	93
3.8.2	Flächenfunktionen und Rauschquellen stimmloser Frikative	95
3.8.3	Modell für die Rauschquellen in dieser Arbeit	98
3.9	Zusammenfassung und Diskussion	100
4	Steuerung der Artikulation	103
4.1	Einleitung	103
4.2	Steuermodell dieser Arbeit	105
4.2.1	Grundprinzip	105
4.2.2	Supraglottale Steuerung	106
4.2.3	Laryngale und pulmonale Steuerung	111
4.2.4	Gesturaler Ablaufplan	114
4.2.5	Technische Realisierung	115
4.3	Zusammenfassung und Diskussion	115
5	Synthese	117
5.1	Zielformen des Vokaltrakts	117
5.1.1	Erstellung einer Lauttabelle	117
5.1.2	Evaluierung des Dominanzmodells anhand von Röntgenbildern	121
5.2	Regeln für die Erstellung gesturaler Ablaufpläne	124
5.2.1	Vokale	125
5.2.2	Sonoranten	126
5.2.3	Obstruenten	128
5.2.4	Synthese längerer Äußerungen	133
5.3	Zusammenfassung und Diskussion	134
6	Zusammenfassung und Diskussion	137
6.1	Zusammenfassung der Beiträge	137
6.2	Diskussion der Ergebnisse	139
6.3	Ausblick	139
A	Programmüberblick	141
B	Konstanten und Symbole	147
B.1	Physikalische Konstanten	147
B.2	Phonemsymbole des Deutschen nach IPA und SAMPA	147

C Lebenslauf, Veröffentlichungen, Vorträge	151
Literaturverzeichnis	153

Kapitel 1

Einleitung

1.1 Motivation

Die verbale Verständigung ist für die meisten Menschen die wichtigste Form der Kommunikation. Durch die zunehmende Integration technischer Geräte in unseren Alltag, insbesondere des Computers, wird auch die Erforschung von Methoden für eine natürliche Mensch-Maschine-Kommunikation immer wichtiger. Ein Teilbereich davon ist die verbale Kommunikation, die u. a. die maschinelle Erkennung und Synthese von Sprache erfordert. Diese Arbeit widmet sich der Sprachsynthese.

Sprachsynthetisatoren werden heute bereits vielfach in verschiedenen Anwendungen eingesetzt. Das klassische Anwendungsfeld ist das automatische Vorlesen von geschriebenen Texten, seien es Geschichten, e-mails, Nachrichten oder Wettervorhersagen. Solche Anwendungen werden als textgesteuerte Sprachsynthese (*text-to-speech*, TTS) bezeichnet. Darüber hinaus ist die Sprachsynthese auch ein wesentlicher Bestandteil automatischer Dialogsysteme, z. B. von automatischen Bestell-Hotlines über das Telefon. Weitere Anwendungsfelder, die in Zukunft an Bedeutung zunehmen werden, sind Übersetzungssysteme für gesprochene Sprache (z. B. zwischen Deutsch und Englisch) und die natürliche Kommunikation zwischen Menschen und Servicerobotern.

Aktuelle kommerzielle Sprachsynthesysteme basieren fast ausschließlich auf der Verkettung (Konkatention) vorab aufgenommener Sprachbausteine. Mit dieser Synthesetechnik hat man bereits Synthetisatoren entwickelt, die relativ verständlich und natürlich klingen. Da diese Systeme aber mit den vorab aufgenommenen Bausteinen arbeiten, sind sie nicht sehr flexibel. So bereitet z. B. die Synthese generischer Stimmen, unterschiedlicher Emotionen und anderer Sprachen als die, die zuvor aufgenommen wurden, große Schwierigkeiten. In dieser Arbeit verfolge ich den Ansatz, den Mechanismus der Spracherzeugung auf der physiologischen und akustischen Ebene direkt zu simulieren. Diese Art der Synthese wird als *artikulatorische Sprachsynthese* bezeichnet und unterliegt den Einschränkungen der konkatenativen Syntheseverfahren prinzipiell nicht.

Die Forschung zur künstlichen Erzeugung von Sprache reicht bereits mehrere Jahrhunderte zurück. Die frühesten dokumentierten mechanischen Apparate für die Erzeugung einzelner Sprachlaute und zusammenhängender Äußerungen wurden von KRATZENSTEIN (1779) und VON KEMPELEN (1791) entwickelt. Im frühen 20. Jahrhundert wurden dann die ersten Sprachsynthetisatoren als analoge elektrische Schaltungen aufgebaut. Die mechani-

schen und elektrischen Apparate wurden dabei jeweils von einem Menschen „gesteuert“ bzw. ähnlich einem Musikinstrument gespielt, um die gewünschten Laute hervorzubringen. Erst die Entwicklung der Computertechnik hat zu den ersten Synthetisatoren geführt, die künstliche Sprache ausgehend von geschriebenem Text erzeugen konnten¹. Neben der rein computergestützten Simulation von Sprache gibt es aber auch heute wieder – in der Tradition der ersten Sprechapparate – Bemühungen zur mechanischen Nachbildung des Sprechtrakts für die Spracherzeugung [88]. Bei solchen Maschinen handelt es sich also um „echte“ artikulatorische Synthetisatoren, während die Spracherzeugung in dieser Arbeit mit Hilfe geeigneter artikulatorischer und akustischer Computermodelle simuliert wird.

Der Rest dieses Kapitels ist wie folgt gegliedert. Zunächst gibt Abschnitt 1.2 einen kurzen Überblick über den Mechanismus der Spracherzeugung. Anschließend werden in Abschnitt 1.3 die wichtigsten (computerbasierten) Methoden der Sprachsynthese vorgestellt und verglichen. Abschnitt 1.4 geht näher auf die Grundlagen der artikulatorischen Sprachsynthese ein. Im letzten Abschnitt dieses Kapitels wird dann der Beitrag der vorliegenden Arbeit zu diesem Forschungsfeld dargelegt.

1.2 Mechanismus der Spracherzeugung

Sprache ist das akustische Endprodukt kontrollierter Bewegungen des menschlichen Sprechapparats. Als Sprechapparat wird in dieser Arbeit das gesamte System aus der Lunge, den Bronchien, der Luftröhre, den Stimmlippen, und dem Rachen-, Mund- und Nasenraum bezeichnet. Die Stimmlippen bzw. die durch sie begrenzte Stimmritze (Glottis) unterteilen den Sprechapparat in das subglottale und das supraglottale System. Das supraglottale System besteht aus dem Rachen-, Mund- und Nasenraum, und das subglottale System aus der Lunge, den Bronchien und der Luftröhre. Der Rachen und der Mundraum bilden zusammen den Vokaltrakt, der auch als Ansatzrohr bezeichnet wird. Das gesamte supraglottale System aus dem Vokaltrakt und dem Nasenraum wird in dieser Arbeit als Sprechtrakt bezeichnet und ist schematisch in Abb. 1.1 dargestellt. Aus akustischer Sicht handelt es sich beim Sprechapparat um ein verzweigtes Rohrsystem, dass sich von den Lungenbläschen bis zur Stimmritze und von der Stimmritze bis zur Mundöffnung und den Nasenlöchern erstreckt. In einigen Bereichen, z. B. in der Luftröhre, ist der Rohrquerschnitt weitestgehend konstant, während er in anderen Bereichen, insbesondere im Rachen- und Mundraum, von den Artikulatoren (Zunge, Velum, Lippen, Unterkiefer) bestimmt wird. Im Vokaltrakt kann der Rohrquerschnitt von 0 cm^2 (kompletter Verschluss) bis etwa 20 cm^2 variiert werden. Der Nasenraum ist ein Seitenzweig des Vokaltrakts und erstreckt sich vom Velum bis zu den Nasenlöchern. Die akustische Kopplung zwischen dem Vokaltrakt und dem Nasenraum wird durch die Größe der Öffnung im Bereich des Velums zwischen dem Rachen- und Nasenraum gesteuert. Diese Öffnung wird als velopharyngale Pforte bezeichnet und ist in Abb. 1.1 für einen geschlossenen Zustand (durchgezogener Umriss) und einen offenen Zustand (gestrichelter Umriss) dargestellt. Für einen männlichen Erwachsenen kann die velopharyngale Öffnung zwischen 0 und etwa 5 cm^2 variiert werden.

Die Energie für die Spracherzeugung liefern Brust- und Bauchmuskulatur. Durch sie wird ein Überdruck in der Lunge aufgebaut, der wiederum einen Luftstrom durch den Sprech-

¹Für eine umfangreiche Darstellung der Anfänge der Sprachsyntheseforschung empfehle ich die Monographie von FLANAGAN [38].

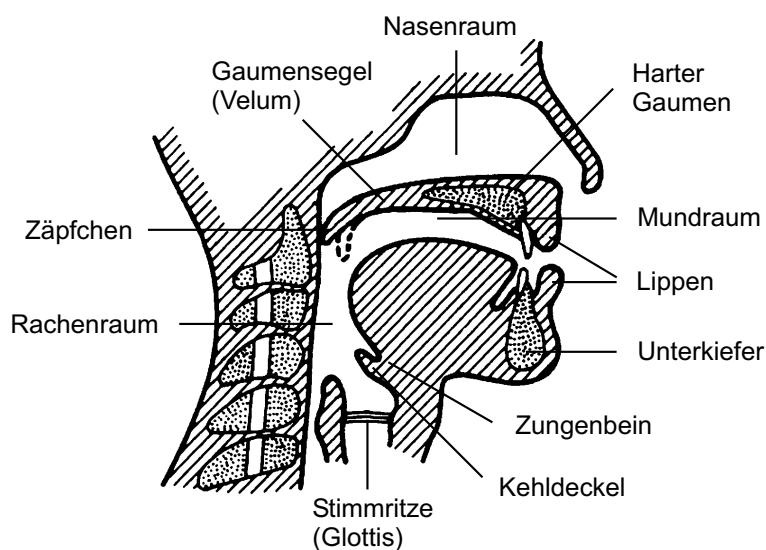


Abbildung 1.1: Schematische Darstellung des Sprechtrakts. Quelle: VARY *et al.* [123].

apparat verursacht. Beim Sprechen bildet der Mensch mit Hilfe der Artikulatoren an bestimmten Stellen des Sprechtrakts Konstriktionen aus, an denen ein Teil der Energie des Luftstroms in Schall umgeformt wird. Dieser wird im Sprechtrakt akustisch gefiltert und an der Mund- und/oder Nasenöffnung in die Umgebung abgestrahlt.

Man unterscheidet hauptsächlich drei Arten der Anregung des Sprechtrakts durch Schallquellen: Die stimmhafte, die stimmlose und die transiente Anregung. Die stimmhafte Anregung entsteht durch die Schwingung der Stimmlippen und wird als Phonation bezeichnet. Die Schwingung resultiert aus dem komplexen Zusammenspiel zwischen dem Luftstrom aus der Lunge und dem Gewebe der Stimmlippen und ist dadurch gekennzeichnet, dass sich die Glottis in schneller Abfolge quasi-periodisch öffnet und schließt. Während der offenen Phasen der Schwingungszyklen entweichen kurze Luftpulse aus dem subglottalen System in den Sprechtrakt, die sich dort als akustische Wellen ausbreiten.

Eine stimmlose Anregung des Sprechtrakts entsteht, wenn die Luft aus der Lunge bei geöffneten Stimmlippen eine Engstelle im Rachen- oder Mundraum durchströmt und dabei aufgrund von Reibung Turbulenzen entstehen. Die Turbulenzen erzeugen ein breitbandiges Rauschen, das sich als Schallwellen im Vokaltrakt ausbreitet. Die stimmhafte und stimmlose Anregung schließen einander nicht aus. Wenn beide gleichzeitig auftreten, so spricht man von gemischter Anregung.

Eine transiente Anregung des Sprechtrakts entsteht, wenn im supraglottalen System ein Verschluss gebildet wird, so dass sich die Luft dahinter aufstauen kann. Wenn der Verschluss plötzlich gelöst wird, entweicht die angestaute Luft schlagartig und erzeugt ein kurzes Plosionsgeräusch. Dieses ist sehr kurz und kann, gegenüber den Geräuschen der stimmhaften und stimmlosen Anregung, nicht über längere Zeit aufrechterhalten werden.

Die Schallquellen, die die unterschiedlichen Arten der Anregung verursachen, sind jeweils durch eine bestimmte spektrale Zusammensetzung gekennzeichnet. Grundsätzlich sind diese Quellen breitbandig, d. h. ihre Energie verteilt sich auf breite Frequenzbereiche. Im Sprechtrakt werden diese Schallquellen akustisch gefiltert. Dabei werden bestimmte Fre-

quenzen verstärkt und andere unterdrückt. Eine Verstärkung der spektralen Amplituden entsteht an den Resonanzfrequenzen des Vokaltrakts, deren Positionen von der Form des Sprechtrakts abhängen. Die Bereiche erhöhter Amplituden im Spektrum werden auch als Formanten bezeichnet. Im wichtigen Frequenzbereich bis 5 kHz besitzt der Vokaltrakt in der Regel 4-6 Formanten. Außerdem findet man in den Spektren einiger Laute Antiformanten, also Bereiche mit besonders niedrigen Amplituden. Antiformanten entstehen durch Seitenzweige bzw. Nebenhöhlen des Vokaltrakts (z. B. wenn der Nasenraum akustisch angekoppelt ist) und durch die hinter der Schallquelle liegenden Hohlräume des Sprechapparats.

In Abhängigkeit von der Art der Anregung und der Form des Sprechtrakts unterscheidet man verschiedene Lautklassen. In Abschnitt B.2 sind die Laute der deutschen Sprache nach Klassen geordnet aufgelistet.

Eine große Klassen von Lauten bilden die Vokale, wie z. B. [a], [i] und [u]. Sie werden normalerweise stimmhaft angeregt, während der Vokaltrakt eine durchgängige offene Passage bildet. Für die deutschen Vokale ist die velopharyngale Pforte dabei in der Regel nicht oder nur geringfügig geöffnet, so dass die stimmhafte Schallquelle hauptsächlich durch den Rachen- und Mundraum gefiltert wird. In Abhängigkeit von der Position der Artikulatoren werden dem Quellspektrum dadurch für jeden Vokal charakteristische Formanten aufgeprägt.

Eine andere große Klasse von Lauten bilden die Konsonanten, die sich in Frikative, Plosive, Nasale und Liquide einteilen lassen. Frikative sind durch eine stimmlose Anregung gekennzeichnet. Das Geräusch entsteht durch Turbulenzen an einer Engstelle im Vokaltrakt. Die Engstelle wird in Abhängigkeit vom konkreten Laut mit unterschiedlichen Artikulatoren und an unterschiedlichen Stellen im Vokaltrakt ausgebildet, z. B. mit der Zungenspitze am Zahndamm für das [s] oder mit der Unterlippe an den oberen Schneidezähnen für das [f]. Die Frikative können stimmhaft oder stimmlos erzeugt werden, d. h. mit oder ohne Beteiligung der stimmhaften Quelle an der Glottis.

Plosive werden erzeugt, indem ein Verschluss im Vokaltrakt gebildet und dahinter ein Überdruck aufgebaut wird. Dieser Verschluss wird dann plötzlich gelöst, so dass eine transiente Anregung entsteht. Damit der Überdruck aufgebaut werden kann, muss die velopharyngale Pforte fest verschlossen sein. Es werden stimmhafte und stimmlose Plosive unterschieden. Die stimmhaften Plosive [b, d, g] sind dadurch gekennzeichnet, dass die Schwingung der Stimmlippen während der oralen Verschlussphase noch teilweise aufrechterhalten wird. Bei den stimmlosen Plosiven [p, t, k] ist die Glottis dagegen während der gesamten oralen Verschlussphase geöffnet. Die Verschlüsse für die deutschen Plosive werden für [p, b] mit den Lippen, für [t, d] mit der Zungenspitze am Zahndamm und für [k, g] mit dem Zungenrücken am Gaumen gebildet.

Die Nasale sind durch eine stimmhafte Anregung, einen oralen Verschluss und eine geöffnete velopharyngale Pforte gekennzeichnet. Dadurch wird der Schall (abgesehen von geringen Anteilen über die Haut) von den Nasenlöchern abgestrahlt. Die Verschlusspositionen für die Nasale [m], [n] und [ɳ] entsprechen denen der Plosive [p], [t] und [k]. Das Spektrum nasaler Laute weist in der Regel mehrere ausgeprägte Antiresonanzen auf. Diese entstehen dadurch, dass der Mundraum zwischen dem Velum und dem oralen Verschluss einen Seitenzweig zum Schallpfad durch den Rachen- und Nasenraum bildet. Dieser Seitenzweig weist jeweils bestimmte Eigenfrequenzen auf, an denen er der Schallwelle im Rachen-Nasen-Raum Energie entzieht.

Die kleine Klasse der Liquide enthält im Deutschen die Laute [l] und [ʀ]. Sie sind durch eine stimmhafte Anregung, eine geschlossene velopharyngale Pforte und die Schallabstrahlung von der Mundöffnung gekennzeichnet. Beim [l] berührt die Zunge den Gaumen, jedoch sind die Zungenseiten abgesenkt, so dass weder ein Verschluss noch eine Rauschquelle aufgrund einer starken Konstriktion entsteht. Der Laut [ʀ] zählt zu den Vibranten und entsteht durch eine strömungsinduzierte flatternde Bewegung des Zäpfchens am Ende des weichen Gaumens.

Einige der Vokale und Konsonanten können so kombiniert werden, dass sie zusammen neue Grundeinheiten der Sprache bilden. Man unterscheidet hier Diphthonge und Affrikative. Diphthonge sind durch die Bewegung der Artikulatoren von einer Vokalposition zu einer anderen gekennzeichnet, z. B. in den Wörtern „Mai“ oder „Auto“. Affrikative sind dagegen die Kombination aus einem Plosiv und einem Frikativ, z. B. in den Wörtern „Pfanne“ und „Matsch“.

Für eine ausführlichere Einführung in die Spracherzeugung verweise ich auf die Bücher von VARY *et al.* [123] und STEVENS [110].

1.3 Methoden der Sprachsynthese

Die Methoden zur Sprachsynthese werden üblicherweise in parametrische (regelbasierte) Verfahren einerseits und datenbasierte (konkatenative) Verfahren andererseits untergliedert [123].

Am häufigsten werden gegenwärtig konkatenative Synthesysteme eingesetzt. In diesen Systemen wird das synthetische Sprachsignal direkt aus Bausteinen zusammengesetzt, die vorher aus natürlichen Sprachaufnahmen herausgeschnitten wurden. Als Bausteine werden in der Regel Phone (kleinste lautsprachliche Einheiten), Diphone (Zeitabschnitt von der Mitte eines Lautes bis zur Mitte des folgenden Lautes), Halbsilben, Silben, ganze Wörter oder Kombinationen aus diesen Elementen verwendet. Grundsätzlich ist dabei die Anzahl der benötigten Bausteine umso größer, je länger sie sind bzw. je mehr Lautkombinationen sie zulassen. Bei der Verkettung der Grundbausteine werden spezielle Techniken angewendet, um an den Verkettungsstellen Sprünge im Signal zu vermeiden. Zusätzlich werden die Grundbausteine vor der Verkettung hinsichtlich der prosodischen Parameter Grundfrequenz, Amplitude und Dauer manipuliert. Man versucht, diese nachträglichen Manipulationen an den Originalbausteinen möglichst gering zu halten, da sie sich in der Regel qualitätsmindernd auf das synthetische Sprachsignal auswirken. Der Trend geht deshalb dahin, sehr große Datenbasen mit Sprachbausteinen zu verwenden, in denen zu jedem Baustein jeweils mehrere Instanzen mit unterschiedlichen prosodischen Parametern abgespeichert sind. Der Schwerpunkt bei der Synthese liegt dann in der Auswahl der Bausteine, die am besten den prosodischen Anforderungen der Zieläußerung genügen und gleichzeitig die notwendigen Manipulationen an den Verkettungsstellen minimieren. Mit dieser Technik lässt sich sowohl eine hohe Verständlichkeit als auch eine gute Qualität der syntetischen Sprache erreichen (z. B. mit dem TTS-System Natural VoicesTM von AT&T). Die typische Größe der Datenbasis für eine einzelne Stimme beträgt dafür 100-500 MB.

Die wichtigsten Vertreter der parametrischen Syntheseverfahren sind die Formantsynthese und die artikulatorische Synthese. Die Formantsynthese stützt sich auf das Quelle-Filter-Modell [33] als Prinzip der Spracherzeugung. Dabei werden die Filtereigenschaften und

die Anregung des Sprechtrakts unabhängig voneinander modelliert, so als wenn sie sich nicht gegenseitig beeinflussen. Diese Annahme ist nicht ganz realistisch, da teilweise zeitvariante und nichtlineare Wechselwirkungen zwischen den Quellen und dem Sprechtrakt existieren [110]. Die Filtereigenschaften werden direkt durch die Formanten und Antiformanten definiert, während die Schallquellen für die stimmhafte und stimmlose Anregung durch einfache parametrisierte Funktionen bzw. weißes Rauschen simuliert werden. Einer der bekanntesten Formantsynthetisatoren wurde in den 1970er und 80er Jahren von DENNIS H. KLATT entwickelt [58, 59] und wurde als das Produkt DECTALK einer der ersten kommerziell verwendbaren Sprachsynthetisatoren.

Artikulatorische Synthetisatoren simulieren direkt den Prozess der Spracherzeugung und benötigen dazu ein Modell des Sprechapparats, ein Verfahren für die akustische Simulation und ein Konzept für die Steuerung der artikulatorischen Bewegungsabläufe. Im nächsten Abschnitt wird genauer auf diese Synthesetechnik eingegangen, die der Hauptgegenstand der vorliegenden Arbeit ist. Aufgrund der Komplexität artikulatorischer Synthetisatoren und dem damit verbundenen Aufwand bei der Erstellung solcher Systeme wurden sie bisher hauptsächlich für die Grundlagenforschung entwickelt. Die Erreichung einer hohen Synthesequalität mit solchen Systemen, die auch den Anforderungen im praktischen Einsatz genügt, hat bisher ein großes Problem dargestellt.

Um die in diesem Abschnitt vorgestellten Syntheseverfahren zu vergleichen, werde ich sie nachfolgend an den Anforderungen eines „idealen“ Synthetisators messen. In Anlehnung an SHADLE *et al.* [102] sollte die ideale Sprachsynthese

1. so verständlich sein wie ein Mensch.
2. natürlich klingen.
3. in der Lage sein, wie viele verschiedene Sprechertypen zu klingen: männlich, weiblich, alt, jung und dazwischen; wie eine hohe, tiefe, helle oder dunkle Stimme, usw.
4. in der Lage sein, in einer beliebigen Sprache zu sprechen.
5. wie ein bestimmter Sprecher klingen können, und nicht nur wie ein generischer.
6. wie ein außerordentlicher Sprecher klingen können, z. B. ein Sänger mit einem Stimmumfang von 7 Oktaven, wie jemand mit einer Sprachstörung, oder wie ein Alien mit zusätzlichen Nebenhöhlen im Sprechtrakt.
7. in der Lage sein, den Sprechertyp oder die Stimmqualität eines Sprechers zu ändern, ohne den gleichen Aufwand wie für die erste Stimme betreiben zu müssen.
8. intuitiv nachvollziehbare Parameter auf allen Ebenen besitzen, so dass dann, wenn etwas falsch oder schlecht klingt, die Ursache dafür schnell gefunden und der Fehler behoben werden kann.
9. uns neue Einsichten in die Spracherzeugung eröffnen, die über die Erstellung eines kommerziell verwertbaren Systems hinausgehen.

Ein konkatenativer Synthetisator mit einer hinreichend großen Datenbasis kann es schaffen, die ersten vier Anforderungen zu erfüllen. Ein Formantsynthetisator mit einem entsprechend großen Regelwerk kann diese vier Anforderungen theoretisch auch erfüllen, wobei

er der konkatenativen Synthese jedoch in den Punkten 2 und 3 unterlegen ist. Beide Synthetisatoren können die fünfte Anforderung erfüllen, wenn die spezielle Stimme für die konkatenative Synthese aufgenommen wurde bzw. in Form von Regeln für die Formantsynthese beschrieben wurde. Um die Punkte 6 bis 9 zu erreichen, ist die artikulatorische Sprachsynthese den anderen Verfahren deutlich überlegen. Auch wenn man bezüglich der 9. Anforderung von allen Syntheseverfahren etwas lernen kann, so ist der Gewinn bei den artikulatorischen Synthetisatoren durch den engen Bezug zur realen Spracherzeugung in der Regel am größten. Durch eine genaue Analyse der Vorteile, Nachteile und Möglichkeiten der unterschiedlichen Verfahren kommen SHADLE *et al.* [102] zu dem Schluss, dass die artikulatorischen Synthetisatoren langfristig gesehen den anderen Ansätzen klar überlegen sind.

1.4 Artikulatorische Sprachsynthese

Die artikulatorische Sprachsynthese beruht auf der Simulation der bei der Sprachproduktion ablaufenden Prozesse. Dazu wird einerseits ein Modell des Sprechapparats benötigt und andererseits Methoden zur Steuerung der Artikulation und zur Simulation der Akustik und Aerodynamik. Die Modellierung realer Vorgänge erfordert stets ein gewisses Maß an Abstraktion, damit die Modelle und Simulationen beherrschbar, nachvollziehbar und in einem vernünftigen Zeitrahmen berechnbar bleiben. Besonders hinsichtlich der aerodynamischen und akustischen Simulation werden in der Regel die folgenden vereinfachenden Annahmen gemacht:

1. Die Schallwellen im Sprechtrakt breiten sich als *ebene Wellen* aus. Dadurch muss das Schallfeld nur entlang einer Dimension simuliert werden, und zwar entlang der Längsachse des Sprechtrakts. Diese Annahme ist bis ca. 4 kHz gut erfüllt. Für höhere Frequenzen werden die Schallwellenlängen zunehmend kürzer, so dass an Stellen im Sprechtrakt mit einem großen Querschnitt auch stehende *Querwellen* entstehen können.
2. Die Biegung des Vokaltrakts zwischen dem Rachen- und Mundraum ist aus akustischer Sicht vernachlässigbar, d. h. der Vokaltrakt kann durch ein gerades Rohr angenähert werden. SONDHI [105] hat nachgewiesen, dass die Abweichungen der Formantfrequenzen zwischen einem geraden und einem stark gebogenen Rohr mit den Abmessungen des Vokaltrakts nur zwischen 4% für den 1. Formanten und 8% für den 4. Formanten liegen.
3. Der Sprechapparat kann durch ein Rohr repräsentiert werden, dass aus vielen kurzen Rohrabschnitten mit jeweils konstantem Querschnitt zusammengesetzt ist. Meistens werden zylindrische Rohrabschnitte verwendet. Hier gilt: Je mehr Rohrabschnitte für die diskrete Approximation eingesetzt werden, desto genauer wird die akustische Simulation. Praktisch genügen für den Vokaltrakt 20-40 Rohrabschnitte. Eine größere Anzahl bringt nur noch minimale Verbesserungen der Genauigkeit im Frequenzbereich bis 5 kHz.

Mit diesen Annahmen kann der Sprechtrakt als ein verzweigtes Rohr mit einem stückweise konstanten Querschnittsverlauf modelliert werden, wie es schematisch in Abb. 1.2

dargestellt ist. Den wichtigsten Einfluss auf das akustische Verhalten des Rohrmodells hat

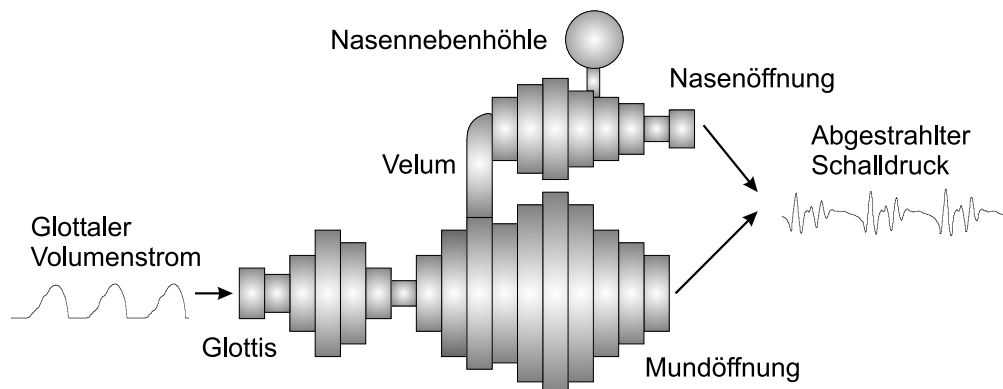


Abbildung 1.2: Rohrmodell des Vokaltrakts.

der Verlauf der Querschnittsfläche entlang der Rohrachse. Dieser Verlauf kann durch eine Funktion $A(x)$ beschrieben werden, wobei x die Position auf der Rohrachse ist. Nachfolgend wird $A(x)$ als *Flächenfunktion* (engl.: *area function*) bezeichnet.

Ein artikulatorischer Sprachsynthesator ist in der Regel wie folgt aufgebaut. Es gibt ein geometrisches, biomechanisches oder statistisches Modell, durch das der Sprechtrakt entweder zweidimensional als Sagittalschnitt oder in seiner vollständigen dreidimensionalen Form repräsentiert wird. Die Form des Sprechtrakts wird durch die Modellartikulatoren bestimmt, die sich mit Hilfe mehrerer Parameter steuern lassen. Um eine bestimmte Äußerung zu synthetisieren, werden zuerst die Zeitverläufe der Steuerparameter von einem vorgeschalteten Steuermodell berechnet und damit die erforderlichen Artikulationsbewegungen generiert. Die kontinuierliche Bewegung des Sprechtraktmodells wird dann auf das eindimensionale Rohrmodell übertragen. Dazu muss in kurzen zeitlichen Abständen die Flächenfunktion des Sprechtraktmodells ermittelt werden. Das Rohrmodell und seine zeitlichen Änderungen werden dann dem akustischen Simulationsverfahren zugeführt, das das Schallfeld im Rohr simuliert. Dabei werden spezielle Modelle benötigt, die die stimmhafte und stimmlose Anregung des Sprechtrakts simulieren. Aus den Schallfeldgrößen an der Mund- und Nasenöffnung des Rohrmodells wird das abgestrahlte Sprachsignal berechnet. Bereits an dieser kurzen Darstellung wird die Komplexität eines artikulatorischen Synthesystems ersichtlich. Für ein funktionierendes System müssen also Modelle für den Sprechtrakt, die aerodynamisch-akustische Simulation, die stimmhafte und stimmlose Anregung und die Steuerung der Modellparameter entwickelt und integriert werden. Viele Forschungsinstitute auf diesem Gebiet konzentrieren sich hauptsächlich auf bestimmte Aspekte der artikulatorischen Synthese, z. B. auf die Weiterentwicklung von Glottismodellen für die stimmhafte Sprechtraktanregung. Es gibt nur sehr wenige artikulatorische Synthesatoren, die die vollständige Prozesskette von der Steuerung der Modellartikulatoren bis hin zur akustischen Synthese des Sprachsignals realisieren. Ein Synthesesystem mit dem Namen ASY („Artikulatory Synthesizer“), dessen Wurzeln bis in die späten 1970er Jahre zurückreichen [97], wird an den Haskins Laboratories in den USA entwickelt. Ein System für die deutsche Sprache wurde am phonetischen Institut der Universität Köln realisiert [66].

Auch am ICP in Grenoble (Frankreich) [6, 76] und am ATR-I in Kyoto (Japan) [26] wird an artikulatorischen Synthetisatoren gearbeitet.

Das Hauptproblem dieser Systeme ist die Erreichung einer hohen Qualität des synthetischen Sprachsignals hinsichtlich der Verständlichkeit und besonders der Natürlichkeit. Deshalb werden sie hauptsächlich als Hilfsmittel für die phonetische Forschung verwendet, bei der es nicht vorrangig auf die Natürlichkeit der synthetischen Sprache ankommt.

1.5 Beitrag und Aufbau der Arbeit

Das Ziel der vorliegenden Arbeit war die Entwicklung eines artikulatorischen Sprachsynthetisators. Angesteuert werden sollte der Synthetisator mit Hilfe artikulatorischer Gesten, die aus der Sicht der Sprachproduktion die grundlegenden Einheiten der motorischen Steuerung darstellen. Das Hauptinteresse bei der Entwicklung des Systems war es, eine möglichst hohe Qualität des synthetischen Sprachsignals zu erzielen. Die Herausforderung dabei war es, geeignete Modelle und Simulationmethoden für die Nachbildung der Spracherzeugung zu entwickeln, zu integrieren und in einem Computerprogramm umzusetzen. Mit „geeigneten“ Modellen sind solche gemeint, die einerseits nicht so komplex sind, dass sie auf aktueller PC-Hardware mehrere Minuten oder Stunden für die Berechnung kurzer Äußerungen benötigen, aber andererseits ausgefeilt genug, um das Potenzial artikulatorischer Sprachsynthese ausschöpfen und demonstrieren zu können. Die integrierten Teilmodelle sind entweder Neuentwicklungen oder Weiterentwicklungen bestehender Ansätze aus der Literatur. Den Hauptbeitrag dieser Arbeit sehe ich in der Entwicklung eines dreidimensionalen Vokaltraktmodells und eines einfachen und effektiven Steuerkonzepts, sowie in der Umsetzung einer vereinheitlichten Methode für die aerodynamisch-akustische Simulation, die weiterentwickelte Modelle für die stimmhafte und stimmlose Anregung integriert. Den Schwerpunkt dieser Arbeit habe ich auf die Entwicklung und Umsetzung dieser Modelle gelegt und nicht auf die Beschaffung artikulatorischer Messdaten oder eine formelle Evaluierung der Synthesergebnisse. Die letzten beiden Punkte sollten in weiterführenden Projekten realisiert werden.

Die vorliegende Arbeit ist wie folgt gegliedert:

- In Kapitel 2 wird die Modellierung des Sprechapparats behandelt. Hier werden die Teilmodelle für den Vokaltrakt, den Nasenraum, die Stimmlippen und das subglottale System vorgestellt und ihre Überführung in das diskrete Rohrmodell erläutert.
- Kapitel 3 beschäftigt sich mit der aerodynamisch-akustischen Simulation. Hier wird ein akustisches Netzwerk entwickelt, das das strömungsakustische Verhalten des Rohrmodells aus Kapitel 2 beschreibt. Auf der Grundlage dieses Netzwerks werden Methoden für die aerodynamisch-akustische Simulation im Zeitbereich und im Frequenzbereich abgeleitet. Hier wird auch ein neues Verfahren für die Simulation der stimmlosen Sprechtraktanregung durch supraglottale Geräuschquellen vorgestellt.
- Kapitel 4 behandelt ein neues Steuermodell für die Generierung artikulatorischer Bewegungsabläufe auf der Grundlage von Gesten. Dieses Modell integriert klassische gestische Steuermodelle mit dominanzbasierten Mechanismen für die realitätsnahe Simulation von Koartikulation.

- Kapitel 5 geht auf die Synthese zusammenhängender Äußerungen auf der Grundlage des Steuermodells ein. Hier wird die Einstellung der Vokaltraktparameter für die Laute der deutschen Sprache behandelt und es werden Regeln für die Erstellung gesturaler Ablaufpläne formuliert. Diese Regeln werden anhand zahlreicher Synthesebeispiele demonstriert, die als Audiodateien auf der beiliegenden CD zu finden sind.
- Kapitel 6 fasst die Beiträge und Ergebnisse der Arbeit zusammen und gibt einen Ausblick auf mögliche weiterführende Forschungsarbeiten.
- Im Anhang befindet sich ein weiteres Kapitel, das einen Überblick über die programmtechnische Realisierung des Synthetisators und die grafische Benutzerschnittstelle gibt.

Die Kapitel 2 bis 4 enthalten jeweils einen Literaturüberblick über die dort behandelten Themen, um die in dieser Arbeit entwickelten Modelle und Methoden einordnen zu können.

Kapitel 2

Modellierung des Sprechapparats

2.1 Einleitung

Dieses Kapitel behandelt die Modellierung der einzelnen Teile des Sprechapparats und ihre einheitliche Darstellung in Form eines verzweigten diskreten Rohrmodells. Das Rohrmodell ist aus vielen kurzen Rohrabschnitten zusammengesetzt, von denen jeder durch seine Länge, seinen Umfang, seine Querschnittsfläche und seine Position auf der Rohrachse gekennzeichnet ist. In Abb. 2.1 ist das Rohrmodell am Beispiel des Vokals [o:] schematisch dargestellt. Hierbei wurden die Querschnittsflächen der Rohrabschnitte über der Position auf der Rohrachse abgetragen.

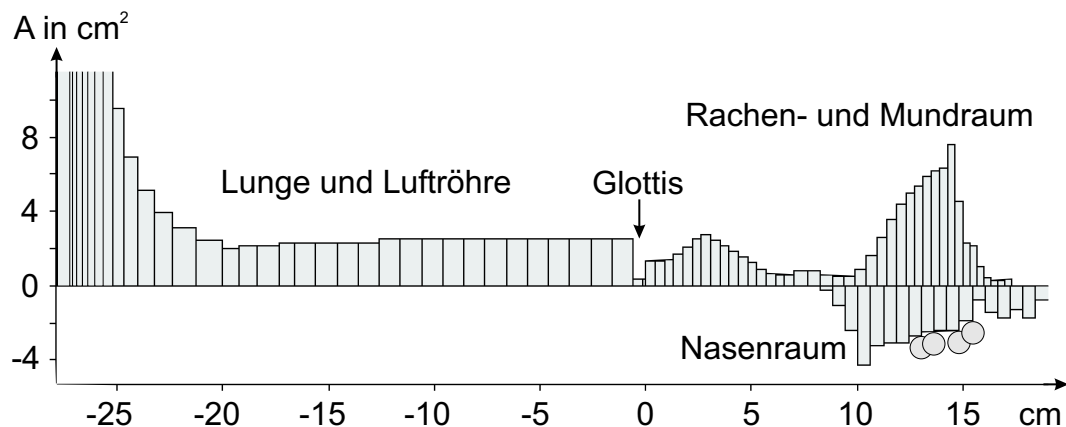


Abbildung 2.1: Flächenfunktion des Sprechapparats für den Vokal [o:].

Die Rohrabschnitte sind verschiedenen Teilen des Sprechapparats zugeordnet. Der subglottale Teil aus der Lunge und der Luftröhre wird durch 34 Rohrabschnitte repräsentiert, der Rachen- und Mundraum durch 40 Abschnitte und der Nasenraum durch 19 Abschnitte (Querschnittsflächen sind „nach unten geklappt“ dargestellt). An die Rohrabschnitte des Nasenraums sind vier als Kreise dargestellte Nebenhöhlen angekoppelt. Die Glottis besteht aus zwei Rohrabschnitten, von denen einer den unteren und einer den oberen Teil darstellt. Zwei weitere Rohrabschnitte sind parallel zu den glottalen Rohrabschnitten an-

geordnet und repräsentieren einen parallelen Spalt zwischen den Stellknorpeln, der nicht mit der eigentlichen Stimmritze verbunden ist (siehe Abschnitt 2.5). In Abb. 2.1 sind diese beiden Abschnitte nicht sichtbar, da ihr Flächeninhalt für den gezeigten Laut auf Null gesetzt wurde. Am Eingang und am Ausgang der Glottis, an der velopharyngealen Pforte und an den Eingängen zu den Nasennebenhöhlen besitzt das Rohrmodell jeweils eine Verzweigungsstelle.

Das hier verwendete Rohrmodell zeichnet sich gegenüber den Rohrmodellen in den meisten anderen Arbeiten (z. B. [66, 78, 97, 99]) dadurch aus, dass für die Rohrabschnitte kein kreisrunder, sondern eine elliptischer Querschnitt angenommen wurde, und dass die Länge der Rohrabschnitte variabel ist. Aufgrund der ersten Eigenschaft ist der Umfang eines Rohrabschnitts, der sich bei der akustischen Simulation im Reibungswiderstand bemerkbar macht, vom Flächeninhalt des Rohrquerschnitts unabhängig. Die zweite Eigenschaft ermöglicht die einheitliche aerodynamisch-akustische Behandlung des gesamten Sprechapparats (incl. der Glottis mit ihren sehr kurzen Rohrabschnitten) und eine stufenlose Änderung der Länge des Ansatzrohres bei der Sprachsynthese.

Der nächste Abschnitt dieses Kapitels gibt einen Überblick über die Modellierung der einzelnen Komponenten des Sprechapparats in anderen Studien. In den darauf folgenden Abschnitten werden die Modellierung des Vokaltrakts, der Stimmlippen, des Nasenraums und des subglottalen Systems in der vorliegenden Arbeit vorgestellt. Der Vokaltrakt und die Stimmlippen werden dabei zunächst als dreidimensionale Objekte modelliert, die anschließend in die Segmente des Rohrmodells überführt werden, während der Nasenraum und das subglottale System direkt durch die Parameter ihrer Rohrabschnitte definiert werden.

2.2 Literaturüberblick

Dieser Abschnitt gibt einen Überblick über bisherige Studien zur Modellierung des Sprechapparats. Analog zum Aufbau dieses Kapitels werden die Modelle für den Vokaltrakt (Rachen- und Mundraum), den Nasenraum, die Stimmlippen und das subglottale System getrennt behandelt. Der Überblick ist aufgrund der Breite dieses Forschungsgebiets darauf ausgerichtet, verschiedene *methodische Ansätze* für die Modellierung aufzuzeigen, ohne zu detailliert auf einzelne Realisierungen der Ansätze einzugehen.

Vokaltrakt. Die meisten der bisher entwickelten Vokaltraktmodelle können den folgenden Kategorien zugeordnet werden: den *geometrischen*, den *statistischen* oder den *biomechanischen* Modellen. Weiterhin lassen sie sich in eindimensionale, zweidimensionale und dreidimensionale Modelle unterteilen. Die 3D-Modelle ermöglichen die genaueste Nachbildung des Vokaltrakts, während 2D-Modelle die Artikulatoren und Sprechtraktwände nur durch ihre Konturlinien in der mediosagittalen Ebene repräsentieren. Bei den 1D-Modellen wird der Vokaltrakt von vornherein gleich durch die Flächenfunktion dargestellt.

Bei den geometrischen und statistischen Artikulatormodellen definieren die Vokaltraktparameter unmittelbar die Position und Form einzelner Artikulatoren (2D- und 3D-Modelle) bzw. die Flächenfunktion des Ansatzrohres (1D-Modelle). Die beiden Kategorien unterscheiden sich dadurch, dass die Freiheitsgrade der geometrischen Modelle *a priori* und daher willkürlich festgelegt werden, während sie für die statistischen Modelle aus einer Analyse

artikulatorischer Messdaten resultieren. Die Analyse erfolgt mit statistischen Methoden, die Korrelationen in den Messdaten auswerten und die tatsächliche Anzahl *statistisch unabhängiger* Freiheitsgrade aufdecken, die den Daten zugrunde liegen. Die Anzahl dieser unabhängigen Variablen ist in der Regel gering, was zu einer kompakten Repräsentation der Messdaten führt. Sehr häufig wird als statistische Methode die Hauptkomponentenanalyse (*principal component analysis* bzw. PCA) verwendet. Die Grundidee der PCA besteht darin, eine Vielzahl gemessener Funktionen bzw. Datenvektoren als Linearkombination *weniger* orthogonaler Basisfunktionen bzw. Basisvektoren (die Hauptkomponenten) auszudrücken, die mit Hilfe einer Eigenwertanalyse ermittelt werden.

Es existieren ein-, zwei- und dreidimensionale statistische Artikulatoremodelle. MEYER *et al.* [78] beschreiben ein eindimensionales Modell, mit dem die Flächenfunktion durch die Summe aus einer „neutralen“ Flächenfunktion und zwei gewichteten Hauptkomponenten gebildet wird. Die zwei Hauptkomponenten decken bereits 93% der Varianz in den Ausgangsdaten ab und genügen für die Nachbildung der meisten deutschen Vokale. Zweidimensionale statistische Modelle, die auf der Hauptkomponentenanalyse von Zungenkonturen beruhen, die wiederum aus Röntgenaufnahmen gewonnen wurden, haben u. a. MAEDA [74] und BEAUTEMPS *et al.* [9] vorgestellt. Der Ansatz von BEAUTEMPS *et al.* wurde später für die Konstruktion eines 3D-Artikulatoremodells genutzt (siehe BADIN *et al.* [6]), das auf der statistischen Analyse von 3D-MRT-Aufnahmen basiert. Etwa im gleichen Zeitraum wurde auch von ENGWALL [30] ein statistisches 3D-Modell des Sprechtrakts mit Hilfe von MRT-Daten entwickelt. Allen statistischen Modellen ist gemein, dass sie auf große Mengen von Messdaten angewiesen sind, die aufwendig erhoben und ausgewertet werden müssen. Dafür weisen diese Modelle in der Regel nur eine geringe Anzahl statistisch unabhängiger Freiheitsgrade auf.

Bei der Entwicklung eines geometrischen Modells ist man dagegen selbst für die Definition der Freiheitsgrade zuständig und benötigt weitaus weniger Messdaten. Jedoch lässt sich nicht von vornherein ausschließen, dass einige der Parameter miteinander korreliert sind und das Modell somit „zu viele“ Freiheitsgrade besitzt. Wenn korrelierte Freiheitsgrade unabhängig voneinander eingestellt werden können, so erhöht sich unter Umständen der Aufwand bei der Ansteuerung des Modells, und es können Vokaltraktformen oder Artikulatorbewegungen entstehen, die in der Realität nicht vorkommen. Das passiert bei den statistischen Modellen nicht und muss bei den geometrischen Modellen durch geeignete Maßnahmen verhindert werden.

Auch bei den geometrischen Modellen existieren ein-, zwei- und dreidimensionale Ansätze. Eindimensionale Modelle, die mit Hilfe weniger Variablen die Flächenfunktion des Vokaltrakts beschreiben, wurden z. B. von FLANAGAN *et al.* [32] und FANT *et al.* [34] vorgestellt. Beide Modelle basieren auf der Feststellung, dass sich die Flächenfunktionen von Vokalen hauptsächlich durch den Grad und die Position einer charakteristischen Engstelle unterscheiden, die mit der Zunge und der Rachenwand bzw. dem Gaumen gebildet wird [129, 109]. Daher besitzen beide Modelle jeweils einen Parameter für den Ort und einen für den Querschnitt dieser Konstriktion. Abhängig vom Modell gibt es weitere Parameter, die z. B. die Gesamtlänge des Vokaltrakts oder den Querschnitt im Bereich der Lippen definieren. Zweidimensionale geometrische Modelle wurden u. a. von MERMELSTEIN [77], COKER [19], KRÖGER [66] und LINDBLOM *et al.* [69] vorgeschlagen. In den ersten beiden Modellen wird die Kontur der Zunge explizit aus einfachen geometrischen

Formen zusammengesetzt (Linienabschnitte, Kreis- und Parabelbögen), während sie in den Modellen von KRÖGER und LINDBLOM *et al.*, wenn auch auf unterschiedliche Weise, aus der Interpolation zwischen vordefinierten Extremalkonturen der Eckvokale [a:], [i:] und [u:] hervorgehen. Das Modell von KRÖGER wurde später auf der Basis von MRT-Daten zu einem visuellen Hilfsmittel in der Therapie von Sprechstörungen weiterentwickelt [67]. Das Modell von LINDBLOM *et al.*, das ursprünglich nur für die Artikulation von Vokalen entwickelt wurde, haben STARK *et al.* [107] um Parameter für konsonantische Konstruktionen erweitert und zu einem Werkzeug für die Untersuchung artikulatorisch-akustischer Beziehungen weiterentwickelt. Das Modell von MERMELSTEIN wird z. B. in den Arbeiten [14] und [97] für die artikulatorische Sprachsynthese eingesetzt. Dreidimensionale geometrische Modelle wurde bisher nur von BIRKHOLZ *et al.* [13] und ENGWALL [29] vorgestellt. Beide Ansätze basieren auf dem 2D-Modell von MERMELSTEIN [77] und wurde um Parameter zur Berücksichtigung der dritten Dimension erweitert.

Biomechanische Modelle des Sprechtrakts basieren auf der Simulation des Zusammenspiels von Muskeln und artikulatorischen Strukturen mit Hilfe von Methoden der mechanischen Analyse (Finite-Elemente Methoden, FEM) und der numerischen Simulation. Diese Modelle besitzen eine intrinsische Dynamik und Struktur und werden mit Hilfe von Muskelkommandos angesteuert. Aufgrund der Vielzahl der Muskeln, die für die Artikulation eine Rolle spielen, besitzen diese Modelle sehr viele Freiheitsgrade und sind entsprechend aufwendig anzusteuern. Verglichen mit den geometrischen oder statistischen Modellen sind biomechanische Modelle in der Simulationen sehr rechenaufwändig.

Die meisten Arbeiten zur biomechanischen Modellierung konzentrieren sich auf die Nachbildung der Zunge. Das erste biomechanische 2D-Zungenmodell wurde 1969 von PERKELL [95] entwickelt und neuere 3D-Zungenmodelle u. a. von WILHELMS-TRICARICO [128] und GERARD *et al.* [46]. Ein vollständiges physiologisches Artikulationsmodell, das sowohl die Zunge als auch die Lippen durch finite Elemente repräsentiert und für eine schnelle Berechnung optimiert wurde, haben 2004 DANG und HONDA [26] vorgestellt.

Nasenraum. Der Nasenraum hat anatomisch gesehen eine sehr komplexe Gestalt. Aus akustischer Sicht kann er jedoch analog zum Rachen- und Mundraum näherungsweise wie ein Rohr mit variablem Querschnitt entlang der Rohrachse behandelt werden. Da sich die Form des Nasenraums beim Sprechen außerdem nur geringfügig ändert, wird er in den meisten Studien zur artikulatorischen Sprachsynthese deshalb auch direkt als Rohr mit einer vorgegebenen Flächenfunktion repräsentiert [14, 78]. In verschiedenen Untersuchungen wurde erkannt, dass sich die komplexen akustischen Eigenschaften nasaler Laute nur dann realistisch simulieren lassen, wenn zusätzlich die Nasennebenhöhlen berücksichtigt werden, die an mehreren Stellen vom Nasenraum abzweigen [70, 73]. In der Arbeit von DANG *et al.* [24] wurde erstmals die Magnetresonanztomographie (MRT) eingesetzt, um eine genaue Flächenfunktion des Nasenraums und die Abmessungen der Nasennebenhöhlen an *lebenden* Personen zu bestimmen. Aus den MRT-Daten haben DANG *et al.* in einer späteren Arbeit [25] ein Modell des Nasenraums mit vier angekoppelten Nebenhöhlen entwickelt, welches die komplexen akustischen Eigenschaften nasaler Übertragungsfunktionen nachbilden kann.

Stimmlippen. Ein Modell der Stimmlippen dient der Simulation der zeitvariablen

Form der glottalen Öffnung, die direkt mit der akustischen Anregung des Vokaltrakts in Beziehung steht. Die akustische Anregung entscheidet zum großen Teil über die Natürlichkeit der künstlichen Stimme und damit über die Qualität der gesamten Sprachsynthese. Eine große Klasse von Stimmlippenmodellen bilden die *selbstschwingenden Modelle*. Bei ihnen wird die Form der Stimmlippen im ungestörten Zustand (keine Schwingung) modelliert und zusätzlich werden ihre physikalischen Eigenschaften wie Dämpfung, Trägheit und Elastizität definiert. Die Schwingung der Stimmlippen wird bei diesen Modellen – so wie beim realen Vorbild – durch die zeitvariablen Strömungs- und Druckverhältnisse in der Glottis induziert, d. h. durch die Übertragung von Strömungsenergie auf die Bewegung der Stimmlippen. Daher besitzen diese Modelle theoretisch das größte Potenzial für die Erzeugung natürlicher Stimmen. Andererseits können sie sehr komplex und rechenintensiv werden. Außerdem sind genaue Kenntnisse über die physikalischen Parameter der Gewebeschichten (Muskelgewebe, Stimmbänder, Schleimhaut) erforderlich, aus denen die Stimmlippen aufgebaut sind (siehe [120] für den Aufbau der Stimmlippen).

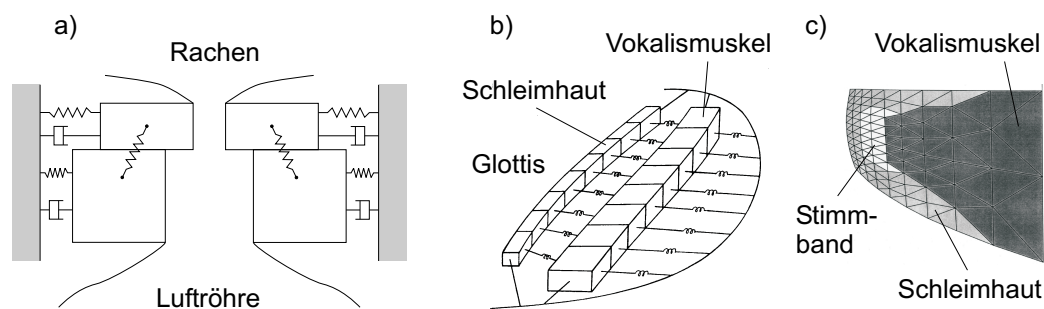


Abbildung 2.2: Modellierung der Stimmlippen mit unterschiedlicher Komplexität. (a) Zwei-Massen-Modell von ISHIZAKA *et al.* [51] mit einem unteren und oberen Oszillator pro Stimmlippe, (b) 16-Massen-Modell von TITZE [113, 114] mit jeweils acht gekoppelten Massen für den Vokalmuskel und die Schleimhaut pro Stimmlippe, (c) Finite-Elemente-Modell von ALIPOUR *et al.* [2] mit 1500 Knoten und drei Schichten mit unterschiedlichen physikalischen Eigenschaften pro Stimmlippe.

Für die artikulatorische Sprachsynthese versucht man selbstschwingende Stimmlippenmodelle zu entwickeln, die einerseits einfach aufgebaut und schnell zu berechnen sind, andererseits aber in ihrer Dynamik die wesentlichen Eigenschaften realer Stimmlippen aufweisen. Dieser Kompromiss ist ISHIZAKA und FLANAGAN [51] 1972 mit dem klassischen Zwei-Massen-Modell gelungen, welches in Abb. 2.2 (a) dargestellt ist und eine weite Verbreitung gefunden hat. In diesem Modell wird jede Stimmlippe durch zwei übereinanderliegende mechanische Oszillatoren repräsentiert. Jeder Oszillator besitzt eine bestimmte Masse, Steifigkeit und Dämpfung, und kann sich in lateraler Richtung bewegen. Der obere und untere Oszillator jeder Stimmlippe sind über eine Feder miteinander gekoppelt. Mit diesem Modell war es erstmals möglich, die Phasenverschiebung zwischen dem oberen und unteren Rand der Stimmlippen zu simulieren, die sowohl aus akustischer Sicht als auch für die effiziente Energieübertragung von der Strömung auf das Gewebe eine wichtige Rolle spielt [66, S. 25]. Das Zwei-Massen-Modell wurde in einer Vielzahl von Studien modifiziert oder weiterent-

wickelt. KOIZUMI *et al.* [63] haben allein drei modifizierte Varianten des Modells vorgeschlagen und sie hinsichtlich der Natürlichkeit ihrer Stimme untersucht. PELORSON *et al.* [93] haben für eine genauere aerodynamische Simulation die quaderförmigen Oszillatoren des Modells durch Zylinder ersetzt. DE VRIES *et al.* [125, 124] haben auf der Basis einer Finiten-Elemente-Simulation der Stimmlippen realistischere Werte für die physikalischen Parameter des Zwei-Massen-Modells abgeleitet. KRÖGER [66] und CRANEN *et al.* [21] haben das Modell um einen glottalen Bypass zwischen den Stellknorpeln erweitert, um den unvollständigen Verschluss der Stimmritze während der eigentlichen Verschlussphase der Schwingungszyklen zu simulieren, der in mehreren Studien festgestellt wurde [100, 22]. Ein Modell, in dem jede Stimmlippe durch nur einen einzigen Oszillator repräsentiert wird, der dafür aber einen zusätzlichen Rotationsfreiheitsgrad besitzt, wurde von LILJENCRAANTS [68] entwickelt. Von STORY *et al.* [112] wird ein Modell aus *drei* gekoppelten Oszillatoren vorgeschlagen, um die unterschiedlichen physikalischen Eigenschaften des Vokalismuskels (1 Oszillator) und der Schleimhaut (2 Oszillatoren) zu berücksichtigen. Ein Modell aus 16 gekoppelten Oszillatoren pro Stimmlippe wurde von TITZE [113, 114] vorgestellt und ist in Abb. 2.2 (b) dargestellt. Dabei wurden der Vokalismuskels und die Schleimhaut in Längsrichtung jeweils durch 8 gekoppelte Oszillatoren approximiert. Eine noch genauere Simulation der Stimmlippenbewegung ermöglichen kontinuumsmechanische Modelle wie der Ansatz von ALIPOUR *et al.* [2] in Abb. 2.2 (c) oder die Modelle von ROSA *et al.* [91] und DE VRIES *et al.* [125]. Diese Modelle sind aus finiten Elementen aufgebaut und erfordern daher eine sehr hohe Berechnungszeit für die Simulation.

Eine andere Möglichkeit, als selbstschwingende Modelle der Stimmlippen für die Sprachsynthese zu verwenden, besteht darin, die Öffnungsfläche der Glottis unabhängig von den konkreten Druck- und Strömungswerten in der Glottis direkt vorzugeben. Dadurch besitzt man eine unmittelbare Kontrolle über die Eigenschaften der Stimmlippenschwingung, z. B. die F_0 und die Schwingungsamplitude. Solche Modelle wurden u. a. von TITZE [115, 117] und CHILDERS *et al.* [18] vorgestellt. Die genannten Modelle wurden nicht nur entwickelt, um die zeitvariable glottale Öffnung zu simulieren, sondern auch die Kontaktfläche zwischen den beiden Stimmlippen. Die Simulation der Kontaktfläche ist für die Elektroglossographie von Bedeutung (eine nichtinvasive Methode zur abstrakten Visualisierung der Stimmlippenbewegung). Von CRANEN *et al.* [23] wurde das Modell von TITZE [115] um einen Bypass zwischen den Stellknorpeln erweitert. Die Autoren haben die akustischen Auswirkungen untersucht, die sich ergeben, wenn man den Bypass entweder als einen *parallelen* Spalt zur eigentlichen Glottis behandelt oder aber als ein mit der Glottis verbundenes Leck. Der Nachteil dieser „parametrischen“ Stimmlippenmodelle ist, dass die Rückwirkung der Aerodynamik auf das Schwingungsverhalten nicht modelliert wird und daher mit Hilfe zusätzlicher Regeln in die Simulation eingehen muss.

Subglottale Luftwege. Wie der Vokaltrakt selbst spielen auch die subglottalen Luftwege (Lufttröhre und Lunge) eine wichtige Rolle bei der Sprachproduktion. Einerseits ist die Lunge die Druck- bzw. Energiequelle für die Spracherzeugung, und andererseits besitzt das subglottale System spezielle akustische Eigenschaften, die sich in Abhängigkeit von der glottalen Öffnungsfläche unterschiedlich stark im Sprachschall manifestieren. Bei einer relativ weiten Stimmritze, wie bei der behauchten Phonation oder während der Erzeugung von Frikativen, ist der Vokaltrakt akustisch stark mit dem subglottalen

System gekoppelt. Die Kopplung führt sowohl zu einer stärkeren Dämpfung der Vokaltraktformanten [110] als auch zum Erscheinen zusätzlicher subglottaler Formanten im Sprachschall [35]. Bei kleineren Öffnungsflächen der Glottis, wie bei der modalen Phonation, spielen die akustischen Eigenschaften des subglottalen Systems dagegen eine untergeordnete Rolle. Deshalb wird in vielen Arbeiten zur artikulatorischen Synthese auf eine Modellierung der subglottalen Luftwege verzichtet und statt dessen eine Quelle zur Erzeugung des subglottalen Drucks direkt unter der Stimmritze angeordnet [106, 72, 40]. Dennoch kann man eine höhere Natürlichkeit synthetischer Sprache erwarten, wenn die subglottalen Luftwege in die Simulation einbezogen werden.

Eine Möglichkeit der Modellierung besteht darin, das subglottale System, genauso wie den Vokaltrakt, durch ein Rohr mit variablem Querschnitt zu repräsentieren. Da der Bronchialbaum ein stark verzweigtes Rohrsystem ist, welches nur mit hohem Rechenaufwand akustisch zu simulieren wäre, modelliert man das System in der Regel mit einer *einzi-gen* Flächenfunktion, in der die Querschnittsflächen an verschiedenen Abständen von der Glottis aufsummiert sind [14, 35, 53]. ISHIZAKA *et al.* [53] haben das subglottale System z. B. nach den anatomischen Daten von WEIBEL [127] modelliert und die Materialeigenschaften (Masse, Steifigkeit, Dämpfung) des Lungengewebes so angepasst, dass ihr Modell die gleichen Resonanzfrequenzen aufweist, die am Menschen gemessen wurden. Eine einfachere Möglichkeit besteht darin, das subglottale System direkt durch eine Kaskade von RLC-Parallelschwingkreisen zu realisieren, wobei jeder Schwingkreis einen Formanten erzeugt [4, 7]. Darüber hinaus gibt es außerdem Modelle, die nicht nur die aerodynamisch-akustische Seite des subglottalen Systems nachbilden, sondern auch die physiologisch-dynamische Seite, d. h. das Zusammenspiel zwischen der Atemmuskulatur und dem Lungengewebe. Eine Zusammenfassung hierzu findet man in KRÖGER [66].

2.3 Geometrische Modellierung des Vokaltrakts

Ich habe mich in dieser Arbeit für die Entwicklung eines *dreidimensionalen geometrischen* Vokaltraktmodells entschieden. Ein geometrisches Modell kommt einerseits mit relativ wenigen artikulatorischen Messdaten aus, und andererseits ist man durch die freie Wahl der Modellparameter sehr flexibel. Ein statistisches Modell ist dagegen in aller Regel aus den konkreten artikulatorischen Daten eines Sprechers abgeleitet. Dadurch besitzt es automatisch eine eingeschränkte Menge an Freiheitsgraden, was der Steuerung des Modells zugute kommt, aber die Flexibilität einschränkt. Bestimmte Artikulationen (z. B. Laute einer Fremdsprache), die nicht von vornherein als Messdaten von dem Sprecher aufgenommen wurden, können daher mit hoher Wahrscheinlichkeit auch nicht exakt vom Vokaltraktmodell generiert werden. Bei einem geometrischen Vokaltraktmodell hängen die Beschränkungen in Hinsicht auf die Artikulationsvielfalt nur von der Wahl der Parameter ab. Ein biomechanisches Vokaltraktmodell kam aufgrund der Zielstellung dieser Arbeit nicht in Frage, da es auf heutiger PC-Hardware zu rechenaufwendig für die Sprachsynthese wäre und die Ansteuerung beim gegenwärtigen Stand der Forschung noch problematisch ist.

Für ein *dreidimensionales* Modell habe ich mich entschieden, weil damit im Gegensatz zu den 1D- und 2D-Modellen, sowohl im Hinblick auf die Akustik als auch auf die visuelle Darstellung, die exakteste Repräsentation der Sprechtraktform möglich ist. Ein Nachteil der 2D-Modelle ist z. B., dass die Details der Artikulation in lateraler Richtung nicht erfasst

werden. Für die meisten Phoneme kann zwar aus der mediosagittalen Darstellung sehr gut auf die Vokaltraktform geschlossen werden, aber die freien Passagen an den abgesenkten Zungenrändern beim [l] müssen bei diesen Modellen z. B. durch eine veränderte Artikulation kompensiert werden. 3D-Modelle sind auch für weiterführende Forschungsarbeiten im Bereich der Sprachakustik interessant. Sie können z. B. für die Untersuchung dreidimensionaler Simulationsmethoden der (Strömungs-)Akustik verwendet werden, welche im Umfeld der Spracherzeugung erst seit kurzem erforscht werden.

2.3.1 Grundkonzepte

Das größte Problem bei der Entwicklung eines geometrischen Artikulatormodells besteht darin, geeignete Parameter auszuwählen und deren Einfluss auf die Vokaltraktform zu definieren. Die Entscheidung darüber sollte in Abhängigkeit von den konkreten Anforderungen getroffen werden, die man an das Modell stellt. Im Idealfall sollte das Modell alle Sprechtraktformen erzeugen können, zu denen auch ein Mensch in der Lage ist (um grundsätzlich die Laute aller Sprachen dieser Welt erzeugen zu können), jedoch keine anatomisch unmöglichen Formen. Der erste Punkt ist eher ein Argument für viele Parameter und der zweite Punkt für wenige Parameter bzw. eine realistische Beschränkung ihrer Wertebereiche. Daher ist ein Kompromiss gefragt, der sowohl einer hohen Flexibilität als auch einer einfachen Handhabung des Vokaltraktmodells gerecht wird.

Die Grundidee bei der Entwicklung des Modells bestand darin, dass jeder Vokaltraktparameter nur lokal einen bestimmten Bereich des Modells (und damit indirekt der Flächenfunktion) beeinflusst. Die Koordination der Parameter zur Erzeugung von Synergieeffekten (z. B. die gegenseitige Beeinflussung der Unterkiefer- und Zungenposition) sollte durch eine übergeordnete Kontrollschicht realisiert werden und nicht durch die komplexe Beeinflussung der Vokaltraktform an einer bestimmten Stelle durch mehrere Parameter. Dies erfolgt auf der Ebene der Steuerung (siehe Kapitel 4). Ich habe bei der Entwicklung nicht versucht, die Anzahl der Parameter von vornherein auf ein Minimum zu reduzieren, sondern habe im Zweifelsfall lieber einen Parameter (Freiheitsgrad) mehr als zu wenig zugelassen. Wenn nachträglich Korrelationen zwischen den Parametern festgestellt werden, so können die redundanten Parameter im gewählten Ansatz problemlos automatisch berechnet werden.

Weiterhin habe ich besonderen Wert darauf gelegt, dass die Geometrie im vorderen Bereich des Vokaltrakts möglichst gut wiedergegeben wird, da dieser Bereich eine wichtige Rolle bei der Erzeugung der Frikative spielt. Das erfordert insbesondere eine genaue Modellierung der Zähne und Lippen. Die Synthese guter Frikative hängt empfindlich von der konkreten Form der Flächenfunktion in diesem Bereich ab (siehe Abschnitt 3.8.2).

Die Zunge sollte durch eine geeignete Parametrisierung so flexibel sein, dass sie neben der Bewegung in der sagittalen Ebene auch das Heben und Senken der Zungenseiten erlaubt. Die Anhebung der Zungenseiten sollte entlang der Zunge von der Wurzel bis zur Spitze variabel sein, um die dreidimensionalen Zungenformen unterschiedlicher Laute möglichst genau nachbilden zu können (siehe z. B. STONE *et al.* [111] für dreidimensionale Zungenformen).

Neben den Anforderungen im Hinblick auf die akustische Simulation sollte das Modell auch so gestaltet werden, dass es eine ansprechende visuelle Darstellung erlaubt.

2.3.2 Geometrie und Parameter des Vokaltrakts

Das Vokaltraktmodell ist aus mehreren Gitternetzen aufgebaut, die die Oberflächen der Artikulatoren und Sprechtraktwände im dreidimensionalen Raum repräsentieren. Jedes der Gitter ist durch eine zweidimensionale Matrix von Punkten $P_{i,j}$ definiert, wobei $i = 0 \dots M-1$ und $j = 0 \dots N-1$. Die Anzahl $M \times N$ der Punkte hängt vom betrachteten Gitter ab. Zwischen jeweils vier benachbarten Gitterpunkten $P_{i,j}$, $P_{i+1,j}$, $P_{i,j+1}$ und $P_{i+1,j+1}$ ist die Oberfläche durch die zwei Dreiecke mit den Eckpunkten $(P_{i,j}, P_{i+1,j+1}, P_{i+1,j})$ und $(P_{i,j}, P_{i+1,j+1}, P_{i,j+1})$ definiert.

Das Modell ist aus den folgenden Gittern zusammengesetzt: Eines der Gitter bildet die Kehlkopf- und Rachenrückseite sowie den weichen und harten Gaumen und wird als *obere Hülle* bezeichnet. Ein anderes Gitter bildet die Kehlkopf- und Rachenvorderseite sowie den Unterkiefer und wird als *untere Hülle* bezeichnet. Die oberen Zähne, die unteren Zähne, die Oberlippe, die Unterlippe und die Zunge werden jeweils durch ein eigenes Gitter repräsentiert.

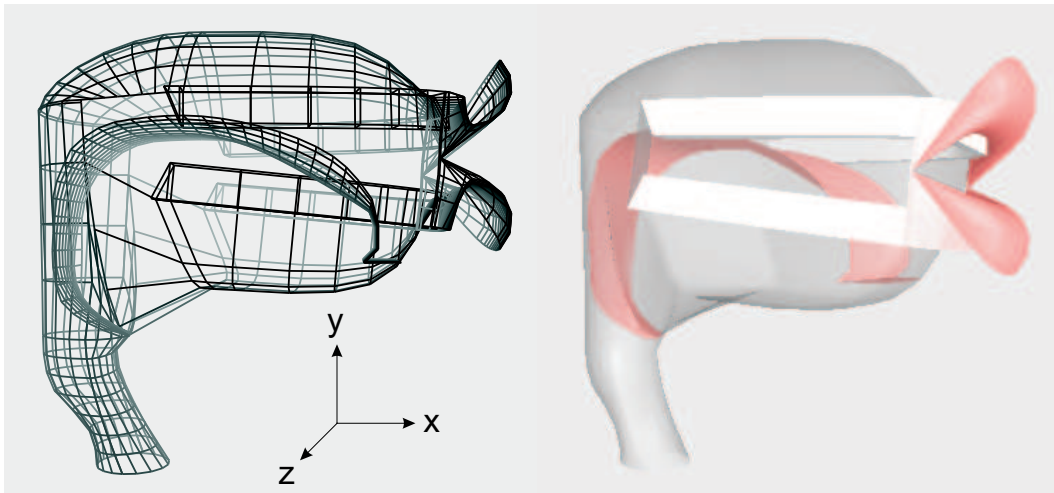


Abbildung 2.3: Das dreidimensionale Vokaltraktmodell in Drahtgitter- und Flächendarstellung.

Abbildung 2.3 zeigt das Modell für den Vokal [a:] einmal als Drahtgittermodell und einmal mit ausgefüllten Flächen. Die obere und untere Hülle sind in der rechten Abbildung durchscheinend dargestellt, so dass man die Zunge erkennen kann. Im linken Bild sind zusätzlich die Achsen des kartesischen Koordinatensystems eingezeichnet, in dem die Gitter definiert sind. Die xy -Ebene entspricht somit der sagittalen Ebene.

Um die folgende Beschreibung der Gitterformen zu erleichtern, wird die stückweise lineare Verbindung zwischen den Punkten $P_{i,0} \dots P_{i,N-1}$ eines Gitternetzes als *Rippe* bezeichnet. Im Vokaltraktmodell liegen alle Gitterpunkte einer Rippe räumlich jeweils *in einer Ebene*. In Abb. 2.3 sind die Rippen der beiden Hüllen und der Zunge die Gitterlinien, die von der linken zur rechten Seite des Modells verlaufen, also in lateraler Richtung. Sie sind jeweils senkrecht zur xy -Ebene ausgerichtet. Die Linien, die vom Kehlkopf zu den Schneidezähnen bzw. von der Zungenwurzel zur Zungenspitze verlaufen, verbinden jeweils korrespondieren-

de Gitterpunkte auf benachbarten Rippen.

Um den Rechenaufwand bei der Simulation des Modells zu reduzieren, habe ich von der oberen und unteren Hülle sowie von den Zähnen und den Lippen jeweils nur die linke Hälfte tatsächlich modelliert. Die rechte Hälfte entsteht durch Spiegelung an der mediosagittalen Ebene. Die Zunge wurde dagegen vollständig modelliert, um auch asymmetrische Zungenformen erzeugen zu können. Die Größe der einzelnen Gitter ist in Tab. 2.1 zusammengestellt. Außer für die Zunge gilt die angegebene Anzahl an Rippenpunkten jeweils nur für die linke Hälfte des Modells.

Gitter	Anzahl Rippen	Anzahl Rippenpunkte
Obere Hülle	28	6
Untere Hülle	20	5
Obere und untere Zähne (jeweils)	13	5
Ober- und Unterlippe (jeweils)	13	9
Zunge	37	11

Tabelle 2.1: Größe der Gitter des Modells. Außer für die Zunge bezieht sich die Anzahl der Rippenpunkte nur auf die linke Hälfte des Modells.

Für die Definition der Gitterformen ist es zweckmäßig, die obere und untere Hülle in separate *Teilgitter* zu zerlegen. Die Teilgitter der oberen Hülle sind: die Kehlkopfrückseite, die Rachenrückseite sowie der weiche und der harte Gaumen. Die Teilgitter der unteren Hülle sind: die Kehlkopfvorderseite, die Rachenvorderseite und der Unterkiefer. Einen Überblick über alle Teilgitter des Modells gibt Abb. 2.4.

