Institut für
Technische Informatik und
Kommunikationsnetze

Schätzung des bewegten Vokaltraktes aus dem Sprachsignal

Simon Simonet

Masterarbeit MA-2009-14

Herbstsemester 2009

Institut für Technische Informatik
und Kommunikationsnetze

Betreuer: Dr. B. Pfister und M. Gerber

Verantwortlicher: Prof. Dr. L. Thiele

Zusammenfassung

Die vorliegende Arbeit befasst sich mit der Frage, wie sich die Bewegungen des Vokaltraktes aus einem Sprachsignal schätzen lassen. Die akustisch-artikulatorische Inversion ist aufgrund der Nichtlinearität und Nicht-Eindeutigkeit der Abbildung als schwieriges Problem bekannt. In der Arbeit werden zwei Lösungsansätze vorgestellt, die durch Einführung unterschiedlich starker Einschränkung eine Schätzung ermöglichen. Der Sprechapparat wird dazu auf der akustischen und auf der artikulatorischen Ebene modelliert.

Für die akustische Simulation wird der Sprechapparat durch zwei unterschiedlich aufwändige Rohrsysteme repräsentiert. Im ersten Fall wird der Vokaltrakt zwischen der Stimmritze und den Lippen durch ein unverzweigtes verlustloses Rohrmodell nachgebildet. Die Grundlage des zweiten akustischen Modells ist ein verzweigtes verlustbehaftetes Rohrsystem, bei welchem neben Rachen- und Mundraum auch der Nasenraum und die subglottalen Luftwege berücksichtigt werden. Die Geometrie des Vokaltraktes wird mit Hilfe eines linearen statistischen Artikulatormodells beschrieben. Schliesslich werden zwei an die akustischen Modelle angepasste Schätzmethoden vorgestellt, um vom Sprachsignal zu den artikulatorischen Parametern zu gelangen.

Die Modelle und Verfahren wurden in einem interaktiven Tool integriert, mit welchem sich die Schätzungen durchführen und die Ergebnisse auf verschiedene Arten (statisch und dynamisch) visualisieren lassen.

Eine qualitative Beurteilung der Schätzmethoden anhand kurzer Sprachproben hat ergeben, dass die Berücksichtigung der Energieverluste und Seitenzweige im Sprechtrakt die Schätzergebnisse klar verbessern. Gute Ergebnisse konnten für die Lautklassen Vokale und Frikative erzielt werden.

Abstract

The present thesis deals with the question of how to estimate the articulators' positions and motions from the speech acoustics. The acoustic-to-articulatory mapping is known as a difficult problem due to its non-linear and one-to-many characteristics. The thesis presents two approaches using different constraints, which allow an estimation. The speech apparatus is therefore modeled on the acoustic and articulatory level.

First, a brief overview over the mechanisms of speech production and the involved organs will be given. The acoustics of the vocal-tract are modeled by two tube systems of different complexity. In the first case, the vocal-tract from the glottis to the lips is represented by a lossless unbranched tube model. The second model is based on a lossy branched tube model including the nasal cavity and the subglottal airways. The vocal-tract geometry is described by a statistic articulatory model. Finally, two adapted estimation methods are presented to retrieve the articulatory parameters and trajectories from the speech signal.

The applied models and methods have been integrated into an interactive tool, which allows to perform estimations and the visualisation of the results in different ways (static and dynamic).

A qualitative evaluation of the inversion methods based on real speech data has shown that the estimation results are clearly improved if the losses and side-branches in the vocal-tract are taken into account. Good results were achieved for vowels and fricatives.

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

Sprache ist der gebräuchlichste Weg der zwischenmenschlichen Kommunikation. Als Kommunikationsprozess dient sie der Übermittlung von Informationen. Diese Informationen sind als akustisches Signal kodiert, an dessen Entstehung immer der menschliche Sprechapparat, mit dem Vokaltrakt und den beweglichen Artikulatoren, beteiligt ist.

Die artikulatorische Repräsentation der Sprache, als Beschreibung des Sprachprozesses in einem früheren Stadium, bietet gegenüber der akustischen Beschreibung des Sprachsignals einige interessante Vorteile. Die Artikulatoren ändern ihre Position und ihre Bewegungen nur langsam im Vergleich zu den Änderungen des resultierenden Sprachsignals. Eine Beschreibung des Sprachprozesses auf der artikulatorischen Ebene würde daher eine erhebliche Datenreduktion durch Verringerung redundanter Informationen erlauben.

Abgesehen vom offensichtlichen Nutzen im Bereich der Sprachkodierung sind zahlreiche andere Anwendungsgebiete denkbar, die von einer artikulatorischen Beschreibung profitieren könnten. Eine der grossen Schwierigkeiten bei der Analyse und Verarbeitung von Sprache besteht in der hohen Variabilität der Sprachsignale. Die Ursachen der Variabilität sind vielfältiger Natur. Sie können auf äusserliche Einflüsse, wie z.B. Umgebungsgeräusche, und auf Variationen, die durch den Sprecher selbst bedingt sind, zurückgeführt werden. Die Variabilität aufgrund äusserer Einflüsse kann durch eine geeignete Vorverarbeitung reduziert werden. Die durch den Sprecher bedingten Variationen bleiben jedoch bestehen. Hier erhofft man sich durch die Analyse im artikulatorischen Raum eine Verringerung der Variabilität basierend auf der Beobachtung, dass für die Erzeugung der Laute einer Sprache verschiedene Sprecher sehr ähnliche artikulatorische Bewegungen vollführen. Die verbleibende Variabilität rührt dann vor allem von der Koartikulation und der Prosodie her.

Auch für das Sprachtraining von Menschen mit einer Hör- oder Sprechbehinderung würde eine artikulatorische Beschreibung des Sprachprozesses in Verbindung mit einer geeigneten Visualisierung hilfreich sein.

Trotz allen Vorteilen hat die artikulatorische Repräsentation der Sprache bis heute keine praktische Bedeutung erlangt. Der Hauptgrund dafür liegt darin, dass eine Bestimmung der Vokaltraktform und der Bewegung der Artikulatoren sich sehr schwierig gestaltet. Obschon grosse Fortschritte bei den bildgebenden Verfahren der Medizin gemacht wurden, ist diese Art der Datengewinnung immer noch mit einem erheblichen Aufwand und Kosten verbunden und nicht alltagstauglich. Aus diesem Grunde wäre ein Verfahren, welches die Vokaltraktform direkt aus dem Sprachsignal schätzen kann, wünschenswert.

Das Problem, welches sich bei der Schätzung stellt, wird als inverses Problem bzw. als Inversionsproblem bezeichnet. Gegeben ist eine akustische Beschreibung des Sprachsignals (z.B. eine extrahierte Merkmalssequenz) und gesucht wird die äquivalente artikulatorische Repräsentation. Die Schwierigkeit bei der inversen Transformation vom akustischen in den artikulatorischen Raum besteht in der Nichtlinearität und der Surjektivität der Abbildung. FLANAGAN [13] und ATAL *et al.* [2] haben gezeigt, dass unterschiedliche Konfigurationen des Vokaltraktes und der Artikulatoren zum selben akustischen Ergebnis führen können. Diese Tatsache wissen beispielsweise Bauchredner für ihre Kunst sehr gut einzusetzen.

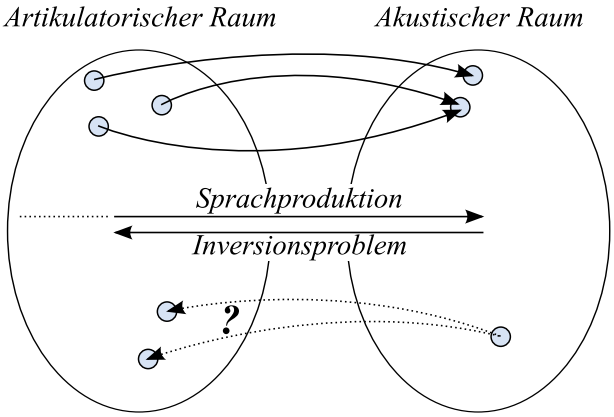


Abbildung 1.1: *Artikulatorisch-akustische Transformation und Inversionsproblem.*

1.2 Überblick der Verfahren zur artikulatorischen Rekonstruktion anhand akustischer Daten

Die Idee der inversen Transformation vom akustischen in den artikulatorischen Raum ist nicht neu. Seit mehr als 30 Jahren sind dazu diverse Lösungsansätze erarbeitet worden, wobei das inverse Problem als Ganzes immer noch als ungelöst gilt. Bis heute ist keine befriedigende Lösung gefunden worden, die für alle Lautklassen praktisch brauchbare Ergebnisse liefert.

KRSTULOVIĆ [20] nimmt eine Einteilung der verschiedenen Ansätze in fünf Gruppen vor:

- **Analytische Ansätze:** Die akustisch-artikulatorische Beziehung wird explizit durch einen analytischen Ausdruck angegeben. Möglich wird dies, indem das inverse Problem durch die Vorgabe von bestimmten Nebenbedingungen eingegrenzt wird, so dass eine eindeutige Lösung vorhanden ist.

  MERMELSTEIN und SCHROEDER [31] haben ein Verfahren vorgestellt, bei welchem die Parameter einer Fourierreihen-Entwicklung der logarithmierte Flächenfunktion mit den Formantenfrequenzen in Verbindung gesetzt werden können.

  Eine ähnliche Methode wurde von SHIRAI und HONDA [42] entwickelt. Für die Modellierung der nichtlinearen Abbildung zwischen Formantfrequenzen und den artikulatorischen Parametern nutzen sie Polynome 3. Ordnung und Kalman-Filter, um eine eindeutige Schätzung der Artikulatorenbewegungen zu erhalten.

  Ein neues Verfahren zur Schätzung der Vokaltrakflächen basierend auf der inversen Filterung des Sprachsignals wurde von WAKITA [50] vorgeschlagen. Er hat einen direkten Zusammenhang zwischen dem Modell der inversen Filterung und einem einfachen akustischen Rohrmodell aufgezeigt und macht sich bei der Schätzung der Modellparameter bekannte Verfahren (LPC-Analyse) zu Nutze.

- **Stochastische Ansätze:** Der Lösungsraum wird mit Hilfe von kontinuierlichen Wahrscheinlichkeitsverteilungen beschrieben, die eine effiziente Modellierung der surjektiven Abbildung ermöglichen. Ein grosser Vorteil der stochastischen Verfahren liegt darin, dass die Modelle gewisse Effekte auch ohne Vorwissen direkt aus Trainingsdaten erlernen können. Voraussetzung dafür ist jedoch, dass genügend Trainingsdaten zur Verfügung stehen.

Die verschiedenen Ansätze orientieren sich hierbei stark an Verfahren, wie sie auch bei der statistischen Spracherkennung eingesetzt werden. RAMSAY und DENG [38] nutzen Markov-Ketten zur statistischen Beschreibung der Zustände der Artikulatoren.

- **Neuronale Netzwerke:** Die neuronalen Netze werden trainiert, um die nichtlineare Beziehung zwischen den artikulatorischen und akustischen Parametern zu approximieren. Die Trainingsdaten können von einem artikulatorischen Synthesemodell stammen (SHIRAI und KOBAYASHI [43]) oder aus Messdaten von bildgebenden Verfahren gewonnen werden. (PAPCUN *et al.* [36]).

- **Ansätze mit Codebüchern:** Die Erstellung eines Codebuches umfasst die Quantisierung des artikulatorischen und des akustischen Raums und die Generierung einer Liste mit zueinander gehörenden Vektorpaaren. Das Lösen des inversen Problems besteht in der Durchsuchung des Codebuches nach dem optimalen Vektorpaar, wobei Einschränkungen den möglichen Suchraum eingrenzen können. Ein Verfahren zur inversen Transformation basierend auf einem Codebuch mit knapp über 30'000 Vektorpaaren wurde von ATAL *et al.* [2] entwickelt. Der artikulatorische Raum wurde hierfür gleichmässig abgetastet und die akustischen Merkmalsvektoren (Frequenzen und Bandbreiten der ersten fünf Formanten) mit Hilfe eines artikulatorischen Synthesemodells berechnet. Einen ähnlichen Weg sind LARAR *et al.* [23] gegangen, mit dem Unterschied, dass sie den artikulatorischen Raum nur zwischen vorgegebenen Stammformen des Vokaltraktes abgetastet haben. Ihre Absicht war, unrealistische Vokaltraktformen von Anfang an auszuschliessen.

- **Ansätze basierend auf Optimierungsverfahren:** Anstatt im Voraus Codebücher zu generieren kann ein passendes Vektorpaar auch mit Hilfe eines Synthesemodells und einem allgemeinen Optimierungsverfahren gefunden werden. Die artikulatorischen Parameter des Synthesemodells werden so optimiert, dass ein akustisches Abstandsmass zwischen dem synthetisierten und dem vorgegebenen Signal minimiert wird. Zusätzliche Bedingungen können genutzt werden, um beispielsweise geeignete Startvektoren für den Optimierungsvorgang zu erhalten. Ein Verfahren dieser Art wurde von SOROKIN und TRUSHKIN [47] entwickelt. Schwierigkeiten bereitet hierbei vor allem der Zeitaufwand für den Optimierungsvorgang, welcher mit zunehmender Komplexität des Synthesemodells stark ansteigt.

1.3 Aufbau der Arbeit

Die vorliegende Arbeit ist wie folgt gegliedert:

- Kapitel 2 gibt einen Überblick über die Physiologie des menschlichen Sprechapparates und den Mechanismus der Sprachproduktion. Für die Inversion findet eine Modellierung des Sprechapparates auf zwei Ebenen statt:

- In Kapitel 3 werden zwei unterschiedlich komplexe Modelle zur akustischen Simulation des Sprechtraktes vorgestellt. Für Analysen im Frequenzbereich werden Verfahren zur Berechnung der Übertragungsfunktionen erläutert.

- Kapitel 4 beschäftigt sich mit der geometrischen Modellierung des Sprechtraktes. Ein artikulatorisches Modell und die Verknüpfung zwischen akustischer und artikulatorischer Ebene werden vorgestellt.
- In Kapitel 5 werden an die akustischen Modelle angepasste Verfahren behandelt, um die Parameter des artikulatorischen Modells aus einem Sprachsignal zu schätzen.
- Eine Evaluation der Schätzmethoden erfolgt in Kapitel 6 anhand von kurzen Sprachproben.
- In Kapitel 7 werden die erzielten Ergebnisse zusammengefasst und Möglichkeiten zu weiteren Verbesserungen diskutiert.

2 Menschliche Spracherzeugung

2.1 Physiologie

Sprache ist das akustische Ergebnis der Bewegungen des menschlichen Sprachapparates. Zum Sprachapparat gehören die Lunge, die Bronchien, die Luftröhre, die Stimmlippen, der Rachen-, Mund- und Nasenraum. Die Stimmlippen bzw. die dadurch gebildete Stimmritze unterteilen das gesamte System in einen subglottalen und supraglottalen Bereich. Das supraglottale System wird häufig auch als Vokal- oder Sprechtrakt bezeichnet.

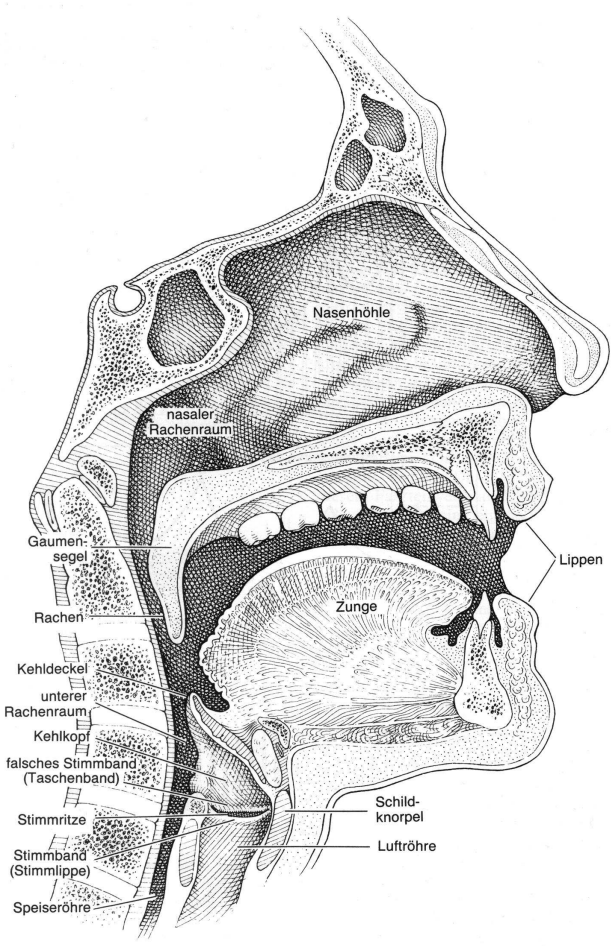


Abbildung 2.1: *Mediodsagittal-Ansicht des Vokaltraktes (Quelle: [14]).*

2.1.1 Subglottale Luftwege

Am Eingang des subglottalen Bereichs befindet sich die Luftröhre (Trachea). Sie ist eine etwa 9-14 cm lange Röhre, die von 16-20 hufeisenförmigen Knorpelspangen versteift wird, so dass sie immer offen bleibt. Die Luftröhre spaltet sich hinter dem Brustbein in den linken und rechten Hauptbronchius, die sich einige Zentimeter weiter in die Lappenbronchien verzweigen. Diese Verästelung setzt sich bis zu den Bronchiolen fort, wobei der Durchmesser von ungefähr 20 mm am Anfang der Luftröhre auf weniger als 0.5 mm abnimmt. Auch die Bronchiolen teilen sich

noch einige Male bis sie in den Lungenbläschen (Alveolen) enden, wo der Gasaustausch stattfindet. Die subglottalen Luftwege entsprechen daher, wenn man einen symmetrischen Aufbau annimmt, einem vollständigen Binärbaum mit 2^n Zweigen auf der Höhe n.

WEIBEL [52] unterteilt die subglottalen Luftwege basierend auf einem solchen Binärbaum in 2 unterschiedliche Zonen mit insgesamt 24 Teilungsstufen (siehe Abbildung 2.2). Die Leitungszone reicht vom Anfang der Luftröhre bis zu den Terminalbronchiolen und hat eine Pfadlänge von ungefähr 25 cm (bei 150 ml Volumen). Im Anschluss kommt eine kurze Übergangszone, die jedoch bereits ein Volumen von 1500 ml aufweist, gefolgt von der eigentlichen Respirationszone mit den Lungenbläschen und einem Volumen von 3150 ml. Die Grössenangaben des Modells entsprechen den Werten eines durchschnittlichen Erwachsenen bei normaler Atmung. Die gesamte Querschnittsfläche auf einer Verzweigungsstufe als Funktion vom Abstand zum oberen Luftröhreneingang ist in Abbildung 2.3 dargestellt. Das elastische Lungengewebe zieht sich bei Erschlaffung der Atemmuskulatur zusammen und führt zu einem von der Lunge kommenden (pulmonalen) Luftstrom.

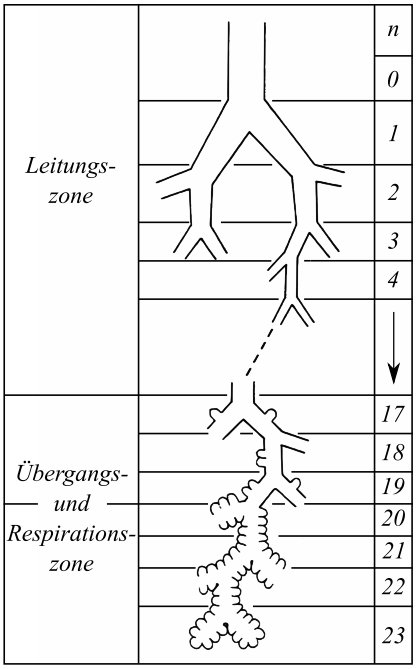


Abbildung 2.2: *Schematische Ansicht der Baumstruktur der subglottalen Luftwege nach dem Modell von* WEIBEL *[52].*

2.1.2 Kehlkopf und Stimmlippen

Der Kehlkopf (Larynx) ist das Verbindungslement zwischen Luftröhre und Rachenraum. Seine Hauptfunktionen bestehen darin, die Luftröhre beim Schlucken von Speisestücken und Flüssigkeiten zu schützen und den pulmonalen Luftstrom mit den Stimmlippen zu regulieren. Der Kehlkopf besteht aus mehreren Knorpelelementen, die durch Bänder zusammengehalten werden und so ein äusseres Gerüst formen. Der innere Aufbau des Kehlkopfes wird aufgrund der Engstellen in drei Stockwerke gegliedert.

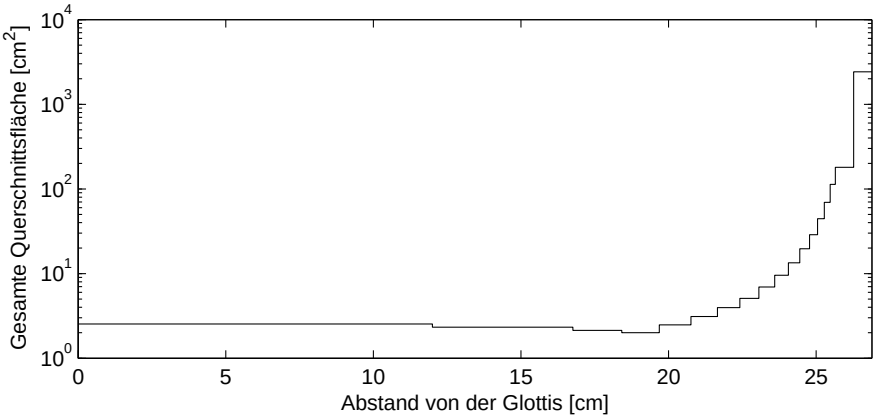


Abbildung 2.3: *Querschnittsfunktion der subglottalen Luftwege des Modells von* WEIBEL *[52].*

Im unteren Übergangsbereich von der Luftröhre zu den Stimmlippen verengt sich der Innenraum konisch bis zu den Stimmlippen. Die Wand ist mit einer Schleimhaut ausgekleidet, welche bei Kontakt mit Fremdkörpern (z.B. beim Verschlucken) zu starkem Hustenreiz führt. Die Stimmbandmuskeln regulieren die Spannung der Stimmlippen, welche die Stimmritze (Glottis) auf beiden Seiten begrenzen. Die Öffnungsfläche der Stimmritze wird zudem durch die Position der hinteren beweglichen Stellknorpel bestimmt, mit denen die Stimmlippen verbunden sind. Die Stellknorpel sind in der Mitte durch Zwischenknorpelmuskeln miteinander verbunden und zusätzlich seitlich durch einen weiteren Muskel mit dem äusseren Ringknorpel. Die Kontraktion der mittleren Muskulatur führt die Stimmlippen näher aneinander heran, während die äusseren Muskeln sie auseinander ziehen, so dass ein grösserer Luftstrom passieren kann.