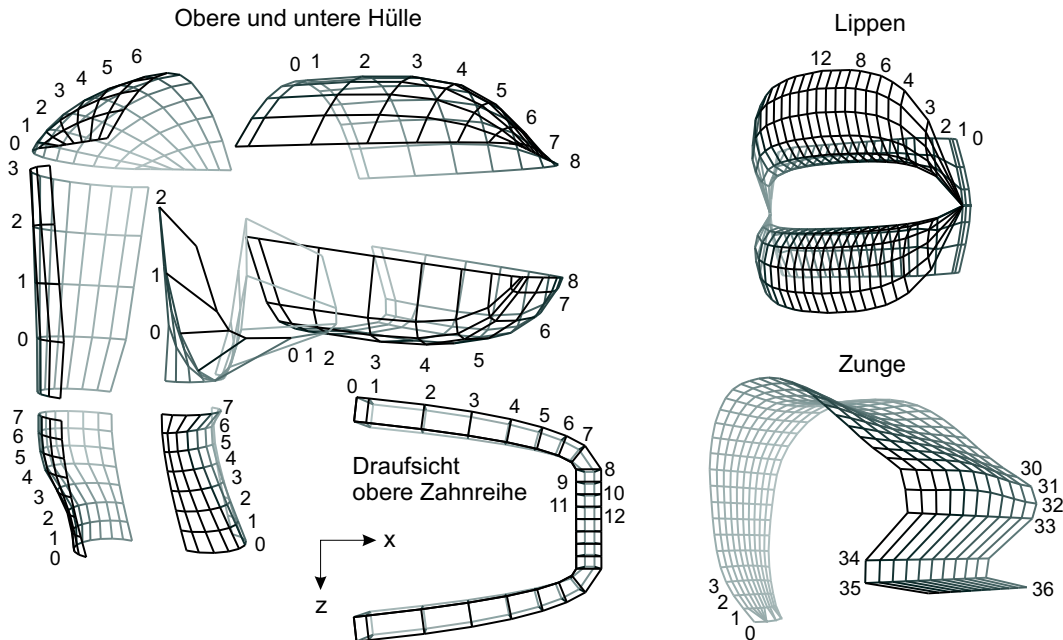


Abbildung 2.4: Zerlegung des Vokaltraktmodells in seine Teilgitter.

Die räumliche Orientierung der dargestellten Gitter in Abb. 2.4 wurde so gewählt, dass man ihre Struktur gut erkennen kann. Die Zahlen an den Gittern dienen der Nummerierung ihrer Rippen. Daran erkennt man, dass sich die 28 Rippen der oberen Hülle in 8 Rippen für die Kehlkopfrückseite, 4 Rippen für die Rachenrückseite, 7 Rippen für den weichen Gaumen und 9 Rippen für den harten Gaumen aufteilen. Die 20 Rippen der unteren Hülle teilen sich auf in 8, 3 und 9 Rippen für die Kehlkopfvorderseite, die Rachenvorderseite und den Unterkiefer. Von den Teilgittern der oberen und unteren Hülle sind die erste Rippe der Rachenrückseite, die letzte Rippe des weichen Gaumens und die erste und letzte Rippe der Rachenvorderseite nicht nummeriert. Das liegt daran, dass diese Rippen im Modell genau den ersten bzw. letzten Rippen der Nachbargitter entsprechen und dadurch in der Zeichnung redundant sind. Von den Zähnen ist nur die obere Zahnreihe dargestellt, da die oberen und unteren Zähne identisch aufgebaut sind und sich im fertigen Modell nur durch ihre Position und Ausrichtung unterscheiden. Die obere und untere Lippe sind zusammen dargestellt, wobei sich die entsprechenden Gitter an den Mundwinkeln berühren. Die Zunge wird von schräg vorne gezeigt, d. h. die Rippe 0 befindet sich an der Zungenwurzel und die Rippe 33 an der Zungenspitze. Die Rippen 34–36 stellen die Verbindung zwischen der Zunge und dem Unterkiefer (Mundboden) her. Die Formen und Abmessungen der Gitter orientieren sich hauptsächlich an anatomischen Daten aus der Literatur und an eigenen Messungen an Röntgenaufnahmen des Vokaltrakts [86].

Wir wenden uns nun der konkreten geometrischen Beschaffenheit der einzelnen Gitter zu. Als wichtigstes Element zur Gestaltung der Gitterformen habe ich rationale Bézierkurven verwendet. Diese Kurven sind sowohl ein einfaches wie auch flexibles Werkzeug für die geometrische Modellierung [37]. Die *rationalen* Bézierkurven haben gegenüber den normalen Bézierkurven den Vorteil, dass sich mit ihnen alle Arten von Kegelschnitten exakt darstellen lassen, d. h. sowohl Kreis- und Ellipsenbögen als auch Parabel- und Hyperbelbögen. Eine rationale Bézierkurve wird durch $n + 1$ Kontrollpunkte P_i mit $i = 0 \dots n$ definiert, denen jeweils ein Gewicht w_i zugeordnet ist. Die Gleichung für die Kurve lautet

$$Q(t) = \frac{\sum_{i=0}^n P_i \cdot w_i \cdot B_i^n(t)}{\sum_{i=0}^n w_i \cdot B_i^n(t)} \quad \text{mit} \quad B_i^n(t) = \binom{n}{i} \cdot t^i \cdot (1-t)^{n-i},$$

wobei $B_i^n(t)$ als Bernsteinpolynom und t als Kurvenparameter bezeichnet wird. Wenn t zwischen 0 und 1 läuft, dann liegt die Kurve $Q(t)$ in der konvexen Hülle der Kontrollpunkte (bzw. des Kontrollpolygons), und insbesondere gilt $Q(0) = P_0$ und $Q(1) = P_n$. Für die Modellierung des Vokaltrakts habe ich zumeist Kurven vom Grad $n = 2$ oder $n = 3$ verwendet. Abbildung 2.5 zeigt, wie mit einer rationalen Bézierkurve zweiten Grades verschiedene Kegelschnitte konstruiert werden können. Dabei sind die Gewichte an den Kontrollpunkten P_0 und P_2 jeweils 1, und das Gewicht des Kontrollpunktes P_1 entscheidet darüber, welcher Kegelschnitt entsteht. Für $w_1 > 1$ entsteht ein Hyperbelbogen, für $w_1 = 1$ ein Parabelbogen und für $w_1 < 1$ eine Ellipsenbogen. Man erkennt, dass die Kurve umso stärker von P_1 „angezogen“ wird, je größer das Gewicht w_1 ist.

Obere Hülle, untere Hülle, Zähne.

In Abbildung 2.6 sind die Teilgitter der oberen und unteren Hülle sowie der Zähne maßstabsgetreu in orthogonalen Schnitten dargestellt. Jedes Teilgitter ist hier in einem eigenen *lokalen Koordinatensystem* definiert. Später werden sie durch affine Transformationen in

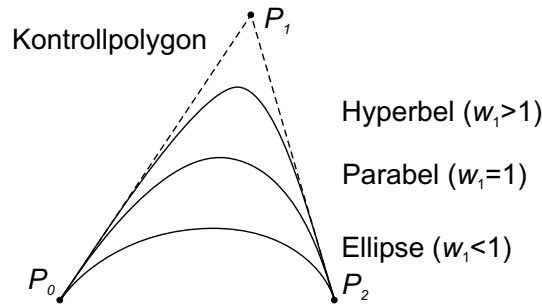


Abbildung 2.5: Darstellung unterschiedlicher Klassen von Kegelschnitten mit rationalen Bézierkurven durch die Variation des Gewichts von P_1 .

das *Modellkoordinatensystem* überführt und zusammengesetzt. Die breiten durchgezogenen Linien stellen je nach Ansicht die Konturlinie eines Gitters oder die Form einer Rippe dar. Die Koordinatenachsen und die orthogonalen Projektionen einzelner Rippen sind als schmale Linien dargestellt. Die gestrichelten Linien definieren die Kontrollpolygone für die Rippen und Konturlinien, die als Bézierkurven modelliert wurden. Die Zahlen an den Eckpunkten eines Kontrollpolygons geben das Gewicht an den entsprechenden Kontrollpunkten an. Wenn an einem Kontrollpunkt keine Zahl angegeben ist, dann hat er ein Gewicht von 1. Auch wenn für eine bessere Anschaulichkeit in einigen der Ansichten die linke *und* rechte Gitterhälfte abgebildet sind, so wurde, wie bereits erwähnt, nur die linke Seite tatsächlich als Gitter modelliert. Bei der Darstellung einzelner Rippen in den Ansichten (a) bis (d) und (m) ist die Verteilung der Gitterpunkte entlang der Rippen für die linke Modellhälfte dargestellt (die linke Hälfte befindet sich im Halbraum mit negativen z -Koordinaten).

Die Form von einigen Teilgittern hängt von den Vokaltraktparametern ab. In Tab. 2.2 sind alle Parameter des Vokaltraktmodells zusammengestellt, von denen jeder mit einem Kürzel aus Großbuchstaben bezeichnet wird und einen bestimmten Wertebereich besitzt.

In Abb. 2.7 ist schematisch der Einfluss der Parameter dargestellt, die sich auf die Form der Hülle und der Lippen des Modells auswirken. Wie die Parameter die Teilgitter im Einzelnen beeinflussen, wird nachfolgend erläutert.

Die Form des harten Gaumens, des Unterkiefers sowie der oberen und unteren Zahnreihe ist unveränderlich und kann unmittelbar aus den Ansichten (b), (c), (f) und (i) rekonstruiert werden. Die Zähne sind zur Abgrenzung gegenüber den Gittern des harten Gaumens und des Unterkiefers grau hinterlegt. Sowohl die mediosagittalen Konturen als auch die Rippen des harten Gaumens und Unterkiefers sind als Bézierkurven definiert. Die Rippen des harten Gaumens und des Unterkiefers sind alle senkrecht zur xy -Ebene ausgerichtet und die Rippen der Zähne senkrecht zur xz -Ebene.

Die Form des Velums wird durch den Parameter VO bestimmt, der ein Maß dafür ist, wie weit das Gaumensegel abgesenkt ist. Der Wertebereich von VO ist das Intervall $[-1, 1]$, wobei die Werte -1, 0 und 1 speziellen Zuständen entsprechen. Für $VO = 1$ ist das Velum maximal abgesenkt, für $VO = -1$ ist es maximal angehoben und für $VO = 0$ ist es nur soweit angehoben, dass die velopharyngale Pforte gerade so verschlossen ist. Diese drei Zustände spiegeln die Tatsache wider, dass die Höhe des Velums bei einer verschlossenen velopharyngalen Pforte in gewissen Grenzen variabel ist [12, 83]. Abbildung 2.6 (e) zeigt die

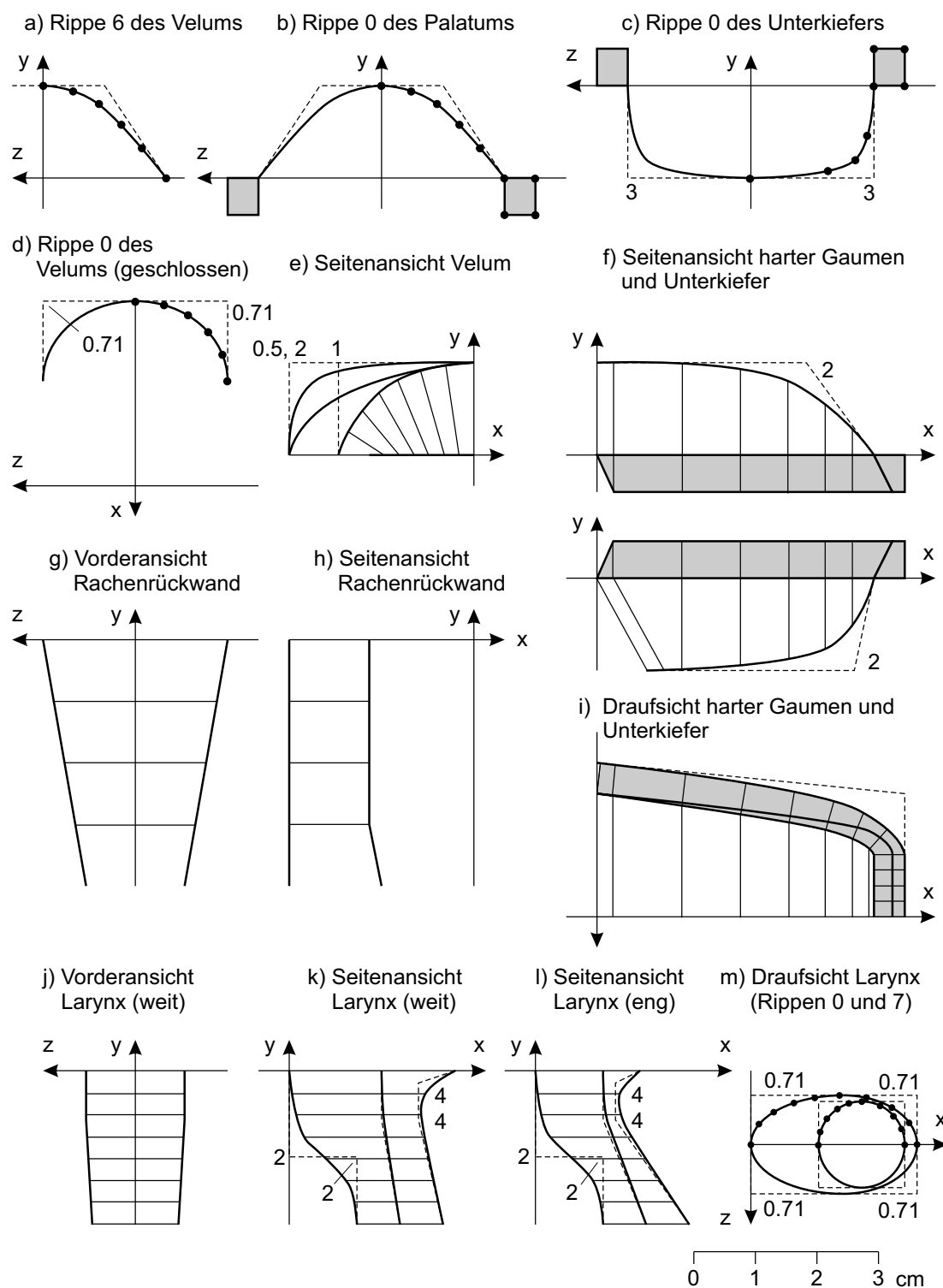


Abbildung 2.6: Geometrie der oberen Hülle, der unteren Hülle und der Zahngitter.

Bezeichnung	Wertebereich	Einheit	Bestimmung
<i>HX</i>	[0, 1]		Maß für die horiz. Zungenbeinposition
<i>HY</i>	[−5.5, −3.5]	cm	<i>y</i> -Koordinate des Zungenbeins
<i>JX</i>	[−0.5, 0.5]	cm	<i>x</i> -Koordinate des Unterkiefers
<i>JY</i>	[−1.8, −1.2]	cm	<i>y</i> -Koordinate des Unterkiefers
<i>JA</i>	[−0.2, 0.0]	rad	Öffnungswinkel des Unterkiefers
<i>LP</i>	[−1.0, 1.0]		Maß für die Vorstülpung der Lippen
<i>LH</i>	[−1.0, 4.0]	cm	Vert. Abstand zw. Ober- und Unterlippe
<i>VO</i>	[−1.0, 1.0]		Maß für die Öffnung des Velums
<i>TCX</i>	[−3.0, 4.0]	cm	<i>x</i> -Koordinate des Zungenkörpers
<i>TCY</i>	[−3.0, 1.5]	cm	<i>y</i> -Koordinate des Zungenkörpers
<i>TCR</i>	[1.0, 2.0]	cm	Radius des Zungenkörpers
<i>TTX</i>	[1.5, 4.5]	cm	<i>x</i> -Koordinate der Zungenspitze
<i>TTY</i>	[−3.0, 1.5]	cm	<i>y</i> -Koordinate der Zungenspitze
<i>TBX</i>	[−3.0, 4.0]	cm	Horizontale Zungenblattposition
<i>TBY</i>	[−3.0, 5.0]	cm	Vertikale Zungenblattposition
<i>TRE</i>	[−4.0, 4.0]	cm	Krümmung an der Zungenwurzel
<i>TS1</i>	[−1.4, 1.0]	cm	Zungenseitenhöhe bei der Wurzel
<i>TS2</i>	[−1.4, 1.0]	cm	Zungenseitenhöhe am Rücken
<i>TS3</i>	[−1.4, 1.0]	cm	Zungenseitenhöhe am Blatt
<i>TS4</i>	[−1.4, 1.0]	cm	Zungenseitenhöhe bei der Spitze
<i>MA1</i>	[0.0, 0.3]	cm ²	Min. Querschnitt hinter der Zungenspitze
<i>MA2</i>	[0.0, 0.3]	cm ²	Min. Querschnitt an der Zungenspitze
<i>MA3</i>	[0.0, 0.3]	cm ²	Min. Querschnitt vor der Zungenspitze

Tabelle 2.2: Parameter des Vokaltraktmodells.

mediosagittalen Konturen des Velums, die den drei Zuständen entsprechen. Die Konturen wurden wiederum mit Bézierkurven zweiter Ordnung modelliert, wobei sich die beiden angehobenen Zustände nur durch das Gewicht des mittleren Kontrollpunkts unterscheiden. Für Werte des Parameters *VO*, die zwischen -1, 0 und 1 liegen, wird die Kontur linear zwischen diesen Zuständen interpoliert. Die Rippen des Velums sind senkrecht zur *xy*-Ebene ausgerichtet. Die erste und letzte Rippe sind in Abb. 2.6 (d) und (a) dargestellt. Das Gewicht des mittleren Kontrollpunktes der Bézierkurve von Rippe 0 ist $1/\sqrt{2} \approx 0.71$. Dadurch bekommt die Rippe die Form einer Ellipse. Das Gewicht und die Position des mittleren Kontrollpunkts für die Rippen 1...5 wird linear zwischen denen der Rippen 0 und 6 interpoliert.

Unter dem Velum befindet sich das Gitter für die Rachenrückwand, welches in der Vorder- und Seitenansicht in Abb. 2.6 (g) und (h) dargestellt ist. Die Form der Rippen ist analog zur Ansicht (d) ellipsenförmig. Die dargestellte Länge der Rachenrückwand beträgt 4 cm. Tatsächlich wird diese Länge auf $|HY|$ skaliert, d. h. die Rachenlänge nimmt zu oder ab, wenn das Zungenbein abgesenkt oder angehoben wird.

Die Form des Larynx wird durch den Parameter *HX* bestimmt, der ein Maß für die horizontale Position des Zungenbeins ist. Laut MERMELSTEIN [77] besteht eine Korrelation zwischen der Zungenbeinposition und der Weite des Larynx: Je weiter vorne sich das Zungenbein befindet, desto größer ist der Hohlraum im Larynx. Um diese Beziehung nachzu-

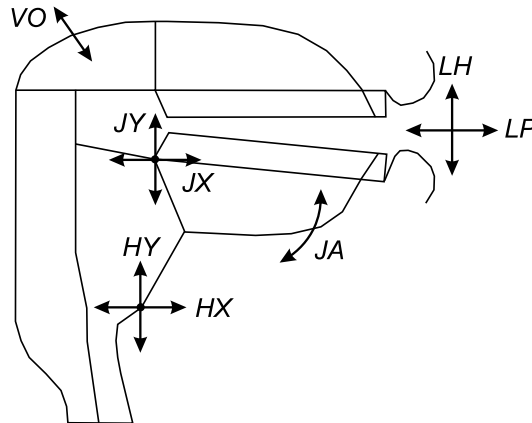


Abbildung 2.7: Schematische Darstellung der Beeinflussung der Vokaltraktkontur durch die Parameter für die Hülle, die Zähne und die Lippen.

bilden, wird die Form des Larynx in Abhängigkeit vom Parameter HX zwischen zwei vordefinierten Formen interpoliert: Eine Form für einen maximal engen Hohlraum ($HX = 0$) und eine Form für einen maximal weiten Hohlraum ($HX = 1$). Die Seitenansichten der Larynxgitter für die beiden Zustände sind in den Ansichten (k) und (l) dargestellt. Genaugenommen handelt es sich in jeder der Abbildungen um ein vorderes und ein hinteres Teilgitter. Das vordere Teilgitter gehört zur unteren Hülle und das hintere Teilgitter zur oberen Hülle. In der Draufsicht (m) sind die untersten und obersten Rippen der beiden Teilgitter des weiten Larynx dargestellt. Sie haben die Form von Ellipsenbögen.

In Abb. 2.8 ist dargestellt, wie die Teilgitter aus Abb. 2.6 in das Modellkoordinatensystem überführt werden. Das Velum, der harte Gaumen und die oberen Zähne brauchen nicht transformiert werden, da der lokale Koordinatenursprung dieser Gitter mit dem Koordinatenursprung des Modellkoordinatensystems (Punkt A) übereinstimmt. Der Unterkiefer und die unteren Zähne werden dagegen zunächst um den Winkel JA rotiert und anschließend in den Punkt $B = (JX, JY)^T$ verschoben. Die Parameter JA , JX und JY definieren also den Öffnungswinkel und die Position des Unterkiefers. Das hintere und vordere Larynxgitter werden an den Punkt C mit den Koordinaten $(-3, HY)^T$ verschoben. Der Parameter HY bestimmt also die Höhe des Larynx bzw. des Zungenbeins. Die gepunkteten Linien in Abb. 2.8 zeigen die Orientierung der drei Rippen für die Rachenvorderseite. Sie sind das Bindeglied zwischen dem vorderen Larynxgitter, dem Unterkiefer und der Rachenrückwand. In der mediosagittalen Projektion verlaufen sie jeweils vom Zungenbein (Punkt D) zu den Rippen 0, 1 und 2 der Rachenrückwand. Die Form dieser Rippen ist elliptisch, wie man an der perspektivischen Projektion in Abb. 2.4 erkennt.

Lippen.

Die geometrische Modellierung der Lippen baut auf einer Untersuchung von ABRY *et al.* [1] auf. In ihrer Arbeit wurden Abhängigkeiten zwischen verschiedenen geometrischen Abmessungen der Lippen bei der Produktion französischer Vokale und Konsonanten untersucht. Die untersuchten Abmessungen waren u. a. die horizontale Breite der Mundöffnung, der vertikale Abstand zwischen den Lippen sowie die Vorstülpung der Lippenwinkel und der

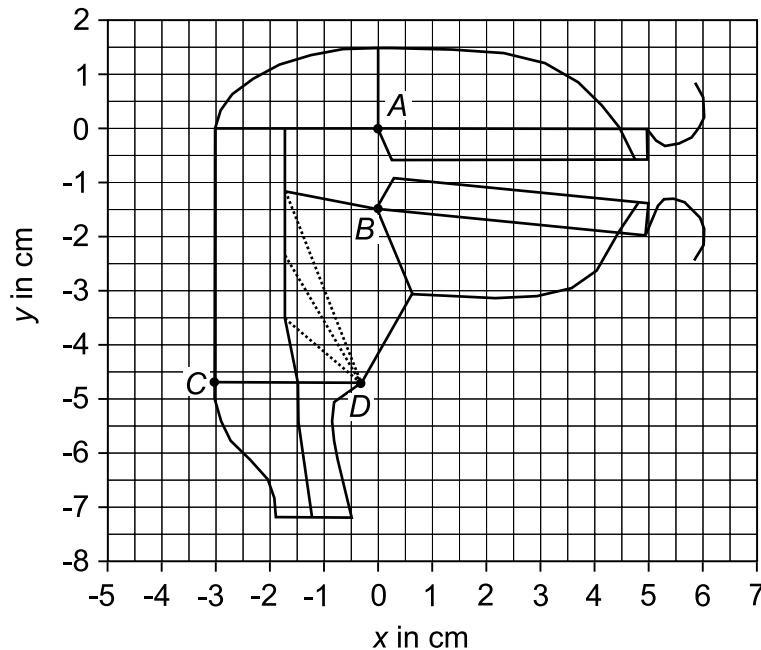


Abbildung 2.8: Komplette mediosagittale Kontur des Vokaltrakts im Modellkoordinatensystem.

Vorderseite der Ober- und Unterlippe. Die Untersuchung hat eine starke Korrelation zwischen der Vorstülpung der Lippenwinkel einerseits und der Vorstülpung der Oberlippe, der Unterlippe und der Breite der Mundöffnung andererseits ergeben. Die Korrelation zwischen der Vorstülpung der Lippenwinkel und dem vertikalen Lippenabstand war dagegen gering. Auf dieser Grundlage habe ich zwei Parameter für die Lippen des Vokaltraktmodells definiert: Die Vorstülpung LP der Mundwinkel und der vertikale Abstand LH zwischen der Ober- und Unterlippe. Aufgrund der oben genannten Korrelationen lassen sich mit LP die Breite der Mundöffnung und die Vorstülpung der Ober- und Unterlippe vorhersagen. Dadurch kann mit LP und LH auf alle wichtigen Abmessungen der Lippenform geschlossen werden.

Die labiodentalen Laute [f] und [v] waren in dem von ABRY *et al.* verwendeten Korpus nicht enthalten, da sich ihre Lippenstellung von den anderen Vokalen und Konsonanten besonders unterscheidet. In dem statistischen Vokaltraktmodell von BEAUTEMPS *et al.* [9] wird aus diesem Grund ein spezieller Lippenparameter eingeführt, der die Lippen so verformt, dass eine labiodentale Konstriktion entsteht. In der vorliegenden Arbeit realisiere ich die labiodentale Konstriktion mit dem Vokaltraktmodell durch eine Verschiebung des Unterkiefers nach hinten (Parameter JX). Der Unterkiefer und die Unterlippe sind im Modell so miteinander verbunden, dass die Unterlippe der Bewegung des Unterkiefers folgt und dadurch zum Teil unabhängig von der Oberlippe bewegt werden kann.

Abbildung 2.9 zeigt das Profil der Lippen in der Seitenansicht. Dort sind mehrere Punkte eingezeichnet, die bei der Konstruktion der Lippengitter verwendet werden. Die Punkte T_0 , T_1 , F_0 und F_1 liegen in der mediosagittalen Ebene und markieren die obere Kante der oberen Schneidezähne, die untere Kante der unteren Schneidezähne, den tiefsten Punkt der

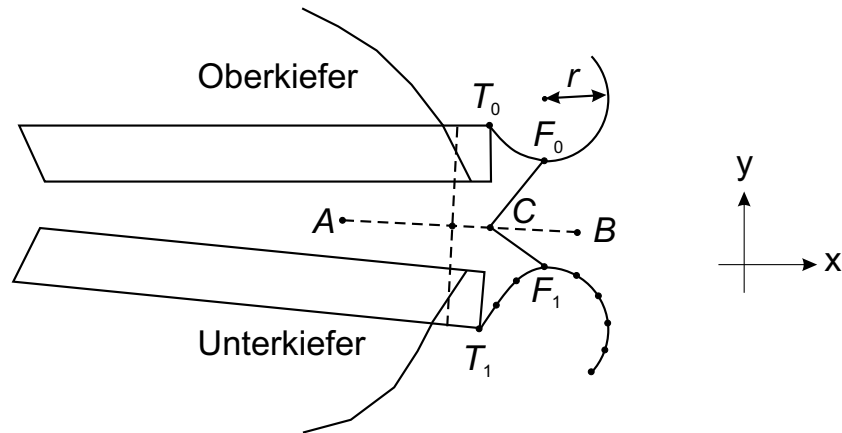
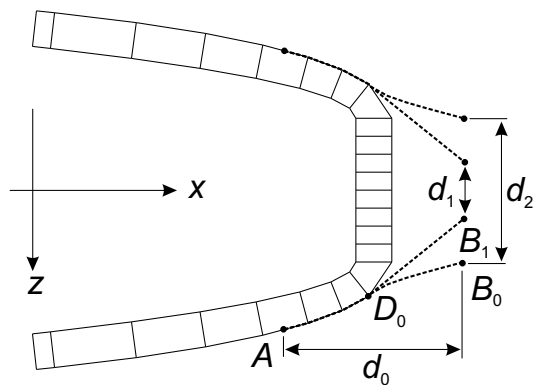


Abbildung 2.9: Lippenprofil in der Seitenansicht.

Oberlippe und den höchsten Punkt der Unterlippe. Der Punkt C markiert den Lippenwinkel und liegt auf einer Linie, die die Punkte A und B miteinander verbindet. Die Linie AB wird nachfolgend als *Führungslinie* bezeichnet.

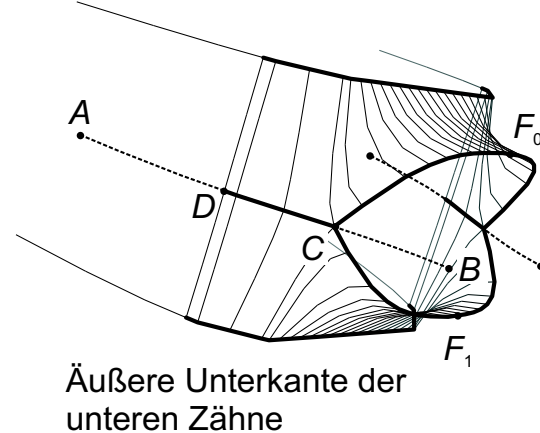
Die Lage von T_0 und T_1 ist von den Gittern der oberen und unteren Zähne bereits bekannt. Die Position des Mundwinkels C auf der Führungslinie hängt dagegen vom Parameter LP ab. Für $LP = LP_{\min}$ befindet er sich im Punkt A , für $LP = LP_{\max}$ im Punkt B und für Werte zwischen $LP_{\min}$ und $LP_{\max}$ entsprechend zwischen A und B ($LP_{\min}$ und $LP_{\max}$ sind die Intervallgrenzen in Tab. 2.2). In der mediasagittalen Parallelprojektion ist die Führungslinie stets so ausgerichtet, dass sie in jedem Punkt den gleichen vertikalen Abstand zu den oberen und unteren Zähnen besitzt. Wenn sich die oberen und unteren Zähne berühren, dann ist sie also parallel zur x -Achse ausgerichtet. Für diesen Fall sind zwei Varianten

a) Draufsicht obere Zähne



b)

Äußere Oberkante der oberen Zähne



Äußere Unterkante der unteren Zähne

Abbildung 2.10: (a) Führungslinien für den Lippenwinkel. (b) Innere Rippenabschnitte der Ober- und Unterlippe.

der Führungslinie in der Draufsicht in Abb. 2.10 (a) dargestellt (AB_0 und AB_1), wobei $d_0 = 2.5$ cm, $d_1 = 0.8$ cm und $d_2 = 2.0$ cm. Für beide Führungslinien nimmt die Mundbreite mit zunehmender Vorstülpung der Lippenwinkel ab (negative Korrelation zwischen der Mundbreite und der Vorstülpung der Lippenwinkel entsprechend ABRY *et al.* [1]). Die zwei verschiedenen Linien verwende ich, um die Veränderung der Mundbreite zu modellieren, die entsteht, wenn sich bei gleichbleibender Vorstülpung des Lippenwinkels der Abstand zwischen Ober- und Unterlippe ändert¹. Daher verwende ich die Linie AB_0 als Führungslinie, wenn $LH = LH_{\max}$ und die Linie AB_1 , wenn $LH = LH_{\min}$. Wenn der Lippenabstand zwischen $LH_{\min}$ und $LH_{\max}$ liegt, dann wird die Führungslinie entsprechend zwischen AB_0 und AB_1 interpoliert.

Sobald die Position von C bekannt ist, können damit die Punkte F_0 und F_1 berechnet werden. Die x -Koordinate der Punkte berechne ich mit den Gleichungen

$$\begin{aligned} F_{0,x} &= \max\{C_x + L, T_{0,x} + 1 \text{ mm}\} \\ F_{1,x} &= \max\{C_x + L, T_{1,x} + 1 \text{ mm}\}, \end{aligned}$$

d. h. F_0 und F_1 liegen entweder um die Länge L vor dem Lippenwinkel oder befinden sich direkt 1 mm vor den Schneidezähnen. Die Länge L nimmt laut ABRY *et al.* [1] mit zunehmender Vorstülpung der Lippenwinkel ab. Ich verwende die Beziehung $L = 1 - t \cdot 0.7$ cm, wobei $t = (LP - LP_{\min}) / (LP_{\max} - LP_{\min})$. Die y -Koordinaten von F_0 und F_1 berechne ich mit den Gleichungen

$$F_{0,y} = y_c + LH/2 \quad \text{und} \quad F_{1,y} = y_c - LH/2,$$

wobei $y_c = (T_{0,y} + T_{1,y})/2 + (1-t) \cdot 0.35$ cm die vertikale Position ist, an der sich die Lippen beim Verschluss berühren. Wenn LH negativ ist, dann ist $F_{0,y} = F_{1,y} = y_c$.

Nachdem auch die Positionen von F_0 und F_1 bekannt sind, können die Rippen der Lippengitter berechnet werden. Jede der Rippen besteht aus einem „inneren“ Abschnitt und einem „äußeren“ Abschnitt. Die inneren Rippenabschnitte repräsentieren den nach innen bzw. hinten gewandten Teil der Lippenoberfläche, der sich unmittelbar vor den Schneidezähnen befindet. Die äußeren Rippenabschnitte repräsentieren dagegen den nach außen bzw. vorne gewandten Teil der Lippenoberfläche. Jede Rippe liegt in einer Ebene parallel zur xy -Ebene. Sie sind so entlang der z -Achse verteilt, dass sie eine lückenlose Verbindung zu den entsprechenden Zahngittern herstellen. Die inneren Rippenabschnitte sind in Abb. 2.10 (b) dargestellt und verbinden entweder die Oberkante der oberen Zähne mit der Linie DCF_0 oder die Unterkante der unteren Zähne mit der Linie DCF_1 . Die Position des Punktes D hängt von der Lippenwinkelposition C ab. Wenn C sich vor dem „Ablösepunkt“ D_0 in Abb. 2.10 (a) befindet, so ist $D = D_0$. Ansonsten, also bei zurückgezogenen Mundwinkeln, befindet sich D an der gleichen Position wie C .

Der Abschnitt DC der Linien DCF_0 und DCF_1 liegt direkt auf der Führungslinie, während die Abschnitte CF_0 und CF_1 jeweils eine Bézierkurve zweiten Grades sind. Auch die inneren Abschnitte der Rippen selbst wurden mit Bézierkurven zweiten Grades modelliert. Die äußeren Rippenabschnitte haben die Form eines Kreisbogens. Der Radius

¹Als Lippenwinkel werden hier nicht die tatsächlichen Mundwinkel betrachtet, sondern die Punkte, an denen sich die Ober- und Unterlippe voneinander trennen. Beim [u] liegen diese Trennpunkte z. B. weiter vorne als die „echten“ Lippenwinkel.

der Kreisbögen beträgt für die am weitesten nach vorne reichenden Rippen 0.7 cm und nimmt für die Rippen, die im Lippenwinkel oder dahinter enden, auf Null ab. In Abb. 2.9 ist die Form der in der mediosagittalen Ebene liegenden Rippen der Ober- und Unterlippe dargestellt. Für die Unterlippe sind auch die Rippenpunkte eingezeichnet, von denen die vier Punkte zwischen T_1 und F_1 zum inneren Rippenabschnitt und die restlichen fünf Punkte zum äußeren Rippenabschnitt gehören.

Zunge.

Die Modellierung der Zunge erfolgt, indem zuerst die Kontur der Zunge in der medio-sagittalen Ebene berechnet wird und anschließend die entlang der Kontur angeordneten Rippen geformt und ausgerichtet werden. Der Aufbau der Zungenkontur ist in Abb. 2.11

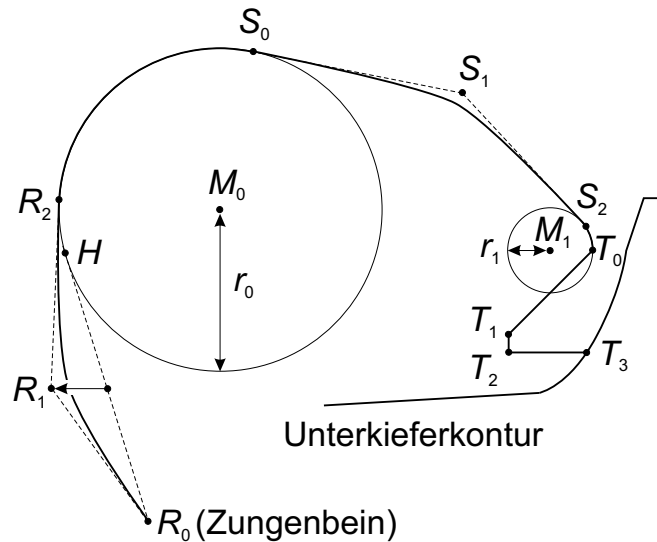


Abbildung 2.11: Kontur der Zunge in der mediosagittalen Ebene.

dargestellt. Sie besteht aus den zwei rationalen quadratischen Bézierkurven mit den Kontrollpolygonen $R_0R_1R_2$ und $S_0S_1S_2$, aus den zwei Kreisbögen R_2S_0 und S_2T_0 und dem Linienzug $T_0T_1T_2T_3$. Der Kreisbogen R_2S_0 liegt auf einem Kreis mit dem Mittelpunkt M_0 und dem Radius r_0 . Dieser Kreis repräsentiert den Hauptanteil bzw. das Zentrum der Zunge und wird nachfolgend als Zungenkörper bezeichnet. Der Kreisbogen S_2T_0 liegt auf dem Kreis mit dem Mittelpunkt M_1 und dem Radius r_1 , der die Zungenspitze repräsentiert. Der Mittelpunkt und Radius des Zungenkörpers sind direkt durch die Parameter TCX , TCY und TCR gegeben, d. h. $M_0 = (TCX, TCY)^T$ und $r_0 = TCR$. Der Mittelpunkt der Zungenspitze wird von den Parametern TTX und TTY bestimmt, d. h. $M_1 = (TTX, TTY)^T$, während der Radius r_1 mit 0.4 cm als konstant angenommen wird.

Die Kurve $R_0R_1R_2$ formt die Zungenrückseite und verbindet das Zungenbein R_0 (entspricht dem Punkt D in Abb. 2.8) mit dem Zungenkörper. Die Position der Punkte R_1 und R_2 ergibt sich wie folgt: Zuerst wird von R_0 aus eine Tangente an die Hinterseite des Zungenkörpers gelegt, die diesen im Punkt H berührt. Die Position von R_1 ist dann $\frac{1}{2}(R_0 + H) + (TRE, 0)^T$. Der Parameter TRE gibt also an, wie stark die Zungenrückseite

gekrümmt ist. R_2 ist der Berührungspunkt einer Tangente, die von R_1 aus an die Rückseite des Zungenkörpers gelegt wird.

Die Kurve $S_0S_1S_2$ bildet das Zungenblatt und verbindet die Kreise für den Zungenkörper und die Zungenspitze. Die Position von S_1 wird direkt durch die Parameter TBX und TBY vorgegeben, d. h. $S_1 = (TBX, TBY)^T$. S_0 und S_2 sind die Punkte, an denen die von S_1 aus angelegten Tangenten die Oberseiten der beiden Kreise berühren. Die Gewichte der Kontrollpunkte der beiden rationalen Bézierkurven wurden jeweils auf 1, 2, 1 gesetzt, so dass sich die Kurven dem mittleren Punkt jeweils etwas stärker annähern als bei einfachen Bézierkurven.

Der Linienzug $T_0T_1T_2T_3$ hat die Aufgabe, die Verbindung zwischen der Zungenspitze und dem Unterkiefer bzw. dem Mundboden herzustellen. T_0 liegt im Winkel von Null Grad auf dem Zungenspitzenkreis, und T_1 befindet sich in einem Abstand von 1.1 cm links unter T_0 . T_2 und T_3 befinden sich auf dem Mundboden, der 0.5 cm über dem tiefsten Punkt des Unterkiefers liegt. Wenn die Zungenspitze sehr tief liegt, dann wird der Punkt T_1 so angepasst, dass er nicht unterhalb des Zungenbodens zu liegen kommt.

In Tab. 2.2 ist für jeden der Zungenparameter der erlaubte Wertebereich als ein Intervall angegeben. Aufgrund der gegenseitigen Abhängigkeit der erlaubten Parameterwerte und der gekrümmten Form der oberen und unteren Hülle reichen die statischen Intervallgrenzen jedoch nicht aus. Deshalb werden die Parameterwerte durch zusätzliche Maßnahmen eingeschränkt. In Abb. 2.12 (a) ist als fette Linie der Bereich dargestellt, der von den beiden Zungenkreisen nicht „übertreten“ werden darf. Vorne ist der Bereich durch die Schneidezäh-

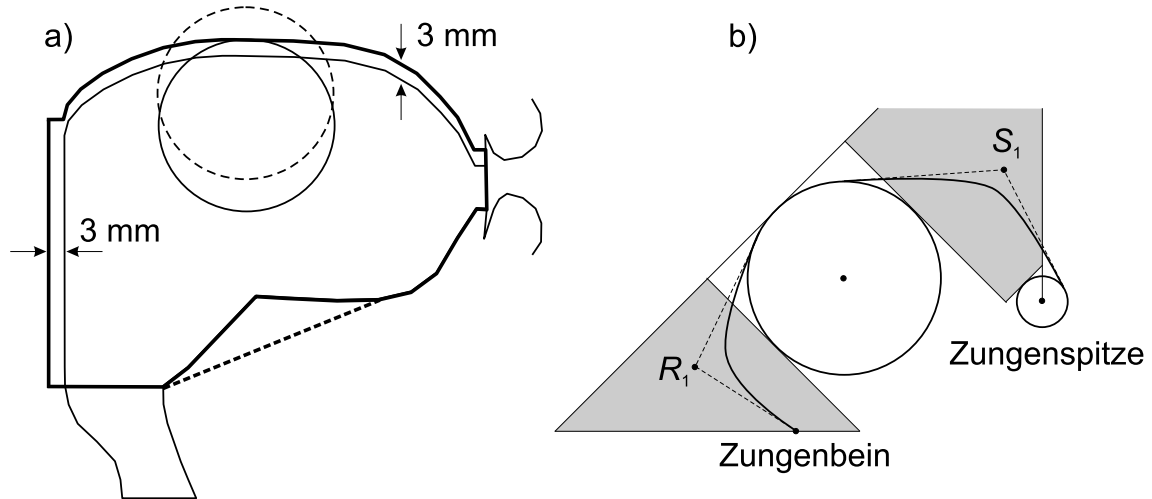


Abbildung 2.12: Beschränkungen der Zungenparameter.

ne begrenzt und unten durch die Konturlinie des Unterkiefers. Für den Zungenkörper ist der Bereich nach unten hin etwas größer und wird durch die gestrichelte Linie begrenzt. Oben und links wird der Bereich durch die Kontur des Gaumens und der Rachenrückwand eingeschränkt. Diese Grenzlinie folgt jedoch nicht direkt der Kontur, sondern lässt nach oben und links jeweils 3 mm mehr Platz für die Kreise. Dadurch können die Kreise geringfügig über die äußere Kontur des Vokaltrakts herübertreten. Die überstehenden Teile des Kreises

bzw. der daraus resultierenden Zungenkontur werden später der äußeren Kontur angepasst, so dass damit eine „Abflachung“ der Zunge an der Kontaktstelle mit der Vokaltraktwand modelliert wird. Diese Verformung ist oft auf Röntgenaufnahmen vom Vokaltrakt zu erkennen, wenn bestimmte Zungenteile gegen die äußeren Sprechtraktwände gepresst werden und sich dabei ihrer Form anpassen (z. B. beim [k] oder [n]). Die Einschränkung der Zungenkreiskoordinaten erfolgt zuerst durch die vertikalen linken und rechten Begrenzungslinien und anschließend durch die oberen und unteren Bereichsgrenzen. Abbildung 2.12 (a) zeigt z. B., wie der Zungenkörper von seiner ursprünglichen Position (gestrichelter Kreis) durch eine Translation in y -Richtung an eine erlaubte Position (durchgezogener Kreis) verschoben wird. Der Kreis für die Zungenspitze wird analog dazu behandelt. Außerdem darf sich sein Mittelpunkt M_1 nicht weiter links befinden als die x -Koordinate $M_{0,x} + r_0$.

Abbildung 2.12 (b) zeigt grau hinterlegt die Bereiche, in denen sich die mittleren Punkte R_1 und S_1 der Bézierkurven aufhalten dürfen. Die Bereiche ändern sich dynamisch mit der Position der Zungenkreise, wobei die Winkel und Berührungspunkte der Begrenzungslinien mit den Kreisen erhalten bleiben. Aus diesen Aufenthaltsbereichen ergeben sich die Intervallgrenzen für die Parameter TRE , TBX und TBY .

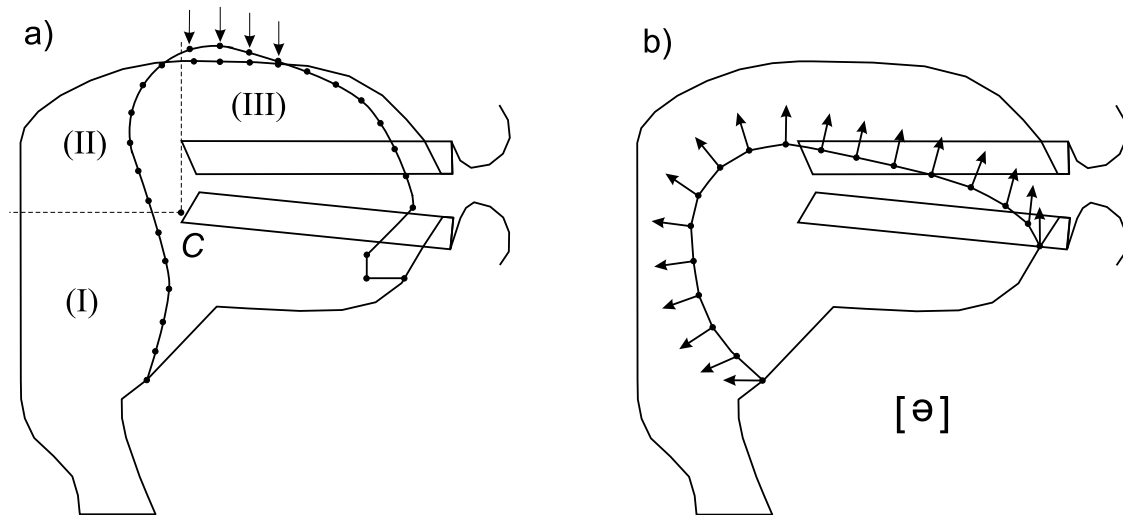


Abbildung 2.13: (a) Abtastung der mediosagittalen Zungenkontur und Verschiebung der Abtastpunkte an erlaubte Positionen. (b) Normalenvektoren der Zungenkontur für den neutralen Vokal [ə].

Für die Berechnung des Zungengitters wird die Zungenkontur gleichmäßig abgetastet und in jedem Abtastpunkt der Normalenvektor der Kontur bestimmt. Die Abtastpunkte entsprechen den Gitterpunkten der einzelnen Zungenrippen in der mediansagittalen Ebene, und die Normalen definieren die Ausrichtung der Rippen. Abbildung 2.13 (a) zeigt beispielhaft die Abtastung der Kontur. Vor der Berechnung der Normalen werden die Abtastpunkte, die sich außerhalb der Hüllkontur befinden, so verschoben, dass sie genau *auf* der Hülle liegen. Der dabei verwendete Verschiebungsvektor hängt von der Position des betrachteten Abtastpunktes ab. Dazu werden die drei Bereiche (I), (II) und (III) in Abb. 2.13 (a) unterschieden, die sich im Punkt $C = (0, -1.5)^T$ berühren. Im Bereich (I) werden die

betroffenen Punkte nach rechts verschoben, im Bereich (III) nach unten und im Bereich (II) radial in Richtung des Punktes C . Abbildung 2.13 (b) zeigt die Normalenvektoren an den Abtastpunkten der Zungenkontur für den Vokal [ə]. Um einen nahtlosen Übergang der Zungenrippen zur unteren Hülle und zum Mundboden zu gewährleisten, wurden die Normalenvektoren der Rippen 0 und 32 (siehe Abb. 2.4) auf $(-1, 0)^T$ und $(0, 1)^T$ gesetzt und die Rippen unmittelbar danach bzw. davor so angepasst, dass die Ausrichtung benachbarter Rippen nicht zu stark variiert.

Anschließend wird die Position einiger Abtastpunkte der Zungenkontur gegebenenfalls durch zwei weitere Parameter beeinflusst. Es handelt sich um die Parameter $MA1$ und $MA2$, die den minimalen Querschnitt des Ansatzrohres hinter der Zungenspitze und direkt an der Zungenspitze vorgeben. Diese Parameter ermöglichen die genaue Kontrolle über den Inhalt der Querschnittsflächen von Konstriktionen, die mit der Zunge und der oberen Hülle gebildet werden. Mit den geometriestimmenden Parametern allein sind *präzise* Flächenquerschnitte an supraglottalen Konstriktionen nämlich oft schwierig zu realisieren, da diese von der Genauigkeit der Kontaktbehandlung zwischen den beteiligten Gitternetzen abhängen. Diese Genauigkeit ist aber aufgrund der begrenzten Auflösung der Flächengitter beschränkt. Da das Geräusch von Frikativen aber empfindlich von den Querschnittsflächen der Konstriktionen abhängt [5], wurden die Parameter $MA1$ und $MA2$ eingeführt. Die Punkte der Zungenkontur sind insofern davon betroffen, dass sie ein kleines Stück entgegengesetzt der Richtung ihres Normalenvektors verschoben werden, wenn sie sich für den eingestellten minimalen Querschnitt zu dicht an der Kontur der oberen Hülle befinden. Die Weite der Korrekturbewegung wird konservativ über eine empirische Beziehung zwischen dem mediasagittalen Abstand Zungenpunkt-Hüllenkontur und der entsprechenden Querschnittsfläche abgeschätzt. Die *exakte* Berücksichtigung der minimalen Querschnittsflächen erfolgt erst bei der Berechnung der Flächenfunktion in Abschnitt 2.3.3. Neben den Parametern $MA1$ und $MA2$ gibt es noch einen dritten Parameter $MA3$, der den minimalen Querschnitt im Bereich der Schneidezähne und Lippen für labiodentale Konstriktionen vorgibt. Die Berücksichtigung dieses Parameters erfolgt jedoch ausschließlich durch eine Manipulation der Flächenfunktion.

Wir wenden uns nun den Rippen des Zungengitters zu, die der Zunge ihre endgültige dreidimensionale Form verleihen. Die Form jeder Rippe ist durch drei Punkte vollständig definiert: Den Abtastpunkt auf der mediasagittalen Kontur sowie jeweils einen Punkt für den linken und rechten Zungenrand. Die Randpunkte können relativ zum Punkt in der Zungenmitte höher oder tiefer liegen. Vorerst verwendet das Modell für die Höhe der beiden Ränder jeweils den gleichen Wert, so dass die Zunge symmetrisch aufgebaut ist. In Abb. 2.14 ist der Verlauf der Zungenränder (gestrichelte Linie) im Vergleich zur mediasagittalen Kontur der Zunge (durchgezogene Linie) für den Vokal [i:] dargestellt. An den Punkten $Q_0 \dots Q_4$ auf der Zungenkontur ist die Höhe der Zungenränder relativ zur Zungenmitte explizit vorgegeben, und zwischen diesen Punkten wird sie linear interpoliert. Mit Bezug auf Abb. 2.11 befinden sich Q_0 und Q_4 an den Punkten R_0 und T_0 und die Punkte $Q_1 \dots Q_3$ in der Mitte der Kurve $R_0R_1R_2$, des Kreisbogens R_2S_0 und der Kurve $S_0S_1S_2$. Die Randhöhe im Punkt Q_0 ist konstant, während sie in den Punkten $Q_1 \dots Q_4$ von den Parametern $TS1 \dots TS4$ vorgegeben wird.

Für den Vokal [i:] in Abb. 2.14 (a) sind die Parameter so eingestellt, dass die Zungenoberfläche im hinteren Teil eine Rille bildet (angehobene Ränder), während sie im vorderen Teil

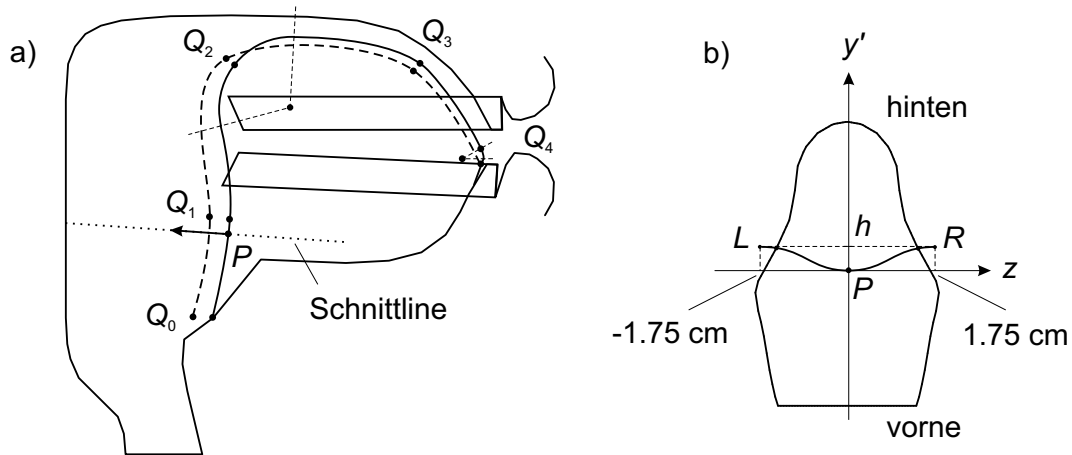


Abbildung 2.14: (a) Höhe der Zungenränder (gestrichelte Linie) für den Vokal [i:]. (b) Schnittbild durch das Modell in (a) im Punkt P für die Konstruktion der dortigen Zungenrippe.

konvex ist (abgesenkte Ränder). Das sind laut einer Untersuchung von *STONE et al.* [111] typische Formmerkmale der Zunge für vordere hohe Vokale. Hintere Vokale weisen demgegenüber eine typische konvexe Form im Bereich des Zungenrückens auf und tiefe Vokale eine durchgehende Rille in der Mitte der Zunge. Konsonanten zeigen darüber hinaus noch weitere typische Muster für die Höhe der Zungenränder.

Alle diese Formen lassen sich mit den Parametern $TS1 \dots TS4$ nachbilden. Zwar ist die Höhe der Zungenränder teilweise mit den Merkmalen der mediosagittalen Zungenkontur korreliert, solange diese Korrelationen aber nicht quantitativ untersucht wurden, halte ich die explizite Einstellung der Zungenrandhöhe für einen sinnvollen Ansatz.

Abbildung 2.14 (b) verdeutlicht am Beispiel eines Schnittes durch das Modell in Abb. 2.14 (a) entlang der gepunkteten Linie die Konstruktion der Zungenrippen (Die Berechnung der Schnittbilder wird in Abschnitt 2.3.3 erörtert). Der Koordinatenursprung in Abb. 2.14 (b) befindet sich im Punkt P auf der mediosagittalen Kontur und die Achse y' ist entlang der Normalen im Punkt P ausgerichtet. Die Punkte L und R markieren den linken und rechten Zungenrand und sind gegenüber der Zungenmitte um die Höhe h angehoben. Die Form der Rippe zwischen den Punkten L und R ist im Schnittbild durch eine Kurve gegeben, deren erste Ableitung nach z in den beiden Randpunkten und im Koordinatenursprung Null ist. Damit die Zungenrippen nicht durch die obere und untere Hülle hindurchragen, werden die Gitterpunkte der Rippen nur für den Bereich der Kurve abgetastet, der sich innerhalb der Schnittkontur befindet. Wenn dieser Bereich insgesamt schmäler als 1 cm werden sollte, so wird die Kurve trotzdem zwischen $z = -0.5$ und $z = 0.5$ abgetastet, und die Punkte, die über die Schnittkontur hinausragen, werden auf die Schnittkontur verschoben.

2.3.3 Berechnung der Mittellinie und der Flächenfunktion

Nach der Konstruktion der einzelnen Oberflächengitter kann die Flächenfunktion des Vokaltrakts berechnet werden. Dazu ist es zuerst erforderlich, die Mittellinie des Vokaltrakts zu bestimmen, d. h. die Linie, entlang der sich die Schallwellen ausbreiten. Anschließend müssen die Querschnitte durch das Modell senkrecht zur Mittellinie berechnet werden und die resultierenden Flächeninhalte über der Position auf der Mittellinie abgetragen werden. Die Vorgehensweise bei der Berechnung der Mittellinie ist in den Abbildungen 2.15 und 2.16 dargestellt. Abbildung 2.15 (a) zeigt eine erste grobe Approximation der Mittellinie,

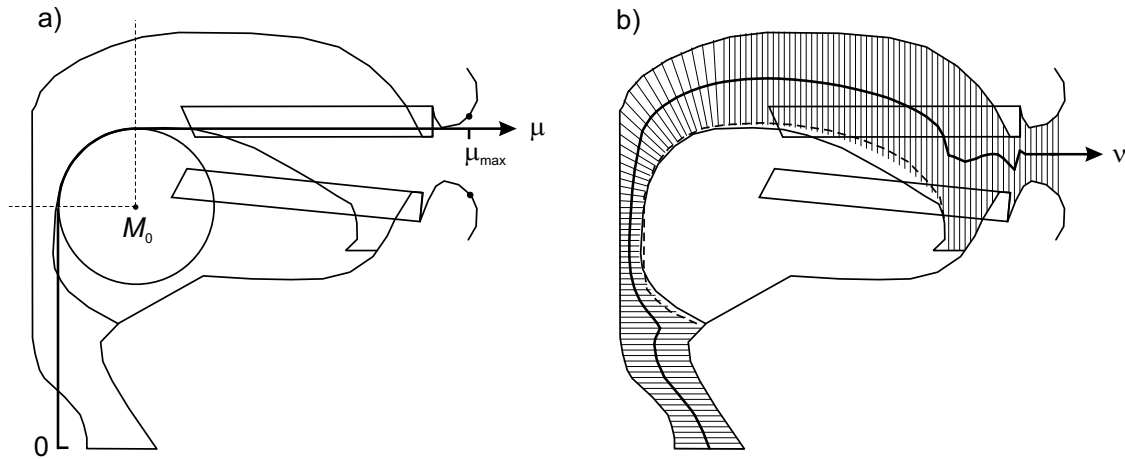


Abbildung 2.15: Erste und zweite Näherung der Mittellinie.

die aus einem vertikalen Streckenabschnitt, einem Kreisbogen um den Mittelpunkt M_0 und einem horizontalen Streckenabschnitt zusammengesetzt ist. Diese Linie wird in gleichmäßigen Abständen zwischen den Positionen 0 und $\mu_{\max}$ mit 128 Punkten abgetastet. Die Position $\mu_{\max}$ auf der Linie befindet sich kurz hinter der Vorderseite der Ober- und Unterlippe. Jeder der Abtastpunkte ist der Ausgangspunkt einer Schnittpunktlinie, die orthogonal zur Mittellinie in Abb. 2.15 (a) ausgerichtet ist. Für jede der orthogonalen Linien werden die Schnittpunkte mit den Konturlinien des Vokaltrakts bestimmt. Die Mittelpunkte der Schnittpunkte mit der oberen Kontur einerseits und der unteren Kontur bzw. der Zungenkontur andererseits ergeben miteinander verbunden die nächste Näherung für die Mittellinie, die in Abb. 2.15 (b) dargestellt ist (zusammen mit den Schnittpunkten).

Das bis hierher beschriebene Verfahren zur Berechnung der Mittellinie ist prinzipiell identisch mit dem von MERMELSTEIN [77]. Da die Mittellinie jedoch noch einige deutliche „Knicke“ in ihrem Verlauf aufweist, verwende ich zusätzlich einen Tiefpassfilter, um die Linie zu glätten. Das Prinzip der Filterung ist in Abb. 2.16 (a) dargestellt und besteht darin, dass ein Fenster mit der Länge von 2 cm (Linienabschnitt zwischen den Punkten S und T) in gleichmäßigen Abständen über die Mittellinie geschoben wird, und an jeder betrachteten Position der Schwerpunkt (Punkt Q) des Linienabschnitts berechnet wird. Die Berechnung des Schwerpunktes wird in 128 äquidistanten Fenstern auf der Mittellinie durchgeführt, und ihre Verbindung ergibt die endgültige Form der Mittellinie in Abb. 2.16 (b). Für jeden Punkt der endgültigen Mittellinie ist zusätzlich die orthogonale Schnittpunktlinie dargestellt.

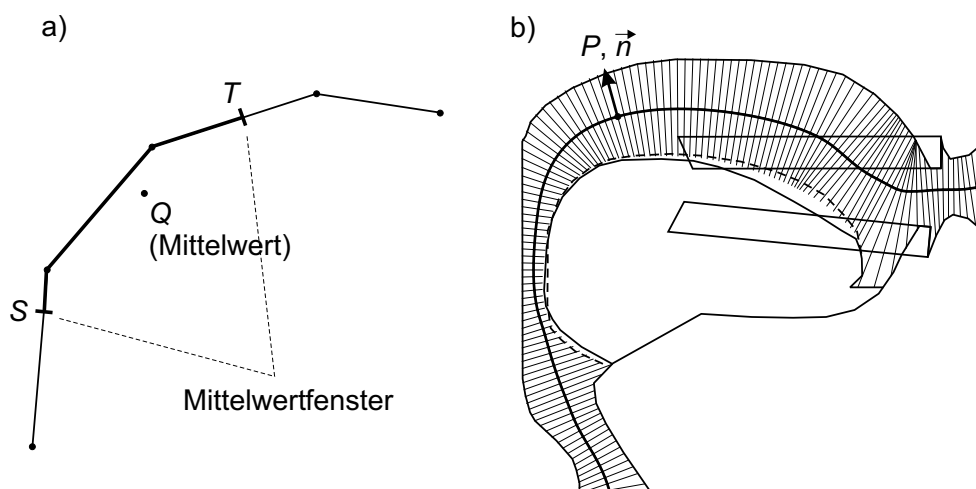


Abbildung 2.16: (a) Prinzip der Glättung der Mittellinie und (b) Ergebnis der Glättung.

Abbildung 2.17 zeigt am Beispiel des Vokals [a:] die Schnittbilder, die sich an drei verschiedenen Positionen entlang der Mittellinie ergeben. Die Berechnung der Schnittbilder wird

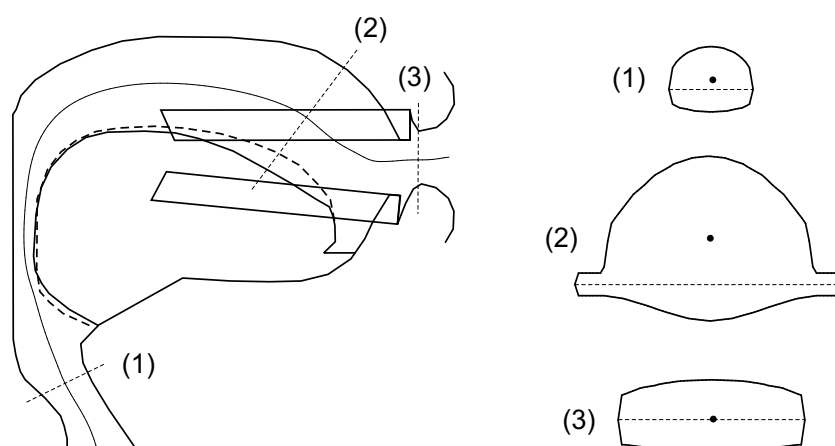


Abbildung 2.17: Querschnitte an verschiedenen Stellen für den Vokal [a:]. Die Punkte in den Schnittbildern zeigen die Position der Mittellinie, und die gestrichelten Linien unterteilen die Schnitte in einen oberen und unteren Umriss.

nachfolgend erläutert.

Jeder Querschnitt durch den Vokaltrakt lässt sich grundsätzlich in eine obere und eine untere Umrisslinie zerlegen. In Abb. 2.17 ist die Grenze zwischen dem oberen und unteren Umriss jeweils durch eine gestrichelte Linie eingezeichnet. Entsprechend Abb. 2.18 (a) ist sowohl der obere als auch der untere Umriss eine Funktion $y' = f(z)$, wobei y' die Umriss-

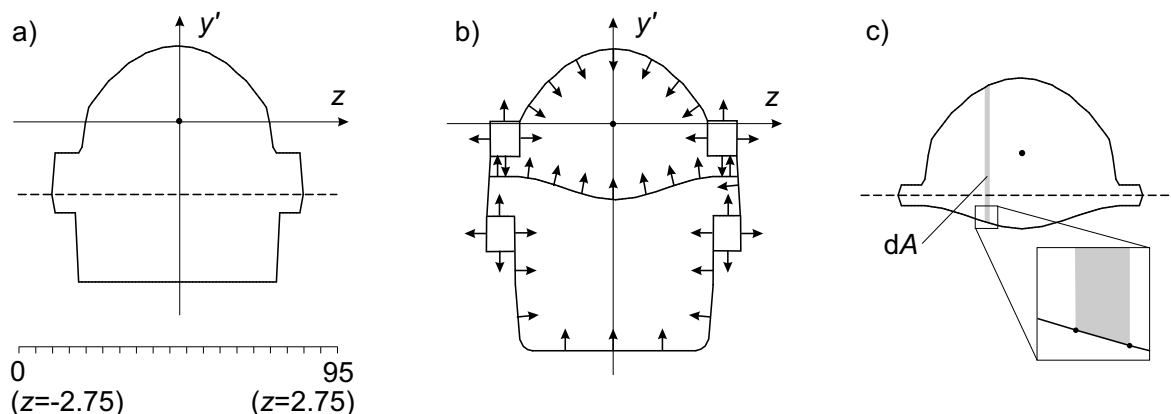


Abbildung 2.18: Berechnung der oberen und unteren Umrisslinien sowie des Flächeninhalts.

höhe im Schnittbild kennzeichnet². Jede der Umrissfunktionen wird durch eine abgetastete Funktion mit 96 Punkten repräsentiert, die gleichmäßig zwischen $z = -2.75$ und $z = 2.75$ verteilt sind. Die maximal mögliche Breite eines Querschnitts beträgt damit 5.5 cm und das Abtastintervall etwas mehr als 0.5 mm. Die Funktionswerte der abgetasteten Umrisse ergeben sich aus den Schnittlinien zwischen den Dreiecksflächen der Gitternetze des Vokaltraktmodells und den senkrecht zur Mittellinie ausgerichteten Schnittebenen. Jede Schnittlinie wird im oben genannten Raster abgetastet und in den oberen oder unteren Umriss integriert.

In welchen der beiden Umrisse die Schnittlinie mit einem Dreieck integriert wird, entscheidet die Ausrichtung der auf die Schnittebene projizierte Oberflächennormale des Dreiecks. Zeigt die Normale im Schnittbild nach unten (y' -Komponente des Vektors ist kleiner als Null), so wird die Linie in den oberen Umriss integriert und ansonsten in den unteren Umriss. Damit das funktioniert, wurde die Richtung der Normalen bei der Initialisierung der Oberflächengitter so gewählt, dass sie in den „Innenraum“ des Vokaltrakts zeigen. In Abb. 2.18 (b) ist für einen Querschnitt durch die Mundhöhle von jedem geschnittenen Dreieck der auf die Schnittebene projizierte Normalenvektor dargestellt. In diesem Bild zeigen sowohl die Normalen der Dreiecke des Unterkiefergitters als auch die Normalen der Dreiecke des Zungengitters in positive y' -Richtung. Durch die Abtastung der Dreiecksschnitte gibt es also für einige Abtastpunkte des unteren Umrisses zwei „konkurrierende“ Funktionswerte, von denen natürlich jeweils der obere von der Zunge ausgewählt werden muss. Analog muss bei mehreren potenziellen Funktionswerten für einen Abtastpunkt des oberen Umrisses jeweils der unterste ausgewählt werden.

Abbildung 2.18 (c) veranschaulicht, wie aus einem oberen und unteren Umriss der Flächeninhalt des Querschnitts berechnet wird. Dazu müssen lediglich die Trapezflächen dA zwischen jeweils benachbarten Abtastpunkten der beiden Umrisse aufsummiert werden. Außerdem wird von jedem Querschnitt der Umfang berechnet, indem die Länge des oberen und unteren Umrisses addiert wird.

Abb. 2.19 (a) zeigt die Flächenfunktion für den Vokal [a:], die man erhält, wenn die Flächen-

²Die Zunge wurde in diesem Bild nicht geschnitten.

inhalte der Vokaltraktschnitte an allen Abtastpunkten der Mittellinie über der Position auf der Mittellinie abgetragen und linear miteinander verbunden werden. Diese Funktion wird

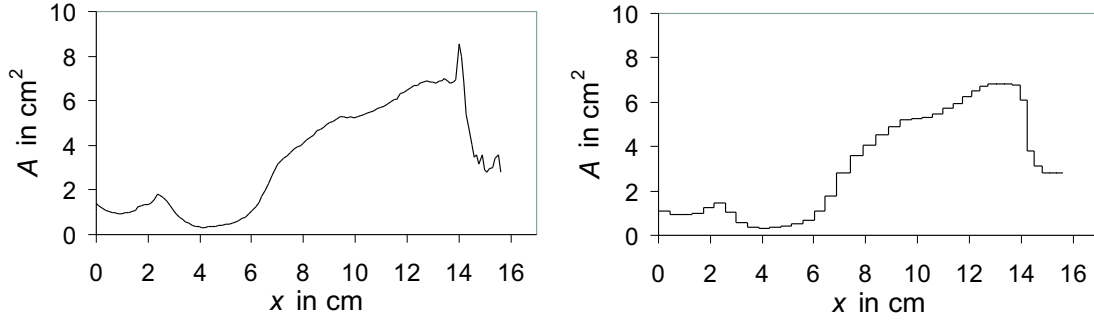


Abbildung 2.19: Stückweise lineare (links) und stückweise konstante (rechts) Flächenfunktion des Vokals [a:].

gegebenenfalls noch durch die Parameter $MA1$, $MA2$ und $MA3$ modifiziert, die einen minimalen Rohrquerschnitt im Bereich hinter der Zungenspitze, direkt an der Zungenspitze, und davor (Schneidezähne, Lippen) angeben. Sollte in den Bereichen der Flächenfunktion, die diesen Abschnitten des Vokaltrakts entsprechen, ein Abtastpunkt unter dem erlaubten Minimalwert liegen, so wird er genau auf diesen Minimalwert gesetzt.

Mit dieser stückweise linearen Flächenfunktion $A(x)$ werden im nächsten Schritt die Flächeninhalte A_i des diskreten Rohrmodells bestimmt. Wenn die Position und Länge der einzelnen Rohrabschnitte jeweils mit x_i und l_i bezeichnet werden, so besteht die Aufgabe darin, die Funktionswerte von $A(x)$ im Intervall $[x_i, x_i + l_i]$ jeweils auf einen einzelnen Wert A_i abzubilden. Die Berechnung von A_i als Mittelwert der Funktion $A(x)$ in dem entsprechenden Intervall hat sich als nicht so günstig herausgestellt, wie ich ursprünglich vermutet habe. Denn dadurch werden sehr kurze enge Passagen oder Verschlüsse (etwa von der Länge eines Rohrabschnitts oder kürzer) „aufgeweitet“, da die umliegenden größeren Querschnitte zum Teil in die Mittelwerte einfließen. Dadurch kann es passieren, dass Verschlüsse im Vokaltrakt im Rohrmodell keine echten Verschlüsse mehr sind.

Eine bessere Abbildung der Funktion $A(x)$ auf die Flächeninhalte A_i ist deshalb die Minimalwertbildung nach der Formel

$$A_i = \min_{x=x_i}^{x_i+l_i} \{A(x)\}.$$

Analog zur Flächenfunktion werden auch die Umfangsfunktion $S(x)$ und die Umfangslängen S_i der einzelnen Rohrabschnitte berechnet. In fast allen Studien zu Rohrmodellen des Vokaltrakts haben alle Rohrabschnitte des Rachen- und Mundraums die gleiche Länge. In der vorliegenden Arbeit wird dagegen der Vokaltrakt nur von der Glottis bis zur velopharyngealen Pforte durch 16 gleichlange Rohrabschnitte repräsentiert und von dort an bis zur Mundöffnung durch 24 Rohrabschnitte mit abnehmender Länge. Dadurch sind die Rohrabschnitte im Bereich der Schneidezähne und Lippen am kürzesten, so dass die Flächenfunktion dort sehr genau repräsentiert wird. Das ist für die Synthese der Frikative von

Vorteil, da ihre spektrale Zusammensetzung sensibel von den Details der Flächenfunktion stromabwärts von der Konstriktion abhängt.

2.4 Nasenraum

Der Nasenraum wird in dieser Arbeit nicht geometrisch modelliert, sondern direkt durch seine stückweise konstante Flächen- und Umfangsfunktion repräsentiert. Die Daten dazu stammen aus einer morphologischen Studie von DANG und HONDA [24] und sind in Tab. 2.3 angegeben. Der Rohrabchnitt $i = 0$ bildet das Rohrende an der velopharyngalen Pforte und der Abschnitt $i = 18$ entspricht den Nasenlöchern. Jeder der 19 Rohrabchnitte hat eine Länge von 0.6 cm, so dass sich eine Gesamtlänge von 11.4 cm ergibt. Entsprechend

i	0	1	2	3	4	5	6	7	8	9
A_i (cm ²)	1.63	2.07	2.72	3.59	4.24	3.26	3.04	3.04	2.72	2.5
S_i (cm)	6.67	7.33	8.00	8.00	8.00	9.33	14.67	22.67	24.67	25.33

i	10	11	12	13	14	15	16	17	18
A_i (cm ²)	2.39	2.39	1.85	0.76	1.41	1.74	1.30	1.74	0.76
S_i (cm)	24.67	26.67	18.67	6.67	12.00	12.00	9.33	8.00	6.67

Tabelle 2.3: Flächen- und Umfangsfunktion des Nasenraums.

einer weiteren Studie von DANG und HONDA [25] wird der akustische Einfluss der Nasennebenhöhlen durch vier Helmholtzresonatoren nachgebildet, die am rechten Ende der Rohrabchnitte mit den Indizes 8, 9, 11 und 12 an den Nasenraum angekoppelt sind. Die morphologischen und akustischen Daten zu den vier Nebenhöhlen sind in Tab. 2.4 zusammengestellt.

	l. Keilbeinh.	r. Keilbeinh.	Kieferhöhle	Stirnhöhle
Volumen V (cm ³)	11.3	6.8	33	6.2
Halslänge l (cm)	0.3	0.3	0.45	1.0
Halsquerschnitt A (cm ²)	0.185	0.185	0.145	0.11
Abstand der Öffnung von den Nasenlöchern (cm)	6.5	6.5	5.0	4.0
Resonanzfrequenz (Hz)	1305	1682	552	749
Bandbreite (Hz)	188	216	108	124
Kopplungsabschnitt	8	9	11	12

Tabelle 2.4: Daten zu den vier Nasennebenhöhlen des Modells von DANG und HONDA [25].

Die Flächenfunktion des Nasenraums ist zusammen mit den Nebenhöhlen grafisch in Abb. 2.1 dargestellt. Die Querschnittsflächen und Umfangslängen der ersten vier Rohrabchnitte des Nasenraums sind in dieser Arbeit variabel und passen sich dem Öffnungsquerschnitt A_{vp} und dem Umfang S_{vp} der velopharyngalen Pforte an. A_0 und S_0 werden direkt auf die Werte A_{vp} und S_{vp} gesetzt, und die Werte für die Rohrabchnitte 1...3 werden

linear zwischen denen der Abschnitte 0 und 4 interpoliert. A_{vp} und S_{vp} sind der Inhalt und der Umfang der Fläche, die von der Rippe 3 der Rachenrückwand und der Rippe 0 des Velums begrenzt wird (siehe Abb. 2.4). Für eine geschlossene Pforte sind beide Rippen identisch, so dass die von ihnen eingeschlossene Fläche Null ist. Dadurch wird der akustische Widerstand am Eingang des Nasenraums unendlich hoch und trennt ihn somit akustisch vom Rachen- und Mundraum ab.

2.5 Stimmlippen

Die Simulation der Stimmlippen basiert in dieser Arbeit auf einem Modell von TITZE [115], bei dem die geometrische Form der zeitvariablen glottalen Öffnungsfläche am Eingang und Ausgang der Stimmritze direkt vorgegeben wird. Nach meiner Erfahrung erzeugt dieses Modell einen subjektiv natürlicheren Stimmklang als das selbstschwingende Zwei-Massen-Modell von ISHIZAKA und FLANAGAN [51] und verwandte Ansätze. Das hier verwendete Modell ermöglicht durch die parametrische Beschreibung der glottalen Öffnungsfläche eine bessere Einflussnahme auf die Stimmmerkmale, als dies bei den selbstschwingenden Modellen möglich ist.

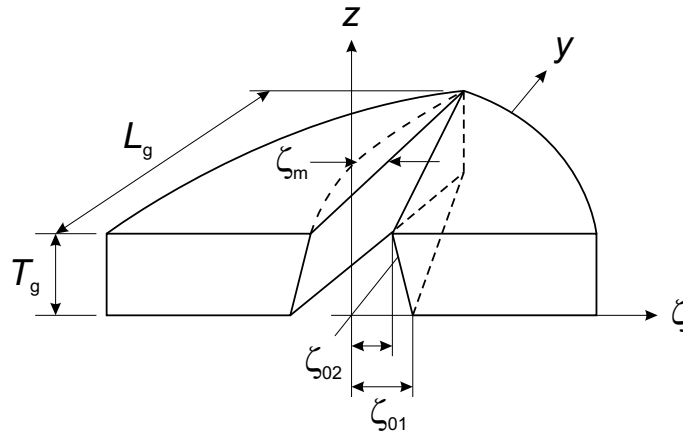


Abbildung 2.20: Modell der Glottis nach TITZE [115].

Abbildung 2.20 zeigt das Glottismodell in seiner präphonatorischen Stellung. Darin sind T_g und L_g die Dicke und Länge des glottalen Spalts, und ζ_{01} und ζ_{02} geben die Auslenkung der Stimmlippen bezüglich der glottalen Mittellinie an ihrem unteren und oberen Rand an der Verbindungsstelle zu den Stellknorpeln an. Von den Stellknorpeln bis zum Winkel der Schildknorpel geht die Weite der präphonatorischen Glottis linear gegen Null. In der $z\zeta$ -Ebene ist die Glottis trapezförmig. Daher lässt sich die statische Auslenkung der rechten Stimmlippe in präphonatorischer Stellung wie folgt formulieren:

$$\zeta_{\text{stat}}(y, z) = \left(\zeta_{01} + \frac{z}{T_g} \cdot (\zeta_{02} - \zeta_{01}) \right) \cdot \left(1 - \frac{y}{L_g} \right).$$

Es wird angenommen, dass die linke Stimmlippe spiegelbildlich dazu aufgebaut ist. Die

Schwingung der Stimmlippen wird simuliert, indem der statischen Konfiguration eine dynamische Auslenkung in der Form

$$\zeta_{\text{dyn}}(y, z, t) = \zeta_{\text{m}} \cdot \sin\left(\pi \cdot \frac{y}{L_g}\right) \cdot \sin\left(\psi(t) - \phi \cdot \frac{z}{T_g}\right)$$

überlagert wird, wobei $\psi(t) = \int_0^t 2 \cdot \pi \cdot F_0(\tau) \cdot d\tau$ der Phasenwinkel der Schwingung am unteren Rand ist und $F_0(t)$ die Grundfrequenz. ζ_{m} ist die Amplitude der Schwingung und ϕ die Phasendifferenz zwischen dem oberen und unteren Rand der Stimmlippen. Mit dem Winkel ϕ kann die Verzögerung eingestellt werden, mit der der obere Teil der Stimmlippen der Bewegung des unteren Teils nacheilt.

Die gesamte Auslenkung von der Mitte der Stimmritze ist somit

$$\zeta(y, z, t) = \max\{0, \zeta_{\text{stat}}(y, z) + \zeta_{\text{dyn}}(y, z, t)\}.$$

Die Kontaktbehandlung zwischen der linken und rechten Stimmlippe erfolgt hier einfach dadurch, dass negative Auslenkungen auf Null gesetzt werden.

Die Länge und Dicke der Stimmlippen sind laut TITZE [117] eine Funktion der Grundfrequenz:

$$\begin{aligned} L_g &= 0.038 \cdot L_{g0}^{1.6} \cdot F_0^{0.5}, \\ T_g &= T_{g0} \cdot L_{g0}/L_g. \end{aligned}$$

Die Referenzabmessungen L_{g0} und T_{g0} sind 1.6 cm und 0.45 cm [117, 119]. Für die Schwingungsamplitude wird von TITZE [118] die Gleichung

$$\zeta_{\text{m}} = \left(\alpha - \beta \cdot \frac{L_g}{L_{g0}}\right) \cdot \sqrt{p_{\text{sub}}} \quad (2.1)$$

angegeben, wobei p_{sub} der subglottale Druck in cm Wassersäule ist. Für die empirisch bestimmten Konstanten α und β gilt $\alpha = 0.0862$ und $\beta = 0.0824$. Mit dieser Gleichung ist die Schwingungsamplitude offensichtlich nur eine Funktion des subglottalen Drucks. Die Änderung der Schwingungsamplitude durch Abduktion, durch Adduktion und mediale Kompression oder durch die Zunahme des supraglottalen Drucks bei einem oralen Verschluss wird in dieser Gleichung nicht berücksichtigt. Da das Verhalten der Stimmlippen in diesen Situationen aber sehr wichtig für die Sprachsynthese ist, habe ich Gl. (2.1) in dieser Arbeit leicht modifiziert:

$$\zeta_{\text{m}} = f \cdot \left(\alpha - \beta \cdot \frac{L_g}{L_{g0}}\right) \cdot \sqrt{p_{\text{trans}}}. \quad (2.2)$$

Diese Gleichung unterscheidet sich in zwei Punkten von Gl. (2.1). Einerseits verwende ich statt des subglottalen Drucks den transglottalen Druck p_{trans} , also die Differenz zwischen dem subglottalen und dem supraglottalen Druck. Wenn durch einen oralen Verschluss, z. B. im Laut [t], der supraglottale Druck ansteigt, so wird p_{trans} automatisch kleiner und die Schwingung der Stimmlippen versiegt. In der Simulation wird p_{trans} als Differenz zwischen dem pulmonalen Druck p_{Lunge} und dem tiefpassgefilterten Druck p_{supra} im ersten Rohrabschnitt oberhalb der Glottis berechnet³.

³Die Knickfrequenz des Tiefpassfilters wurde auf 25 Hz gesetzt.

Der zweite Unterschied zwischen den Gleichungen ist der Faktor f in Gl. (2.2). Dieser Faktor bewirkt eine Abnahme der Schwingungsamplitude, wenn die Stimmritze weit geöffnet ist oder wenn die Stimmlippen zusammengepresst werden. Er wird nach der folgenden Formel berechnet:

$$f = \begin{cases} \cos\left(\frac{\pi}{2} \cdot \frac{\zeta_{02}}{\zeta_{02,\max}}\right) & \text{für } \zeta_{02} > 0 \\ \cos\left(\frac{\pi}{2} \cdot \frac{\zeta_{02}}{\zeta_{02,\min}}\right) & \text{für } \zeta_{02} \leq 0. \end{cases}$$

Darin ist $\zeta_{02,\max} = 3$ mm der Maximalwert und $\zeta_{02,\min} = -0.5$ mm der Minimalwert des Parameters ζ_{02} .

Für die Simulation glottaler Leckströme durch eine permanente Öffnung im Bereich der Stellknorpel wurde das Glottismodell in dieser Arbeit analog zu CRANEN *et al.* [22, 23] um den Bereich zwischen den Stellknorpeln erweitert. Dieser Bereich wird direkt durch seinen Flächeninhalt repräsentiert, der sich aus zwei Teilen zusammensetzt. Ein Teil repräsentiert die passive Öffnung A_{passiv} zwischen den Stellknorpeln aufgrund der Abduktion der Stimmlippen. Hierbei gehe ich von einer rechteckigen Öffnungsfläche mit den Kantenlängen $2 \cdot \zeta_{01}$ und 2 mm aus. Der andere Teil repräsentiert die aktive Veränderung der Öffnungsfläche durch das Kippen oder Drehen der Stellknorpel und wird mit ΔA_{Spalt} bezeichnet. Die Gesamtfläche zwischen den Stellknorpeln ist folglich $A_{\text{Spalt}} = A_{\text{passiv}} + \Delta A_{\text{Spalt}}$. Falls $\zeta_{01} \leq 0$, d. h. wenn die Stimmbänder am hinteren Ende eng zusammengeführt sind, so wird A_{Spalt} als ein *paralleler Spalt* behandelt, der nicht mit der Öffnung zwischen dem membranartigen Teil der Glottis verbunden ist. Falls jedoch $\zeta_{01} > 0$, so wird die Öffnung zwischen dem membranartigen und dem knorpelartigen Teil als ein zusammenhängender Spalt behandelt.

Für die akustische Simulation muss die Geometrie des Glottismodells auf die Parameter der glottalen Rohrabschnitte des Rohrmodells abgebildet werden. Die Flächeninhalte und Umfangslängen des unteren und oberen glottalen Rohrabschnitts werden durch die Auswertung der Stimmlippenform in der $y\zeta$ -Ebene an den Stellen $z = 0$ und $z = T_g$ berechnet. Die Summe der Länge der beiden Rohrabschnitte ist T_g und ihr Längenverhältnis wird auf 4:1 gesetzt, d. h. der untere Abschnitt ist viermal so lang wie der obere. Wenn die Öffnung zwischen den Stellknorpeln mit der Öffnung im membranartigen Teil verbunden ist, so wird A_{Spalt} zu den Flächen der beiden Abschnitte dazuaddiert. Ansonsten werden die Flächen der beiden Rohrabschnitte, die den parallelen Spalt repräsentieren, auf A_{Spalt} gesetzt.

Bezeichnung	Wertebereich	Einheit	Bestimmung
F_0	[50, 400]	Hz	Grundfrequenz
p_{Lunge}	[0, 1200]	Pa	Pulmonaler Druck
ζ_{01}	[-0.5, 3]	mm	Abduktion am unteren Rand
ζ_{02}	[-0.5, 3]	mm	Abduktion am oberen Rand
ΔA_{Spalt}	[-5, 5]	mm ²	Aktive Öffnung im Bereich der Stellknorpel
ϕ	[0, π]	rad	Phasendiff. zw. oberem und unterem Rand

Tabelle 2.5: Parameter des Glottismodells.

In Tab. 2.5 sind alle Parameter zusammengefasst, die einen Einfluss auf die zeitvariable Form der Stimmlippen bzw. die glottale Anregung haben. Zusätzlich sind ihre jeweiligen

Wertebereiche angegeben. Die Parameter in Tab. 2.5 werden in dieser Arbeit zusammen als „Anregungsparameter“ bezeichnet.

Bei Syntheseexperimenten mit dem Glottismodell hat sich herausgestellt, dass die synthetische Stimme am natürlichsten klingt, wenn ζ_{01} und ζ_{02} gleich groß sind. Abhängig von der Wertebelegung von ζ_{01} und ζ_{02} kann eine gepresste, modale oder behauchte Stimme simuliert werden. Die drei Stimmregister liegen dabei auf einem Kontinuum zwischen kleinen und großen ζ -Werten. Für die Simulation modaler Phonation ist der Wert $\zeta_{01} = \zeta_{02} = 0.3$ mm geeignet. Die Phasendifferenz ϕ wurde in allen Simulationen auf $\phi = 2 \cdot \pi \cdot 0.14$ rad = 50.4° gesetzt. Für eine höhere Natürlichkeit der synthetischen Stimme werden der vorgegebenen F_0 -Kontur zusätzlich kleine quasi-zufällige Schwankungen $\Delta F_0(t)$ („Jitter“) nach einer Formel von KLATT & KLATT [60] überlagert.

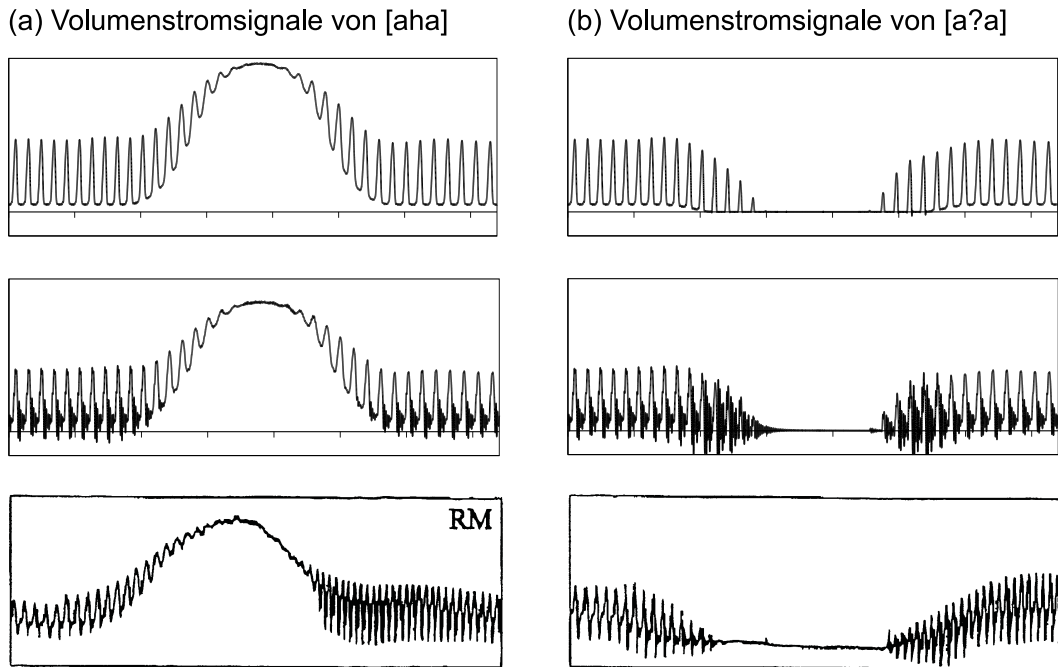


Abbildung 2.21: Volumenstrom-Zeitfunktionen für die Äußerungen [aha] (links) und [a?a] (rechts). Die oberen Kurven zeigen den simulierten glottalen Volumenstrom, die mittleren Kurven den simulierten Volumenstrom durch die Mundöffnung und die unteren Kurven entsprechende gemessene Signale für den Volumenstrom durch die Mundöffnung (von LÖFQVIST *et al.* [71]). Die simulierten Signale haben jeweils eine Dauer von 370 ms.

Das Verhalten des Stimmlippenmodells bei Abduktion und Adduktion ist in Abb. 2.21 anhand von Volumenstromkurven der Äußerungen [aha] und [a?a] dargestellt. Die oberen und mittleren Kurven zeigen jeweils den simulierten Volumenstrom durch die Glottis und durch die Mundöffnung. Für beide Äußerungen ist die realistische Unterbindung der Stimmlippenschwingung in der Mitte der Konsonanten erkennbar. Zum Vergleich mit den simulierten Volumenstromkurven in der Mitte zeigen die unteren Kurven Volumenstromsignale, die von LÖFQVIST *et al.* [71] an der Mundöffnung von einem Sprecher *gemessen*

wurden.

2.6 Subglottale Luftwege

In diesem Abschnitt stelle ich das Modell des subglottalen Systems vor. Zunächst folgt eine Zusammenfassung über den Aufbau und anschließend die Vorstellung des verwendeten Modells. Als subglottales System bzw. als untere Luftwege wird der Teil des Sprechapparats bezeichnet, der sich unterhalb der Stimmlippen befindet. Den oberen Teil des Systems bildet die Luftröhre (Trachea). Sie beginnt direkt unterhalb des Kehlkopfes, ist etwa 12 cm lang und hat einen Durchmesser von 1.3 bis 2 cm. Die Luftröhre endet in der Brusthöhle mit ihrer Teilung in die zwei Hauptbronchien, die zur linken und rechten Lunge führen. Der rechte Hauptbronchus teilt sich in drei Äste, die zu den drei Lungenlappen der rechten Körperhälfte führen. Der linke Hauptbronchus teilt sich dagegen in zwei Äste für die zwei linken Lungenlappen. Die Verzweigung der Luftwege setzt sich über mehrere Stufen fort und ist schematisch in Abb. 2.22 (a) dargestellt. Die Astabschnitte der letzten Teilungsstufen sind traubenförmig von den Lungenbläschen (Alveolen) umgeben, die für den Gasaustausch sorgen.

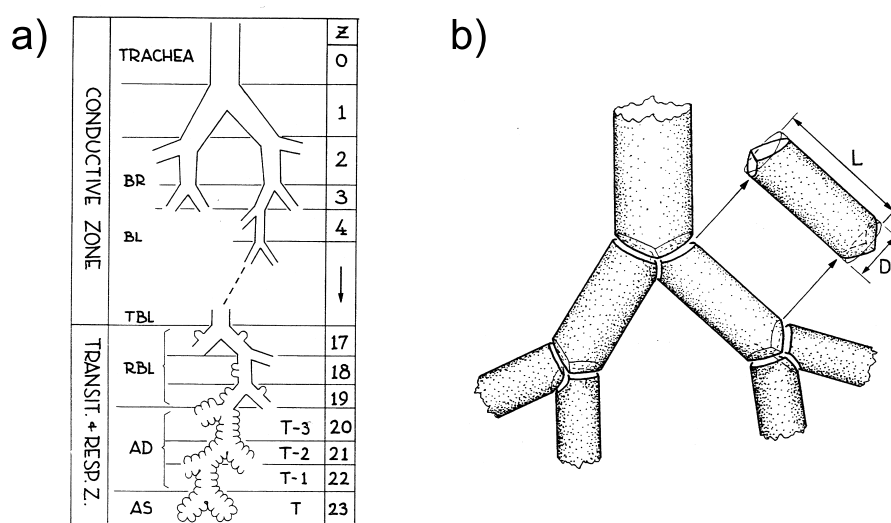


Abbildung 2.22: a) Verzweigung der unteren Luftwege: Z = Teilungsstufe, BR = Bronchien, BL = Bronchiolen, TBL = Terminalbronchiolen, RBL = Bronchioli respiratorii, AD = Alveolargänge, AS = Alveolarsäcke. b) Vereinfachte Darstellung der Luftwegsabschnitte. Quelle: [127].

Anhand der Teilungsstufen lassen sich die subglottalen Luftwege bezüglich ihrer strukturellen Eigenschaften in drei Zonen einteilen [127]. Die Stufen 0–16 mit der Luftröhre, den Bronchien, den Bronchiolen und den Terminalbronchiolen werden als Leitungszone (conductive zone) bezeichnet. Die drei darauf folgenden Stufen 17–19 bilden die Übergangszone (transitory zone). Die Astabschnitte dieser Zone heißen Bronchioli respiratorii und sind bereits teilweise mit Lungenbläschen besetzt. Die letzte Zone mit den Teilungsstufen 20–23 ist die Respirationszone (respiratory zone). Ihre Astabschnitte sind vollständig von Alveolen

umgeben.

Auf der Grundlage einer umfassenden Ausmessung der unteren Luftwege schlägt WEIBEL [127] ein vereinfachtes Modell der Lunge vor, das auch in der vorliegenden Arbeit verwendet wird. Es entspricht der Lunge eines Erwachsenen durchschnittlicher Größe bei einem Grad der Inflation von 75%. WEIBEL geht bei der Konstruktion des Modells davon aus, dass sich jeder Astabschnitt einer Teilungsstufe am Anfang der nächsten Stufe in exakt zwei gleiche Unteräste verzweigt. Diese Situation ist in Abb. 2.22 (b) dargestellt. Die Gesamtanzahl der Teilrohre in jeder Stufe z ist demnach $n_z = 2^z$. Die Astabschnitte werden als Hohlzylinder modelliert, von denen alle innerhalb derselben Stufe die gleichen Abmessungen besitzen. Diesem Modell entsprechend lässt sich jede Teilungsstufe z als ein Rohrbündel mit 2^z gleichförmigen, parallel verlaufenden Zylinderrohrsegmenten auffassen. Eine Stufe z ist durch die Länge l_z , den Gesamtquerschnitt $A_{\text{ges},z}$ und die Anzahl n_z der parallelen Rohrsegmente gekennzeichnet. Im Rahmen dieser Arbeit werden die Rohrbündel der 24 Teilungsstufen in insgesamt 34 *Rohrverbundabschnitte* (analog zu den „einfachen“ Rohrab-schnitten des Vokaltrakts) überführt. Dabei werden die Rohrbündel der ersten drei relativ langen Teilungsstufen (Luftröhre, Hauptbronchien, Nebenbronchien) durch jeweils mehrere Rohrverbundabschnitte repräsentiert, und die sehr kurzen Rohrbündel der Übergangs- und Respirationszone ($z=17-23$) werden zu einem einzigen gleichwertigen Rohrverbundabschnitt zusammengefasst. Die Rohrbündel der Stufen 3–16 entsprechen jeweils genau einem Rohrverbundabschnitt. Durch diese Maßnahmen liegt die Länge aller Abschnitte zwischen 0.165 und 1 cm, was für die numerische Simulation des Schallfeldes günstiger ist, als sehr kurze oder sehr lange Abschnitte. Die Abmessungen der Rohrverbundabschnitte sind in Tab. 2.6 angegeben. Der Index i kennzeichnet die Reihenfolge der Abschnitte und beginnt mit dem gemeinsamen Abschnitt für die Respirations- und Transitionszone. z ist jeweils die Teilungsstufe und n_i die Anzahl der parallelen Einzelrohre für den Rohrverbundabschnitt i . Der Querschnitt eines Einzelrohres ist jeweils $A_{\text{ges},i}/n_i$.

Abbildung 2.1 zeigt den Verlauf der Flächenfunktion $A_{\text{ges},i}$. Die Gesamtlänge der Leitungszone in dem Modell beträgt 26.6 cm und fasst ein Volumen von ca. 150 ml. Die Übergangszone und die Respirationszone sind zusammen nur 0.619 cm lang und fassen ohne die Alveolen ca. 1500 ml. Die Alveolen haben ein Volumen von 3150 ml und stellen damit ca. 66% des gesamten Lungenvolumens von 4800 ml. Das beträchtliche Volumen der Alveolen ist in Tab. 2.6 nicht enthalten und entspricht einer so großen akustischen Kapazität, dass sie bei der akustischen Simulation wie ein offenes Rohrende (aber ohne eine Strahlungsimpedanz) behandelt werden kann (siehe Abschnitt 3.5.3).

2.7 Zusammenfassung und Diskussion

In diesem Kapitel wurde die Modellierung des Vokaltrakts, des Nasenraums, der Stimmlippen und der subglottalen Luftwege vorgestellt. Für den Vokaltrakt wurde ein dreidimensionales geometrisches Modell entwickelt, das die Artikulatoren und Sprechtraktwände mit Hilfe mehrerer Dreiecksnetze repräsentiert. Die Form des Vokaltrakts wird durch 23 Parameter bestimmt. Die Auswahl der Parameter orientiert sich zum Teil an dem 2D-Modell von MERMELSTEIN [77], wurde aber zusätzlich um Parameter zur Berücksichtigung der dritten Dimension ergänzt. In Verbindung mit dem Vokaltraktsmodell habe ich spezielle Methoden für die Kontaktbehandlung zwischen den Artikulatoren und den Sprechtrakt-

i	l_i in cm	$A_{\text{ges},i}$ in cm ²	n_i	z	i	l_i in cm	$A_{\text{ges},i}$ in cm ²	n_i	z
0	0.619	2423.0	131072	17–23	17	0.76	2.33	2	1
1	0.165	180.0	65536	16	18	1.0	2.33	2	1
2	0.2	113.0	32768	15	19	1.0	2.33	2	1
3	0.23	69.4	16384	14	20	1.0	2.33	2	1
4	0.27	44.5	8192	13	21	1.0	2.33	2	1
5	0.33	28.8	4096	12	22	1.0	2.54	1	0
6	0.39	19.6	2048	11	23	1.0	2.54	1	0
7	0.46	13.4	1024	10	24	1.0	2.54	1	0
8	0.54	9.56	512	9	25	1.0	2.54	1	0
9	0.64	6.95	256	8	26	1.0	2.54	1	0
10	0.76	5.1	128	7	27	1.0	2.54	1	0
11	0.9	3.96	64	6	28	1.0	2.54	1	0
12	1.07	3.11	32	5	29	1.0	2.54	1	0
13	1.27	2.48	16	4	30	1.0	2.54	1	0
14	0.76	2.0	8	3	31	1.0	2.54	1	0
15	0.9	2.13	4	2	32	1.0	2.54	1	0
16	1.0	2.13	4	2	33	1.0	2.54	1	0

Tabelle 2.6: Abmessungen der Rohrverbundabschnitte der subglottalen Luftwege. l_i , $A_{\text{ges},i}$ und n_i sind die Länge, der Gesamtquerschnitt und die Anzahl paralleler Rohrsegmente für den Rohrverbundabschnitt i , und z ist die Verzweigungsstufe des Bronchialbaums, zu dem der Abschnitt gehört.

wänden sowie für die Berechnung der Flächen- und Umfangsfunktion entwickelt. Das hier vorgestellte Vokaltraktmodell eignet sich nicht nur für die akustische Synthese, sondern auch für die visuelle Darstellung der Spracherzeugung.

Die Simulation der Stimmlippen basiert auf einem Modell von TITZE [115] und wurde um ein realitätsnahes Verhalten bei der Abduktion und der Adduktion erweitert, das für die Sprachsynthese sehr wichtig ist. Zusätzlich wurde das Modell um den Bereich zwischen den Stellknorpeln erweitert, um glottale Leckströme simulieren zu können. Die Modellierung des subglottalen Systems und des Nasenraums orientiert sich an früheren Arbeit von DANG *et al.* [24, 25] und WEIBEL [127]. Alle hier vorgestellten Teilmodelle wurden zu einem gemeinsamen verzweigten Rohrmodell des Sprechapparats integriert, das die Grundlage für die aerodynamisch-akustische Simulation bildet.

Kapitel 3

Aerodynamisch-akustische Simulation

3.1 Grundlagen

Dieses Kapitel behandelt die Berechnung des vom Vokaltrakt abgestrahlten Schalldrucksignals. Die Berechnung erfolgt durch eine Simulation der aerodynamisch-akustischen Luftbewegung innerhalb des Vokaltrakts. Grundsätzlich basiert die Simulation der Bewegung von Fluiden (Flüssigkeiten oder Gase) auf den allgemeinen Grundgleichungen der Strömungsmechanik. Diese drücken die drei physikalischen Gesetze von der Erhaltung des Impulses (Impulssatz), der Erhaltung der Masse (Kontinuitätsgleichung) und der Erhaltung der Energie (Energiesatz) aus. In Abhängigkeit vom Anwendungsfall können diese Grundgleichungen durch bestimmte Annahmen vereinfacht werden. Die wichtigste Vereinfachung, die häufig im Rahmen der artikulatorischen Sprachsynthese und auch in der vorliegenden Arbeit gemacht wird, ist die eindimensionale Behandlung der Strömung und des Schallfeldes im Vokaltrakt. Es wird also angenommen, dass die Fluidbewegung nur entlang einer einzigen Ortskoordinate erfolgt.

Bei der eindimensionalen Simulation genügt zur Beschreibung der Rohrgeometrie die Flächenfunktion $A(x)$ und die Rohrumfangsfunktion $S(x)$. Diese Funktionen sind die Schnittstelle zwischen der Modellierung des Sprechapparats im vorangegangenen Kapitel und der aerodynamisch-akustischen Simulation in diesem Kapitel.

Obwohl der Vokaltrakt beim Sprechen in ständiger Bewegung ist, können die Flächen- und Umfangsfunktion für die Herleitung der Simulationsgleichungen als zeitlich konstant angenommen werden. Das ist dadurch gerechtfertigt, dass die Änderung der Rohrgeometrie im Vergleich zu den Änderungen der Schwankungsgrößen des Schallfeldes in der Regel sehr langsam erfolgt. Durch die zeitunabhängige Behandlung der Geometriefunktionen vereinfachen sich die Grundgleichungen deutlich.

Die physikalischen Größen, die für die Fluidbewegung eine Rolle spielen, sind der absolute Druck $p_a(x, t)$, die absolute Dichte $\varrho_a(x, t)$, die Strömungsgeschwindigkeit $v(x, t)$ und der Volumenstrom $u(x, t) = v(x, t) \cdot A(x)$. Die im Vokaltrakt vorkommenden Abweichungen der Druck- und Dichtewerte vom atmosphärischen Druck p_0 und der mittleren Dichte ϱ_0 sind sehr klein. Der atmosphärische Druck beträgt beispielsweise ca. 100 kPa, während der subglottale Überdruck bei lauter Phonation nur etwa 1 kPa beträgt. Diese geringen

Abweichungen ermöglichen in der Akustik die Linearisierung der Grundgleichungen, indem der absolute Druck und die absolute Dichte jeweils als Summe eines großen statischen Anteils und eines kleinen Schwankungsanteils behandelt werden:

$$\varrho_a(x, t) = \varrho_0 + \varrho(x, t) \quad (3.1)$$

$$p_a(x, t) = p_0 + p(x, t). \quad (3.2)$$

$p(x, t)$ und $\varrho(x, t)$ werden als Schalldruck und Schalldichte bezeichnet. Das komplexe Verhältnis vom Schalldruck zum Schwankungsanteil des Volumenstroms (im Sinne der komplexen Zeigerarithmetik) ist die akustische Impedanz Z . Eine rein reelle Impedanz wird auch als akustischer Widerstand R bezeichnet.

Der Rest des Kapitels ist wie folgt gegliedert. Zunächst werde ich in Abschnitt 3.2 einen kurzen Überblick über die Simulationstechniken anderer Studien geben. In Abschnitt 3.3 werde ich, ausgehend von elementaren physikalischen Gesetzen, die Grundgleichungen für die aerodynamisch-akustische Simulation in dieser Arbeit ableiten. Die Herleitung erfolgt für eine reibungsfreie, instationäre, kompressible Fluidbewegung durch ein Rohr mit elastischen Wänden. In Abschnitt 3.4 werden die im Vokaltrakt auftretenden Energieverluste diskutiert und in Abschnitt 3.5 werden die Grundgleichungen unter Berücksichtigung der Verluste integriert und örtlich diskretisiert. Daraus entsteht eine Repräsentation des Sprechapparats in Form eines analogen elektrischen Netzwerks, genauer gesagt, in Form einer verzweigten inhomogenen Übertragungsleitung mit konzentrierten Elementen. Diese schematische Repräsentation dient in den Abschnitten 3.6 und 3.7 als Grundlage für die numerische Simulation der Fluidbewegung im Frequenz- und im Zeitbereich. In Abschnitt 3.8 beschreibe ich Versuche, mit deren Ergebnissen ich ein Modell für die Vorhersage der Merkmale von Rauschquellen an supraglottalen Konstriktionen ableite. Die Ergebnisse dieses Kapitels werden in Abschnitt 3.9 zusammengefasst und diskutiert.

3.2 Literaturüberblick

Ein sehr schnelles Verfahren für die akustische Simulation eines eindimensionalen ortsdiskreten Rohrmodells wurde 1962 von KELLY und LOCHBAUM [54] vorgestellt. Das Verfahren wird heute als Wellendigitalfilter-Methode bezeichnet und in einer Vielzahl von Studien verwendet, z. B. von KOB [61], KRÖGER [66], MEYER *et al.* [78], RUBIN *et al.* [97] und SCHNELL [99]. Das Verfahren beruht auf der Berechnung von vor- und zurücklaufenden Wellen in den Rohrabschnitten, die mit Hilfe von Streugleichungen an den Querschnittsprüngen ermittelt werden. Die Länge Δx der einzelnen Rohrabschnitte ist dabei durch die Beziehung $\Delta x = c/(2 \cdot f_A)$ fest mit der Abtastrate f_A und der Schallgeschwindigkeit c verknüpft. Dadurch kann die Gesamtlänge des Rohres jeweils nur ein ganzzahliges Vielfaches von Δx betragen. Für den Einsatz in der Sprachsynthese ist dies ein Nachteil, da sich die Vokaltraktlänge beim Sprechen kontinuierlich ändert.

Eine andere Möglichkeit der akustischen Simulation beruht auf der Analogie des Vokaltrakts zu einem elektrischen Netzwerk, und zwar zu einer inhomogenen Übertragungsleitung mit konzentrierten Elementen [51, 40, 73]. Hierbei werden die Netzwerkgleichungen zeitdiskret approximiert und damit die Zweigströme und Knotenspannungen im Netzwerk berechnet. Das Verfahren ist rechenintensiver als die Wellendigitalfilter-Methode, lässt aber

variable Längen für die Rohrabschnitte zu. Dieser Ansatz wird daher auch in dieser Arbeit verfolgt.

Zeitdiskrete Simulationsverfahren wie die Netzwerksimulation oder Wellendigitalfilter haben Schwierigkeiten damit, frequenzabhängige Verluste (z. B. durch Reibung oder Wärmeleitung) im Vokaltrakt zu berücksichtigen. Daher haben SONDHI und SCHROETER [106] 1987 ein hybrides Verfahren vorgestellt, das die Übertragungsfunktion des Vokaltrakts im Frequenzbereich berechnet (incl. aller frequenzabhängigen Verluste) und den Volumenstrom durch die Stimmritze im Zeitbereich. Der abgestrahlte Schalldruck wird durch die digitale Faltung des glottalen Volumenstroms mit der Impulsantwort des Vokaltrakts berechnet, die sich durch die inverse Fourier-Transformation aus der Übertragungsfunktion ergibt. Ein vergleichbarer Ansatz wurde kurze Zeit vorher auch von ALLEN *et al.* [3] benutzt.

Neben den erwähnten eindimensionalen Ansätzen gibt es auch dreidimensionale Simulationsverfahren für das Schallfeld im Vokaltrakt, z. B. von EL-MASRI *et al.* [28] und MATSUZAKI *et al.* [75]. Aufgrund ihres hohen Rechenaufwandes sind sie praktisch nicht für die Sprachsynthese geeignet. Jedoch kann mit ihrer Hilfe der Fehler abgeschätzt werden, der mit den eindimensionalen Simulationsverfahren gemacht wird.

Konkrete Umsetzungen einzelner Simulationsverfahren unterscheiden sich unabhängig vom gewählten Ansatz in der Art und Weise, wie und ob akustische und strömungsmechanische Energieverluste berücksichtigt werden, wie die Wechselwirkung zwischen dem Fluid und den Vokaltraktwänden behandelt wird, mit welchen Modellen die primäre und sekundäre Sprechtraktanregung erfolgt und welche numerischen Techniken für die Simulation eingesetzt werden. Eine diesbezügliche Einordnung der in diesem Kapitel vorgestellten Simulationmethoden erfolgt in den jeweiligen Unterabschnitten.

3.3 Grundgleichungen

3.3.1 Impulssatz

Der Impulssatz in allgemeiner Form lautet: „Die zeitliche Änderung des Impulses eines Körpers oder Fluidteilchens ist gleich der Summe der an ihm angreifenden Kräfte.“ Als Fluidteilchen betrachten wir eine infinitesimal kleine Stoffmenge der Luft mit der konstanten Masse dm , welche sich zum Zeitpunkt der Betrachtung t an der Stelle x befindet und sich mit der Geschwindigkeit $v(x, t)$ bewegt. Der Inhalt des Impulssatzes wird in differenzieller Form durch die folgende Gleichung ausgedrückt:

$$dm \cdot \frac{dv}{dt} = dF_{\text{sum}}, \quad (3.3)$$

worin dF_{sum} die Summe der angreifenden Kräfte ist. Das Fluidteilchen hat die Form einer dünnen Scheibe mit dem mittleren Querschnitt $A(x)$ und der Dicke dx . Das Volumen des Fluidteilchens beträgt somit $dV = A \cdot dx$ und seine Masse $dm = \varrho_a \cdot A \cdot dx$. In Abb. 3.1 sind das Fluidteilchen und die angreifenden Kräfte schematisch dargestellt.

Bei den angreifenden Kräften handelt es sich um die Druckkräfte F_1 und F_2 auf die linke und rechte Stirnfläche der Scheibe sowie die von der Mantelfläche ausgeübte Druckkraft dF_M . Die Gravitationskraft kann gegenüber diesen Kräften vernachlässigt werden.

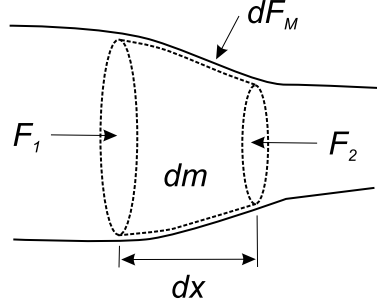


Abbildung 3.1: Kräfte, die an einem Fluidelement der Masse dm angreifen.

Die Druckkräfte auf die Stirnflächen sind

$$F_1 = p_a \cdot A \quad \text{und} \quad F_2 = - \left[p_a \cdot A + \frac{\partial(p_a \cdot A)}{\partial x} \cdot dx \right].$$

Das negative Vorzeichen vor F_2 zeigt an, dass diese Kraft entgegengesetzt zu F_1 gerichtet ist. Für die Mantelfläche des Fluidteilchens wird angenommen, dass auf ihr der mittlere Druck

$$p_M = \frac{1}{2} \left[p_a + \left(p_a + \frac{\partial p_a}{\partial x} \cdot dx \right) \right] = p_a + \frac{1}{2} \cdot \frac{\partial p_a}{\partial x} \cdot dx$$

lastet. Die x -Komponente der Druckkraft ist

$$dF_M = \left[p_a + \frac{1}{2} \cdot \frac{\partial p_a}{\partial x} \cdot dx \right] \left[A - \left(A + \frac{\partial A}{\partial x} \cdot dx \right) \right] \approx p_a \cdot \frac{\partial A}{\partial x} \cdot dx$$

Die rechte Seite der Gleichung erhält man durch die Vernachlässigung des Anteils, der klein von höherer Ordnung ist. Die Summe der angreifenden Kräfte beträgt

$$dF_{\text{sum}} = F_1 + F_2 + dF_M \approx -A \cdot \frac{\partial p_a}{\partial x} \cdot dx.$$

Wenn dF_{sum} und dm in Gl. (3.3) eingesetzt werden, so erhält man

$$-A \cdot \frac{\partial p_a}{\partial x} \cdot dx = \varrho_a \cdot A \cdot dx \cdot \frac{dv}{dt},$$

und nach dem Kürzen durch $(A \cdot dx)$:

$$-\frac{\partial p_a}{\partial x} = \varrho_a \cdot \frac{dv}{dt}. \quad (3.4)$$

Wenn wir das totale Differenzial dv der Geschwindigkeit als $dv = (\partial v / \partial x) \cdot dx + (\partial v / \partial t) \cdot dt$ ausschreiben und durch dt dividieren, so erhalten wir für den Differenzialquotienten

$$\frac{dv}{dt} = v \cdot \frac{\partial v}{\partial x} + \frac{\partial v}{\partial t}. \quad (3.5)$$

Wenn (3.5) in (3.4) eingesetzt wird, so erhält man die sog. *Euler'sche Bewegungsgleichung*:

$$-\frac{\partial p_a}{\partial x} = \varrho_a \cdot v \cdot \frac{\partial v}{\partial x} + \varrho_a \cdot \frac{\partial v}{\partial t}. \quad (3.6)$$

Wenn wir darin den absoluten Druck p_a und die absolute Dichte ϱ_a jeweils als Summe des statischen Anteils und des Schwankungsanteils schreiben und weiterhin die Ersetzung $v \cdot \partial v / \partial x = (1/2) \cdot \partial v^2 / \partial x$ vornehmen, so erhalten wir

$$-\frac{\partial(p_0 + p)}{\partial x} = (\varrho_0 + \varrho) \cdot \frac{1}{2} \cdot \frac{\partial v^2}{\partial x} + (\varrho_0 + \varrho) \cdot \frac{\partial v}{\partial t}. \quad (3.7)$$

Die Bewegungsgleichung für das *lineare Schallfeld* erhält man, wenn alle Produkte kleiner Größen (p , ϱ und v), d. h. alle Terme zweiter oder höherer Ordnung, vernachlässigt werden:

$$-\frac{\partial p}{\partial x} = \varrho_0 \cdot \frac{\partial v}{\partial t}. \quad (3.8)$$

Bei der Fluidbewegung im Vokaltrakt sind die Geschwindigkeiten in der Regel jedoch nicht klein, da die Luftmoleküle nicht nur akustische Schwingungsbewegungen um ihre Ruhepositionen ausführen, sondern im Zuge der Ausatmung gleichzeitig aus der Lunge durch den Vokaltrakt nach außen strömen. Die Strömungsgeschwindigkeit ist umso größer, je kleiner der betrachtete Rohrquerschnitt ist. Bei der Erzeugung der Frikative kann die Strömungsgeschwindigkeit durch die Glottis und die supraglottale Konstriktion z. B. Mach 0.1 erreichen, also 1/10 der Schallgeschwindigkeit [104]. Daher darf im allgemeinen Fall zumindest der Term $\varrho_0 \cdot (\partial v^2 / \partial x)$ in Gl. (3.7), in dem die Geschwindigkeit in zweiter Ordnung vorkommt, nicht vernachlässigt werden. Der Ausgangspunkt für die folgenden Betrachtungen ist somit die Bewegungsgleichung

$$-\frac{\partial p}{\partial x} = \varrho_0 \cdot \frac{\partial v}{\partial t} + \frac{\varrho_0}{2} \cdot \frac{\partial v^2}{\partial x}. \quad (3.9)$$

Ersetzt man weiterhin alle Vorkommen von $v(x, t)$ durch $u(x, t)/A(x)$, so lautet die Gleichung

$$-\frac{\partial p}{\partial x} = \frac{\varrho_0}{A} \cdot \frac{\partial u}{\partial t} + \frac{\varrho_0}{2} \cdot \frac{\partial}{\partial x} \left(\frac{u^2}{A^2} \right). \quad (3.10)$$

Im Summanden mit der zeitlichen Ableitung konnte A^{-1} vor den Differenzialquotienten gezogen werden, da A nur von x aber nicht von t abhängt.

3.3.2 Kontinuitätsgleichung

Die Kontinuitätsgleichung drückt das Prinzip der Erhaltung der Masse innerhalb eines ortsfesten Kontrollvolumens aus. In unserem Fall ist das Kontrollvolumen eine infinitesimal dünne Scheibe des Rohres mit der Breite dx an der *festen* Position x . Eine Änderung der Fluidmasse innerhalb des Kontrollraums ist nur dann möglich, wenn durch seine linke oder rechte Stirnseite Massenströme ein- oder austreten. Die Massenbilanzgleichung lautet

$$\frac{\partial(dm)}{\partial t} = \dot{m}_1 - \dot{m}_2, \quad (3.11)$$

wobei dm die infinitesimale Masse im Kontrollraum ist und $\dot{m}_1$ und $\dot{m}_2$ die durch die linke bzw. rechte Seitenfläche tretenden Massenströme sind. In Abb. 3.2 ist der beschriebene Sachverhalt verdeutlicht. Die Massenströme sind

$$\dot{m}_1 = \varrho_a \cdot A \cdot v \quad \text{und} \quad \dot{m}_2 = \varrho_a \cdot A \cdot v + \frac{\partial(\varrho_a \cdot A \cdot v)}{\partial x} \cdot dx$$

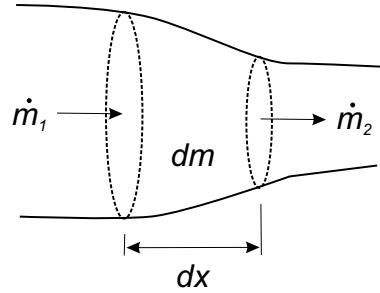


Abbildung 3.2: Massenströme in und aus einem ortsfesten Kontrollraum.

und die Fluidmasse im Kontrollraum ist $dm = \varrho_a \cdot A \cdot dx$. Wenn $\dot{m}_1$, $\dot{m}_2$ und dm in Gl. (3.11) eingesetzt werden, so ergibt sich

$$\frac{\partial(\varrho_a \cdot A \cdot dx)}{\partial t} = -\frac{\partial(\varrho_a \cdot A \cdot v)}{\partial x} \cdot dx.$$

Da die Breite dx des Kontrollvolumens konstant ist, kann sie auf beiden Seiten gekürzt werden:

$$\frac{\partial(\varrho_a \cdot A)}{\partial t} = -\frac{\partial(\varrho_a \cdot A \cdot v)}{\partial x}. \quad (3.12)$$

An dieser Stelle muss berücksichtigt werden, dass die Rohrwände in der Simulation in gewissen Grenzen elastisch auf die auftretenden Druckverhältnisse reagieren sollen. Wenn sich bei Überdruck das Rohr entlang seines Umfangs ausdehnt, so vergrößert sich sein Querschnitt geringfügig, und umgekehrt verringert er sich beim Zusammenziehen durch einen Unterdruck. Wenn die Wandauslenkung $y(x, t)$ senkrecht zur Rohrwand klein gegenüber dem Rohrdurchmesser ist, dann kann die daraus resultierende Querschnittsänderung $\Delta A(x, t)$ als $y(x, t) \cdot S(x)$ geschrieben werden, wobei $S(x)$ der Rohrumfang ist. Unter Berücksichtigung von ΔA wird Gl. (3.12) zu

$$\frac{\partial(\varrho_a \cdot A)}{\partial t} + \frac{\partial(\varrho_a \cdot S \cdot y)}{\partial t} = -\frac{\partial(\varrho_a \cdot A \cdot v)}{\partial x}$$

bzw. zu

$$A \cdot \frac{\partial \varrho_a}{\partial t} + S \cdot \frac{\partial(\varrho_a \cdot y)}{\partial t} = -\frac{\partial(\varrho_a \cdot A \cdot v)}{\partial x}, \quad (3.13)$$

wenn man die zeitunabhängigen Größen $A(x)$ und $S(x)$ auf der linken Seite der Gleichung vor die Differenzialquotienten zieht. Wenn weiterhin $A \cdot v$ durch den Volumenstrom u ersetzt wird und die absolute Dichte als Summe des statischen und dynamischen Anteils geschrieben wird, so ergibt sich

$$A \cdot \frac{\partial(\varrho_0 + \varrho)}{\partial t} + S \cdot \frac{\partial[(\varrho_0 + \varrho) \cdot y]}{\partial t} = -\frac{\partial[(\varrho_0 + \varrho) \cdot u]}{\partial x}.$$

Wenn auch hier wieder die Terme, die klein von höherer Ordnung sind, vernachlässigt werden, so erhält man

$$A \cdot \frac{\partial \varrho}{\partial t} + \varrho_0 \cdot S \cdot \frac{\partial y}{\partial t} = -\varrho_0 \cdot \frac{\partial u}{\partial x},$$

bzw. umgestellt

$$\frac{\partial \varrho}{\partial t} = \frac{\varrho_0}{A} \cdot \left[-\frac{\partial u}{\partial x} - S \cdot \frac{\partial y}{\partial t} \right]. \quad (3.14)$$

3.3.3 Zustandsgleichung

Wenn man eine konstante Menge Luft mit der Masse m komprimiert, so erhöhen sich Druck und Temperatur. Das Umgekehrte passiert bei der Dekompression. Wenn die Kompression und Dekompression wie bei einer Schallwelle genügend schnell ablaufen, so kann praktisch keine Wärme abfließen, weil Luft ein guter Isolator ist. Für diese sog. adiabatische Zustandsänderung gilt die Poisson-Gleichung (Adiabatengleichung)

$$p_a \cdot V^\gamma = \text{const.}, \quad (3.15)$$

wobei V das von der betrachteten Luftmenge eingenommene Volumen und γ der Adiabatenexponent ist. Der Adiabatenexponent von Luft beträgt $\gamma = 1.4$. Das totale Differenzial von Gl. (3.15) lautet

$$\frac{dp_a}{p_a} = -\gamma \cdot \frac{dV}{V}. \quad (3.16)$$

Weiterhin lässt sich die Beziehung $\rho_a \cdot V = m = \text{const.}$ zwischen Masse, Dichte und Volumen der betrachteten Luftmenge in differentieller Form wie folgt schreiben:

$$\frac{dV}{V} = -\frac{d\rho_a}{\rho_a}. \quad (3.17)$$

Setzt man Gl. (3.17) in (3.16) ein, so erhält man

$$\frac{dp_a}{p_a} = \gamma \cdot \frac{d\rho_a}{\rho_a}. \quad (3.18)$$

Da $p_a(t) = p_0 + p(t)$ und $\rho_a(t) = \rho_0 + \rho(t)$ jeweils die Summe aus einem großen statischen Anteil und einem kleinen Schwankungsanteil sind, wobei $\rho \ll \rho_0$ und $p \ll p_0$, kann man für Gl. (3.18) näherungsweise auch

$$\frac{p}{p_0} = \gamma \cdot \frac{\rho}{\rho_0} \quad \text{bzw.} \quad p = \gamma \cdot \frac{p_0}{\rho_0} \cdot \rho$$

schreiben. Der Term $\gamma \cdot p_0/\rho_0$ ist per Definition das Quadrat der Schallgeschwindigkeit c , so dass man die einfache Beziehung

$$p = c^2 \cdot \rho \quad (3.19)$$

erhält. Die durch eine Schallwelle verursachten Druck- und Dichteschwankungen sind folglich direkt proportional zueinander.

3.3.4 Schwingungsgleichung für die Rohrwand

Die Wände des Vokaltrakts sind ein verteiltes mechanisches System mit Eigenschaften wie Trägheit und Elastizität. Bei niedrigen Frequenzen, bei denen die Wände einen maßgeblichen Einfluss auf die akustischen Eigenschaften des Vokaltrakts haben, verhalten sie sich in erster Näherung wie ein gedämpftes Feder-Masse-System [52]. Die Luftdruckänderungen im Vokaltrakt sind die treibende Kraft für das Ausdehnen und Zusammenziehen der Wände. Dadurch wird dem Schallfeld Energie entzogen, was zu einer Dämpfung und Verschiebung der Formanten bei niedrigen Frequenzen führt.

Mathematisch kann die Ausdehnung bzw. das Zusammenziehen der Wände durch die Auslenkung $y(x, t)$ senkrecht zu ihrer Oberfläche erfasst werden. Wir betrachten dazu ein infinitesimales Flächenelement dA_w der Vokaltraktwand, auf die der lokale Schalldruck $p(x, t)$ als äußere Kraft einwirkt. Wenn wir von der vereinfachenden Annahme ausgehen, dass die Vokaltraktwand nur lokal auf den dortigen Druck reagiert und keine Kräfte durch die Kopplung zu den benachbarten Wandabschnitten erfährt, so lautet die allgemeine Schwingungsgleichung

$$p \cdot dA_w = M \cdot dA_w \cdot \frac{\partial^2 y}{\partial t^2} + B \cdot dA_w \cdot \frac{\partial y}{\partial t} + K \cdot dA_w \cdot y,$$

wobei $M(x)$, $B(x)$ und $K(x)$ jeweils die Masse, Dämpfungskonstante und Federkonstante (Steifigkeit) der Wand *pro Flächeneinheit* sind. Die Division der Gleichung durch dA_w ergibt

$$p = M \cdot \frac{\partial^2 y}{\partial t^2} + B \cdot \frac{\partial y}{\partial t} + K \cdot y. \quad (3.20)$$

3.4 Verluste im Sprechtrakt

Im Sprechtrakt entstehen aufgrund verschiedener Ursachen Energieverluste, die sich auf die Frequenzen und die Bandbreiten der Formanten auswirken. Den auditiv wichtigsten Effekt hat die Vergrößerung der Bandbreiten bzw. die Dämpfung der Formanten. Die Ursachen für die Verluste sind die Reibung der Luft an der Rohrwand, die Entstehung von Wirbelgebieten an plötzlichen Rohrerweiterungen und -verengungen, die Abstrahlung von Schallwellen an der Mund- und Nasenöffnung, die Absorption von Energie durch die elastischen Sprechtraktwände und die Ableitung von Kompressionswärme über die Rohrwand. Von STEVENS [110] wird der Einfluss der einzelnen Verlustmechanismen auf die Formantbandbreiten im Einzelnen diskutiert. Für mittlere bis hohe Frequenzen dominiert der Verlust durch Schallabstrahlung, gefolgt von den Verlusten durch Reibung und Wärmeleitung. Bei niedrigen Frequenzen dominiert der Verlust durch die elastischen Wände, der sich hauptsächlich auf die Bandbreite des ersten Formanten auswirkt. Der Einfluss der Verluste durch plötzliche Querschnittsänderungen hängt von den Querschnittsflächen der Engstellen im Vokaltrakt ab.

In diesem Abschnitt wird der theoretische Hintergrund und die Umsetzung der Verluste im Rahmen dieser Arbeit sowie in früheren Studien diskutiert. Die Verluste durch Wärmeleitung werden aufgrund ihres geringen Effekts auf die Formantbandbreiten nicht berücksichtigt.

3.4.1 Verluste durch Reibung an der Rohrwand

Theoretischer Hintergrund.

Jede Fluidbewegung durch eine Rohrleitung ist aufgrund der Viskosität des Fluids einem Widerstand R ausgesetzt, der als das Verhältnis von Druckverlust zu Volumenstrom definiert ist, d. h. $R = \Delta p / u$. Der Widerstand hängt sowohl von der Rohrgeometrie und -beschaffenheit als auch von den Strömungsverhältnissen ab. Bei den folgenden Betrachtungen beziehen wir uns jeweils auf ein kurzes Rohrstück mit der Länge Δx , dem Querschnitt A und dem Umfang S .

Wenn das Rohrstück von einer Schallwelle durchlaufen wird, so entsteht der Reibungswiderstand hauptsächlich durch die Schubspannungen in einer dünnen Grenzschicht direkt an der Rohrwand. Weiter in Richtung Rohrmitte ist die Strömungsgeschwindigkeit annähernd gleichförmig, so dass dort kaum noch Schubspannungen auftreten. Die Dicke der Grenzschicht und damit der Widerstand hängen neben der Rohrgeometrie auch von der Frequenz der Schallwelle ab. In [38] wurde die folgende Formel für den Grenzschichtwiderstand abgeleitet:

$$R_{\text{grenz}} = \Delta x \cdot \frac{S}{A^2} \cdot \sqrt{\frac{\omega \cdot \varrho_0 \cdot \mu}{2}}. \quad (3.21)$$

Darin ist ω die Kreisfrequenz und μ die Viskosität des Fluids (siehe Tab. B.1).

Anders verhält sich der Rohrwiderstand im Fall einer voll ausgebildeten *stationären* Strömung. Dabei muss man wiederum zwischen einer laminaren und einer turbulenten Strömung unterscheiden. Ob eine Strömung laminar oder turbulent verläuft, lässt sich anhand der sogenannten Reynoldszahl Re entscheiden. Das ist eine dimensionslose Kennzahl, die in der Strömungsmechanik eine breite Anwendung findet und wie folgt berechnet wird:

$$Re = v \cdot d / \nu. \quad (3.22)$$

Darin ist v die mittlere Strömungsgeschwindigkeit, d der Rohrdurchmesser und ν die kinematische Viskosität (siehe Tab. B.1). Für nicht-kreisförmige Rohre bezeichnet d die „wirksame Weite“ des Rohres, die für elliptische Querschnitte laut [79] wie folgt berechnet wird:

$$d = \frac{4 \cdot b \cdot a}{(b + a) \cdot \left[1 + \frac{1}{4} \cdot \left(\frac{b-a}{b+a} \right)^2 \right]}. \quad (3.23)$$

Darin ist $2 \cdot b$ der längste und $2 \cdot a$ der kürzeste Durchmesser der Ellipse. Mittels Gl. (3.23) ergibt sich immer eine wirksame Weite d , die zwischen dem kürzesten und dem längsten Ellipsendurchmesser liegt. Der Umschlag von einer laminaren zu einer turbulenten Strömung erfolgt, wenn Re eine kritische Reynoldszahl Re_{krit} überschreitet, die typischerweise als $Re_{\text{krit}} \approx 2300$ angegeben wird [121, S. 19]. Tatsächlich erfolgt der Umschlag jedoch keineswegs plötzlich, sondern die Entstehung erster vereinzelter Wirbel beginnt bereits bei $Re \approx 1000$ und nimmt für größere Re -Zahlen kontinuierlich zu, bis der Umschlag bei $Re \approx 2300$ dann vollständig erfolgt ist (siehe [49, S. 203], [96, S. 525]).

Für eine stationäre, *laminare* Rohrströmung in einem kreisrunden Rohr lautet die Formel für den Reibungswiderstand

$$R_{\text{lam}} = \Delta x \cdot \frac{8 \cdot \pi \cdot \mu}{A^2}. \quad (3.24)$$

Diese Strömung ist durch ein parabolisches Geschwindigkeitsprofil gekennzeichnet und wird Hagen-Poiseuille-Strömung genannt. Für den allgemeineren Fall einer stationären, laminaren Strömung durch ein Rohr mit *elliptischem* Querschnitt hat das Geschwindigkeitsprofil die Form eines elliptischen Paraboloids. Der Reibungswiderstand lässt sich aus dem Geschwindigkeitsprofil, das in [49] angegeben ist, ableiten, und lautet

$$R_{\text{lam}} = \Delta x \cdot \frac{4 \cdot \mu \cdot (a^2 + b^2)}{\pi \cdot a^3 \cdot b^3}, \quad (3.25)$$

wobei $2 \cdot a$ und $2 \cdot b$ der längste und kürzeste Durchmesser der Ellipse sind. Für den Spezialfall $a = b$ erhält man wieder die Gl. (3.24) für den Widerstand eines kreisrunden Rohres.

Für eine voll ausgebildete *turbulente* Strömung lässt sich der Druckverlust in einem hydraulisch glatten Rohr laut HUTTER [49, S. 203] und TRUCKENBRODT [121, S. 101] mit der Formel

$$\Delta p = \frac{\varrho_0}{2} \cdot v^2 \cdot \frac{\Delta x}{d} \cdot \lambda$$

berechnen. Darin ist d der Rohrdurchmesser (bzw. die charakteristische Weite für einen elliptischen Querschnitt), v die mittlere Strömungsgeschwindigkeit und λ die dimensionslose Rohrreibungszahl, die für die turbulente Strömung nach dem empirischen Gesetz von Blasius mit Hilfe der Reynolds-Zahl ausgedrückt werden kann: $\lambda = 0.316 \cdot Re^{-1/4}$. Für den Widerstand ergibt sich mit $v = u/A$:

$$R_{\text{turb}} = \Delta x \cdot \frac{\lambda}{d} \cdot \frac{\varrho_0}{2} \cdot \frac{|u|^2}{A^2} \quad (3.26)$$

Die typischen Re -Zahlen für turbulente Strömung in der Sprachproduktion liegen in etwa zw. 1800 und 10000 und die Rohrreibungszahlen λ damit zwischen 0.049 und 0.032.

Behandlung des Reibungswiderstandes in anderen Studien.

In mehreren Studien wird für die akustische Simulation der Grenzschichtwiderstand nach Gl. (3.21) verwendet (z. B. in [33, 51, 126]), denn diese Formel gibt korrekt die Zunahme des Widerstandes mit ansteigender Frequenz wieder. Für eine Simulation des Schallfeldes im Zeitbereich benötigt man jedoch *frequenzunabhängige* Komponenten für diese Widerstände. Um einen frequenzunabhängigen Widerstand zu erhalten, wird von vielen Autoren einfach eine feste Frequenz ω_{fest} in die Gleichung eingesetzt [14, 51]. WAKITA *et al.* [126] haben aber gezeigt, dass dadurch relativ große Fehler bei den Bandbreiten der Formanten entstehen. Für ein hohes ω_{fest} werden die Formanten mit niedrigen Frequenzen zu stark gedämpft, und für ein niedriges ω_{fest} werden die Formanten mit hohen Frequenzen zu wenig gedämpft. MAEDA [72] verwendet in seiner Zeitbereichssimulation statt des Grenzschichtwiderstandes den frequenzunabhängigen Widerstand für die stationäre Rohrströmung nach Gl. (3.24). Außer für sehr kleine Rohrquerschnitte ist dieser Widerstand jedoch grundsätzlich kleiner als der Grenzschichtwiderstand, so dass die Dämpfung aller Formanten zu gering ausfällt. Der Widerstand der Stimmritze wird in der Literatur häufig anders behandelt als der des restlichen Vokaltrakts. Meistens wird dafür der Widerstand einer stationären, laminaren *Spaltströmung* verwendet [11, 51]:

$$R_{\text{spalt}} = \Delta x \cdot \frac{12 \cdot \mu}{a \cdot b^3}. \quad (3.27)$$

Dabei bezeichnet a die Länge und b die Breite des (rechteckigen) Spalts bzw. der Glottis.

Berücksichtigung des Reibungswiderstandes in dieser Arbeit.

In dieser Arbeit erfolgt die Sprachsynthese mit Hilfe einer akustischen Simulation im Zeitbereich, so dass frequenzunabhängige Widerstände benötigt werden. Da ich von elliptischen Rohrquerschnitten ausgegangen bin, verwende ich zur Berechnung der Reibungswiderstände Gl. (3.25), die streng genommen nur für eine stationäre, laminare Strömung gilt. Die Verwendung dieser Gleichung hat den Vorteil, dass damit auch der Widerstand von elliptischen Rohrquerschnitten mit hoher Exzentrizität, wie der der Glottis, berechnet werden kann. Für eine typische glottale Querschnittsfläche von 10 mm^2 während der Phonation

unterscheiden sich R_{lam} nach Gl. (3.25) und R_{spalt} nach Gl. (3.27) nur um ca. 20%, wenn man sowohl für die elliptische Öffnung als auch die Spaltöffnung eine Länge von 10 mm ansetzt. Der Widerstand für turbulente Strömungsverhältnisse wird in dieser Arbeit nicht gesondert berücksichtigt. Weiter oben wurde bereits angemerkt, dass der Widerstand für eine stationäre Strömung außer für sehr kleine Rohrquerschnitte geringer ausfällt als der Grenzschichtwiderstand, und dadurch die Formanten zu wenig gedämpft werden. In Abschnitt 3.7 werde ich eine Möglichkeit aufzeigen, die Bandbreiten durch *numerische Dämpfung* zu vergrößern und den Fehler somit klein zu halten.

Für die akustische Simulation im Frequenzbereich in Abschnitt 3.6 wird statt R_{lam} der Grenzschichtwiderstand R_{grenz} verwendet, um die daraus resultierenden Formantbandbreiten mit denen der zeitdiskreten Simulation zu vergleichen.

3.4.2 Verluste durch Wirbelbildung

Theoretischer Hintergrund.

In Abschnitt 3.3.1 wurde die Gleichung

$$-\frac{\partial p}{\partial x} = \varrho_0 \cdot \frac{\partial v}{\partial t} + \frac{\varrho_0}{2} \cdot \frac{\partial v^2}{\partial x} \quad (3.28)$$

für die eindimensionale, reibungsfreie Fluidbewegung abgeleitet. In dieser Gleichung steckt implizit die Annahme, dass sich der Strömungsquerschnitt in jeder Situation exakt dem Rohrquerschnitt anpasst, dass also die Stromlinien jeder Änderung des Rohrquerschnitts folgen. Für stetige Querschnittsänderungen entlang der Rohrachse ist die Annahme in aller Regel erfüllt. Bei plötzlichen Änderungen des Querschnitts kann sich die Strömung jedoch von der Rohrwand ablösen, was zur Entstehung von Wirbelgebieten führt und einen fluidmechanischen Energieverlust zur Folge hat. Im Vokaltrakt findet man solche plötzlichen Querschnittsänderungen hauptsächlich am Eingang und/oder Ausgang von Engstellen wie der Glottis oder einer supraglottalen Konstriktion. Für die Untersuchung der fluidmechanischen Energieverluste an diesen Stellen ist es zweckmäßig, die örtliche Druckänderung $-\partial p/\partial x$ aufgrund des Gradienten von v^2 unabhängig von der Druckänderung durch den Beschleunigungsterm $\varrho_0 \cdot \partial v/\partial t$ zu betrachten. Dadurch erhält man die Gleichung

$$-\frac{\partial p}{\partial x} = \frac{\varrho_0}{2} \cdot \frac{\partial v^2}{\partial x}.$$

Sie entspricht dem Fall einer reibungsfreien, *stationären* Strömung. Wenn man beide Seiten der Gleichung mit dem infinitesimalen Wegelement dx multipliziert und zwischen zwei Positionen x_1 und x_2 auf der Rohrachse integriert, so erhält man die Druckgleichung von Bernoulli:

$$p_2 - p_1 = \frac{\varrho_0}{2} \cdot v_1^2 - \frac{\varrho_0}{2} \cdot v_2^2 \quad \text{bzw.} \quad p_1 + \frac{\varrho_0}{2} \cdot v_1^2 = p_2 + \frac{\varrho_0}{2} \cdot v_2^2 \quad (3.29)$$

Darin sind v_1 und v_2 die Strömungsgeschwindigkeiten und p_1 und p_2 die Drücke an den Positionen x_1 und x_2 . Die Terme $\varrho_0 \cdot v_1^2/2$ und $\varrho_0 \cdot v_2^2/2$ werden als Geschwindigkeitsdruck oder Staudruck bezeichnet. Wenn die Positionen x_1 und x_2 nicht zu weit auseinanderliegen bzw. die Kompression des Fluids zwischen ihnen vernachlässigt werden kann, so gilt außerdem $v_1 \cdot A_1 = v_2 \cdot A_2$ (Kontinuität des Volumenstroms), wobei A_1 und A_2 die Rohrquerschnitte

an den entsprechenden Positionen sind. Gleichung (3.29) drückt damit aus, dass für den verlustfreien Fall (1) der Druck durch eine Querschnittserweiterung bzw. eine Abnahme der Strömungsgeschwindigkeit zunimmt und (2) der Druck durch eine Querschnittsverengung bzw. eine Zunahme der Strömungsgeschwindigkeit abnimmt. Diese Druckänderungen aufgrund von Geschwindigkeitsänderungen werden als Bernoulli-Effekt bezeichnet.

Bei einer *plötzlichen* Rohrerweiterung oder -verengung kann sich die Strömung von der Rohrwand ablösen. Für eine Rohrerweiterung ist der Fall in Abb. 3.3 (a) dargestellt.

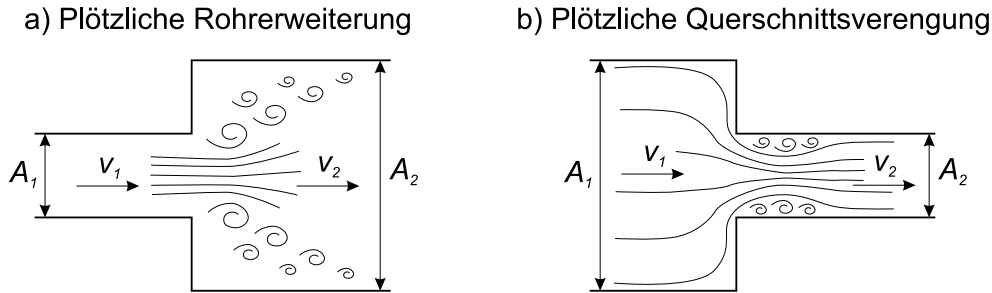


Abbildung 3.3: (a) Plötzliche Rohrerweiterung (Stufendiffusor) und (b) plötzliche Rohrverengung mit Strahleinschnürung (Stufendüse).

In der Fluidmechanik spricht man in dieser Situation von einem Stufendiffusor oder Stoßdiffusor [121]. Dabei tritt das strömende Fluid mit der mittleren Geschwindigkeit v_1 zunächst aus dem engeren in den weiteren Querschnitt ein und vermischt sich weiter stromabwärts infolge der Reibung unter starker Wirbelbildung mit dem umgebenden Fluid. Nach einer gewissen Übergangsströmung legt sich das Fluid wieder an die Rohrwand an. Die Druckänderung $p_1 - p_2$ lässt sich als

$$p_1 - p_2 = \frac{\rho_0}{2} \cdot (v_2^2 - v_1^2) + \underbrace{\frac{\rho_0}{2} \cdot (v_1 - v_2)^2}_{\Delta p_{\text{Carnot}}} = \rho_0 \cdot (v_2^2 - v_1 \cdot v_2) \quad (3.30)$$

schreiben [49, S. 108]. Von der reibungsfreien Bernoulli-Gleichung (3.29) unterscheidet sich (3.30) nur durch den zusätzlichen Verlustterm¹ Δp_{Carnot} . Definiert man das Verhältnis des tatsächlichen Druckanstiegs $p_2 - p_1$ nach Gl. (3.30) zum theoretisch größtmöglichen Druckanstieg nach Gl. (3.29) als Wirkungsgrad η , so kann man dafür schreiben

$$\eta = \frac{2 \cdot (v_1 - v_2) \cdot v_2}{v_1^2 - v_2^2} = \frac{2}{1 + A_2/A_1}.$$

η ist also ein Maß dafür, wieviel von dem Staudruck im engeren Rohrabschnitt durch die sprunghafte Rohrerweiterung zurückgewonnen wird. Wenn $A_1 = A_2$, so ist $\eta = 1$. Wenn sich das Flächenverhältnis A_2/A_1 jedoch vergrößert, nimmt der Wirkungsgrad immer weiter ab und damit der Energieverlust zu. Der Wirkungsgrad am Ausgang der Glottis lässt sich abschätzen, wenn man für eine weite Glottisöffnung $A_1 = 0.2 \text{ cm}^2$ und für den Querschnitt direkt oberhalb der Glottis $A_2 = 2.0 \text{ cm}^2$ annimmt. Daraus folgt $\eta \approx 0.18$, d. h. nur 18% des Staudrucks in der Glottis werden zurückgewonnen. Während eines großen Teils

¹Der fluidmechanische Energieverlust an einem Stufendiffusor wird auch als Carnotscher Stoßverlust bezeichnet.

des glottalen Schwingungszyklus' ist A_1 aber deutlich kleiner als 0.2 cm^2 , so dass der Wirkungsgrad in der Regel kleiner als 10% ist [116]. An der Mundöffnung und an den Nasenlöchern tritt die Strömung aus dem Ansatzrohr aus, d. h. $A_2 \gg A_1$. Damit geht η gegen Null (der gesamte Staudruck geht verloren) und $p_1 \approx p_2$.

Wenn sich der Rohrquerschnitt in Strömungsrichtung plötzlich verengt, so spricht man von einer Stufendüse [121]. Auch diese Situation kann einen fluidmechanischen Energieverlust verursachen und ist in Abb. 3.3 (b) dargestellt. Dabei wird der Strahl am Eingang des engeren Rohres ein wenig eingeschnürt. Zwischen dem eingeschnürten Strahl und der Rohrwand entsteht ein Wirbelgebiet (Totraum). Den dabei verursachten Energieverlust kann man durch einen Verlustfaktor k in der Bernoulli-Gleichung erfassen:

$$p_1 - p_2 = k \cdot \frac{\rho}{2} \cdot (v_2^2 - v_1^2) \quad \text{mit} \quad k \geq 1. \quad (3.31)$$

$k = 1$ entspricht einem verlustfreien Übergang.

Behandlung von Wirbelverlusten in anderen Studien.

Außer am Eingang und Ausgang der Glottis wird der Bernoulli-Effekt im Vokaltrakt von fast allen Autoren vernachlässigt. Die Druckänderung am Eingang der Glottis wird normalerweise durch Gl. (3.31) mit $k = 1.37$ erfasst (siehe z. B. [51, 20, 23]). Dieser Wert für k geht auf Messungen des Druckverlustes an einem Modell der Glottis durch VAN DEN BERG *et al.* [11] zurück. Von einigen Autoren [93, 94] wird die Einschnürung des Luftstrahls am Eingang der Glottis jedoch bezweifelt und weggelassen ($k = 1$), da sie theoretisch nur an *plötzlichen* Übergängen von einem weiten in einen engen Rohrquerschnitt auftreten sollte. Der Einlass zur Glottis erfolgt aber relativ allmählich. Es ist daher möglich, dass der in [11] gemessene Verlust auf die Wandreibung im Eingangsbereich statt auf die Einschnürung des Luftstrahls zurückzuführen ist. Bei der plötzlichen Querschnittserweiterung am Ausgang der Glottis wird der Druckanstieg in der Regel entweder nach Gl. (3.30) berechnet [23, 51, 112, 113] oder ein Totalverlust des Staudrucks in der Glottis angenommen ($p_1 = p_2$) [21, 116].