Der mittlere Bereich oberhalb der Stimmlippen weist auf beiden Seiten kleine Ausbuchtungen auf, die die Stimmlippen von den Taschenfalten (oft als falsche Stimmlippen bezeichnet) trennen. Ihre Ausdehnung und Form ist variabel. Der genaue Einfluss dieser Ausbuchtungen auf die Stimmbildung ist derzeit noch unklar. Die Taschenfalten weisen eine ähnliche Form wie die Stimmlippen auf, sind normalerweise jedoch nicht an der Phonation beteiligt.

Das obere Stockwerk reicht von der Engstelle der Taschenfalten bis zum Kehlkopfeingang mit dem Kehlkopfdeckel zum Verschliessen des Einganges.

2.1.3 Rachenraum

Der Rachen (Pharynx) ist ein etwa 12-15 cm langer muskulärer Schlauch, welcher sich vom Kehlkopf bis zur Schädelbasis hinter Mund- und Nasenhöhle erstreckt. In ihm überkreuzen sich die Atem- und Speisewege. Die Rachenrückwand ist wenig beweglich. Die Form der vorderen Wand wird insbesondere durch das Zungenbein bestimmt, an welchem der Kehlkopf durch eine Membran befestigt ist. Durch eine Vielzahl von Muskeln, die an dem Zungenbein festgemacht sind, wird seine Position bestimmt. Beim Öffnen des Mundes (Herabziehen des Kiefers) wird das Zungenbein durch das Zusammenspiel mehrerer Muskelgruppen nach hinten und nach oben gezogen und führt deshalb zu einer Verengung im Rachenraum. Beim Schlucken oder wenn die Zunge bei der Artikulation sich weit nach vorne bewegt, wird das Zungenbein nach vorne und nach oben gezogen. Im Rachenraum führt diese Bewegung zu einer Vergrösserung der Querschnittsfläche.

Etwa 2 cm oberhalb des Kehlkopfes auf Höhe des Kehlkopfeinganges befinden sich zwei Schleimhautausbuchtungen, die als Schluckrinne oder Sinus piriformis bezeichnet werden. Die Ausbuchtungen sind etwa 1.6 bis 2 cm tief und weisen eine kegelförmige Form auf. Ihr Volumen beträgt zwischen 2 und 3 cm^3. DANG und HONDA [7] konnten in ihren Experimenten zeigen, dass bei offenen Sinus piriformis im Bereich von 4 bis 5 kHz spektrale Minima zu beobachten sind und zudem eine signifikante Verminderung der Formantfrequenzen bei Vokalen feststellbar ist. In Abbildung 2.4 sind 2 dreidimensionale Rekonstruktionen des Sprechtraktes während der Artikulation eines [i] bzw. eines [a] zu sehen. Diese Aufnahmen sind mit Hilfe eines bildgebenden Verfahrens (Computertomographie) im Rahmen von Messungen der Querschnittsfunktion des Sprechtraktes entstanden [48]. Die Sinus piriformis sind die taschenförmigen Gebilde beidseits des Kehlkopfes. Ob die Hohlräume der Sinus piriformis auf dem rechten Bild tatsächlich abgekoppelt sind, oder ob es sich dabei um ein Bewegungsartefakt während der Aufnahmen handelt, ist nicht klar.

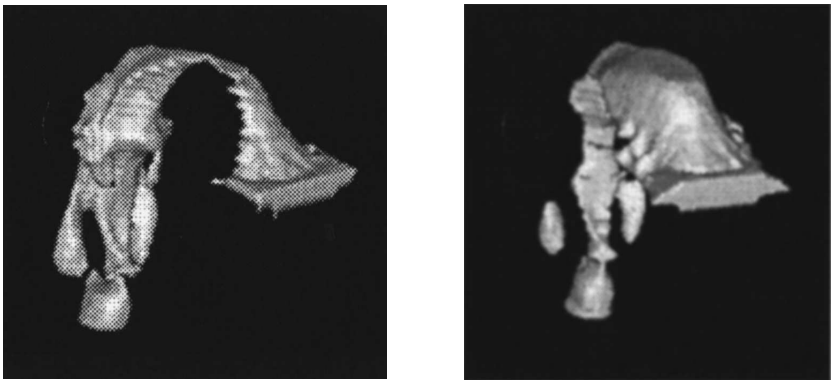


Abbildung 2.4: *Dreidimensionale Rekonstruktion (CT-Aufnahmen) des Sprechtraktes mit den Sinus piriformis. Links während der Artikulation eines [i] und rechts während eines [a] (Quelle: [48]).*

Zwischen Mund- und Nasenrachenraum liegt der bewegliche weiche Gaumen (Velum). In Ruhestellung hängt das Gaumensegel schräg Richtung Zungenwurzel herab. Beim Schlucken und bei der Artikulation nicht nasaler Laute wird der weiche Gaumen bis an die Rachenrückwand gehoben und trennt den Nasenraum vom Mund-Rachen-Raum (velopharyngealer Verschluss).

2.1.4 Mundhöhle

Die Mundhöhle wird nach vorne durch die Lippen, nach oben durch den Gaumen, seitlich durch die Wangen und nach unten durch die Zunge und den Mundboden begrenzt. Oft wird zwischen Mundvorhof, bestehend aus dem Zwischenraum zwischen den Lippen, den Wangen und den Zahnreihen, und der eigentlichen Mundhöhle innerhalb der Zahnreihen unterschieden.

Der vordere harte Gaumen entlang des Oberkieferknochens trennt die Mundhöhle von der Nasenhöhle. Der Unterkiefer kann durch die Kiefermuskulatur gehoben und gesenkt werden, nach vorne und nach hinten verlagert werden und seitlich verschoben werden. Im Zusammenhang mit der Lautbildung ist vor allem das Heben und Senken entscheidend.

Die Lippen, unterteilt in Ober- und Unterlippe, bestehen zu einem grossen Teil aus Muskel-

und Bindegewebe und sind daher sehr beweglich. Bei der Sprachproduktion dienen die Lippen zur Formung der Öffnungsfläche am Ausgang des Sprechapparates. Zudem variiert die Länge des Sprechtraktes durch Vorstülpen oder Spreizen der Lippen.

Auf dem Mundboden liegt die Zunge auf, die bei geschlossenem Kiefer fast die ganze Mundhöhle ausfüllt. Sie ist im Wesentlichen aus Muskelfasern aufgebaut und mit einer Schleimhaut überzogen, in der sich die Geschmacksknopsen befinden. Die Zunge wird grob in Zungenspitze, Zungenrücken und Zungenwurzel unterteilt. Die Zungenwurzel im Rachenraum ist durch Muskeln fest mit dem Zungenbein verbunden, weshalb die Zunge indirekt auch in Verbindung mit dem Kehlkopf steht. Der vordere Bereich der Zunge in der Mundhöhle ist dank der Anordnung der Muskelfasern äusserst beweglich und ein unverzichtbarer Bestandteil bei der Lautbildung.

2.1.5 Nasenhöhle

Die Nasenhöhle ist der obere Teil der Atemwege und verbindet sie durch die Nasenlöcher mit der Aussenwelt. Die Nasenhöhle wird durch die Nasenscheidewand in zwei ungefähr gleich grosse Kammern getrennt. Im hinteren Nasenrachenraum vereinen sich die zwei getrennten Hohlräume wieder. Die beiden Hohlräume zu jeder Seite werden durch die drei übereinander liegenden Nasenmuscheln gestützt und unterteilt. Zwischen ihnen liegen die Nasengänge. Der unterste Nasengang zwischen Gaumen und der ersten Nasenmuschel wird als Atemgang bezeichnet, da der Luftaustausch über ihn erfolgt. Am mittleren Nasengang zwischen den unteren zwei Nasenmuscheln sind die Nasennebenhöhlen angeschlossen. Im hinteren Teil des oberen Nasenganges sitzt das Geruchsorgan.

Bei den Nasennebenhöhlen handelt es sich um Hohlräume bestimmter Schädelknochen. Sie sind mit einer Schleimhaut überzogen und mit Luft gefüllt. Durch kleine Zugänge sind sie mit der Nasenhöhle verbunden.

Beim Menschen unterscheidet man 4 Nasennebenhöhlen (Abbildung 2.5): Die Kieferhöhle (Sinus maxillaris), die Stirnhöhle (Sinus frontalis), die Keilbeinhöhle (Sinus sphenoidalis) und die Siebbeinzellen (Cellulae ethmoidales). Die ersten 3 weisen relativ grosse zusammenhängende Hohlräume auf, wohingegen die Siebbeinzellen aus einer grösseren Anzahl kleiner miteinander verbundener Hohlräume bestehen.

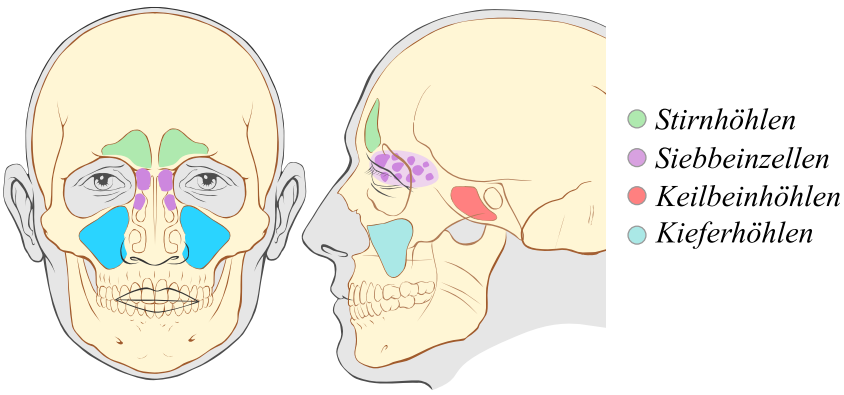


Abbildung 2.5: *Schematische Darstellung der Nasennebenhöhlen (nach: [25]).*

DANG und HONDA [9] haben die Nasenhöhlen verschiedener Probanden vermessen. Die Gesamtlänge der Nasenhöhlen lag bei allen im Bereich zwischen 11.4 und 11.7 cm. Beim Umfang und bei der Querschnittsfläche der Nasenhöhlen zeigte sich zum Teil eine ausgeprägte Asymmetrie zwischen der rechten und der linken Seite der Nasenhöhle. Auch waren beträchtliche Unterschiede dieser Werte von einer Versuchsperson zur nächsten festzustellen. So lag das Volumen der kleinsten Nasenhöhle bei 15.4 cm^3 und das der grössten bei 39.7 cm^3, wobei die unterschiedlichen Werte vor allem auf den mittleren Teil der Nasenhöhlen zurückzuführen sind. Die in der vorliegenden Arbeit verwendeten festen geometrischen Grössen des Nasenraums (Flächenfunktion und Umfangsfunktion) richten sich nach den Messergebnissen der Arbeit von DANG und HONDA (Proband 1).

Ähnliche Beobachtungen konnten in einer weiteren Arbeit von DANG und HONDA [6] auch bei den geometrischen Grössen der Nasennebenhöhlen gemacht werden. Es wurden zudem die akustischen Eigenschaften der Nasennebenhöhlen untersucht und festgestellt, dass sie als akustische Resonatoren bei der Produktion von nasalierten Lauten wirken und durch Helmholtz-Resonatoren modelliert werden können. In Tabelle 2.1 sind die Daten des Modells von DANG und HONDA aufgeführt, die auch in dieser Arbeit verwendet werden.

	Keilbeinhöhle		Kieferhöhle	Stirnhöhle
	grosse	kleine		
Volumen [cm^3]	11.3	6.8	33	6.2
Halslänge [cm]	0.3	0.3	0.45	1
Halsquerschnitt [cm^2]	0.185	0.185	0.145	0.11
Abstand der Öffnung von den Nasenlöchern [cm]	6.2	6.2	5.1	4.3
Resonanzfrequenz [Hz]	1305	1682	552	749
Bandbreite [Hz]	188	216	108	124

Tabelle 2.1: *Daten der Nasennebenhöhlen des Modells von* DANG *und* HONDA *[6].*

Die Keilbeinhöhle wird durch eine Scheidewand in zwei unterschiedlich grosse Kammern unterteilt, die einzeln modelliert werden. Aufgrund der feinen Unterteilung der Siebbeinzellen wird ihr Einfluss auf Frequenzen unterhalb von 3 kHz als gering eingeschätzt. Sie werden daher nicht weiter berücksichtigt.

2.1.6 Geometrie des Vokaltraktes

Die momentane Form des Vokaltraktes ist hauptsächlich von der Position der beschriebenen beweglichen Artikulatoren abhängig. Die dadurch verursachten geometrischen Veränderungen des Sprechtraktes können wie folgt gegliedert werden:

- **Querschnittsflächen:** Während die Querschnittsflächen der subglottalen Bereiche oder des Nasaltraktes sich nur geringfügig ändern, sind im Mundraum Querschnittsflächen zwischen 0 cm^2 bei einem kompletten Verschluss und über 15 cm^2 bei offenen Vokalen möglich. Auch die Querschnittsfläche des hinteren Nasaltraktes variiert mit der Position des Gaumensegels.

- **Vokaltraktlänge:** Die Gesamtlänge des Vokaltraktes wird im Wesentlichen von drei Faktoren beeinflusst. Die Positionierung der Lippen und des Kehlkopfes bestimmt die Endpunkte des Vokaltraktes. Die Gesamtlänge variiert aber auch mit der Zungenposition. Die Mittelline des Vokaltraktes wird aufgrund der Biegung länger je näher die Zunge an die Gaumenwand oder an die Rachenrückwand geführt wird.

- **Seitenzweige:** Der Nasaltrakt wird durch das Gaumensegel als Seitenzweig an- bzw. abgekoppelt. Auch im Rachenraum sind mit den Sinus piriformis zwei Seitenzweige vorhanden.

2.2 Mechanismus der Spracherzeugung

Nach dem klassischen Quelle-Filter-Modell unterscheidet man bei der Spracherzeugung zwischen Schallproduktion und Klangformung. Der pulmonale Luftstrom liefert die nötige aerodynamische Energie für die Schallquellen. Die Klangformung erfolgt durch die akustische Filterung des Quellensignals im Sprechtrakt. Es werden hauptsächlich drei Anregungsarten unterschieden:

- **Stimmhafte Anregung:** Durch einen erhöhten subglottalen Druck werden die Stimmlippen von unten her auseinander gedrückt bis eine kleine durchgehende Öffnung entsteht. Aufgrund des Druckunterschiedes strömt ein Luftstrom mit hoher Geschwindigkeit durch diese Engstelle. Die hohe Strömungsgeschwindigkeit führt zu einem Druckverlust senkrecht zur Strömungsrichtung, bekannt als Bernoulli-Effekt. In Verbindung mit den elastischen Rückstellkräften der Stimmlippen schliesst sich die Glottis in der Folge von unten her wieder. Dieser Vorgang wiederholt sich quasi-periodisch und moduliert die glottale Öffnungsfläche und den passierenden Luftstrom zeitlich. Die resultierenden Luftstromimpulse am Ausgang führen dann zur akustischen Anregung des Sprechtraktes. Diese Anregungsart wird als stimmhaft bzw. als Phonation bezeichnet.

- **Stimmlose Anregung:** Die stimmlose Anregung entsteht, wenn ein ausreichend hoher Luftstrom eine Engstelle im Vokaltrakt passiert. Durch die Engstelle erhöht sich die Strömungsgeschwindigkeit soweit, dass am Ausgang der Konstriktion keine laminare Strömung mehr vorherrscht. Stattdessen vermischt sich der austretende Luftstrom unter Verwirbelung mit der umgebenden Luft. Diese Verwirbelungen führen zu unregelmässigen Geschwindigkeitsschwankungen, die sich als Schallsignal in beide Richtungen der Sprechtraktes ausbreiten. Beim Auftreffen des Lufstrahls auf ein Hindernis stromabwärts kommt es zu zufälligen Druckschwankungen die sich ebenfalls als Schallsignal ausbreiten. Je nach Lage der Engstelle weisen die Schallquellen ein unterschiedliches Spektrum auf (siehe Abschnitt 3.4).

- **Plosive Anregung:** Die plosive Anregung entsteht, indem im Vokaltrakt für eine kurze Zeitspanne (40-100 ms) ein vollständiger Verschluss gebildet wird. Hinter diesem wird ein Überdruck aufgebaut, bis durch plötzliches Öffnen des Verschlusses eine impulsartige transiente Anregung entsteht. Anders als die stimmhafte oder stimmlose Anregung weist die plosive Anregung aufgrund des instationären Verhaltens kein typisches Quellenspektrum auf.

In Abhängigkeit von der Art der Anregung und der Form des Vokaltraktes unterscheidet man die verschiedene Lautklassen:

- **Vokale:** Die Vokale zeichnen sich dadurch aus, dass der Vokaltrakt durchgängig offen ist und stimmhaft angeregt wird. Je nachdem ob ein velopharyngealer Verschluss besteht oder nicht, können die Vokale weiter in nicht-nasalierte und nasalierte Vokale unterteilt werden.

- **Konsonanten:** Allen Konsonanten ist gemeinsam, dass im Vokaltrakt an einer bestimmten Stelle eine Verengung besteht, die den Luftstrom teilweise oder ganz blockiert (vgl. Abbildung 2.6).

  Frikative: Frikative entstehen durch eine starke Verengung im Vokaltrakt. Je nach Laut wird die Konstriktion durch unterschiedliche Artikulatoren und an verschiedenen Stellen im Vokaltrakt gebildet. Sie können stimmlos oder auch in Kombination mit der stimmhaften Anregung durch die Glottis erzeugt werden. Die Artikulationsorte sind: [f,v] labiodental, [s,z] alveolar, [ʃ, ʒ] postalveolar, [ç, j] palatal, [x] velar und [h] glottal.

  Plosive: An einer Stelle im Vokaltrakt wird ein vollständiger Verschluss gebildet. Der Nasenraum muss durch das Gaumensegel verschlossen sein, damit der Druckaufbau erfolgen kann. Bei stimmhaften Plosiven schwingen die Stimmlippen auch noch während der Verschlussphase: [b,p] bilabial, [d,t] alveolar, [ɡ,k] velar und [ʔ] glottal.

  Laterale und Vibranten: Die Laute dieser Gruppen werden stimmhaft angeregt. Der Nasaltrakt ist vom Mund-Rachen-Raum abgetrennt. Beim [l] berührt die Zunge den Zahndamm (alveolar), dabei sind die Zungenseiten jedoch abgesenkt, so dass es weder zu einem vollständigen Verschluss noch zu einer Konstriktion mit rauschhafter Anregung kommt. [l] wird daher als lateraler Laut bezeichnet. [r,ʀ] zählen zu den Vibranten, bei denen ein flexibles Organ wiederholt gegen ein anderes schlägt. Das [r] wird alveolar mit der Zungenspitze gebildet, beim [ʀ] vibriert das Gaumenzäpfchen (uvular). Wie bei den Stimmlippen ist auch in diesem Fall der Bernoulli-Effekt für die Schwingungen mitverantwortlich.

  Nasale: Die Nasale sind durch eine stimmhafte Anregung und einem vollständigen Verschluss im Mundraum gekennzeichnet. Die velopharyngeale Pforte ist offen, so dass sich der Schall im Nasaltrakt ausbreiten kann und über die Nasenlöcher abgestrahlt wird. Das Spektrum der Nasale ist durch Antiresonanzen geprägt, die durch die Seitenzweige der Mundhöhle und der Nasennebenhöhlen entstehen: [m] bilabial, [n] alveolar und [ŋ] velar.

Daneben gibt es noch Kombination aus diesen Vokalen und Konsonanten. So zeichnen sich die Diphtonge [a͡i, a͡u, o͡i] durch die Bewegung der Artikulatoren von einer Position zur nächsten aus. Sie werden als eigenständige Phoneme angesehen. Affrikate sind Kombinationen aus einem Plosiv und einem Frikativ [p͡f, t͡s,t͡ʃ].

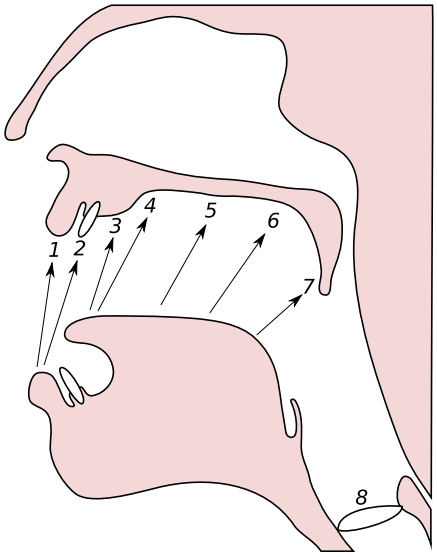


Abbildung 2.6: *Schematische Darstellung der Artikulationsorte: 1. bilabial [b,ɒ,m], 2. labiodental [f,v], 3. alveolar [s,z,d,t,l,r,n], 4. postalveolar [ʃ, ʒ], 5. palatal [ç, j], 6. velar [x,g,k,ŋ], 7. uvular [ʀ] und 8. glottal [h] (nach: [33]).*

3 Akustische Modellierung des Vokaltraktes

Die entscheidenden Grundlagen zur akustischen Modellierung des Sprechapparates wurden von FANT [10] mit der akustischen Theorie der Sprachproduktion gelegt. Sie beschreibt die Spracherzeugung als ein zweistufiger Prozess, der aus den Komponenten Schallproduktion und Klangformung besteht (bekannt als Quelle-Filter-Modell). Das Sprachsignal wird darin unter Anwendung der Theorie der Wellenausbreitung in einer ersten Approximation durch ebene Wellen beschrieben, die sich im Vokaltrakt ausbreiten und an den Lippen abgestrahlt werden. Der Sprechapparat wird als ein (verzweigtes) Rohrsystem modelliert.

In diesem Kapitel werden zwei akustische Modelle des Sprechapparates vorgestellt. Beide basieren auf den selben Grundgleichungen, die durch Annahmen aber unterschiedlich stark vereinfacht werden.