Der Carnotsche Stoßverlust am Ausgang einer supraglottalen Konstriktion wird in den meisten Studien durch einen einzelnen punktuellen Widerstand R_c in der Konstriktion repräsentiert, über dem der dortige Staudruck abfällt [5, 106, 108, 110]. Wenn A_c der Querschnitt der Konstriktion ist und u_c der Volumenstrom, dann lautet die Gleichung für den Widerstand

$$R_c = k_c \cdot \frac{\rho_0}{2} \cdot \frac{|u_c|}{A_c^2}.$$

Die Konstante k_c hängt von dem Verhältnis von A_c zu den Querschnittsflächen vor und hinter der Konstriktion ab. Für einen vollständigen Verlust der kinetischen Energie gilt $k_c \approx 1$. Grundsätzlich wird der Druckverlust durch den einzelnen Widerstand R_c in der Konstriktion quantitativ richtig erfasst. Die tatsächliche *Druckverteilung* im Vokaltrakt wird damit jedoch nur sehr grob approximiert, da der gesamte Abfall des Geschwindigkeitsdrucks an einer einzigen Stelle (an der Position von R_c) im Vokaltrakt erfolgt. Tatsächlich nimmt der Druck im Eingangsbereich der Konstriktion dagegen aufgrund des Bernoulli-Effekts mit abnehmendem Querschnitt kontinuierlich ab und steigt am Ende der Konstriktion durch den Stoßverlust nicht oder kaum wieder an. Abbildung 3.4 verdeutlicht den Unterschied zwischen einem plötzlichen Druckabfall am Widerstand R_c und

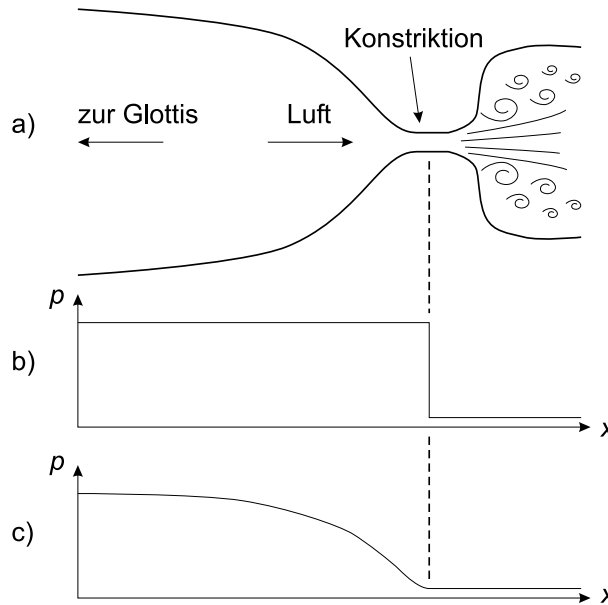


Abbildung 3.4: (a) Luftstrom durch eine supraglottale Konstriktion. (b) Druckverlauf mit einem konzentrierten Verlustwiderstand in der Konstriktion. (c) Tatsächlicher Druckverlauf mit einem Stoßverlust am Ausgang der Konstriktion.

dem allmählichen Druckabfall am Eingang der Konstriktion aufgrund des Bernoulli-Effekts.

Behandlung der Wirbelverluste in der vorliegenden Arbeit.

In der vorliegenden Arbeit wurde ein einfaches wie auch effektives Verfahren zur Behandlung der Verluste durch die Wirbelbildung an plötzlichen Querschnittsänderungen umgesetzt. Es unterscheidet sich von den meisten anderen Arbeiten dadurch, dass der Bernoulli-Effekt auf der Grundlage der Bewegungsgleichung (3.28) grundsätzlich im gesamten Vokaltrakt berücksichtigt wird. Das Verfahren beruht auf zwei einfachen Prinzipien²:

(1) Die Fluidbewegung von einem weiten in einen engen Rohrquerschnitt erfolgt stets verlustfrei, also ohne Einschnürung des Luftstrahls. Daraus folgt, dass der Druck im engen Querschnitt entsprechend der Bernoulli-Gleichung um den Betrag $(\rho_0/2) \cdot (v_2^2 - v_1^2)$ niedriger ist als im weiten Querschnitt.

(2) Die Fluidbewegung von einem engen in einen weiten Rohrquerschnitt ist stets mit dem Totalverlust der überschüssigen Geschwindigkeitsenergie im engen Rohrquerschnitt verbunden. Daraus folgt, dass der Druck an einer Rohrerweiterung nicht durch die Zurückgewinnung von fluidmechanischer Geschwindigkeitsenergie ansteigt. Der Ausdruck $(\rho_0/2) \cdot (v_2^2 - v_1^2)$ in Gl. (3.29) ist somit Null und folglich gilt $p_1 = p_2$.

Das erste Prinzip ist unmittelbar einsichtig, wenn man ausschließlich von *allmählichen* Querschnittsverengungen im Vokaltrakt ausgeht, auch beim Eingang zur Glottis. Die Druckverteilung am Eingang einer Konstriktion ähnelt damit Abb. 3.4 (c). Das erste Prinzip hat auch praktische Auswirkungen auf die Qualität der Synthese dynamischer Vokal-Frikativ-Vokal-Sequenzen. Wenn man jedoch entgegen dem ersten Prinzip den Druckverlust

²Der Einfluss der Wandreibung wird in dieser Diskussion nicht berücksichtigt.

an einer Konstriktion durch einen punktuellen Widerstand simuliert (siehe Abb. 3.4 (b)), so kann aufgrund der Approximation des Vokaltrakts durch kurze Rohrabschnitte die Position des Verlustwiderstands bei einer geringen Verschiebung des Konstriktionsortes plötzlich von einem Rohrabschnitt zu einem benachbarten Rohrabschnitt „springen“. Diese diskrete Änderung der Position des Verlustwiderstands ist im synthetischen Sprachsignal als leises aber störendes Knackgeräusch hörbar. Bei der kontinuierlichen Druckabnahme nach Abb. 3.4 (c) treten solche Störgeräusche bei einer Verschiebung der Konstriktion nicht auf. Das zweite Prinzip unterstellt, dass der Wirkungsgrad η für die Zurückgewinnung von Geschwindigkeitsdruck an allen Querschnittserweiterungen Null ist. Für die Querschnittserweiterung am Ausgang der Glottis wurde gezeigt, dass η die meiste Zeit während einer Schwingungsperiode kleiner als 0.1 ist. Bei der Querschnittserweiterung nach einer supraglottalen Konstriktion kann man aufgrund ähnlicher Flächenverhältnisse von einem ebenso kleinen Wirkungsgrad ausgehen. An den offenen Rohrenden (Mund- und Nasenöffnung) geht der Wirkungsgrad auch theoretisch gegen Null. Von daher ist auch das zweite Prinzip plausibel.

Beide Prinzipien gelten auch für die Querschnittsänderung *innerhalb* der Stimmritze. Diese hat je nach Phasenlage eine eher konvergente Form (Querschnittsverengung) während des Öffnens und eine eher divergente Form (Querschnittserweiterung) während des Schließens. Die obigen Prinzipien drücken dabei aus, dass sich der gebündelte Luftstrahl durch die Glottis stets an der Position des engsten Querschnitts ablöst, also entweder im unteren oder oberen Teil der Glottis. Dieses Verhalten ist laut [93, 94, 112] realistischer als die Annahme von ISHIZAKA *et al.* [51], dass sich der Luftstrahl stets am oberen Ende der Glottis ablöst und innerhalb der Stimmritze die Bernoulli-Gleichung gilt. In meinen Syntheseversuchen hat sich gezeigt, dass ein wechselnder Ablösepunkt für den glottalen Luftstrahl eine wesentlich natürlichere Stimme hervorbringt als ein konstanter Ablösepunkt am oberen Ende der Stimmritze. Andererseits hat sich kein wahrnehmbarer Unterschied im Stimmklang gezeigt, wenn bei der Querschnittserweiterung von der Glottis in den Rachen statt des zweiten Prinzips die Gleichung (3.30) für den Stoßverlust verwendet wurde.

3.4.3 Verluste durch Schallabstrahlung

Neben den Verlusten durch Wandreibung und Querschnittsänderungen entstehen zusätzliche Verluste durch die Abstrahlung von Schallwellen an der Mundöffnung und den Nasenlöchern. Die Verluste durch die Abstrahlung lassen sich in Form von Strahlungsimpedanzen erfassen, die den Sprechtrakt an den offenen Enden abschließen. In erster Näherung sind sie mit der Strahlungsimpedanz eines vibrierenden Kolbens in einer Kugel vergleichbar (siehe Abb. 3.5 (a)). Die Kolbenquerschnittsfläche entspricht dabei der Öffnungsfläche vom Mund bzw. von den Nasenlöchern und die Kugel repräsentiert den Kopf. In [85] wurde die Strahlungsimpedanz genau für einen solchen Fall abgeleitet. Sie ist sowohl eine Funktion von der Frequenz als auch von der relativen Größe zwischen Kolben und Kugel. Die Funktion lässt sich jedoch nicht als geschlossener analytischer Ausdruck darstellen.

Ein Grenzfall liegt laut [38] aber vor, wenn die Größe des Kolbens sehr klein im Vergleich zum Kugeldurchmesser ist. Dann ist die Strahlungsimpedanz mit der eines Kolbens in einer unendlich ausgedehnten Wand (siehe Abb. 3.5 (b)) vergleichbar. Sie lässt sich in

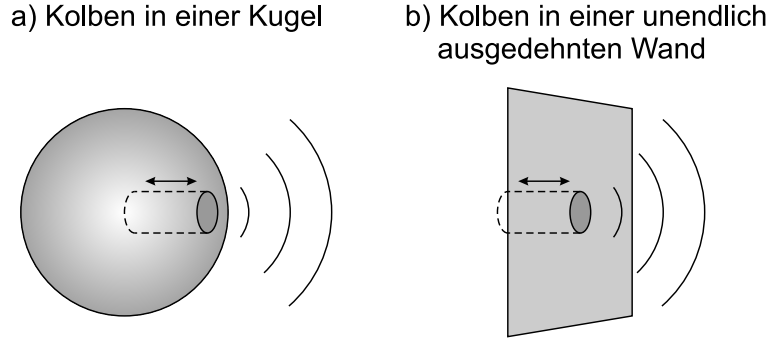


Abbildung 3.5: Schallabstrahlung von einem schwingenden Kolben in einer Kugel (a) und in einer unendlich ausgedehnten Wand (b). Die Doppelpfeile geben die Schwingungsrichtung der Kolben an.

geschlossener Form als

$$Z = \frac{\varrho_0 \cdot c}{A} \cdot \left[1 - \frac{J_1(2 \cdot k \cdot a)}{k \cdot a} + j \cdot \frac{S_1(2 \cdot k \cdot a)}{k \cdot a} \right]$$

aufschreiben, wobei a der Radius des Kolbens, $k = \omega/c$ die Wellenzahl und A die Kolbenfläche ist. $J_1(x)$ ist die Besselfunktion erster Ordnung und $S_1(x)$ die Struve-Funktion erster Ordnung:

$$\begin{aligned} J_1(x) &= \frac{x}{2} - \frac{x^3}{2^3 \cdot 1! \cdot 2!} + \frac{x^5}{2^5 \cdot 2! \cdot 3!} - \frac{x^7}{2^7 \cdot 3! \cdot 4!} + \dots \\ S_1(x) &= \frac{2}{\pi} \cdot \left(\frac{x^2}{3} - \frac{x^4}{3^2 \cdot 5} + \frac{x^6}{3^2 \cdot 5^2 \cdot 7} - \frac{x^8}{3^2 \cdot 5^2 \cdot 7^2 \cdot 9} + \dots \right). \end{aligned}$$

Für kleine Werte für $k \cdot a$ kann man Z durch die ersten zwei Terme von $J_1(x)$ und den ersten Term von $S_1(x)$ approximieren, und man erhält

$$Z = \frac{\varrho_0 \cdot c}{A} \cdot \left[\frac{(k \cdot a)^2}{2} + j \cdot \frac{8 \cdot (k \cdot a)}{3 \cdot \pi} \right].$$

Wenn man $k = \omega/c$ und $a = \sqrt{A/\pi}$ in die Gleichung einsetzt und ausmultipliziert, so ergibt sich

$$Z = \frac{\varrho_0 \cdot \omega^2}{2 \cdot c \cdot \pi} + j \cdot \frac{8 \cdot \varrho_0 \cdot \omega}{3 \cdot \pi \cdot \sqrt{A \cdot \pi}}. \quad (3.32)$$

Diese Formel für die Strahlungsimpedanz wird häufig verwendet. FANT [33] und WAKITA *et al.* [126] gehen noch einen Schritt weiter und kompensieren den Fehler, der durch die Verwendung der Formel für einen vibrierenden Kolben in einer Wand (statt in einer Kugel) gemacht wurde, durch einen Korrekturfaktor vor dem Realteil der Impedanz. Ihre Formel lautet

$$Z = K(\omega) \cdot \frac{\varrho_0 \cdot \omega^2}{4 \cdot c \cdot \pi} + j \cdot \frac{8 \cdot \varrho_0 \cdot \omega}{3 \cdot \pi \cdot \sqrt{A \cdot \pi}},$$

wobei

$$K(\omega) = \begin{cases} \frac{0.6 \cdot \omega}{2 \cdot \pi \cdot 1600} + 1 & \text{für } 0 \leq \omega < 2 \cdot \pi \cdot 1600 \\ 1.6 & \text{für } \omega \geq 2 \cdot \pi \cdot 1600. \end{cases}$$

Aus Gründen, die in Abschnitt 3.5.2 dargelegt werden, verwende ich in der vorliegenden Arbeit jedoch die Strahlungsimpedanz nach Gl. (3.32).

3.4.4 Verluste durch elastische Rohrwände

In Abschnitt 3.3.4 wurden die elastischen Wände des Vokaltrakts als ein gedämpftes Feder-Masse-System modelliert, das lokal auf den Schalldruck im Sprechtrakt reagiert. Die Parameter des Systems sind die Masse M , die Dämpfungskonstante B und die Federkonstante K jeweils pro Flächeneinheit. Direkte Messungen dieser Parameter innerhalb des Vokaltrakts wurden bisher nicht angestellt, sie lassen sich aber anhand von Messungen der mechanischen Impedanz an anderen Körperteilen abschätzen. Solche Messungen wurden z. B. von ISHIZAKA *et al.* [52] für den Unterarm, den Hals und die Wangen im entspannten und angespannten Zustand vorgenommen. Außerdem hat STEVENS [110] auf der Basis verschiedener Studien Wertebereiche für die Parameter angegeben. In Tab. 3.1 sind die Wertebereiche von STEVENS und die konkreten Messergebnisse von ISHIZAKA *et al.* zusammengestellt.

	M in kg/m ²	B in kg/(m ² · s)	K in N/m ³
Wangen (entspannt)	21	8000	845000
Wangen (angespannt)	15	10600	333000
Hals	24	23200	4910000
Unterarm	18	16200	6150000
Wertebereiche nach STEVENS	10 – 20	8000 – 20000	330000 – 1000000

Tabelle 3.1: Parameterwerte der mechanischen Impedanz der Vokaltraktwände. Die ersten vier Zeilen sind konkrete Messergebnisse aus [52] und die letzte Zeile zeigt die von STEVENS [110] angegebenen Wertebereiche.

Für den Rachen-, Mund- und Nasenraum verwende ich in der vorliegenden Arbeit die Daten für die entspannten Wangen. Die Parameter für die subglottalen Luftwege basieren dagegen auf den experimentellen Ergebnisse der Studie [53]. Dort wurden zunächst die subglottalen Resonanzfrequenzen von mehreren Versuchspersonen gemessen. Die ersten drei Resonanzen lagen im Mittel bei 640, 1400 und 2100 Hz. Anschließend wurde versucht, diese Frequenzen mit Hilfe eines Computermodells des subglottalen Systems zu erzeugen. Das dazu verwendete Modell basiert auf den gleichen anatomischen Daten wie die vorliegende Arbeit (siehe Abschnitt 2.6). Die Studie hat gezeigt, dass die mechanische Impedanz der Rohrwände einen wesentlichen Einfluss auf die Resonanzfrequenzen und -bandbreiten des subglottalen Systems besitzt. Für eine gute Übereinstimmung von real gemessenen und simulierten Resonanzen wurden die folgenden Daten experimentell ermittelt:

$$\begin{aligned} \text{Wandmasse:} \quad & M = 2.5 \dots 3.0 \text{ kg} \cdot \text{m}^{-2}, \\ \text{Dämpfungskoeffizient:} \quad & B/(2 \cdot M) = 250 \text{ (Luftröhre)} \dots 2500 \text{ s}^{-1} \text{ (Alveolen)}. \end{aligned}$$

Der Dämpfungskoeffizient hängt von der Position im subglottalen System ab und ist für den peripheren Bereich der Lunge etwa um den Faktor 10 größer als für die Luftröhre. In der vorliegenden Studie verwende ich für die Wandmasse einen konstanten Wert von $M = 3.0 \text{ kg} \cdot \text{m}^{-2}$ und für B einen Wert, der linear von $10000 \text{ kg} \cdot \text{m}^{-2} \cdot \text{s}^{-1}$ an den Alveolen bis auf $1000 \text{ kg} \cdot \text{m}^{-2} \cdot \text{s}^{-1}$ für den obersten subglottalen Rohrabchnitt abfällt. Werte für die

Wandsteifigkeit K wurden in [53] nicht angegeben, da diese ohnehin nur einen sehr geringen Einfluss auf die Resonanzfrequenzen besitzen. In der vorliegenden Arbeit verwende ich eine konstante Steifigkeit von $K = 845000 \text{ N} \cdot \text{m}^{-3}$. Die ersten drei Resonanzfrequenzen, die ich mit diesen Parametern aus der Eingangsimpedanz des subglottalen Systems ermittelt habe, betragen 651, 1313 und 1983 Hz. Damit weichen sie von den in [53] gemessenen Werten um maximal 7% ab.

3.5 Der Sprechapparat als akustisches Netzwerk

3.5.1 Örtliche Diskretisierung der Grundgleichungen

Die Grundgleichungen aus Abschnitt 3.3 lauten zusammengefasst

$$-\frac{\partial p}{\partial x} = \frac{\rho_0}{A} \cdot \frac{\partial u}{\partial t} + \frac{\rho_0}{2} \cdot \frac{\partial}{\partial x} \left(\frac{u^2}{A^2} \right) \quad (\text{Bewegungsgleichung}), \quad (3.33)$$

$$\frac{\partial p}{\partial t} = \frac{\rho_0 \cdot c^2}{A} \cdot \left[-\frac{\partial u}{\partial x} - S \cdot \frac{\partial y}{\partial t} \right] \quad (\text{Kontinuitätsgleichung}), \quad (3.34)$$

$$p = M \cdot \frac{\partial^2 y}{\partial t^2} + B \cdot \frac{\partial y}{\partial t} + K \cdot y. \quad (\text{Wandschwingungsgleichung}) \quad (3.35)$$

In der Kontinuitätsgleichung wurde die Schalldichte ρ mit Hilfe der Zustandsgleichung durch p/c^2 ersetzt. Damit reduzieren sich die unbekannten Variablen in den Gleichungen auf den Volumenstrom $u(x, t)$, den Schalldruck $p(x, t)$ und die Auslenkung der Rohrwand $y(x, t)$. Für die numerische Simulation müssen diese Größen entlang der Rohrachse abgetastet werden. Dazu bedienen wir uns der Unterteilung des Rohrmodells des Sprechapparats in kurze Rohrabschnitte mit jeweils konstantem Querschnitt (siehe Abschnitt 2). Die Abtastung der Funktionen $p(x, t)$ und $y(x, t)$ erfolgt jeweils in der Mitte der einzelnen Rohrabschnitte, und der Volumenstrom $u(x, t)$ wird jeweils am Ein- bzw. Ausgang der Rohrabschnitte abgetastet. Rohrabschnitte werden im Folgenden durch einen tiefgestellten Index gekennzeichnet (meistens i , j oder k). Für den Schalldruck und die Wandauslenkung in einem Abschnitt i schreiben wir somit $p_i(t)$ und $y_i(t)$. Die Länge, der Querschnitt und der Umfang des Abschnitts sind l_i , A_i und S_i . Der Volumenstrom, der von einem Abschnitt i in einen Abschnitt j fließt, wird als $u_{ij}(t)$ bezeichnet. Abbildung 3.6 verdeutlicht die Abtastung anhand der drei benachbarten Rohrabschnitte i , j und k .

Benachbarte Rohrabschnitte werden bewusst nicht mit fortlaufenden Indizes wie i und $i+1$ versehen, da diese Zuordnung an Rohrverzweigungen unvorteilhaft wäre. Weiterhin werden Volumenströme, die in einen Abschnitt i hineinfließen, für die der Ausgangspunkt aber keine Rolle spielt, als $u_{e,i}$ bezeichnet. Analog steht $u_{a,i}$ für den Volumenstrom, der aus dem Abschnitt i austritt, für den das Ziel aber keine Rolle spielt.

Ausgehend von diesen Überlegungen können nun die Grundgleichungen örtlich diskretisiert werden. Die diskrete Bewegungsgleichung erhält man, wenn man Gl. (3.33) mit dem differenziellen Wegelement dx multipliziert und über den Weg zwischen den Mittelpunkten zweier benachbarter Rohrabschnitte integriert:

$$-\int_{x_i+l_i/2}^{x_j+l_j/2} \frac{\partial p}{\partial x} \cdot dx = \int_{x_i+l_i/2}^{x_j+l_j/2} \frac{\rho_0}{A} \cdot \frac{\partial u}{\partial t} \cdot dx + \frac{\rho_0}{2} \int_{x_i+l_i/2}^{x_j+l_j/2} \frac{\partial}{\partial x} \left(\frac{u^2}{A^2} \right) \cdot dx.$$

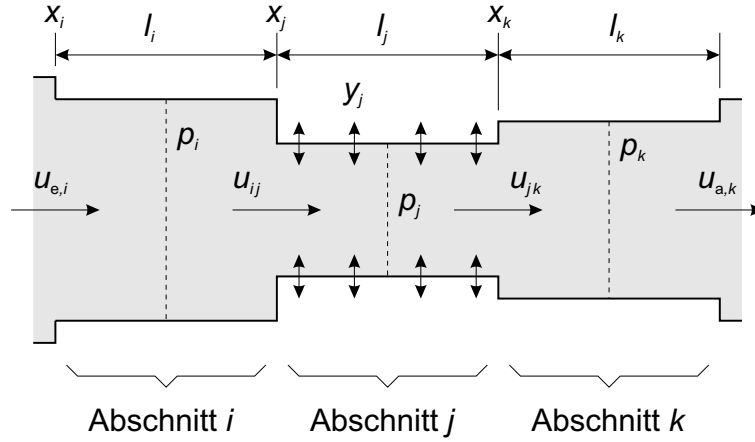


Abbildung 3.6: Örtliche Abtastung von Schalldruck p , Volumenstrom u und Auslenkung y der Rohrwand.

Zwischen den Integrationsgrenzen $x_i + l_i/2$ und $x_j + l_j/2$ ist der Volumenstrom nur eine Funktion der Zeit, weil dort genau ein örtlicher Abtastwert liegt. Deshalb gilt innerhalb der Integrationsgrenzen $u(x, t) \equiv u_{ij}(t)$ und $\partial u(x, t)/\partial t \equiv \dot{u}_{ij}(t)$. Damit kann man die Gleichung als

$$-\int_{x_i+l_i/2}^{x_j+l_j/2} \frac{\partial p}{\partial x} \cdot dx = \dot{u}_{ij} \cdot \int_{x_i+l_i/2}^{x_j+l_j/2} \frac{\rho_0}{A} \cdot dx + \frac{\rho_0}{2} \cdot u_{ij}^2 \cdot \int_{x_i+l_i/2}^{x_j+l_j/2} \frac{\partial}{\partial x} \left(\frac{dx}{A^2} \right)$$

umschreiben und dann die Integration ausführen. Das Ergebnis lautet

$$p_i - p_j = \dot{u}_{ij} \cdot \rho_0 \cdot \left(\frac{l_i}{2 \cdot A_i} + \frac{l_j}{2 \cdot A_j} \right) + u_{ij}^2 \cdot \frac{\rho_0}{2} \cdot \left(\frac{1}{A_i^2} - \frac{1}{A_j^2} \right). \quad (3.36)$$

Die örtliche Diskretisierung der Kontinuitätsgleichung (3.34) erfolgt analog zur Bewegungsgleichung durch die Multiplikation mit dem Wegelement dx und der Integration zwischen den beiden Enden des Rohrabschnitts i in Abb. 3.6:

$$\int_{x_i}^{x_j} \frac{\partial p}{\partial t} \cdot dx = \int_{x_i}^{x_j} \frac{\rho_0 \cdot c^2}{A} \cdot \left[-\frac{\partial u}{\partial x} - S \cdot \frac{\partial y}{\partial t} \right] \cdot dx.$$

Zwischen den Integrationsgrenzen ist der Schalldruck nur eine Funktion der Zeit, da sich dort genau ein Abtastwert befindet. Daher gilt $p(x, t) \equiv p_i(t)$ und $\partial p(x, t)/\partial t \equiv \dot{p}_i(t)$. Auch die Wandauslenkung wird für $x_i \leq x \leq x_j$ nur durch den einen Abtastwert $y_i(t)$ repräsentiert, so dass $\partial y(x, t)/\partial t \equiv \dot{y}_i(t)$. Weiterhin sind der Querschnitt $A(x)$ und der Umfang $S(x)$ für den betrachteten Rohrabschnitt gleich A_i bzw. S_i , so dass sich die folgende vereinfachte Gleichung ergibt:

$$\dot{p}_i \cdot \int_{x_i}^{x_j} 1 \cdot dx = \frac{\rho_0 \cdot c^2}{A_i} \cdot \int_{x_i}^{x_j} \left[-\frac{\partial u}{\partial x} \cdot dx - S_i \cdot \dot{y}_i \cdot dx \right].$$

Führt man die Integration aus und berücksichtigt, dass $x_j - x_i = l_i$, so erhält man

$$\dot{p}_i \cdot l_i = \frac{\rho_0 \cdot c^2}{A_i} \cdot (u_{e,i} - u_{a,i} - S_i \cdot l_i \cdot \dot{y}_i).$$

Das Produkt $l_i \cdot S_i$ ist genau die Mantelfläche des Rohrabchnitts und $\dot{y}_i(t)$ die Geschwindigkeit, mit der sich die Wand senkrecht zu ihrer Oberfläche bewegt. Der Ausdruck $S_i \cdot l_i \cdot \dot{y}_i(t)$ ist daher der Volumenstrom, der durch das „Atmen“ der Wand entsteht. Wenn man ihn mit $u_{w,i}(t)$ bezeichnen und beide Seiten der Gleichung durch l_i dividieren, so folgt daraus

$$\dot{p}_i = \frac{\rho_0 \cdot c^2}{A_i \cdot l_i} \cdot (u_{e,i} - u_{a,i} - u_{w,i}). \quad (3.37)$$

Die Wandschwingungsgleichung (3.35) für den gemeinsamen Abtastwert in der Mitte des Rohrabchnittes i lautet

$$p_i = M_i \cdot \ddot{y}_i + B_i \cdot \dot{y}_i + K_i \cdot y_i,$$

wobei M_i , B_i und K_i die Parameter der mechanische Wandimpedanz an dieser Stelle sind. Wenn nun auch hier $y_i(t)$ durch den zur Wand gerichteten Volumenstrom $u_{w,i}(t) = S_i \cdot l_i \cdot \dot{y}_i(t)$ ausgedrückt wird, so lautet die Schwingungsgleichung

$$p_i = \frac{M_i}{S_i \cdot l_i} \cdot \dot{u}_{w,i} + \frac{B_i}{S_i \cdot l_i} \cdot u_{w,i} + \frac{K_i}{S_i \cdot l_i} \cdot \int_0^t u_{w,i} \cdot dt. \quad (3.38)$$

Die ortsdiskreten Gleichungen (3.36), (3.37) und (3.38) können jetzt wie folgt zusammengefasst werden:

$$p_i - p_j = \dot{u}_{ij} \cdot (L_i + L_j) + u_{ij} \cdot (\bar{R}_{\text{bern},i} + R_{\text{bern},j}), \quad (3.39)$$

$$\dot{p}_i = \frac{1}{C_i} \cdot (u_{e,i} - u_{a,i} - u_{w,i}), \quad (3.40)$$

$$p_i = L_{w,i} \cdot \dot{u}_{w,i} + R_{w,i} \cdot u_{w,i} + \frac{1}{C_{w,i}} \cdot \int_0^t u_{w,i} \cdot dt. \quad (3.41)$$

Darin wurden die folgenden Abkürzungen verwendet:

$$\begin{aligned} C_i &= A_i \cdot l_i / (\rho_0 \cdot c^2), & C_{w,i} &= S_i \cdot l_i / K_i, \\ L_i &= \rho_0 \cdot l_i / (2 \cdot A_i), & L_{w,i} &= M_i / (S_i \cdot l_i), \\ R_{\text{bern},i} &= |u_{e,i}| \cdot \rho_0 / (2 \cdot A_i^2), & R_{w,i} &= B_i / (S_i \cdot l_i), \\ \bar{R}_{\text{bern},i} &= -|u_{a,i}| \cdot \rho_0 / (2 \cdot A_i^2). \end{aligned} \quad (3.42)$$

Anhand von Abb 3.7 ist ersichtlich, dass die Beziehungen zwischen Schalldruck und Volumenstrom in den Gleichungen (3.39) bis (3.41) exakt mit den Beziehungen zwischen Spannung und Strom an einer inhomogenen elektrischen Übertragungsleitung mit konzentrierten Elementen übereinstimmen. Diese Übereinstimmung beruht auf einer grundsätzlichen Analogie zwischen elektrischen und akustischen Systemen [10]. Sie rührt von der gleichen Art der mathematischen Gleichungen her, mit denen solche Systeme beschrieben werden. Dadurch lässt sich die Theorie elektrischer Netzwerke weitestgehend auf akustische Probleme anwenden. Die Voraussetzung dafür ist, dass ein akustisches System wie der Vokaltrakt so wie in Abb. 3.7 schematisch in Form eines Netzwerks repräsentiert wird. Der elektrische Strom entspricht in dieser Analogie dem Volumenstrom u und die elektrische Spannung dem Schalldruck p . Spule, Kondensator und Widerstand repräsentieren die *akustische Masse* (Induktivität), die *akustische Federung* (Kapazität) und den *akustischen Widerstand*. Analog zu elektrischen Schaltkreisen gelten für sie die Gleichungen

$$p(t) = L \cdot \frac{du(t)}{dt}, \quad p(t) = \frac{1}{C} \cdot \int u(t) \cdot dt \quad \text{und} \quad p(t) = R \cdot u(t),$$

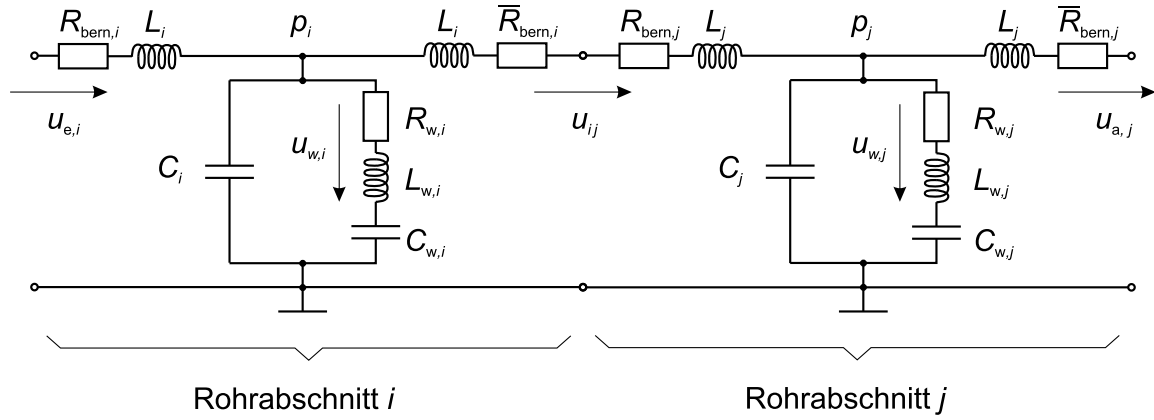


Abbildung 3.7: Zwei Abschnitte einer inhomogenen elektrischen Übertragungsleitung mit konzentrierten Elementen. Die Strom-Spannungs-Beziehungen sind den Volumenstrom-Schalldruck-Beziehungen in einer akustischen Rohrleitung mit stückweise konstantem Querschnitt analog.

wobei $p(t)$ der Druck(-abfall) über einem Element ist und $u(t)$ der Volumenstrom durch das Element. Die Netzwerkelemente in Abb. 3.7 können demnach wie folgt interpretiert werden: C_i repräsentiert die Nachgiebigkeit bzw. die Kompressibilität der Luft im Rohrabschnitt i , und die L_i 's repräsentieren die Masseträgheit der Luft in jeweils einer Hälfte des Abschnitts i . Die Widerstände $R_{\text{bern},i}$ bzw. $\bar{R}_{\text{bern},i}$ verursachen den in Abschnitt 3.4.2 diskutierten Bernoulli-Effekt. Die Komponenten $R_{w,i}$, $L_{w,i}$ und $C_{w,i}$ repräsentieren die mechanische Impedanz der elastischen Rohrwände.

3.5.2 Verluste und Energiequellen

Die Gleichungen (3.39) bis (3.41) bzw. das Netzwerk in Abb. 3.7 berücksichtigen bisher weder dissipative Verluste durch Reibung oder Wirbelbildung noch die Verluste durch Schallabstrahlung an den offenen Rohrenden. Diese Verlustmechanismen wurden in Abschnitt 3.4 diskutiert und werden hier nun in die Netzwerkstruktur integriert.

Das in Abschnitt 3.4.2 vorgestellte Verfahren zur Berücksichtigung von Wirbelverlusten sieht vor, dass an jedem Übergang von einem engen in einen weiten Rohrquerschnitt die überschüssige kinetische Druckenergie $(\rho_0/2) \cdot (v_1^2 - v_2^2)$ dissipativ umgesetzt wird. Das heißt, dass bei einer Rohrerweiterung keine Druckerhöhung erfolgt, die im verlustfreien Fall aufgrund des Bernoulli-Effekts zu erwarten wäre. In der Bewegungsgleichung und dem akustischen Netzwerk müssen dazu die Widerstände $\bar{R}_{\text{bern},i}$ und $R_{\text{bern},j}$ auf Null gesetzt werden. Die Bewegungsgleichung (3.39) lautet für diesen Fall

$$p_i - p_j = \dot{u}_{ij} \cdot (L_i + L_j).$$

Für eine stationäre Strömung gilt folglich $p_i = p_j$.

Die Verluste durch Reibung an der Rohrwand werden in der Bewegungsgleichung durch zwei zusätzliche Widerstände $R_{\text{reib},i}$ und $R_{\text{reib},j}$ erfasst:

$$p_i - p_j = \dot{u}_{ij} \cdot (L_i + L_j) + u_{ij} \cdot (\bar{R}_{\text{bern},i} + R_{\text{bern},j} + R_{\text{reib},i} + R_{\text{reib},j}).$$

$R_{\text{reib},i}$ ist der Widerstand der rechten Hälfte des Abschnitts i und $R_{\text{reib},j}$ der Widerstand der linken Hälfte des Abschnitts j . Entsprechend den Überlegungen in Abschnitt 3.4.1 werden sie wie folgt berechnet:

$$R_{\text{reib},i} = l_i \cdot \frac{2 \cdot \mu \cdot (a_i^2 + b_i^2)}{\pi \cdot a_i^3 \cdot b_i^3}.$$

Im akustischen Netzwerk erscheinen sie als zwei zusätzliche Längswiderstände in Reihe mit den Induktivitäten. In dem T-Glied für einen einzelnen Rohrabschnitt in Abb. 3.8 wurden die Reibungswiderstände mit den Widerständen $R_{\text{bern},i}$ und $\bar{R}_{\text{bern},j}$ zusammengefasst:

$$R_i = R_{\text{bern},i} + R_{\text{reib},i} \quad (3.43)$$

$$\bar{R}_i = \bar{R}_{\text{bern},i} + R_{\text{reib},i}. \quad (3.44)$$

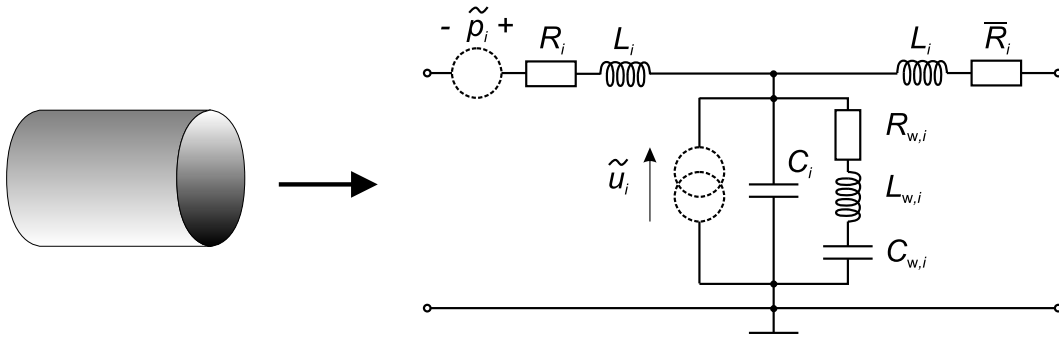


Abbildung 3.8: Vollständiges Vierpol-Netzwerk zu einem einfachen Rohrabschnitt.

Damit sich die Luft im Vokaltrakt überhaupt bewegt, muss ihr Energie zugeführt werden. In dem akustischen Netzwerk kann die Energieversorgung entweder durch Schalldruck- oder Volumenstromquellen erfolgen. Um den subglottalen Druck zu erzeugen, wird beispielsweise eine Druckquelle im ersten Rohrabschnitt der Lunge eingesetzt. Außerdem wird die sekundäre Geräuscherzeugung durch supraglottale Rauschquellen simuliert (mehr dazu in Abschnitt 3.8). Da die Position supraglottaler Rauschquellen nicht konstant ist, habe ich die Möglichkeit vorgesehen, in jedem Rohrabschnitt i im Bedarfsfall eine Schalldruckquelle mit der Amplitude $\tilde{p}_i$ und eine Volumenstromquelle mit der Amplitude $\tilde{u}_i$ einzusetzen. In der Netzwerkdarstellung erscheint die Schalldruckquelle als Reihenelement am Eingang des Rohrabschnitts und die Volumenstromquelle in einem Querzweig in der Mitte des Abschnitts (siehe Abb. 3.8). Dadurch bleibt für $\tilde{p}_i = 0$ und $\tilde{u}_i = 0$ die ursprüngliche Struktur des Netzwerks erhalten. Außerdem lassen sich die Quellen auf diese Weise leicht in der Bewegungsgleichung und der Kontinuitätsgleichung erfassen. Mit allen hier gemachten Ergänzungen lauten die Gleichungen

$$\begin{aligned} p_i - p_j &= \dot{u}_{ij} \cdot (L_i + L_j) + u_{ij} \cdot (R_{\text{reib},i} + R_{\text{reib},j} + \bar{R}_{\text{bern},i} + R_{\text{bern},j}) - \tilde{p}_j, \\ \dot{p}_i &= \frac{1}{C_i} \cdot (u_{e,i} - u_{a,i} - u_{w,i} + \tilde{u}_i). \end{aligned}$$

Der jeweils letzte Rohrabschnitt des Mund- und Nasenraums ist durch eine Strahlungsimpedanz abgeschlossen, die den Energieverlust durch die Abstrahlung von Schallwellen

in die Umgebung repräsentieren. Die Formel für die Strahlungsimpedanz wurde in Abschnitt 3.4.3 diskutiert. Für die Behandlung der Impedanz in der zeitdiskreten Simulation ist es erforderlich, sie mit konzentrierten Netzwerkelementen (Spule, Kondensator, Widerstand) zu approximieren. Laut FLANAGAN [38] lässt sich die Impedanz nach Gl. (3.32) für die wichtigen Frequenzen bis 4 kHz mit hoher Genauigkeit durch die Parallelschaltung von einem Widerstand R_{rad} und einer Induktivität L_{rad} annähern. Die Formeln für R_{rad} und L_{rad} lauten

$$R_{\text{rad}} = \frac{128 \cdot \varrho_0 \cdot c}{9 \cdot \pi^2 \cdot A} \quad \text{und} \quad L_{\text{rad}} = \frac{8 \cdot \varrho_0}{3 \cdot \pi \cdot \sqrt{\pi \cdot A}}. \quad (3.45)$$

3.5.3 Netzwerkstruktur des gesamten Sprechapparats

Das akustische Netzwerk für den gesamten Sprechapparat entsteht durch die Zusammenschaltung der Vierpole für alle Rohrabschnitte des Rohrmodells und ist schematisch in Abb. 3.9 dargestellt. Das Netzwerk beginnt links mit den Rohrverbundabschnitten des

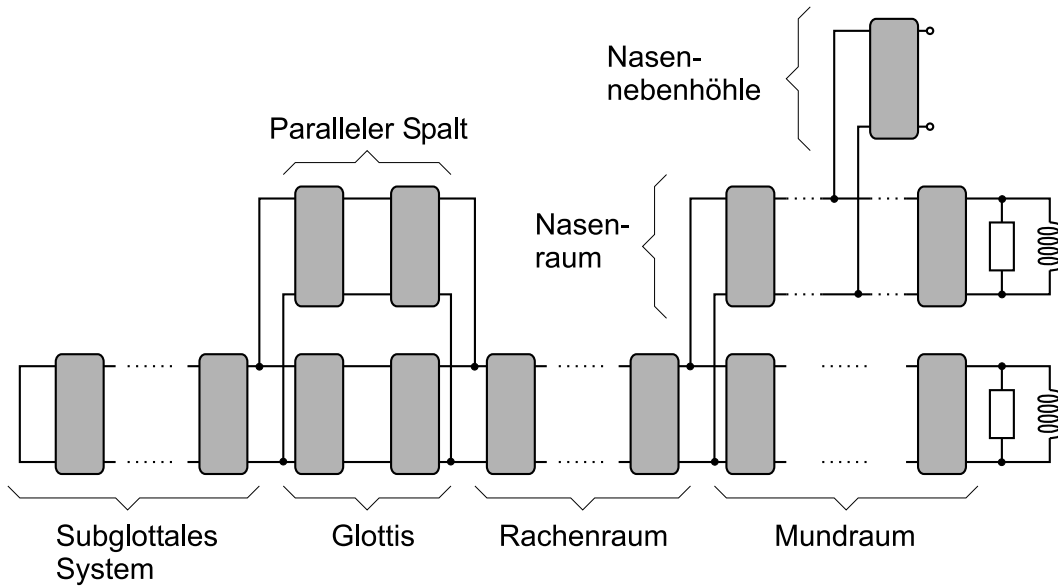


Abbildung 3.9: Struktur des Netzwerks für den gesamten Sprechapparat.

subglottalen Systems. Von denen ist der erste Abschnitt, der die feinste Verzweigungsstufe des Bronchialbaums repräsentiert, mit einem Kurzschluss abgeschlossen. Streng genommen müsste hier statt dessen die akustische Kapazität C_0 der Alveolen angeschlossen werden. Jedoch ist das Gesamtvolumen V_0 der Lungenbläschen von etwa 3150 ml so groß, dass der Blindwiderstand der Kapazität $C_0 = V_0/(\varrho_0 \cdot c^2)$ vernachlässigt werden kann. Am Ausgang des obersten subglottalen Abschnitts verzweigen sich die Luftwege in zwei Teilrohre – die eigentliche Stimmritze und einen parallelen Spalt, der unter bestimmten Umständen zwischen den Stellknorpeln entsteht. Jeder Zweig wird durch einen unteren und einen oberen Rohrabschnitt repräsentiert. Wenn kein paralleler Spalt vorhanden ist, dann wird die akustische Impedanz dieses Zweigs automatisch so hoch, dass der Volumenstrom durch ihn versiegt. Die oberen Abschnitte der Glottis und des parallelen Spalts münden gemeinsam

in den untersten Abschnitt des Rachens. Der Rachen verzweigt sich an seinem Ende wiederum in den Mundraum und den Nasenraum, die jeweils durch eine Strahlungsimpedanz aus Widerstand und Spule abgeschlossen sind (siehe Gl. (3.45)). Innerhalb des Nasenraums existieren zusätzlich noch die Abzweige zu den vier Nasennebenhöhlen. Die Rohrabschnitte des Nasenraums, an denen die Nebenhöhlen abzweigen, sind in Tab. 2.4 angegeben.

Grundsätzlich verbirgt sich hinter jedem der Vierpole in Abb. 3.9 ein Netzwerk wie in Abb. 3.8. Für die Nasennebenhöhlen und die Rohrverbundabschnitte des Bronchialbaums weichen die Netzwerkkomponenten jedoch von denen eines „einfachen“ Rohrabschnitts ab. Als nächstes werden die Netzwerkkomponenten eines einfachen Rohrabschnitts zusammengefasst und anschließend die Komponenten der Nasennebenhöhlen und Rohrverbundabschnitte diskutiert.

Einfacher Rohrabschnitt. Für einen einfachen Rohrabschnitt, der als Teil der Luftröhre, der Glottis und des Rachen-, Mund- und Nasenraums vorkommt, gilt:

$$\begin{aligned}
 C_i &= A_i \cdot l_i / (\varrho_0 \cdot c^2) & C_{w,i} &= S_i \cdot l_i / K_i \\
 L_i &= \varrho_0 \cdot l_i / (2 \cdot A_i) & L_{w,i} &= M_i / (S_i \cdot l_i) \\
 R_{\text{reib},i} &= l_i \cdot 2 \cdot \mu \cdot (a_i^2 + b_i^2) / (\pi \cdot a_i^3 \cdot b_i^3) & R_{w,i} &= B_i / (S_i \cdot l_i) \\
 R_{\text{bern},i} &= \begin{cases} |u_{e,i}| \cdot \varrho_0 / (2 \cdot A_i^2) & \text{wenn } A_i < A_{\text{vor}} \\ 0 & \text{sonst} \end{cases} \\
 \bar{R}_{\text{bern},i} &= \begin{cases} -|u_{a,i}| \cdot \varrho_0 / (2 \cdot A_i^2) & \text{wenn } A_{\text{nach}} < A_i \\ 0 & \text{sonst} \end{cases} \\
 M_i &= 21 \text{ kg/m}^2 \\
 B_i &= 800 \text{ kg/(m}^2 \cdot \text{s)} \\
 K_i &= 845000 \text{ N/m}^3
 \end{aligned} \tag{3.46}$$

Dabei sind a_i und b_i die Halbachsen des elliptischen Querschnitts des Rohrabschnitts i , und A_{vor} und A_{nach} stehen für die Querschnittsflächen des vorangegangenen bzw. nachfolgenden Rohrabschnitts. Wenn i der letzte Rohrabschnitt des Mund- oder Nasenraums ist, so ist A_{nach} stets größer als A_i , so dass $\bar{R}_{\text{bern},i} = 0$. Wenn sich der Rohrabschnitt i nach links oder rechts verzweigt, so sind A_{vor} oder A_{nach} jeweils die Summe der Querschnittsflächen der Zweigrohre.

Nasennebenhöhle. Jede Nasennebenhöhle wird in dieser Arbeit analog zu [25, 106] wie ein diskreter Helmholtz-Resonator behandelt. Geometrisch ist jeder Helmholtz-Resonator i durch die Halslänge l_i , den Halsquerschnitt A_i und das Hohlraumvolumen V_i gekennzeichnet (siehe Abb. 3.10 und Tab. 2.4). Aus akustischer Sicht entspricht der Resonator im einfachsten Fall einer R - L - C -Reihenschaltung. R_i und L_i repräsentieren den Reibungswiderstand und die akustische Masse des Resonatorhalses, und C_i repräsentiert die akustische Federung der Luft im Resonatorhohlraum. Für die Netzwerkelemente gelten die folgenden Formeln:

$$R_i = l_i \cdot \frac{8 \cdot \pi \cdot \mu}{A_i^2} \quad L_i = \frac{\varrho_0 \cdot l_i}{A_i} \quad C_i = \frac{V_i}{\varrho_0 \cdot c^2}.$$

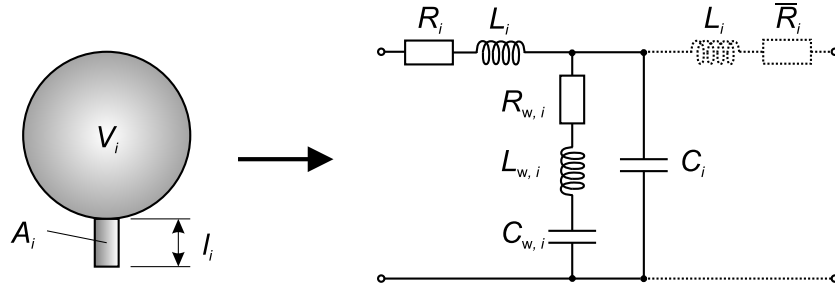


Abbildung 3.10: Nasennebenhöhle und entsprechendes akustisches Netzwerk.

Der Widerstand $R_{\text{bern},i}$ fällt für die Nasennebenhöhlen weg, da sie nur einen Eingang, aber keinen Ausgang besitzen, und sie somit nur rein akustisch wirksam sind. Die Impedanz der elastischen Hohlraumwand eines Resonators kann parallel zu C_i geschaltet werden. Für den mechanischen Widerstand der Wand pro Flächeneinheit wird von DANG *et al.* [25] der Wert $B_i = 65000 \text{ kg}/(\text{s} \cdot \text{m}^2)$ vorgeschlagen, der auch in dieser Arbeit verwendet wird. Für die Steifigkeit und die Masse verwende ich dieselben Werte wie für den restlichen Sprechtrakt, d. h. $K_i = 845000 \text{ N}/\text{m}^3$ und $M_i = 21 \text{ kg}/\text{m}^2$. Die Komponenten der Wandimpedanz sind folglich

$$R_{w,i} = B_i/A_{O,i} \quad L_{w,i} = M_i/A_{O,i} \quad C_{w,i} = A_{O,i}/K_i,$$

wobei $A_{O,i} = 4 \cdot \pi \cdot [3 \cdot V_i/(4 \cdot \pi)]^{2/3}$ die Oberfläche des kugelförmigen Hohlraums ist.

Rohrverbundabschnitt. Bei den subglottalen Luftwegen handelt es sich um ein stark verzweigtes Luftkanalsystem, welches in Abschnitt 2.6 durch 34 aneinandergereihte Rohrverbundabschnitte modelliert wurde. Rohrverbundabschnitte unterscheiden sich von den normalen Rohrabschnitten dadurch, dass sie statt eines einzelnen Rohrstücks jeweils ein Bündel aus mehreren gleichartigen, parallel verlaufenden Rohrstücken repräsentieren. Diese Situation ist schematisch in Abb. 3.11 dargestellt.

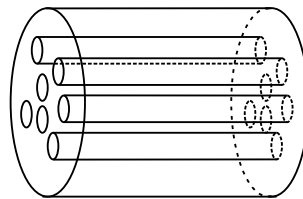


Abbildung 3.11: Schematische Darstellung eines Rohrverbundabschnitts.

Jeder Rohrverbundabschnitt i ist durch eine Gesamtquerschnittsfläche $A_{\text{ges},i}$, eine Länge l_i und die Anzahl paralleler Einzelrohre n_i gekennzeichnet. Die Querschnittsfläche eines Einzelrohres aus einem Verbund ist folglich $A_{\text{ges},i}/n_i$. Der Einfachheit halber wird angenommen, dass sich alle Einzelrohre eines Rohrverbundabschnitts exakt gleich verhalten. Die Ströme $u_{e,i}$ und $u_{a,i}$ sind jeweils die Summe aller Teilvolumenströme, die durch die linke bzw. rechte Stirnseite des Verbundabschnitts fließen, und p_i ist der Schalldruck, der gleichzeitig in der Mitte aller Einzelrohre herrscht. Die Netzwerkstruktur eines Rohrverbundabschnitts ist dieselbe wie in Abb. 3.8. Die Netzwerkkomponenten unterscheiden sich

jedoch von denen eines einfachen Rohrabschnitts und können wie folgt zusammengefasst werden:

$$\begin{aligned} C_i &= A_i \cdot n_i \cdot l_i / (\varrho_0 \cdot c^2), & C_{w,i} &= S_i \cdot l_i \cdot n_i / K_i, \\ L_i &= \varrho_0 \cdot l_i / (2 \cdot A_i \cdot n_i), & L_{w,i} &= M_i / (S_i \cdot l_i \cdot n_i), \\ R_i &= 4 \cdot \pi \cdot \mu \cdot l_i / (A_i^2 \cdot n_i), & R_{w,i} &= B_i / (S_i \cdot l_i \cdot n_i) \end{aligned} \quad (3.47)$$

Dabei ist $A_i = A_{\text{ges},i}/n_i$ der Querschnitt und $S_i = 2 \cdot \sqrt{\pi \cdot A_i}$ der Umfang eines Einzelrohres. Die akustische Federung C_i und Masse L_i entsprechen genau der eines einfachen Rohrabschnitts mit der Querschnittsfläche $A_{\text{ges},i}$. Der Widerstand R_i ergibt sich aus der Parallelschaltung der Widerstände für die n_i (halben) Einzelrohre. Die Elemente der Wandimpedanz $C_{w,i}$, $L_{w,i}$ und $R_{w,i}$ berücksichtigen die n_i -fach größere Mantelfläche gegenüber einem einzelnen Rohr aus dem Verbund. Die Widerstände $R_{\text{bern},i}$ und $\bar{R}_{\text{bern},i}$, die für den Bernoulli-Effekt verantwortlich sind, habe ich bei den subglottalen Luftwegen vernachlässigt, da hier kein Stoßverlust durch eine plötzliche Rohrerweiterung zu erwarten ist. Wenn n_i wie im Fall der Luftröhre gleich Eins ist, so entsprechen die Netzwerkkomponenten denen eines einfachen Zylinderrohres. Die Parameter $C_{w,i}$, $L_{w,i}$ und $R_{w,i}$ der elastischen Wände des subglottalen Systems sind in Abschnitt 3.4.4 angegeben.

3.6 Simulation im Frequenzbereich

In dieser Arbeit wurden zwei Verfahren zur Simulation des Vokaltrakt-Netzwerks umgesetzt. Ein Verfahren operiert im Zeitbereich und wird in Abschnitt 3.7 besprochen. Das andere Verfahren operiert im Frequenzbereich und wird in diesem Abschnitt vorgestellt. Beide Verfahren haben individuelle Stärken und Schwächen. Die Simulation im Frequenzbereich eignet sich besonders für eine Analyse der spektralen Eigenschaften stationärer Sprachlaute. Die Methode ist sehr schnell und sehr genau, da frequenzabhängige Netzwerkkomponenten, wie der Grenschichtwiderstand, berücksichtigt werden können. Nichtlineare Druckverluste sind dagegen schwieriger zu realisieren. Die Simulationsmethode im Zeitbereich eignet sich besonders für die dynamische Sprachsynthese. Dabei werden die Volumenstrom- und Schalldruckwerte im Vokaltrakt zeitdiskret durch ein iteratives Verfahren berechnet. Der Vokaltrakt kann dabei stetig seine Form verändern und die Berücksichtigung nichtlinearer Druckverluste ist problemlos möglich. Durch die zeitdiskrete Approximation und die Verwendung frequenzunabhängiger Rohrreibungswiderstände entstehen jedoch Fehler, die gegenüber der Frequenzbereichssimulation zu kleinen Abweichungen der Formantfrequenzen und -bandbreiten führen.

Grundsätzlich ist das Ziel beider Simulationsverfahren die Erzeugung bzw. Charakterisierung des Sprachschalls, der bei einer bestimmten Form und Anregung des Sprechtrakts entsteht. Als Sprachschall wird normalerweise der Schalldruck $p_{\text{rad}}(t)$ bezeichnet, der in einem gewissen Abstand von der Mundöffnung durch das Ohr eines Hörers oder durch ein Mikrofon registriert wird. $p_{\text{rad}}(t)$ kann man berechnen, wenn man die Mundöffnung und gegebenenfalls die Nasenlöcher als *Punktschallquellen* approximiert, von denen aus die Schallwellen in die Umgebung abgestrahlt werden. Für die interessierenden Frequenzen bis ca. 4 kHz ist diese Näherung hinreichend genau. Eine Punktschallquelle, die den Volumenstrom $u_0(t)$ abgibt, erzeugt in der Entfernung r von der Quelle laut STEVENS [110] den Schalldruck

$$p_{\text{rad}}(t) = \frac{\varrho_0}{4 \cdot \pi \cdot r} \cdot \frac{\partial u_0(t - r/c)}{\partial t}. \quad (3.48)$$

Der abgestrahlte Schalldruck ist also proportional zur zeitlichen Ableitung des von der Quelle abgegebenen Volumenstroms zur Zeit $t - r/c$ und umgekehrt proportional zum Abstand r . Wenn Schalldruck und Volumenstrom mit Hilfe komplexer, frequenzabhängiger Zeigergrößen ausgedrückt werden, d. h. $p_{\text{rad}}(t) = P_{\text{rad}}(\omega) \cdot e^{j \cdot \omega \cdot t}$ und $u_0(t) = U_0(\omega) \cdot e^{j \cdot \omega \cdot t}$, so erhält man

$$P_{\text{rad}}(\omega) = \frac{j \cdot \omega \cdot \varrho_0}{4 \cdot \pi \cdot r} \cdot U_0(\omega) \cdot e^{-j \cdot \omega \cdot r/c}.$$

Die Funktion

$$R(\omega) = \frac{P_{\text{rad}}(\omega)}{U_0(\omega)} = \frac{j \cdot \omega \cdot \varrho_0}{4 \cdot \pi \cdot r} \cdot e^{-j \cdot \omega \cdot r/c} \quad (3.49)$$

wird als Strahlungscharakteristik bezeichnet. Der abgestrahlte Schalldruck hängt also direkt vom Volumenstrom durch die Mund- und Nasenöffnung ab. Die Volumenstromschwankungen an diesen Stellen entstehen wiederum durch die akustische Filterung der Schallquellen innerhalb des Vokaltrakts. Im akustischen Netzwerk wurde in jedem Rohrabschnitt der Platz für eine Schalldruckquelle und eine Volumenstromquelle vorgesehen. Der Beitrag jeder Schallquelle zum Volumenstrom an der Mund- und Nasenöffnung kann mit Hilfe frequenzabhängiger Übertragungsfunktionen bestimmt werden. Die Berechnung dieser Übertragungsfunktionen für die Schalldruck- und Volumenstromquellen wird in diesem Abschnitt erläutert. Am Ende des Abschnitts wird auf zwei spezielle Anwendungen der Analyse im Frequenzbereich eingegangen:

- (1) Die Berechnung der Vokaltrakt-Übertragungsfunktion zwischen der Glottis und der Mund- und Nasenöffnung. Diese werden in Abschnitt 3.7.4 für die Abschätzung der Fehler herangezogen, die bezüglich der Formantfrequenzen und -bandbreiten durch die *zeitdiskrete Simulation* gemacht werden.
- (2) Die schnelle und genaue Berechnung der Schalldruckspektren simulierter stimmloser Frikative. Damit wird in Abschnitt 3.8 die Platzierung und Parametrisierung von Geräuschquellen für die Synthese von Frikativen untersucht.

3.6.1 Theoretische Grundlagen

Ein effizienter Zugang zur Analyse des Vokaltrakts im Frequenzbereich ist die Theorie der Zweitore bzw. Vierpole, die analog für elektrische und akustische Netzwerke gilt. Ein Zweitor ist ein beliebig kompliziertes Netzwerk oder System, das ein Eingangstor und ein Ausgangstor mit jeweils zwei Anschlussklemmen besitzt. Die Beschreibung des Systems

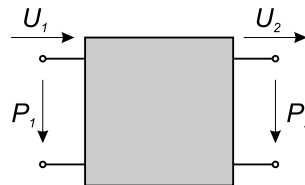


Abbildung 3.12: Schematische Darstellung eines akustischen Vierpols.

erfolgt durch die Beziehung zwischen seinen Eingangs- und Ausgangsgrößen. Bei einem elektrischen Vierpol sind die Eingangs- und Ausgangsgrößen jeweils Strom und Spannung, und bei einem akustischen Netzwerk analog dazu Volumenstrom und Schalldruck. Abbildung 3.12 zeigt einen akustischen Vierpol mit den Eingangsgrößen $U_1(\omega)$ und $P_1(\omega)$ und

den Ausgangsgrößen $U_2(\omega)$ und $P_2(\omega)$. In Hinblick auf eine kettenartige Verbindung bzw. Hintereinanderschaltung der Zweitore beschreibt man die Beziehung zwischen den vier Größen in der Form

$$\begin{pmatrix} P_1 \\ U_1 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \cdot \begin{pmatrix} P_2 \\ U_2 \end{pmatrix} = K \cdot \begin{pmatrix} P_2 \\ U_2 \end{pmatrix}.$$

$A(\omega)$, $B(\omega)$, $C(\omega)$ und $D(\omega)$ sind komplexe, frequenzabhängige Funktionen, die das System vollständig charakterisieren. Die Matrix K wird als Kettenmatrix bezeichnet.

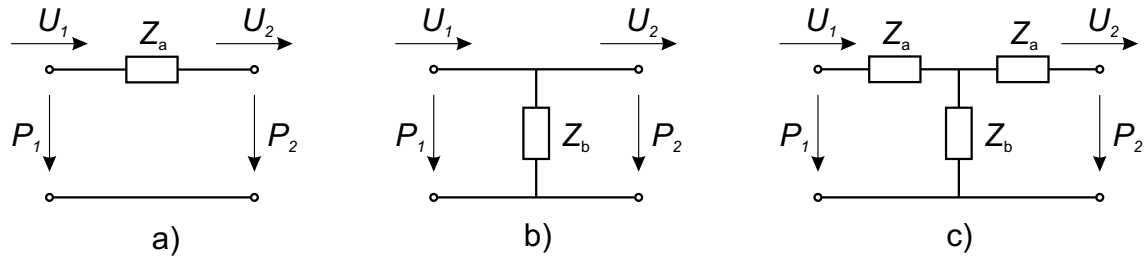


Abbildung 3.13: Verschiedene einfache Vierpole.

Abbildung 3.13 zeigt drei häufig vorkommende Vierpol-Netzwerke, für die die Kettenmatrizen einfach zu berechnen sind. Für den Vierpol mit einer einzelnen Seriennimpedanz Z_a gelten die Beziehungen

$$\begin{aligned} P_2 &= P_1 - U_2 \cdot Z_a \\ U_1 &= U_2, \end{aligned}$$

so dass sich die Matrix

$$K_a = \begin{pmatrix} 1 & Z_a \\ 0 & 1 \end{pmatrix} \quad (3.50)$$

ergibt. Für den Vierpol mit einer einzelnen Querimpedanz Z_b gilt dagegen

$$\begin{aligned} P_1 &= P_2 \\ U_1 &= U_2 + P_2/Z_b. \end{aligned}$$

Die Kettenmatrix ist für diesen Fall

$$K_b = \begin{pmatrix} 1 & 0 \\ 1/Z_b & 1 \end{pmatrix}. \quad (3.51)$$

Das T-Glied in Abb. 3.13 (c) kann man als Verkettung von drei einzelnen Vierpolen mit den Kettenmatrizen K_a , K_b und K_a betrachten. Die Kettenmatrix des T-Glieds erhält man durch die Multiplikation der Einzelmatrizen, also

$$\begin{pmatrix} P_1 \\ U_1 \end{pmatrix} = \begin{pmatrix} 1 & Z_a \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 1/Z_b & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & Z_a \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} P_2 \\ U_2 \end{pmatrix}$$

Die resultierende Matrix für den Vierpol lautet

$$K_T = \begin{pmatrix} 1 + Z_a/Z_b & 2 \cdot Z_a + Z_a^2/Z_b \\ 1/Z_b & 1 + Z_a/Z_b \end{pmatrix}. \quad (3.52)$$

Im allgemeinen Fall ist bei der Verkettung von Vierpolen die Reihenfolge der Matrizenmultiplikation entscheidend, die mit der Reihenfolge der Verkettung übereinstimmen muss. Wenn also das Ausgangstor eines Systems mit der Matrix K_1 mit dem Eingangstor eines Systems mit der Matrix K_2 verbunden wird, so ist die Matrix des zusammengesetzten Systems $K_1 \cdot K_2$. Für ein zusammengesetztes System aus N verketteten Zweitoren mit den Matrizen $K_1 \dots K_N$ gilt analog

$$K_{\text{ges}} = \prod_{i=1}^N K_i.$$

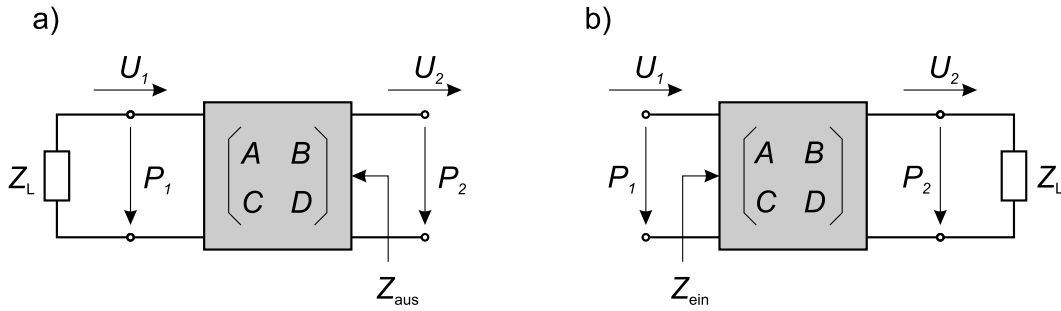


Abbildung 3.14: Definition der Eingangs- und Ausgangsimpedanz.

Ist das Eingangstor eines Systems wie in Abb. 3.14 (a) mit einer Impedanz Z_L verbunden, so ist die *Ausgangsimpedanz* des Systems

$$Z_{\text{aus}} = \frac{P_2}{-U_2} = \frac{D \cdot Z_L + B}{C \cdot Z_L + A}. \quad (3.53)$$

Ist ein Zweitor-System wie in Abb. 3.14 (b) am Ausgang mit einer Impedanz Z_L belastet, so ist die Impedanz seines Eingangstores (*Eingangsimpedanz*)

$$Z_{\text{ein}} = \frac{P_1}{U_1} = \frac{A \cdot Z_L + B}{C \cdot Z_L + D}. \quad (3.54)$$

Die *Übertragungsfunktion* eines Systems ist das Verhältnis der Frequenzfunktion einer Ausgangsgröße zur Frequenzfunktion einer Eingangsgröße. Da je zwei Eingangs- und Ausgangsgrößen existieren, gibt es vier Übertragungsfunktionen. An dieser Stelle interessiert uns aber nur die Übertragungsfunktion zwischen dem Volumenstrom U_1 am Eingangstor und dem Volumenstrom U_2 am Ausgangstor. Die Gleichung hierfür lautet

$$H = \frac{U_2}{U_1} = \frac{1}{C \cdot Z_L + D}. \quad (3.55)$$

3.6.2 Vierpolanalyse des Netzwerks für den Sprechapparat

Bei der Simulation des Netzwerks für den Sprechapparat im Frequenzbereich wurden gegenüber dem Netzwerk in Abschnitt 3.5.3 zwei Änderungen vorgenommen. Einerseits wurde für die Reibungswiderstände die frequenzabhängige Formel (3.21) für den Grenzschichtwiderstand verwendet, da diese der realen Situation besser entspricht als die Formel für die stationäre Strömung. Somit gilt

$$R_{\text{reib},i} = \frac{l_i \cdot S_i}{2 \cdot A_i^2} \cdot \sqrt{\frac{\omega \cdot \varrho_0 \cdot \mu}{2}}.$$

Andererseits wurden die Widerstände $R_{\text{bern},i}$ und $\bar{R}_{\text{bern},i}$ im gesamten Netzwerk vernachlässigt, da diese Komponenten selbst Funktionen des Volumenstroms sind und dadurch im Frequenzbereich nicht berechnet werden können. Wie bereits erläutert wurde, sind sie für eine korrekte Simulation der aerodynamischen Stoßverluste notwendig. Für einen statischen Vokaltrakt, wie er im Frequenzbereich untersucht wird, können die Stoßverluste jedoch bei Bedarf in Form *lokaler* Druckabfälle simuliert werden (siehe weiter unten). Mit diesen Modifikationen kann jeder Rohrabchnitt durch ein *symmetrisches* T-Glied repräsentiert werden, für das die Kettenmatrix nach Gl. (3.52) gilt. Die Widerstände Z_a und Z_b sind entsprechend Abb. 3.8

$$\begin{aligned} Z_{a,i} &= R_i + j \cdot \omega \cdot L_i \\ Z_{b,i} &= \frac{1}{j \cdot \omega \cdot C_i + 1/[R_{w,i} + j \cdot \omega \cdot L_{w,i} + 1/(j \cdot \omega \cdot C_{w,i})]} \end{aligned}$$

Wir wenden uns nun der Berechnung der Übertragungsfunktion von einer Volumenstromquelle $\tilde{U}(\omega)$ oder Schalldruckquelle $\tilde{P}(\omega)$ innerhalb des Vokaltrakts zum Volumenstrom $U_M(\omega)$ an der Mundöffnung oder zum Volumenstrom $U_N(\omega)$ an den Nasenlöchern zu. Wir betrachten nur Schallquellen, die sich an der Glottis, im Rachen oder im Mundraum befinden. Wenn der Nasenraum durch ein angehobenes Gaumensegel vom Mundraum und dem Rachen abgekoppelt ist, so braucht jeweils nur die Übertragungsfunktion von der Quelle zur Mundöffnung betrachtet werden, also $U_M/\tilde{P}$ für eine Schalldruckquelle und $U_M/\tilde{U}$ für eine Volumenstromquelle. Bei einem abgesenkten Gaumensegel kommt die Übertragungsfunktion zwischen der Quelle und der Nasenöffnung hinzu, also $U_N/\tilde{P}$ bzw. $U_N/\tilde{U}$. Zur Berechnung einer Übertragungsfunktion ist es zweckmäßig, die Rohrabchnitte des Vokaltrakts hinter der Quelle (in Richtung Lunge) in einer Kettenmatrix K_h zusammenzufassen und die Rohrabchnitte vor der Quelle (in Richtung Mund- oder Nasenöffnung) in einer Matrix K_v . Die Zusammenfassung erfolgt durch die Multiplikation der Matrizen der einzelnen Rohrabchnitte entlang des betrachteten Schallpfades. Rohrverzweigungen müssen bei der Zusammenfassung der Kettenmatrizen gesondert berücksichtigt werden. Wenn beispielsweise die Übertragungsfunktion $U_M/\tilde{U}$ für eine Volumenstromquelle an der Glottis berechnet werden soll und der weiche Gaumen abgesenkt ist, so muss der Abzweig in den Nasenraum mit Hilfe einer Kopplungsmatrix erfasst werden. Die Kopplungsmatrix entspricht derjenigen für eine einzelne Querimpedanz nach Gl. (3.51), wobei die Querimpedanz genau die Eingangsimpedanz des abzweigenden Nasenraums ist. Die Gesamtmatrix für den Schallpfad zwischen der Quelle und der Mundöffnung ist somit

$$K_v = K_{\text{Rachen}} \cdot \begin{pmatrix} 1 & 0 \\ 1/Z_{\text{ein}} & 1 \end{pmatrix} \cdot K_{\text{Mund}},$$

wobei K_{Rachen} und K_{Mund} die zusammengefassten Matrizen für den Rachen und den Mundraum sind und Z_{ein} die Eingangsimpedanz des Nasenraums ist. Wenn die Rohrab-schnitte des Nasenraums zu einer Matrix

$$K_{\text{Nase}} = \begin{pmatrix} A_{\text{Nase}} & B_{\text{Nase}} \\ C_{\text{Nase}} & D_{\text{Nase}} \end{pmatrix}$$

zusammengefasst wurden, so gilt nach Gl. (3.54)

$$Z_{\text{ein}} = \frac{A_{\text{Nase}} \cdot Z_{\text{rad}} + B_{\text{Nase}}}{C_{\text{Nase}} \cdot Z_{\text{rad}} + D_{\text{Nase}}},$$

wobei $Z_{\text{rad}} = R_{\text{rad}} \parallel j \cdot \omega \cdot L_{\text{rad}}$ die Strahlungsimpedanz der Nasenöffnung ist. Wenn die Übertragungsfunktion $U_N/\tilde{U}$ für eine Volumenstromquelle an der Glottis berechnet werden soll, so ergibt sich K_v analog durch die Verkettung der Matrix für den Rachenraum, die Kopplungsmatrix mit der Eingangsimpedanz des Mundraums und der Matrix für den Nasenraum. Auch bei der Berechnung der Gesamtmatrix K_{Nase} für den Nasenraum müssen die Abzweige zu den vier Nasennebenhöhlen durch Kopplungsmatrizen berücksichtigt werden. Die Eingangsimpedanz einer Nasennebenhöhle mit der Matrix

$$K_{\text{Sinus}} = \begin{pmatrix} A_{\text{Sinus}} & B_{\text{Sinus}} \\ C_{\text{Sinus}} & D_{\text{Sinus}} \end{pmatrix}$$

ist

$$Z_{\text{ein}} = \frac{A_{\text{Sinus}}}{C_{\text{Sinus}}},$$

da sie durch einen Leerlauf bzw. eine unendlich große Lastimpedanz abgeschlossen ist.

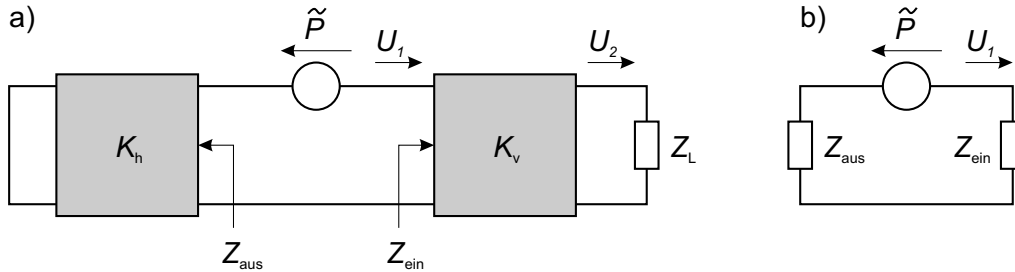


Abbildung 3.15: Repräsentation einer Schalldruckquelle $\tilde{P}(\omega)$ innerhalb des Netzwerks.

Die Berechnung der Übertragungsfunktion von einer Schalldruckquelle zum Volumenstrom an der Mund- oder Nasenöffnung wird anhand von Abb. 3.15 erläutert. K_h und K_v sind die Matrizen der Hohlräume hinter und vor der Quelle, und U_2 steht entweder für U_M oder U_N . Der Volumenstrom U_1 in den vorderen Hohlraum kann mit Hilfe der Ausgangsimpedanz des hinteren und der Eingangsimpedanz des vorderen Hohlraums als

$$U_1 = \frac{\tilde{P}}{Z_{\text{ein}} + Z_{\text{aus}}}$$

geschrieben werden. Mit Gl. (3.55) erhält man somit für die gesuchte Übertragungsfunktion

$$\frac{U_2}{\tilde{P}} = \frac{1}{C \cdot Z_L + D} \cdot \frac{1}{Z_{\text{ein}} + Z_{\text{aus}}}. \quad (3.56)$$

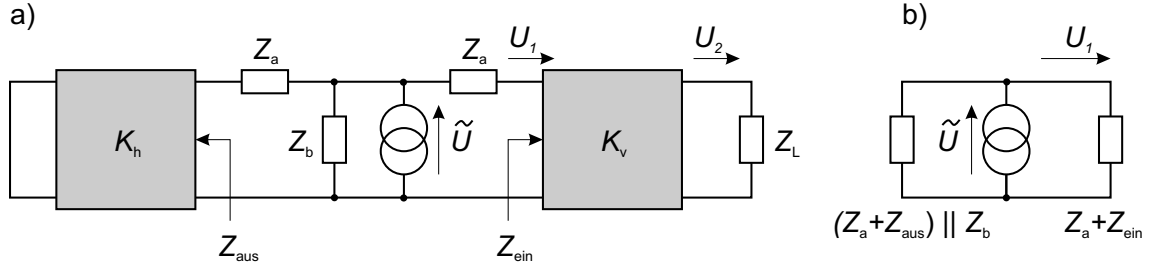


Abbildung 3.16: Repräsentation einer Volumenstromquelle $\tilde{U}(\omega)$ innerhalb des Netzwerks.

Für eine Volumenstromquelle $\tilde{U}$ ist die Situation in Abb. 3.16 dargestellt. Eine Volumenstromquelle befindet sich immer in der Mitte eines Rohrabchnitts, wo sie parallel zur Querimpedanz Z_b angeordnet ist. Der Abschnitt mit der Quelle ist in Abb. 3.16 daher nicht Bestandteil der Matrix K_v , sondern direkt durch seine Impedanzen Z_a und Z_b repräsentiert. Der Volumenstrom U_1 in den vorderen Hohlraum mit der Matrix K_v ergibt sich nach der Stromteilerregel wie folgt:

$$U_1 = \tilde{U} \cdot \frac{(Z_a + Z_{aus}) \parallel Z_b}{(Z_a + Z_{aus}) \parallel Z_b + Z_a + Z_{ein}}.$$

Mit Gl. (3.55) kann daraus wieder die gesuchte Übertragungsfunktion gebildet werden:

$$\frac{U_2}{\tilde{U}} = \frac{1}{C_v \cdot Z_L + D_v} \cdot \frac{(Z_a + Z_{aus}) \parallel Z_b}{(Z_a + Z_{aus}) \parallel Z_b + Z_a + Z_{ein}}. \quad (3.57)$$

Übertragungsfunktion von Vokalen. Die Übertragungsfunktion von Vokalen kann unmittelbar mit Gl. (3.57) berechnet werden, sofern sich die Quelle $\tilde{U}$ im ersten Rohrabchnitt über der Glottis befindet. Die davor liegenden Rohrabchnitte des Vokaltrakts müssen zur Matrix K_v zusammengefasst werden und die Rohrabchnitte des subglottalen Systems und der Glottis selbst zur Matrix K_h . Meistens werden Übertragungsfunktionen von Vokalen ohne Berücksichtigung der subglottalen Luftwege berechnet, da die Glottis bei modaler Phonation eine sehr hohe Impedanz besitzt. In dem vorgestellten Modell ergibt sich diese hohe Impedanz automatisch, wenn die Querschnittsflächen der Rohrabchnitte der Glottis auf sehr kleine Werte gesetzt werden (im Grenzfall 0.0 cm^2). Alternativ kann man direkt die Ausgangsimpedanz Z_{aus} auf einen unendlich hohen Wert setzen, was einer Abkopplung des subglottalen Systems gleichkommt. Mit der vorgestellten Methode ist es aber auch möglich, die Auswirkungen der Kopplung zwischen dem Vokaltrakt und dem subglottalen System auf die Übertragungsfunktion zu untersuchen.

Schalldruckspektrum von stimmlosen Frikativen. Stimmlose Frikative entstehen durch die Verwirbelung des Luftstroms an einer supraglottalen Konstriktion. Damit ein genügend hoher Luftstrom entsteht, ist das Gaumensegel angehoben und die Glottis weit geöffnet. Die durch die Verwirbelung verursachten Geräusche können durch ein oder mehrere Schalldruck- oder Volumenstrom-Rauschquellen in der Nähe der Konstriktion simuliert werden (siehe Abschnitt 3.8). Jede Rauschquelle ist durch eine individuelle spektrale Zusammensetzung gekennzeichnet (Quellspektrum) und liefert aufgrund des

Überlagerungsprinzips einen bestimmten additiven Beitrag zum Volumenstrom U_M bzw. zum abgestrahlten Schalldruck P_{rad} . Wenn jede beteiligte Rauschquelle durch einen Index i gekennzeichnet wird, und $Q_i(\omega)$ und $T_i(\omega)$ das Quellspektrum und die Übertragungsfunktion für die Quelle i sind, so kann das abgestrahlte Schalldruckspektrum als

$$P_{\text{rad}}(\omega) = R(\omega) \cdot \sum_i Q_i(\omega) \cdot T_i(\omega)$$

geschrieben werden, wobei $R(\omega)$ die Abstrahlungscharakteristik nach Gl. (3.49) ist. Für eine Volumenstromquelle ist $T_i = U_M/\tilde{U}$ (siehe Gl. (3.57)) und für eine Schalldruckquelle gilt $T_i = U_M/\tilde{P}$ (siehe Gl. (3.56)).

Bei der Simulation der Frikative darf man die Druckverluste durch die Verwirbelung an der glottalen und supraglottalen Konstriktion nicht vernachlässigen. Diese können, da es sich um eine statische Rohrgeometrie handelt, der Einfachheit halber als punktuelle Verluste behandelt werden. Die Druckverluste an den Konstriktionen sind näherungsweise

$$\Delta p_g = \frac{\varrho_0}{2} \cdot \left(\frac{u_g}{A_g} \right)^2 \quad \text{und} \quad \Delta p_c = \frac{\varrho_0}{2} \cdot \left(\frac{u_c}{A_c} \right)^2,$$

wobei der Index „g“ für die glottale Konstriktion und der Index „c“ für die supraglottale Konstriktion steht. Da der Volumenstrom u quadratisch in die Druckverluste eingeht, lassen sie sich im Frequenzbereich nicht direkt durch lineare Widerstände in das Vokaltrakt-Netzwerk integrieren. Man kann jedoch eine lineare Approximation der Widerstände vornehmen. Dazu nähert man die quadratischen Gleichungen für die Druckverluste durch ihre Tangenten in einem „Arbeitspunkt“ an, der durch den mittleren Volumenstrom $\bar{u}$ durch den Vokaltrakt gegeben ist. Die differentiellen Kleinsignalwiderstände sind allgemein

$$R = \left. \frac{\partial(\Delta p)}{\partial u} \right|_{\text{Arbeitspunkt}} = \left. \frac{\partial [\varrho_0 \cdot u^2 / (2 \cdot A^2)]}{\partial u} \right|_{u=\bar{u}} = \varrho_0 \cdot \frac{\bar{u}}{A^2}$$

bzw. konkret

$$R_g = \varrho_0 \cdot \frac{\bar{u}}{A_g^2} \quad \text{und} \quad R_c = \varrho_0 \cdot \frac{\bar{u}}{A_c^2}.$$

Den mittleren Volumenstrom $\bar{u}$ erhält man mit dem Ansatz $p_{\text{sub}} = \Delta p_c + \Delta p_g$, wobei p_{sub} ein vorgegebener subglottaler Druck ist (z. B. 800 Pa). Daraus folgt

$$\bar{u} = \sqrt{\frac{2 \cdot p_{\text{sub}}}{\varrho_0 \cdot (1/A_g^2 + 1/A_c^2)}}. \quad (3.58)$$

R_g und R_c müssen als Längswiderstände direkt oberhalb der Glottis und am Ende der supraglottalen Konstriktion in das Vokaltrakt-Netzwerk eingefügt werden. Bei der Berechnung der Matrizen K_h und K_v können beide Widerstände als zusätzliche Vierpole in der Kette der Rohrabschnitte berücksichtigt werden, deren Kettenmatrix durch Gl. (3.50) gegeben ist.

3.7 Simulation im Zeitbereich

Dieser Abschnitt behandelt das Verfahren für die aerodynamisch-akustische Simulation im Zeitbereich, das ich für die Synthese zusammenhängender Äußerungen umgesetzt habe. So wie die Simulation im Frequenzbereich basiert auch dieses Verfahren auf dem akustischen Netzwerk, welches in Abschnitt 3.5 entwickelt wurde.

3.7.1 Methoden der zeitdiskreten Approximation

Allgemeine Betrachtung. Die Grundlage für die Simulation im Zeitbereich ist die Abtastung der Schalldruck- und Volumenstromfunktionen im Vokaltrakt zu den Zeitpunkten $n \cdot \Delta t$, wobei $n = 0, 1, 2, \dots$ der Abtastindex und Δt der Zeitschritt ist. Der Funktionswert einer Funktion $f(t)$ zum Zeitpunkt $n \cdot \Delta t$ wird dabei als $f[n]$ geschrieben. Die Simulation besteht darin, iterativ für jeden Abtastzeitpunkt alle relevanten Schalldruck- und Volumenstromwerte im Netzwerk zu berechnen. Dazu ist eine zeitdiskrete Approximation der Netzwerkgleichungen erforderlich. In den Gleichungen kommen sowohl zeitliche Ableitungen in der Form $\dot{f}(t)$ als auch Integrale über die Zeit in der Form $\int_0^t f(t) \cdot dt$ vor, die durch numerische Verfahren angenähert werden müssen.

Eine einfache Möglichkeit zur numerischen Integration besteht darin, das Integral einer Funktion $f(t)$ zwischen zwei aufeinanderfolgenden Abtastpunkten $n - 1$ und n durch eine Rechteckfläche mit der Breite Δt und der Höhe $f[n - 1]$ oder $f[n]$ zu approximieren, also

$$\int_{(n-1)\Delta t}^{n\Delta t} f(t) \cdot dt \approx \Delta t \cdot f[n - 1] \quad (3.59)$$

oder

$$\int_{(n-1)\Delta t}^{n\Delta t} f(t) \cdot dt \approx \Delta t \cdot f[n]. \quad (3.60)$$

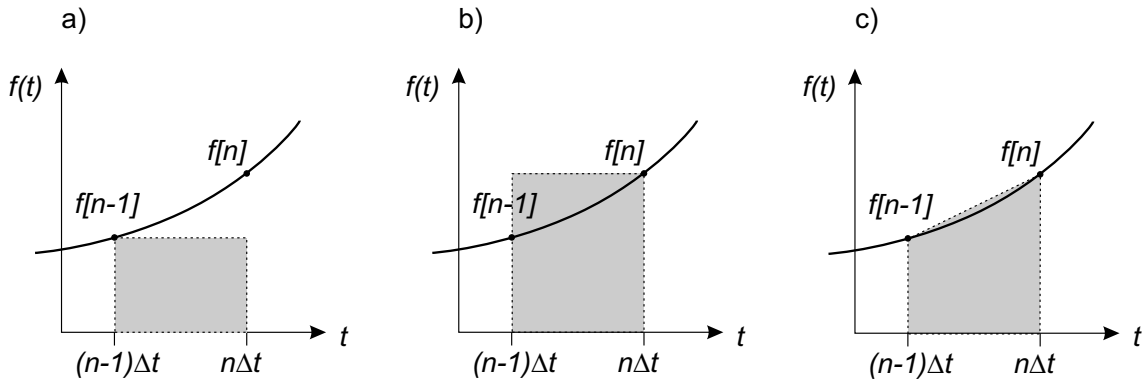


Abbildung 3.17: Näherungsverfahren für die zeitliche Diskretisierung der Differenzialgleichungen.

Diese zwei Fälle sind in Abb. 3.17 (a) und (b) dargestellt. Eine genauere Methode besteht darin, das Integral zwischen den Abtastpunkten $n - 1$ und n durch eine Trapezfläche nach der Formel

$$\int_{(n-1)\Delta t}^{n\Delta t} f(t) \cdot dt \approx \frac{\Delta t}{2} \cdot (f[n - 1] + f[n]) \quad (3.61)$$

anzunähern, wie es in Abb. 3.17 (c) dargestellt ist. Weiterhin gibt es die Möglichkeit, die Funktion $f(t)$ zwischen den beiden Abtastwerten durch eine Funktion zweiter oder höherer Ordnung zu approximieren. Diese Methoden sind aber für den hier betrachteten Fall unzuverlässig. Genau genommen sind die Gleichungen (3.59) bis (3.61) nur Spezialfälle der *allgemeinen* Gleichung

$$\int_{(n-1)\cdot\Delta t}^{n\cdot\Delta t} f(t) \cdot dt \approx \Delta t \cdot [(1 - \theta) \cdot f[n - 1] + \theta \cdot f[n]], \quad (3.62)$$

wobei der Parameter θ zwischen 0 und 1 variieren kann. Die Fälle $\theta = 0$, $\theta = 1$ und $\theta = 0.5$ entsprechen den Gleichungen (3.59), (3.60) und (3.61). Wenn man in Gl. (3.62) alle Vorkommen der Funktion f durch ihre Ableitung $df/dt = \dot{f}$ ersetzt, so erhält man damit ein allgemeines Verfahren zur Approximation der Ableitung. Die Substitution $f \rightarrow df/dt$ ergibt

$$\int_{(n-1)\cdot\Delta t}^{n\cdot\Delta t} \frac{df(t)}{dt} \cdot dt = f[n] - f[n - 1] \approx \Delta t \cdot [(1 - \theta) \cdot \dot{f}[n - 1] + \theta \cdot \dot{f}[n]]. \quad (3.63)$$

Die Umstellung nach $\dot{f}[n]$ führt zu

$$\dot{f}[n] = \frac{1}{\Delta t \cdot \theta} \cdot (f[n] - f[n - 1]) - \frac{1 - \theta}{\theta} \cdot \dot{f}[n - 1]. \quad (3.64)$$

Die Ableitung an der Stelle n kann also mit Hilfe des aktuellen Abtastwerts $f[n]$ und den zurückliegenden Abtastwerten $f[n - 1]$ und $\dot{f}[n - 1]$ approximiert werden. Zwei häufig in der Literatur beschriebene Näherungen für den Differenzialquotienten $\dot{f}[n]$ sind der *vordere Differenzenquotient* $(f[n + 1] - f[n])/\Delta t$ und der *hintere Differenzenquotient* $(f[n] - f[n - 1])/\Delta t$. Beide Differenzenquotienten ergeben sich unmittelbar aus Gl. (3.63) für die Spezialfälle $\theta = 0$ und $\theta = 1$.

Zeitdiskrete Approximation in anderen Studien. Von ISHIZAKA *et al.* [51] und MAEDA [72] wurden zeitdiskrete Simulationsmethoden für akustische Netzwerke beschrieben, die eine ähnliche Struktur wie das Netzwerk in dieser Arbeit besitzen. ISHIZAKA *et al.* verwenden für die Approximation der Differenzialquotienten hintere Differenzenquotienten und nähern die Integrale entsprechend Gl. (3.59) bzw. Abb. 3.17 (a) an. MAEDA verwendet in seiner Studie für die Integration die Trapezregel und nähert die Differenzialquotienten dementsprechend nach Gl. (3.64) mit $\theta = 0.5$ an. In den Arbeiten [3] und [106] werden Differenzialgleichungen numerisch approximiert, um den Volumenstrom durch die Glottis zu simulieren. In beiden Studien werden zeitliche Ableitungen mit Hilfe von hinteren Differenzenquotienten berechnet. BOERSMA [14] verwendet für seine Simulation des Vokaltrakts ein *mehrstufiges* Zeitschrittverfahren, welches die Integration und Differentiation in jedem Zeitschritt mit Hilfe von Zwischengrößen annähert.

Zeitdiskrete Approximation in dieser Arbeit. In dieser Arbeit verwende ich zur Approximation der Integrale und Differenzialquotienten die Gleichungen (3.62) und (3.64) mit dem Diskretisierungsparameter θ . Obwohl man davon ausgehen kann, dass der Diskretisierungsfehler für $\theta = 0.5$ am geringsten ist, habe ich mir trotzdem die Möglichkeit offengehalten, die Simulation mit anderen Werten für θ zu untersuchen. Wie

in Abschnitt 3.7.4 gezeigt wird, ist tatsächlich der Wert $\theta = 0.51$ in Hinblick auf die Formantbandbreiten günstiger als 0.5.

Mit den Abkürzungen $Q[n] = \int_0^{n \cdot \Delta t} f(t) \cdot dt$ und $\bar{\theta} = 1 - \theta$ lauten die Näherungsformeln

$$\begin{aligned} Q[n] &= Q[n-1] + \Delta t \cdot (\bar{\theta} \cdot f[n-1] + \theta \cdot f[n]), \\ \dot{f}[n] &= \frac{1}{\Delta t \cdot \theta} \cdot (f[n] - f[n-1]) - \frac{\bar{\theta}}{\theta} \cdot \dot{f}[n-1]. \end{aligned}$$

Wenn man zur weiteren Vereinfachung der Schreibweise für aktuelle Abtastwerte $f[n] \equiv f$ schreibt und für die unmittelbar zurückliegenden Werte $f[n-1] \equiv f'$ (analog für Q und $\dot{f}$), so erhält man

$$Q = Q' + \Delta t \cdot (\bar{\theta} \cdot f' + \theta \cdot f), \quad (3.65)$$

$$\dot{f} = \frac{1}{\Delta t \cdot \theta} \cdot (f - f') - \frac{\bar{\theta}}{\theta} \cdot \dot{f}'. \quad (3.66)$$

3.7.2 Diskretisierung der Netzwerkgleichungen

Der Kern der Simulation besteht darin, dass in jedem Zeitschritt ein lineares Gleichungssystem (LGS) für die Volumenströme, die in die einzelnen Rohrabschnitte hinein- und aus ihnen herausfließen, aufgestellt und gelöst wird. Konkret handelt es sich dabei um die einzelnen Zweigströme u_{ij} , die jeweils vom Netzwerkknoten in der Mitte eines Rohrabschnitts i zum Knoten in der Mitte des folgenden Rohrabschnitts j fließen (siehe Abb. 3.7). Rohrverzweigungen stellen jeweils zwei gesuchte Zweigströme des LGS. Auch die offenen Rohrenden stellen jeweils zwei unbekannte Zweigströme: einen durch den Strahlungswiderstand R_{rad} und einen durch die Strahlungsinduktivität L_{rad} . Diese Zweigströme werden im folgenden mit u_{iL} und u_{iR} bezeichnet, wobei i der letzte Rohrabschnitt des Mund- oder Nasenraums ist. Der letzte unbekannte Zweigstrom des LGS ist der Strom, der in den ersten Rohrabschnitt i der Lunge fließt und mit u_{Ai} bezeichnet wird.

In diesem Abschnitt werden die Gleichungen des LGS für die gesuchten Zweigströme formuliert. Dazu müssen zunächst die Drücke an den Verbindungsknoten in der Mitte der einzelnen Rohrabschnitte selbst als Funktion der jeweiligen Zweigströme an den Knoten ausgedrückt werden. Dies erreicht man durch die folgende Diskretisierung und Zusammenfassung der Kontinuitätsgleichung und der Wandschwingungsgleichung.

Die Kontinuitätsgleichung und die Wandschwingungsgleichung für einen Rohrabschnitt i wurden in den vorangegangenen Abschnitten abgeleitet und lauten

$$\dot{p}_i = \frac{1}{C_i} \cdot (u_{e,i} - u_{a,i} - u_{w,i} + \tilde{u}_i) \quad (3.67)$$

$$p_i = L_{w,i} \cdot \dot{u}_{w,i} + R_{w,i} \cdot u_{w,i} + \frac{1}{C_{w,i}} \cdot Q_{w,i}, \quad (3.68)$$

wobei $Q_{w,i} = \int_0^t u_{w,i} \cdot dt$. Diese beiden Gleichungen werden nun zu einer einzelnen zeitdiskreten Gleichung kombiniert. Dazu werden zunächst $\dot{u}_{w,i}$ und $Q_{w,i}$ in (3.68) durch die zeitdiskreten Näherungen ersetzt und die Gleichung dann nach $u_{w,i}$ umgestellt. Das Ergebnis lautet

$$u_{w,i} = p_i \cdot \alpha_{w,i} + \beta_{w,i}, \quad (3.69)$$

wobei

$$\begin{aligned}\alpha_{w,i} &= 1 / \left(\frac{L_{w,i}}{\Delta t \cdot \theta} + R_{w,i} + \frac{\Delta t \cdot \theta}{C_{w,i}} \right), \\ \beta_{w,i} &= \alpha_{w,i} \cdot \left[L_{w,i} \cdot \left(\frac{u'_{w,i}}{\Delta t \cdot \theta} + \frac{\bar{\theta}}{\theta} \cdot \dot{u}'_{w,i} \right) - \frac{Q'_{w,i}}{C_{w,i}} - \frac{\Delta t \cdot \bar{\theta}}{C_{w,i}} \cdot u'_{w,i} \right].\end{aligned}$$

Wenn man Gl. (3.69) in die Kontinuitätsgleichung (3.67) einsetzt, darin die Ableitung $\dot{p}_i$ durch die numerische Näherungsformel ersetzt und nach p_i umstellt, so erhält man

$$p_i = E_i \cdot u_{e,i} - E_i \cdot u_{a,i} + D_i \quad (3.70)$$

mit

$$E_i = 1 / \left(\alpha_{w,i} + \frac{C_i}{\Delta t \cdot \theta} \right) \quad \text{und} \quad D_i = E_i \cdot \left(\frac{C_i}{\Delta t \cdot \theta} \cdot p'_i + C_i \cdot \frac{\bar{\theta}}{\theta} \cdot \dot{p}'_i - \beta_{w,i} + \tilde{u}_i \right).$$

Der Druck p_i in der Mitte eines Rohrabschnitts i kann somit also als Linearkombination der Ströme geschrieben werden, die in den Abschnitt hinein und aus ihm heraus fließen. Als nächstes stellen wir das LGS für die Berechnung der gesuchten Zweigströme auf. Dazu wird für jeden Zweigstrom genau eine Netzwerkgleichung formuliert, in der die Beziehung zu jeweils benachbarten Strömen hergestellt wird. Je nachdem, für welchen der Zweigströme die Gleichung aufgestellt wird, muss man verschiedene Fälle unterscheiden.

Fall 1: Einfacher Strom ohne Abzweig. Für den Strom u_{ij} , der vom Abschnitt i in den Abschnitt j fließt, ist die Netzwerkgleichung genau die Bewegungsgleichung

$$p_i - p_j = \dot{u}_{ij} \cdot L_{ij} + u_{ij} \cdot R_{ij} - \tilde{p}_j,$$

wobei $L_{ij} = L_i + L_j$ und $R_{ij} = \bar{R}_i + R_j$. Wenn man die Ableitung des Volumenstroms durch die zeitdiskrete Näherung ersetzt, so erhält man die Gleichung

$$p_i - p_j = u_{ij} \cdot G + H$$

mit

$$G = \frac{L_{ij}}{\Delta t \cdot \theta} + R_{ij} \quad \text{und} \quad H = -L_{ij} \cdot \left(\frac{u'_{ij}}{\Delta t \cdot \theta} + \dot{u}'_{ij} \cdot \frac{\bar{\theta}}{\theta} \right) - \tilde{p}_j.$$

Wenn man weiterhin p_i und p_j jeweils durch Gl. (3.70) ersetzt, so erhält man als Gleichung für das LGS

$$u_{e,i} \cdot E_i + u_{ij} \cdot (-E_i - E_j - G) + u_{a,j} \cdot E_j = H + D_j - D_i.$$

Fall 2: Strom in den untersten subglottalen Rohrabschnitt. Der Strom u_{Aj} von den Lungenbläschen in den ersten subglottalen Rohrabschnitt j fließt nur durch die Netzwerkkomponenten L_j und R_j . Die Netzwerkgleichung lautet daher

$$0 - p_j = \dot{u}_{Aj} \cdot L_j + u_{Aj} \cdot R_j - \tilde{p}_j,$$

wobei $\tilde{p}_j = p_{\text{Lunge}}$ die pulmonale Druckquelle ist. Wenn man die Ableitung des Volumens durch die zeitdiskrete Näherung ersetzt, so erhält man die Gleichung

$$0 - p_j = u_{Aj} \cdot G + H$$

mit

$$G = \frac{L_j}{\Delta t \cdot \theta} + R_j \quad \text{und} \quad H = -L_j \cdot \left(\frac{u'_{Aj}}{\Delta t \cdot \theta} + \dot{u}'_{Aj} \cdot \frac{\bar{\theta}}{\theta} \right) - \tilde{p}_j.$$

Wenn man weiterhin p_j durch Gl. (3.70) ersetzt, so erhält man für das LGS die Gleichung

$$u_{Aj} \cdot (-E_j - G) + u_{a,j} \cdot E_j = H + D_j.$$

Fall 3: Stromverzweigung. Wenn ein Zweigstrom u_{ij} von einem Rohrabschnitt i in einen Rohrabschnitt j fließt, und gleichzeitig ein abzweigender Strom vom Abschnitt i in den Abschnitt k , so lautet die Netzwerkgleichung

$$p_i - p_j = (\dot{u}_{ij} + \dot{u}_{ik}) \cdot L_i + (u_{ij} + u_{ik}) \cdot \bar{R}_i + \dot{u}_{ij} \cdot L_j + u_{ij} \cdot R_j - \tilde{p}_j,$$

wie sich anhand von Abb. 3.18 nachvollziehen lässt. Wenn man die Ableitung der Volu-

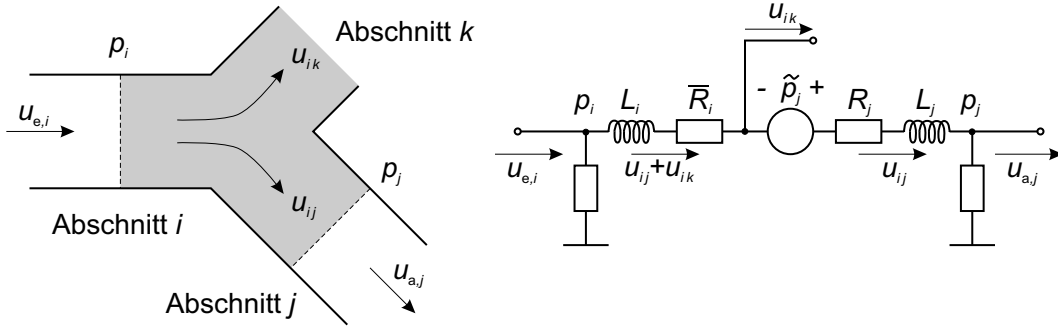


Abbildung 3.18: Abzweigender Strom.

menströme durch die zeitdiskreten Näherungen ersetzt, so erhält man die Gleichung

$$p_i - p_j = u_{ij} \cdot F + u_{ik} \cdot G + H$$

mit

$$\begin{aligned} F &= \frac{L_{ij}}{\Delta t \cdot \theta} + R_{ij}, \\ G &= \frac{L_i}{\Delta t \cdot \theta} + \bar{R}_i \quad \text{und} \\ H &= -\frac{1}{\Delta t \cdot \theta} \cdot (u'_{ij} \cdot L_{ij} + u_{ik} \cdot L_i) - \frac{\bar{\theta}}{\theta} \cdot (\dot{u}'_{ij} \cdot L_{ij} + \dot{u}_{ik} \cdot L_i) - \tilde{p}_j. \end{aligned}$$

Wenn man wiederum p_i und p_j jeweils durch Gl. (3.70) ersetzt, so ergibt sich die folgende Gleichung für das LGS:

$$u_{e,i} \cdot E_i + u_{ij} \cdot (-E_i - E_j - F) + u_{a,j} \cdot E_j + u_{ik} \cdot (-E_i - G) = H + D_j - D_i.$$

Fall 4: Einmündender Volumenstrom. Wenn ein Volumenstrom u_{ij} von einem Rohrabschnitt i in einen Rohrabschnitt j fließt, und gleichzeitig ein zusätzlicher Strom u_{kj} aus dem Abschnitt k in den Abschnitt j fließt, so lautet die Netzwerkgleichung

$$p_i - p_j = \dot{u}_{ij} \cdot L_i + u_{ij} \cdot \bar{R}_i + (\dot{u}_{ij} + \dot{u}_{kj}) \cdot L_j + (u_{ij} + u_{kj}) \cdot R_j - \tilde{p}_j,$$

wie sich anhand von Abb. 3.19 nachvollziehen lässt. Wenn man die Ableitung der Volu-

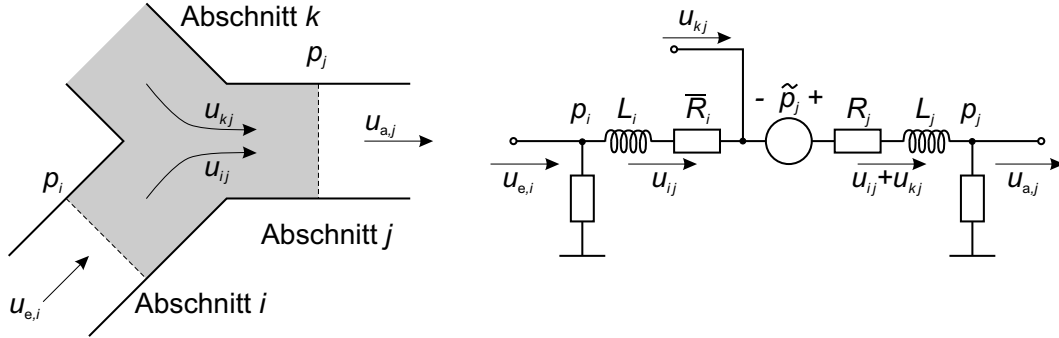


Abbildung 3.19: Einmündender Strom.

menströme durch die zeitdiskreten Näherungen ersetzt, so erhält man die Gleichung

$$p_i - p_j = u_{ij} \cdot F + u_{kj} \cdot G + H$$

mit

$$\begin{aligned} F &= \frac{L_{ij}}{\Delta t \cdot \theta} + R_{ij}, \\ G &= \frac{L_j}{\Delta t \cdot \theta} + R_j \quad \text{und} \\ H &= -\frac{1}{\Delta t \cdot \theta} \cdot (u'_{ij} \cdot L_{ij} + u_{kj} \cdot L_j) - \frac{\bar{\theta}}{\theta} \cdot (\dot{u}'_{ij} \cdot L_{ij} + \dot{u}_{kj} \cdot L_j) - \tilde{p}_j. \end{aligned}$$

Wenn man wiederum p_i und p_j jeweils durch Gl. (3.70) ersetzt, so ergibt sich die folgende Gleichung für das LGS:

$$u_{e,i} \cdot E_i + u_{ij} \cdot (-E_i - E_j - F) + u_{a,j} \cdot E_j + u_{kj} \cdot (-E_j - G) = H + D_j - D_i.$$

Fall 5 und 6: Ströme durch die Komponenten der Strahlungsimpedanz. Abbildung 3.20 zeigt die Netzwerkstruktur an einem offenen Rohrende, d. h. der Mund- oder Nasenöffnung. Dabei fließen zwei Zweigströme aus dem Rohrabschnitt i durch die Komponenten R_{rad} und L_{rad} der Strahlungsimpedanz. Für den Strom u_{iR} durch den Widerstand der Strahlungsimpedanz gilt die Gleichung

$$p_i = (\dot{u}_{iR} + \dot{u}_{iL}) \cdot L_i + (u_{iR} + u_{iL}) \cdot \bar{R}_i + u_{iR} \cdot R_{\text{rad}} - \tilde{p}_L. \quad (3.71)$$

Für den Strom u_{iL} durch die Induktivität der Strahlungsimpedanz gilt entsprechend

$$p_i = (\dot{u}_{iR} + \dot{u}_{iL}) \cdot L_i + (u_{iR} + u_{iL}) \cdot \bar{R}_i + \dot{u}_{iL} \cdot L_{\text{rad}} - \tilde{p}_L. \quad (3.72)$$

Wenn die Ableitungen des Volumenstroms durch die zeitdiskrete Näherung ersetzt werden, so können beide Gleichungen in die Form

$$p_i = u_{iR} \cdot F + u_{iL} \cdot G + H \quad (3.73)$$

gebracht werden, wobei die Größen F , G und H für Gl. (3.71) die Werte

$$\begin{aligned} F &= \frac{L_i}{\Delta t \cdot \theta} + R_i + R_{\text{rad}}, \\ G &= \frac{L_i}{\Delta t \cdot \theta} + \bar{R}_i \quad \text{und} \\ H &= -(u'_{iR} + u'_{iL}) \cdot \frac{L_i}{\Delta t \cdot \theta} - (\dot{u}'_{iR} + \dot{u}'_{iL}) \cdot L_i \cdot \frac{\bar{\theta}}{\theta} - \tilde{p}_L \end{aligned}$$

annehmen, und für Gl. (3.72) die Werte

$$\begin{aligned} F &= \frac{L_i}{\Delta t \cdot \theta} + \bar{R}_i, \\ G &= \frac{L_i + L_{\text{rad}}}{\Delta t \cdot \theta} + \bar{R}_i \quad \text{und} \\ H &= -\frac{[L_i \cdot (u'_{iR} + u'_{iL}) + L_{\text{rad}} \cdot u'_{iL}]}{\Delta t \cdot \theta} - \frac{\bar{\theta}}{\theta} \cdot [L_i \cdot (\dot{u}'_{iR} + \dot{u}'_{iL}) + L_{\text{rad}} \cdot \dot{u}'_{iL}] - \tilde{p}_L. \end{aligned}$$

Wenn in Gl. (3.73) p_i wieder durch Gl. (3.70) ersetzt wird, so erhält man jeweils eine Gleichung für u_{iR} und u_{iL} in der Form

$$u_{e,i} \cdot E_i + u_{iR} \cdot (-E_i - F) + u_{iL} \cdot (-E_i - G) = H - D_i.$$

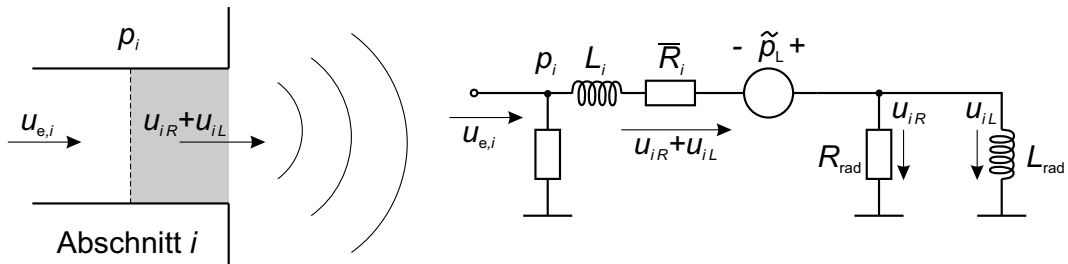


Abbildung 3.20: Abgestrahlte Zweigströme.

3.7.3 Simulation

Die Simulation des Netzwerks beruht darauf, dass in jedem Zeitschritt die Abtastwerte der Signale u_{ij} , $\dot{u}_{ij}$, p_i , $\dot{p}_i$, $u_{w,i}$, $\dot{u}_{w,i}$ und $Q_{w,i}$ (für alle i, j) berechnet werden. Zu Beginn der Simulation werden alle Signalwerte mit Null initialisiert, so dass sich das Netzwerk in Ruhe befindet. Anschließend werden in jedem Zeitschritt nacheinander die folgenden Punkte abgearbeitet:

1. Berechnung der Parameter des Rohrmodells mit Hilfe der Modelle für den 3D-Vokaltrakt und die Stimmlippen.
2. Berechnung der Komponenten und Schallquellen des akustischen Netzwerks.
3. Berechnung der Koeffizientenmatrix und des Lösungsvektors des LGS für die gesuchten Zweigströme u_{ij} .
4. Auflösung des LGS nach den Zweigströmen.
5. Berechnung der Abtastwerte der Signale $\dot{u}_{ij}$, p_i , $\dot{p}_i$, $u_{w,i}$, $\dot{u}_{w,i}$ und $Q_{w,i}$ mit Hilfe der Zweigströme.
6. Berechnung des abgestrahlten Schalldruckwertes p_{rad} .

Punkt 1 liefert für jeden Rohrabschnitt i die Länge l_i , den Querschnitt A_i und den Umfang S_i . Damit können in Punkt 2 die passiven Netzwerkkomponenten mit den Gleichungen in Abschnitt 3.5.3 berechnet werden. Damit bei der Berechnung der Komponenten kein numerischer Überlauf auftritt („Division durch Null“), wurde für den Querschnitt und den Umfang aller Rohrabschnitte ein Mindestwert eingeführt. Der Mindestquerschnitt A_{min} wurde mit $1/1000 \text{ mm}^2$ so klein gewählt, dass er einen Volumenstrom praktisch unterbindet, aber groß genug ist, um einen numerischen Überlauf zu vermeiden. Für S_{min} wird der Wert $2 \cdot \sqrt{\pi \cdot A_{\text{min}}}$ verwendet. Für die Berechnung der Widerstände R_{bern} und $\bar{R}_{\text{bern}}$ werden die Werte der Zweigströme zum aktuellen Abtastzeitpunkt benötigt. Da diese noch nicht bekannt sind, werden hierfür die Werte vom vorangegangenen Zeitpunkt eingesetzt. Die Schallquellen, die in der Simulation verwendet werden, sind die Schalldruckquelle p_{Lunge} am Eingang des ersten Rohrabschnitts des subglottalen Systems, sowie supraglottale Rauschquellen, die in Abschnitt 3.8.3 diskutiert werden.

Mit Hilfe der Netzwerkkomponenten können die Gleichungen für die Zweigströme aufgestellt werden (siehe Abschnitt 3.7.2), die sich in Form einer Koeffizientenmatrix und eines Lösungsvektors zusammenfassen lassen. Bei der Berechnung der Koeffizienten wird jeweils nur auf die unmittelbar zurückliegenden Abtastwerte der Schalldrücke, Volumenströme, usw. zugegriffen. Es genügt daher, von jeder dieser zeitdiskreten Funktionen immer nur den aktuellen und den unmittelbar vorangegangenen Abtastwert im Speicher zu halten. Für die Auflösung des Gleichungssystems habe ich ein iteratives Lösungsverfahren implementiert, welches als SOR-Verfahren (successive over relaxation) bezeichnet wird [15]. Dieses Verfahren besitzt einen Relaxationsparameter ω , der die Konvergenzeigenschaften bei der Lösung beeinflusst und zwischen 1 und 2 liegen darf. Für $\omega = 1$ erhält man das bekannte Gauß-Seidel Verfahren. Der Wert des Parameters für eine möglichst schnelle Konvergenz der Lösung muss experimentell ermittelt werden. In der vorliegenden Arbeit verwende ich den Wert $\omega = 1.25$. Gegenüber dem Gauß-Seidel-Verfahren werden damit für die Lösung des LGS im Durchschnitt nur 18.7 statt 34.3 Iterationen für eine bestimmte Genauigkeit der Lösung benötigt. Die Implementierung des Lösungsverfahrens habe ich weiterhin dadurch optimiert, dass nur diejenigen Koeffizienten der Matrix in der Berechnung verwendet werden, die von Null verschieden sind. Da die Matrix sehr dünn besetzt ist, ergibt sich dadurch ein großer Geschwindigkeitsgewinn.

Nach der Berechnung der Zweigströme u_{ij} können die Abtastwerte der Funktionen $\dot{u}_{ij}$, p_i , $\dot{p}_i$, $u_{w,i}$, $\dot{u}_{w,i}$ und $Q_{w,i}$ ausgerechnet werden. Zuerst werden die Ableitungen $\dot{u}_{ij}$ nach

Gl (3.66) berechnet. Anschließend werden die Druckwerte p_i in der Mitte der einzelnen Rohrabschnitte mit Gl. (3.70) ermittelt und dann die Ableitungen $\dot{p}_i$ wieder nach Gl. (3.66). Zuletzt werden die Werte für $u_{w,i}$, $\dot{u}_{w,i}$ und $Q_{w,i}$ nach den Gleichungen (3.69), (3.66) und (3.65) berechnet.

Der abgestrahlte Schalldruck $p_{\text{rad}}[n]$ wird entsprechend Gl. (3.48) durch die zeitliche Ableitung des Volumenstroms $u_{\text{rad}}[n]$ durch die Mund- und die Nasenöffnung approximiert. Als Ableitung wird in diesem Fall der hintere Differenzenquotient verwendet, d. h. $p_{\text{rad}}[n] \sim (u_{\text{rad}}[n] - u_{\text{rad}}[n-1])/\Delta t$. Für $u_{\text{rad}}[n]$ gilt

$$u_{\text{rad}}[n] = u_{iR}[n] + u_{iL}[n] + u_{jR}[n] + u_{jL}[n],$$

wobei i und j die letzten Rohrabschnitte des Mund- und Nasenraums sind. Um auch den von der Haut abgestrahlten Schall zu berücksichtigen, kann zu $u_{\text{rad}}[n]$ noch der Volumenstrom $c \cdot u_{w,i}[n]$ in Richtung der Rohrwand addiert werden, wobei c ein empirisch zu bestimmender Wichtungsfaktor ist und i den ersten Rohrabschnitt des Rachens direkt über der Glottis kennzeichnet. Dieser Summand ist eine angemessene Approximation der Abstrahlung über die Haut, da diese in der Nähe der Glottis laut FANT *et al.* [36] am stärksten ausgeprägt ist.

Als Abtastfrequenz f_A für die Simulation habe ich 44.1 kHz verwendet, da diese eine Standardfrequenz gängiger Audioformate ist. Für den Zeitschritt gilt somit $\Delta t = 1/44100$ s. Die Frequenz von 44.1 kHz ist ein Kompromiss zwischen der Genauigkeit und der Rechenzeit der Simulation. Grundsätzlich treten bei der zeitdiskreten Simulation eines kontinuierlichen Systems spektrale Verzerrungen auf, die mit denen bei der Transformation eines analogen Filters in einen digitalen Filter (z. B. mit Hilfe der bilinearen Transformation) vergleichbar sind. Eine doppelt so hohe Abtastrate bedeutet zwar eine Verringerung dieser spektralen Verzerrungen im unteren Frequenzbereich, aber auch eine doppelt so lange Rechenzeit. Mit der nächstniedrigeren Abtastfrequenz von 22.05 kHz würde die Rechenzeit halbiert werden, jedoch wären dann die Verschiebungen der Formantfrequenzen durch die Frequenzverzerrung bereits deutlich wahrnehmbar.

Neben der vollständigen Simulation des gesamten Netzwerks für die Sprachsynthese wurde auch die Möglichkeit implementiert, nur die Volumenstrom-Impulsantwort des Vokaltrakts zu berechnen. Dazu werden die Rohrabschnitte der Glottis verschlossen, und der Vokaltrakt wird durch einen Dirac-Stoß der Volumenstromquelle im ersten Rohrabschnitt des Rachens angeregt. Der daraus resultierende Volumenstrom an der Mund- und Nasenöffnung wird solange aufgezeichnet, bis die Impulsantwort weitestgehend abgeklungen ist (z. B. 1 s lang). Die Fourier-Transformierte des aufgezeichneten Signals entspricht dann genau der Volumenstrom-Übertragungsfunktion des Vokaltrakts und kann mit den Ergebnissen der Simulation im Frequenzbereich verglichen werden.

3.7.4 Vergleich mit der Simulation im Frequenzbereich

In diesem Abschnitt werden die Simulationen für den Zeitbereich und den Frequenzbereich miteinander verglichen. Die Simulation im Frequenzbereich (FBS) ist einerseits sehr genau, weil sie frequenzabhängige Netzwerkkomponenten, speziell die Rohrreibungswiderstände, korrekt berücksichtigt. Andererseits eignet sie sich nur für die Analyse statischer Netzwerke.

Die Zeitbereichssimulation (ZBS) synthetisiert das Sprachsignal dagegen als zeitliche Folge von Abtastwerten, wobei der Vokaltrakt fortwährend seine Form ändern kann. Dafür können im Zeitbereich keine frequenzabhängigen Netzwerkkomponenten verwendet werden, so dass statt der Grenzsichtwiderstände die Widerstände für die stationäre Strömung eingesetzt wurden. Diese sind jedoch normalerweise kleiner als die Grenzsichtwiderstände, so dass die Formanten zu schwach gedämpft werden. Außerdem mussten in der ZBS die Netzwerkgleichungen durch numerische Methoden approximiert werden. Diese Approximation verursacht eine spektrale Verzerrung, die sich in einer Verschiebung der Formanten zu tieferen Frequenzen hin äußert (siehe auch MAEDA [72]). Die ZBS unterscheidet sich von der FBS also durch zu geringe Bandbreiten und zu niedrige Frequenzen der Formanten. Für jedes der beiden Probleme habe ich mir eine Möglichkeit überlegt, um die entsprechenden Abweichungen zu reduzieren.

Formantbandbreiten. Als Lösung für die zu geringe Dämpfung der Formanten schlage ich vor, die Bandbreiten durch die *numerische* Dämpfung zu erhöhen. Die Grundlage dafür war die Beobachtung, dass die Resonanzen bei der Verwendung von Vorwärtsdifferenzen ($\theta = 1$) für die finiten Differenzengleichungen deutlich stärker gedämpft wurden als bei der Verwendung der Trapezregel ($\theta = 0.5$). Bei der Herleitung der finiten Differenzengleichungen wurde deshalb mit dem Diskretisierungsparameter θ ein Kontinuum zwischen der Trapezregel ($\theta = 0.5$) und den Vorwärtsdifferenzen ($\theta = 1$) geschaffen. Um herauszufinden, welcher θ -Wert die bestmögliche Übereinstimmung zwischen den Bandbreiten der ZBS und der FBS herstellt, habe ich für $\theta \in \{0.5, 0.51, 0.52, 0.53, 0.54\}$ die Bandbreiten der ersten vier Formanten der synthetischen Vokale [a:], [ɛ:], [i:] und [u:] gemessen³. Für $\theta > 0.54$ wurde die Dämpfung der Formanten bereits unnatürlich groß. Abbildung 3.21 zeigt am Beispiel des Vokals [a:] den Anstieg der Bandbreiten mit zunehmendem θ . In Abb. 3.22 sind die Bandbreiten $B1$, $B2$, $B3$ und $B4$ für alle vier Vokale

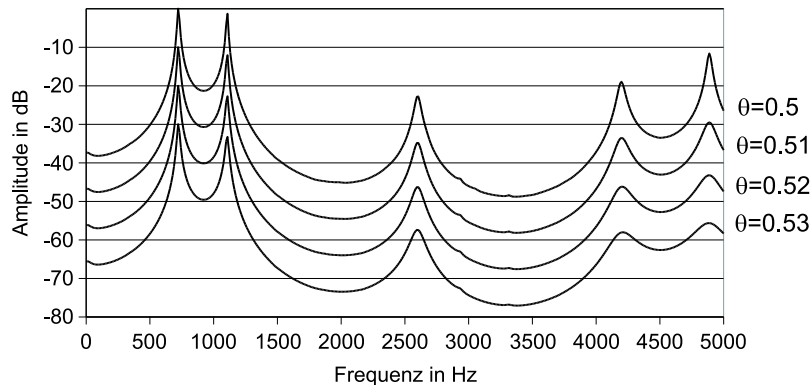


Abbildung 3.21: Übertragungsfunktion des Vokals [a:] für verschiedene Werte des Diskretisierungsparameters θ . Mit zunehmendem θ werden die Formanten stärker gedämpft.

³Die Messungen wurden an der FFT der Volumenstrom-Impulsantwort durchgeführt, die mit der ZBS berechnet wurde.

als Funktion von θ dargestellt. Man erkennt hier, dass der Anstieg der Bandbreiten mit

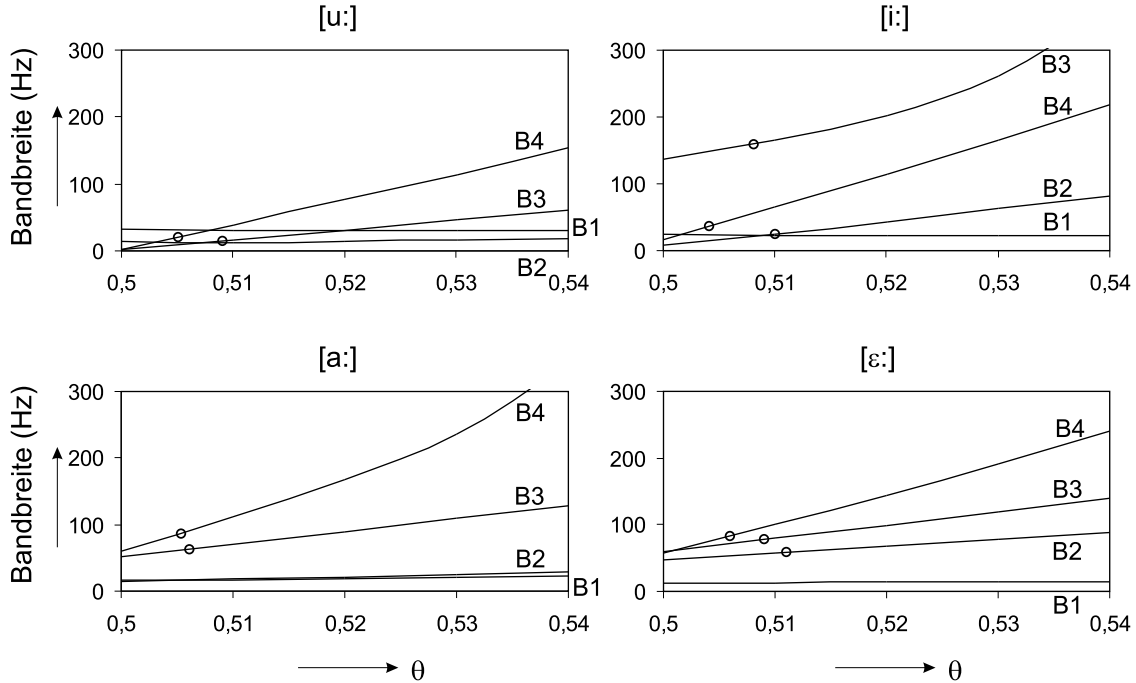


Abbildung 3.22: Die Formantbandbreiten B_1, B_2, B_3 und B_4 als Funktion des Diskretisierungsparameters θ für vier Vokale. Die Kreise markieren die Stellen, an denen die Bandbreiten mit denen aus der Simulation im Frequenzbereich übereinstimmen.

zunehmender Ordnungszahl der Resonanzen zunimmt. Die kleinen Kreise auf den Kurven markieren die Bandbreiten, die für die entsprechenden Formanten der FBS gemessen wurden. Daran lässt sich ablesen, welchen Wert θ aus der Sicht der einzelnen Formanten annehmen müsste, damit die Bandbreiten zwischen der ZBS und FBS übereinstimmen. Die Werte liegen in etwa zwischen 0.505 und 0.51. Ich habe mich in der Simulation für die Verwendung des Wertes 0.51 entschieden. Damit ist sichergestellt, dass alle Formanten entweder genauso stark oder ein wenig stärker gedämpft werden als in der FBS.

Formantfrequenzen. Neben den Bandbreiten wurden auch die Formantfrequenzen für die oben genannten Vokale gemessen, die sich in der ZBS mit verschiedenen Werten für θ ergeben haben. In Tab. 3.2 sind die Messwerte für $\theta \in \{0.5, 0.52, 0.54\}$ gemeinsam mit den Referenzwerten aus der FBS zusammengestellt. Daraus ist ersichtlich, dass die Formantfrequenzen unabhängig vom Parameter θ in der ZBS in aller Regel niedriger sind als die entsprechenden Frequenzen der FBS. Diese Abweichungen entstehen durch die Verzerrung der Frequenzachse aufgrund der zeitdiskreten Approximation der Netzwerkgleichungen. Einen entgegengesetzten Effekt auf die Formantfrequenzen hat eine Erhöhung der Schallgeschwindigkeit. Um herauszufinden, inwiefern sich dieser Effekt nutzen lässt, um der Verschiebung der Formantfrequenzen in der ZBS entgegenzuwirken, habe ich die folgende Untersuchung angestellt. Ich habe die Schallgeschwindigkeit für die ZBS gegenüber der FBS um 0.5, 1, 1.5 und 2% erhöht und die daraus resultierende

		$\theta = 0.5$	$\theta = 0.52$	$\theta = 0.54$	FBS
[a:]	F1	723.54	723.55	723.57	724.35
	F2	1106.24	1106.07	1105.76	1108.23
	F3	2600.6	2599.84	2598.4	2630.76
	F4	4197.82	4204.35	4217	4328.2
[ɛ:]	F1	565.4	565.38	565.35	568.44
	F2	1892.29	1892.51	1892.84	1904.03
	F3	2581.47	2579.69	2577	2610.81
	F4	3895.07	3896.87	3899.75	3998.52
[i:]	F1	318.32	318.24	318.21	319.92
	F2	2518.25	2519.28	2521.49	2545.99
	F3	2964.16	2957.69	2947.8	3004.9
	F4	4192.27	4191.94	4190.84	4321.05
[u:]	F1	301.52	301.45	301.42	302.88
	F2	779.22	779.16	778.99	778.69
	F3	2305.84	2305.7	2305.15	2326.38
	F4	3664.21	3664.35	3665.01	3748.77

Tabelle 3.2: Formantfrequenzen für verschiedene θ -Werte bei 44.1 kHz. Die Referenzwerte von der Frequenzbereichssimulation sind zum Vergleich angegeben.

Änderung der Formantfrequenzen für die gleichen vier Vokale wie oben ausgewertet. Das linke Diagramm in Abb. 3.23 zeigt, wie sich die prozentuale Abweichung der jeweils ersten vier Formantfrequenzen jedes Vokals (insg. 16 Linien) von den entsprechenden Referenzfrequenzen der FBS mit zunehmender Schallgeschwindigkeit entwickelt. Bei einer Übereinstimmung der Schallgeschwindigkeiten liegen fast alle Frequenzen der ZBS um bis zu 3% unter denen der FBS⁴. Wenn die Schallgeschwindigkeit in der ZBS aber erhöht wird, so nehmen die Abweichungen der Formantfrequenzen betragsmäßig zunächst kontinuierlich ab, bis sie früher oder später die 0%-Marke schneiden. Anschließend nimmt die Abweichung wieder zu, weil die Frequenzen dann größer werden als die Referenzwerte. Bei welcher Zunahme der Schallgeschwindigkeit die Linien die 0%-Marke schneiden hängt offensichtlich stark vom betrachteten Formanten ab. Deshalb sind im rechten Diagramm in Abb. 3.23 für jede prozentuale Erhöhung der Schallgeschwindigkeit die Mittelwerte der Absolutwerte der Frequenzabweichungen jeweils aller 16 Formanten dargestellt. Offensichtlich besitzt die mittlere Abweichung bei einer um etwa 1% erhöhten Schallgeschwindigkeit ein lokales Minimum. Deshalb verwende ich für die ZBS statt der in Tab. B.1 angegebenen Schallgeschwindigkeit c , die in der FBS verwendet wird, die Schallgeschwindigkeit $c' = 1.01 \cdot c$.

Um die Fehler zu verringern, die in der ZBS durch die zeitliche Diskretisierung und die Verwendung frequenzunabhängiger Reibungswiderstände gemacht werden, wurden hier-

⁴FLANAGAN [38] gibt an, dass Verschiebungen von Formantfrequenzen um weniger als 3% vom Ohr nicht wahrnehmbar sind. Diese Unterscheidungsschwelle wurde jedoch für die Verschiebung von jeweils nur einem einzelnen Formanten gemessen. Für die simultane Verschiebung mehrerer Formanten könnte diese Schwelle aber tiefer liegen, so dass eine „Korrektur“ der Verschiebungen auch aus der Sicht der Perzeption gerechtfertigt ist.

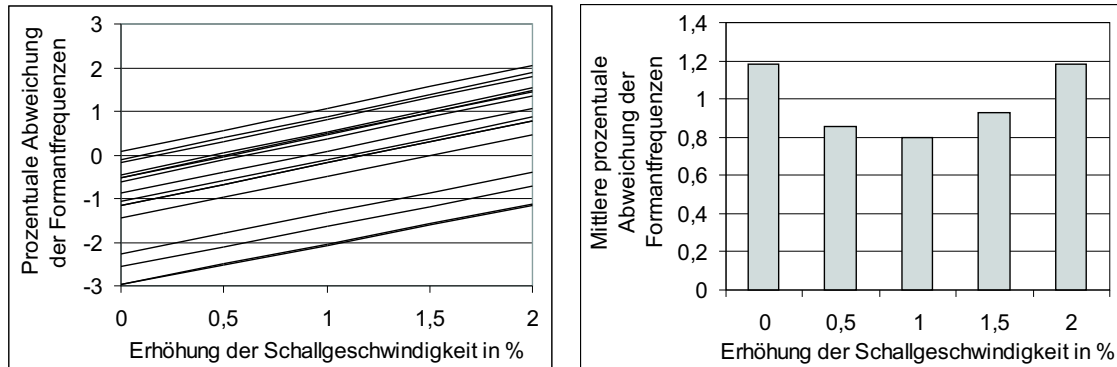


Abbildung 3.23: Links: Prozentuale Abweichung der Formantfrequenzen (4 Vokale mit je 4 Formanten ergibt 16 Linien) der ZBS ($\theta = 0.51$) von denen der FBS für eine prozentuale Vergrößerung der Schallgeschwindigkeit in der ZBS. Rechts: Mittlere betragsmäßige Abweichung der 16 Formantfrequenzen.

mit zwei Möglichkeiten vorgeschlagen: Die Verwendung des Diskretisierungsparameters $\theta = 0.51$ statt $\theta = 0.5$ und die Erhöhung der Schallgeschwindigkeit in der Simulation um 1%. Wie diese beiden Maßnahmen die Übertragungsfunktion für den Vokal [o:] beeinflussen, ist in Abb. 3.24 dargestellt. Die obere durchgezogene Kurve zeigt das Amplituden-

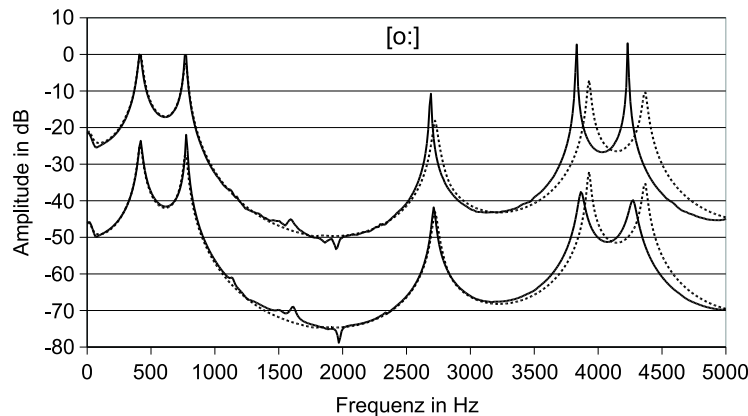


Abbildung 3.24: Vergleich der Übertragungsfunktionen für den Vokal [o:] aus der FBS und ZBS mit und ohne Maßnahmen zur Korrektur der Formantfrequenzen und -bandbreiten. Die gestrichelten Kurven zeigen die Übertragungsfunktion der FBS. Die durchgezogenen Kurven zeigen die Übertragungsfunktion der ZBS ohne Korrektur (oben) und mit Korrektur (unten).

spektrum *ohne* die beschriebenen Maßnahmen ($\theta = 0.5$ und $c' = c$) und die untere Kurve zeigt das Amplitudenspektrum für $\theta = 0.51$ und $c' = 1.01 \cdot c$. Die gestrichelten Kurven zeigen zum Vergleich das Referenzspektrum der FBS. Man erkennt, dass sowohl die Frequenzen als auch die Bandbreiten der Formanten durch die zwei Maßnahmen besser mit

den entsprechenden Werten der FBS übereinstimmen, was sich analog auf die anderen Vokale überträgt und einen wichtigen Beitrag zur guten Synthesequalität stimmhafter Laute leistet. Meine Syntheseexperimente haben gezeigt, dass die geringen Formantbandbreiten bei einigen Vokalen ohne die vorgeschlagenen Korrekturen der ZBS zu einem unnatürlichen hochfrequenten Störgeräusch führen.

3.8 Modellierung sekundärer Schallquellen

3.8.1 Mechanismus der Sekundärschallerzeugung

Wenn der Luftstrom durch eine Konstriktion im Vokaltrakt groß genug ist, so löst sich die Strömung am Ausgang der Konstriktion von der Rohrwand ab und tritt als gebündelter Luftstrahl in den stromabwärts liegenden Hohlraum ein. Dort vermischt sich der Luftstrahl unter Wirbelbildung mit der umliegenden Luft. In Abschnitt 3.4.2 wurde der fluidmechanische Energieverlust diskutiert, der dadurch entsteht. Gleichzeitig erzeugen die Turbulenzen aber auch Geräusche, die als Sekundärschall bezeichnet werden. Die Geräusche haben im Wesentlichen zwei Ursachen. Einerseits wirken die zufälligen Geschwindigkeitsschwankungen bei der Vermischung der Luft als akustische Monopolquellen. Andererseits entstehen zufällige Druckschwankungen, wenn der Luftstrahl auf feste Hindernisse wie die Rohrwand oder die Zähne trifft. Die Druckschwankungen wirken als akustische Dipolquellen. Sowohl die Geschwindigkeits- als auch die Druckschwankungen erstrecken sich jeweils entlang eines bestimmten Bereichs des Vokaltrakts. Streng genommen handelt es sich hierbei also um *verteilte* Schallquellen.

Für die Simulation von Sekundärschall müssen die verteilten Monopol- und Dipolquellen durch *konzentrierte* Quellen approximiert werden. In einem akustischen Netzwerk entspricht die Monopolquelle einer Volumenstromquelle und die Dipolquelle einer Schalldruckquelle. In den meisten Simulationen wird das Turbulenzgeräusch auf eine einzige konzentrierte Quelle reduziert. Für gewöhnlich wird dafür eine Schalldruckquelle verwendet [5, 14, 39], da die turbulenten Schalldruckschwankungen in aller Regel ein lauterer Geräusch verursachen als die Geschwindigkeitsschwankungen und somit den größten Anteil am Sekundärschall ausmachen [87]. Besonders entscheidend für gute Simulationsergebnisse sind die Positionen, Amplituden und spektralen Zusammensetzungen der eingesetzten Quellen. Diese Parameter hängen jeweils von mehreren Faktoren ab, wie z. B. der Länge und dem Querschnitt der Konstriktion, der Strömungsgeschwindigkeit in der Konstriktion, der Entfernung, bis der Luftstrahl auf ein Hindernis trifft, und wahrscheinlich auch von der konkreten dreidimensionalen Geometrie des Vokaltrakts stromabwärts von der Konstriktion [101]. Da die genauen Abhängigkeiten aufgrund der Komplexität des Turbulenzproblems noch nicht vollständig verstanden sind, ist man hierbei hauptsächlich auf empirisch gefundene Zusammenhänge angewiesen.

Eine der ältesten quantitativen Studien zum Sekundärschall ist die Arbeit von MEYER-EPPLER [79]. Der Autor hat anhand von Messungen an Rohrmodellen herausgefunden, dass die Amplitude des abgestrahlten Schalldrucks p_s eine Funktion der Reynolds-Zahl Re in der Konstriktion ist:

$$p_s = \begin{cases} \alpha \cdot (Re^2 - Re_{\text{krit}}^2), & \text{wenn } Re > Re_{\text{krit}} \\ 0 & \text{sonst} \end{cases} \quad (3.74)$$

Für die kritische Reynoldszahl Re_{krit} wurde der Wert 1800 angegeben. Ein Geräusch entsteht demnach erst dann, wenn die Reynoldszahl in der Konstriktion den Wert Re_{krit} übersteigt, d. h. wenn die Strömung turbulent geworden ist. Oberhalb des Schwellwerts wächst der Schalldruck, wie der Vergleich mit Gl. (3.22) lehrt, proportional zu $v_c^2 \cdot A_c$ an, wobei v_c die Strömungsgeschwindigkeit in der Konstriktion ist und A_c ihre Querschnittsfläche. Gleichung (3.74) wurde in verschiedenen Arbeiten aufgegriffen, um die Amplitude einer Schalldruckquelle an einer supraglottalen Konstriktion zu berechnen [14, 66, 106]. STEVENS [108, 110] kommt dagegen aufgrund experimenteller Befunde und theoretischer Überlegungen zu dem Schluss, dass die Amplitude p_s der Schalldruckquelle proportional zu $v_c^3 \cdot \sqrt{A_c}$ ist. BADIN *et al.* [8] wiederum geben für p_s eine Proportionalität zu $v_c^2 \cdot A_c^{0.4}$ an, die auf aerodynamischen und akustischen Daten gehaltener Frikative eines Sprechers basiert. Die Proportionalität von p_s zu v_c^2 wird laut SINDER [104, S. 14] auch durch die Theorie zur strömungsakustischen Schallerzeugung von HOWE [48] unterstützt. Unabhängig von der Proportionalitätsbeziehung weisen experimentelle Daten außerdem darauf hin, dass die absolute Amplitude der Schalldruckquelle umso geringer ist, je weiter das Hindernis im Luftstrom von der Konstriktion entfernt ist [87].

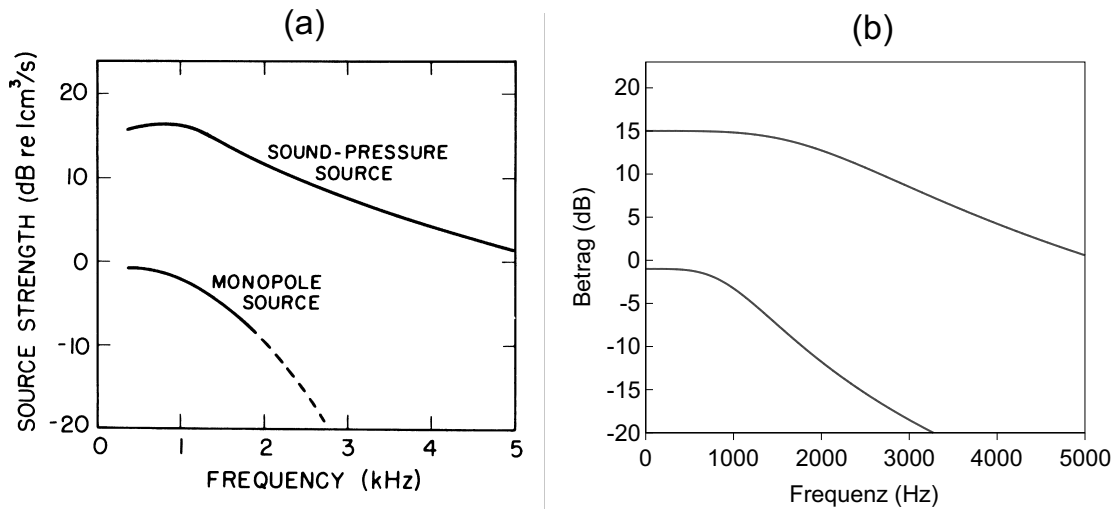


Abbildung 3.25: Links: Spektren einer Monopolquelle (unten) und einer Dipolquelle (oben) nach STEVENS [110]. Rechts: Nachbildung der Spektren durch die Übertragungsfunktion von Butterworth-Tiefpassfiltern zweiter Ordnung. Die Knickefrequenzen der Filter sind 1100 Hz (unten) und 2200 Hz (oben).

Das Spektrum einer Schalldruckquelle, die durch das Auftreffen eines gebündelten Luftstrahls auf ein Hindernis entsteht, besitzt laut STEVENS [110] eine breite Spitze bei einer Frequenz von $f_s = k \cdot v_c / d$, wobei v_c die Geschwindigkeit des Luftstrahls ist, d der Durchmesser der Konstriktion und k eine Konstante, die zwischen 0.1 und 0.2 liegt. Oberhalb der Frequenz f_s fällt das Spektrum langsam ab. Abbildung 3.25 (a) (siehe [110]) zeigt das Spektrum einer Dipolquelle zusammen mit dem Spektrum einer Monopolquelle. Beide Spektren wurden experimentell durch Versuche an Rohrmodellen bestimmt. Das Dipolspektrum bestätigt das oben genannte Verhalten. Das Monopolspektrum zeigt im Vergleich dazu eine geringere Amplitude und einen stärkeren Abfall bei hohen Frequenzen.

Was die Positionen der Geräuschquellen betrifft, so sollte eine Schalldruckquelle stets in unmittelbarer Nähe des Hindernisses liegen, auf das der gebündelte Luftstrahl trifft. Monopolquellen müssten streng genommen an mehreren Stellen entlang des sich auflösenden Luftstrahls platziert werden, um der verteilten Natur der Quellen gerecht zu werden. In mehreren Studien wird nur eine einzelne Schalldruckquelle für die Modellierung der Geräuschaute verwendet, die ein kleines Stück stromabwärts von der Konstriktion eingesetzt wird [5, 14, 66].

Ein neuerer Ansatz zur Simulation von Sekundärschall basiert erstmalig auf einer genauen physikalischen Modellierung des Turbulenzproblems [104]. Dabei wird direkt die Konvektion einzelner Wirbel im Luftstrom simuliert und damit automatisch die räumliche Verteilung der Quellen berücksichtigt. Gleichzeitig wird die Amplitude der Quellen mit Hilfe strömungsakustischer Gleichungen berechnet. Die Ergebnisse des neuen Simulationsmodells müssen aber noch stärker anhand von Experimenten an realen Rohrmodellen verifiziert werden.

3.8.2 Flächenfunktionen und Rauschquellen stimmloser Frikative

Für die Synthese natürlich klingender Frikative ist neben den Parametern der eingesetzten Schallquellen auch die Form des Vokaltrakts von entscheidender Bedeutung. Bei Frikativen ist besonders die Geometrie stromabwärts der Konstriktion wichtig, denn die Hohlräume dahinter haben aufgrund der hohen akustischen Impedanz der Konstriktion nur eine geringe Bedeutung für die Filterung der Geräuschquellen. Im Rahmen dieser Arbeit wurde versucht, für mehrere deutsche Frikative die wesentlichen Merkmale der Flächenfunktionen vor der Konstriktion zusammen mit den Parametern der Geräuschquellen zu bestimmen. Die Vorgehensweise bestand darin, die Spektren künstlicher Frikative durch die *interaktive* Anpassung der Geräuschquellenparameter und Flächenfunktionen in eine möglichst gute Übereinstimmung mit den Spektren natürlicher Frikative zu bringen.

Untersucht wurden die Frikative [f], [s], [ç], [x] sowie das harte und das weiche [ʃ]. Die natürlichen Frikative wurden von vier Personen (zwei Männer und zwei Frauen) aufgenommen und mit 22 kHz und 16 Bit/Sample digitalisiert. Jeder Frikativ wurde mindestens 500 ms lang gehalten. Die Spektren wurden berechnet, indem jeweils 40 Kurzzeit-Amplitudenspektren von überlappenden Signalfenstern gemittelt wurden (23.2 ms Fensterlänge, 11 ms Überlappung, Wichtung mit Hamming-Fenster). Von den vier berechneten Spektren pro Frikativ wurde jeweils eines für die Untersuchung ausgewählt.