3.1 Akustische Grundlagen

Die theoretischen Grundlagen der Wellenausbreitung im Vokaltrakt basieren auf den allgemeinen Gleichungen der Strömungslehre. Diese drücken die physikalischen Gesetze von der Erhaltung des Impulses (Bewegungsgleichung), der Erhaltung der Masse (Kontinuitätsgleichung) und der Erhaltung der Energie (Zustandsgleichung) aus [24]. Aus diesen Gesetzen lassen sich die drei Grundgleichungen in differenzieller Form zur Beschreibung der Schallausbreitung im Sprechtrakt herleiten. Eine vollständige Beschreibung der Schallausbreitung umfasst auch Effekte wie

- Verluste durch Schallabstrahlung an den Lippen und Nasenlöchern
- Verluste durch Reibung an den Rohrwänden
- Verluste durch elastische Rohrwände
- Verluste durch Wärmeleitung
- Verluste und Schallerzeugung an Konstriktionen
- Zeitliche Variation der Vokaltraktform

Da diese Gleichungen und deren Lösungen unter Berücksichtigung aller im Sprechtrakt auftretenden Effekte äusserst kompliziert sind, werden sie für gewöhnlich durch Annahmen weiter vereinfacht:

- Die wichtigste Einschränkung wird durch die Annahme einer ebenen Wellenausbreitung gemacht. Die Ausbreitung soll nur entlang einer Achse (im Weiteren ist dies die x-Achse) erfolgen. Für Frequenzen unterhalb von 4 kHz ist diese Annahme gut erfüllt. Mit zunehmender Frequenz werden die Wellenlängen immer kürzer, so dass sich bei grossen Querschnittsflächen im Vokaltrakt auch Querwellen ausbilden können.
- Der Sprechapparat wird für die akustische Simulation durch ein Rohrsystem repräsentiert, welches in kurze kreisrunde Abschnitte unterteilt wird. Die Querschnittsflächen

sind abschnittsweise konstant. Damit der Fehler durch die Diskretisierung vernachlässigbar bleibt, muss für die Länge Δl der Rohrabschnitte gelten: $\Delta l \ll \frac{c}{F_{max}}$. F_{max} stellt die obere Grenze des untersuchten Frequenzbereiches dar und c entspricht der Schallgeschwindigkeit.

- Die Krümmung des Vokaltraktes wird vernachlässigt. Gemäss SONDHI [45] ist diese Vereinfachung gerechtfertigt, da die Formantfrequenzen eines gebogenen und eines geraden Rohres unterhalb von 4 kHz um wenige Prozent voneinander abweichen.

Zur Beschreibung der Geometrie genügt deshalb eine Funktion $A(x)$, welche die Querschnittsfläche als Funktion vom Abstand von der Glottis angibt.

Die in den Gleichungen relevanten physikalischen Grundgrössen sind der Druck $p(x,t)$, die Dichte des Mediums $\rho(x,t)$, die mittlere Geschwindigkeit der Teilchen des Mediums $v(x,t)$ und der Schallfluss $u(x,t) = v(x,t) \cdot A(x)$.

Die im Vokaltrakt durch die Sprachproduktion entstehenden Abweichungen des Luftdruckes und der Luftdichte vom atmosphärischen Druck und von der mittleren Dichte sind sehr gering. Schallereignisse treten dann auf, wenn die Druckverteilung nicht gleichmässig ist und es zu einem Ausgleich kommt. Daher sind in erster Linie Druckdifferenzen entscheidend, weshalb der absolute Druck und die absolute Dichte in einen konstanten und einen zeitlich variierenden Term getrennt werden:

$$\rho_a(x,t) = \rho_0 + \rho(x,t) \quad \text{wobei} \quad \rho(x,t) \ll \rho_0, \tag{3.1}$$

$$p_a(x,t) = p_0 + p(x,t) \quad \text{wobei} \quad p(x,t) \ll p_0. \tag{3.2}$$

3.1.1 Grundgleichungen

Die Bewegungsgleichung

Der Impulserhaltungssatz besagt, dass der Gesamtimpuls in einem abgeschlossenen System konstant ist. Die Übertragung des Gesetzes auf ein infinitesimal kleines Luftvolumen unter Anwendung der genannten Vereinfachungen führt zur sog. Euler'schen Bewegungsgleichung:

$$-\frac{\partial p_a}{\partial x} = \rho_a \cdot v \cdot \frac{\partial v}{\partial x} + \rho_a \cdot \frac{\partial v}{\partial t} + r \cdot v. \tag{3.3}$$

r entspricht dabei einem Strömungswiderstand. Mit dem Druck und der Dichte als Summe gemäss (3.1) und (3.2) ergibt sich:

$$-\frac{\partial (p_0 + p)}{\partial x} = (\rho_0 + \rho) \cdot v \cdot \frac{\partial v}{\partial x} + (\rho_0 + \rho) \cdot \frac{\partial v}{\partial t} + r \cdot v. \tag{3.4}$$

Vernachlässigt man die Terme höherer Ordnung in (3.4) und den Strömungswiderstand und berücksichtigt die Linearisierungen, so erhält man die Bewegungsgleichung für das lineare Schallfeld:

$$-\frac{\partial p}{\partial x} = \rho_0 \cdot \frac{\partial v}{\partial t} = \frac{\rho_0}{A} \cdot \frac{\partial u}{\partial t}. \tag{3.5}$$

Hierbei ist anzumerken, dass die Vernachlässigung des Terms $v \cdot \frac{\partial v}{\partial x}$ sowie des Terms $r \cdot v$ nur bei kleinen Schallschnellen gemacht werden kann. Treten im Vokaltrakt hingegen Konstriktionen auf, so ergeben sich in diesen verhältnismässig grosse Strömungsgeschwindigkeiten. Daher darf für ein Modell, welches beispielsweise auch Frikativlaute zulassen soll, diese Vereinfachungen nicht vorgenommen werden. Die Bewegungsgleichung für diesen Fall lautet:

$$-\frac{\partial p}{\partial x} = \rho_0 \cdot \frac{\partial v}{\partial t} + \rho_0 \cdot v \cdot \frac{\partial v}{\partial x} + r \cdot v = \frac{\rho_0}{A} \cdot \frac{\partial u}{\partial t} + \frac{\rho_0}{2} \cdot \frac{\partial}{\partial x}\left(\frac{u^2}{A^2}\right) + \frac{r \cdot u}{A}. \tag{3.6}$$

Die Kontinuitätsgleichung

Die Kontinuitätsgleichung beschreibt das Prinzip der Massenerhaltung. Betrachtet man einen infinitesimal kurzen Rohrabschnitt, so sind Änderungen der darin befindlichen Masse nur möglich, wenn ein Massenstrom hinein- bzw. herausfliesst. Die aus dem Massenerhaltungsgesetz resultierende zweite Grundgleichung in differenzieller Form lautet:

$$\frac{\partial \rho_a \cdot A'}{\partial t} = -\frac{\partial \rho_a \cdot A \cdot v}{\partial x}. \tag{3.7}$$

Unter der Annahme, dass die Rohrwände bei erhöhtem oder erniedrigtem Druck leicht nachgeben, muss in diesem Fall auch die zeitliche Variabilität der Querschnittsfläche berücksichtigt werden. Sofern diese Auslenkung der Wände klein gegenüber dem Durchmesser des Rohres ist, kann eine weitere Linearisierung vorgenommen werden:

$$A'(x,t) = A(x) + S(x) \cdot y(x,t). \tag{3.8}$$

$S(x)$ entspricht dem Rohrumfang und $y(x,t)$ der Wandauslenkung. Mit den Linearisierungen und durch das Vernachlässigen der Terme höherer Ordnung kann die Gleichung (3.7) geschrieben werden als:

$$A \cdot \frac{\partial \rho}{\partial t} + \rho_0 \cdot S \cdot \frac{\partial y}{\partial t} = -\rho_0 \cdot \frac{\partial u}{\partial x}. \tag{3.9}$$

Falls man starre Wände annimmt vereinfacht sich die Gleichung zu:

$$A \cdot \frac{\partial \rho}{\partial t} = -\rho_0 \cdot \frac{\partial u}{\partial x}. \tag{3.10}$$

Die Schwingungsgleichung

Bei elastischen Wänden, die aufgrund des Druckes eine Auslenkung erfahren, ist eine weitere Gleichung notwendig, um die daraus resultierenden Effekte zu beschreiben. In einer ersten Näherung verhält sich die Wand wie ein gedämpftes Masse-Feder-System, welches durch die Gleichung:

$$p = M(x) \cdot \frac{\partial^2 y}{\partial t^2} + B(x) \cdot \frac{\partial y}{\partial t} + K(x) \cdot y \tag{3.11}$$

beschrieben werden kann [27]. $M(x)$ entspricht dabei der Masse, $B(x)$ dem Dämpfungskoeffizienten und $K(x)$ der Federkonstante, jeweils pro Flächeneinheit. Je nach Beschaffenheit der Wand und des dahinter liegenden Gewebes oder Muskeln können diese Grössen variieren. Der Schwingungseffekt führt zu einer Dämpfung und frequenzmässigen Verschiebung der Formanten bei tiefen Frequenzen.

Die Zustandsgleichung

Wird eine konstante Menge Luft komprimiert oder dekomprimiert, so ändern sich Druck und Temperatur. Wenn dieser Vorgang genügend schnell abläuft, wie bei Schallwellen, so findet in der Regel kein Temperaturaustausch mit der Umgebung statt. Diese Zustandsänderung wird auch als adiabatisch bezeichnet und durch die Poisson'sche Gleichung:

$$p \cdot V^{\gamma} = const. \tag{3.12}$$

beschrieben, wobei V dem Volumen und γ dem Adiabatenexponent (~ 1.4 für Luft bei Raumtemperatur) entspricht. Für die zwei Volumina V_0 und $V' = V_0 + V$ folgt daher unter der Annahme eines adiabatischen Prozesses:

$$p_0 \cdot V_0^{\gamma} = (p_0 + p) \cdot (V_0 + V)^{\gamma}. \tag{3.13}$$

Setzt man in Gleichung (3.13) $V = k/\rho$ ein, wobei k eine Konstante ist, und trennt die Gleichung nach Druck und Dichte ergibt sich:

$$(1 + \frac{p}{p_0}) = (1 + \frac{\rho}{\rho_0})^{\gamma} \simeq 1 + \gamma\frac{\rho}{\rho_0}. \tag{3.14}$$

Der rechte Term folgt durch eine Taylor-Reihenentwicklung ($\rho_0 \ll \rho$) ohne die Terme höherer Ordnung. Gleichung (3.14) liefert die wichtige Beziehung zwischen Druck und Dichte:

$$p = c^2 \cdot \rho \quad \text{mit} \quad c = \sqrt{\frac{\gamma \cdot p_0}{\rho_0}} \tag{3.15}$$

mit dem Quadrat der Schallgeschwindigkeit als Proportionalitätskonstante.

3.2 Verluste im Vokaltrakt

Verluste verschiedener Ursachen haben Energieverluste im Sprechtrakt zur Folge. Diese führen zu Änderungen sowohl der Bandbreiten als auch der Frequenzen der Formanten. Die Auswirkungen der einzelnen Verluste sind unterschiedlich stark und sind im Allgemeinen auch frequenzabhängig.

3.2.1 Verluste durch Schallabstrahlung

Die bedeutendsten Verluste treten bei der Schallabstrahlung an den Lippen und den Nasenlöchern auf. Diese Verluste können in Form von Strahlungsimpedanzen erfasst werden, welche die Ausgänge des Sprechtraktes abschliessen.

Eine recht genaue Approximation erhält man durch einen vibrierenden Kolben in einer Kugel. Die Herleitung der Strahlungsimpedanz dieses Modells ist jedoch sehr schwierig und lässt sich nicht als geschlossener analytischer Ausdruck darstellen. Für den Fall, dass der Kolbendurchmesser klein ist gegenüber der Kugel, liefert das Modell einer Kolbenmembran, die sich in einer unendlich ausgedehnten schallharten Fläche befindet und darin schwingt, eine gute Näherung [13]. Die resultierende Strahlungsimpedanz lässt sich als Summe einer Besselfunktion $J_1(z)$ und einer Struve-Funktion $S_1(z)$ erster Ordnung schreiben:

$$Z_n(\omega) = \frac{\rho_0 c}{\pi R^2}(1 - \frac{J_1(2 \cdot k \cdot R)}{kR} + j\frac{S_1(2 \cdot k \cdot R)}{k \cdot R}) \quad \text{mit} \quad k = \frac{\omega}{c} \tag{3.16}$$

mit der Besselfunktion

$$J_1(z) = \frac{z}{2} - \frac{z^3}{2^3 \cdot 1! \cdot 2!} + \frac{z^5}{2^5 \cdot 2! \cdot 3!} - \frac{z^7}{2^7 \cdot 3! \cdot 4!} \pm \cdots \tag{3.17}$$

und der Struve-Funktion

$$S_1(z) = \frac{2}{\pi}\left(\frac{z^2}{1^2 \cdot 3} - \frac{z^4}{1^2 \cdot 3^2 \cdot 5} + \frac{z^6}{1^2 \cdot 3^2 \cdot 5^2 \cdot 7} \mp \cdots\right). \tag{3.18}$$

R bezeichnet den Radius des Kolbens. Die Strahlungsimpedanz ist sowohl eine Funktion der Frequenz als auch der Kolbenfläche und somit der Öffnungsfläche an den Lippen und Nasenlöchern.

Für kleine Werte von kR kann die Gleichung (3.16) vereinfacht werden, indem die Terme höherer Ordnung für $J_1(z)$ und $S_1(z)$ vernachlässigt werden [51]. In der selben Arbeit wird zudem ein Korrekturterm vorgeschlagen, der für eine bessere Anpassung der Strahlungsimpedanz an das Modell eines Kolbens in einer Kugel sorgt. Diese Approximation ist, vermutlich wegen ihrer Einfachheit, häufig anzutreffen. Die sich daraus ergebende Gleichung lautet:

$$Z_f(\omega) = \frac{\rho_0 \cdot c}{\pi \cdot R^2}\left(\frac{(k \cdot R)^2}{4} \cdot K_s(\omega) + j\frac{8 \cdot k \cdot R}{3 \cdot \pi}\right). \tag{3.19}$$

$K_s(\omega)$ ist der Korrekturterm, welcher wie folgt definiert ist:

$$K_s(\omega) = \begin{cases} \frac{0.6 \cdot \omega}{2 \cdot \pi \cdot 1600} + 1 & 0 \leq \omega < 2 \cdot \pi \cdot 1600 \\ 1.6 & \omega \geq 2 \cdot \pi \cdot 1600 \end{cases}. \tag{3.20}$$

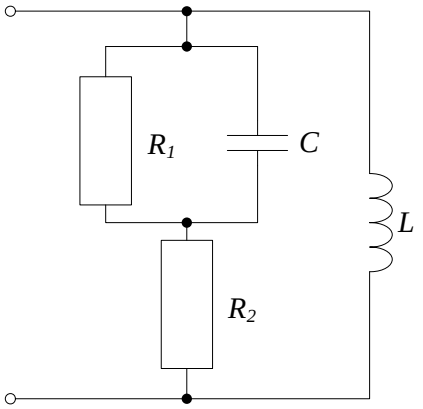
Ein Modell der Strahlungsimpedanz als elektrisches Netzwerk wird von VOJNOVIC und MIJIC in [49] vorgestellt. Die Schaltung, dargestellt in Abbildung 3.1, zeigt für Frequenzen bis 10 kHz eine gute Näherung. Die Elemente der Schaltung wurden durch ein Optimierungsverfahren ermittelt, wobei der mittlere quadratische Fehler zum Modell des Kolbens in einer Kugel minimiert wurde. Die daraus resultierende Abstrahlungsimpedanz kann geschrieben werden als:

$$Z_p(\omega) = R_p + jX_p, \tag{3.21}$$

wobei

$$R_p = \frac{\omega^2 \cdot L^2 \cdot (R_1 + R_2 + \omega^2 \cdot R_1^2 \cdot R_2 \cdot C^2)}{(R_1 + R_2 - \omega^2 \cdot R_1 \cdot L \cdot C)^2 + \omega^2 \cdot (L + R_1 \cdot R_2 \cdot C)^2},$$
$$X_p = \frac{\omega \cdot L \cdot [(R_1 + R_2)^2 - \omega^2 \cdot R_1^2 \cdot C \cdot (L - R_2^2 \cdot C)^2]}{(R_1 + R_2 - \omega^2 \cdot R_1 \cdot L \cdot C)^2 + \omega^2 \cdot (L + R_1 \cdot R_2 \cdot C)^2}.$$

In Abbildung 3.2 sind die verschiedenen Approximationen der normalisierten Strahlungsimpedanzen für eine Fläche von 4 cm^2 zum Vergleich dargestellt. Die Normalisierung erfolgt durch die Multiplikation mit $(\pi \cdot R^2)/(\rho_0 \cdot c)$. Die Strahlungsimpedanz eines Kolbens in einer unendlichen Fläche gilt dabei als Referenz. Die grobe Vereinfachung gemäss Gleichung (3.19) weist eine gute Übereinstimmung auf für Werte von $kR < 1$. Dies entspricht für eine Öffnungsfläche von 8 cm^2 Frequenzen bis 3.5 kHz oder bei einer Fläche von 3 cm^2 einem Frequenzbereich bis 5.7 kHz. Bei Werten von $kR > 1$ nimmt der Fehler jedoch sehr schnell zu. Das Modell von Gleichung (3.21) weist hingegen bis $kR \cong 3$ eine gute Approximation auf, was Frequenzen bis 10 kHz bei einer Fläche von 8 cm^2 einschliesst.



$$R_1 = \frac{30.09}{A^{1.08}}$$
$$R_2 = \frac{26.4}{A^{0.89}}$$
$$L = \frac{5.25 \cdot 10^{-4}}{A^{0.53}}$$
$$C = 1.56 \cdot 10^{-7} \cdot A^{1.74}$$

Abbildung 3.1: *Modell der Strahlungsimpedanz als elektrisches Netzwerk nach [49].*

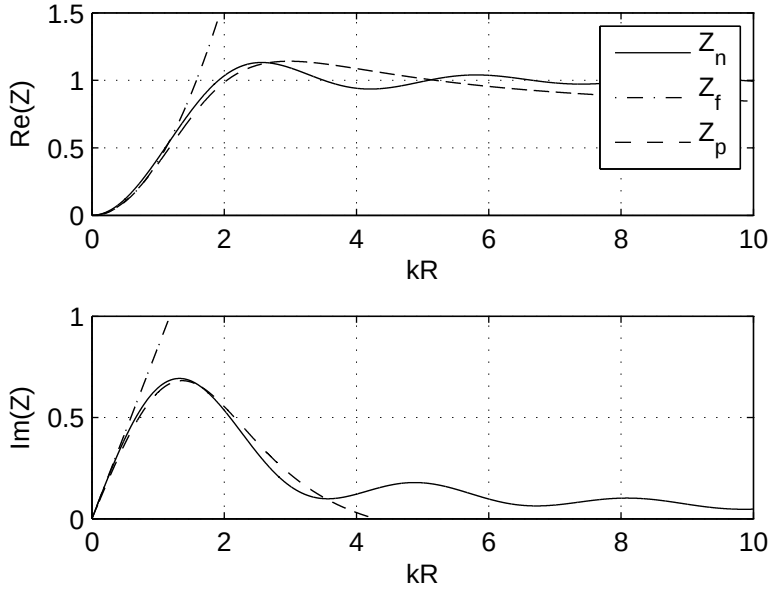


Abbildung 3.2: *Real- und Imaginärteil der normalisierten Strahlungsimpedanzen bei einer Öffnungsfläche von 4 cm*^{2}*: Z_n ist die Strahlungsimpedanz eines Kolbens in einer unendlichen Fläche. Z_f ist die Approximation nach* WAKITA *und Z_p nach* VOJNOVIC *und* MIJIC.

	M [kg/m^2]	B [kg/(m^{2}s)]	K [N/m^3]	F0 [Hz]
Wangen (entspannt)	2.1	8000	84500	32
Wangen (angespannt)	1.5	10600	33300	60
Hals	2.4	23200	491000	72
Unterarm	1.8	16200	615000	93

Tabelle 3.2: *Messwerte der ermittelten mechanischen Impedanzen einiger Köperbereiche aus [17].*

3.2.2 Verluste durch elastische Rohrwände

Wie aus der Schwingungsgleichung ersichtlich, können die elastischen Wände als gedämpftes Masse-Feder-System angesehen werden [13]. Die Grössen des Systems sind abhängig vom Gewebe, welches modelliert werden soll und variieren daher entlang des Sprechtraktes. Eine Reihe von Messungen, die eine Abschätzung der Parameter ermöglichen, wurden von ISHIZAKA *et al.* [17] gemacht. Die Parameter wurden dabei aus Messungen der mechanischen Impedanz einiger Köperbereiche abgeleitet. In Tabelle 3.2 sind diese Werte aufgelistet. Welche Parameterwerte für welchen Bereich des Sprechtraktes angemessen sind, ist indes nicht klar. Bei Sprachsynthese-Systemen werden für den Sprechtrakt die Werte der entspannten Wangen bevorzugt verwendet [27, 3].