Die synthetischen Spektren wurden direkt im Frequenzbereich mit der in Abschnitt 3.6 beschriebenen Methode berechnet. Als Grundlage für die Flächenfunktionen der Frikative dienten die Messdaten von FANT [33, S. 172]. Jede der Funktionen wurde durch die 40 Rohrabchnitte des Rachen- und Mundraums approximiert, wobei die Länge der Rohrabchnitte für einen bestimmten Frikativ jeweils gleich war. Der Nasenraum war während der Simulationen akustisch abgekoppelt, und der Querschnitt der Stimmritze wurde auf 0.3 cm^2 gesetzt, was ein typischer Wert für die Erzeugung von Zischlauten ist [110]. Dadurch wurde eine realistische Schallabstrahlung in den subglottalen Raum ermöglicht, und somit eine natürliche Dämpfung der Formanten und Antiformanten herbeigeführt. Den pulmonalen Druck zur Berechnung des mittleren Volumenstroms nach Gl. (3.58) habe ich auf 800 Pa gesetzt.

Für die Berechnung der Geräuschquellenspektren bin ich dem Vorschlag von NARAYANAN

et al. [87] gefolgt und habe für jede Quelle entweder die Übertragungsfunktionen eines Tiefpassfilters erster Ordnung (ein einzelner Pol auf der reellen Achse in der Laplace-Ebene) oder zweiter Ordnung (ein konjugiert komplexes Polpaar) verwendet. Diese Filter bieten die Wahl einer beliebigen Knickfrequenz und Amplitude, und eines asymptotischen spektralen Abfalls nach der Knickfrequenz von entweder -6 dB/oct (erste Ordnung) oder -12 dB/oct (zweite Ordnung). Die Filter zweiter Ordnung besitzen als zusätzlichen Parameter die Polgüte Q . In Abb. 3.25 (b) ist dargestellt, wie mit Hilfe von stark gedämpften Tiefpässen zweiter Ordnung die in Abb. 3.25 (a) gegebenen Quellspektren approximiert werden können. Beide Filter besitzen eine Polgüte von $1/\sqrt{2}$. Dadurch haben sie ein maximal flaches Passband und gehören zur Gruppe der Butterworth-Filter.

Um eine gute Übereinstimmung zwischen den synthetischen und den natürlichen Spektren herbeizuführen, habe ich die folgenden Maßnahmen zugelassen: Die Variation der Gesamtlänge des Vokaltrakts zwischen Glottis und Mundöffnung, die Veränderung der Querschnittsflächen stromabwärts von der Konstriktion und die Platzierung und Parametrisierung beliebiger Rauschquellen. Ich habe versucht, für jeden Frikativ möglichst wenige Quellen einzusetzen und die Flächenfunktionen von FANT gleichzeitig möglichst wenig zu verändern. Als Flächenwerte für die Rohrabsnitte wurden – in Anlehnung an FANT [33] – nur die folgenden diskreten Werte zugelassen: 0.16–0.32–0.65–1–1.3–1.6–2–2.6–3.2–4–5–6.5–8–10.5–13–16 cm².

Das Ergebnis der manuellen spektralen Anpassung für die sechs Frikative ist in Abb. 3.26 dargestellt. Die linke Seite zeigt die diskreten Flächenfunktionen, wobei die Rohrabsnitte stromabwärts von der Konstriktion dunkelgrau gezeichnet sind. Die vermutete Position der Zähne habe ich jeweils durch ein kleines Dreieck gekennzeichnet, und die Position der Monopol- und Dipolquellen mit einem **M** bzw. **D**. Die Monopolquellen befinden sich aufgrund der Struktur des akustischen Netzwerks (siehe Abschnitt 3.5.2) jeweils in der Mitte eines Rohrabchnitts und die Dipolquellen an den Übergängen zwischen zwei Abschnitten. Auf der rechten Seite der Abbildung sind für jeden Frikativ das natürliche Spektrum (oben) und das synthetische Spektrum (unten) dargestellt. Man erkennt, dass für alle Frikative eine verhältnismäßig gute Übereinstimmung zwischen den Spektren hergestellt werden konnte. Auch auditiv war eine starke Ähnlichkeit zwischen den natürlichen und den synthetischen Frikativen festzustellen. Zur Berechnung der Audiosignale aus den synthetischen Spektren habe ich gaußverteiltes weißes Rauschen mit den Impulsantworten (inverse Fourier-Transformation der Spektren) der Frikative gefaltet.

Wie man in Abb. 3.26 sieht, war es möglich, jeden Frikativ mit einer Monopolquelle in der Konstriktion und einer Dipolquelle an einem Hindernis weiter stromabwärts zu synthetisieren. Diese Erkenntnis wird durch die Studie von NARAYANAN *et al.* [87] bestätigt. Für die Quellspektren wurden ausschließlich Butterworth-Tiefpässe zweiter Ordnung verwendet ($Q = 1/\sqrt{2}$ und -12 dB/oct nach der Knickfrequenz). Die Knickfrequenzen aller Monopolspektren wurden weiterhin auf 1100 Hz festgesetzt (siehe Abb. 3.25 (b)). Die Knickfrequenzen f_c der Dipolspektren waren jedoch variabel und sind in Abb. 3.26 für die einzelnen Quellen angegeben. Die Amplitude aller Dipolquellen wurde auf einen konstanten Wert gesetzt und die Amplitude der Monopolquellen jeweils relativ dazu skaliert. Die relativen Amplituden sind unter die Buchstaben **M** geschrieben. Ein Wert von $a = 0.2$ bedeutet beispielsweise, dass die absolute Amplitude $\tilde{u}$ der Volumenstromquelle gleich $\tilde{u} = 0.2 \cdot \tilde{p}/R_c$ ist, wobei $\tilde{p}$ die Amplitude der Dipolquelle und R_c der Widerstand für den Stoßverlust an

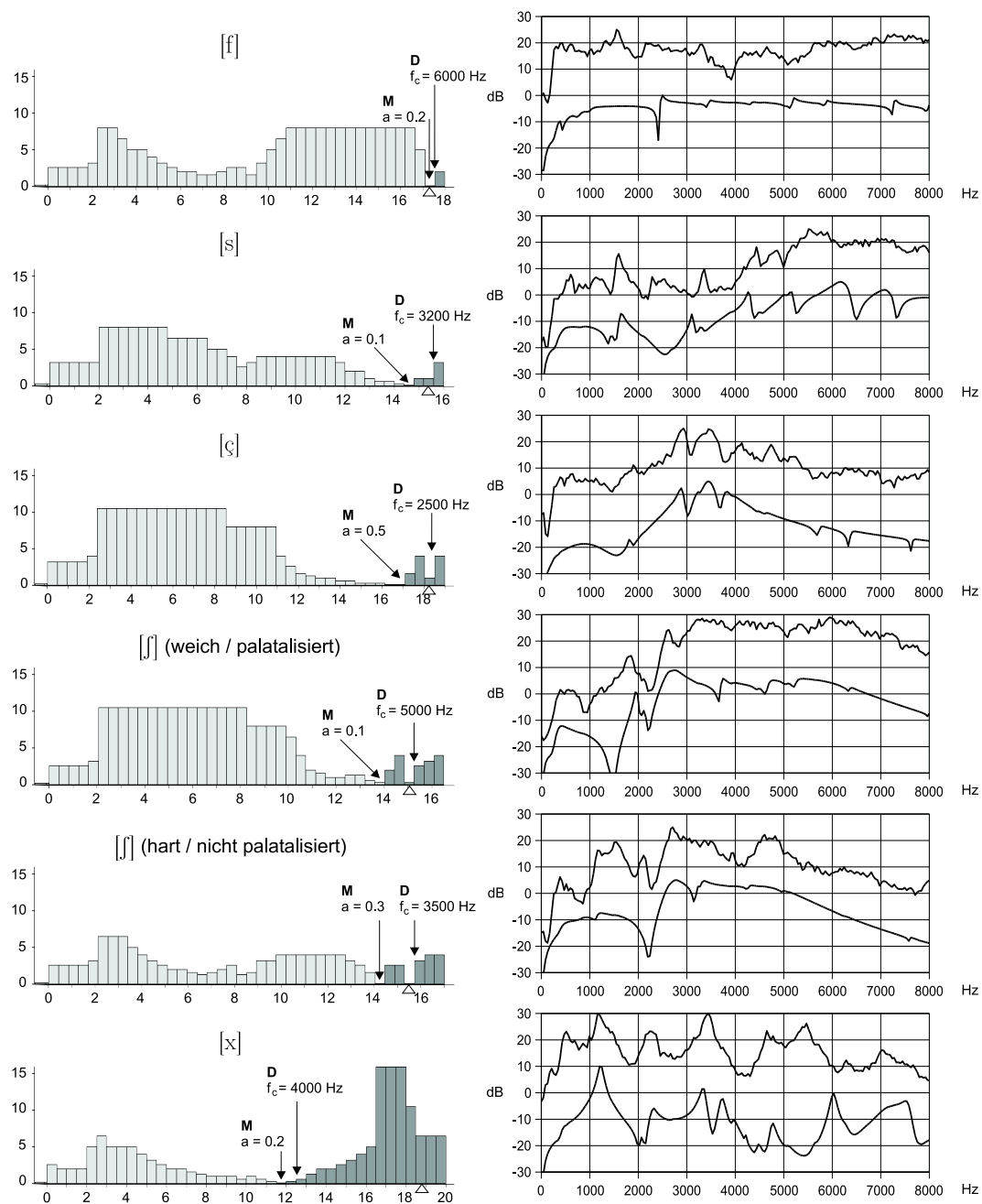


Abbildung 3.26: Ergebnisse der Anpassung synthetischer Frikativspektren an reale Spektren. Die linke Seite zeigt die Flächenfunktionen der künstlichen Frikative. Die Rohrschnitte stromabwärts der Konstriktion sind dunkelgrau gezeichnet. Die Positionen der Geräuschquellen sind mit Pfeilen gekennzeichnet. Die rechte Seite des Bildes zeigt zu jedem Frikativ das synthetische Spektrum (unten) und das natürliche Spektrum (oben).

der Konstriktion ist. Generell war der Einfluss bzw. die Amplitude der Monopolquellen im Vergleich zu den Dipolquellen gering. Für eine gute Übereinstimmung der zwei Spektren des Frikativs [ç] bei tiefen Frequenzen war die Monopolquelle aber unerlässlich. Neben den Eigenschaften der Quellen waren auch die Querschnittsflächen stromabwärts von der Konstriktion entscheidend für eine gute spektrale Übereinstimmung. In Tab. 3.3 sind sie jeweils von der engsten Stelle der Konstriktion an aufgelistet.

Frikativ	Querschnittsfläche in cm ²
[f]	0.16 – 2
[s]	0.16 – 1 – 1 – 3.2
[ç]	0.16 – 1.6 – 4 – 1 – 4
[ʃ] (weich)	0.32 – 2 – 4 – 0.32 – 2.6 – 3.2 – 4
[ʃ] (hart)	0.16 – 2.6 – 2.6 – 0.16 – 3.2 – 4 – 4
[x]	0.16 – 0.32 – 0.65 – 1.3 – 2 – 2 – ...

Tabelle 3.3: Querschnittsflächen von der Konstriktion aus stromabwärts.

Wir wenden uns nun einer Diskussion der Ergebnisse für die einzelnen Frikative zu. Die Monopolquelle für das [f] befindet sich in der labio-dentalen Konstriktion und die Dipolquelle eine halbe Rohrlänge weiter an den Lippen. Für [s], [ç] und [ʃ] wird die Konstriktion mit der Zungenspitze bzw. dem Zungenblatt gebildet, während die Zähne das wichtigste Hindernis für den Luftstrahl bilden. Für das weiche und das harte [ʃ] sind jeweils die Konstriktionen an den Zähnen genauso eng wie diejenigen mit der Zunge. Die relativ großen Querschnittsflächen zwischen den Zähnen und der Konstriktion mit der Zunge für [ç] und [ʃ] lassen sich durch einen sublingualen Hohlraum erklären, der anhand von MRT-Messungen in [87] bestätigt wird. Die Konstriktion für das [x] wird mit dem Zungendorsum gebildet, wobei die Vokaltraktwände hier als Hindernis für den gebündelten Luftstrahl wirksam werden. Die Knickfrequenzen der Dipolquellen variieren zwischen 2500 und 6000 Hz und scheinen kein vorhersagbares Muster zu zeigen. Hervorstechend ist nur die hohe Knickfrequenz für das [f], die nötig war, um ein hinreichend flaches Spektrum zu erhalten.

3.8.3 Modell für die Rauschquellen in dieser Arbeit

Mit dem theoretischen Hintergrund aus Abschnitt 3.8.1 und der Untersuchung aus Abschnitt 3.8.2 habe ich ein Verfahren für die automatische Generierung und Parametrisierung von Geräuschquellen entwickelt, welches für die zeitdiskrete Simulation in dieser Arbeit eingesetzt wird. Das Verfahren beruht auf der Annahme, dass zu jedem Zeitpunkt maximal eine supraglottale Konstriktion existiert, von der ein gebündelter Luftstrahl ausgeht. Sofern eine solche Konstriktion existiert, werden zwei Geräuschquellen eingesetzt: Eine Monopolquelle $\tilde{u}$ am Ende der Konstriktion und eine Dipolquelle $\tilde{p}$ an einem Hindernis, auf das der Luftstrahl trifft. In jedem Zeitschritt der Simulation muss also bestimmt werden, ob es eine supraglottale Konstriktion gibt und wo sie sich befindet, wo sich ein potenzielles Hindernis für den Luftstrahl befindet und welche Parameter die Geräuschquellen annehmen.

Die naheliegenste Methode zur Bestimmung des Konstriktionsortes ist die Auswahl des Rohrabschnitts mit dem geringsten Querschnitt [66, 14]. Bei einigen Frikativen, insbesondere [ʃ], kann das dreidimensionale Vokaltraktmodell jedoch Flächenfunktionen liefern, bei

denen der Rohrquerschnitt an der Position der Schneidezähne kleiner ist als der Querschnitt an der Position der Zungenkonstriktion. In diesem Fall sollte trotzdem die Konstriktion mit der Zunge als Ausgangspunkt für den gebündelten Luftstrahl gewählt werden, bei einem [f] jedoch die Konstriktion an den Zähnen. Um das zu entscheiden, kann man ausnutzen, dass eine Konstriktion an den Zähnen häufig *kürzer* ist als eine Zungenkonstriktion. Dies wurde wie folgt im Algorithmus zur Bestimmung des Konstriktionsortes bedacht. Wenn die Rohrabschnitte von der Glottis bis zur Mundöffnung mit 1 bis N durchnummeriert sind, und die Schneidezähne sich im Bereich des Abschnitts i befinden, so wird sowohl für den hinteren Bereich mit den Abschnitten $1 \dots i - 1$ der Rohrabschnitt mit dem kleinsten Querschnitt bestimmt, als auch für den vorderen Bereich mit den Abschnitten $i \dots N$. Dann wird für beide Bereiche die Länge der Konstriktion bestimmt. Die Rohrabschnitte links und rechts von dem Abschnitt mit dem geringsten Querschnitt $A_{\min}$ werden jeweils solange zur Konstriktion gezählt, wie deren Querschnitt kleiner als $A_{\min} + 0.2 \text{ cm}^2$ ist. Wenn der erste und letzte Rohrabschnitt der Konstriktion im hinteren Bereich mit j und k bezeichnet werden und die des vorderen Bereiches mit m und n , so können die folgenden zwei „Qualitätsmaße“ definiert werden:

$$Q_{\text{hinten}} = \sum_{i=j}^k l_i / A_i \quad \text{und} \quad Q_{\text{vorne}} = \sum_{i=m}^n l_i / A_i.$$

Der Q -Werte einer Konstriktion ist also umso höher, je länger und enger sie ist. Wenn $Q_{\text{hinten}} > Q_{\text{vorne}}$, so entscheidet sich der Algorithmus für die hintere Konstriktion als Ausgangspunkt für den gebündelten Luftstrahl und ansonsten für die vordere Konstriktion. Eine Berechnung der Geräuschquellen wird jedoch nur dann vorgenommen, wenn $A_{\min}$ der ausgewählten Konstriktion nicht größer als 0.5 cm^2 ist. Als konkreter Ablösepunkt x_S für den Luftstrahl wird die Mitte des letzten, stromabwärts liegenden Rohrabschnitts der Konstriktion gewählt. Dorthin wird die Monopolquelle gesetzt. Die Position der Dipolquelle hängt davon ab, ob bei der vorangegangenen Entscheidung die Konstriktion im vorderen Bereich (Zähne und Lippen) oder im hinteren Bereich gewählt wurde. Im ersten Fall wird die Dipolquelle unmittelbar stromabwärts von der Monopolquelle platziert (analog zu den Quellen für das [f] in Abb. 3.26). Dabei wird die Unterlippe als das dominante Hindernis im Luftstrahl angesehen. Wenn die Konstriktion im hinteren Bereich ausgewählt wurde, so werden entweder die Schneidezähne oder die Vokaltraktwand 0.5 cm stromabwärts der Monopolquelle als Hindernis aufgefasst. Die Schneidezähne an der Position x_Z werden immer dann ausgewählt, wenn $x_Z - x_S < 4 \text{ cm}$. Diese Bedingung ist in aller Regel für die Frikative [s], [ʃ] und [ç] erfüllt. Ansonsten wird als Hindernis die Vokaltraktwand angesehen, wie es für den Frikativ [x] der Fall ist. Die Dipolquelle wird am rechten Ende des Rohrabschnitts platziert, in dem sich das Hindernis befindet (siehe Abb. 3.26).

Das Spektrum der beiden Quellen entspricht jeweils der Übertragungsfunktion eines Butterworth-Tiefpassfilters zweiter Ordnung, d. h.

$$H(\omega) = A \cdot \frac{\omega_0^2}{(j \cdot \omega - \omega_0 \cdot e^{j \cdot 3 \cdot \pi / 4}) \cdot (j \cdot \omega - \omega_0 \cdot e^{-j \cdot 3 \cdot \pi / 4})},$$

wobei A die Amplitude ist, $\omega_0 = 2 \cdot \pi \cdot f_0$ die Knickfrequenz und $j = \sqrt{-1}$. Die Berechnung der Amplituden $\tilde{u}$ und $\tilde{p}$ der beiden Quellen erfolgt in Anlehnung an MEYER-EPPLER [79]

mit den Gleichungen

$$\begin{aligned}\Phi &= \begin{cases} \alpha \cdot (Re^2 - Re_{\text{krit}}^2) & \text{für } Re > Re_{\text{krit}} \\ 0 & \text{sonst} \end{cases} \\ \tilde{p} &= \Phi \cdot \eta \cdot e^{-d/\tau} \\ \tilde{u} &= \beta \cdot \Phi,\end{aligned}$$

wobei $Re = v_c \cdot d_c / \nu$ die Reynoldszahl der Strömung in der Konstriktion ist und $Re_{\text{krit}} = 1800$ die kritische Reynoldszahl. v_c und d_c sind die Strömungsgeschwindigkeit und die charakteristische Weite des *engsten* Rohrabschnitts der Konstriktion (Querschnitt A_{min}). Die Variable Φ kann als Amplitude einer Dipolquelle interpretiert werden, die entsteht, wenn der gebündelte Luftstrahl senkrecht auf ein Hindernis trifft, welches sich direkt stromabwärts vom Ablösepunkt befindet. Diese Größe geht zusammen mit den Faktoren η und $e^{-d/\tau}$ in die Berechnung von $\tilde{p}$ ein. Der Faktor η dient der Abschwächung des Geräusches, wenn der Luftstrahl nicht senkrecht, sondern schräg auf ein Hindernis trifft. Wenn die Zähne das Hindernis sind, so wird η auf 1 gesetzt, und wenn es die Lippen oder die Wände des Vokaltrakts sind, auf 0.5. Diese Annahmen sorgen dafür, dass für die Reibelaute [f] und [x] ein leiseres Geräusch als für die Zischlaute [s] und [ʃ] entsteht. Der Faktor $e^{-d/\tau}$ verursacht eine Abschwächung der Geräuschamplitude mit zunehmendem Abstand d zwischen dem Ablösepunkt des Luftstrahls und dem Hindernis. Der Bezugswert τ wurde auf 1.23 cm gesetzt. Die Amplitude der Monopolquelle ist stets proportional zu Φ . Die Konstanten α und β wurden auf die empirisch bestimmten Werte $\alpha = 4 \cdot 10^{-6}$ und $\beta = 5 \cdot 10^{-8}$ gesetzt.

Die Knickfrequenz f_M der Monopolquelle beträgt konstant 1100 Hz, und die Knickfrequenz f_D der Dipolquelle wird analog zu STEVENS [110] mit der Formel $f_D = \gamma \cdot v_c / d_c$ berechnet. Die Proportionalitätskonstante γ wurde auf den Wert 0.4 gesetzt. Für eine vordere Konstriktion wie beim [f] wird für f_D jedoch ein konstanter Wert von 6000 Hz angesetzt. In der zeitdiskreten Simulation werden die Abtastwerte für $\tilde{p}[n]$ und $\tilde{u}[n]$ durch die Filterung von weißem, gaußverteilten Rauschen (Mittelwert 0 und Standardabweichung 1) mit den entsprechenden Butterworth-Tiefpassfiltern gewonnen. Für die Filterung verwende ich FIR-Filter, da diese gegenüber IIR-Filtern auch bei zeitlich veränderlichen Knickfrequenzen bzw. Koeffizienten unbedingt stabil sind. Bei der Implementierung der Gleichungen für Re und f_D muss darauf geachtet werden, dass für v_c ein tiefpass-gefilterter Abtastwert verwendet wird, um die Stabilität der Simulation zu gewährleisten. Ich habe dazu einen digitalen Tiefpassfilter erster Ordnung mit einer Knickfrequenz von 500 Hz implementiert.

Neben den Geräuschquellen an der supraglottalen Konstriktion wird zusätzlich eine einzelne Dipolquelle zwei Rohrabschnitte oberhalb der Glottis eingesetzt, um die Aspiration an der Glottis zu simulieren. Die Skalierung der Amplitude erfolgt mit der Reynoldszahl im oberen Rohrabschnitt der Glottis. Die Experimente in Kapitel 5 haben gezeigt, dass sich mit dem hier beschriebenen Modell für die Geräuschquellen nicht nur Frikative, sondern auch Plosive und Hauchlaute in hoher Qualität simulieren lassen.

3.9 Zusammenfassung und Diskussion

In diesem Kapitel habe ich das Rohrmodell für den Sprechapparat in eine schematische Netzwerkdarstellung überführt und auf dieser Grundlage zwei Methoden für die

aerodynamisch-akustische Simulation vorgestellt. Eine Methode dient der Analyse des Vokaltrakts im Frequenzbereich und die andere Methode der Simulation im Zeitbereich. Das Netzwerk repräsentiert in dieser Arbeit den gesamten Sprechapparat aus dem Vokaltrakt, der Stimmritze, den subglottalen Luftwegen, dem Nasenraum und den Nasennebenhöhlen. Die Nasennebenhöhlen wurden in effizienter Weise als Helmholtz-Resonatoren modelliert und die Rohrabschnitte des stark verzweigten Bronchialbaums in Form von speziellen Rohrverbundabschnitten.

Die für die Simulation im Zeitbereich verwendeten Netzwerkkomponenten unterscheiden sich zum Teil von denen früherer Arbeiten [51, 40, 72]. Der wichtigste Unterschied besteht darin, dass ich in konsistenter Weise sowohl den Bernoulli-Effekt als auch fluidmechanische Energieverluste an plötzlichen Rohrerweiterungen berücksichtige. Dies führt einerseits zu einem natürlicheren Stimmklang durch einen wechselnden Ablösepunkt der Strömung in der Stimmritze [94] und andererseits zur Vermeidung akustischer Artefakte, wenn sich der Konstriktionsort in Frikativen während der kontinuierlichen Artikulation zwischen zwei Rohrabschnitten verschiebt. Ein weiterer Unterschied zu den bisherigen Studien liegt in der einheitlichen Behandlung der Reibungsverluste. Als Reibungswiderstand verwende ich den Strömungswiderstand einer Hagen-Poiseuille-Strömung für *elliptische* Rohrquerschnitte, der sowohl für langgestreckte Rohrquerschnitte, wie die Glottis, als auch für eher runde bzw. kompakte Querschnitte eingesetzt wird. Für hohe Frequenzen und mittlere bis große Rohrquerschnitte ist der Hagen-Poiseuille-Widerstand jedoch kleiner als der exakte frequenzabhängige Grenzschichtwiderstand. Die daraus resultierende zu geringe Dämpfung der höheren Formanten habe ich durch den kontrollierten Einsatz *numerischer Dämpfung* ausgeglichen. Die Verluste durch elastische Rohrwände und die Schallabstrahlung an der Mund- und Nasenöffnung werden analog zu früheren Studien behandelt.

Mit Hilfe der Frequenzbereichsanalyse des Netzwerks habe ich untersucht, welche Eigenschaften der Flächenfunktionen und Geräuschquellen für die Synthese der Frikative wichtig sind. Mit Hilfe der Ergebnisse dieser Untersuchung habe ich ein Verfahren entwickelt, welches automatisch unter geeigneten Bedingungen Geräuschquellen in den Vokaltrakt einfügt. Dieses neue Verfahren ermöglicht die Synthese der stimmlosen und stimmhaften Frikative sowie der Burst- und Aspirationsphase stimmloser Plosive in hoher Qualität.

Kapitel 4

Steuerung der Artikulation

4.1 Einleitung

In diesem Kapitel stelle ich das Steuermodell für die Generierung der artikulatorischen Bewegungsabläufe vor. Der Ansatz basiert auf dem Konzept der artikulatorischen Geste und wurde mit einem Dominanzmodell kombiniert.

Bevor ich das Modell im nächsten Abschnitt ausführlich beschreiben werde, folgt hier zunächst ein kurzer Überblick über bisherige Konzepte zur Steuerung. Ein umfangreicher Literaturüberblick zu Steuermodellen wurde von KRÖGER [66] zusammengestellt. Hier werde ich hauptsächlich die Kernaussagen davon zusammenfassen, um das in dieser Arbeit entwickelte Steuermodell einordnen zu können.

Die Bewegungen der Artikulatoren eines Vokaltraktmodells entstehen durch die zeitabhängige Änderung seiner Steuerparameter. Die Aufgabe eines Steuermodells ist daher die Generierung der Steuerparameter-Zeit-Funktionen. KRÖGER [66] teilt die Steuermodelle anhand der Art, wie die Artikulatorbewegungen beschrieben werden, in *kinematische Modelle*, *zeitinvariante dynamische Modelle* und *zeitvariante dynamische Modelle* ein.

Kinematische Modelle erzeugen die Steuerparameter-Zeit-Kurven direkt. In diesen Modellen werden zu bestimmten Zeitpunkten in einer Äußerung die Steuerparameterwerte vorgegeben (z. B. an den Umkehrpunkten artikulatorischer Bewegungen), und dazwischen wird durch vorgegebene Funktionen interpoliert. Die Unterschiede zwischen kinematischen Modellen liegen hauptsächlich in der Art der Funktionen, die für die Interpolation verwendet werden.

Von den dynamischen zeitinvarianten Steuermodellen wird die Bewegung der Artikulatoren mit Hilfe von Bewegungsgleichungen beschrieben. Den einzelnen Artikulatoren werden dabei bestimmte Ruhepositionen und gleichbleibende (zeitinvariante) Eigenschaften wie Masse, Elastizität und Dämpfung (analog zu einem gedämpften Feder-Masse-System) zugeordnet. Aufgrund dieser Eigenschaften wirken Trägheits-, Reibungs- und Rückstellkräfte auf die Artikulatoren, die mit Hilfe der Bewegungsgleichungen in Beziehung zu den angreifenden Muskelkräften gesetzt werden. Aus der Lösung der Bewegungsgleichungen resultieren die Steuerparameter-Zeit-Funktionen. Artikulatorische Bewegungen werden in diesen Modellen also durch die angreifenden Muskelkräfte gesteuert. Aufgrund der gleichbleibenden physikalischen Eigenschaften der Artikulatoren können die Parameter-Zeit-Funktionen durch die Filterung der angreifenden Kraft-Zeit-Kurven mit einem *zeitinvarianten* Filter

berechnet werden. Die zeitvarianten Steuermodelle verzichten dagegen auf die Annahme der Invarianz der Eigenschaften der Artikulatoren, insbesondere der Elastizität.

Das Prinzip der Beschreibung von Bewegungsabläufen durch dynamische Systeme lässt sich verallgemeinern, indem man nicht die Dynamik einzelner Artikulatoren modelliert, sondern die Dynamik artikulatorischer *Gesten*. Gesten sind laut BROWMAN *et al.* [16] abstrakte Beschreibungen diskreter artikulatorischer Ereignisse, die während der Sprachproduktion auftreten und sich in der Bewegung von Artikulatoren widerspiegeln. Diese Ereignisse bestehen in der Bildung und Lösung von Konstriktionen im Vokaltrakt. Gesten können auch als elementare Einheiten der motorischen Planung angesehen werden. Im Rahmen der *artikulatorischen Phonologie* [16] sind sie außerdem die Grundeinheiten für die Unterscheidung zwischen lexikalischen Elementen.

Wenn Gesten als Grundelemente für die Steuerung eines artikulatorischen Synthetisators verwendet werden, müssen sie einerseits quantitativ beschrieben werden und andererseits mit geeigneten Mitteln in die Steuerparameter-Zeit-Funktionen überführt werden können. Zu diesem Zweck wurde von SALTZMAN *et al.* [98] (siehe auch KELSO *et al.* [55]) das sog. „aufgabendynamische“ Modell (*task-dynamic model*) entwickelt.

Die Simulation koordinierter Artikulatorbewegungen mit dem aufgabendynamischen Modell erfolgt in drei Schritten. Der erste Schritt besteht in der Spezifikation der Gesten. Jede Geste wird explizit als ein Ereignis verstanden, das in der Bildung und Lösung einer Sprechtraktkonstriktion besteht, die durch zwei Parameter beschrieben wird: Die *Position* der Konstriktion im Vokaltrakt und den *Grad* der Enge. Eine komplette Äußerung wird durch eine Konstellation von mehreren dieser Gesten beschrieben.

Der zweite Schritt im aufgabendynamischen Modell besteht in der kinematischen Transformation der abstrakt definierten Gesten in den Raum der Traktvariablen (*tract variables*). Jede Traktvariable charakterisiert eine Dimension einer Konstriktion, also entweder ihre Position oder den Grad der Enge. Im aufgabendynamischen Modell existieren jeweils zwei Traktvariablen für die Konstriktionen mit den Lippen, der Zungenspitze und dem Zungenkörper, und jeweils eine Traktvariable für die glottale und die velische Konstriktion. Ein wichtiges Merkmal des aufgabendynamischen Modells ist, dass nicht die Bewegung einzelner Artikulatoren dynamisch beschrieben wird, sondern die Bewegung der Traktvariablen. Für die Charakterisierung der Bewegungen werden Bewegungsgleichungen kritisch gedämpfter Systeme zweiter Ordnung verwendet.

Der letzte Schritt der Aufgabendynamik besteht in der Überführung der Traktvariablen in die Parameter der Modellartikulatoren. Für jede Traktvariable gibt es jeweils mehrere Modellartikulatoren, mit deren Hilfe die durch die Traktvariable beschriebene Konstriktion realisiert werden kann. So wird z. B. der Grad der Lippenöffnung, der durch eine Traktvariable definiert ist, durch drei Artikulatoren beeinflusst: Den Unterkiefer, die Oberlippe und die Unterlippe. Dadurch gibt es immer mehrere Möglichkeiten für die Positionierung der Modellartikulatoren, um eine vorgegebene Konfiguration der Traktvariablen zu realisieren. Wie stark die unterschiedlichen Artikulatoren bei der Ausbildung einer Konstriktion letztendlich beteiligt werden, wird mit Hilfe einer Artikulator-Gewichtungsmatrix entschieden. Ein anderes Steuermodell, das auch die Geste als grundlegende Einheit der Steuerung ansieht, wurde von KRÖGER [66] entwickelt und implementiert. Die Umsetzung der Gesten in Artikulatorbewegungen unterscheidet sich jedoch von dem oben beschriebenen aufgabendynamischen Ansatz. Aufgrund der Eigenschaften des von KRÖGER verwendeten Artiku-

latformodells wurde auf den Zwischenraum der Traktvariablen verzichtet und die Dynamik der Gesten direkt auf die Modellartikulatoren übertragen. Die Dynamik wurde auch hier mit kritisch gedämpften linearen Systemen zweiter Ordnung modelliert.

4.2 Steuermodell dieser Arbeit

4.2.1 Grundprinzip

In der vorliegenden Arbeit wird analog zu den Ansätzen von SALTZMAN *et al.* [98] und KRÖGER [66] die Geste als grundlegende Einheit der Steuerung angesehen. Die Modellierung der Dynamik der Gesten sowie die Berechnung der Steuerparameter-Zeit-Funktionen unterscheidet sich jedoch von den früheren Ansätzen. Die wichtigsten Unterschiede werden am Ende des Kapitels zusammengefasst.

Die Generierung einer sprachlichen Äußerung erfordert in der Regel mehrere Gesten, die in einer bestimmten Weise zeitlich angeordnet bzw. koordiniert sind. In dieser Arbeit wird die zeitliche Anordnung der Gesten in Form eines *gesturalen Ablaufplans* beschrieben. Der gesturale Ablaufplan wird mit Hilfe des Steuermodells in die zeitlichen Funktionsverläufe der artikulatorischen Steuerparameter überführt. Ausgehend von den Steuerparametern wird dann das akustische Sprachsignal erzeugt. Abbildung 4.1 zeigt das Flussdiagramm für die Berechnung einer Sprachäußerung. In diesem Abschnitt geht es um die Repräsen-



Abbildung 4.1: Flussdiagramm für die Sprachsynthese.

tation einer Äußerung im gesturalen Ablaufplan und die Überführung des Plans in die Steuerparameter-Zeit-Funktionen.

Unter dem Begriff Steuerparameter werden in dieser Arbeit die Parameter des Vokaltraktmodells (Abschnitt 2.3) und die Parameter des Stimmlippenmodells („Anregungsparameter“, siehe Abschnitt 2.5) zusammengefasst. Beide Gruppen von Parametern werden jeweils nur durch bestimmte Typen von Gesten gesteuert. Die Vokaltraktparameter betreffen ausschließlich die supraglottale Artikulation und werden mit Hilfe der *vokalischen*, *konsonantischen* und *velischen* Gesten gesteuert. Die Anregungsparameter definieren dagegen den Grad der glottalen Abduktion/Adduktion, die Grundfrequenz F_0 und den pulmonalen Druck. Sie werden mit Hilfe von *glottalen* Gesten, *pulmonalen* Gesten und *F_0 -Phrasen- und Akzentgesten* gesteuert.

Jede Geste ist durch ein *Aktivitätsintervall* und eine daraus resultierende *gestische Realisierungsfunktion* gekennzeichnet. Die Aktivitätsintervalle können im Sinne der motorischen Planung als (zeitlich ausgedehnte) „Kommandos“ für die Ausführung der Gesten interpretiert werden. Sie werden als Rechteckimpulse mit einem Start- und Endzeitpunkt und einer Amplitude modelliert¹. Eine gestische Realisierungsfunktion beschreibt den zeitlichen Verlauf der Ausprägung einer Geste bzw. den Verlauf des gestischen Realisierungsgrads. Die

¹Einen Spezialfall bilden hier die Gesten für die Phrasenkomponente der Grundfrequenz, die mit Dirac-Impulsen ausgelöst werden.

Realisierungsfunktionen entstehen durch die Tiefpassfilterung der Rechteckimpulse beim Durchlaufen eines abstrakten kausalen Systems. Als Länge und Amplitude einer Geste werden die zeitliche Dauer und die Amplitude des Gestenkommandos bzw. Rechteckimpulses bezeichnet. In den folgenden Unterabschnitten wird die Berechnung der gestischen Realisierungsfunktionen und die Überführung der Gesten in die Steuerparameter-Zeit-Funktionen für die verschiedenen Gestentypen beschrieben.

4.2.2 Supraglottale Steuerung

Ein wichtiges Merkmal der Spracherzeugung ist die Koartikulation. Sie zeigt sich dadurch, dass die artikulatorische Realisierung eines Lautes zum Teil durch die umgebenden Laute beeinflusst wird. Ein offensichtliches Beispiel dafür ist, dass der Lippenverschluss für das [p] in der Äußerung [upu] mit stärker gerundeten Lippen erfolgt als in [ipi]. Der Grad der Lippenrundung spielt also offenbar keine Rolle für die Produktion des [p]. Genausowenig spielt die konkrete Form der Zunge während des [p] eine Rolle (solange mit ihr kein zweiter Verschluss im Vokaltrakt gebildet wird), so dass sie in [ipi] weitestgehend die Form für das [i] beibehalten kann, und in [upu] die Form für das [u]. Wichtig ist für die Produktion des [p] nur, dass die Lippen vollständig verschlossen sind, dass die Glottis geöffnet ist und dass zwischen der Glottis und den Lippen kein weiterer Verschluss existiert.

Was hier für den Laut [p] erläutert wurde, gilt analog für andere Konsonanten. Für jeden Konsonanten gibt es bestimmte Artikulatoren bzw. Freiheitsgrade des Vokaltrakts, die für seine Realisierung entscheidend sind und wiederum andere, die keine oder eine untergeordnete Rolle spielen und dadurch vom Lautkontext beeinflusst werden. Für die Vokale ist dagegen die Form des gesamten Vokaltrakts von Bedeutung und somit die konkrete Form und Position aller Artikulatoren.

Aufgrund des Unterschieds in der Artikulation zwischen Vokalen und Konsonanten werden im Steuermodell dieser Arbeit vokalische und konsonantische Gesten unterschieden (siehe auch BROWMAN *et al.* [16], ÖHMAN [89, 90]). Die vokalischen Gesten realisieren eine grundlegende Bewegung zwischen den invarianten artikulatorischen Zielpositionen der Vokale in einer Äußerung, und die konsonantischen Gesten realisieren die Ausbildung von Konstriktionen durch die Bewegungen jeweils bestimmter Artikulatoren, die sich der Grundbewegung überlagern.

Jede Geste wird durch sieben Parameter vollständig charakterisiert: Ihren Typ (z. B. vokalisches oder konsonantisches), den Startzeitpunkt t_0 , die Amplitude A , die Länge l , die Dauer des Ankling- und Abklingintervalls (T_{an} und T_{ab}) und einen Deskriptor. Die Parameter t_0 , A , l , T_{an} und T_{ab} definieren die gestische Realisierungsfunktion $y(t)$. Diese Funktion ist ein Maß für die artikulatorische Ausprägung einer Geste und ist für zwei Fälle in Abb. 4.2 dargestellt. Die linke Grafik zeigt den Verlauf einer gestischen Realisierungsfunktion mit den Parametern $t_0 = 50$ ms, $A = 1$, $l = 250$ ms, $T_{\text{an}} = 200$ ms und $T_{\text{ab}} = 100$ ms. Sie ist aus einem Anklingintervall, einer stationären Phase und einem Abklingintervall zusammengesetzt. Das Ankling- und Abklingintervall besteht jeweils aus einer halben Sinuswelle mit unterschiedlicher Anfangsphase, und die stationäre Phase verbindet die beiden Intervalle. Das Aktivitätsintervall bzw. das Kommando für die Geste ist als gestrichelter Rechteckimpuls dargestellt.

Die rechte Grafik in Abb. 4.2 zeigt eine Geste, deren Länge kürzer ist als das Anklingintervall ($T_{\text{an}} = 200$ ms und $l = 100$ ms). Dadurch wird das Anklingintervall bereits nach

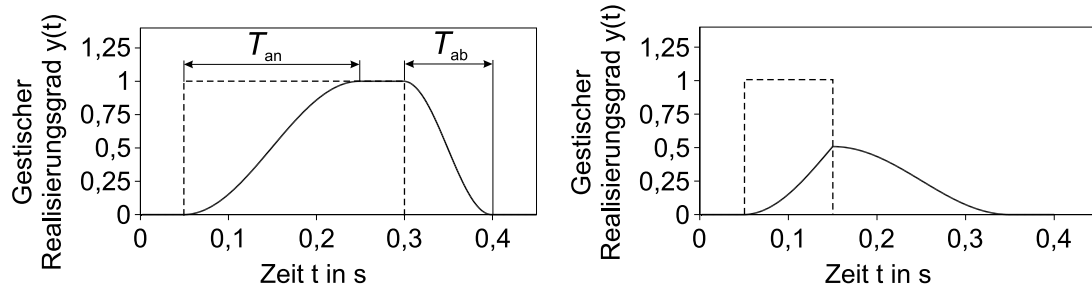


Abbildung 4.2: Zeitverlauf des Realisierungsgrads von zwei Gesten. Die gestrichelten Rechteckimpulse kennzeichnen die gestischen Aktivitätsintervalle.

der Länge l der Geste abgeschnitten, und es folgt ohne eine stationäre Phase direkt das Abklingintervall. Durch die reduzierte Dauer des Anklingintervalls wird die Geste nicht vollständig realisiert. Dieses Prinzip kann unabhängig von der Normamplitude eine reduzierte Ausprägung einer Geste simulieren, die nur aufgrund der Verkürzung der Geste entsteht ($l < T_{an}$).

In beiden Beispielen verhält sich die Funktion $y(t)$ wie die Systemantwort eines kausalen (aber nicht notwendigerweise linearen) Filters auf den Rechteckimpuls. Die Realisierungsfunktion $y(t)$ habe ich wie folgt mathematisch definiert:

$$y(t) = \begin{cases} \frac{A}{2} \cdot \left[1 + \sin\left(-\frac{\pi}{2} + \pi \cdot \frac{t-t_0}{T_{an}}\right) \right], & \text{wenn } 0 \leq t - t_0 \leq \min\{l, T_{an}\} \\ A, & \text{wenn } \min\{l, T_{an}\} \leq t - t_0 \leq l \\ \frac{A}{2} \cdot \left[1 + \sin\left(\frac{\pi}{2} + \pi \cdot \frac{t-t_0-l}{T_{ab}}\right) \right], & \text{wenn } l \leq t - t_0 \leq l + T_{ab} \\ 0 & \text{sonst.} \end{cases} \quad (4.1)$$

Die Amplitude A der vokalischen, konsonantischen und velischen Gesten kann zwischen 0 und 1 eingestellt werden. Dadurch nehmen auch die Realisierungsfunktionen nur Werte zwischen 0 und 1 an.

Jede vokalische und konsonantische Geste ist über ihren Deskriptor mit einer bestimmten *Zielform* des Vokaltrakts verbunden. Eine Zielform ist durch eine feste Belegung der Vokaltraktparameter definiert und entspricht somit einer bestimmten geometrischen Form des Vokaltrakts. Die Zielform eines Vokals entspricht der Vokaltraktform bei der gehaltenen Phonation des Lautes. Für einen Konsonanten definiert die Zielform eine bestimmte „Idealkonfiguration“ der Artikulatoren zum Zeitpunkt der maximalen Ausbildung der konsonantischen Konstriktion. Die Idealkonfiguration entspricht in etwa der Form des Vokaltrakts in der Mitte der Konstriktionsphase, wenn der Konsonant im Kontext des neutralen Vokals [ə] produziert wird. In Abschnitt 5.1 wird näher auf die Erstellung des Inventars von Zielformen und die Einstellung ihrer Parameter eingegangen. Jede Zielform wird über einen Bezeichner (eine kurze ASCII-Zeichenkette) identifiziert. Der Bezeichner entspricht in den meisten Fällen dem phonetischen Symbol für den mit einer Vokaltraktform gebildeten Laut (oder Lautgruppe) in der SAMPA-Notation (siehe Abschnitt B.2). Der Deskriptor einer vokalischen oder konsonantischen Geste enthält jeweils die Bezeichnung der mit ihr verbundenen Zielform.

Wie bereits erwähnt wurde, spielen die einzelnen Artikulatoren bzw. Vokaltraktparameter

für die Generierung eines bestimmten Konsonanten jeweils eine unterschiedlich wichtige Rolle. Je unwichtiger ein Parameter für die Ausbildung der konsonantischen Konstriktion ist, desto stärker wird er durch Koartikulation von den angrenzenden Lauten beeinflusst. Um dieses Prinzip zu simulieren, ist jedem Vokaltraktparameter einer konsonantischen Zielform zusätzlich ein *Dominanzwert* aus dem Intervall $[0, 1]$ zugeordnet. Wenn ein Parameter die Dominanz 0 besitzt, so ist er für den entsprechenden Konsonanten nicht von Bedeutung und wird vollständig durch die überlagerten vokalischen Gesten bestimmt. Ein Dominanzwert von 1 heißt dagegen, dass die Erreichung des Parameterwertes für den Konsonanten unbedingt erforderlich ist. Für die Zielform des Konsonanten $[t]$ besitzen z. B. die Parameter der Zungenspitze (TTX und TTY) die Dominanz 1, damit die alveolare Konstriktion in jedem Fall ausgebildet wird. Die Position des Zungenbeins ist für das $[t]$ dagegen irrelevant, so dass die Dominanz der entsprechenden Parameter (HX und HY) 0 ist. Den Dominanzen für die Konsonanten können auch beliebige Werte zwischen 0 und 1 zugewiesen werden. Ein Wert von 0.4 bedeutet z. B., dass der Parameter zu 60% von den überlagerten Gesten beeinflusst wird. Nach welchen Prinzipien die Dominanzwerte für die konkreten konsonantischen Zielformen eingestellt wurden, wird in Abschnitt 5.1 erläutert.

Außer von den vokalischen und konsonantischen Gesten wird die supraglottale Artikulation zusätzlich von den velischen Gesten gesteuert, die die Position des Velums kontrollieren. Für die individuelle Steuerung des Velums gibt es zwei Gründe. Erstens bewegt sich das Velum oft nicht synchron mit den restlichen supraglottalen Artikulatoren. In einer Plosiv-Vokal-Nasal-Sequenz beginnt z. B. die Absenkbewegung des Velums oft bereits bei der Verschlusslösung des Plosivs und erstreckt sich über den Vokal bis in den Nasal [83]. Die gerichtete Bewegung der an der oralen Verschlussbildung für den Nasal beteiligten Artikulatoren beginnt demgegenüber oft erst später, z. B. in der Mitte des Vokals. Die Bewegung des Velums sollte also zeitlich unabhängig von den anderen Artikulatoren gesteuert werden können. Der zweite Grund für die separate Ansteuerung des Velums ist die Reduktion der Redundanz im Inventar für die konsonantischen Zielformen. Bis auf den Zustand des Velums unterscheiden sich z. B. die Vokaltraktkonfigurationen für die Laute $[p, b, m]$ in der Verschlussphase kaum. Daher können sie durch eine gemeinsame Zielform statt durch drei einzelne Zielformen repräsentiert werden. Welcher der drei Laute mit der gemeinsamen Zielform konkret realisiert wird, hängt dann davon ab, ob gleichzeitig eine velische oder glottale Öffnungsgeste aktiv ist.

Für die Spezifikation der supraglottalen Artikulation einer Sprachäußerung müssen die dazu benötigten vokalischen, konsonantischen und velischen Gesten zeitlich in einer bestimmten Weise angeordnet werden. In Abb. 4.3 ist als Beispiel solch eine Anordnung für das Wort „Einbau“ dargestellt. Zusammen mit den laryngalen und pulmonalen Gesten (siehe Abschnitt 4.2.3) bildet diese Anordnung den gesturalen Ablaufplan. Der Ablaufplan für das Wort „Einbau“ enthält vier vokalische Gesten für die Laute $[a]$, $[i]$, $[a]$ und $[u]$, zwei konsonantische Gesten für die Laute $[n]$ und $[b]$, und eine velische Öffnungsgeste für die Ankopplung des Nasenaums während des Nasals. Die Gesten sind in Form ihrer Realisierungsfunktionen $y_x(t)$ dargestellt, wobei der Index x die einzelnen Gesten und Gestentypen unterscheidet.

Die Überführung des gesturalen Ablaufplans in die Steuerparameter-Zeit-Funktionen erfolgt, indem jeweils in kurzen Abständen von 10 ms die Wertebelegung der Steuerparameter berechnet und dazwischen linear interpoliert wird. Die Berechnung der Vokaltraktparame-

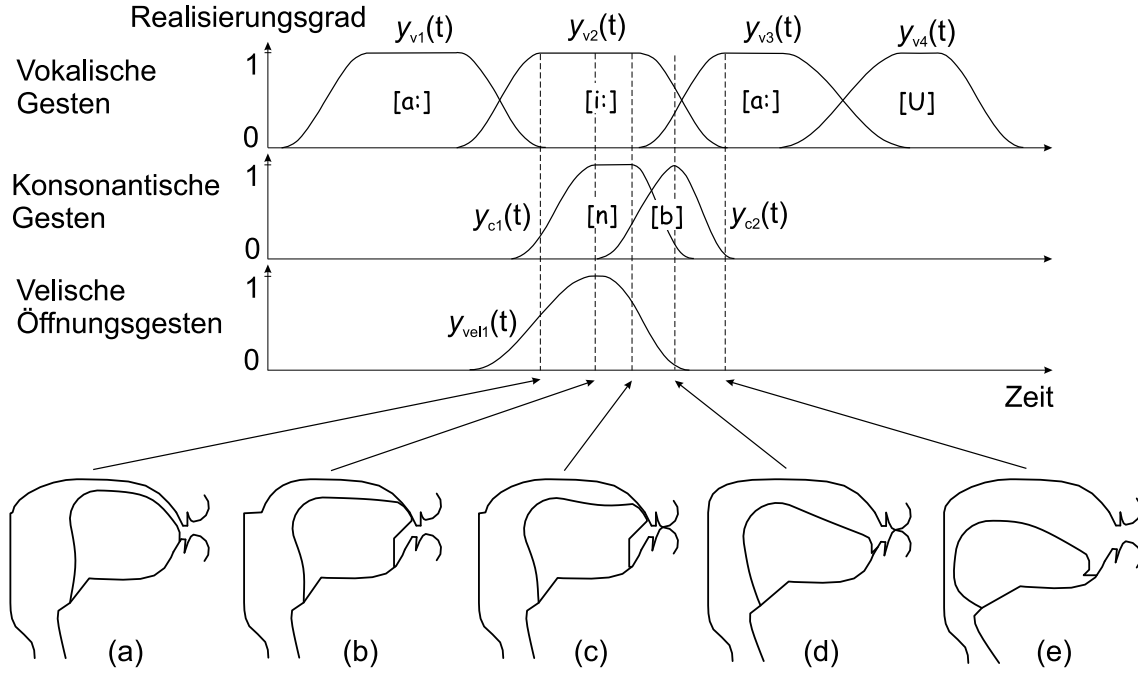


Abbildung 4.3: Anordnung der vokalischen, konsonantischen und velischen Gesten für das Wort „Einbau“. Unten ist die resultierende Kontur des Vokaltrakts an fünf Positionen im Bereich des Konsonantenclusters [nb] dargestellt.

ter mit Hilfe der supraglottalen Gestentypen wird nachfolgend erläutert.

Angenommen, zum betrachteten Zeitpunkt t sind M Vokale mit den gestischen Realisierungsgraden $y_{v_1}(t) \dots y_{v_M}(t)$ überlagert (siehe Abb. 4.3). Die Vokaltraktparameterwerte der M Vokale sind als $v_1(i) \dots v_M(i)$ durch die mit den Gesten verbundenen Zielformen gegeben, wobei $i = 1 \dots N$ der Parameterindex und N die Parameteranzahl ist. Die aus der Überlagerung resultierenden Vokaltraktparameter $v(i, t)$ ergeben sich durch die gewichtete Mittelwertbildung der beteiligten Vokale, wobei das Gewicht jedes Vokals durch seinen gestischen Realisierungsgrad gegeben ist:

$$v(i, t) = \left[\sum_{k=1}^M y_{v_k}(t) \cdot v_k(i) \right] / \left[\sum_{k=1}^M y_{v_k}(t) \right].$$

In dieser Berechnungsvorschrift steckt die vereinfachende Annahme, dass sich die supraglottalen Artikulatoren stets synchron bewegen. Eine Ausnahme bildet das Velum, das durch separate Gesten gesteuert wird.

Wenn keine vokalischen Gesten aktiv sind ($y(t) = 0$), so ist die Form des Vokaltrakts durch die Parameter $n(i)$ des neutralen Vokals [ə] gegeben. Wenn jedoch vokalische Gesten aktiv sind, so wird der neutrale Vokal mit zunehmendem Realisierungsgrad dieser Gesten ersetzt. Die resultierenden Parameter dieser Überlagerung werden mit $p^{(0)}(i, t)$ bezeichnet und wie folgt berechnet:

$$\begin{aligned} p^{(0)}(i, t) &= y_v(t) \cdot v(i, t) + [1 - y_v(t)] \cdot n(i) \\ y_v(t) &= \min\{ y_{v_1}(t) + y_{v_2}(t) + \dots + y_{v_M}(t), 1 \}. \end{aligned}$$

$y_v(t)$ ist dabei der kummulierte Realisierungsgrad der vokalischen Gesten zum betrachteten Zeitpunkt t .

Die vokalische Grundform des Vokaltrakts mit den Parametern $p^{(0)}(i, t)$ wird ihrerseits durch eventuell vorhandene konsonantische Gesten überformt. Inwiefern eine konsonantische Geste die vokalische Grundform verändert, hängt sowohl vom Realisierungsgrad der Geste als auch von den Dominanzwerten der einzelnen Parameter der konsonantischen Zielform ab. Hierbei muss auch der Fall berücksichtigt werden, dass sich zum betrachteten Zeitpunkt zwei oder mehr konsonantische Gesten überlagern. In diesem Fall ist es erforderlich, die Koartikulation von Konsonanten untereinander realitätsnah zu simulieren! Was diese Anforderung konkret bedeutet, habe ich anhand des Wortes „Einbau“ in Abb. 4.3 veranschaulicht. Dort ist die Kontur des Vokaltrakts zu fünf Zeitpunkten im Bereich des Konsonantenclusters [nb] dargestellt. Die erste Kontur zeigt den Vokaltrakt kurz nach dem Beginn der Geste für das [n] und entspricht noch weitestgehend der Kontur für den Vokal [i]. In der Kontur (b) ist das Velum vollständig abgesenkt und der alveolare Verschluss im Mundraum hergestellt, so dass ein [n] generiert wird. Damit ein lückenloser Übergang zwischen dem [n] und dem [b] entsteht, darf nun aber nicht zuerst der alveolare Verschluss vom [n] gelöst und anschließend der bilabiale Verschluss für das [b] hergestellt werden, sondern es muss zuerst der bilabiale Verschluss gebildet werden, und anschließend darf der alveolare Verschluss gelöst werden (Konturen (c) und (d)). Ansonsten entstünde kurzzeitig eine durchgängige Passage im Mundraum, die einen kurzen vokalähnlichen Laut zwischen [n] und [b] verursachen würde. Um das zu verhindern, werden überlappende konsonantische Gesten wie folgt behandelt.

Für jeden einzelnen Vokaltraktparameter werden die gleichzeitig aktiven konsonantischen Gesten in aufsteigender Reihenfolge nach den Dominanzwerten für den aktuellen Parameter sortiert. Angefangen mit dem kleinsten Dominanzwert werden die Parameterwerte der mit den Gesten verbundenen Zielformen den Parametern $p^{(0)}(i, t)$ der vokalischen Grundform des Vokaltrakts überlagert. Ein alter Parameterwert wird dabei umso mehr durch den Parameterwert einer konsonantischen Zielform ersetzt, je größer das Produkt aus dem Realisierungsgrad der konsonantischen Geste und dem Dominanzwert des entsprechenden Parameters ist. An der Position (c) in Abb. 4.3 werden z. B. die Vokaltraktparameter der Zungenspitze durch die [n]-Geste dominiert und die Parameter für die Lippen durch die [b]-Geste. Dadurch werden beide zu dem Zeitpunkt erforderlichen Verschlüsse im Vokaltrakt ausgebildet. Der folgende Algorithmus formalisiert die Berechnungsvorschrift.

```

for  $i = 1$  to  $N$                                 Durchlauf der Parameter
    Sortiere die konsonantischen Gesten in aufsteigender Reihenfolge
    nach den Dominanzwerten des aktuellen Parameters  $i$ , und lege die
    Indizes der Gesten in der Liste  $j(k)$  mit  $k = 1 \dots K$  ab.

    for  $k = 1$  to  $K$                                 Durchlauf der konsonantischen Gesten
         $\mu = y_{c_{j(k)}}(t) \cdot d_{j(k)}(i)$ 
         $p^{(k)}(i, t) = (1 - \mu) \cdot p^{(k-1)}(i, t) + \mu \cdot c_{j(k)}(i)$ 
    Ende der  $k$ -Schleife

Ende der  $i$ -Schleife

```

Darin ist K die Anzahl der aktiven konsonantischen Gesten. $c_1(i) \dots c_K(i)$ sind die Para-

meter und $d_0(i) \dots d_K(i)$ die Dominanzen der mit diesen Gesten verbundenen Zielformen. Der gestische Realisierungsgrad wurde mit $y_{c_1}(t) \dots y_{c_K}(t)$ bezeichnet (vgl. Abb. 4.3). Die endgültigen Parameterwerte stehen in den Variablen $p^{(K)}(i, t)$.

Der Wert des Parameters VO , der die Öffnung der velopharyngealen Pforte steuert, wird außer durch die vokalischen und konsonantischen Gesten auch noch durch die velischen Gesten beeinflusst. Wenn der Wert von VO nach der Auswertung der vokalischen und konsonantischen Gesten mit VO_{alt} bezeichnet wird² und $y_{\text{vel}}(t)$ die Summe der Realisierungsgrade der velischen Gesten zum Zeitpunkt t ist, dann wird der endgültige Parameterwert VO_{neu} mit der Gleichung

$$VO_{\text{neu}} = VO_{\text{alt}} + y_{\text{vel}}(t) \cdot (VO_{\text{max}} - VO_{\text{alt}})$$

berechnet. VO_{max} ist der Parameterwert für die maximale Öffnung der velopharyngealen Pforte.

4.2.3 Laryngale und pulmonale Steuerung

Für die laryngale und pulmonale Ansteuerung des Synthetisators wurden vier Gestentypen eingeführt: Glottale Gesten, F_0 -Phrasen- und Akzentgesten und pulmonale Gesten. Sie wirken ausschließlich auf die Anregungsparameter (siehe Abschnitt 2.5).

Der Realisierungsgrad der pulmonalen Gesten wird analog zu den supraglottalen Gesten berechnet und steuert unmittelbar den pulmonalen Druck p_{Lunge} . Die Amplitude einer pulmonalen Geste liegt zwischen 0 und 1200 Pa, wobei 800 Pa der typische Druck für eine mittlere Sprechlautstärke ist. Wenn sich zwei oder mehr pulmonale Gesten überlagern, so addieren sich ihre Realisierungsgrade. In aller Regel genügt für den pulmonalen Druck jedoch eine einzelne Geste, die sich vom Anfang bis zum Ende der Äußerung erstreckt.

Mit den glottalen Gesten werden die Parameter ζ_{01} und ζ_{02} des Stimmlippenmodells gesteuert. Diese Parameter definieren die Auslenkung des unteren und oberen Randes der Stimmlippen an ihrem hinteren Ende und kontrollieren somit die Weite der glottalen Öffnung. In dieser Arbeit werden zwei Arten von glottalen Gesten unterschieden: Abduktionsgesten und Adduktionsgesten. Der Deskriptor einer glottalen Geste entscheidet, um welche Art es sich handelt. Für Abduktionsgesten enthält der Deskriptor die Zeichenkette „open“ und für Adduktionsgesten die Zeichenkette „close“. Die Amplitude der Gesten kann zwischen 0 und 1 eingestellt werden, und ihre Realisierungsfunktionen werden analog zu den supraglottalen Gesten berechnet. Wenn zu einem bestimmten Zeitpunkt weder eine Abduktions- noch eine Adduktionsgeste aktiv ist, so wird davon ausgegangen, dass sich die Stimmlippen in Phonationsstellung befinden. Wenn der Realisierungsgrad einer Abduktionsgeste jedoch größer als 0 ist, so wird die Weite der Glottis proportional zum Realisierungsgrad bis auf einen Maximalwert erhöht. Analog wird die Weite der Glottis bei einer Adduktionsgeste in Abhängigkeit vom Realisierungsgrad bis auf einen Minimalwert reduziert. Die Gleichung zur Berechnung der Auslenkung der Stimmlippen lautet

$$\zeta_{01} = \zeta_{02} = \zeta_{\text{pho}} + y_{\text{open}}(t) \cdot (\zeta_{\text{max}} - \zeta_{\text{pho}}) + y_{\text{close}}(t) \cdot (\zeta_{\text{min}} - \zeta_{\text{pho}}),$$

wobei $y_{\text{open}}(t)$ und $y_{\text{close}}(t)$ jeweils die (kummulierten) Realisierungsgrade der Abduktions- bzw. Adduktionsgesten sind³. ζ_{pho} , ζ_{max} und ζ_{min} sind die Auslenkungen für die Phonati-

²Die Zielformen der vokalischen und konsonantischen Gesten haben jeweils eine geschlossene velopharyngale Pforte.

³In der Regel ist zu einem bestimmten Zeitpunkt immer nur maximal eine glottale Geste aktiv.

onsstellung, die maximale Öffnung der Glottis und den maximalen Verschluss der Glottis. Sie wurden auf die Werte $\zeta_{\text{pho}} = 0.3$ mm, $\zeta_{\text{max}} = 3$ mm und $\zeta_{\text{min}} = -0.5$ mm gesetzt. Bei der fortgeschrittenen Realisierung einer Adduktionsgeste wird die Auslenkung der Stimmlippen *negativ*, was letztendlich zu einem festen Verschluss der Stimmritze führt. Auf diese Weise können glottale Verschlusslaute erzeugt werden (siehe Abschnitt 2.5).

Die Parameter für die Größe des zusätzlichen Spalts zwischen den Stellknorpeln und die Phasendifferenz zwischen dem unteren und oberen Rand der Stimmlippen werden gegenwärtig nicht aktiv von Gesten gesteuert. Für diese Parameter wurden bei den Simulationen die konstanten Werte $\Delta A_{\text{Spalt}} = 0.0$ cm² und $\phi = 50.4^\circ$ verwendet.

Die Generierung der Grundfrequenzkontur basiert auf dem Überlagerungsmodell von FUJISAKI *et al.* [42]. Dieses Modell wurde ursprünglich für die Nachbildung und Synthese japanischer F_0 -Konturen entwickelt und später auch auf andere Sprachen wie Englisch, Deutsch und Schwedisch angewendet [43, 44]. Eine F_0 -Kontur wird in diesem Modell durch die Überlagerung eines asymptotischen Basiswertes F_{min} mit *Phrasen- und Akzentkomponenten* generiert. Die Phrasenkomponenten dienen der Modellierung des langsamen Abfalls der F_0 während einzelner Phrasen („Deklinaton“), und die Akzentkomponenten dienen der Modellierung der Wortakzente.

Der Logarithmus der F_0 wird in dem Modell wie folgt berechnet:

$$\ln F_0(t) = \ln F_{\text{min}} + \sum_{i=1}^M y_{\text{p},i}(t) + \sum_{i=1}^N y_{\text{a},i}(t).$$

Darin sind die Funktionen $y_{\text{p},i}(t)$ und $y_{\text{a},i}(t)$ die Phrasen- und Akzentkomponenten. Jede Phrasen- und Akzentkomponente ist die Antwort eines kritisch gedämpften linearen Systems zweiter Ordnung auf ein bestimmtes Eingangssignal. Die Eingangssignale sind Dirac-Impulse für die Phrasenkomponenten und Rechteckimpulse für die Akzentkomponenten. Die Impulsantwort eines kritisch gedämpften linearen Systems zweiter Ordnung mit der Kennfrequenz α_i lautet

$$g_i(t) = \begin{cases} \alpha_i^2 \cdot t \cdot e^{-\alpha_i \cdot t} & \text{für } t > 0 \\ 0 & \text{sonst,} \end{cases}$$

und die Sprungantwort für solch ein System mit der Kennfrequenz β_i ist

$$h_i(t) = \begin{cases} 1 - (1 + \beta_i \cdot t) \cdot e^{-\beta_i \cdot t} & \text{für } t > 0 \\ 0 & \text{sonst.} \end{cases}$$

Eine Phrasenkomponente ist direkt eine verschobene und skalierte Impulsantwort, d. h.

$$y_{\text{p},i}(t) = A_{\text{p},i} \cdot g_i(t - t_{\text{p},i}),$$

wobei $A_{\text{p},i}$ die Amplitude und $t_{\text{p},i}$ der Zeitpunkt des auslösenden Dirac-Impulses sind. Eine Akzentkomponente ist dagegen die Systemantwort auf einen Rechteckimpuls und kann als Summe zweier verschobener Sprungantworten berechnet werden:

$$y_{\text{a},i}(t) = A_{\text{a},i} \cdot [h_i(t - t_{\text{a},i}) - h_i(t - t_{\text{a},i} - l_{\text{a},i})].$$

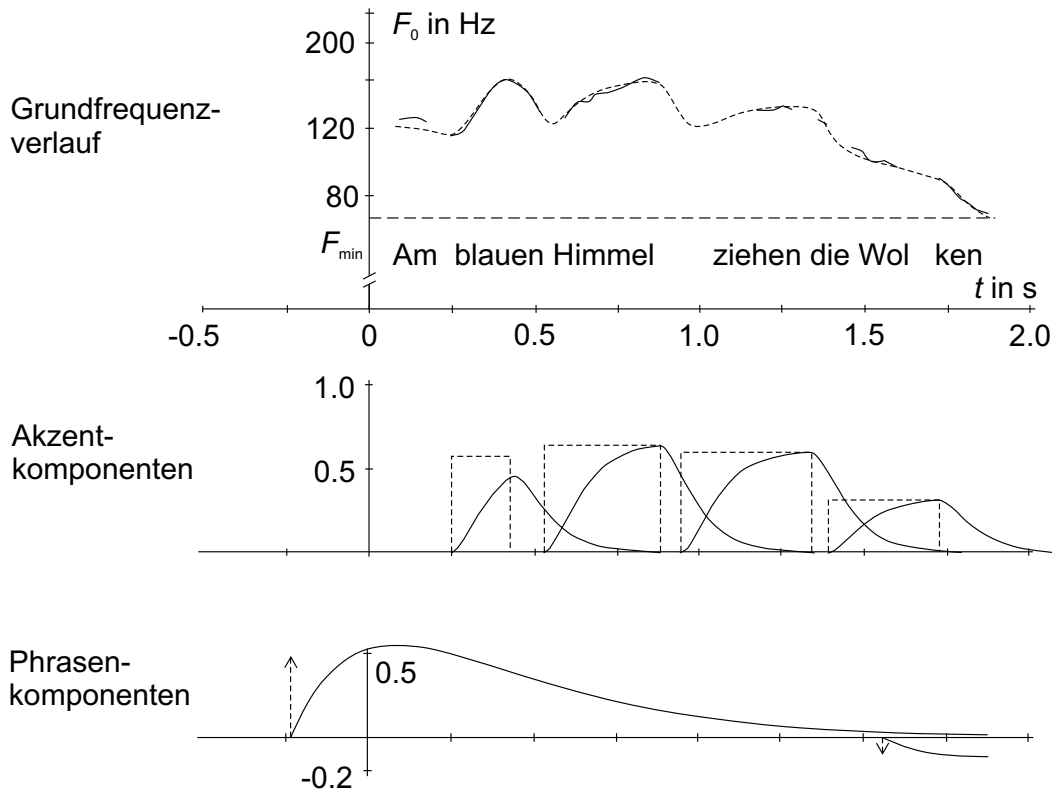


Abbildung 4.4: Überlagerungsmodell zur Berechnung der Grundfrequenzkontur von FUJISAKI *et al.* [42]. Dargestellt ist die Nachbildung der F_0 -Kontur für den Satz „Am blauen Himmer ziehen die Wolken.“ nach MÖBIUS [81].

Darin ist $A_{a,i}$ die Amplitude, $t_{a,i}$ der Startzeitpunkt und $l_{a,i}$ die Länge des Rechteckimpulses. Die Kennfrequenzen α_i und β_i der Phrasen- und Akzentkomponenten sind im FUJISAKI-Modell prinzipiell frei wählbar.

Abbildung 4.4 zeigt, wie durch die Überlagerung der Phrasen- und Akzentkomponenten die F_0 eines Satzes nachgebildet werden kann. Im oberen Graphen sind die durchgezogenen Linienabschnitte die gemessene F_0 -Kontur des Satzes, und die gestrichelte Linie ist die synthetische F_0 -Kontur. Darunter sind jeweils die dafür aufsummierten Phrasen- und Akzentkomponenten zusammen mit den auslösenden Rechteck- und Dirac-Impulsen dargestellt.

Neben dem FUJISAKI-Modell gibt es zahlreiche andere Ansätze für die Generierung von F_0 -Konturen für die Sprachsynthese (siehe z. B. MIXDORFF [80] für eine Zusammenfassung). Dass ich gerade dieses Modell für die vorliegende Arbeit ausgewählt habe, hat mehrere Gründe. Erstens wurde in verschiedenen Veröffentlichungen gezeigt, dass sich das FUJISAKI-Modell grundsätzlich für die Nachbildung der F_0 in vielen verschiedenen Sprachen eignet. Zweitens wurden von MÖBIUS [81] und MIXDORFF [80] bereits speziell für die deutsche Sprache Regeln formuliert, mit denen die Parameter der Phrasen- und Akzentkomponenten automatisch berechnet werden können. Dies ist in Hinblick auf eine Weiterentwicklung des Synthetisators zu einem vollständigen TTS-System interessant. Drittens

ist das Modell neben der mathematischen Beschreibung auch physiologisch interpretierbar und somit artikulatorisch basiert [80]. Weiterhin wird eine leicht modifizierte Variante des FUJISAKI-Modells in einem der momentan am natürlichsten klingenden konkatentativen Sprachsynthesatoren verwendet (das Bell-Labs-TTS-System). Letztendlich passt das Modell konzeptionell zu dem in dieser Arbeit gewählten gestischen Ansatz zur Steuerung. Dazu können die Rechteckimpulse und Dirac-Impulse analog zu den anderen Gesten im Steuermodell als die Aktivitätsintervalle bzw. Kommandos der Gesten angesehen werden und die Signale am Ausgang der kritisch gedämpften Systeme zweiter Ordnung als gestische Realisierungsfunktionen.

In der vorliegenden Arbeit werden die F_0 -Akzentkomponenten nicht nur wie im Originalmodell von FUJISAKI *et al.* [42] zur Modellierung der Wortakzente benutzt, sondern auch zur Nachbildung mikroprosodischer Grundfrequenzvariationen (siehe Abschnitt 5.2).

4.2.4 Gesturaler Ablaufplan

Ein gesturaler Ablaufplan enthält die Informationen zu allen an einer Äußerung beteiligten Gesten. Die Ablaufpläne müssen gegenwärtig manuell mit Hilfe einer grafischen Bedienoberfläche erstellt werden. In Zukunft wäre aber eine automatische, z. B. regelbasierte, Generierung von Ablaufplänen denkbar. Im nächsten Kapitel sind einige Regeln zusammengestellt, die die manuelle Erstellung von Ablaufplänen unterstützen.

Abbildung 4.5 zeigt den Ablaufplan für das Wort „tanzen“. Die Äußerung wurde mit einer

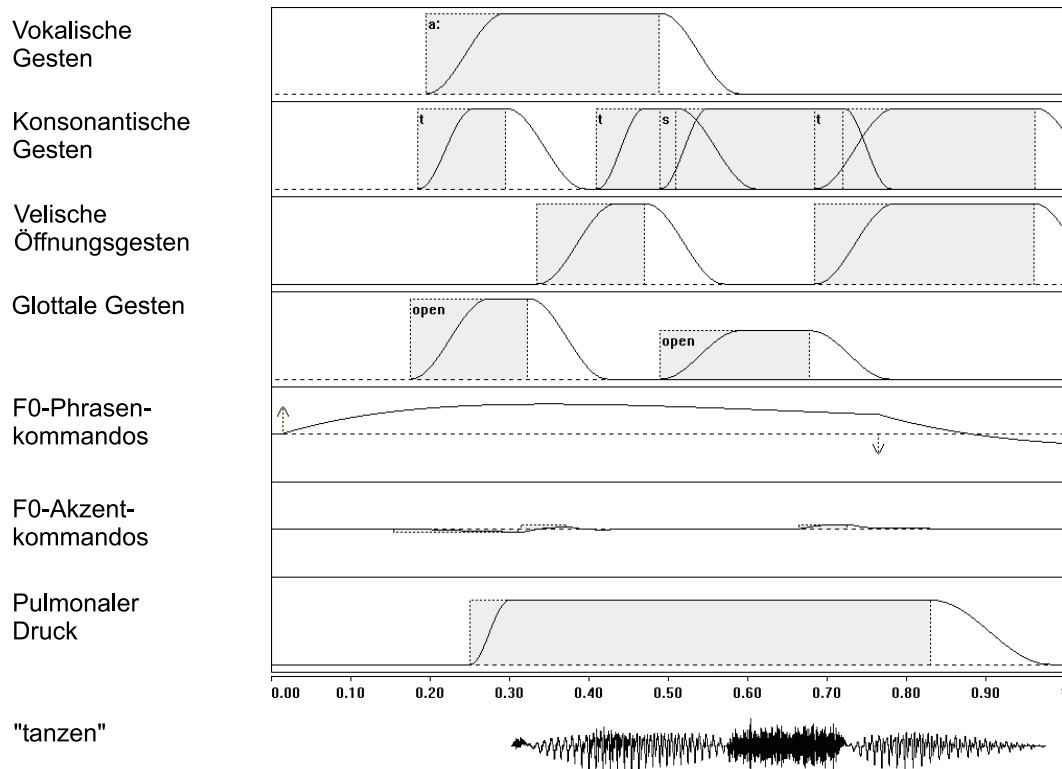


Abbildung 4.5: Gesturaler Ablaufplan für das Wort „tanzen“.

vokalischen Geste für den Vokal [a:] und vier konsonantischen Gesten für die Konsonanten [t], [n], [s] und [ŋ] modelliert. Die drei Konsonanten am Wortende überlappen sich zeitlich derart, dass sie einen zusammenhängenden Konsonantencluster bilden. Man beachte, dass für die beiden [n] die gleiche Zielform wie für das [t] verwendet wurde (Bezeichner in der linken oberen Ecke der Rechteckimpulse), da sie sich bis auf den Zustand des Velums kaum unterscheiden. Die Öffnung der velopharyngealen Pforte wurde mit Hilfe von zwei velischen Öffnungsgesten modelliert, die sich zeitlich mit den konsonantischen Gesten für die Nasale überlappen. Für die Realisierung des [t] und des [s] wurden zwei glottale Abduktionsgesten eingesetzt, wobei die maximale Abduktion für den Plosiv größer ist als für den Frikativ. Die F_0 -Kontur wurde mit 2 Phrasen- und 3 Akzentkomponenten modelliert. Die Akzentkomponenten dienen hier nicht nur der Nachbildung des Akzents auf der ersten Silbe, sondern auch der Modellierung koartikulatorischer F_0 -Variationen (siehe Abschnitt 5.2). Für den pulmonalen Druck wurde eine Geste verwendet, die sich über die gesamte Äußerung erstreckt und einen Druck von 800 Pa erzeugt.

4.2.5 Technische Realisierung

Für die Synthese eines Sprachsignals werden alle 10 ms die Steuerparameter aus dem gesturalen Ablaufplan berechnet und damit, über den Zwischenschritt des Vokaltrakt- und Glottismodells, die Geometrie des Rohrmodells für den Sprechapparat. Für die aerodynamisch-akustische Simulation werden die Parameter des Rohrmodells (Länge, Querschnitt und Umfang der einzelnen Rohrabschnitte) zu jedem Abtastzeitpunkt neu durch die lineare Interpolation zwischen den Rohrgeometrien an den nächstgelegenen linken und rechten 10 ms-Positionen interpoliert. Durch die Überführung des 3D-Vokaltraktmodells in das diskrete Rohrmodell kann es in bestimmten Situationen passieren, dass die Änderungsrate des Querschnitts einzelner Rohrabschnitte zu bestimmten Zeitpunkten einer Äußerung unnatürlich groß wird. Diese schnellen Änderungen können Störungen in Form von Klicktönen im akustischen Ausgabesignal erzeugen. Um das zu verhindern, habe ich für jeden einzelnen Rohrabschnitt einen digitalen Tiefpassfilter erster Ordnung mit einer Kennfrequenz von 100 Hz eingeführt, der sehr plötzliche Änderungen des Querschnitts zwischen benachbarten Abtastzeitpunkten „glättet“.

4.3 Zusammenfassung und Diskussion

In diesem Kapitel wurde das Steuermodell für die Generierung der artikulatorischen Bewegungsabläufe vorgestellt. Die grundlegende Einheit der Steuerung ist in diesem Modell die Geste. Die Eingabe für das Steuermodell ist der gesturale Ablaufplan, in dem die Gesten für eine Äußerung zeitlich miteinander koordiniert sind. Die Ausgabe des Steuermodells sind die zeitlichen Funktionsverläufe der Steuerparameter für das Vokaltrakt- und Stimmlippenmodell. Das in dieser Arbeit entwickelte Konzept zur Steuerung kombiniert Elemente früherer gestischer Steuermodelle (aufgabendynamisches Modell von SALTZMAN *et al.* [98], gestisches Modell von KRÖGER [66]) mit dem Prinzip der Dominanz einzelner Artikulatoren⁴. Es wurden Mechanismen entwickelt, um auf der Basis der gestischen Spezifikation

⁴Dominanzmodelle wurden bisher vorrangig verwendet, um die Koartikulation in der visuellen Sprachsynthese, d. h. der Gesichtsanimation, zu steuern (siehe z. B. FAGEL [31] und KING [57]).

einer Äußerung sowohl die Koartikulation zwischen Vokalen und Konsonanten als auch die Koartikulation zwischen Konsonanten untereinander zu simulieren.

Nachfolgend werden kurz die wichtigsten Gemeinsamkeiten und Unterschiede des Steuermodells dieser Arbeit im Vergleich zu den verwandten gestischen Steuermodellen von SALTZMAN *et al.* [98] und KRÖGER [66] zusammengefasst. Alle drei Steuermodelle überführen eine bestimmte Anordnung artikulatorischer Gesten in die Steuerparameter-Zeit-Funktionen eines Vokaltraktmodells. Dabei wird jede Geste als koordinierte Bewegung von einem oder mehreren Artikulatoren realisiert. In allen betrachteten Steuermodellen werden verschiedene Gestentypen unterschieden, insbesondere vokalische, konsonantische, velische und glottale Gesten.

Das hier vorgestellte Modell unterscheidet sich von den beiden früheren Ansätzen dadurch, dass die Dynamik einer Geste nicht als Reaktion eines kritisch gedämpften Systems zweiter Ordnung auf einen Rechteckimpuls modelliert wird (Ausnahme: F_0 -Akzentgesten). KRÖGER [66] hat gezeigt, dass es mit den dabei entstehenden exponentiellen Zeitfunktionen nur sehr ungenau gelingt, gemessene Auslenkungs-Zeit-Kurven von Artikulatoren nachzubilden. In der vorliegenden Arbeit wird der Realisierungsgrad einer Geste statt dessen direkt als Funktionsverlauf vorgegeben. Die Parameter für die Spezifikation der Realisierungsfunktion sehen dabei im Unterschied zu den Steuermodellen von SALTZMAN *et al.* und KRÖGER auch unterschiedliche zeitliche Längen für die An- und Abklingintervalle einer Geste vor. Dadurch können mit einer *einzelnen* Geste unterschiedliche Geschwindigkeiten bei der Bildung und Lösung einer Konstriktion simuliert werden (siehe Abschnitt 5.2).

Ein weiterer wesentlicher Unterschied zu den früheren Ansätzen ist die dominanzbasierte Simulation von Koartikulation und die Behandlung von zeitlich überlappenden Gesten, die auf den gleichen Artikulator wirken. Im KRÖGER-Modell wird die Steuerung eines Artikulators bei der zeitlichen Überlappung zweier Gesten jeweils von einer Geste vollständig dominiert, wobei die dominierende Geste mit Hilfe von Prioritätsregeln ermittelt wird. Im aufgabendynamischen Modell tritt bei der zeitlichen Überlappung von zwei denselben Artikulator steuernden Gesten dagegen eine Mischung ein. In der vorliegenden Arbeit wird der Einfluss, den überlappende Gesten auf die Steuerung desselben Artikulators ausüben, mit Hilfe der Dominanzwerte bestimmt.

Wie auch der Ansatz von KRÖGER verzichtet das vorliegende Steuermodell im Gegensatz zum aufgabendynamischen Modell auf die Ebene der Traktvariablen. Statt dessen wird die dynamische Aktivität der Gesten direkt in die Steuerparameter-Zeit-Funktionen überführt. Dadurch bleibt das Steuermodell verhältnismäßig einfach und übersichtlich. Durch die Integration eines Überlagerungsmodells für die Synthese der Grundfrequenzkonturen wurde der gestische Ansatz auch auf die Intonationssteuerung ausgedehnt.

Für die Erstellung gesturaler Ablaufpläne wurde eine grafische Schnittstelle entwickelt, mit der die Parameter der Gesten und ihre zeitliche Anordnung interaktiv eingestellt werden können.

Kapitel 5

Synthese

Dieses Kapitel behandelt zwei Themen mit Bezug zur Synthese zusammenhängender Sprachäußerungen. Das erste Thema ist der Erstellung der Zielformen für die vokalischen und konsonantischen Gesten und wird in Abschnitt 5.1 behandelt. Das zweite Thema ist die Erstellung gesturaler Ablaufpläne, d. h. die zeitliche Koordinierung der unterschiedlichen Gesten für eine Äußerung und die Einstellung ihrer Parameter. In Abschnitt 5.2 werden Regeln für die Erstellung von Ablaufplänen formuliert und anhand von Beispielen demonstriert.

5.1 Zielformen des Vokaltrakts

5.1.1 Erstellung einer Lauttabelle

Jede vokalische und konsonantische Geste ist über einen Deskriptor mit einer bestimmten *Zielform* des Vokaltrakts verbunden, auf die bei der Berechnung der Parameter-Zeit-Funktionen im Steuermodell zugegriffen wird. Jede Zielform ist durch eine bestimmte Belegung der Vokaltraktparameter definiert, und konsonantische Zielformen besitzen zusätzlich einen Dominanzwert pro Parameter. Alle Zielformen wurden mit ihrem Bezeichner als Schlüssel in einer Tabelle abgelegt.

Für jeden der deutschen Vokal [a:, e:, i:, o:, u:, ɛ:, ø:, y:, ɪ, ɛ, a, ɔ, ʊ, ʏ, œ, ə, ɐ] wurde genau eine Zielform erstellt. Als Bezeichner für die Vokale wurden ihre entsprechenden SAMPA-Symbole (Symbole aus ASCII-Zeichen für die Laute des internationalen Phonetischen Alphabets; siehe Abschnitt B.2) verwendet. Für die deutschen Konsonanten wurden teilweise mehrere Laute gruppiert und durch eine gemeinsame Zielform repräsentiert. Tabelle 5.1 zeigt die Konsonanten zusammen mit den Symbolen für ihre Zielformen. Von den Frika-

Gruppe	Symbol	Gruppe	Symbol	Gruppe	Symbol
[p, b, m]	p	[f, v]	f	[ç, j]	C
[t, d, n]	t	[s, z]	s	[x, ʀ]	x
[k, g, ŋ]	k	[ʃ, ʒ]	S	[l]	l

Tabelle 5.1: Die Konsonanten der deutschen Sprache und die Symbole für die entsprechenden Zielformen des Vokaltrakts.

tiven wurden jeweils die stimmhaften und stimmlosen Laute derselben Artikulationsstelle durch eine gemeinsame Vokaltraktform repräsentiert, da sie sich in ihrer supraglottalen Artikulation nur unwesentlich unterscheiden¹. Jede Form wurde mit dem SAMPA-Symbol für den stimmlosen Frikativ in der Gruppe bezeichnet. Weiterhin habe ich jeweils die Plosive und Nasale derselben Artikulationsstelle zusammengefasst. Auch sie unterscheiden sich, abgesehen vom Velum, nicht wesentlich in der supraglottalen Artikulation und wurden durch eine Vokaltraktform mit angehobenem Velum modelliert. Die Steuerung des Velums für die Nasale erfolgt im gesturalen Ablaufplan unabhängig von den konsonantischen Gesten durch separate velische Öffnungsgesten. Als Bezeichner für die Zielformen wurden hier die SAMPA-Symbole der stimmlosen Plosive in den Gruppen gewählt.

Als Vorlage für die Erstellung der Vokaltraktformen dienten sagittale Konturbilder des Sprechtrakts, die von KONECZNA *et al.* [65] mit Hilfe von Röntgenaufnahmen erstellt wurden. Die Sammlung von KONECZNA *et al.* enthält insgesamt 262 Bilder von vier russischen Sprechern bei der Artikulation russischer Vokale und Konsonanten in verschiedenen Kontexten. Davon wurden die subjektiv „besten“ Aufnahmen eines Sprechers als Vorlagen ausgewählt.

Die Zielformen habe ich erstellt, indem ich durch die interaktive Veränderung der Vokaltraktparameter versucht habe, Deckungsgleichheit zwischen den Konturlinien des Modells und den Konturlinien der Vorlagebilder herbeizuführen. Durch die Flexibilität des Vokaltraktmodells ist dies in den meisten Fällen gut gelungen. Eine gute *geometrische* Übereinstimmung zwischen dem Vokaltraktmodell und den Vorlagebildern hat jedoch nicht immer auch zu guten akustischen Resultaten geführt. Vielfach klangen die nach den Vorlagen eingestellten Vokale unnatürlich, meistens zu zentralisiert. Zu diesem Problem sind u. a. die folgenden Ursachen denkbar:

- Die geometrische Anpassung zwischen dem Modell und den Vorlagen war nur in der sagittalen Ebene möglich. Informationen über angehobene oder abgesenkte Zungenränder, die sich auf die Flächenfunktion und somit das akustische Resultat auswirken, waren in den Vorlagen nicht vorhanden.
- Die Vorlagen stammen von einem russischen Sprecher. Es muss angenommen werden, dass sich die phonetische Realisierung einzelner Vokale im Russischen und Deutschen unterscheidet.
- Die Konturbilder der Aufnahmen sind vermutlich mit kleinen Ungenauigkeiten behaftet, da die Konturen der Artikulatoren auf Röntgenbildern teilweise aufgrund unscharfer Übergänge und schlechter Kontraste schwer zu bestimmen sind.
- Die aerodynamisch-akustische Simulation in der vorliegenden Arbeit ist nicht exakt. So können z. B. verschobene Formantfrequenzen aufgrund eines zu hohen oder zu niedrigen Wertes für die Schallgeschwindigkeit bereits zu wahrnehmbaren Klangänderungen führen.

Da man den möglichen Fehlerquellen kaum im Einzelnen entgegenwirken kann, habe ich mich dafür entschieden, die Vokaltraktformen aus der geometrischen Anpassung in einem zweiten Schritt in Hinblick auf eine akustische Verbesserung abzuändern.

¹Der uvulare Vibrant [ʀ] wird hier der Einfachheit halber als stimmhaftes Gegenstück zum velaren Frikativ [x] behandelt.

Für die vokalischen Zielformen habe ich die Parameter durch „Versuch-und-Irrtum“ interaktiv so abgeändert, dass sich die resultierenden ersten drei Formantfrequenzen möglichst gut an vorgegebene Werte angenähert haben. Die vorgegebenen Formantfrequenzen habe ich durch eine LPC-Analyse aus selbst aufgenommenen Vokalen ermittelt. Um zu verhindern, dass durch das akustisch orientierte Vorgehen phonetisch unplausible Vokaltraktformen entstehen, habe ich darauf geachtet, dass die Parameter dabei nicht zu stark von den Werten aus der geometrischen Anpassung abweichen. Bei der Einstellung der Parameter für die Höhe der Zungenränder ($TS1 \dots TS4$) habe ich mich an den Ergebnissen der Arbeit von STONE *et al.* [111] orientiert. Folglich wurde z. B. der tiefe Vokale [a:] mit durchgängig angehobenen Zungenrändern erstellt und der hohe vordere Vokal [i:] mit einer konvexen Krümmung im Bereich vom Zungenblatt und einer konkaven Krümmung für den hinteren Teil der Zunge. Die Vokaltraktparameter für die Frikative wurden analog eingestellt. Hier wurde die Qualität der Einstellung nach jeder inkrementellen Änderung eines Parameters hauptsächlich perzeptiv beurteilt und nicht durch den Vergleich von Formantfrequenzen.

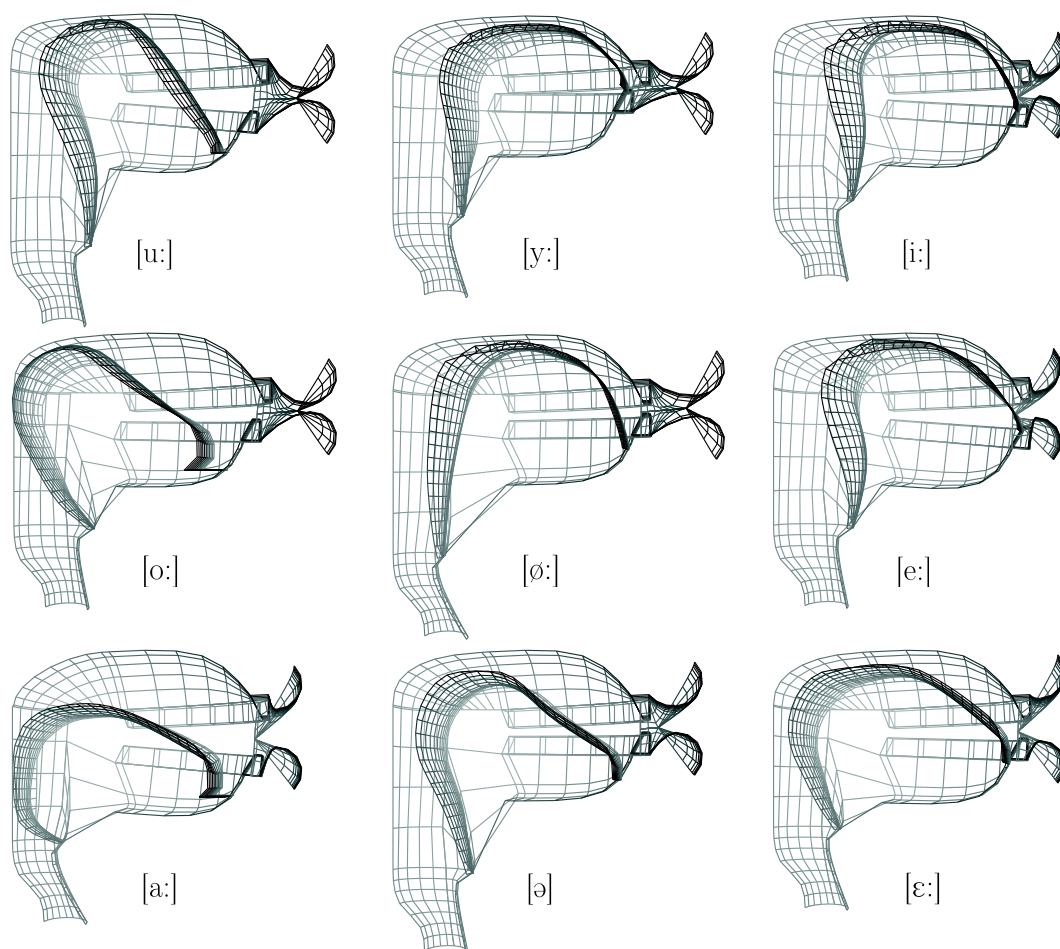


Abbildung 5.1: Zielformen für die langen Vokale und das [ə].

Abbildung 5.1 zeigt die eingestellten Vokaltraktformen für die langen Vokale und den neutralen Vokal, und Abb. 5.2 zeigt die Zielformen für die Konsonanten. Die Flächenfunktionen

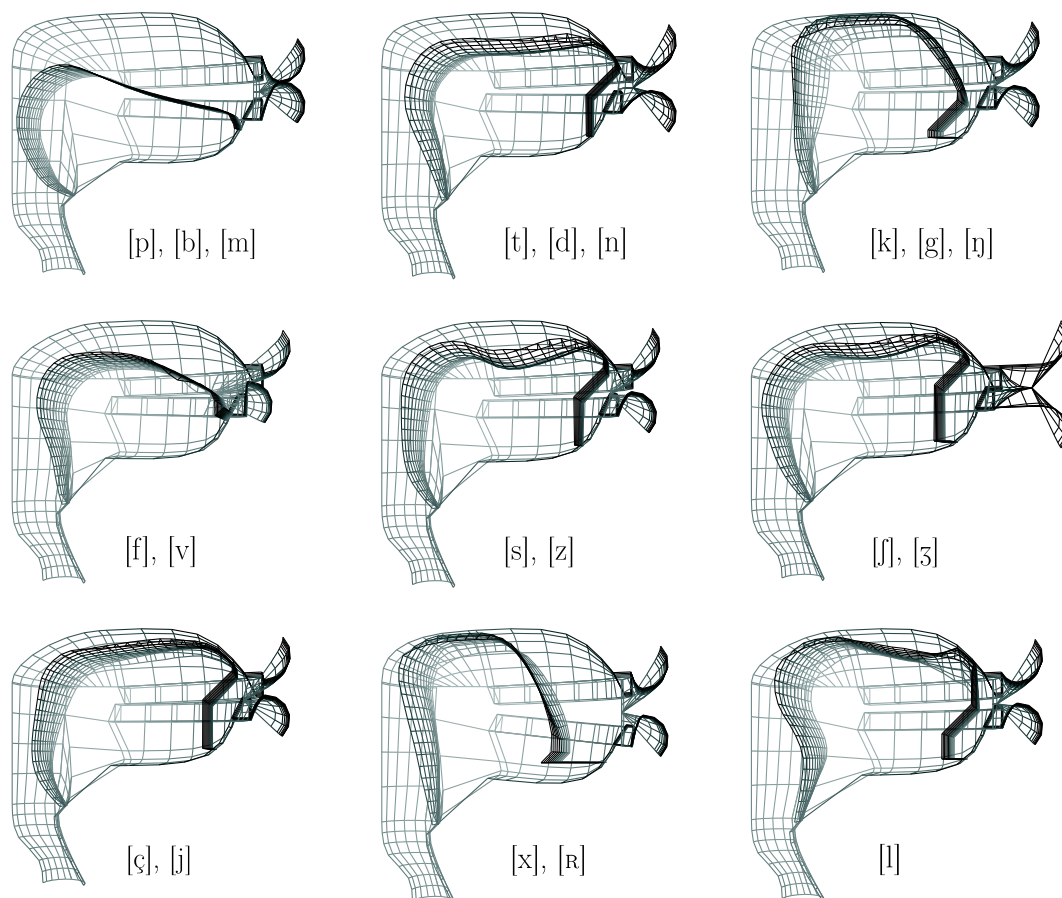


Abbildung 5.2: Zielformen für die Konsonanten.

der eingestellten Laute stimmen in ihren wesentlichen Eigenschaften mit den Flächenfunktionen überein, die in der Literatur veröffentlicht wurden. Jedoch ist damit nicht bewiesen, dass die dargestellten Zielformen die Laute in allen drei Dimensionen *physiologisch* richtig widerspiegeln. Deshalb wäre in weiterführenden Arbeiten die Anpassung der Vokaltraktformen an dreidimensionale Messdaten eines Sprechers (z. B. MRT-Voxeldaten) wünschenswert.

Für die konsonantischen Zielformen war neben der Einstellung der Vokaltraktparameter auch die Einstellung der Dominanzwerte erforderlich. Grundsätzlich ist der Dominanzwert eines Parameters ein Maß dafür, welche Bedeutung er für die Ausbildung der wesentlichen Formmerkmale des entsprechenden Konsonanten (hauptsächlich der konsonantischen Konstriktion) besitzt. Durch das Fehlen geeigneter Messdaten wurden auch hier wieder empirische Regeln angewendet. Durch die empirische Vorgehensweise macht es keinen Sinn, beliebige Dominanzwerte aus dem Intervall von 0 bis 1 bzw. 0% bis 100% zuzuweisen. Statt dessen habe ich die Dominanzen auf die diskreten Werte 0%, 25%, 50%, 75% und 100% beschränkt.

Eine Dominanz von 100% wurde jeweils den Parametern zugewiesen, die unverzichtbar für die Ausbildung der konsonantischen Konstriktion sind. Für die Zielform der Gruppe [k, g,

η] ist dies z. B. die Höhe des Zungenkörpers (*TCY*) und für die Zielform der Gruppe [t, d, n] die Position der Zungenspitze. Denjenigen Parametern, die die Vokaltraktform räumlich weit entfernt von der Konstriktion beeinflussen, wurde ein Dominanzwert von 0% zugewiesen (z. B. die Position des Zungenbeins für die Konsonanten [b, p, m]). Für die restlichen Parameter wurden die Dominanzwerte in Abhängigkeit von ihrem vermuteten Einfluss auf 25%, 50% oder 75% gesetzt. Teilweise wurden auch Informationen aus der Literatur herangezogen, um auf die Dominanzwerte zu schließen. So wird z. B. von MOOSHAMMER *et al.* [84] angegeben, dass sich die Position des Unterkiefers für die Laute [s, ʃ, t] nur sehr wenig ändert, wenn man den Vokalkontext variiert. Das bedeutet, dass die Unterkieferparameter für diese Laute kaum vom Lautkontext beeinflusst werden und eine hohe Dominanz besitzen.

5.1.2 Evaluierung des Dominanzmodells anhand von Röntgenbildern

Inwiefern das hier verwendete Dominanzmodell tatsächlich in der Lage ist, koartikulatorische Effekte realitätsnah zu simulieren, wurde durch den Vergleich von berechneten mit gemessenen Vokaltraktformen für die Plosive [g] und [d] in unterschiedlichen Vokalkontexten untersucht. Bei den Messdaten handelt es sich um Vokaltraktkonturen von Röntgenaufnahmen aus der Arbeit von ÖHMAN [90] für die Vokale [a], [i], [u] und die Konsonanten [g] (in den Sequenzen [igi] und [ugu]) und [d] (in den Sequenzen [ada] und [udu]). Aufgrund der Koartikulation unterscheidet sich sowohl die Realisierung des [g] im Kontext von [i] und [u] als auch die Realisierung des [d] im Kontext von [a] und [u]. Das Ziel dieser Untersuchung war es, Vokaltraktparameter für die Zielformen der Laute [a, i, u, g, d] und Dominanzwerte für die Parameter von [b] und [g] zu finden, so dass die berechneten Vokaltraktkonturen mit den Aufnahmen von ÖHMAN übereinstimmen.

Die Einstellung der Vokaltraktparameter für die Vokale wurde interaktiv durch eine Konturanpassung realisiert und hat keine Schwierigkeiten bereitet. Abb. 5.3 zeigt die Konturen der Vokale von ÖHMAN (links) und ihre Nachbildung mit Hilfe des Vokaltraktmodells (rechts). Die rechten Abbildungen sind zum besseren Vergleich mit den gemessenen Konturen grau hinterlegt. Bis auf den weichen Gaumen und den Larynx stimmen die gemessenen und berechneten Konturen weitestgehend überein. Für die beiden Konsonanten habe ich die Vokaltraktparameter und Dominanzwerte durch systematisches ausprobieren gefunden. Die Ergebnisse der Untersuchung sind in Abb. 5.4 dargestellt. Die Bilder auf der linken Seite zeigen die gemessenen Vokaltraktkonturen von ÖHMAN [90] für das [g] im [i]- und [u]-Kontext (oben) und das [d] im [a]- und [u]-Kontext (unten). Hier ist deutlich zu erkennen, dass der Verschluss für das [g] im [i]-Kontext aufgrund der Koartikulation weiter vorne gebildet wird als im [u]-Kontext. Für das [d] unterscheiden sich die Verschlusspositionen im [a]- und [u]-Kontext kaum, jedoch befindet sich der Zungenrücken im [u]-Kontext an einer höheren Position.

Die Bilder auf der rechten Seite in Abb. 5.4 zeigen die entsprechenden berechneten Vokaltraktkonturen. Sie wurden im stationären Intervall der konsonantischen Gesten berechnet, mit denen die Kontextvokale überlagert wurden. Grau hinterlegt sind hier zum Vergleich wieder die gemessenen Konturen. Sowohl für das [d] als auch für das [g] stimmen die gemessenen und berechneten Konturen, insbesondere für die Zunge, in den wesentlichen Merkmalen überein. Für das [g] im [u]-Kontext ist die Übereinstimmung nicht ganz so gut wie für die anderen Konturen, da der weiche Gaumen im gemessenen Vokaltrakt tiefer liegt

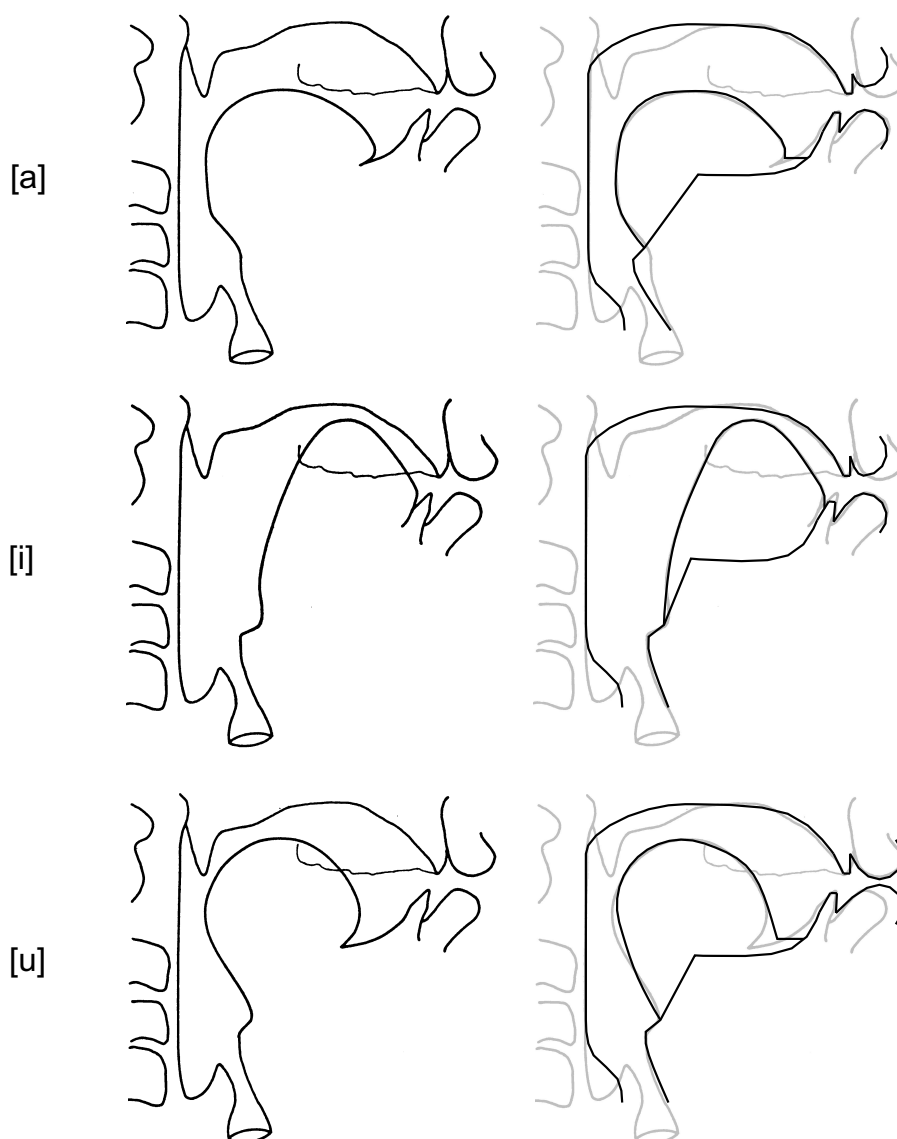


Abbildung 5.3: Vokaltraktformen für die Vokale [a], [i] und [u] nach ÖHMAN [90] (links) und ihre Nachbildung mit Hilfe des Vokaltraktsmodells dieser Arbeit (rechts).

als im Modell.

Tabelle 5.2 gibt die eingestellten Dominanzwerte der Zungenparameter für das [d] und das [g] wieder (für eine Erläuterung der Zungenparameter siehe Abschnitt 2.3). Für das [d] wurden die Dominanzwerte der Koordinaten TTX und TTY für die Zungenspitze auf 100% gesetzt, so dass die alveolare Konstriktion unabhängig vom Vokalkontext immer an der gleichen Stelle ausgebildet wird. Die restlichen Zungenparameter haben Dominanzwerte von 50% oder weniger, so dass sie zu mindestens 50% von den überlagerten Vokalen beeinflusst werden.

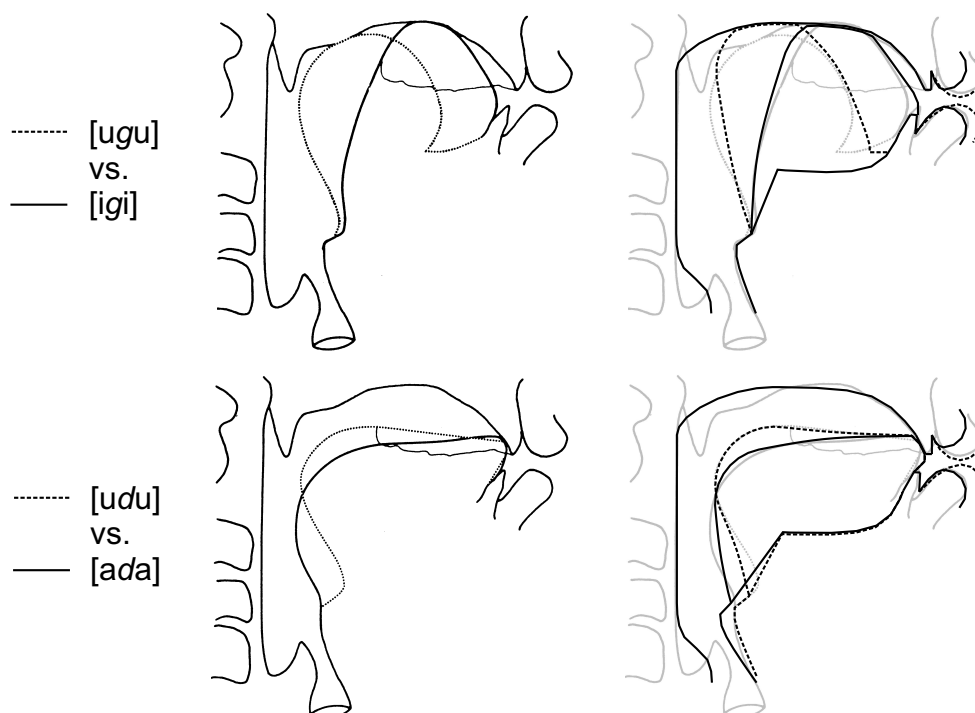


Abbildung 5.4: Koartikulatorische Beeinflussung der Zungenformen für den Konsonanten [g] in den Äußerungen [ugu] und [igi] (oben) und den Konsonanten [d] in den Äußerungen [udu] und [ada] (unten). Die linken Bilder zeigen die Kontur von Röntgenaufnahmen nach ÖHMAN [90] und die rechten Bilder ihre Simulation mit dem Steuermodell dieser Arbeit.

Laut	<i>TCX</i>	<i>TCY</i>	<i>TCR</i>	<i>TTX</i>	<i>TTY</i>	<i>TBX</i>	<i>TBY</i>	<i>TRE</i>
[d]	50%	25%	50%	100%	100%	50%	50%	50%
[g]	0%	100%	100%	25%	25%	25%	25%	25%

Tabelle 5.2: Dominanzwerte der Zungenparameter für die Konsonanten [d] und [g].

Für das [g] wurden die Parameter *TCY* und *TCR* für die *y*-Koordinate und den Radius des Zungenkörpers auf 100% gesetzt. Dadurch wird für diesen Konsonanten stets eine Konstriktion mit dem Zungenkörper (Zungenrücken) ausgebildet, indem dieser gegen den Gaumen gepresst wird. Die *horizontale* Position des Zungenkörpers (Parameter *TCX* mit 0% Dominanz) wird dagegen zu 100% vom überlagerten Vokal bestimmt und ist somit im [i]-Kontext weit vorne und im [u]-Kontext weit hinten.

Die Untersuchung in diesem Abschnitt hat ergeben, dass koartikulatorische Effekte mit dem gestischen Dominanzmodell in dieser Arbeit – zumindest für die untersuchten Fälle – realitätsnah nachgebildet werden können. Aufgrund des begrenzten Umfangs der von ÖHMAN [90] veröffentlichten Aufnahmen wurden die Parameter und Dominanzwerte der Zielformen aber letztendlich mit Hilfe der Aufnahmen von KONECZNA *et al.* [65] (siehe vorangegangener Abschnitt) eingestellt. Die hier gewonnenen Erkenntnisse über die Domi-

nanzwerte für die Laute [d] und [g] wurden auf die Zielformen der Konsonantengruppen [t, d, n] und [k, g, ŋ] übertragen.

5.2 Regeln für die Erstellung gesturaler Ablaufpläne

Die Erstellung von Äußerungen mit mehreren Lauten erfolgt in dieser Arbeit auf der Basis des gesturalen Ablaufplans. Der Ablaufplan muss manuell erstellt werden, indem die artikulatorischen Gesten zur Realisierung der Äußerung zeitlich koordiniert und ihre Parameter eingestellt werden. Die manuelle Erstellung ist nur dann mit einem vertretbaren Zeitaufwand möglich, wenn ihr gewisse Regeln, oder zumindest qualitative Anhaltspunkte, zugrunde liegen. Der Begriff „Regel“ ist in diesem Abschnitt also nicht als eine streng quantitative Aussage zu verstehen.

Innerhalb einer längeren Äußerung hängt die konkrete Realisierung eines einzelnen Lautes von sehr vielen Einflussfaktoren ab, z. B. der Position des Lautes im Wort, der Stellung und Funktion des Wortes im Satz, der Betonung der Silbe und der Sprechgeschwindigkeit. Die quantitative Untersuchung und Beschreibung all dieser Einflüsse auf die Gestenparameter geht weit über den Rahmen dieser Arbeit hinaus.

Die nachfolgend vorgeschlagenen Regeln zur Erstellung eines Ablaufplans betreffen im Wesentlichen die artikulatorische Dynamik auf der Ebene von Silben und einzelnen Lauten und Lautübergängen. Auf dieser Ebene wurden in zahlreichen Studien Untersuchungen anhand von Logatomen der Form VCV oder CVC durchgeführt (z. B. [89, 45, 92, 122]), wobei die V's für Vokale und die C's für Konsonanten stehen. Diese Untersuchungen geben für den jeweils untersuchten phonetischen Kontext u. a. Auskunft über die typische Dauer von Lauten und Lautübergängen sowie über die zeitabhängige Bewegungsgeschwindigkeit und Auslenkung einzelner Artikulatoren. Die für die vorliegende Arbeit wichtigen Ergebnisse dieser Untersuchungen werden in diesem Abschnitt mit Bezug auf die gestische Realisierung der Laute zusammengestellt.

Zu einigen Fragestellungen waren jedoch nur spärliche Informationen in der Literatur zu finden, so dass ich diese durch eigene akustische Messungen ergänzt habe oder indirekt durch Syntheseexperimente auf unbekannte Gestenparameter geschlossen habe. Hierbei muss betont werden, dass die ergänzenden akustischen Messungen, z. B. von der Länge einiger Laute, nur jeweils mit Sprachaufnahmen vom Autor in relativ begrenztem Umfang durchgeführt wurden. Die Messergebnisse sind daher nur als Anhaltspunkte zu verstehen, da sie nicht notwendigerweise repräsentativ für eine größere Gruppe von Sprechern sind. Auch die Bewertung der Sprachqualität bei den ergänzenden Syntheseexperimenten wurde bisher nur vom Autor der Arbeit im Eigenversuch durchgeführt. Die aus den Syntheseexperimenten gefolgerten Parameter für bestimmte artikulatorische Gesten sind daher auch nur als grobe Abschätzungen zu verstehen.

Neben der realistischen Bewegungssteuerung der supraglottalen Artikulatoren trägt auch die Nachbildung *mikroprosodischer Grundfrequenzvariationen* zur Natürlichkeit und Verständlichkeit synthetischer Sprache bei [27, 103]. Mikroprosodische Variationen der Grundfrequenz unterliegen nicht der aktiven Kontrolle eines Sprechers und sind von der segmentalen Struktur einer Äußerung abhängig. Sie sind der eigentlichen Grundfrequenzkontur überlagert, die aktiv vom Sprecher im Sinne der beabsichtigten Intonation erzeugt wird.

Man unterscheidet zwei Arten mikroprosodischer Grundfrequenzänderungen: die *intrinsic*-

sche F_0 -Variation IF_0 und die *koartikulatorische* F_0 -Variation CF_0 [82]. Die IF_0 drückt charakteristische Grundfrequenzunterschiede zwischen hohen und tiefen Vokalen aus. Die grundsätzliche Tendenz ist, dass hohe (geschlossene) Vokale mit einer höheren F_0 erzeugt werden als tiefe (offene) Vokale und dass der IF_0 -Unterschied zwischen hohen und tiefen Vokalen 1 bis 2 Halbtöne beträgt [27].

Die CF_0 charakterisiert den Einfluss von Frikativen und Plosiven auf die F_0 des nachfolgenden Vokals. Der Trend hier ist, dass stimmlose Obstruenten die F_0 am Anfang des Vokals um bis zu 3 Halbtöne anheben. Der Effekt ist unabhängig von der Artikulationsstelle und klingt im Laufe von 30-130 ms ab. Bei stimmhaften Obstruenten ist die Anhebung der F_0 im folgenden Vokal deutlich kleiner oder nicht vorhanden. Sonorante Konsonanten verursachen keine koartikulatorischen F_0 -Abweichungen [27].

Mikroprosodische F_0 -Variationen sind ein universelles sprachenübergreifendes Phänomen. Die Ausprägung der Effekte kann jedoch zwischen verschiedenen Sprechern sehr unterschiedlich sein und hängt außerdem vom prosodischen Kontext und der Silbenposition in einer Äußerung ab [27]. Trotz der Unterschiede in der Ausprägung gibt es klare Hinweise darauf, dass zumindest die CF_0 perzeptiv eine Rolle bei der Unterscheidung von stimmhaften und stimmlosen Plosiven spielt [103]. Im Rahmen des gesturalen Ablaufplans lassen sich mikroprosodische F_0 -Variationen am einfachsten mit zusätzlichen F_0 -Akzentgesten realisieren, die im Vergleich zu den intonatorischen Akzentgesten durch eine geringere Amplitude und steilere Flanken gekennzeichnet sind. In den nachfolgenden Beispielen für gesturale Ablaufpläne habe ich mich auf die Nachbildung koartikulatorischer F_0 -Variationen beschränkt, da ihre perzeptive Relevanz nachgewiesen ist.

Die hier zusammengestellten Informationen sind auch als Grundlage für die Erstellung eines Regelsystems gedacht, das die automatische Umwandlung einer (linguistischen) Textrepräsentation in einen gesturalen Ablaufplan ermöglicht. Mit solch einem Regelsystem könnte das Synthesemodell in dieser Arbeit zu einem vollständigen TTS-System erweitert werden.

5.2.1 Vokale

Vokale und Halbvokale werden im Ablaufplan als vokalische Gesten realisiert. Für einen nahtlosen Übergang zwischen benachbarten Vokalen sollten die Gestenkommandos so aneinandergereiht werden, dass eine neue Vokalgeste jeweils genau am Ende des vorangegangenen Gestenkommandos beginnt. Dadurch überlagern sich jeweils das Abklingintervall der zeitlich früheren und das Anklingintervall der zeitlich späteren Geste.

Laut STEVENS [110] dauert die Bewegung der Zunge zwischen zwei Vokalkonfigurationen rund 100 ms. Daher können T_{an} und T_{ab} in erster Näherung jeweils auf 100 ms gesetzt werden. Die Länge eines vokalischen Gestenkommandos und damit die Lautdauer ist stark kontextabhängig und sollte zwischen 100-200 ms liegen. Mit der Amplitude der vokalischen Gestenkommandos lässt sich einstellen, inwieweit die ideale Zielposition für einen Vokal erreicht wird. Für $A = 1$ wird die Zielposition vollständig erreicht, während kleinere Werte eine Zentralisierung des Vokals verursachen, d. h. eine Verschiebung in Richtung des neutralen Vokals.

Bei der Überlagerung von vokalischen und konsonantischen Gesten in VCV-Sequenzen stellt sich die Frage nach der zeitlichen Koordination der Gesten. GAY [45] hat gezeigt, dass die Bewegung der Artikulatoren vom ersten zum zweiten Vokal, die *nicht* an der Bildung der konsonantischen Konstriktion beteiligt sind, erst beginnt, sobald die Konstriktion ausgebil-

det wurde. Zum Beispiel beginnt bei der Äußerung [ita] die Bewegung des ungebundenen Zungenkörpers vom [i] zum [a] erst dann, wenn der Verschluss für den Plosiv [t] mit der Zungenspitze vollständig hergestellt wurde. Daraus lässt sich schließen, dass die Position einer überlagerten konsonantischen Geste so gewählt werden sollte, dass der Beginn des zweiten Vokals etwa mit dem Zeitpunkt übereinstimmt, an dem die vollständige Ausbildung der konsonantischen Konstriktion erreicht wird.

Es muss beachtet werden, dass die vollständige Ausbildung einer konsonantischen Konstriktion in der Regel nicht genau mit dem Ende des konsonantischen Anklingintervalls zusammenfällt, sondern oft bereits zu einem früheren Zeitpunkt erreicht wird, z. B. bei einem Realisierungsgrad der Geste von 0.8 oder früher. Das hängt mit den Zielpositionen der konstriktionsbildenden Artikulatoren für die Konsonanten zusammen. Der Kreis für die Zungenspitze ragt an seiner Zielposition für den Laut [t] z. B. teilweise nach oben über die Konturlinie des harten Gaumens hinaus (siehe Kapitel 2), und die endgültige Form der Zunge ergibt sich erst durch das „Abschneiden“ der Zungenkontur an der Gaumenkontur. Im Zuge der Bewegung der Zungenspitze aus einer Vokalposition zur Zielposition für den Konsonanten wird der Verschluss schon gebildet, sobald der Kreis für die Zungenspitze den Gaumen *berührt*. Der Realisierungsgrad der konsonantischen Geste ist hier aber noch kleiner als 1, da die Zielposition der Zungenspitze noch nicht vollständig erreicht ist. Aus dem gleichen Grund erfolgt auch die Verschlusslösung erst kurz *nach* dem Beginn der konsonantischen Abklingphase. Insgesamt ist dadurch die Verschlussdauer (bei Plosiven) oder die Lautdauer (bei Frikativen) jeweils etwas länger als die stationäre Phase der konsonantischen Geste.

5.2.2 Sonoranten

Zu den Sonoranten zählen im Hochdeutschen u. a. die Nasale [m], [n] und [ŋ] und der Lateral [l]. Sie sind dadurch ausgezeichnet, dass bei ihrer Erzeugung kein Druck hinter der konsonantischen Engstelle im Vokaltrakt aufgebaut wird und die Stimmlippen dadurch ungehindert (weiter-)schwingen können.

Nasale. Die Nasale werden im gesturalen Ablaufplan mit einer konsonantischen Geste für den oralen Verschluss und einer velischen Geste zur Absenkung des Velums realisiert. Die typische Dauer der Verschlussphase eines Nasals liegt nach eigenen Messungen in Abhängigkeit von der Sprechgeschwindigkeit, dem Kontext und der Artikulationsstelle zwischen 70 und 130 ms. Die genaue Abhängigkeit der Lautdauer von diesen Faktoren muss noch untersucht werden. Ein auffälliger Trend bei meinen Analysen war aber, dass die Lautdauer von [n] unter sonst gleichen Bedingungen kürzer war als die von [m] oder [ŋ].

Zur Dauer der artikulatorischen Transitionen zwischen einem Nasal und den angrenzenden Vokalen gibt es in der Literatur kaum Informationen. Laut STEVENS [110, S. 45] ist die Öffnungsgeschwindigkeit der Lippen nach dem bilabialen Verschluss für ein [m] etwa 30 cm/s und für ein [p] etwa 75 cm/s. Der Rückschluss auf die Transitionsdauer zwischen einem [m] und dem folgenden Vokal ist damit nur bedingt möglich. STEVENS gibt außerdem an, dass die Bewegung des Zungenkörpers für einen velaren Konsonanten langsamer ist als die Bewegung der Zungenspitze oder der Lippen für einen alveolaren oder labialen Konsonanten, und dass die Geschwindigkeit im gewissen Maße vom Vokalkontext abhängt.

Um Referenzwerte für die Transitionsdauer von und zu einem nasalen Konsonanten zu erhalten, habe ich deshalb synthetische Vokal-Nasal-Vokal-Stimuli mit unterschiedlichen Werten für die An- und Abklingdauer der konsonantischen Geste untersucht. Eine gute perzeptive Unterscheidung der Nasale war unabhängig vom Vokalkontext möglich, wenn T_{an} und T_{ab} für [n] jeweils auf 60 ms, für [m] auf 80 ms und für [ŋ] auf 100 ms gesetzt wurden.

Untersuchungen zur Bewegung des Velums und ihrer Koordination mit der Zungen- und Lippenbewegung wurden u. a. von MOLL *et al.* [83] und KOLLIA *et al.* [64] durchgeführt. Die Ergebnisse lassen sich knapp wie folgt zusammenfassen. Die *Öffnungsbewegung* des Velums im CVN- und CVVN-Kontext beginnt oft bereits zum Zeitpunkt der oralen Verschlusslösung des (nicht-nasalen) Konsonanten C oder kurze Zeit später und geht dadurch der oralen Verschlussbildung für das N voraus. Dadurch ist die velopharyngale Pforte während der V's bereits immer schon zum Teil oder vollständig geöffnet. Die *Schließbewegung* des Velums im NC- oder NVC-Kontext beginnt oft schon zum Zeitpunkt der oralen Verschlussbildung für den Nasal und spätestens zum Zeitpunkt der oralen Verschlusslösung, aber auf alle Fälle während des Nasals. Bezüglich der Dauer der velischen Bewegung zwischen einem abgesenkten und einem angehobenen Zustand wird von STEVENS [110] angegeben, dass die *minimale* Dauer für einen kompletten Zyklus von einer geschlossenen zu einer offenen und zurück zu einer geschlossenen velopharyngalen Pforte mindestens 200-300 ms beträgt. Die typische Dauer für eine unidirektionale Bewegung des Velums scheint somit 150 ms zu sein (T_{an} und T_{ab} der velischen Geste), was anhand der Messkurven von MOLL *et al.* [83] und KOLLIA *et al.* [64] bestätigt wird.

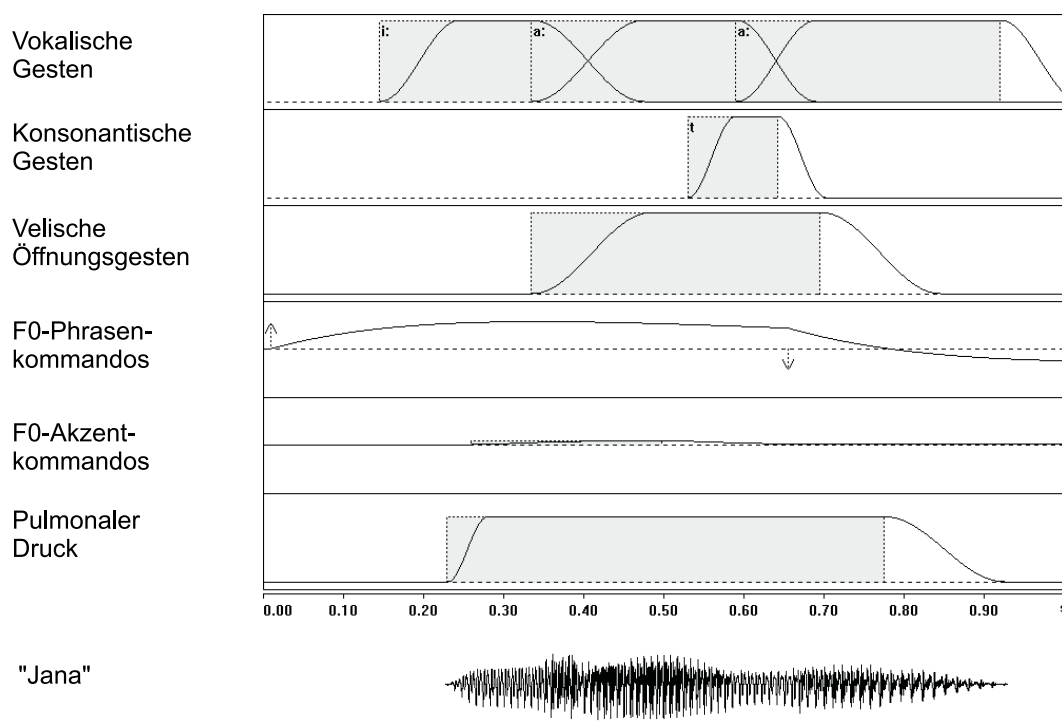


Abbildung 5.5: Gesturaler Ablaufplan für das Wort „Jana“.

Der gesturale Ablaufplan in Abb. 5.5 zeigt beispielhaft die zeitliche Koordination der Gesten für das Wort „Jana“. Deutlich erkennbar ist dabei die Vorwegnahme der velischen Öffnungsbewegung gegenüber der konsonantischen Geste für den alveolaren Verschluss ([t]-Geste).

Lateral. Der Lateral [l] lässt sich im Ablaufplan durch die konsonantische [l]-Geste realisieren. Die Lautdauer hängt vom Kontext und der Sprechgeschwindigkeit ab, und eine An- und Abklingdauer von jeweils 100 ms für die Geste liefert perzeptiv gute Ergebnisse.

5.2.3 Obstruenten

Zu den Obstruenten zählen die Frikative und die Plosive. Ihre Artikulation ist dadurch gekennzeichnet, dass hinter der konsonantischen Konstriktion im Vokaltrakt ein Überdruck entsteht. Mit Ausnahme der stimmhaften Plosive wird der Aufbau des Überdrucks durch die aktive Öffnung der Stimmritze unterstützt, die im gesturalen Ablaufplan mit Hilfe der glottalen Abduktionsgeste realisiert wird. Die maximale Amplitude des glottalen Gestenkommandos $A = 1$ entspricht einer glottalen Öffnungsfläche von etwa 0.3 cm^2 . Dieser Wert ist eine typische Spitzenöffnungsfläche für stimmlose Frikative und Plosive [110]. Die Zeit für die Adduktion der Glottis von der Spitzenöffnungsfläche der Obstruenten in die Phonationsstellung beträgt ungefähr 100 ms [56, 50]. Laut HUTTERS [50], HIROSE [47] und STEVENS [110] sind die Öffnungs- und Schließbewegungen der Stimmlippen zumindest für die stimmlosen Obstruenten annähernd symmetrisch. Daraus folgt sowohl für die Ankling- als auch für die Abklingphase der glottalen Gesten eine typische Dauer von 100 ms.

Stimmlose aspirierte Plosive. Das Sprachsignal eines stimmlosen aspirierten Plosivs (im Deutschen $[p^h]$, $[t^h]$ und $[k^h]$) kann in eine Verschlussphase, einen kurzen Burst (Verschlusslösungsgeräusch) und eine Frikations- und Aspirationsphase unterteilt werden, die zeitlich aufeinander folgen. Die Zeit für den Burst und die Frikations- und Aspirationsphase wird als *voice onset time* (VOT) bezeichnet, d. h. als die Zeit, die zwischen der Verschlusslösung des Plosivs und dem Beginn des anschließenden stimmhaften Lautes liegt.

Nach eigenen akustischen Messungen der Verschlussdauer von Plosiven in VCV-Logatomen variiert die Verschlusszeit in Abhängigkeit von der Sprechgeschwindigkeit, dem Vokalkontext und der Artikulationsstelle etwa zwischen 50 und 120 ms. Dabei weisen Plosive im Kontext tiefer Vokale eine kürzere Verschlussphase auf als im Kontext hoher Vokale (Unterschied zwischen [a]- und [u]-Kontext von etwa 20 ms), und für die Plosive [p] und [k] scheint die Verschlussdauer etwas größer zu sein als für [t].

Bezüglich der Dauer der artikulatorischen Transition von einem vorangegangenen Vokal zum Verschluss für einen Plosiv gibt es in der Literatur Hinweise darauf, dass die Verschlussbildung für stimmlose Plosive schneller erfolgt als für stimmhafte [17, 62, 122]. Laut CHEN [17] ist dieser Umstand auch zum Teil dafür verantwortlich, dass Vokale vor stimmhaften Plosiven sprachenübergreifend eine längere Dauer aufweisen als Vokale vor stimmlosen Plosiven. In der Untersuchung von VAN SUMMERS [122] dauert z. B. die Transition vom [a] zum [b] in [bab] signifikant länger als die Transition vom [a] zum [p] in [bap] (Faktor von ca. 1.4). KOHLER [62] hat ein dazu reziprokes Verhalten nach der Verschlusslösung von Plosiven festgestellt. Daraufhin formulierte er die Hypothese, dass für Plosive vom

Gehirn entweder eine schnelle Verschlussbildung und eine langsame Verschlusslösung „programmiert“ wird (stimmlose Plosive) oder aber eine langsame Verschlussbildung und eine schnelle Verschlusslösung (stimmhafte Plosive). In Syntheseexperimenten habe ich festgestellt, dass die konkreten Zeiten für das Anklingen und Abfallen der gestischen Realisierungsfunktionen auch von der artikulatorischen Distanz zwischen den Ziellauten abhängen. So ist beispielsweise die Distanz des Zungenrückens zwischen [u] und [k] deutlich geringer als zwischen [a] und [k]. Die An- und Abklingphasen in [aka] müssen deshalb länger dauern als in [uku], um eine hohe Verständlichkeit bei der Synthese der Logatome zu gewährleisten. Tabelle 5.3 enthält die Werte für T_{an} und T_{ab} für die Plosive [p, t, k] im Kontext der Vokale [a, i, u], mit denen ich eine subjektiv hohe Verständlichkeit erzielt habe.

VpV	T_{an}	T_{ab}	VtV	T_{an}	T_{ab}	VkV	T_{an}	T_{ab}
[apa]	70	100	[ata]	70	100	[aka]	70	100
[ipi]	50	70	[iti]	50	70	[iki]	50	70
[upu]	35	50	[utu]	70	100	[uku]	35	50

Tabelle 5.3: Dauer der Ankling- und Abklingintervalle für stimmlose Plosive im Kontext verschiedener Vokale (Zeitangaben in ms).

Im Verhältnis zur konsonantischen Geste beginnt die glottale Öffnungsbewegung für stimmlose Plosive zum Zeitpunkt der oralen Verschlussbildung oder kurz danach [47, 41]. Der Beginn der glottalen Schließbewegung hängt dagegen mit der gewünschten Dauer der Aspirationsphase bzw. der VOT zusammen. Laut dem vielzitierten Artikel von KIM [56] gibt es eine direkte Korrelation zwischen der glottalen Öffnungsfläche zum Zeitpunkt der konsonantischen Verschlusslösung und der Aspirationsdauer bzw. VOT. Für stark aspirierte Plosive fällt der Zeitpunkt der maximalen glottalen Öffnung ungefähr mit der konsonantischen Verschlusslösung zusammen. Übertragen auf den gesturalen Ablaufplan bedeutet dies, dass die Abklingphase der glottalen Geste etwa zum Zeitpunkt der konsonantischen Verschlusslösung beginnt. Eine kürzere Aspirationsphase erhält man, wenn die Abklingphase der glottalen Geste früher beginnt, da dann zum Zeitpunkt der oralen Verschlusslösung die glottale Öffnung bereits reduziert ist.

Meine eigenen Messungen der VOT stimmloser Plosive im Kontext der Vokale [a, i, u] haben gezeigt, dass sie nach den hohen Vokalen deutlich länger ist als nach tiefen Vokalen. In Abhängigkeit von der Sprechgeschwindigkeit wurden für die VOT im [a]-Kontext 40-80 ms gemessen und im [i]- und [u]-Kontext 60-120 ms. Die Artikulationsstelle des Konsonanten hatte keinen offensichtlichen Einfluss auf die VOT.

Bezüglich der koartikulatorischen F_0 sind stimmlose Plosive dadurch gekennzeichnet, dass sie die F_0 im folgenden Vokal um bis zu 3 Halbtöne anheben. Der Effekt klingt in ca. 30-130 ms ab und kann durch ein F_0 -Akzentkommando modelliert werden.

In Abb. 5.6 oben ist beispielhaft ein gesuraler Ablaufplan für die Äußerung „apa“ dargestellt. Man erkennt hier die unterschiedliche Steilheit der Flanken der [p]-Geste und die zeitliche Verschiebung zwischen der glottalen Abduktionsgeste und der konsonantischen Geste.

Stimmlose unaspirierte Plosive. Stimmlose unaspirierte Plosive unterscheiden sich von den stimmlosen aspirierten Plosiven hauptsächlich durch die Parameter der glot-

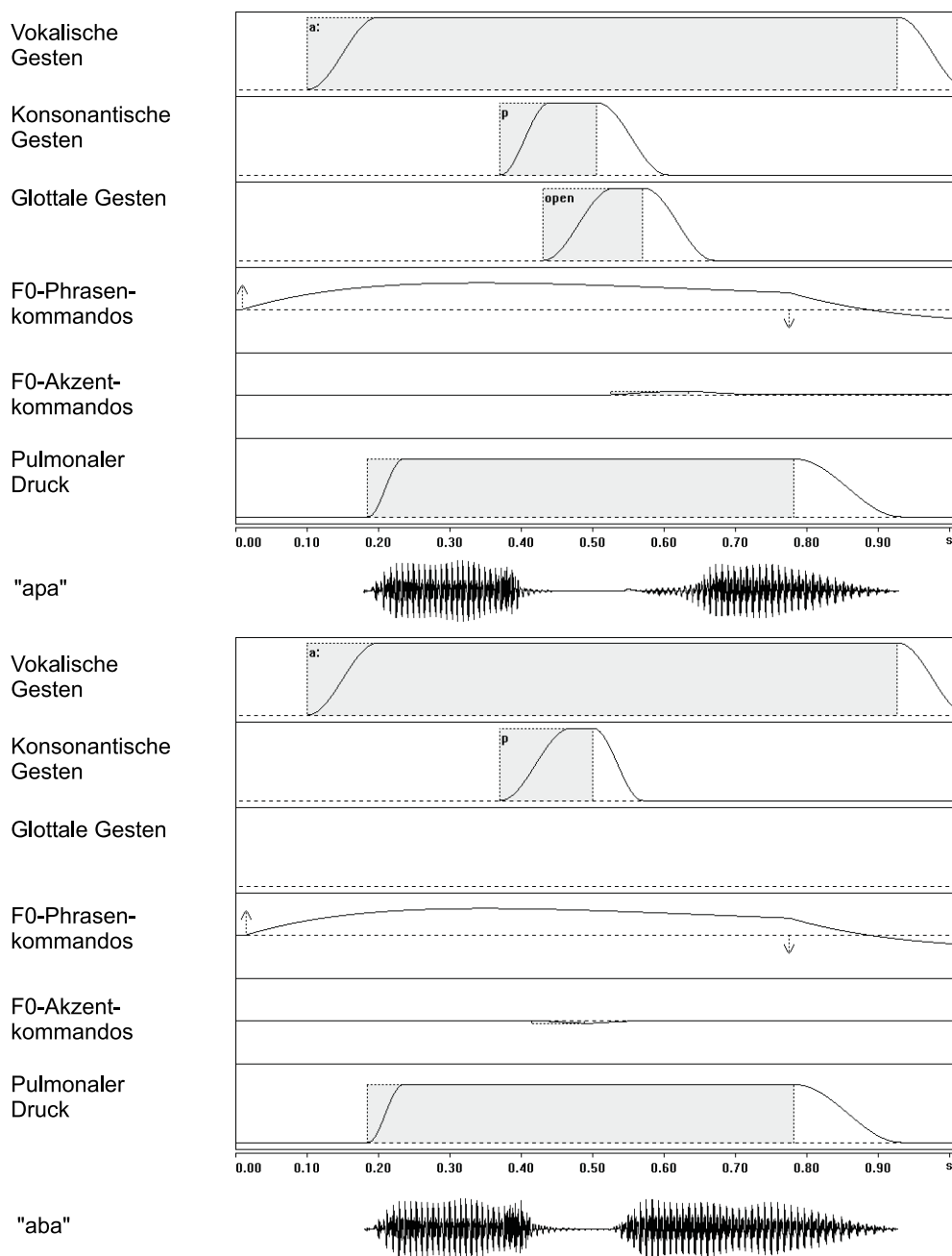


Abbildung 5.6: Gesturale Ablaufpläne für die Äußerungen „apa“ und „aba“.

talen Geste. Laut HUTTERS [50] erreicht die maximale glottale Öffnung für unaspirierte Plosive nur etwa 25% der Öffnung für die aspirierten Plosive, die Dauer der glottalen Geste ist nur halb so lang und das Maximum der glottalen Öffnungsfläche wird bereits etwa 50 ms vor der oralen Verschlusslösung erreicht. Dadurch ist die Glottis bei der konsonantischen Verschlusslösung schon fast wieder in Phonationsstellung, und die VOT beträgt nur etwa 30 ms. Die konsonantische Verschlussphase dauert (im Dänischen) für

unaspirierte Plosive etwa 25 ms länger als für aspirierte Plosive.

Stimmhafte Plosive. Stimmhafte Plosive sind dadurch gekennzeichnet, dass der Stimmton während der Verschlussphase möglichst lange aufrechterhalten wird. Dabei verbleibt die Glottis in Phonationsstellung. Im gesturalen Ablaufplan wird daher neben der konsonantischen Geste keine glottale Abduktionsgeste benötigt. Laut KOHLER [62] erfolgt die Verschlussbildung bei stimmhaften Plosiven relativ langsam und die Verschlusslösung relativ schnell. Die Dauer der Ankling- und Abklingphasen der Gesten für stimmhafte Plosive verhalten sich somit reziprok zu denen der stimmlosen Plosive. Als quantitative Anhaltspunkte können die Werte in Tab. 5.3 verwendet werden, wobei T_{an} und T_{ab} jeweils vertauscht werden müssen. Die Verschlussphase stimmhafter Plosive ist nach eigenen Messung unter sonst gleichen Bedingungen (Kontext, Sprechgeschwindigkeit, Artikulationsstelle) jeweils etwas kürzer als die Verschlussphase stimmloser Plosive.

Anders als die stimmlosen Plosive beeinflussen die stimmhaften Plosive die F_0 des folgenden Sonoranten nicht oder kaum [103]. Statt dessen weisen sie einen zunehmenden Abfall der F_0 (des über die Haut abgestrahlten Stimmtons) während der Verschlussphase auf. Dieser Abfall ist vermutlich auf die fortwährende Abnahme der transglottalen Druckdifferenz während der Verschlussphase zurückzuführen, die laut TITZE [118] zu einer Abnahme der Grundfrequenz führt. Nach meinen eigenen Messungen ist die F_0 am Ende gegenüber dem Anfang der Verschlussphase in der Regel um 10-20 Hz abgefallen. Im gesturalen Ablaufplan kann dieses Verhalten mit einem F_0 -Akzentkommando mit negativer Amplitude nachgebildet werden.

Abbildung 5.6 (unten) zeigt den Ablaufplan für die Äußerung „aba“. Die konsonantische Geste (hier auch mit „p“ bezeichnet, da für [b] und [p] die gleiche Vokaltraktform angesetzt wird) ist dabei gegenüber der entsprechenden Geste im Ablaufplan für „apa“ durch einen langsameren Anstieg, einen schnelleren Abfall und eine kürzere Dauer gekennzeichnet. Außerdem fehlt die glottale Geste, und das F_0 -Akzentkommando für den mikroprosodischen Einfluss unterscheidet sich.

Stimmlose Frikative. Von stimmlosen Frikativen liegt die Lautdauer nach eigenen Messungen anhand von VCV-Logatomen zwischen 100-240 ms und hängt vom Vokalkontext, der Sprechgeschwindigkeit und der Artikulationsstelle ab. Die Messungen haben ergeben, dass die Lautdauer im Kontext von hohen Vokalen (z. B. in „isi“) um bis zu 30% größer ist als im Kontext tiefer Vokale (z. B. in „asa“). Demgegenüber scheint der Einfluss der Artikulationsstelle auf die Lautdauer eher gering zu sein (maximale Unterschiede von 20...30 ms). Bezüglich der An- und Abklingdauer der konsonantischen Geste haben Syntheseexperimente ergeben, dass man gute Ergebnisse erzielt, wenn T_{an} und T_{ab} unabhängig vom Vokalkontext für [f] und [v] auf 40 ms, für [s], [z], [ʃ], [ʒ], [ç] und [j] auf 60 ms und für [x] und [ʁ] auf 80 ms gesetzt werden.

Die glottale Abduktion beginnt für Frikative gegenüber den Plosiven bereits *vor* der vollständigen Ausbildung der oralen Konstriktion [41]. Entsprechend der Abb. (8.2) in [110] scheint die glottale Öffnungs- und Schließbewegung außerdem symmetrisch zu sein und phasensynchron zur supraglottalen Konstriktionsgeste abzulaufen. Für die zeitliche Koordination der konsonantischen und der glottalen Geste empfiehlt sich daher, dass sowohl die Startzeitpunkte der Anklingphasen als auch die Endzeitpunkte der Abklingphasen in Über-

einstimmung gebracht werden. Diese Regel hat sich in Syntheseexperimenten bestätigt und sehr verständliche Ergebnisse geliefert.

Stimmlose Frikative sind ähnlich den stimmlosen Plosiven dadurch gekennzeichnet, dass sie die F_0 am Anfang des folgenden Vokals anheben [27]. Dieser koartikulatorische Effekt kann wiederum mit Hilfe eines F_0 -Akzentkommandos mit geringer positiver Amplitude am Ende des Frikativs nachgebildet werden.

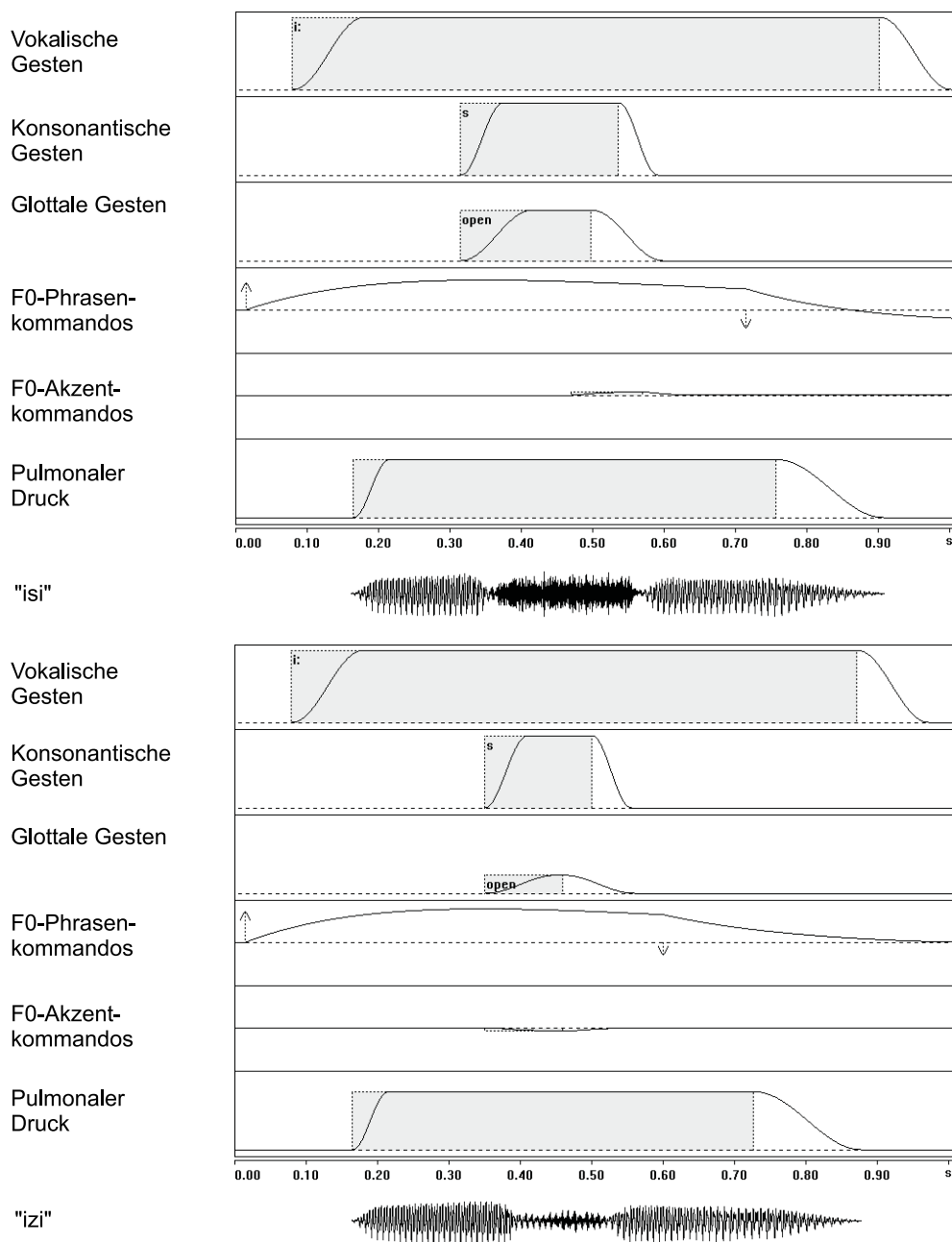


Abbildung 5.7: Gesturale Ablaufpläne für die Äußerungen „isi“ und „izi“.

Abb. 5.7 (oben) zeigt als Beispiel den gesturalen Ablaufplan für die Äußerung „isi“. Als typische Merkmale für einen stimmlosen Frikativ im Vokalkontext erkennt man die vollständige zeitliche Überlappung der Realisierungsfunktionen der konsonantischen und glottalen Geste sowie das Akzentkommando für die mikroprosodische F_0 -Variation.

Stimmhafte Frikative. Stimmhafte Frikative haben unter sonst gleichen Bedingungen eine deutlich kürzere Lautdauer als die stimmlosen Frikative. Nach eigenen Messungen ist ihre Länge oft nur 50-70% der Länge stimmloser Frikative. Die Ankling- und Abklingdauer der konsonantischen Gesten kann für die stimmhaften Frikative auf die gleichen Werte wie für die stimmlosen Korrelate gesetzt werden, und auch die zeitliche Koordinierung der konsonantischen und glottalen Gesten kann analog erfolgen. Damit einerseits ein genügend hoher Luftstrom für die Erzeugung von Friktionsrauschen entsteht und andererseits der Stimmtton erhalten bleibt, muss die mittlere glottale Öffnung größer als für die modale Phonation aber kleiner als für einen stimmlosen Frikativ sein. In Syntheseversuchen hat eine Amplitude von 0.2-0.3 für das glottale Gestenkommando für alle Frikative ein natürliches Verhältnis von Rauschanteil zu harmonischem Anteil ergeben. Nach eigenen akustischen Analysen von VCV-Logatomen fällt die F_0 in stimmhaften Frikativen in der Regel auf Werte ab, die deutlich unter denen der angrenzenden Vokale liegen. Dieser Effekt ist z. B. auch in Abb. (4.8) in der Dissertation von MIXDORFF [80] erkennbar. Im gesturalen Ablaufplan kann der Abfall durch ein F_0 -Akzentkommando mit negativer Amplitude simuliert werden, bei dem das Minimum der gestischen Realisierungsfunktion zeitlich etwa in der Mitte des Frikativs liegt.