3.2.3 Verluste durch Reibung an den Rohrwänden

Ein weiterer Verlust, der an den Rohrwänden auftritt, entsteht durch Reibung. Die Sprechtraktwände sind nicht perfekt glatt, sondern weisen gewisse Unebenheiten auf. Diese Unebenheiten führen zu Störungen der Strömung in einer Grenzschicht entlang der Rohrwände. Die Strömungsgeschwindigkeit in der Mitte des Rohres wird davon wenig beeinflusst. Die Dicke der Grenzschicht ist von der Viskosität μ, von der Dichte ρ_0 des Ausbreitungsmediums und von der Frequenz abhängig:

$$\delta_v = \sqrt{\frac{2 \cdot \mu}{\omega \cdot \rho}}. \tag{3.22}$$

In [13] ist für diese Grenzschicht ein äquivalenter elektrischer Widerstand pro Längeneinheit hergeleitet:

$$R_i = \sqrt{\frac{2 \cdot \pi \cdot \omega \cdot \rho \cdot \mu}{A_i^3}}. \tag{3.23}$$

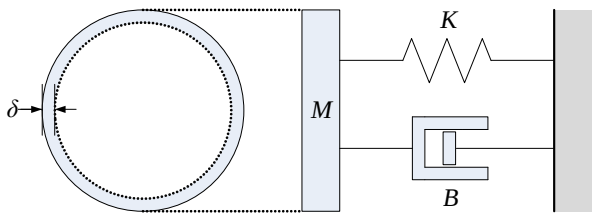


Abbildung 3.3: *Grenzschicht und Modell der elastischen Rohrwand.*

3.2.4 Verluste durch Wärmeleitung

Ein ähnlicher Effekt wie bei der Reibung ist zu beobachten, wenn die Wände thermische Energie aufnehmen. Die Wärme entsteht durch Kompression beim Durchgang der Schallwelle. Auch hier entsteht eine Grenzschicht entlang der Rohrwände, die zu einer Dämpfung aufgrund der thermalen Verluste führt. Die Dicke dieser Grenzschicht hängt von der Wärmeleitfähigkeit κ, dem Adiabatenexponen γ, der Dichte ρ_0, der spezifischen Wärmekapazität C_p und der Frequenz ab:

$$\delta_h = \sqrt{\frac{2 \cdot \kappa}{\omega \cdot \rho_0 \cdot C_p}}. \tag{3.24}$$

Der äquivalente elektrische Leitwert lautet [13]:

$$G_i = \frac{S_i \cdot (\gamma - 1)}{\rho_0 \cdot c^2} \sqrt{\frac{\kappa \cdot \omega}{2 \cdot C_p \cdot \rho_0}}. \tag{3.25}$$

S_i entspricht dem Rohrumfang. Die Verluste aufgrund der Wärmeleitung und damit ihr Effekt auf die Formantbandbreiten sind vergleichsweise gering.

3.2.5 Verluste an Konstriktionen

Bei Frikativen erhöht sich die Teilchengeschwindigkeit im Bereich der Verengung sehr stark, was mit einem Druckverlust einher geht, bekannt als Bernoulli-Effekt. Dieser Druckverlust wird am Ausgang der Konstriktionsstellen (Glottis und Verengung im Vokaltrakt) nicht mehr vollständig kompensiert. Als Ursache vermutet man die entstehenden Verwirbelungen nach den Konstriktionen. Eine Möglichkeit, die Druckverluste zu berücksichtigen, ist ein serieller Widerstand am Ausgang der Konstriktion [44]:

$$R_c = \frac{\rho_0 \cdot \mid \bar{u} \mid}{2 \cdot A_c}. \tag{3.26}$$

$\bar{u}$ ist der mittlere Volumenstrom. Für eine Simulation im Frequenzbereich kann dieser aus dem subglottalen Druck p_{sub} (typischer Wert ∼800 Pa) und den Querschnittsflächen an den Konstriktionen der Glottis A_g und des Vokaltraktes A_v geschätzt werden [3]:

$$\bar{u} = \sqrt{\frac{2 \cdot p_{sub}}{\rho \cdot (1/A_g^2 + 1/A_v^2)}}. \tag{3.27}$$

3.3 Verlustloses zeitdiskretes Rohrmodell

Im Jahr 1962 wurde von KELLY und LOCHBAUM [19] erstmals ein Verfahren zur akustischen Simulation eines zeitdiskreten Rohrsystems vorgestellt. Für die Modellierung des Sprechtraktes wurde dieses Verfahren seither vielfach aufgegriffen (z.B. [50, 32, 21, 20, 39]). In seiner einfachsten Form wird der Sprechtrakt damit als ein System von miteinander verbundenen Rohrabschnitten gleicher Länge, aber mit variablen Durchmessern angesehen.

Für dieses Modell wird angenommen, dass keine Verluste im Sprechtrakt auftreten und dass die Rohrwände starr sind.

3.3.1 Grundgleichungen im verlustlosen Fall

Die Wellengleichung für die verlustlosen Schallausbreitung lässt sich aus der Bewegungsgleichung (3.5), der Kontinuitätsgleichung (3.10) und der Zustandsgleichung (3.15) herleiten. Das Einsetzen von (3.15) in die Gleichung (3.10) und die Elimination des Schalldruckes mit Hilfe von (3.5) ergibt:

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2} \cdot \frac{\partial^2 u}{\partial t^2} = 0. \tag{3.28}$$

Diese Wellengleichung gilt ebenso für das Schnellepotential $\Phi(x,t)$:

$$\frac{\partial^2 \Phi}{\partial x^2} - \frac{1}{c^2} \cdot \frac{\partial^2 \Phi}{\partial t^2} = 0, \tag{3.29}$$

welches mit der Schallschnelle und dem Schalldruck über folgende Beziehungen verknüpft ist:

$$u(x,t) = -A \cdot \frac{\partial \Phi(x,t)}{\partial x} \quad \text{und} \quad p(x,t) = \rho \cdot \frac{\partial \Phi(x,t)}{\partial t}. \tag{3.30}$$

Für den Fall einer sinusförmigen Funktion lautet die allgemeine Lösung der Gleichung (3.29):

$$\Phi(x,t) = C \cdot e^{j\omega \cdot (t - x/c)} + D \cdot e^{j\omega \cdot (t + x/c)}, \tag{3.31}$$

wobei C und D zwei Konstanten sind. Das Einsetzen der allgemeinen Lösung in die Wellengleichung des Schnellepotentials zeigt, dass die Ausbreitung der jeweiligen Schallfeldgrösse als eine Überlagerung einer vor- und einer zurücklaufenden Welle beschrieben werden kann. Der Schalldruck und der Schallfluss in einem Rohrabschnitt mit dem Index i können entsprechend als Summe geschrieben werden:

$$u_i(t) = u_i^+(t) - u_i^-(t), \tag{3.32}$$

$$p_i(t) = \frac{\rho \cdot c}{A_i} \cdot [u_i^+(t) - u_i^-(t)]. \tag{3.33}$$

Die zeitdiskrete Betrachtung des Prozesses hat zur Folge, dass die Schallfeldgrössen nur an Orten berechnet werden können, in der sich die Schallwellen nach einem Vielfachen des Abtastintervalls befinden können. Die Länge der Rohrabschnitte Δl wird also durch die Abtastfrequenz Fs und durch die Ausbreitungsgeschwindigkeit c bestimmt:

$$\Delta l = \frac{c}{2 \cdot Fs}. \tag{3.34}$$

Anstatt der x-Koordinate wird deshalb der betreffende Rohrabschnitt im Weiteren mit dem Index i und zusätzlich noch mit l, r für die linke bzw. rechte Abschnittsgrenze spezifiziert, wie in Abbildung 3.4 dargestellt ist.

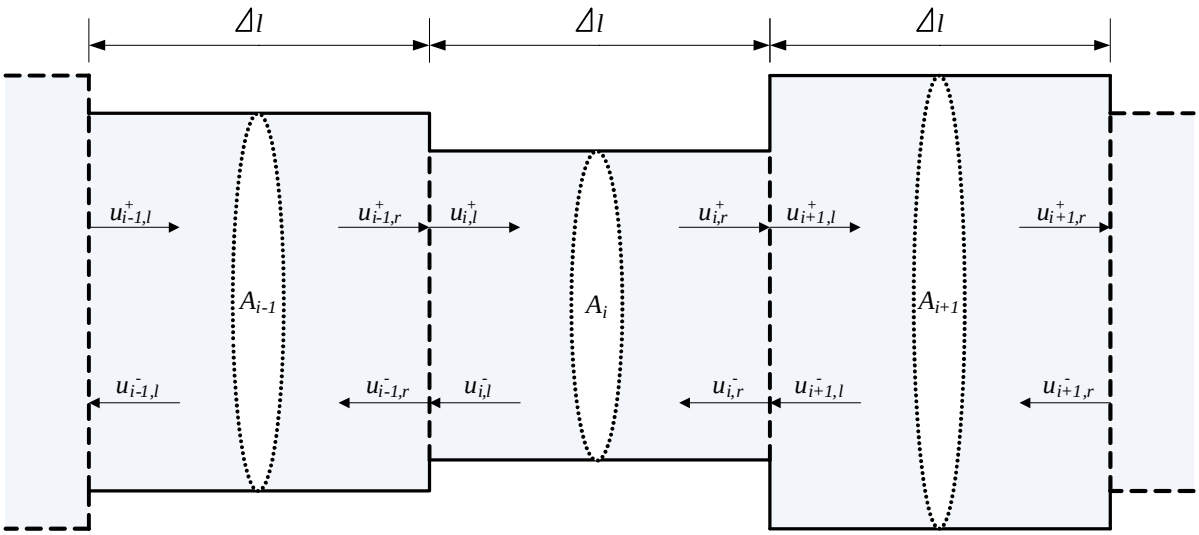


Abbildung 3.4: *Schallfluss im zeitdiskreten verlustlosen Rohrmodell.*

Da es gemäss Annahmen keine Verluste innerhalb eines Rohrabschnittes gibt, gelten die Gleichungen:

$$u^+_{i,l}(t - \frac{\Delta l}{c}) = u^+_{i,r}(t), \tag{3.35}$$

$$u^-_{i,l}(t + \frac{\Delta l}{c}) = u^-_{i,r}(t). \tag{3.36}$$

An den Grenzen zweier Rohrabschnitte muss zudem die Kontinuität der Schallfeldgrössen gewährleistet sein. Die daraus resultierenden Bedingungen lauten:

$$u_{i-1,r}(t) = u_{i,l}(t), \tag{3.37}$$

$$p_{i-1,r}(t) = p_{i,l}(t). \tag{3.38}$$

Mit diesen Bedingungen und den Gleichungen (3.32) und (3.33) kann der Zusammenhang der Schallfeldgrössen an der Schnittstelle zweier Rohrabschnitte als Funktion der Querschnittsflächen ausgedrückt werden. Wenn ξ_i dem Reflexionsfaktor

$$\xi_i = \frac{A_{i-1} - A_i}{A_{i-1} + A_i} \tag{3.39}$$

an einem Übergang entspricht, lautet der Schallfluss der vorwärts- und rückwärtslaufenden Wellen:

$$u^+_{i-1,l}(t - \frac{\Delta l}{c}) = \frac{1}{1 - \xi_i} \cdot [u^+_{i,l}(t) + \xi_i u^-_{i,l}(t)], \tag{3.40}$$

$$u^-_{i-1,l}(t + \frac{\Delta l}{c}) = \frac{1}{1 - \xi_i} \cdot [\xi_i u^+_{i,l}(t) + u^-_{i,l}(t)]. \tag{3.41}$$

Eine Umformung dieser Gleichungen und die Transformation in den z-Bereich führt zu einer praktischen Matrix-Form, die weit verbreitet ist:

$$\begin{bmatrix} U^+_{i-1,l} \\ U^-_{i-1,l} \end{bmatrix} = z^{\frac{1}{2}} \cdot \frac{1}{1-\xi_i} \begin{bmatrix} 1 & \xi_i z^{-1} \\ \xi_i & z^{-1} \end{bmatrix} \cdot \begin{bmatrix} U^+_{i,l} \\ U^-_{i,l} \end{bmatrix}. \tag{3.42}$$

Die Verzögerungen von $z^{\pm\frac{1}{2}}$ ergeben sich durch die Laufzeiten pro Rohrabschnitt, wobei z, anders als üblich, als $z = e^{j\omega \cdot 2 \cdot \Delta l/c}$ definiert ist. Dies dient der Anpassung an die Ergebnisse des Modells der inversen Filterung [50], wie später in Abschnitt 5.1 ersichtlich wird. In dieser Form sind auch die Indices l und r nicht mehr nötig.

Der Vorteil der Matrix-Darstellung liegt darin, dass die Beziehungen der Schallfeldgrössen über mehrere Rohrelemente hinweg durch das Produkt der Matrizen dargestellt werden können. Für ein Rohrsystem bestehend aus N Rohrabschnitten gilt dann:

$$\begin{aligned} \begin{bmatrix} U^+_0 \\ U^-_0 \end{bmatrix} &= z^{\frac{N}{2}} \cdot \prod_{i=1}^{N} \frac{1}{1-\xi_i} \begin{bmatrix} 1 & \xi_i z^{-1} \\ \xi_i & z^{-1} \end{bmatrix} \cdot \begin{bmatrix} U^+_N \\ U^-_N \end{bmatrix} \\ &= z^{\frac{N}{2}} \cdot \prod_{i=1}^{N} T_i \cdot \begin{bmatrix} U^+_N \\ U^-_N \end{bmatrix}. \end{aligned} \tag{3.43}$$

T_i wird als Betriebskettenmatrix bezeichnet. Der Term $z^{\frac{N}{2}}$ entspricht nur einer Gesamtverzögerung und wird im Weiteren vernachlässigt. Der Vorfaktor $\frac{1}{1-\xi_i}$ variiert, je nachdem welche Schallfeldgrösse betrachtet wird. Für Schalldruckwellen lautet der Vorfaktor $\frac{1}{1+\xi_i}$ [22].

Das Produkt $\Gamma = \prod_{i=1}^{N} \frac{1}{1-\xi_i}$ dieser Vorfaktoren stellt lediglich einen Verstärkungsfaktor dar und wird für die weiteren Betrachtungen ebenfalls weggelassen.

3.3.2 Übertragungsfunktion

Für ein zeitdiskretes Rohrmodell bestehend aus N Rohrelementen mit den Betriebsmatrizen T_i und einem Rohrabschluss C am Systemausgang, wie in Abbilung 3.5 zu sehen ist, kann die Übertragungsfunktion mit Hilfe der Gleichung (3.43) berechnet werden. X steht dabei für die gewählte Schallfeldgrösse.

Die Betriebskettenmatrizen lassen sich durch Multiplikation miteinander zur einer Matrix zusammenfassen:

$$T_K = \begin{bmatrix} T_{K,11} & T_{K,12} \\ T_{K,21} & T_{K,22} \end{bmatrix} = \prod_{i=1}^{N} T_i. \tag{3.44}$$

Damit folgt:

$$\begin{bmatrix} X^+_0 \\ X^-_0 \end{bmatrix} = \begin{bmatrix} T_{K,11} & T_{K,12} \\ T_{K,21} & T_{K,22} \end{bmatrix} \cdot \begin{bmatrix} X^+_N \\ X^-_N \end{bmatrix} = \begin{bmatrix} T_{K,11} & T_{K,12} \\ T_{K,21} & T_{K,22} \end{bmatrix} \cdot \begin{bmatrix} X^+_N \\ X^+_N \cdot C \end{bmatrix}. \tag{3.45}$$

Die Übertragungsfunktion des Systems lautet daher:

$$H(z) = \frac{X^+_N}{X^+_0} = \frac{1}{T_{K,11} + C \cdot T_{K,12}}. \tag{3.46}$$

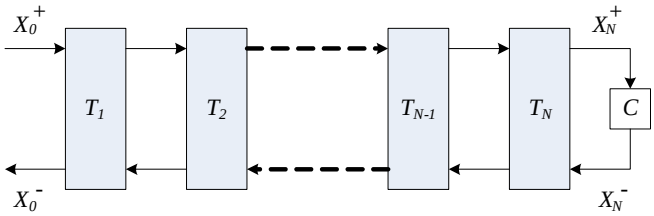


Abbildung 3.5: *Betriebskettenmatrizen des verlustlosen Rohrmodells.*

Unter der Annahme, dass an den Lippen ein idealisierter Übergang von einer endlichen zu einer unendlichen Fläche stattfindet, entspricht C dem Reflexionsfaktor -1 und damit einem schallweichen Abschluss.

3.4 Verlustbehaftetes diskretes Rohrmodell als akustisches Netzwerk

Das akustische System des Sprechtraktes kann auch durch die Übertragung auf ein entsprechendes elektrisches Netzwerk beschrieben werden. Die Analogie zwischen den Grössen und den Impedanzen sind in den Tabellen 3.3 und 3.4 aufgezeigt.

	Potentialgrösse	Flussgrösse
elektrisch	Spannung	Strom
akustisch	Schalldruck	Schallfluss

Tabelle 3.3: *Elektrisch-akustische Analogien.*

	elektrisch	akustisch
$Z = R$	R	$R = R_m \cdot G_1$
$Z = \frac{1}{j\omega c}$	C	$C = \frac{V}{\rho c^2} \cdot G_2$
$Z = j\omega L$	L	$L = \frac{\rho l}{A} \cdot G_3$

Tabelle 3.4: *Vergleich elektrischer und akustischer Impedanzen. V entspricht einem Volumen, l einer Länge und G_i sind Proportionalitätskonstanten.*

Ein frühes Modell für die Simulation des Sprechtraktes im Zeitbereich mit Hilfe eines elektrischen Ersatzschaltbildes stellt MAEDA in [27] vor. Neuere Arbeiten auf dem Gebiet der Sprachsynthese, die auf einer akustischen Zeitbereichssimulation beruhen, sind Weiterentwicklungen dieses Modells ([21, 3]). SONDHI hat ein hybrides System als Abwandlung davon entwickelt [44]. Dabei findet zuerst eine Simulation im Frequenzbereich statt, um die Übertragungsfunktion zu ermitteln. Die Synthese erfolgt dann wiederum mit Hilfe der entsprechenden Impulsantwort und Anregungsfunktion im Zeitbereich.

Die Modellierung des Sprechapparates in Form eines elektrischen Netzwerks bietet einige Vorteile. So lassen sich frequenzabhängige Verluste innerhalb des Sprechtraktes auf einfache Art berücksichtigen. Eine grosse Einschränkung des verlustlosen Rohrmodells ist, dass die Länge der einzelnen Rohrabschnitte fest mit der Abtastfrequenz und der Schallgeschwindigkeit verküpft sind. Die Modellierung mit Hilfe eines elektrischen Netzwerks erlaubt hingegen auch zeitlich variable Längen der Rohrabschnitte und damit auch die Variation der Gesamtlänge des Vokaltraktes unabhängig von der Abtastfrequenz.

3.4.1 Grundgleichungen im verlustbehafteten Fall

Die Grundgleichungen des Systems sind die Bewegungsgleichung (3.6), die Kontinuitätsgleichung (3.9) und die Schwingungsgleichung (3.11). Zusammengefasst mit dem Schalldruck und

dem Schallfluss als Schallfeldgrössen lauten sie:

$$\frac{\partial p}{\partial x} + \frac{\rho_0}{A} \cdot \frac{\partial u}{\partial t} + \frac{\rho_0}{2} \cdot \frac{\partial}{\partial x}\left(\frac{u^2}{A^2}\right) + \frac{r \cdot u}{A} = 0, \tag{3.47}$$

$$\frac{\partial p}{\partial t} + \frac{\rho_0 \cdot c^2}{A}\left(\frac{\partial u}{\partial x} + S \cdot \frac{\partial y}{\partial t}\right) = 0, \tag{3.48}$$

$$M \cdot \frac{\partial^2 y}{\partial t^2} + B \cdot \frac{\partial y}{\partial t} + K \cdot y = p. \tag{3.49}$$

Die drei Unbekannten des Gleichungssystems sind der Schalldruck $p(x,t)$, der Schallfluss $u(x,t)$ und die Auslenkung der Rohrwand $y(x,t)$. Numerische Lösungen können durch eine Zeit- bzw. Ortsdiskretisierung erhalten werden. Dazu wird das Rohrsystem in Abschnitte mit einer konstanten Querschnittsfläche unterteilt. Die Abtastung von $p(x,t)$ und $y(x,t)$ erfolgt in der Mitte jedes Abschnittes. $u(x,t)$ wird am Ein- bzw. Ausgang der Rohrabschnitte abgetastet.

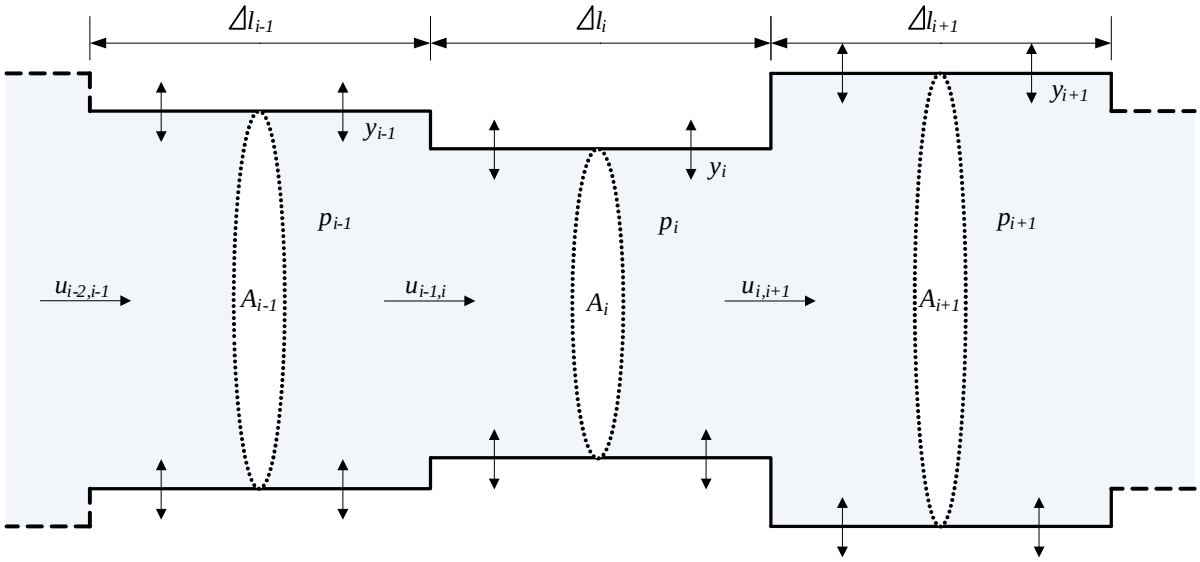


Abbildung 3.6: *Schallgrössen im verlustbehafteten Rohrmodell.*

Die ortsdiskrete Bewegungsgleichung erhält man durch Integration vom Anfang eines Rohrabschnittes bis zum Abschnittsende über ein differenzielles Wegelement. Analog dazu wird die ortsdiskrete Kontinuitätsgleichung von der Mitte eines Abschnittes bis zur Mitte des nächsten Abschnittes berechnet. Die dazu nötigen Berechnungen findet man in einer ausführlichen Form in [27].

3.4.2 Elemente und Struktur des elektrischen Netzwerks

Aus der Diskretisierung folgen die konzentrierten Bauelemente des elektrischen Netzwerks:

$$
\begin{aligned}
R_i &= \sqrt{\frac{l_i^2 \cdot \pi \cdot \omega \cdot \rho \cdot \mu}{2 \cdot A_i^3}} & \qquad R_{w,i} &= \frac{B_i}{O_i} \\
L_i &= \frac{\rho_0 \cdot l_i}{2 \cdot A_i} & L_{w,i} &= \frac{M_i}{O_i} \\
C_i &= \frac{A_i \cdot l_i}{\rho_0 \cdot c^2} & C_{w,i} &= \frac{O_i \cdot l_i}{K_i} \\
G_i &= \frac{O_i \cdot (\gamma - 1)}{\rho_0 \cdot c^2} \sqrt{\frac{\kappa \cdot \omega}{2 \cdot C_p \cdot \rho_0}}
\end{aligned}
\tag{3.50}
$$

R_i repräsentiert den Energieverlust durch Reibung an den Rohrwänden. Der akustische Widerstand eines Rohrelementes wird immer auch von einer akustischen Masse begleitet. Das elektrische Gegenstück dieser akustischen Masse ist die Induktivität L_i. Die Kapazität C_i, welche einer akustischen Feder entspricht, ermöglich die Kompression und Expansion der Luft innerhalb eines Rohrelementes. Die Leitfähigkeit G_i modelliert die Energieverluste durch die Wärmeleitung entlang der Wände. $R_{w,i}$, $L_{w,i}$ und $C_{w,i}$ berücksichtigen die mechanischen Effekte, die durch die elastischen Wände entstehen.

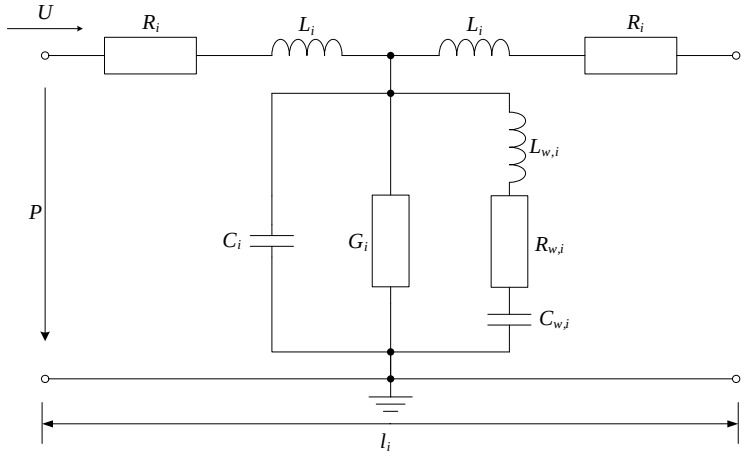


Abbildung 3.7: *Ersatzschaltbild eines einfachen Rohrelements.*

$O_i = \sqrt{4 \cdot \pi \cdot A_i \cdot l_i^2}$ entspricht der Oberfläche eines Rohrelementes mit einem kreisförmigen Querschnitt.

Damit sind alle Elemente eines einfachen Rohrabschnittes gemäss Abbildung 3.7 bestimmt. Die Rohrelemente des Rachen-, des Mund- und des Nasenraumes werden durch eine Verkettung solcher Abschnitte gebildet.

Nasennebenhöhlen

Die Nasennebenhöhlen werden durch diskrete Helmholtz-Resonatoren nachgebildet, wie von DANG und HONDA [6] vorgeschlagen. Diese haben sich gegenüber Seitenzweigen zur Modellierung der Nasennebenhöhlen besser bewährt. Die Helmholtz-Resonatoren sind charakterisiert durch die Länge des Resonatorhalses l_i, durch die Querschnittsfläche A_i des Resonatorhalses und durch das eingeschlossene Volumen V_i. Die geometrischen Grössen für die verschiedenen Nasennebenhöhlen sind in Tabelle 2.1 zu finden. Das resultierende elektrische Netzwerk

besteht aus einer Serieschaltung der Reibungswiderstände und der akustischen Federung, die sich durch das eingeschlossene Volumen ergibt.

$$
\begin{aligned}
R_i &= \sqrt{\frac{2 \cdot l_i^2 \cdot \pi \cdot \omega \cdot \rho \cdot \mu}{A_i^3}} & \qquad R_{w,i} &= \frac{B_i}{O_i} \\
L_i &= \frac{\rho_0 \cdot l_i}{A_i} & \qquad L_{w,i} &= \frac{M_i}{O_i} \\
C_i &= \frac{V_i}{\rho_0 \cdot c^2} & \qquad C_{w,i} &= \frac{O_i}{K_i} \\
G_i &= \frac{O_i \cdot (\gamma - 1)}{\rho_0 \cdot c^2} \sqrt{\frac{\kappa \cdot \omega}{2 \cdot C_p \cdot \rho_0}}
\end{aligned}
\tag{3.51}
$$

Die Elemente zur Berücksichtigung der Wandeffekte des Hohlraumes sind die gleichen wie bei einem einfachen Rohrabschnitt, lediglich die Oberfläche muss entsprechend angepasst werden: $O_i = 4 \cdot \pi \cdot \left(\frac{3 \cdot V_i}{4 \cdot \pi}\right)^{\frac{2}{3}}$. Die Wände des Resonatorhalses werden als starr angenommen.

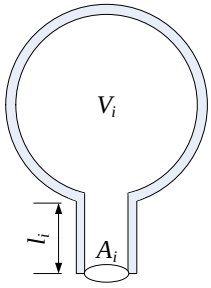


Abbildung 3.8: *Grössen des Helmholtz-Resonators.*

Sinus piriformis

Die Hohlräume der Sinus piriformis werden als zwei Seitenzweige, die an den Rachentrakt angeschlossen sind, modelliert. Die Seitenzweige werden durch die Verkettung einfacher Rohrabschnitte gebildet, wobei der letzte Abschluss wie ein offenes Rohrende behandelt wird. Die verwendeten Querschnittsflächen und die Längen der einzelnen Abschnitte sind in Tabelle 3.5 aufgeführt und sind den Messdaten aus [7] nachempfunden. Der Anschluss der zwei Seitenzweige erfolgt ca. 2 cm oberhalb der Glottis.

Subglottaler Bereich

Der subglottale Bereich des Sprechapparates, bestehend aus der Luftröhre, den Bronchien und der Lunge, wird durch eine Verkettung von Rohrverbundabschnitten gebildet [18, 3]. Die zunehmende Verästelung der Luftwege von der Glottis an abwärts wird durch ein Verbund von n_i homogenen Einzelrohren modelliert, deren Querschnittsflächen sich zusammen auf $A_{ges,i}$ summieren. Die Querschnittsflächen basieren auf dem Modell von WAIBEL [52] (siehe Kapitel 2.1), wobei die ersten 3 langen Rohrabschnitte weiter unterteilt werden und die letzten kurzen Abschnitte der Respirationszone zusammengefasst werden. Die Wandmasse M_i beträgt 3 kg/m^2, die Wandsteifigkeit K_i = 845000 N/m^3 und der Dämpfungskoeffizient nimmt linear von den Alveolen zur Luftröhre von 10000 kg/(m^2·s) auf 1000 kg/(m^2·s) ab. Diese Werte wurden

Index i	$l_{l,i}[cm]$	$A_{l,i}[cm^2]$	$l_{r,i}[cm]$	$A_{r,i}[cm^2]$
1	0.8	1	0.8	1.2
2	0.5	0.8	0.5	0.95
3	0.5	0.6	0.5	0.7
4	0.5	0.4	0.5	0.45
5	0.5	0.2	0.5	0.2

Tabelle 3.5: *Querschnittsflächen $A_{l,i}$, $A_{r,i}$ und Längen $l_{r,i}$, $l_{l,i}$ des linken und rechten Sinus Piriformis.*

von ISHIZAKA *et al.* [18] durch Anpassung von simulierten an gemessenen Resonanzen der subglottalen Luftwege ermittelt.

Die Elemente des elektrischen Netzwerks lauten:

$$\begin{aligned} R_i &= \sqrt{\frac{l_i^2 \cdot \pi \cdot \omega \cdot \rho \cdot \mu}{2 \cdot A_i^3 \cdot n_i^2}} & R_{w,i} &= \frac{B_i}{O_i \cdot n_i} \\ L_i &= \frac{\rho_0 \cdot l_i}{2 \cdot A_i \cdot n_i} & L_{w,i} &= \frac{M_i}{O_i \cdot n_i} \\ C_i &= \frac{A_i \cdot l_i \cdot n_i}{\rho_0 \cdot c^2} & C_{w,i} &= \frac{O_{i \cdot n_i} \cdot l_i}{K_i} \\ G_i &= \frac{O_i \cdot (\gamma - 1)}{\rho_0 \cdot c^2} \sqrt{\frac{\kappa \cdot \omega}{2 \cdot C_p \cdot \rho_0}} \end{aligned} \qquad (3.52)$$

Mit diesen Bausteinen lässt sich nun ein Modell des gesamten Sprechapparates realisieren, wie in Abbildung 3.9 dargestellt.

3.4.3 Übertragungsfunktion

Die Berechnung der Übertragungsfunktion des Netzwerks beruht auf der Vierpoltheorie. Ein Vierpol beschreibt allgemein ein elektrisches Bauelement mit vier Anschlüssen. Zwei Anschlüsse werden jeweils zu einem Klemmenpaar zusammengefasst, und als Eingang bzw. Ausgang bezeichnet. Die Grössen an den Klemmen sind der Schallfluss $U(\omega)$ und der Schalldruck $P(\omega)$. Das Klemmenverhalten eines Vierpoles wird durch seinen Frequenzgang beschrieben. Die Zusammenhänge der Grössen an den Klemmenpaaren können mit Gleichungen in einer kompakten Form als Matrix notiert werden. Durch Permutation sind diverse Darstellungen möglich. Für die Berechnung der Übertragungsfunktion eignet sich die bereits bekannte Darstellung durch Kettenmatrizen gut:

$$\begin{bmatrix} P_{Ein}(\omega) \\ U_{Ein}(\omega) \end{bmatrix} = \begin{bmatrix} A_{11}(\omega) & A_{12}(\omega) \\ A_{21}(\omega) & A_{22}(\omega) \end{bmatrix} \cdot \begin{bmatrix} P_{Aus}(\omega) \\ U_{Aus}(\omega) \end{bmatrix}. \qquad (3.53)$$

Mit den frequenzabhängigen Elementen der Matrix A ist das Verhalten des Systems vollständig charakterisiert. Das gesamte Netzwerk des Sprechapparates lässt sich auf zwei grundlegende Typen von Vierpolen reduzieren.

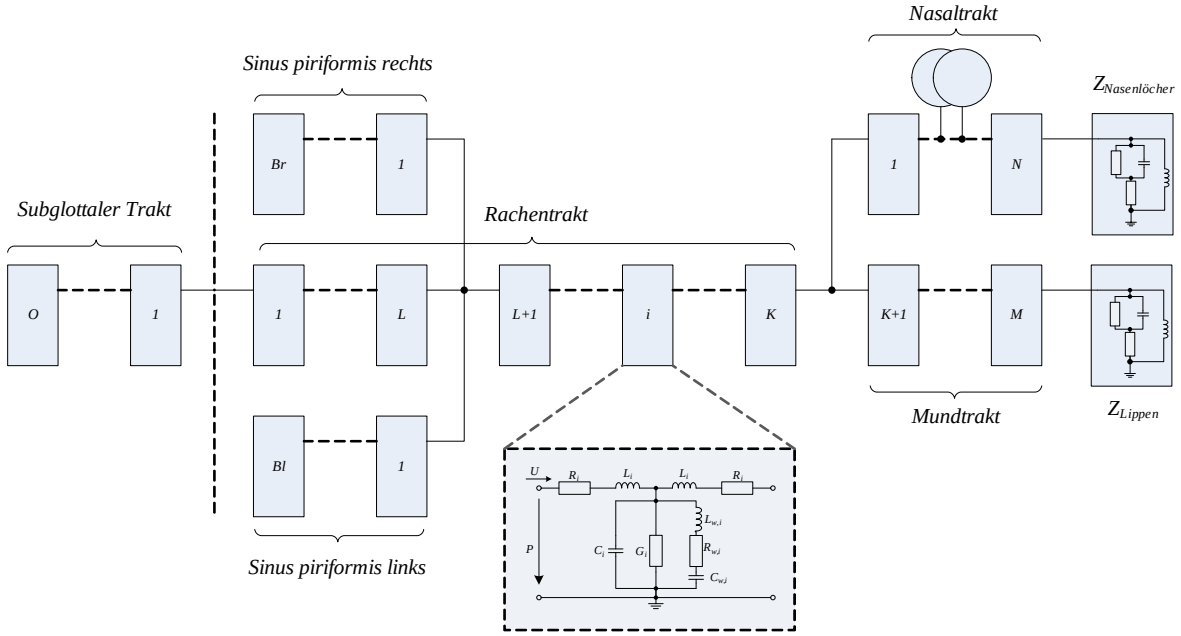


Abbildung 3.9: *Übersicht des Netzwerks des Sprechapparates.*

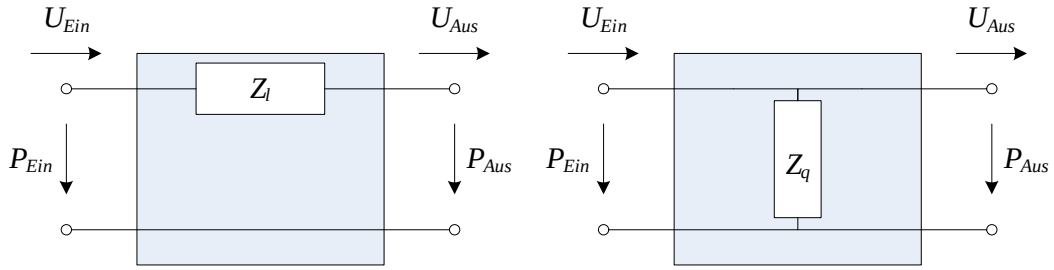


Abbildung 3.10: *Vierpole mit einer Längs- bzw. einer Querimpedanz.*

Der erste Typ weist eine einzelne Längsimpedanz zwischen den Klemmen des Ein- und des Ausganges auf, wie in Abbildung 3.10 gezeigt. Die Beziehungen des Klemmenpaares lauten in Kettenmatrix-Form:

$$\begin{bmatrix} P_{Ein}(\omega) \\ U_{Ein}(\omega) \end{bmatrix} = \begin{bmatrix} 1 & Z_l(\omega) \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} P_{Aus}(\omega) \\ U_{Aus}(\omega) \end{bmatrix}. \tag{3.54}$$

Der zweite vorkommende Typ besitzt eine Querimpedanz zwischen den Klemmenpaaren. Die dazugehörigen Gleichungen mit der Kettenmatrix lauten:

$$\begin{bmatrix} P_{Ein}(\omega) \\ U_{Ein}(\omega) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/Z_q(\omega) & 1 \end{bmatrix} \cdot \begin{bmatrix} P_{Aus}(\omega) \\ U_{Aus}(\omega) \end{bmatrix}. \tag{3.55}$$

Für einen einfachen Rohrabschnitt können die Elemente des elektrischen Ersatzschaltbildes wie folgt zusammengefasst werden:

$$Z_{l,i} = Z_{r,i} = R_i + j\omega \cdot L_i, \tag{3.56}$$

$$Z_{q,i} = \frac{1}{G_i + j\omega \cdot C_i + 1/[R_{w,i} + j\omega \cdot L_{w,i} + \frac{1}{j\omega \cdot C_{w,i}}]}. \tag{3.57}$$

Die Kettenmatrix eines Rohrabschnittes ergibt sich dann zu:

$$\begin{bmatrix} P_{Ein,i}(\omega) \\ U_{Ein,i}(\omega) \end{bmatrix} = \begin{bmatrix} 1 & Z_{l,i}(\omega) \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 1/Z_{q,i}(\omega) & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & Z_{r,i}(\omega) \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} P_{Aus,i}(\omega) \\ U_{Aus,i}(\omega) \end{bmatrix} \\ = K_i \cdot \begin{bmatrix} P_{Aus,i}(\omega) \\ U_{Aus,i}(\omega) \end{bmatrix}. \tag{3.58}$$

Die Berechnung der Schallfeldgrössen über mehrere Rohrabschnitte hinweg erfolgt durch die Multiplikation der Kettenmatrizen. Die Seitenzweige der Sinus piriformis, der Nasaltrakt und die diskreten Helmholtz-Resonatoren werden als zusätzliche Querimpedanzen eingekoppelt. Es bietet sich hier an, die Elemente eines Zweiges zuerst zu einer Matrix $K_s = K_1 \cdot ... \cdot K_N$ zusammenzufassen. Die Eingangsimpedanz eines Zweiges mit der Lastimpedanz Z_L lautet dann:

$$Z_{Ein} = \frac{P_{Ein}}{U_{Ein}} = \frac{K_{s,11} \cdot Z_L + K_{s,12}}{K_{s,21} \cdot Z_L + K_{s,22}}, \tag{3.59}$$

oder in vereinfachter Form

$$Z_{Ein} = \frac{K_{s,11}}{K_{s,21}},$$

wenn der Zweig durch einen Leerlauf abgeschlossen ist, was einer unendlich grossen Lastimpedanz gleichkommt.

Die Einkopplung des Seitenzweiges erfolgt mit der Matrix:

$$\begin{bmatrix} 1 & 0 \\ 1/Z_{Ein}(\omega) & 1 \end{bmatrix}.$$

Die Berechnung der Übertragungsfunktion des gesamten Systems wird von mehreren Faktoren beeinflusst:

- der Flächenfunktion entlang des Sprechtraktes
- der Position der Schallquelle
- der Art der Schallquelle (Schallfluss- oder Schalldruck-Quelle)
- der Position des Gaumensegels (Einkopplung des Nasaltraktes)
- der Art der Phonation (glottale Öffnungsfläche)

Es ist sinnvoll, den Sprechtrakt für die Berechnungen in zwei Bereiche aufzuteilen, nämlich in einen Bereich vor der Quelle und in einen Bereich hinter der Quelle. Durch die Multiplikation der Elemente, ausgehend von der Quelle in beide Richtungen, können die Bereiche zu den Kettenmatrizen K_v und K_h mit den Eingangsimpedanzen $Z_{Ein,v}$ und $Z_{Ein,h}$ zusammengefasst werden. Allfällige Seitenzweige werden direkt an der passenden Stelle durch ihre Kopplungsmatrizen integriert.

Die Übertragungsfunktion für eine Schalldruckquelle P_0 innerhalb des Netzwerks, wie in Abbildung 3.11 veranschaulicht, lautet:

$$\frac{P_{Aus}}{P_0} = \frac{Z_L}{K_{v,21} \cdot Z_L + K_{v,22}} \cdot \frac{1}{Z_{Ein,h} + Z_{Ein,v}}. \tag{3.60}$$

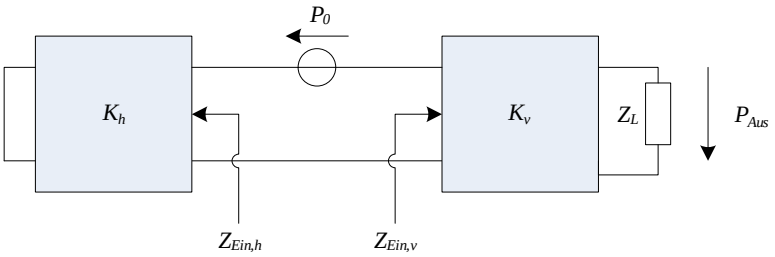


Abbildung 3.11: *Vereinfachtes Netzwerk mit einer Schalldruckquelle.*

Im Falle einer Schallflussquelle U_0 befindet sich diese in der Mitte eines Rohrabschnittes parallel zur Querimpedanz. Die Rohrabschnitte dahinter und die Abschnitte davor Richtung Abstrahlungsöffnung werden wiederum zusammengefasst. Damit kann die Übertragungsfunktion bestimmt werden zu:

$$\frac{P_{Aus}}{U_0} = \frac{Z_L}{K_{v,21} \cdot Z_L + K_{v,22}} \cdot \frac{Z_q \cdot (Z_l + Z_r + Z_q + Z_{Ein,h} + Z_{Ein,v})}{(Z_l + Z_q + Z_{Ein,h}) \cdot (Z_r + Z_q + Z_{Ein,v})}. \tag{3.61}$$

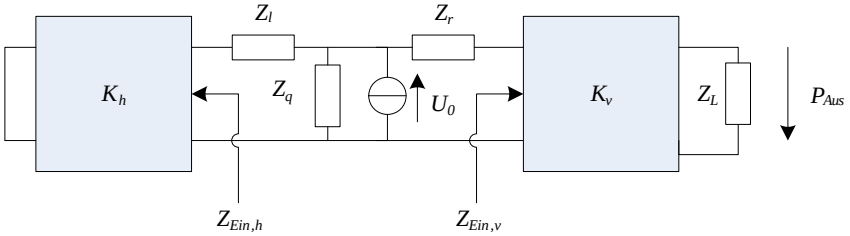


Abbildung 3.12: *Vereinfachtes Netzwerk mit einer Schallflussquelle.*

Für die verschiedenen Lautklassen ergeben sich so typische Konfigurationen.

Der einfachste Fall sind **nicht nasalierte Vokale**. Das Gaumensegel ist bis an die Rachenrückwand angehoben und trennt somit den Nasaltrakt vom Rachen- und Mundtrakt. Durch die stimmhafte Phonation besitzt die Glottis eine sehr hohe Impedanz, so dass der subglottale Trakt kaum einen Einfluss hat. Beides wird erreicht, indem die Öffnungsflächen der Eingangselemente zum Nasaltrakt bzw. zum subglottalen Trakt sehr klein gewählt werden. Unter der Annahme, dass sich die Schallflussquelle U_0 im ersten Rohrabschnitt oberhalb der Glottis befindet, kann die Übertragungsfunktion mit Gleichung (3.61) bestimmt werden.

$$H_M = \frac{P_M}{U_0}.$$

Nasale sind dadurch charakterisiert, dass im Mundraum an einer bestimmten Position ein vollständiger Verschluss vorhanden ist während das Gaumensegel abgesenkt ist. Die Schallabstrahlung erfolgt deshalb hauptsächlich über die Nasenlöcher. Die Übertragungsfunktion wird ebenfalls mit Hilfe der Gleichung (3.61) berechnet, wobei der Mundraum bis zum Verschluss zusätzlich als Seitenzweig eingekoppelt wird.

$$H_N = \frac{P_N}{U_0}.$$

Bei **nasalierten Vokalen** ist das Gaumensegel gesenkt. Es erfolgt daher eine Schallabstrahlung sowohl über die Lippen als auch über die Nasenlöcher. Dazu werden zwei Übertragungsfunktionen für den Schallpfad durch den Mund- (H_M) und durch den Nasenraum (H_N) benötigt. Beim Pfad durch den Mundraum wird der Nasaltrakt als Seitenzweig eingekoppelt und umgekehrt. Die gesamte Übertragungsfunktion des Sprechtraktes ergibt sich aus der Summe dieser zwei Übertragungsfunktionen:

$$H = H_M + H_N = \frac{P_M + P_N}{U_0}.$$

Frikative zeichnen sich durch eine Engstelle im Mundraum aus. Diese Engstelle hat mehrere Effekte zur Folge, die eine Bestimmung der Übertragungsfunktion erheblich erschwert. Gemäss NARAYANAN und ALWAN [35] lassen sich bei Frikativen drei wesentliche Geräuschquellen unterscheiden. Die stimmhafte Anregung kann analog zu den Vokalen durch eine Schallflussquelle oberhalb der Glottis modelliert werden. Die Verwirbelungen am Ausgang der Konstriktion führen zu zufälligen Geschwindigkeitsschwankungen entlang des Schallpfades Richtung Lippen. Diese Geschwindigkeitsschwankungen entsprechen verteilten akustischen Monopolquellen und lassen sich durch eine konzentrierte Schallflussquelle am Ausgang der Engstelle modellieren. Eine weitere Geräuschquelle ist dort zu finden, wo der Luftstrahl auf Hindernisse trifft, wie beispielsweise die Vokaltraktwände oder die Zähne. In der Folge kommt es an diesen Stellen zu Schalldruckschwankungen die wie akustische Dipolquellen wirken. Im akustischen Netzwerk können die Dipolquellen durch Schalldruckquellen simuliert werden. Die verschiedenen Schallquellen weisen zudem eine unterschiedliche spektrale Zusammensetzung auf, die je nach Ort der Konstriktion oder Hindernis variiert.