Abb. 5.7 (unten) zeigt als Beispiel den gesturalen Ablaufplan für die Äußerung „izi“ im Vergleich zur Äußerung „isi“ darüber. Im unteren Ablaufplan ist die Dauer der konsonantischen und glottalen Gesten kürzer und die Amplitude der glottalen Geste kleiner als im oberen Ablaufplan. Die An- und Abklingdauer der jeweiligen Gesten stimmt dagegen für „isi“ und „izi“ überein.

5.2.4 Synthese längerer Äußerungen

Neben kurzen Wörtern und Logatomen, von denen einige als Ablaufplan auf den vorangegangenen Seiten dargestellt wurden, habe ich im Rahmen dieser Arbeit auch die Ablaufpläne für einige vollständige Sätze erstellt. Unter anderem wurden die Äußerungen „Der Zug hat eine Stunde Verspätung.“, „Nächster Halt: Hamburg.“ und „Guten Tag liebe Zuhörer!“ realisiert.

Abbildung 5.8 zeigt beispielhaft den Ablaufplan für die zweite der Äußerungen. In den vorangegangenen Unterabschnitten wurde bereits angesprochen, dass die Einstellung der Dauer für die einzelnen Laute aufgrund komplexer kontextueller und prosodischer Abhängigkeiten ein schwieriges Problem ist. Deshalb habe ich die Äußerungen selbst vorgesprochen und aufgenommen und die Länge der vokalischen und konsonantischen Gesten so eingestellt, dass die resultierenden synthetischen Lautlängen mit denen der natürlichen Äußerung übereinstimmten. Die Einstellung aller anderen Gestenparametern erfolgte in Anlehnung an die hier aufgestellten Regeln.

Die Intonationskontur wurde in den meisten Äußerungen mit Hilfe von jeweils ein oder zwei

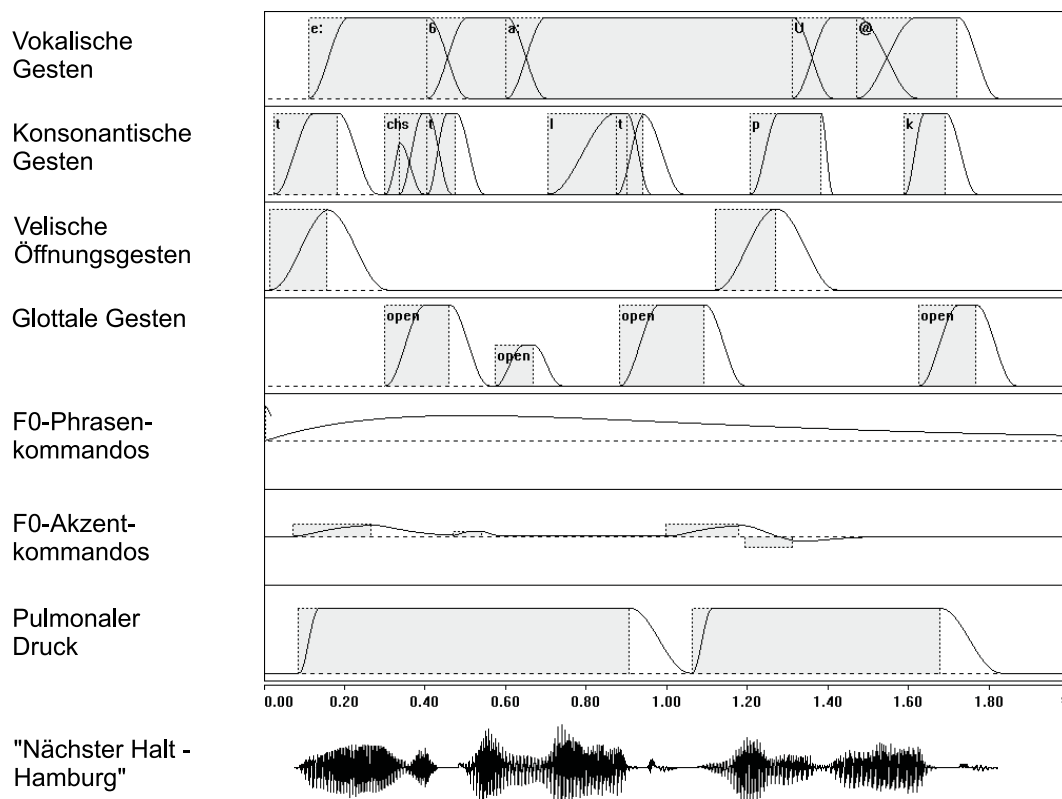


Abbildung 5.8: Gesturaler Ablaufplan für den Satz: „Nächster Halt: Hamburg“.

F_0 -Phrasen- und Akzentkommandos modelliert. Regeln für die Erstellung von Intonationskonturen für die deutsche Sprache auf der Basis des hier verwendeten Überlagerungsmodells wurden in den Arbeiten von MÖBIUS [81] und MIXDORFF [80] aufgestellt.

5.3 Zusammenfassung und Diskussion

In diesem Kapitel wurde zuerst die Erstellung einer Lauttabelle mit Zielformen für den Vokaltrakt und Dominanzwerten für die Vokaltraktparameter erläutert. Anschließend wurden Anhaltspunkte bzw. Regeln für die manuelle Erstellung eines gesturalen Ablaufplans aufgestellt.

Um eine überschaubare Anzahl von Zielformen zu erhalten, wurden teilweise mehrere Konsonanten mit gleichen supraglottalen Formmerkmalen gruppiert und durch eine gemeinsame Vokaltraktform repräsentiert. Die Vokale wurden jeweils einzeln modelliert. Als Vorlagen für die Erstellung der Vokaltraktformen dienten sagittale Röntgenaufnahmen des Sprechtrakts. Nach der geometrischen Anpassung der Vokaltraktformen wurden sie teilweise noch durch eine manuelle Feineinstellung der Parameter in Hinblick auf die akustische Qualität modifiziert. Die Einstellung der Dominanzwerte für die Vokaltraktparameter der konsonantischen Zielformen wurde hauptsächlich von empirischen Regeln geleitet. Die Nachbildung gemessener geometrischer Koartikulationseffekte hat gezeigt, dass das in die-

ser Arbeit verwendete Dominanzmodell seiner Aufgabe gerecht wird. Für zukünftige Arbeiten wäre es wünschenswert, die Einstellung der Vokaltraktparameter und Dominanzwerte anhand einer breiteren Datenbasis (z. B. 3D-MRT-Voxeldaten von Vokalen und Konsonanten in unterschiedlichen Kontexten) vorzunehmen und zu evaluieren.

Die Regeln für die Erstellung gesturaler Ablaufpläne wurden hauptsächlich aus früheren Studien abgeleitet und durch eigene akustische Messungen und Syntheseexperimente ergänzt. Die Regeln wurden als Ausgangspunkt für die Erstellung der Ablaufpläne mehrerer Wörter und Sätze genutzt. Die Äußerungen waren nach meiner Einschätzung durchweg sehr verständlich und klangen vergleichsweise natürlich. Sie sind ein klares Indiz für das gelungene Zusammenspiel der implementierten Teilmodelle auf den Ebenen der Steuerung, der geometrischen Sprechtraktmodellierung und der aerodynamisch-akustischen Simulation sowie für die Stärken der Teilmodelle im Einzelnen. Die Synthesebeispiele aus diesem und dem vorangegangenen Kapitel sind auf der beiliegenden CD als Audiodateien abgespeichert. Zusätzlich sind auf der CD auch einige Videoclips mit Animationen des Vokaltraktmodells zu finden.

Kapitel 6

Zusammenfassung und Diskussion

6.1 Zusammenfassung der Beiträge

In dieser Arbeit wurde ein System für die artikulatorische Sprachsynthese vorgestellt. Das System integriert ein geometrisches 3D-Modell des Vokaltrakts, ein dynamisches Modell der Stimmlippen, ein Steuermodell für die Generierung der artikulatorischen Bewegungsabläufe und ein aerodynamisch-akustisches Simulationsverfahren auf der Basis eines verzweigten Rohrmodells des Sprechapparats.

Das Vokaltraktmodell ist aus mehreren dreidimensionalen Gitternetzen zusammengesetzt, die die Oberflächen der Artikulatoren und der Sprechtraktwände repräsentieren. Die Form der Gitter bzw. der Artikulatoren wird mit Hilfe von 23 Parametern bestimmt. Für die Kontaktbehandlung zwischen den Artikulatoren und den Vokaltraktwänden sowie für die Berechnung der Flächenfunktion habe ich speziell an das Modell angepasste Verfahren entwickelt.

Das Glottismodell baut auf einem früheren Modell von TITZE [115] auf, das die zeitvariable dreidimensionale Form des glottalen Spalts mit Hilfe weniger Parameter beschreibt. Im Rahmen der vorliegenden Arbeit wurde das Modell analog zu CRANEN *et al.* [23] um eine Öffnung im hinteren Bereich zwischen den Stellknorpeln erweitert. Die Größe dieser Öffnung kann mit einem Parameter individuell eingestellt werden und wird strömungsakustisch in Abhängigkeit vom Grad der glottalen Abduktion entweder wie ein „verbundenes Leck“ oder wie ein paralleler Spalt zur glottalen Öffnung zwischen dem membranartigen Teil der Stimmlippen behandelt. Außerdem habe ich das Modell um ein realitätsnahes Verhalten bei Abduktions- und Adduktionsgesten erweitert. Für einen natürlicheren Klang des Stimmtons werden dem vorgegebenen F_0 -Verlauf zusätzlich kleine quasi-zufällige Schwankungen („Jitter“) überlagert.

Das subglottale System und der Nasenraum werden in dieser Arbeit direkt durch ihre diskreten Flächenfunktionen repräsentiert. Für den Nasenraum und die Nasennebenhöhlen wurden die anatomischen Daten von DANG und HONDA [24, 25] verwendet, und das subglottale System wurde entsprechend einer morphologischen Studie von WEIBEL [127] modelliert.

Die Flächenfunktionen für das subglottale System, die Glottis, den Vokaltrakt und den Nasenraum definieren gemeinsam die Geometrie eines verzweigten Rohrmodells, das den gesamten Sprechapparat repräsentiert. Dieses Rohrmodell ist die Grundlage für die

aerodynamisch-akustische Simulation. Sie basiert auf einer Repräsentation des Rohrmodells in Form eines akustischen Netzwerks, das analog zu einer verzweigten elektrischen Übertragungsleitung mit konzentrierten Elementen aufgebaut ist. In dieser Arbeit wurden zwei Simulationsmethoden für das Netzwerk vorgestellt: Eine Methode für die Analyse im Frequenzbereich und eine Methode für die Simulation im Zeitbereich. Für die Synthese zusammenhängender Äußerungen wird die Zeitbereichssimulation des Netzwerks eingesetzt. Die sekundäre Sprechtraktanregung habe ich in der Zeitbereichssimulation mit einem Verfahren realisiert, dass in Abhängigkeit von den Strömungsverhältnissen im Rohr automatisch Rauschquellen mit entsprechenden spektralen Eigenschaften in das Netzwerk einfügt. Dadurch ist die Synthese von Frikativen und aspirierten Plosiven möglich.

Eine besondere Eigenschaft der aerodynamisch-akustischen Simulationsmethode in dieser Arbeit ist, dass die Länge der einzelnen Rohrabschnitte des Rohrmodells variabel ist. Dadurch sind kontinuierliche Längenänderungen des Sprechtrakts während einer Äußerung möglich. Mit den derzeit in anderen Studien häufig für die akustische Simulation verwendeten Wellendigitalfiltern ist die Länge aller Rohrabschnitte dagegen fest mit der Abtastrate und der Schallgeschwindigkeit verknüpft, so dass sich die Rohrlänge hier nur jeweils sprunghaft um eine Rohrabschnittslänge ändern kann.

Eine weitere Besonderheit der Zeitbereichssimulation in dieser Arbeit ist die Ausnutzung der numerischen Dämpfung des zeitlichen Diskretisierungsschemas für die Herbeiführung angemessener Formantbandbreiten. Außerdem basiert die Simulation nicht auf zylindrischen Rohrabschnitten, sondern auf Abschnitten mit elliptischen Querschnitten. Dadurch kann der Reibungswiderstand für enge Rohrabschnitte mit hoher Exzentrizität (z. B. die Glottis) genauer simuliert werden.

Für die Generierung der artikulatorischen Bewegungsabläufe wurde ein Steuermodell entwickelt, das auf früheren Arbeiten von KRÖGER [66] und SALTZMAN *et al.* [98] aufbaut. Die grundlegende Einheit der Steuerung ist in diesem Modell die artikulatorische Geste. Eine Sprachäußerung wird als ein gesturaler Ablaufplan beschrieben, in dem die für die Äußerung erforderlichen Gesten definiert und zeitlich miteinander koordiniert werden müssen. Das Steuermodell überführt den Ablaufplan einer Äußerung in die zeitlichen Funktionsverläufe der Vokaltrakt- und Anregungsparameter. Diese steuern den zeitlichen Verlauf der Artikulation und somit das akustische Verhalten des Rohrmodells.

Das von mir entwickelte Steuermodell unterscheidet sich in mehreren Punkten von den früheren gestischen Modellen von SALTZMAN *et al.* und KRÖGER. Zum Einen wird die Dynamik der Gesten in der vorliegenden Arbeit nicht mit kritisch gedämpften Systemen zweiter Ordnung simuliert. Statt dessen wird der zeitliche Verlauf des gestischen Realisierungsgrads direkt mit Hilfe trigonometrischer Funktionen beschrieben. Weiterhin wird die Koartikulation durch ein Dominanzmodell gesteuert. Dieses Modell regelt die Prioritäten bei der zeitlichen Überlagerung mehrerer Gesten, die auf den gleichen Artikulator wirken. Außerdem integriert das Steuermodell durch die Verwendung des Überlagerungsmodells von FUJISAKI *et al.* [42] in konsistenter Weise die Generierung der Grundfrequenzkontur. Insgesamt ist das Steuermodell in dieser Arbeit konzeptionell einfacher aufgebaut als die früheren gestischen Steuermodelle.

Der Beitrag dieser Arbeit liegt neben der Entwicklung der Teilmodelle für die Synthese auch in ihrer Integration und Implementierung in einem Computerprogramm (siehe Anhang A).

6.2 Diskussion der Ergebnisse

Um die Möglichkeiten des artikulatorischen Synthetisators zu demonstrieren, habe ich mit ihm mehrere Sprachäußerungen generiert. Die Äußerungen finden Sie zur eigenen qualitativen Beurteilung als Audio- und Videodateien auf der beiliegenden CD. Eine formelle Evaluierung der Synthesequalität durch einen Hörtest wurde bisher nicht durchgeführt, wird aber in weiterführenden Arbeiten mit mehr synthetischen Äußerungen angestrebt. Informelle Hörtests mit dem Datenmaterial auf der CD und einer kleinen Gruppe von Hörern bestätigen jedoch bereits eine hohe Verständlichkeit der Syntheseergebnisse.

Die Synthesebeispiele belegen damit grundsätzlich, dass die in dieser Arbeit entwickelten Modelle und Methoden geeignet sind, Sprache von hoher Qualität zu generieren. Dabei muss aber betont werden, dass die gesturalen Ablaufpläne für die Äußerungen manuell erstellt wurden, und nicht automatisch auf der Basis eines Eingabetextes. Jedoch wurde die manuelle Erstellung der Ablaufpläne von Regeln geleitet, die in dieser Arbeit formuliert wurden. Lediglich die Dauer der einzelnen Laute wurde in längeren Äußerungen von natürlichen Sprachaufnahmen kopiert. Da für konkatenative Synthetisatoren bereits gute Vorhersagemodelle für die Lautdauer existieren, die auch für die artikulatorische Synthese adaptiert werden können, ist in Zukunft auch für das hier vorgestellte System eine vollautomatische Generierung gesturaler Ablaufpläne denkbar.

Durch die komplexe Simulation der Spracherzeugung ist der Rechenaufwand für die artikulatorische Sprachsynthese generell relativ hoch. Das im Rahmen dieser Arbeit implementierte System benötigt auf einem Intel Pentium 4 Computer mit 2 GHz Taktfrequenz ca. 14 Sekunden Rechenzeit für die Generierung von einer Sekunde Sprache. Durch algorithmische Optimierungen und die Nutzung erweiterter Befehlssätze aktueller Prozessoren lässt sich die Rechenzeit nach meiner Einschätzung voraussichtlich noch mindestens um die Hälfte reduzieren.

6.3 Ausblick

Die in dieser Arbeit entwickelte Sprachsynthese bietet eine Vielzahl von Ansatzpunkten für die Weiterentwicklung. An dieser Stelle sollen nur einige der Möglichkeiten aufgezählt werden.

Ein wichtiger Schritt in die Richtung eines vollständigen TTS-Systems wäre die Entwicklung eines Moduls, das einen Eingabetext in einen gesturalen Ablaufplan umwandelt. Hierfür müssten insbesondere die Regeln für die zeitliche Koordinierung der Gesten quantitativ formuliert werden. Einige Verarbeitungsschritte, wie die Text-Vorverarbeitung (Konversion von Zahlen, Abkürzungen, ...), die Analyse der Satz- und Wortstruktur und die Generierung prosodischer Parameter sind teilweise mit den aktuellen konkatenativen Syntheseverfahren identisch, so dass dazu auf vorhandene Forschungsergebnisse aufgebaut werden kann.

Unabhängig von der gesturalen Beschreibung von Äußerungen wäre die geometrische Anpassung des Vokaltraktmodells an einen konkreten Sprecher mit Hilfe von MRT-Volumendaten sinnvoll. Damit könnten einerseits die Zielformen des Vokaltrakts für die einzelnen Laute exakt eingestellt und andererseits die Dominanzwerte für die Parameter der konsonantischen Zielformen abgeleitet werden.

Wie im vorangegangenen Abschnitt bereits angemerkt wurde, ist auch eine Verkürzung der

Rechenzeit für die Simulation erstrebenswert. Hier sollten neben den programmiertechnischen Optimierungen auch Vereinfachungen an der Modellierung des Sprechapparats in Betracht gezogen werden. Beispielsweise ließe sich wahrscheinlich die Anzahl der Rohrabchnitte für das (statische) subglottale System auf weniger als die Hälfte reduzieren, ohne Abstriche bei den akustischen Eigenschaften im interessierenden Frequenzbereich machen zu müssen. Auch die Berechnung von weniger Querschnittsflächen des Vokaltraktmodells entlang der Mittellinie würde einen Geschwindigkeitsvorteil bringen, da die Schnittflächenberechnungen am 3D-Modell sehr zeitaufwendig sind.

Das Synthesystem kann auch als Grundlage für die Umsetzung und Evaluierung anderer Teilmodelle verwendet werden, z. B. für die Stimm lippen, die Erzeugung der Frikative oder die akustische Simulation. Dazu müssen die derzeit implementierten Teilmodelle nur gegen die neuen Modelle ausgetauscht werden, die dann im Zusammenspiel mit den anderen Komponenten des Synthesystems untersucht werden können. Ein vielversprechender alternativer Ansatz für die Erzeugung von Friktionsrauschen ist z. B. das Frikativmodell von SINDER [104], das den physikalischen Mechanismus der Geräuscherzeugung tatsächlich modelliert, anstatt ihn nur auf der Basis empirischer Daten nachzubilden. Für die stimmhafte Anregung des Sprechtrakts ließen sich alternativ selbstschwingende Glottismodelle einbinden. Um Verbesserungen oder Verschlechterungen der Synthesequalität beim Austausch von Teilmodellen beurteilen zu können, müssten dann jeweils Hörtests durchgeführt werden.

Mittelfristig und langfristig gesehen werden artikulatorische Sprachsynthetisatoren mit Sicherheit qualitativ an die besten konkatenativen Synthetisatoren heranreichen und darüber hinaus eine Vielzahl an neuen Möglichkeiten eröffnen.

Anhang A

Programmüberblick

Dieses Kapitel gibt einen kurzen Überblick über das Computerprogramm, in dem ich die Konzepte und Modelle dieser Arbeit umgesetzt habe. Das Programm wurde in der Programmiersprache C++ für das Betriebssystem Windows (ab Windows 98) geschrieben. Für die 3D-Grafikausgabe wurde die Grafikbibliothek OpenGL verwendet und die Soundeingebe und -ausgabe wurde mit dem DirectX SDK (DirectSound) realisiert. Das Programm

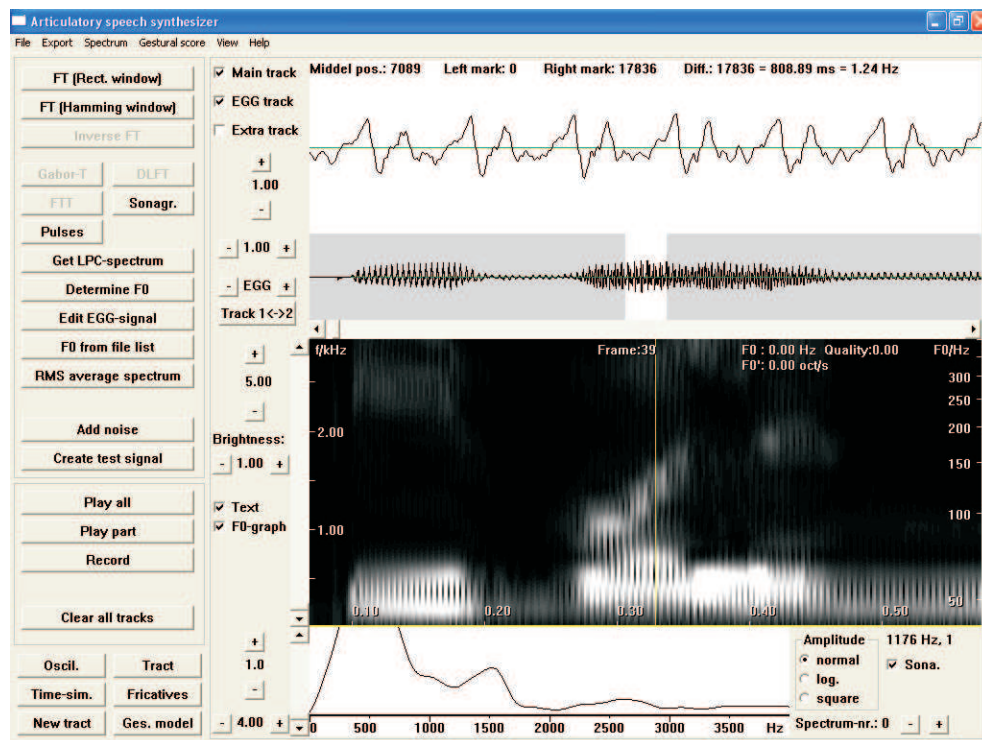


Abbildung A.1: Oszillogramm-Ansicht.

besitzt mehrere „Ansichten“, die jeweils eine grafische Benutzerschnittstelle für die einzelnen Programmteile bereitstellen. Die Funktionalität der wichtigsten Programmteile wird

nachfolgend kurz erläutert.

In Abb. A.1 ist die Oszillogramm-Ansicht dargestellt. Diese Ansicht stellt elementare Funktionen für die Analyse aufgenommenen oder synthetisierter Sprachsignale bereit. Rechts oben im Bild wird der Schalldruckverlauf des aktuellen Signals als Oszillogramm in zwei Zoomstufen angezeigt. Die obere der beiden Kurven repräsentiert dabei jeweils einen kleinen Ausschnitt aus der Mitte des darunter abgebildeten Signals. In der Zeichenfläche unter den Oszillogrammen können verschiedene Zeit-Frequenz-Darstellungen des Sprachsignals erzeugt werden. In Abb. A.1 ist aktuell das Breitband-Spektrogramm dargestellt. In der schmalen Zeichenfläche darunter ist der Frequenzverlauf in einem ausgewählten Zeitpunkt des Spektrogramms dargestellt (vertikale Linie im Spektrogramm). Mit den Schaltflächen links im Ansichtsfenster können verschiedene Analysen auf dem Sprachsignal durchgeführt werden, u. a. die Berechnung der diskreten Fourier-Transformation oder des LPC-Spektrums für einen ausgewählten Signalabschnitt. Desweiteren können hier Funktionen für die Grundfrequenzanalyse, die Analyse von EGG-Signalen und die Erzeugung verschiedener einfacher Testsignale aufgerufen werden.

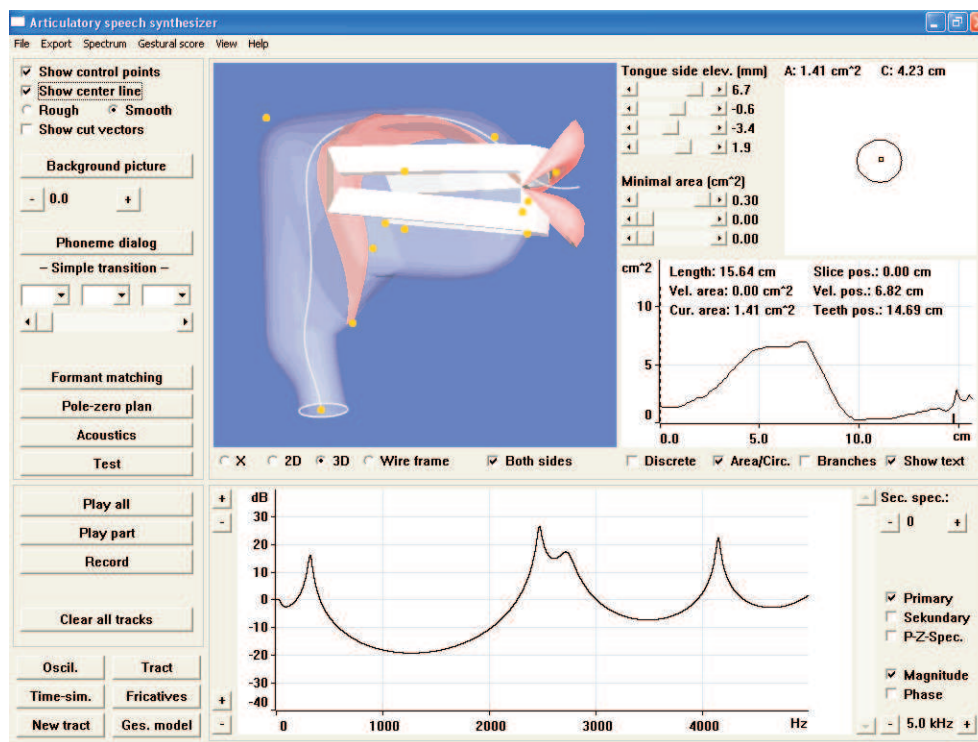


Abbildung A.2: Vokaltrakt-Ansicht.

Abbildung A.2 zeigt die Vokaltrakt-Ansicht, in der das Vokaltraktmodell dargestellt wird und seine Parameter eingestellt werden können. Außerdem wird hier die Flächenfunktion und die Vokaltrakt-Übertragungsfunktion dargestellt. Für die Vokaltraktdarstellung stehen verschiedene Optionen zur Auswahl, z. B. eine 2D-mediosagittale Darstellung oder eine 3D-Darstellung als Drahtgitter- oder Flächenmodell. Weiterhin kann das Modell interaktiv gedreht und gezoomt werden. Einige der Modellparameter können direkt im Dar-

stellungsfenster verändert werden, indem man die gelben Kontrollpunkte mit der Maus verschiebt. So steuert z. B. der gelbe Punkt¹ links oben im Bild die Position des weichen Gaumens. Die Parameter, die die Höhe der Zungenränder und die minimalen Querschnittsflächen bestimmen, können rechts neben dem Vokaltraktbild mit Hilfe von Schiebereglern verändert werden. Bei jeder interaktiven Veränderung eines Parameters wird sofort die Flächenfunktion (rechts in der Mitte) und die Volumenstrom-Übertragungsfunktion mit Hilfe der Frequenzbereichssimulation (unten im Fenster) neu berechnet. In der Vokaltraktdarstellung in Abb. A.2 sind auch die Mittellinie und der Querschnitt an einer bestimmten Stelle (hier direkt über der Stimmritze) eingezeichnet. Die Position des Querschnitts kann entlang der Mittellinie bewegt werden und ist rechts oben im Ansichtsfenster als Schnittbild dargestellt. In dieser Ansicht wurden die unterschiedlichen Zielformen des Vokaltrakts eingestellt. Dazu wurde jeweils ein entsprechendes Röntgen-Konturbild hinter der Darstellung des Modells eingeblendet und durch die interaktive Einstellung der Modellparameter eine geometrische Übereinstimmung zwischen den sagittalen Konturen hergestellt.

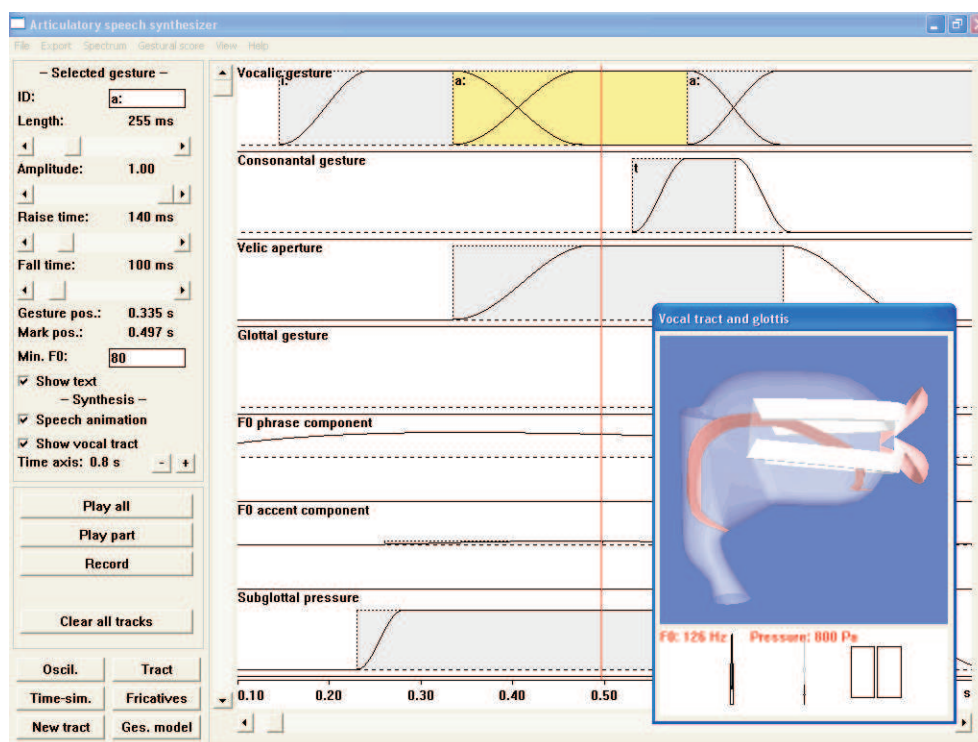


Abbildung A.3: Editor für die Erstellung gesturaler Ablaufpläne.

Abbildung A.3 zeigt den Editor für die Erstellung der gesturalen Ablaufpläne. Die Gesten können hier interaktiv mit der Maus platziert und koordiniert werden. Für die jeweils selektierte Geste (hier gelb dargestellt) können die Länge, Amplitude, Anklindauer und Abklindauer links neben dem Ablaufplan mit Schiebereglern eingestellt werden. Außerdem kann hier die Zeichenkette für den Deskriptor und der asymptotische Basiswert der

¹Alle Abbildungen in diesem Kapitel sind als Farbbilder auf der beiliegenden CD-ROM zu finden.

F_0 eingegeben werden. Wie in Abb. A.3 zu sehen ist, lässt sich über dem Editor ein Dialogfenster einblenden, in dem die Form des Vokaltrakts und die Einstellung der Stimmritze zu einem ausgewählten Zeitpunkt im Ablaufplan dargestellt werden. Der ausgewählte Zeitpunkt ist als lange vertikale Linie im Ablaufplan dargestellt und kann mit der Maus bewegt werden. Nach der Erstellung eines Ablaufplans kann das Sprachsignal mit oder ohne einer Animation der Artikulatorbewegungen synthetisiert werden.

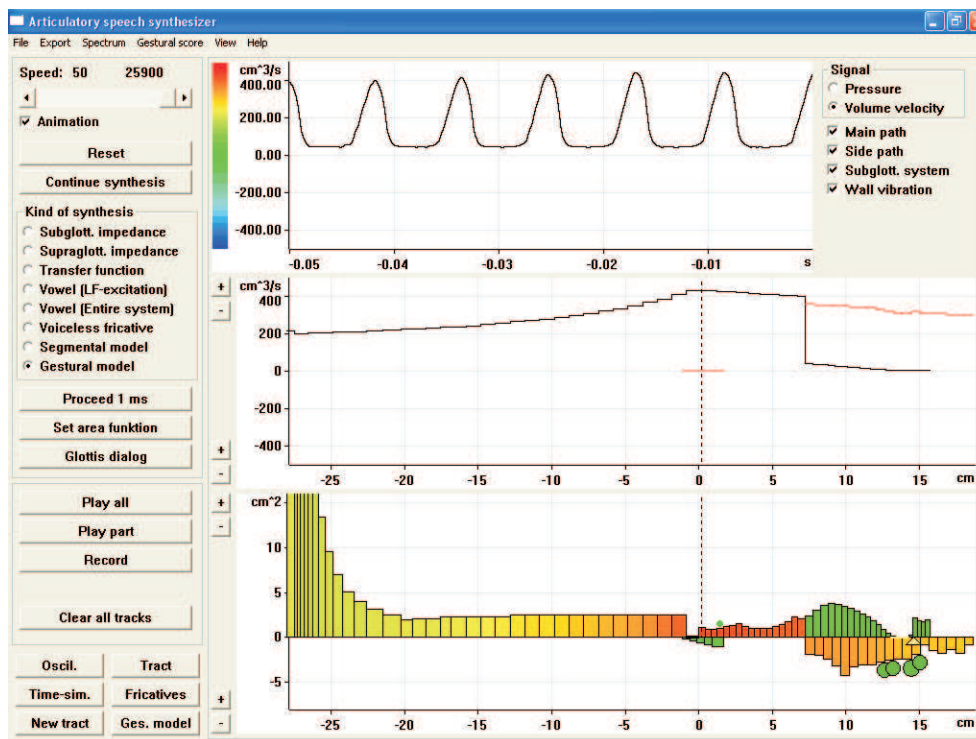


Abbildung A.4: Ansicht der aerodynamisch-akustischen Simulation im Zeitbereich.

Abbildung A.4 zeigt die Synthese-Ansicht, in der die aerodynamisch-akustische Simulation im Zeitbereich visualisiert wird. Während der Simulation werden hier die räumliche Verteilung und die zeitliche Entwicklung von entweder dem Schalldruck oder dem Volumenstrom als Animation dargestellt. In Abb. A.4 wird der Volumenstrom dargestellt.

Im rechten Teil des Ansichtsfensters befinden sich drei Grafiken, von denen die unteren beiden eine Momentaufnahme des Volumenstroms während der Synthese des Lautes [n] zeigen. In der unteren Grafik ist die Flächenfunktion des Sprechapparats mit einer farbcodierten Volumenstromverteilung dargestellt. Gelbe und rote Farben zeigen einen hohen Volumenstrom an und die grüne Farbe einen Volumenstrom von Null. Die grüne Einfärbung hinter der oralen Konstriktion zeigt somit an, dass keine Luft durch den Mundraum strömt. Die Grafik in der Mitte stellt direkt den Funktionsverlauf des Volumenstroms dar, wobei die schwarze Kurve den Strom durch das subglottale System, die Glottis, den Rachen- und den Mundraum zeigt und die rote Kurve den Volumenstrom durch den Nasenraum. In der oberen Grafik ist der zeitliche Verlauf des Volumenstroms direkt oberhalb der Stimmritze für die letzten 50 ms dargestellt. Der Rohrabschnitt, für den hier der Volumenstrom oder der

Schalldruck dargestellt werden soll, kann in der unteren Grafik mit der Maus selektiert werden (vertikale gestrichelte Linie). Im linken Teil des Ansichtsfensters stehen neben der Synthese auf der Basis des gesturalen Ablaufplans noch andere Synthesemodi zur Auswahl. So kann z. B. das subglottale oder das supraglottale System mit einem Volumenstrom-Dirac-Impuls direkt unterhalb bzw. oberhalb der Glottis angeregt werden. Aus dem zeitlichen Verlauf des abklingenden Schalldrucks unterhalb bzw. oberhalb der Stimmritze kann dann die Eingangsimpedanz des subglottalen bzw. supraglottalen Systems berechnet werden.

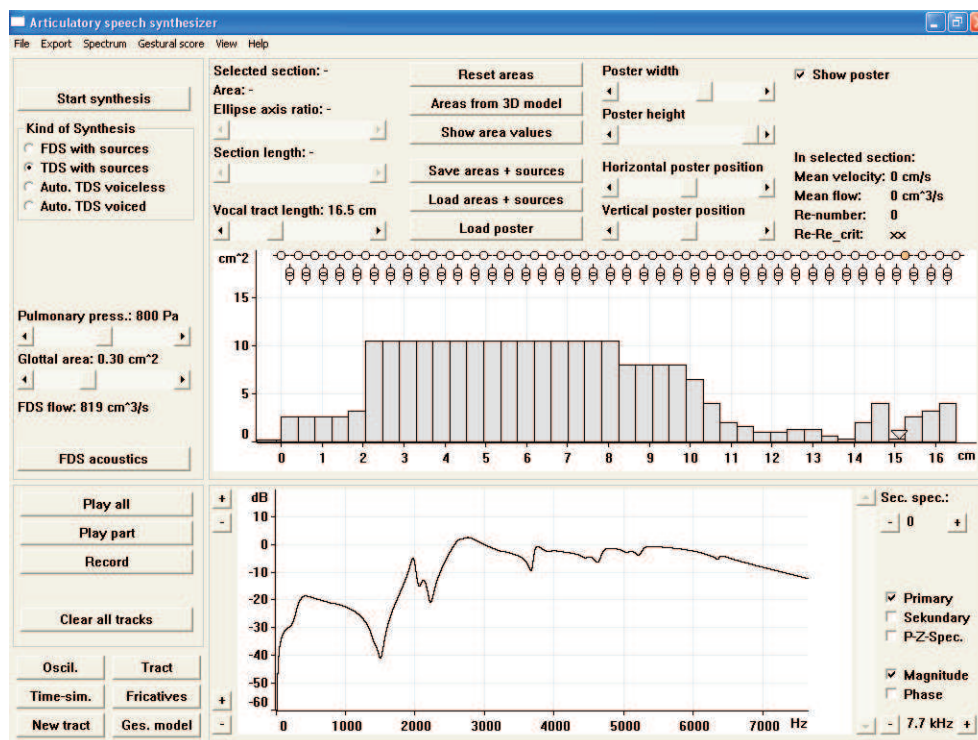


Abbildung A.5: Editor für die manuelle Einstellung von Flächenfunktionen und für die Platzierung von Geräuschquellen.

Abbildung A.5 zeigt die Ansicht, mit der die Flächenfunktionen und die Eigenschaften der Geräuschquellen für die Frikative untersucht wurden (siehe Abschnitt 3.8). Die Grafik in der Mitte des Ansichtsfensters zeigt die diskrete Flächenfunktion des Rachen- und Mundraums. Für jeden Rohrabschnitt kann hier individuell die Länge, der Umfang und die Querschnittsfläche eingestellt werden. Außerdem kann in jedem der Rohrabschnitte eine Schalldruck- und/oder Volumenstromquelle für die sekundäre Anregung aktiviert und parametrisiert (Quellspektrum, Amplitude) werden. Das aus der Flächenfunktion und den Geräuschquellen resultierende Spektrum des von der Mundöffnung abgestrahlten Schalldrucksignals wird in der unteren Grafik dargestellt. In Abb. A.5 ist die Flächenfunktion und das Schalldruckspektrum für den Frikativ [f] dargestellt. Das Spektrum wird sofort nach jeder Änderung der Parameter der Flächenfunktion oder der Geräuschquellen neu berechnet und kann auf Wunsch gleichzeitig mit dem gemessenen Spektrum eines Frikativs dargestellt und verglichen werden. Links neben der Flächenfunktion im Ansichtsfenster

können noch der pulmonale Druck und die Öffnungsfläche des glottalen Spalts eingestellt werden, die zusätzlich in die Berechnung des synthetischen Spektrums eingehen. Über die Schaltfläche „Start synthesis“ wird aus dem Spektrum das Zeitsignal des Schalldrucks berechnet und abgespielt.

Anhang B

Konstanten und Symbole

B.1 Physikalische Konstanten

Die physikalischen Konstanten für feuchte Luft bei Körpertemperatur sind laut FLANAGAN [38]:

Physikalische Größe	Symbol	Wert	Einheit
Atmosphärischer Druck (Normdruck)	p_0	$1.013 \cdot 10^5$	Pa
Viskosität	μ	$1.86 \cdot 10^{-5}$	kg/(m · s)
Kinematische Viskosität	ν	$1.63 \cdot 10^{-5}$	m ² /s
Mittlere Dichte	ϱ_0	1.14	kg/m ³
Schallgeschwindigkeit	c	350.0	m/s
Adiabatenexponent	η	1.4	—

Tabelle B.1: Thermodynamische Konstanten, die in der aerodynamisch-akustischen Simulation verwendet wurden.

B.2 Phonemsymbole des Deutschen nach IPA und SAMPA

IPA	SAMPA	Beschreibung	Beispiel
i:	i:	vorderer, geschlossener, ungerundeter Vokal	Lied
e:	e:	vorderer, halb geschlossener, ungerundeter Vokal	Beet
ɛ:	E:	vorderer, halb offener, ungerundeter Vokal	spät
a:	a:	vorderer, offener, ungerundeter Vokal	Tat
o:	o:	hinterer, halb geschlossener, gerundeter Vokal	rot
u:	u:	hinterer, geschlossener, gerundeter Vokal	Tube
y:	y:	vorderer, geschlossener, gerundeter Vokal	süß
ø:	2:	vorderer, halb geschlossener, gerundeter Vokal	schön

Tabelle B.2: Gespannte Vokale

IPA	SAMPA	Beschreibung	Beispiel
ɪ	I	vorderer, fast geschlossener, ungerundeter Vokal	mit
ɛ	E	vorderer, halb offener, ungerundeter Vokal	Gesetz
a	a	vorderer, offener, ungerundeter Vokal	Satz
ɔ	O	hinterer, halb offener, gerundeter Vokal	Topf
ʊ	U	hinterer, geschlossener, gerundeter Vokal	Mutter
ʏ	Y	vorderer, fast geschlossener, gerundeter Vokal	hübsch
œ	9	vorderer, halb offener, gerundeter Vokal	plötzlich
ə	@	zentraler, neutraler, ungerundeter Vokal (Schwa)	bitte
ɐ	6	zentraler, neutraler, ungerundeter Vokal mit r-Färbung	besser

Tabelle B.3: Ungespannte Vokale

IPA	SAMPA	Beschreibung	Beispiel
aɪ	aI	Gleitbewegung von a nach ɪ	Eis
aʊ	aU	Gleitbewegung von a nach ʊ	Haus
ɔʏ	OY	Gleitbewegung von ɔ nach ʏ	Kreutz

Tabelle B.4: Diphthonge

IPA	SAMPA	Beschreibung	Beispiel
p	p	stimmloser, bilabialer Plosiv	Pein
b	b	stimmhafter, bilabialer Plosiv	Bein
t	t	stimmloser, alveolarer Plosiv	Teich
d	d	stimmhafter, alveolarer Plosiv	Deich
k	k	stimmloser, velarer Plosiv	Kunst
g	g	stimmhafter, velarer Plosiv	Gunst

Tabelle B.5: Plosive

IPA	SAMPA	Beschreibung	Beispiel
f	f	stimmloser, labiodentaler Frikativ	fast
v	v	stimmhafter, labiodentaler Frikativ	was
s	s	stimmloser, alveolarer Frikativ	Tasse
z	z	stimmhafter, alveolarer Frikativ	Hase
ʃ	S	stimmloser, postalveolarer Frikativ	Schuhe
ʒ	Z	stimmhafter, postalveolarer Frikativ	Garage
ç	C	stimmloser, palataler Frikativ	ich
j	j	ungerundeter, vorderer Gleitlaut	ja
x	x	stimmloser, velarer Frikativ	Buch
h	h	stimmloser, glottaler Frikativ	Hand

Tabelle B.6: Frikative und Gleitlaute

IPA	SAMPA	Beschreibung	Beispiel
m	m	bilabialer Nasal	m ein
n	n	dentaler/alveolarer Nasal	n ein
ŋ	N	velarer Nasal	M en ge
l	l	alveolarer Lateral	L eim
r	R	uvularer Vibrant	r eich

Tabelle B.7: Nasale und Liquide

Anhang C

Lebenslauf, Veröffentlichungen, Vorträge

Wissenschaftlicher Lebenslauf von Peter Birkholz

8. April 1978	geboren in Rostock, Deutschland
1984 - 1991	Grund- und Gesamtschule in Wustrow
1991 - 1996	Gymnasium in Ribnitz-Damgarten
1996 - 1997	Bundeswehr
Oktober 1997 - Juli 2002	Studium der Informatik an der Universität Rostock mit Abschluss als Diplom-informatiker (Vordiplom 1999, Diplom 2002)
Oktober 2000 - März 2001	Praktikumssemester bei der Allgemeinen Versicherungs-Software GmbH (AVSG) in Köln
Juni 2001 - Juni 2002	HiWi-Tätigkeit am Fraunhofer Institut für Graphische Datenverarbeitung in Rostock
August 2002 - Juli 2005	Doktorand im Graduiertenkolleg 466 an der Universität Rostock

Liste der Veröffentlichungen

- BIRKHOLZ, P.: *Entwicklung eines dreidimensionalen Artikulatormodells für die Sprachsynthese*. Diplomarbeit, Institut für Informatik, Universität Rostock, 2002.
- BIRKHOLZ, P.: *Grundfrequenzbestimmung unter Berücksichtigung linearer Frequenzänderungen*. In: *Fortschritte der Akustik, DAGA '03*, S. 768–769, 2003.

- BIRKHOLZ, P. und D. JACKÈL: *Sprachgesteuerte Gesichtsanimation*. Rostocker Informatik Berichte, 28, 2003.
- BIRKHOLZ, P. und D. JACKÈL: *A Three-Dimensional Model of the Vocal Tract for Speech Synthesis*. In: *Proceedings of the 15th International Congress of Phonetic Sciences*, S. 2597–2600, Barcelona, Spain, 2003.
- BIRKHOLZ, P. und D. JACKÈL: *Automatic Fricative Noise Generation in the Vocal Tract Based on Aeroacoustic Principles (Abstract only)*. In: *German-French Summerschool on 'Cognitive and Physical Models of Speech Production, Perception and Perception-Production Interaction'*, Lubmin, Germany, 2004.
- BIRKHOLZ, P. und D. JACKÈL: *Boundary-Layer Resistance in Time-Domain Simulations of the Vocal Tract System*. In: *12th European Signal Processing Conference (EUSIPCO-2004)*, S. 999–1002, Vienna, Austria, 2004.
- BIRKHOLZ, P. und D. JACKÈL: *Influence of Temporal Discretization Schemes on Formant Frequencies and Bandwidths in Time Domain Simulations of the Vocal Tract System*. In: *Interspeech 2004-ICSLP*, Jeju, Korea, 2004.
- BIRKHOLZ, P. und D. JACKÈL: *Simulation of Flow and Acoustics in the Vocal Tract*. In: *CFA/DAGA '04*, S. 895–896, Strasbourg, France, 2004.
- BIRKHOLZ, P. und D. JACKÈL: *Artikulatorische Sprachsynthese mit dem Programm TractSyn - Ein Überblick*. In: *Fortschritte der Akustik, DAGA '05*, 2005.

Liste der eingeladenen Fachvorträge

- *Ein geometrisches dreidimensionales Artikulationsmodell für die Sprachsynthese*, Zentrum für allgemeine Sprachwissenschaft, Typologie und Universalienforschung (ZAS), Berlin, 21. März 2002
- *Ein dreidimensionales Artikulatormodell für die Sprachsynthese*, Fraunhofer Institut FIRST, Berlin, 13. September 2002
- *Ein dreidimensionales Sprechtraktmodell für die artikulatorische Sprachsynthese*, Institut für Sprache und Kommunikation, TU Berlin, 28. Januar 2003
- *Audio-Visual Articulatory Speech Synthesis*, Max-Planck-Institut für biologische Kybernetik, Tübingen, 3. Juni 2005

Literaturverzeichnis

- [1] ABRY, C. und L.-J. BOË: *Laws for Lips*. Speech Communication, 5:97–104, 1986.
- [2] ALIPOUR, F., D. A. BERRY und I. R. TITZE: *A Finite-Element Model of Vocal-Fold Vibration*. Journal of the Acoustical Society of America, 108(6):3003–3012, 2000.
- [3] ALLEN, D. R. und W. J. STRONG: *A Model for the Synthesis of Natural Sounding Vowels*. Journal of the Acoustical Society of America, 78(1):58–69, 1985.
- [4] ANANTHAPADMANABHA, T. V. und G. FANT: *Calculation of True Glottal Flow and its Components*. Speech Communication, 1:167–184, 1982.
- [5] BADIN, P.: *Acoustics of Voiceless Fricatives: Production Theory and Data*. STL-QPSR, 3:33–55, 1989.
- [6] BADIN, P., G. BAILLY, L. REVÉRET, M. BACIU, C. SEGEBARTH und C. SAVARIAUX: *Three-Dimensional Linear Articulatory Modeling of Tongue, Lips and Face, based on MRI and Video Images*. Journal of Phonetics, 30:533–553, 2002.
- [7] BADIN, P. und G. FANT: *Notes on Vocal Tract Computation*. STL-QPSR, 2–3:53–108, 1984.
- [8] BADIN, P., C. SHADLE, Y. P. T. NGOC, J. N. CARTER, W. S. C. CHIU, C. SCULLY und K. STROMBERG: *Frication and Aspiration Noise Sources: Contribution of Experimental Data to Articulatory Synthesis*. In: *International Conference on Spoken Language Processing*, S. 163–166, Yokohama, Japan, 1994.
- [9] BEAUTEMPS, D., P. BADIN und G. BAILLY: *Linear Degrees of Freedom in Speech Production: Analysis of Cineradio- and Labio-Film Data and Articulatory-Acoustic Modeling*. Journal of the Acoustical Society of America, 109(5):2165–2180, 2001.
- [10] BERANEK, L. L.: *Acoustics*. McGraw-Hill Book Company, Inc., 1954.
- [11] BERG, J. V. D., J. T. ZANTEMA und P. J. DOORNENBAL: *On the Air Resistance and the Bernoulli Effect of the Human Larynx*. Journal of the Acoustical Society of America, 29(5):626–631, 1957.
- [12] BERRY, D. A., J. B. MOON und D. P. KUEHN: *A Finite Element Model of the Soft Palate*. Cleft Palate - Craniofacial Journal, 36(3):217–222, 1999.

- [13] BIRKHOLZ, P. und D. JACKÈL: *A Three-Dimensional Model of the Vocal Tract for Speech Synthesis*. In: *Proceedings of the 15th International Congress of Phonetic Sciences*, S. 2597–2600, Barcelona, Spain, 2003.
- [14] BOERSMA, P.: *Functional Phonology*. Doktorarbeit, University of Amsterdam, 1998.
- [15] BRONSTEIN, I. N., K. A. SEMENDJAJEW, G. MUSIOL und H. MÜHLIG: *Taschenbuch der Mathematik*. Verlag Harri Deutsch, 2001.
- [16] BROWMAN, C. P. und L. GOLDSTEIN: *Articulatory Phonology: An Overview*. *Phonetica*, 49:155–180, 1992.
- [17] CHEN, M.: *Vowel Length Variation as a Function of the Voicing of the Consonant Environment*. *Phonetica*, 22:129–159, 1970.
- [18] CHILDERS, D. G., D. M. HICKS, G. P. MOORE und Y. A. ALSAKA: *A Model for Vocal Fold Vibratory Motion, Contact Area, and the Electroglottogram*. *Journal of the Acoustical Society of America*, 80(5):1309–1320, 1986.
- [19] COKER, C. H.: *A Model of Articulatory Dynamics and Control*. *Proceedings of the IEEE*, 64(4):452–460, 1976.
- [20] CRANEN, B. und L. BOVES: *Pressure Measurements during Speech Production Using Semiconductor Miniature Pressure Transducers: Impact on Models for Speech Production*. *Journal of the Acoustical Society of America*, 77(4):1543–1551, 1985.
- [21] CRANEN, B. und L. BOVES: *On Subglottal Formant Analysis*. *Journal of the Acoustical Society of America*, 81(3):734–746, 1987.
- [22] CRANEN, B. und J. SCHROETER: *Modeling a Leaky Glottis*. *Journal of Phonetics*, 23:165–177, 1995.
- [23] CRANEN, B. und J. SCHROETER: *Physiologically Motivated Modelling of the Voice Source in Articulatory Analysis/Synthesis*. *Speech Communication*, 19:1–19, 1996.
- [24] DANG, J. und K. HONDA: *Morphological and Acoustical Analysis of the Nasal and the Paranasal Cavities*. *Journal of the Acoustical Society of America*, 96(4):2088–2100, 1994.
- [25] DANG, J. und K. HONDA: *Acoustic Characteristics of the Human Paranasal Sinuses Derived from Transmission Characteristic Measurement and Morphological Observation*. *Journal of the Acoustical Society of America*, 100(5):3374–3383, 1996.
- [26] DANG, J. und K. HONDA: *Construction and Control of a Physiological Articulatory Model*. *Journal of the Acoustical Society of America*, 115(2):853–870, 2004.
- [27] DI CRISTO, A. und D. J. HIRST: *Modelling French Micromelody: Analysis and Synthesis*. *Phonetica*, 43:11–30, 1986.
- [28] EL-MASRI, S., X. PELORSON, P. SAGUET und P. BADIN: *Vocal Tract Acoustics Using the Transmission Line Matrix (TLM) Method*. In: *4th International Conference on Spoken Language Processing*, S. 953–956, Philadelphia, USA, 1996.

- [29] ENGWALL, O.: *Vocal Tract Modeling in 3D*. TMH-QPSR, 1:1–8, 1999.
- [30] ENGWALL, O.: *Tongue Talking: Studies in Intraoral Speech Synthesis*. Doktorarbeit, Royal Institute of Technology, Stockholm, 2002.
- [31] FAGEL, S.: *Audiovisuelle Sprachsynthese: Systementwicklung und -bewertung*. Doktorarbeit, Fakultät Geisteswissenschaften der Technischen Universität Berlin, 2004.
- [32] FANAGAN, J. L., K. ISHIZAKA und K. L. SHIPLEY: *Signal Models for Low Bit-Rate Coding of Speech*. Journal of the Acoustical Society of America, 68(3):780–791, 1980.
- [33] FANT, G.: *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.
- [34] FANT, G. und M. BAVEGARD: *Parameteric Model of VT Area Functions: Vowels and Consonants*. TMH-QPSR, 1:1–21, 1997.
- [35] FANT, G., K. ISHIZAKA, J. LINDQVIST und J. SUNDBERG: *Subglottal Formants*. STL-QPSR, 1:1–12, 1972.
- [36] FANT, G., L. NORD und P. BRANDERUD: *A Note on the Vocal Tract Wall Impedance*. STL-QPSR, 4:13–20, 1976.
- [37] FARIN, G.: *Kurven und Flächen im Computer Aided Geometric Desing*. Academic Press, Inc., 1993.
- [38] FLANAGAN, J. L.: *Speech Analysis, Synthesis and Perception*. Springer-Verlag, Berlin, 1965.
- [39] FLANAGAN, J. L. und L. CHERRY: *Excitation of Vocal-Tract Synthesizers*. Journal of the Acoustical Society of America, 45(3):764–769, 1969.
- [40] FLANAGAN, J. L., K. ISHIZAKA und K. L. SHIPLEY: *Synthesis of Speech From a Dynamic Model of the Vocal Cords and Vocal Tract*. The Bell System Technical Journal, 54(3):485–506, 1975.
- [41] FUCHS, S.: *Articulatory Correlates of the Voicing Contrast in Alveolar Obstruent Production in German*. Doktorarbeit, Queen Margaret University College, Edinburgh, 2004.
- [42] FUJISAKI, H. und K. HIROSE: *Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese*. Journal of the Acoustical Society of Japan (E), 5(4):233–241, 1984.
- [43] FUJISAKI, H., M. LJUNGQVIST und H. MURATA: *Analysis and Modeling of Word Accent and Sentence Intonation in Swedish*. In: *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, Bd. 2, S. 211–214, 1993.
- [44] FUJISAKI, H. und S. OHNO: *Analysis and Modeling of Fundamental Frequency Contours of English Utterances*. In: *Fourth European Conference on Speech Communication and Technology (EUROSPEECH '95)*, S. 985–988, 1995.

- [45] GAY, T.: *Articulatory Movements in VCV Sequences*. Journal of the Acoustical Society of America, 62(1):183–193, 1977.
- [46] GERARD, J.-M., P. PERRIER, Y. PAYAN und R. WILHELMS-TRICARICO: *A New 3D Dynamical Biomechanical Tongue Model*. Journal of the Acoustical Society of America, 115(5):2431 (A), 2001.
- [47] HIROSE, H.: *Laryngeal Adjustments in Consonant Production*. Phonetica, 34:289–294, 1977.
- [48] HOWE, M. S.: *Contributions to the Theory of Aerodynamic Sound, with Application to Excess Jet Noise and the Theory of the Flute*. Journal of Fluid Mechanics, 71(4):625–673, 1975.
- [49] HUTTER, K.: *Fluid- und Thermodynamik*. Springer, 1995.
- [50] HUTTERS, B.: *Vocal Fold Adjustments in Aspirated and Unaspirated Stops in Danish*. Phonetica, 42:1–24, 1985.
- [51] ISHIZAKA, K. und J. L. FLANAGAN: *Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords*. The Bell System Technical Journal, 51(6):1233–1268, 1972.
- [52] ISHIZAKA, K., J. C. FRENCH und J. L. FLANAGAN: *Direct Determination of Vocal Tract Wall Impedance*. IEEE Transactions on Acoustics, Speech, and Signal Processing, 23(4):370–373, 1975.
- [53] ISHIZAKA, K., M. MATSUDAIRA und T. KANEKO: *Input Acoustic-Impedance Measurement of the Subglottal System*. Journal of the Acoustical Society of America, 60(1):190–197, 1976.
- [54] KELLY, J. L. und C. C. LOCHBAUM: *Speech Synthesis*. In: *Proceedings of the Fourth International Congress on Acoustics*, Bd. G42, S. 1–4, Kopenhagen, 1962.
- [55] KELSO, J. A. S., E. L. SALTZMAN und B. TULLER: *The Dynamical Perspective on Speech Production: Data and Theory*. Journal of Phonetics, 14:29–59, 1986.
- [56] KIM, C.-W.: *A Theory of Aspiration*. Phonetica, 21:107–116, 1970.
- [57] KING, S. A. und R. E. PARENT: *Creating Speech-Synchronized Animation*. IEEE Transactions on Visualization and Computer Graphics, 11(3):341–352, 2005.
- [58] KLATT, D. H.: *Software for a Cascade/Parallel Formant Synthesizer*. Journal of the Acoustical Society of America, 67(3):971–1000, 1980.
- [59] KLATT, D. H.: *Review of Text-to-Speech Conversion for English*. Journal of the Acoustical Society of America, 82(3):737–793, 1987.
- [60] KLATT, D. H. und L. C. KLATT: *Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers*. Journal of the Acoustical Society of America, 87(2):820–857, 1990.

- [61] KOB, M.: *Physical Modeling of the Singing Voice*. Doktorarbeit, Rheinisch-Westfälische Technische Hochschule Aachen, 2002.
- [62] KOHLER, K. J.: *Timing of Articulatory Control in the Production of Plosives*. *Phonetica*, 38:116–125, 1980.
- [63] KOIZUMI, T., S. TANIGUCHI und S. HIROMITSU: *Two-Mass Models of the Vocal Cords for Natural Sounding Voice Synthesis*. *Journal of the Acoustical Society of America*, 82(4):1179–1192, 1987.
- [64] KOLLIA, H. B., V. L. GRACCO und K. S. HARRIS: *Articulatory Organization of Mandibular, Labial, and Velar Movements During Speech*. *Journal of the Acoustical Society of America*, 98(3):1313–1324, 1995.
- [65] KONECZNA, H. und W. ZAWADOWSKI: *Obrazy Rentgenograficzne Glosek Rosyjskich*. Panstwowe Wydawnictwo Naukowe, 1956.
- [66] KRÖGER, B. J.: *Ein phonetisches Modell der Sprachproduktion*. Niemeyer, Tübingen, 1998.
- [67] KRÖGER, B. J., P. HOOLE, R. SADER, C. GENG, B. POMPINO-MARSCHALL und C. NEUSCHAEFER-RUBE: *MRT-Sequenzen als Datenbasis eines visuellen Artikulationsmodells*. *Phoniatrie und Pädaudiologie*, 52:837–843, 2004.
- [68] LILJENCRAFTS, J.: *A Translating and Rotating Mass Model of the Vocal Folds*. *STL-QPSR*, 1:1–18, 1991.
- [69] LINDBLOM, B. E. F. und J. E. F. SUNDBERG: *Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement*. *Journal of the Acoustical Society of America*, 50(4):1166–1179, 1971.
- [70] LINDQVIST-GAUFFIN, J. und J. SUNDBERG: *Acoustic Properties of the Nasal Tract*. *Phonetica*, 33:161–168, 1976.
- [71] LÖFQVIST, A., L. L. KOENIG und R. S. MCGOWAN: *Vocal Tract Aerodynamics in /aCa/ Utterances: Measurements*. *Speech Communication*, 16:49–66, 1995.
- [72] MAEDA, S.: *A Digital Simulation of the Vocal-Tract System*. *Speech Communication*, 1:199–229, 1982.
- [73] MAEDA, S.: *The Role of the Sinus Cavities in the Production of Nasal Vowels*. In: *Proceedings of the IEEE International Conference of Acoustics, Speech, Signal Processing*, Bd. 2, S. 911–914, 1982.
- [74] MAEDA, S.: *An Articulatory Model Based on Statistical Analysis*. *Journal of the Acoustical Society of America*, 84(Sup. 1):146, 1988.
- [75] MATSUZAKI, H. und K. MOTOKI: *FEM Analysis of 3D Vocal Tract Model with Asymmetrical Shape*. In: *5th Seminar on Speech Production and CREST Workshop on Models of Speech Production*, S. 329–332, Seon, Bavaria, Germany, 2000.

- [76] MAWASS, K., P. BADIN und G. BAILLY: *Synthesis of French Fricatives by Audio-Video to Articulatory Inversion*. ACUSTICA, 86:136–146, 2000.
- [77] MERMELSTEIN, P.: *Articulatory Model for the Study of Speech Production*. Journal of the Acoustical Society of America, 53(4):1070–1082, 1973.
- [78] MEYER, P., R. WILHELMS und H. W. STRUBE: *A Quasiarticulatory Speech Synthesizer for German Language Running in Real Time*. Journal of the Acoustical Society of America, 86(2):523–540, 1989.
- [79] MEYER-EPPLER, W.: *Zum Erzeugungsmechanismus der Geräuschlaute*. Zeitschrift für Phonetik und allgemeine Sprachwissenschaft, 7:196–212, 1953.
- [80] MIXDORFF, H.: *Intonation Patterns of German - Model-Based Quantitative Analysis and Synthesis of F0 Contours*. Doktorarbeit, Fakultät für Elektrotechnik der Universität Dresden, 1998.
- [81] MÖBIUS, B.: *Ein quantitative Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen*. Max Niemeyer Verlag, Tübingen, 1993.
- [82] MÖBIUS, B., A. ZIMMERMANN und W. HESS: *Untersuchungen zu mikroprosodischen Grundfrequenzvariationen im Deutschen*. In: *Analyse und Synthese gesprochener Sprache*. H. G. Tillmann and G. Willee, Olms, Hildesheim, 1987.
- [83] MOLL, K. L. und R. G. DANILOFF: *Investigation of the Timing of Velar Movements during Speech*. Journal of the Acoustical Society of America, 50(2):678–684, 1971.
- [84] MOOSHAMMER, C., A. GNEUMANN, P. HOOLE, P. ALFONSO, P. VAN LIESHOUT und S. FUCHS: *Coordination of lingual and mandibular gestures for different manners of articulation*. In: *Proceedings of the 15th International Congress of Phonetic Sciences*, S. 81–84, Barcelona, Spain, 2003.
- [85] MORSE, P. und K. U. INGARD: *Theoretical Acoustics*. Princeton University Press, 1987.
- [86] MUNHALL, K. G., E. VATIKIOTIS-BATESON und Y. TOHKURA: *X-Ray Film Database for Speech Research*. ATR Technical Report, ATR Human Information Processing Research Laboratories, 1994.
- [87] NARAYANAN, S. und A. ALWAN: *Noise Source Models for Fricative Consonants*. IEEE Transactions on Speech and Audio Processing, 8(2):328–344, 2000.
- [88] NISHIKAWA, K., K. ASAMA, K. HAYASHI, H. TAKANOBU und A. TAKANISHI: *Development of a Talking Robot*. In: *5th Seminar on Speech Production and CREST Workshop on Models of Speech Production*, S. 329–332, Seon, Bavaria, Germany, 2000.
- [89] ÖHMAN, S. E. G.: *Coarticulation in VCV Utterances: Spectrographic Measurements*. Journal of the Acoustical Society of America, 39(1):151–168, 1966.

- [90] ÖHMAN, S. E. G.: *Numerical Model of Coarticulation*. Journal of the Acoustical Society of America, 41(2):310–320, 1967.
- [91] OLIVEIRA ROSA, M. DE, J. C. PEREIRA, M. GRELLET und A. ALWAN: *A Contribution to Simulating a Three-Dimensional Larynx Model Using the Finite Element Method*. Journal of the Acoustical Society of America, 114(5):2893–2905, 2003.
- [92] PARUSH, A., D. J. OSTRY und K. G. MUNHALL: *A Kinematic Study of Lingual Coarticulation in VCV Sequences*. Journal of the Acoustical Society of America, 74(4):1115–1125, 1983.
- [93] PELORSON, X., A. HIRSCHBERG, R. R. VAN HASSEL, A. P. J. WIJNANDS und Y. AUREGAN: *Theoretical and Experimental Study of Quasisteady-Flow Separation within the Glottis during Phonation. Application to a Modified Two-Mass Model*. Journal of the Acoustical Society of America, 96(6):3416–3431, 1994.
- [94] PELORSON, X., C. VESCOVI, E. CASTELLI, A. HIRSCHBERG, A. P. J. WIJNANDS und H. M. A. BAILLIET: *Description of the Flow Through In-Vitro Models of the Glottis During Phonation. Application to Voiced Sound Synthesis*. Acta Acustica, 82:358–361, 1996.
- [95] PERKELL, J. S.: *A Physiologically-Oriented Model of Tongue Activity in Speech Production*. Doktorarbeit, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1974.
- [96] POZRIKIDIS, C.: *Fluid Dynamics - Theory, Computation and Numerical Simulation*. Kluwer Academic Publishers, 2001.
- [97] RUBIN, P., T. BAER und P. MERMELSTEIN: *An Articulatory Synthesizer for Perceptual Research*. Journal of the Acoustical Society of America, 70(2):321–328, 1981.
- [98] SALTZMAN, E. L. und K. G. MUNHALL: *A Dynamic Approach to Gestural Patterning in Speech Production*. Ecological Psychology, 1:333–382, 1989.
- [99] SCHNELL, K.: *Rohrmodelle des Sprechtraktes – Analyse, Parameterschätzung und Syntheseeexperimente*. Doktorarbeit, Johann Wolfgang Goethe-Universität, Frankfurt am Main, 2003.
- [100] SCHROETER, J. und B. CRANEN: *Physiologically-Motivated Modeling of the Voice Source in Articulatory Analysis/Synthesis*. In: *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech 1993)*, Bd. 1, S. 95–98, Berlin, 1993.
- [101] SHADLE, C. H.: *The Effect of Geometry on Source Mechanisms of Fricative Consonants*. Journal of Phonetics, 19:409–424, 1991.
- [102] SHADLE, C. H. und R. I. DAMPER: *Prospects for Articulatory Synthesis: A Position Paper*. In: *Fourth ISCA ITRW on Speech Synthesis*, 2001.
- [103] SILVERMAN, K.: *F0 Segmental Cues Depend on Intonation: The Case of the Rise after Voiced Stops*. Phonetica, 43:76–91, 1986.

- [104] SINDER, D. J.: *Speech Synthesis Using an Aeroacoustic Fricative Model*. Doktorarbeit, New Brunswick Rutgers, the State University of New Jersey, 1999.
- [105] SONDHI, M. M.: *Resonances of a Bent Vocal Tract*. Journal of the Acoustical Society of America, 79(4):1113–1116, 1986.
- [106] SONDHI, M. M. und J. SCHROETER: *A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer*. IEEE Transactions on Acoustics, Speech and Signal Processing, 35(7):955–967, 1987.
- [107] STARK, J., C. ERICSDOTTER, P. BRANDERUD, J. SUNDBERG, H.-J. LUNDBERG und J. LANDER: *The APEX Model as a Tool in the Specification of Speaker-Specific Articulatory Behavior*. In: *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, California, 1999.
- [108] STEVENS, K. N.: *Airflow and Turbulence Noise for Fricative and Stop Consonants: Static Considerations*. Journal of the Acoustical Society of America, 50(4):1180–1191, 1971.
- [109] STEVENS, K. N.: *On the Quantal Nature of Speech*. Journal of Phonetics, 17:3–45, 1989.
- [110] STEVENS, K. N.: *Acoustic Phonetics*. The MIT Press, 1998.
- [111] STONE, M. und A. LUNDBERG: *Three-Dimensional Tongue Surface Shapes of English Consonants and Vowels*. Journal of the Acoustical Society of America, 99(6):3728–3737, 1996.
- [112] STORY, B. H. und I. R. TITZE: *Voice Simulation with a Body-Cover Model of the Vocal Folds*. Journal of the Acoustical Society of America, 97(2):1249–1260, 1995.
- [113] TITZE, I. R.: *The Human Vocal Cords: A Mathematical Model, Part I*. Phonetica, 28:129–170, 1973.
- [114] TITZE, I. R.: *The Human Vocal Cords: A Mathematical Model, Part II*. Phonetica, 29:1–21, 1974.
- [115] TITZE, I. R.: *Parameterization of the Glottal Area, Glottal Flow, and Vocal Fold Contact Area*. Journal of the Acoustical Society of America, 75(2):570–580, 1984.
- [116] TITZE, I. R.: *The Physics of Small-Amplitude Oscillation of the Vocal Folds*. Journal of the Acoustical Society of America, 83(4):1536–1552, 1988.
- [117] TITZE, I. R.: *A Four-Parameter Model of the Glottis and Vocal Fold Contact Area*. Speech Communication, 8:191–201, 1989.
- [118] TITZE, I. R.: *On the Relation between Subglottal Pressure and Fundamental Frequency in Phonation*. Journal of the Acoustical Society of America, 85(2):901–906, 1989.

- [119] TITZE, I. R.: *Physiologic and Acoustic Differences between Male and Female Voices*. Journal of the Acoustical Society of America, 85(4):1699–1707, 1989.
- [120] TITZE, I. R.: *Principles of Voice Production*. Prentice Hall, 1994.
- [121] TRUCKENBRODT, E.: *Lehrbuch der angewandten Fluidmechanik*. Springer, 1988.
- [122] VAN SUMMERS, W.: *Effects of Stress and Final-Consonant Voicing on Vowel Production: Articulatory and Acoustic Analyses*. Journal of the Acoustical Society of America, 82(3):847–863, 1987.
- [123] VARY, P., U. HEUTE und W. HESS: *Digitale Sprachsignalverarbeitung*. B. G. Teubner, Stuttgart, 1998.
- [124] VRIES, M. P. DE, H. K. SCHUTTE, A. E. P. VELDMAN und G. J. VERKERKE: *Glottal Flow through a Two-Mass Model: Comparison of Navier-Stokes Solutions with Simplified Models*. Journal of the Acoustical Society of America, 111(4):1847–1853, 2002.
- [125] VRIES, M. P. DE, H. K. SCHUTTE und G. J. VERKERKE: *Determination of Parameters for Lumped Parameter Models of the Vocal Folds Using a Finite-Element Method Approach*. Journal of the Acoustical Society of America, 106(6):3620–3628, 1999.
- [126] WAKITA, H. und G. FANT: *Toward a Better Vocal Tract Model*. STL-QPSR, 1:9–29, 1978.
- [127] WEIBEL, E. R.: *Morphometry of the Human Lung*. Springer, Berlin, 1963.
- [128] WILHELMS-TRICARICO, R.: *Physiological Modeling of Speech Production: Methods for Modeling Soft-Tissue Articulators*. Journal of the Acoustical Society of America, 97(5):3085–3098, 1995.
- [129] WOOD, S.: *A Radiographic Analysis of Constriction Locations for Vowels*. Journal of Phonetics, 7:25–43, 1979.