Beiden Quellen ist gemeinsam, dass ab einer bestimmten Knickfrequenz ein spektraler Abfall zu beobachten ist. Die Dipolquellen weisen diesen spektralen Abfall auf beiden Seiten der Knickfrequenz auf. Typische Werte für den spektralen Abfall bei den Dipolquellen reichen von -3 dB pro Oktave bei labiodentalen, über -6 dB bei alveolaren Frikativen bis hin zu -12 dB pro Oktave bei postalveolaren Frikativen. Die Knickfrequenzen liegen zwischen 2 kHz und 6 kHz. Es wird angenommen, dass die Knickfrequenz abhängig ist von der Querschnittsfläche der Konstriktion, der Strömungsgeschwindigkeit, dem Abstand zwischen dem Hindernis und der Engstelle und den Abmessungen des Hindernisses. Die genauen Ursachen und Zusammenhänge, welche für diese Parameter bestimmend sind, sind bisher jedoch noch weitgehend unklar.

Bei den Monopolquellen liegt der spektrale Abfall etwa bei -12 dB pro Oktave und die Knickfrequenz um 1.1 kHz. Hier konnten weniger Variationen festgestellt werden. Die Amplitude der Dipolquelle ist ca. 10 mal grösser als die Amplitude der Monopolquelle. Bei stimmhafter Anregung dominiert die Schallflussquelle oberhalb der Glottis.

Der Grenzwert für die Querschnittsfläche, bei der die turbulente Anregung einsetzt, wurde auf $A_c = 0.3$ cm^2 festgesetzt und den Parametern der Quellen wurden feste Wert zugewiesen:

Für beide Quellen wird ein Tiefpassfilter 2. Ordnung (-12 dB/Oktave) verwendet, wie in [35] vorgeschlagen. Der spektrale Abfall unterhalb der Knickfrequenz bei Dipolquellen wird durch einen zusätzlichen Hochpassfilter 2. Ordnung modelliert. Es werden eine Mono- und eine Dipolquelle eingesetzt um die Geräuschquellen nach der Konstriktion zu simulieren. Die Knickfrequenz der Monopolquelle liegt bei 1.1 kHz, die der Dipolquelle bei 4 kHz. Die Position der Monopoquelle liegt im Rohrabschnitt unmittelbar nach der Konstriktion. Die Dipolquelle wird ca. 0.9 cm vor den Lippen im Mundraum platziert , was in etwa mit der Position der Zähne

übereinstimmt, oder direkt an den Lippen, falls die Verengung dort ist.

Zudem müssen die entstehenden Druckverluste nach der Konstriktion berücksichtigt werden. Dies kann durch einen zusätzlichen Längswiderstand gemäss Gleichung (3.26) jeweils direkt am Ausgang der Engstelle des Vokaltraktes und oberhalb der Glottis geschehen. Die Öffnungsfläche an der Glottis wird bei stimmloser Anregung auf 0.3 cm^2 gesetzt. Damit erhöht sich auch der Einfluss der subglottalen Luftwege erheblich im Vergleich zu den Vokalen.

Die gesamte Übertragungsfunktion setzt sich aus der Summe der Übertragungsfunktionen der einzelnen Quellen zusammen:

$$H_{gesamt}(\omega) = \sum_i H_i(\omega) \cdot Q_i(\omega).$$

Q_i entspricht dabei dem Quellenspektrum und H_i der Übertragungsfunktion der Quelle i.

4 Artikulatorisches Modell des Sprechtraktes

Als artikulatorisches Modell wird ein Modell des Sprechtraktes bezeichnet, welches zum Ziele hat, die Form und allenfalls die Bewegungsabläufe des Sprechapparates zu beschreiben. Solche Modelle wurden insbesondere im Rahmen der artikulatorischen Sprachsynthese entwickelt. Obwohl ganz unterschiedliche Modellansätze existieren, beschreiben alle im Wesentlichen die supraglottale Form des Sprechapparates. Nach KRÖGER [21] lassen sich die verschiedenen methodischen Ansätze in fünf Kategorieren unterteilen:

- **Geometrisch orientierte Ansätze:** Sie beschreiben die Form des Vokaltraktes durch eine direkte Parametrisierung, wie zum Beispiel der Vorgabe der Querschnittsflächen eines Röhrenmodells. Vertreter dieser Kategorie sind das Zwei-Sektionen-Modell und das Vier-Sektionen-Modell von FANT [10]. Für die Parametrisierung des Zwei-Sektionen-Modells reichen der Quotient der Querschnittsflächen und der Quotient der Längen.

- **Effektorisch orientierte Ansätze:** Die Beschreibung der Vokaltraktform erfolgt indirekt über die Parametrisierung der Artikulatoren. Die meisten Modelle beschränken sich auf eine zweidimensionale Beschreibung in der mediosagittalen Ebene. Aus den Konturlinien in dieser Ebene wird dann die Flächenfunktion entlang des Vokaltraktes abgeleitet. Modelle dieser Art stammen zum Beispiel von MERMELSTEIN [30] oder COKER [5]. Basierend auf dem Modell von MERMELSTEIN wurde von BIRKHOLZ [3] ein dreidimensionales Modell des Vokaltraktes entwickelt. Seine Vorteile liegen vor allem darin, dass auch Variationen in lateraler Richtung möglich sind, wie sie zum Beispiel durch Senken und Heben der Zungenränder entstehen und realistischere Querschnittsformen ermittelt werden können.

- **Statistisch orientierte Ansätze:** Die statistischen Modelle beschreiben die Form des Vokaltraktes als eine Kombination artikulatorischer Parameter, die auf der Grundlage einer statistischen Auswertung von Messdaten gewonnen wurden. Modelle dieser Art sind zum Beispiel von MAEDA [26] und MEYER *et al.* [32] entwickelt worden.

- **Physiologisch orientierte Ansätze:** Sie beruhen auf einer physiologischen oder biomechanischen Simulation des Sprechapparates oder Teilen davon mit Hilfe von numerischen Verfahren wie der Finite-Elemente-Methode. Ein Modell zur biomechanischen Simulation der Zunge und der Lippen stammt von DANG und HONDA [8].

- **Akustisch orientierte Ansätze:** Bei den akustisch orientierten Ansätzen werden direkte Zusammenhänge zwischen den Veränderungen der Form des Vokaltraktes und den daraus resultierenden akustischen Merkmalen gesucht. MYRATI *et al.* [34] haben ein Modell basierend auf distinktiven Regionen vorgestellt, welches die Auswirkungen lokaler Änderungen der Querschnittsflächen auf die Formantfrequenzen beschreibt.

Für die geometrische Modellierung des Vokaltraktes in der vorliegenden Arbeit kamen grundsätzlich effektorisch orientierte als auch statistische Modelle in Frage.

Ein wichtiger Aspekt bei den effektorisch orientierten Ansätzen ist, dass die Anzahl der Freiheitsgrade willkürlich gewählt werden kann. So wird beispielsweise im Modell von MERMELSTEIN der Zungenkörper durch einen Kreis mit einem fixen Radius von 2 cm und einem

beweglichen Mittelpunkt, welcher durch zwei Koordinaten festgelegt wird, repräsentiert. Eine Abwägung, welche Parameter ohne Redundanz einzuführen nötig und welche überflüssig sind, ist schwierig. Auch die Wertebereiche, in denen sich die Parameter des Modells bewegen dürfen, müssen bestimmt werden. Sie sollten einerseits alle möglichen Formen des Vokaltraktes, die ein Mensch in der Lage zu erzeugen ist, abdecken und andererseits aber keine anatomisch unmöglichen Formen zulassen.

Bei statistischen Modellen hingegen wird die Anzahl benötigter Freiheitsgrade aus einer Auswertung von Messdaten bestimmt. Damit wird sichergestellt, dass die Parametrisierung die beobachtbare Varianz in den Messdaten erklären kann. Dies ist zugleich auch der Schwachpunkt der statistischen Modelle. Sie sind auf umfangreiche und repräsentative Messdaten angewiesen. Die Erhebung solcher Daten ist mit einem grossen Aufwand verbunden, weshalb sie zumeist nur von wenigen Sprechern stammen. Eine Übertragung auf andere Sprecher oder auf eine Fremdsprache (andere Laute) ist daher nur bedingt möglich.

4.1 Lineares statistisches Artikulatormodell

Da die statistischen Modelle einen guten Kompromiss zwischen Komplexität und Flexibilität darstellen, wurde das Modell von MAEDA ausgewählt. Es bietet darüber hinaus einen entscheidenden Vorteil bei der Schätzung der Modellparameter für einer gegebene Flächenfunktion.

Das Modell von MAEDA basiert auf der Auswertung von Röntgenfilmen, die mit einer Bildfrequenz von 50 Hz aufgezeichnet wurden. In die Auswertung flossen die Daten von etwas mehr als 500 Einzelaufnahmen aus 10 französischen Sätzen einer Sprecherin. Für die Datengewinnung wurde ein halbpolares Koordinatensystem mit dem harten Gaumen als Referenzpunkt über die Röntgenaufnahmen, die die mediosagittale Ebene zeigen, gelegt. Es wurden dann entlang der Konturen des Vokaltraktes 32 Messwerte ermittelt, die sich in 3 Gruppen ordnen lassen (vgl. Abbildung 4.1):

- Die Form der Lippen wird durch 3 Werte repräsentiert, die die Öffnung, die Vorstülpung und die Rundung der Lippen angeben.
- Der Bereich der Zunge wird mit 25 Werten entlang des Koordinatensystems beschrieben, die dem Durchmesser zwischen der Zungenkontur und der fixen Kontur des Gaumens entsprechen. In den zwei linearen Bereichen liegen 0.5 cm zwischen zwei Messwerten und im polaren Bereich beträgt der Winkel 11 Grad.
- Der Kehlkopfbereich wird durch die linearen Koordinaten seiner unteren zwei Eckpunkte beschrieben. Dies ist nötig, da ihre Position während des Sprechens durch Heben und Senken des Kehlkopfes variieren kann und deshalb im allgemeinen Fall nicht auf den Gitternetzlinien zu liegen kommt.

Die obere bzw. hintere Kontur, die den Gaumen und die Wand des Rachens darstellt, wird als fix angenommen und die Position der unteren Kontur relativ zu dieser ermittelt. Alle Messwerte wurden zusätzlich bezüglich ihrem Mittelwert und ihrer Varianz normalisiert. Auf der Grundlange dieser Daten wurden mit einem Faktorenanalyse-Verfahren 7 Kontrollparameter bestimmt. Diese linearen Parameter wurden so gewählt, dass sie auch zugleich eine artikulatorische Interpretation zulassen:

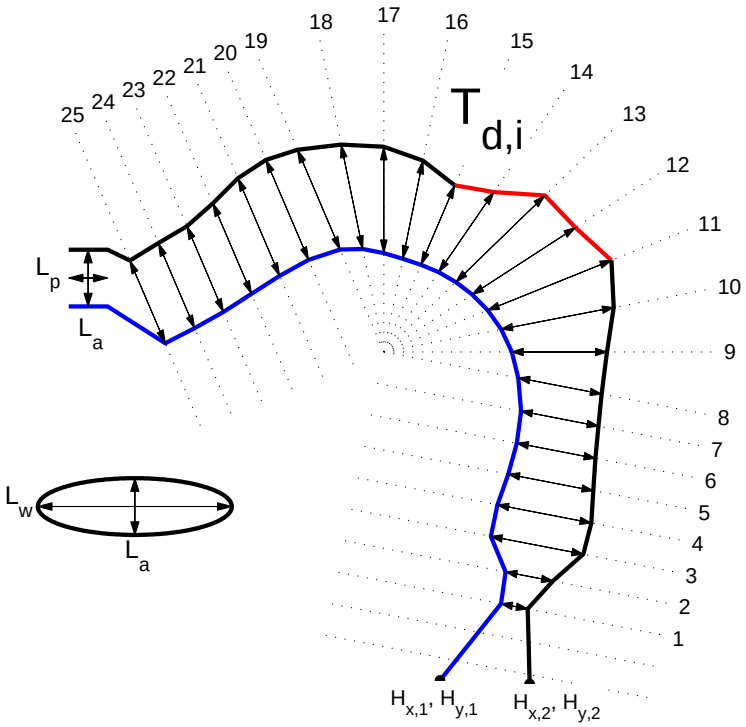


Abbildung 4.1: *Werte der Vokaltraktkontur des linearen Artikulatormodells.*

- JW : Position des Kiefers (offen/geschlossen)
- TP : Position des Zungenrückens (vorne/hinten)
- TS : Form des Zungenrückens (gewölbt/flach)
- TT : Position der Zungespitze (gesenkt/gehoben)
- LA : Lippenöffnung (offen/geschlossen)
- LP : Lippenvorstülpung (gerundet/gespreizt)
- LH : Höhe des Kehlkopfes (gesenkt/gehoben)

In Abbildung 4.2 ist die artikulatorische Bedeutung der Parameter dargestellt.

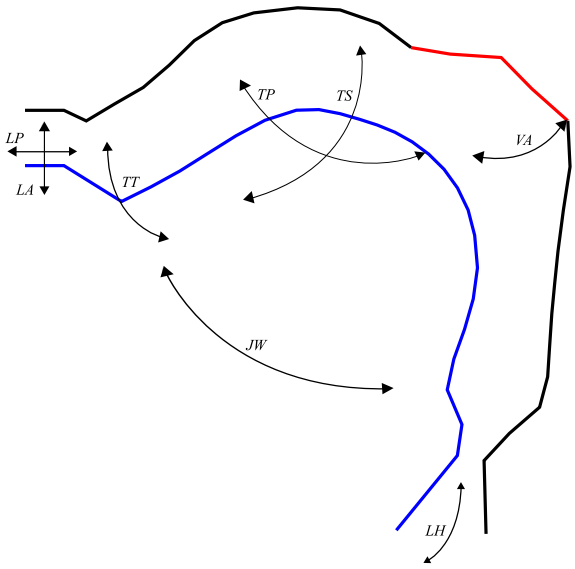


Abbildung 4.2: *Parameter des artikulatorischen Modells und ihr Effekt auf die Vokaltraktform.*

Die Vokaltraktkontur bei einer gegebenen Parameterkonfiguration wird bestimmt durch das Lösen der linearen Gleichung:

$$F \cdot p = b. \tag{4.1}$$

Schliesslich muss noch die Normalisierung mit

$$z_i = b_i \cdot \sigma_i + \mu_i \tag{4.2}$$

rückgängig gemacht werden, wobei σ_i die Standardabweichung und μ_i der Mittelwert ist.

Die vollständige lineare Gleichung sieht folgendermassen aus:

$$\begin{bmatrix} F_{1,1} & & & & F_{1,5} & F_{1,6} & \\ F_{2,1} & & 0 & & F_{2,5} & F_{2,6} & 0 \\ F_{3,1} & & & & F_{3,5} & F_{3,6} & \\ F_{4,1} & F_{4,2} & F_{4,3} & F_{4,4} & & & \\ \vdots & \vdots & \vdots & \vdots & & 0 & \\ F_{28,1} & F_{28,2} & F_{28,3} & F_{28,4} & & & \\ F_{29,1} & & & & & & F_{29,7} \\ F_{30,1} & & & 0 & & & F_{30,7} \\ F_{31,1} & & & & & & F_{31,7} \\ F_{32,1} & & & & & & F_{32,7} \end{bmatrix} \cdot \begin{bmatrix} JW \\ TP \\ TS \\ TT \\ LA \\ LP \\ LH \end{bmatrix} = \begin{bmatrix} L_p \\ L_a \\ L_w \\ T_{d,1} \\ \vdots \\ T_{d,25} \\ H_{x,1} \\ H_{y,1} \\ H_{x,2} \\ H_{y,2} \end{bmatrix}. \tag{4.3}$$

Sämtliche Werte werden vom Parameter des Kiefers beeinflusst und zusätzlich durch die Parameter der jeweiligen Regionen (Lippen, Zunge, Kehlkopf).

Die gültigen Wertebereiche der Parameter liegen zwischen -3 und +3. Die ersten 4 Faktoren zusammen genügen bereits, um annähernd 90% der beobachteten Varianz bezüglich der gemessenen Konturen zu erklären. 43% der Varianz erklärt die Position des Zungenrückens, 23% die Form des Zungenrückens, 15% die Position des Kiefers und die Position der Zungenspitze weitere 7%. Der geringe Einfluss der Parameter der Lippen und des Kehlkopfes liegt an ihrer begrenzten räumlichen Wirkung.

Um die Einkopplung des Nasaltraktes zu steuern, ist das ursprüngliche Modell um einen weiteren Parameter VA erweitert worden, welcher die Öffnung der velopharyngealen Pforte beschreibt. Der gültige Wertebereich des Parameters ist derselbe wie bei den anderen Parametern und wird linear auf eine Fläche zwischen 0 bis 4 cm^2 abgebildet. Diese Fläche entspricht direkt der Querschnittsfläche des ersten Rohrabschnittes des Nasaltraktes. Die Querschnittsflächen der folgenden 3 Abschnitte werden linear zwischen dem Wert des ersten und des fünften Abschnittes des Nasaltraktes interpoliert. In einem letzten Schritt wird der Flächenzuwachs von den gegenüberliegenden Querschnittsflächen des Mundraumes subtrahiert. Damit wird berücksichtigt, dass die Querschnittsfläche im hinteren Mundraum durch das Absenken und Anheben des Velums auch eine Änderung erfährt. Dies betrifft die Abschnitte mit den Bezeichnungen $T_{d,11}$ bis $T_{d,15}$, welche in der Abbildung 4.1 rot eingefärbt sind.

Wird eine konstante Länge des Vokaltraktes vorausgesetzt, wie es für das verlustlose zeitdiskrete Rohrmodell der Fall ist, müssen die Parameter der Lippenvorstülpung und der Höhe des Kehlkopfes fixiert werden. Die Länge des ersten und des letzten Abschnittes werden konstant auf 0.5 cm gesetzt. Das lineare Gleichungssystem kann dann vereinfacht werden zu [20]:

$$\begin{bmatrix} F_{1,1} & & & & F_{1,5} \\ F_{2,1} & F_{4,2} & F_{4,3} & F_{4,4} & \\ \vdots & \vdots & \vdots & \vdots & \\ F_{26,1} & F_{26,2} & F_{26,3} & F_{26,4} & \\ F_{27,1} & & & & \\ F_{28,1} & & & 0 & \\ F_{29,1} & & & & \\ F_{30,1} & & & & \end{bmatrix} \cdot \begin{bmatrix} JW \\ TP \\ TS \\ TT \\ LH \end{bmatrix} = \begin{bmatrix} L_a \\ T_{d,1} \\ \vdots \\ T_{d,25} \\ H_{x,1} \\ H_{y,1} \\ H_{x,2} \\ H_{y,2} \end{bmatrix}. \tag{4.4}$$

Um bei einer gegebenen Vokaltraktkontur die am besten passenden Parameterwerte zu erhalten, gilt es folgendes Minimierungsproblem zu lösen:

$$\min_{p} \|Fp - b\|_2. \tag{4.5}$$

Mit der Methode der kleinsten Quadrate können die optimalen artikulatorischen Parameter $\hat{p}$ auf einfache Weise ermittelt werden:

$$\hat{p} = (F^T F)^{-1} F^T b. \tag{4.6}$$

4.2 $\alpha\beta$-Transformation

Das linearen Modell von MAEDA liefert für eine Parameterkonfiguration die Vokaltraktkonturen im mediosagitallen Schnitt. Aus den Konturen können die Distanzen zwischen einem Artikulator und der Sprechtraktwand oder zwischen zwei Artikulatoren als Funktion vom Abstand zur Glottis abgeleitet werden. Für die Abtastung der Distanzen entlang des Vokaltraktes sind mehrere Varianten denkbar. Eine Möglichkeit besteht in der Definition der Distanz als Abstand zwischen den zwei Schnittpunkten einer Gitternetzlinie mit der inneren und der äusseren Kontur. Die Abstände zwischen zwei benachbarten Gitternetzlinien werden in diesem Fall entlang des gesamten Vokaltraktes als konstant angenommen. Dies stellt jedoch insbesondere im Bereich der Krümmung eine grobe Approximation dar. Die andere Möglichkeit ist, die zweidimensionale Vokaltraktkontur mit Hilfe der Gitternetzlinien in einzelne Vierecke zu unterteilen. Die Vierecke werden in einem nächsten Schritt durch Rechtecke mit der gleichen Fläche ersetzt, so dass man ein gerades Modell erhält. Die Länge der Rechtecke entsprechen den mediosagittalen Distanzen und die Breite der Rechtecke der Länge der Rohrabschnitte, wie in Abbildung 4.3 illustriert.

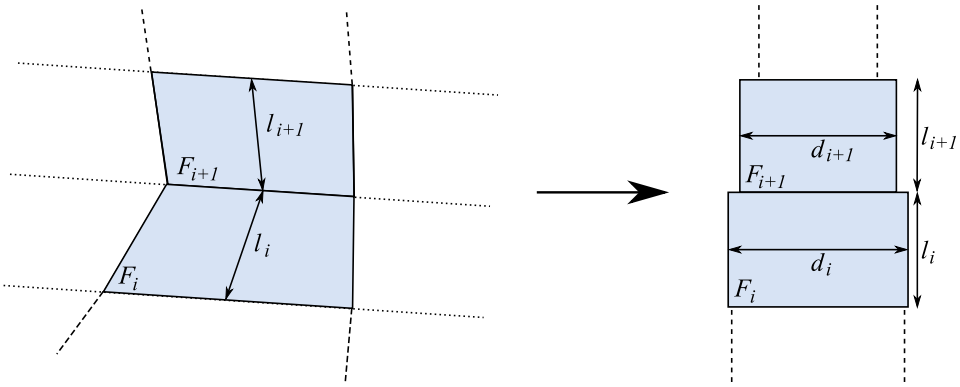


Abbildung 4.3: *Von der Vokaltraktkontur zur Distanzfunktion.*

Es stellt sich im Weiteren die Frage, wie der Zusammenhang zwischen der ermittelten Distanzfunktion und der Flächenfunktion entlang des Vokaltraktes ist. Die naheliegendste Lösung ist es, die Distanz als Durchmesser eines Kreises zu sehen und die Kreisfläche zu berechnen. Vergleiche zwischen gemessenen Querschnittsflächen und berechneten Kreisflächen zeigen jedoch, dass die Abweichungen erheblich sind. Die Gründe dafür sind die stark unterschiedlichen Querschnittsformen entlang des Vokaltraktes. In Abbildung 4.4 sind die Strukturen an

14 Positionen entlang des Vokaltraktes mit Hilfe der Magnetresonanztomographie sichtbar gemacht worden. Man erkennt gut an diesem Beispiel, dass die Querschnittsform im Allgemeinen eher einer Ellipse entspricht und man daher nicht unmittelbar vom Durchmesser auf die Querschnittsfläche schliessen kann.

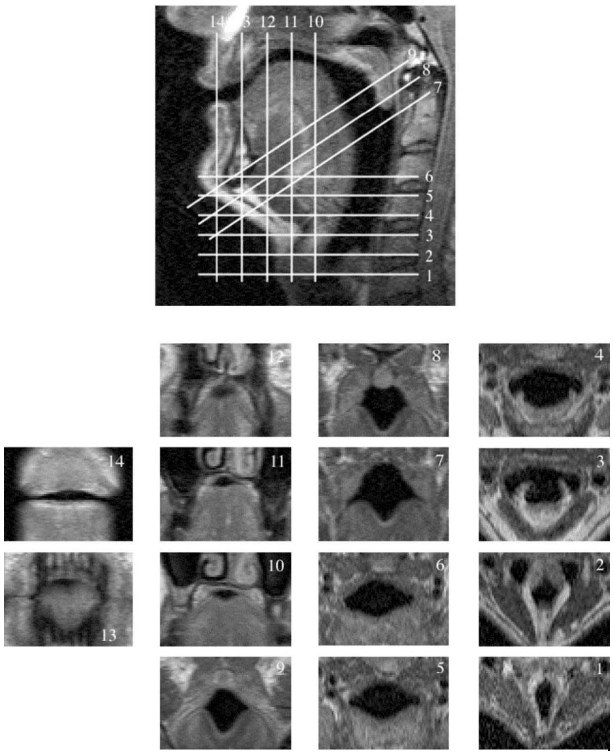


Abbildung 4.4: *14 Querschnittsbilder entlang des Vokaltraktes bei der Äusserung eines [i] (Quelle: [46]).*

HEINZ und STEVENS [15] stellten fest, dass die Beziehung zwischen mediosagitaller Distanz und der Querschnittsfläche besser durch einen exponentiellen Zusammenhang beschrieben werden kann:

$$A(x) = \alpha \cdot d(x)^{\beta}. \tag{4.7}$$

$d(x)$ ist die mediosagitalle Distanz als Funktion des Abstandes von der Glottis und $A(x)$ ist die Querschnittsfläche. α und β sind die zwei Parameter der Transformation. Daher wird sie auch als $\alpha\beta$-Transformation oder als $\alpha\beta$-Modell bezeichnet. Welches die optimalen Werte für α und β sind wurde in diversen Arbeiten untersucht. Die Vorschläge reichen von konstanten Werten über den gesamten Vokaltrakt bis hin zur Definition der Werte in Abhängigkeit vom Glottisabstand, der mediosagitallen Distanz, der Anregungsart und weitere.

Eine vergleichende Studie von SOQUET *et al.* [46], bei der auch andere Modell-Ansätze als das $\alpha\beta$-Modell berücksichtigt wurden, hat ergeben, dass eine Potenzfunktion am besten geeignet ist. Von 4 unterschiedlichen Definitionen der α- und β-Werte wurden die besten Ergebnisse mit den Transformationsparametern von MAEDA erzielt. Diese sind für die Gitternetzlinien, wie sie in Abbildung 4.1 zu sehen sind, definiert. Darüber hinaus bietet diese Definition der Parameter den Vorteil, dass eine Umkehrung der Transformation möglich ist:

$$A_i = \alpha_i \cdot d_i^{\beta_i} \Leftrightarrow d_i = \left(\frac{A_i}{\alpha_i}\right)^{\frac{1}{\beta_i}}. \tag{4.8}$$

Die von MAEDA vorgeschlagenen $\alpha\beta$-Werte entlang des Vokaltraktes sind Abbildung 4.5 dargestellt. In Abbildung 4.6 sieht man an einem Beispiel, wie sich die $\alpha\beta$-Transformation auf die Berechnung der Querschnittsflächen auswirkt. Die Unterschiede zu den Kreisflächen sind zum Teil beträchtlich.

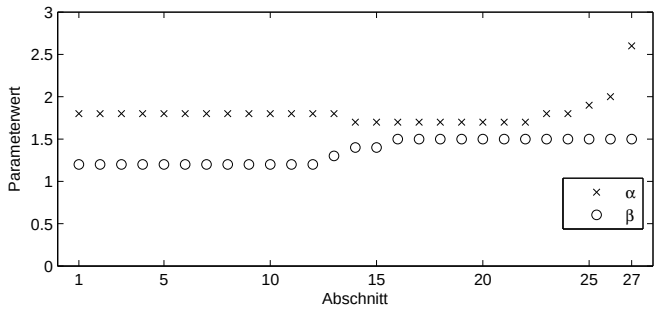


Abbildung 4.5: *Werte der Transformationsparameter des $\alpha\beta$-Modell nach* MAEDA *entlang der Gitternetzlinien des Artikulatormodells.*

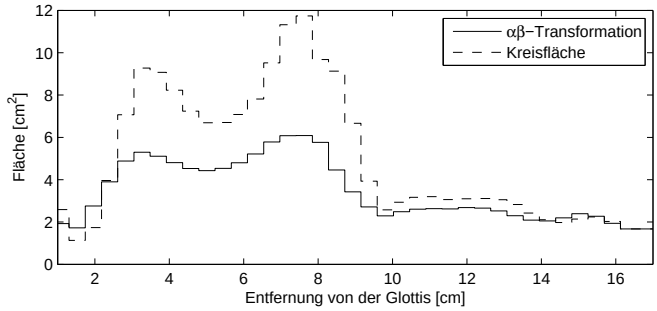


Abbildung 4.6: *Vergleich zwischen den Querschnittsflächen aus der $\alpha\beta$-Transformation und der direkten Berechnung der Kreisflächen aus den mediosagittalen Distanzen.*

Das artikulatorische Modell und das $\alpha\beta$-Modell von MAEDA unterteilen den Sprechtrakt in eine fixe Anzahl Abschnitte. Um eine Flächenfunktion mit einer abweichenden Unterteilung zu erhalten, können die diskreten Flächenwerte interpoliert werden. In dieser Arbeit wird dafür eine lineare Interpolation verwendet.

5 Parameterschätzung

Die unterschiedliche Komplexität der akustischen Modelle, wie sie im Kapitel 3 vorgestellt werden, erfordert angepasste Schätzverfahren. Die Aufgabe der Schätzverfahren ist es, die Parameter der akustischen Modelle so zu bestimmen, dass eine möglichst genaue Übereinstimmung zwischen Modell und Beobachtung erzielt wird. Dieses Kapitel beschreibt die zwei Methoden, welche für die Parameterschätzung des verlustlosen und des verlustbehafteten Rohrmodells eingesetzt werden.

5.1 Parameterschätzung des zeitdiskreten verlustlosen Rohrmodells

WAKITA [50] zeigte erstmals den direkten Zusammenhang zwischen den Koeffizienten eines inversen Filters und den Flächenverhältnissen eines verlustlosen Rohrmodells für die Schätzung der Vokaltraktflächen aus dem Sprachsignal. Die Koeffizienten des inversen Filters werden mit der Methode der linearen Prädiktion ermittelt.

Vorfilterung

Die Grundlage der inversen Filterung bildet das Quelle-Filter-Modell, welches das Sprachsignal S in Anregung G und Schallformung trennt. Die Schallformung geschieht hauptsächlich durch die Filterung H des Anregungssignals im Sprechtrakt und zusätzlich durch die Abstrahlungscharakteristik R an den Lippen bzw. an den Nasenlöchern. Im Spektralbereich kann das Sprachsignal nach dem Quelle-Filter-Modell dargestellt werden als

$$S = G \cdot H \cdot R. \tag{5.1}$$

Es wird vereinfachend angenommen, das weder G noch R Resonanzen im Frequenzbereich zeigen, sondern nur einen spektralen Abfall darstellen. H hingegen zeichnet sich durch Resonanzen aus, weist dafür aber keinen spektralen Abfall oder Anstieg über den gesamten Frequenzbereich auf. Das Spektrum G des Anregungssignals zeigt bei stimmhafter Anregung eine Tiefpass-Charakteristik. Die Abstrahlungscharakteristik der Lippen weist eine Hochpass-Charakteristik auf, die allerdings verglichen mit G deutlich geringer ausfällt. Ein spektraler Abfall kann durch reele Polstellen modelliert werden:

$$G \cdot R = \prod_{i=1}^{N} \frac{1}{1 - k_i \cdot z^{-1}}. \tag{5.2}$$

Soll nun der Einfluss von $G \cdot R$ minimiert werden, kann das Signal S mit den Nullstellen von $1/(G \cdot R)$ vorgefiltert werden (Präemphase). Die Polstellen können nacheinander mit Hilfe der linearen Prädiktion geschätzt werden, wobei abgebrochen wird, falls ein Koeffizient <0 ist, da dies einem spektralen Anstieg entsprechen würde. Oft wird auch ein vordefinierter fester Wert für ein Filter erster Ordnung benutzt (typische Werte liegen um $\sim$0.98):

$$\hat{H} = \prod_{i=1}^{N} (1 - k_i \cdot z^{-1}) \cdot S. \tag{5.3}$$

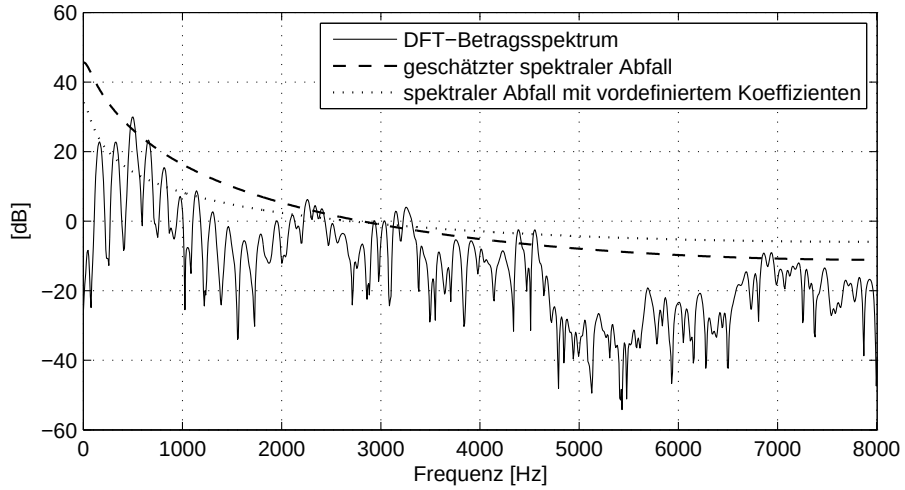


Abbildung 5.1: *DFT-Spektrum eines [a], geschätzter spektraler Abfall ($N = 3$) und spektraler Abfall für den festen Koeffizienten $k_1 = 0.98$ ($N = 1$).*

Lineare Systeme

Bei der linearen Prädiktion wird ein Abtastwert als Linearkombination vergangener Abtastwerte beschrieben. Das allgemeine Modell des Systems für eine Eingangssequenz $x[n]$ und einer Ausgangssequenz $y[n]$ lautet:

$$y[n] = \sum_{l=0}^{L} b_l \cdot x[n-l] - \sum_{k=1}^{K} a_k \cdot y[n-k]. \tag{5.4}$$

Die Transformation in den z-Bereich mit $z = e^{j\omega \cdot T}$, wobei T das Abtastintervall ist, führt zur Übertragungsfunktion $H(z)$ des Systems:

$$H(Z) = \frac{Y(z)}{X(z)} = \frac{\sum\limits_{l=0}^{L} b_l \cdot z^{-l}}{1 + \sum\limits_{k=1}^{K} a_k \cdot z^{-k}}. \tag{5.5}$$

In der Sprachverarbeitung wird jedoch häufig anstatt eines allgemeinen Pol-Nullstellen-Modells ein Nur-Pol-Modell verwendet um die Resonanzen des Sprechtraktes zu beschreiben. Die Differenzgleichung des Systems vereinfacht sich dann zu:

$$y[n] = b_0 \cdot x[n] - \sum_{k=1}^{K} a_k \cdot y[n-k], \tag{5.6}$$

mit der Übertragungsfunktion:

$$H(Z) = \frac{Y(z)}{X(z)} = \frac{b_0}{1 + \sum\limits_{k=1}^{K} a_k \cdot z^{-k}} = \frac{G}{A(z)}. \tag{5.7}$$

$A(z)$ wird als inverses Filter bezeichnet und G entspricht einem Verstärkungsfaktor.

Falls die Eingangssequenz $x[n]$ des Systems unbekannt ist, kann eine Schätzung des Signals $y[n]$ nur aus seinen vergangenen Werten erfolgen. Der Schätzwert $\hat{y}[n]$ für einen Prädiktor der Ordnung K lautet:

$$\hat{y}[n] = -\sum_{k=1}^{K} a_k \cdot y[n-k]. \tag{5.8}$$

Die Koeffizienten a_k des Prädiktors gelten dann als optimal, wenn die Energie des Prädiktionsfehlers $e[n]$ minimiert wird:

$$\min_{a_k} E\{e[n]^2\}, \quad e[n] = y[n] - \hat{y}[n]. \tag{5.9}$$

Die Filterkoeffizienten a_k können mit Hilfe der Autokorrelationsmethode oder der Kovarianzmethode bestimmt werden. Da die Autokorrelationsmethode eine stabile Lösung garantiert [29], wird diese Methode hier bevorzugt. Eine ausführliche Beschreibung der Methode und Lösung mittels des Levinson-Durbin-Algorithmus findet man in [28].

Unter der Voraussetzung dass $H(z)$ stabil ist, kann das inverse Filter $A(z)$ in Kreuzgliedform wie in Abbildung 5.2 realisiert werden. Die Reflexionskoeffizienten r_m der Kreuzglieder sind mit den Filterkoeffizienten der linearen Prädiktion durch folgende Rekursion verknüpft:

$$\begin{cases} a_0^{(m+1)} = 1 \\ a_m^{(m+1)} = r_m \\ a_j^{(m+1)} = a_j^{(m)} + r_m \cdot a_{m-j}^{(m)}, 1 \leq j \leq m \end{cases} \tag{5.10}$$

für $m = 1...K$ Rekursionsschritte und $a_0 = r_0 = 1$. Die Reflexionskoeffizienten können mit den Filterkoeffizienten und der Kurzzeitautokorrelation R_k berechnet werden:

$$r_m = -\frac{\sum_{k=0}^{m} a_k^{(m)} \cdot R_{m+1-k}}{\sum_{k=0}^{m} a_k^{(m)} \cdot R_k}. \tag{5.11}$$

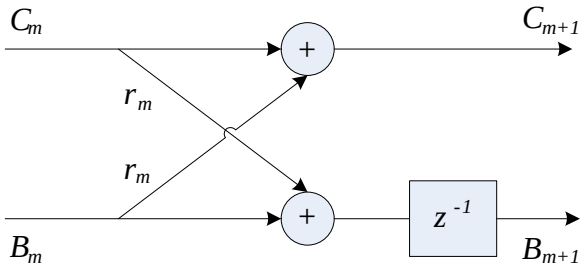


Abbildung 5.2: *Abschnitt eines Kreuzgliedfilters in FIR-Struktur.*

Die Rekursion kann für ein Kreuzgliedfilter in Matrixform geschrieben werden als:

$$\begin{bmatrix} C_{m+1}(z) \\ B_{m+1}(z) \end{bmatrix} = \begin{bmatrix} 1 & r_m \\ r_m \cdot z^{-1} & z^{-1} \end{bmatrix} \cdot \begin{bmatrix} C_m(z) \\ B_m(z) \end{bmatrix}, \tag{5.12}$$

wobei $C_m(z)$ die Übertragungsfunktion des Vorwärtsprediktors:

$$C_m(z) = \sum_{k=0}^{m} a_k^{(m)} \cdot z^{-k} \tag{5.13}$$

und $B_m(z)$ die Übertragungsfunktion des Rückwärtsprediktors ist:

$$B_m(z) = z^{-(m+1)} \cdot C_m(1/z). \tag{5.14}$$

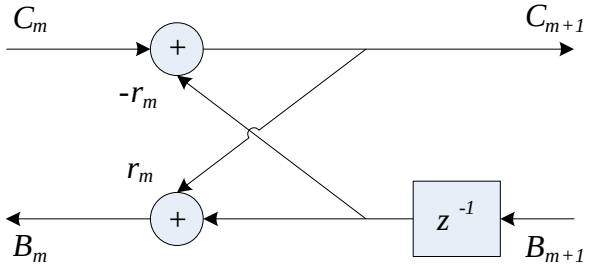


Abbildung 5.3: *Abschnitt eines Kreuzgliedfilters für ein Allpol-System.*

Vertauscht man Ein- und Ausgang des Filters, erhält man ein Allpol-System mit Kreuzgliedern gemäss Abbildung 5.3. Die dazugehörige Kettenmatrix

$$\begin{bmatrix} C_m(z) \\ B_m(z) \end{bmatrix} = \begin{bmatrix} 1 & r_m \cdot z^{-1} \\ r_m & z^{-1} \end{bmatrix} \cdot \begin{bmatrix} C_{m+1}(z) \\ B_{m+1}(z) \end{bmatrix} \tag{5.15}$$

besitzt dieselbe Form wie die Kettenmatrix in Gleichung (3.42) des verlustlosen zeitdiskreten Rohrmodells.

Damit können die Reflexionskoeffizienten beider Matrizen direkt miteinander verknüpft werden und die Verhältnisse der Querschnittsflächen des Rohrmodells als Funktion der Reflexionskoeffizienten der linearen Prädiktion ermittelt werden:

$$\xi_{m+1} = \frac{A_m - A_{m+1}}{A_m + A_{m+1}} = r_m \Rightarrow A_{m+1} = A_m \cdot \frac{1 - r_m}{1 + r_m}. \tag{5.16}$$

Da damit nur die Verhältnisse zwischen den Querschnittsflächen bestimmt sind, muss für eine Flächenfunktion die Grösse einer Fläche initialisiert werden. Dies erfolgt durch die erste Querschnittsfläche nach der Glottis, die für dieses Modell konstant auf 1 cm^2 gesetzt wird.

Damit sind alle Werkzeuge vorhanden, um aus einem Sprachsignal mit Hilfe der linearen Prädiktion die Parameter des verlustlosen Rohrmodells und daraus die Parameter des Artikulatormodells zu bestimmen. Der gesamte Ablauf ist in Abbildung 5.4 zusammengefasst.

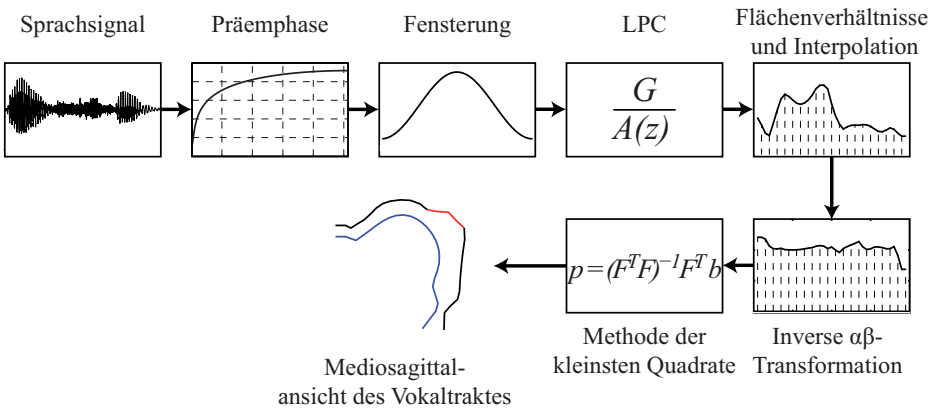


Abbildung 5.4: *Ablauf der Parameterschätzung des verlustlosen Rohrmodells.*

5.2 Parameterschätzung des verlustbehafteten Rohrmodells

Die Schätzung der Parameter des verlustbehafteten Rohrmodells gestaltet sich vergleichsweise schwierig. Durch die Berücksichtigung der Verluste im Vokaltrakt und durch die Seitenzweige kann das Verfahren von WAKITA nicht mehr angewandt werden.

Ein möglicher Ansatz besteht darin, die Parameter des Modells mittels eines allgemeinen gradientenbasierten Optimierungsverfahrens so zu bestimmen, dass ein definierter Fehler minimiert wird. Die Fehlerdefinition kann beispielsweise auf einem spektralen Abstandmass oder Vergleiche der Resonanzen von Modell und Sprachsignal beruhen. Der Rechenaufwand bei jedem Optimierungsschritt ist jedoch erheblich, wobei ein Grossteil auf die Matrixmultiplikationen zur Berechnung der Übertragungsfunktion entfällt. Eine weitere Schwierigkeit, die sich bei der Optimierung ergibt, steht im Zusammenhang mit der Definition der verschiedenen Geräuschquellen des Modells. Ein kleiner Unterschied der Querschnittsflächen entscheidet, ob es an einer Engstelle zu einer turbulenten Anregung kommt oder nicht und hat entsprechend grossen Einfluss auf die resultierende Übertragungsfunktion. Dies kann dazu führen, dass der Optimierungsvorgang in einem lokalen Minimum endet.

5.2.1 Codebuch

Aus diesen Gründen wurde ein Ansatz basierend auf einem vorberechneten Codebuch gewählt. Damit wird einerseits verhindert, dass nur lokale Minima gefunden werden und andererseits reduziert sich der wiederkehrende Rechenaufwand erheblich. Ein Codebuch für die Inversion akustischer in artikulatorischen Informationen besteht aus einer verlinkten Liste von Parametervektoren aus beiden Räumen.

Die Abbildung vom artikulatorischen Raum in den akustischen Raum beschreibt den Prozess der Spracherzeugung oder Synthese und kann in einer allgemeinen Form als eine nichtlineare Funktion g mehrerer Variablen angesehen werden

$$y = g(p), \tag{5.17}$$

wobei p den artikulatorischen Vektor beschreibt und y den akustischen.

Die Datengrundlage eines solchen Codebuches können z.B. geometrische Messdaten aus Röntgenfilmen oder MRI-Aufnahmen sein. Voraussetzung ist, dass das Verfahren gleichzeitig eine Tonaufnahme erlaubt. Die Beschaffung von Messdaten auf diesem Wege ist allerdings sehr umständlich. Als alternative Datenquelle kommen auch synthetische Vektorpaare (p, y) in Frage, die durch die Verknüpfung eines Artikulatormodells und einer akustischen Simulation des Sprechapparates erzeugt werden.

Für die Generierung der Vektorpaare wurden deshalb mit dem Artikulatormodell von MAEDA Vokaltrakformen erzeugt und mit der $\alpha\beta$-Transformation die Flächenfunktionen ermittelt. Das akustische Modell dient der Berechnung der entsprechenden Übertragungsfunktionen. Aus diesen können die akustischen Merkmale für das Codebuch extrahiert werden.

Die 8 Parameter p des Artikulatormodells wurden per Zufall innerhalb ihrer gültigen Bereiche gewählt, um zu gewährleisten, dass der artikulatorische Raum gleichmässig abgedeckt ist. Ein Nachteil dieser Vorgehensweise ist, dass bei Extremwerten auch unrealistische Vokaltraktformen erzeugt werden, die zum Beispiel einen vollständigen Verschluss im Rachenraum

aufweisen (Abbildung 5.5). Vokaltraktformen dieser Art wurden verworfen und die Vektorpaare nicht ins Codebuch aufgenommen.

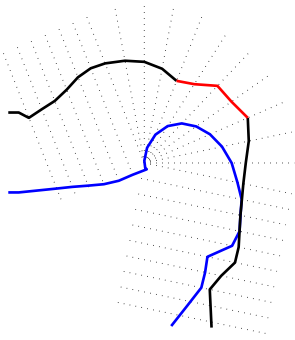


Abbildung 5.5: *Unrealistische Vokaltraktform bei extremen Parameterwerten.*

Als akustische Merkmale y wurden die Koeffizienten des Mel-Cepstrums (MFCC) verwendet. Die MFCC-Merkmalsvektoren haben sich vor allem bei der Spracherkennung als robuste Merkmale bewährt. Sie enthalten genügend Informationen über die Vokaltraktkonfiguration, unabhängig von der Anregungsart, und eignen sich daher sowohl für Vokallaute als auch für Konsonanten. Für die Berechnung der MFCCs wird eine Filterbank mit 24 Dreiecksfiltern verwendet und eine obere Grenzfrequenz von 5 kHz. Es werden die ersten 12 Koeffizienten des Mel-Cepstrums ermittelt ohne den Koeffizienten 0, der ein Mass für die Signalenergie darstellt.

Eine Schwierigkeit, die sich durch die Anzahl der artikulatorischen Parameter ergibt, ist, dass die Grösse des Codebuches stark mit dem gewünschten Abdeckungsgrad des artikulatorischen Raumes wächst. Bei 8 Parametern und jeweils 5 Werten pro Parameter sind bereits annähernd 400'000 Kombinationen möglich. Die Generierung als auch die spätere Durchsuchung des Codebuches nimmt entsprechend viel Zeit in Anspruch.

5.2.2 Dynamische Programmierung

Die Surjektivität der Abbildung $y = g(p)$, die sich auch in den Daten des Codebuches zeigt, macht die Notwendigkeit eines weiteren Kriteriums klar, um eine eindeutige Entscheidung bei der Wahl eines Vektorpaares treffen zu können. Was bisher noch keine Beachtung bei der Schätzung fand, ist die zeitliche Dimension.

Den Bewegungen der Artikulatoren sind, alleine schon wegen ihrer Trägheit, Grenzen gesetzt. So ist in den Bewegungsabläufen ein gewisses Mass an Kontinuität zu erwarten und keine sprunghaften Veränderungen innert kürzester Zeit. Im Rahmen der artikulatorischen Sprachsynthese sind diverse Ansätze erarbeitet worden, um die mögliche Dynamik des Systems zu beschreiben. MEYER *et al.* [32] nutzen hierfür angepasste Kalman-Filter. Das aufgabendynamische Modell von BROWMAN und GOLDSTEIN [4], dessen Grundelemente artikulatorische Gesten sind, modelliert die Dynamik der Gesten durch kritisch gedämpfte Masse-Feder-Systeme. Die Bewegungen der Artikulatoren werden durch untergeordnete dynamische Systeme beschrieben. Die Schwierigkeit dieser Ansätze besteht vor allem darin, dass sie eine genaue Kenntnis über die tatsächliche Dynamik der Systeme voraussetzen.

Ein anderer Ansatz basiert darauf, aus einer Sequenz mit jeweils mehreren möglichen Vokaltraktformen, die zu den akustischen Merkmalen passen würden, diejenigen auszuwählen, welche eine kontinuierliche Bewegung der Artikulatoren ergeben. Es wird keine explizite Kenntnis über die Dynamik der Artikulatoren vorausgesetzt, sondern die Lösung, welche den minimalsten Aufwand darstellt, gesucht. SCHROETER und SONDHI [40] nutzen für diese Aufgabe das Prinzip der dynamischen Progammierung, welches auch in dieser Arbeit verwendet wird.

Die Aufgabe kann man sich in Form eines Trellis-Diagramms wie in Abbildung 5.6 vorstellen. Aus dem Sprachsignal wird mit Hilfe einer Kurzzeitanalyse eine Sequenz $x_1 ... x_t ... x_T$ akustischer Merkmalsvektoren (MFCC) extrahiert. Ziel ist es, eine dazu passende optimale Sequenz von Vokaltraktformen aus den M Einträgen des Codebuches zu finden $(y_1, p_1) ... (y_t, p_t) ... (y_T, p_T)$. Eine vollständige Suche würde die Analyse von insgesamt M^T Kombinationen bedeuten. Da das Codebuch mehrere 10'000 Einträge enthalten kann und T für eine Sekunde Sprachsignal und einer Schrittweite von 10 ms bereits 100 beträgt, ist eine vollständige Suche nicht praktikabel. Durch dynamische Programmierung kann der Aufwand auf $T \cdot M^2$ reduziert werden. Das Auffinden der optimalen Gesamtlösung wird durch das Berechnen und Zusammensetzen von Lösungen kleinerer Teilprobleme vereinfacht.

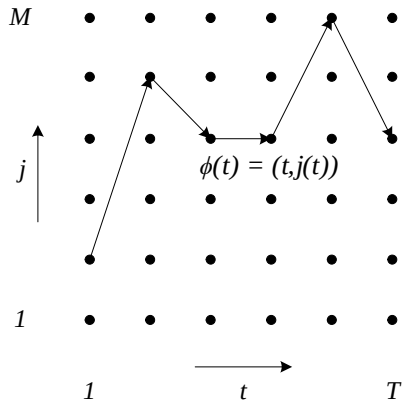


Abbildung 5.6: *Beispiel einer möglichen Warping-Kurve im Trellis-Diagramm.*

Ein ähnliches Problem stellt sich in der Spracherkennung bei der dynamischen Zeitanpassung zweier Signale und wird mit Hilfe des **DTW-Algorithmus** gelöst (siehe [37]). Durch eine Anpassung der Randbedingungen und der Berechnung der Distanzfunktion kann der DTW-Algorithmus auch zum Auffinden der optimalen Sequenz von Vokaltraktformen aus dem Codebuch dienen.

Die angepassten Randbedingungen lauten:

- Monotonie/lokale Kontinuität: Die Sequenz der extrahierten akustischen Merkmalsvektoren wird schrittweise durchlaufen. Die Warping-Kurve $\phi(t) = (t, j(t)), t = 1...T$ enthält daher genau T Punkte.
- Der Anfangs- und der Endpunkt kann ein beliebiger Codebucheintrag sein: $\phi(1) = (1, j(1))$ und $\phi(T) = (T, j(T))$.

Erlaubt sind alle Pfaderweiterungen, welche die Warping-Kurve entlang der Zeitachse um eine Einheit vorwärts bewegen. Die Distanzfunktion setzt sich aus einem akustischen und einem artikulatorischen Distanzmass zusammen.

Die akustische Distanz zwischen einem Merkmalsvektor x_t und y_j mit L Koeffizienten des Mel-Cepstrum wird mit der euklidischen Distanz berechnet:

$$d_{Aku}(t,j) = \sqrt{\sum_{l=1}^{L}(x_t(l) - y_j(l))^2}. \tag{5.18}$$

Die artikulatorische Distanz wird analog dazu aus den Merkmalsvektoren der Artikulatorparameter berechnet:

$$d_{Art}(j_t, j_{t-1}) = \sqrt{\sum_{l=1}^{8}(p_{j_t}(l) - p_{j_{t-1}}(l))^2}. \tag{5.19}$$

Die akkumulierte Distanz $D_A(t,j)$ im Punkt (t,j) kann nun für $t = 2...T$ rekursiv mit den Distanzmassen ermittelt werden:

$$D_A(t,j_t) = \min_{j_{t-1}\epsilon[1...M]}[D_A(t-1,j_{t-1}) + c_{Art}(t,t-1)\cdot d_{Art}(j_t,j_{t-1})] + c_{Aku}(t)\cdot d_{Aku}(t,j_t). \tag{5.20}$$

wobei jeweils die optimale Pfaderweiterung gespeichert wird. $c_{Art}(t,t-1)$ und $c_{Aku}(t)$ sind zusätzliche Gewichtsfunktionen für die Distanzmasse. Die Initialisierung der Matrix mit den akkumulierten Distanzen erfolgt mit $D(1,j) = d_{Aku}(1,j), j = 1...M$.

Beginnend bei der minimalen akkumulierten Distanz im Zeitpunkt T und den gespeicherten optimalen Pfaderweiterungen kann die optimale Warping-Kurve durch Backtraking ermittelt werden.

Die Gewichtsfunktionen wurden aufgrund folgender Überlegungen festgelegt:

- Je unterschiedlicher zwei aufeinanderfolgende akustische Merkmalsvektoren des Sprachsignals sind, desto grösser können auch die Unterschiede der Vokaltraktformen ausfallen.

$$c_{Art}(t,t-1) = \frac{c_0}{\sqrt{\sum_{l=1}^{L}(x_t(l) - x_{t-1}(l))^2}}. \tag{5.21}$$

  c_0 ist ein empirisch ermittelter zusätzlicher Skalierungsfaktor.

- In Zeitabschnitten mit einer geringen Signalenergie, wie zum Beispiel bei einer präplosiven Pause, ist der Informationsgehalt der extrahierten akustischen Merkmalsvektoren bzgl. der Vokaltraktform gering. Daher soll während dieser Zeit vor allem die artikulatorische Distanz für die Wahl ausschlaggebend sein. Dazu wird die Signalenergie E_t jedes Analyseabschnittes berechnet und $E_{max} = \max(E_1,...,E_T)$ ermittelt. Zwischen E_{max} und $E_{min} = E_{max} - E_{diff}$ nimmt der Gewichtsfaktor der akustischen Distanzen von 1 auf 0 ab. Die Differenz von $E_{diff} = 40$dB hat sich für die analysierten Signale bewährt.

$$c_{Aku}(t) = \sqrt{\frac{\max(E_t - E_{min}, 0)}{E_{diff}}}. \tag{5.22}$$

- Es besteht auch die Möglichkeit, zusätzliche Informationen aus dem Sprachsignal in die akustische Gewichtsfunktion einfliessen zu lassen. So können zum Beispiel anhand der Nulldurchgangsrate entweder eher Frikative oder Vokale bevorzugt werden oder die Kosten für Plosivlaute verringert werden je kleiner die Kurzzeitenergie des Signals wird.

5.2.3 Clustering

Die dynamische Programmierung verringert den Suchaufwand nach der optimalen Sequenz enorm. Bei einem Codebuch mit mehreren 10'000 Einträgen dauert die Suche trotzdem noch sehr lange. Um den Rechenaufwand weiter zu reduzieren, wird das Codebuch in Partitionen (Clusters) unterteilt.

Die generierten Vektorpaare $(p_i, y_i), i = 1...M$ des Codebuches werden in einem ersten Schritt in N_y akustische Partitionen unterteilt. Die Unterteilung erfolgt anhand der euklidischen Distanz der MFCC-Merkmalsvektoren mit Hilfe des K-means-Algorithmus [37]. Das Ergebnis der Unterteilung sind die Mittelwertvektoren $y_c = \{y_{c_1}...y_{c_{N_y}}\}$ der akustischen Partitionen. Die Merkmalsvektoren, die einer Gruppe zugeordnet wurden, weisen zu deren Mittelwertvektor eine kleinere Distanz auf als zu allen anderen Mittelwertvektoren.

Das Verfahren führt zu N_y Partitionen mit Vektorpaaren (p, y) deren akustische Merkmale innerhalb einer Gruppe ähnlich sind. Betrachtet man die dazugehörigen artikulatorischen Parameter, so ist diese Ähnlichkeit innerhalb einer Partition im Allgemeinen nicht gegeben. Diese Tatsache ist auf die nicht eindeutige Abbildung zurückzuführen, wodurch unterschiedliche Vokaltraktformen oder artikulatorische Parameterkonfigurationen ähnliche akustische Ergebnisse haben können.

Unter der Annahme, dass nur eine begrenzte Anzahl Grundkonfigurationen zu ähnlichen akustischen Merkmalen führen, wird eine weitere Unterteilung innerhalb der akustischen Gruppen in N_p artikulatorische Partitionen durchgeführt. Als Distanzmass dient die euklidische Distanz zwischen den artikulatorischen Parametervektoren.

Die Daten des Codebuches sind nun ingesamt in $N_y \times N_p$ Partitionen unterteilt worden und die Untergruppen haben die Eigenschaft, dass die akustischen und artikulatorischen Merkmalsvektoren einander ähnlich sind.

Anstatt direkt das gesamte Codebuch für die Suche mittels dynamischer Programmierung zu verwenden, wird zuerst die optimale Sequenz aus den Mittelwertvektoren der Partitionen gesucht. In einem zweiten Schritt wird die Suche nochmals durchgeführt, wobei der Suchraum auf alle Vektoren erweitert wird, die jeweils zur Partition des gewählten Mittelwertvektors gehören.

Der Ablauf der Parameterschätzung für das verlustbehaftete Rohrmodell ist in Abbildung 5.7 zusammengefasst.

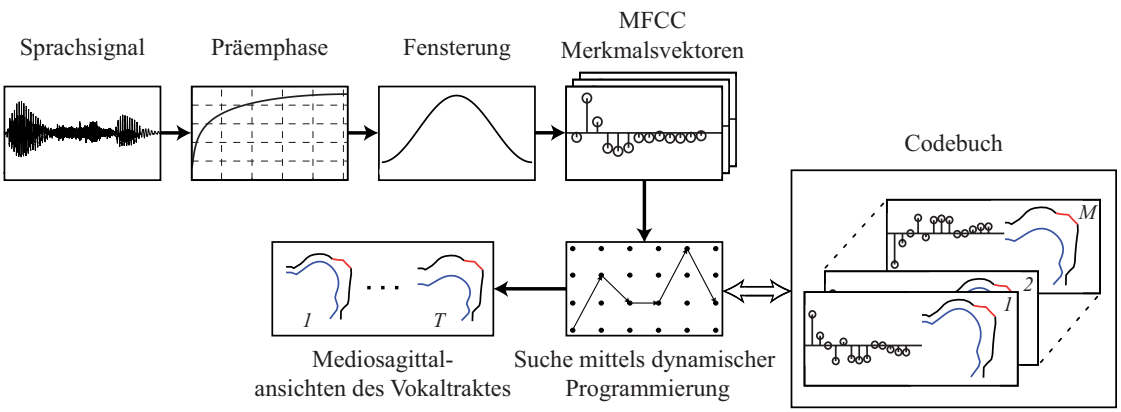


Abbildung 5.7: *Ablauf der Parameterschätzung des verlustbehafteten Rohrmodells.*

6 Evaluation der akustisch-artikulatorischen Inversion

Im diesem Kapitel werden einige experimentelle Ergebnisse der akustisch-artikulatorischen Inversion vorgestellt. Die Auswertung erfolgt anhand eines Diphon-Korpus und kurzen Sprachproben verschiedener Sprecherinnen und Sprecher. Ein direkter Vergleich zwischen den geschätzten und den tatsächlichen Vokaltraktformen kann mangels geeigneter artikulatorischer Daten nicht erfolgen. Stattdessen dienen die typischen Vokaltraktformen für die verschiedenen Laute, wie sie zum Beispiel in [13] zu finden sind oder durch das Vokalviereck beschrieben werden, als Orientierung.

Für einen direkten auditiven Vergleich können zudem synthetische Sprachsignale, basierend auf den geschätzten Parametern, generiert werden. Kurze stationäre Signale lassen sich mit der Übertragungsfunktion des akustischen Modells und einer Anregungsfunktion gewinnen. Dazu wird aus der Übertragungsfunktion die Impulsantwort des System berechnet. Um ein Abklingen der Impulsantwort sicherzustellen, wird sie mit einer Fensterfunktion multipliziert (z.B. rechte Hälfte eines Hamming-Fensters)

Zur Erzeugung des stimmhaften Anregungssignals wird das LF-Modell von FANT *et al.* [11] verwendet. Es modelliert direkt den sich zeitlich verändernden Volumenstrom durch die Glottis als Funktion von 4 Parametern und der Grundfrequenz. Die zeitliche Ableitung des

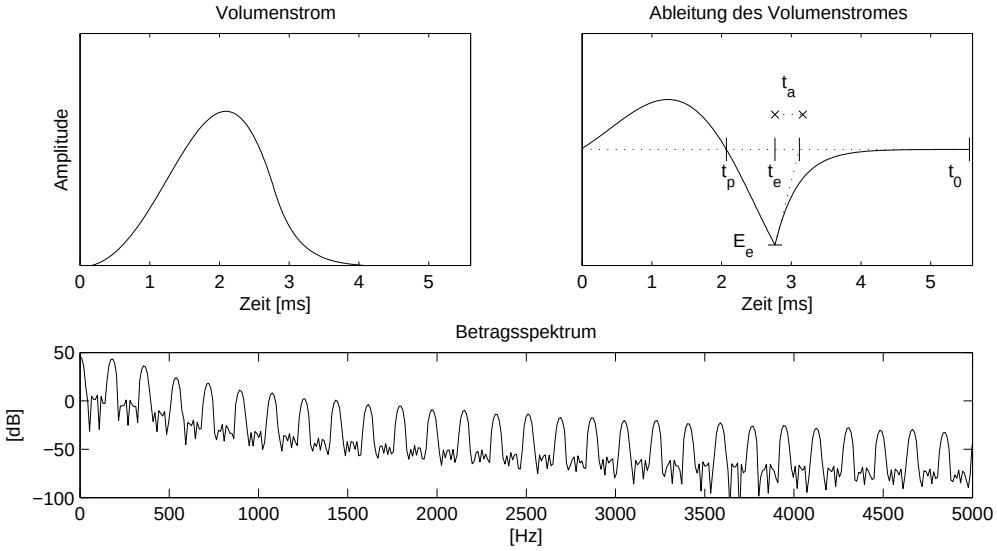


Abbildung 6.1: *Anregungsimpuls (Volumenstrom durch Glottis) nach dem LF-Modell [11], zeitliche Ableitung des Volumenstroms und Betragsspektrum des Anregungssignals bei einer Grundfrequenz von 180 Hz.*

Volumenstromes $\dot{g}(t)$ wird wie folgt berechnet:

$$\dot{g}(t) = \begin{cases} E_0 \cdot e^{\alpha \cdot t} \cdot sin(\frac{\pi \cdot t}{t_p}) & 0 \leq t \leq t_e \\ \frac{-E_e}{\epsilon \cdot t_a} \cdot [e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_0-t_e)}] & t_e < t \leq t_0 \end{cases} . \tag{6.1}$$

Die Parameter ϵ und α ergeben sich aus Gleichung (6.1) für die Bedingungen $\dot{g}(t_e) = -E_e$ und $\int_0^{t_0} \dot{g}(t)dt = 0$. Die stimmlose Anregung für Frikative erfolgt mit einem Rauschsignal

(gaussverteiltes weisses Rauschen). Das synthetische stationäre Sprachsignal ergibt sich dann aus der Faltung des Anregungssignals mit der Impulsantwort.

Für die Synthese eines längeren Signals mit einer zeitlich variierenden Vokaltraktkonfiguration ist eine Interpolation zwischen den einzelnen Schätzungen nötig. Eine Möglichkeit besteht in der Interpolation der artikulatorischen Parameter oder der Querschnittsflächen. Dies hätte allerdings zur Folge, dass die Übertragungsfunktion des Modells in jedem Abtastzeitpunkt neu berechnen werden müsste, was einen enormen rechnerischen Aufwand bedeutet. Alternativ können direkt die Abtastwerte der Impulsantworten zwischen zwei Schätzzeitpunkten linear interpoliert werden. Das Ergebnis dieses Vorgehens entspricht nicht genau der Interpolation der artikulatorischen Parameter oder der Querschnittsflächen, stellt aber laut [44] eine akzeptable Approximation dar.

6.1 Diphon-Korpus

Die Aufnahmen des Diphon-Korpus stammen von einem männlichen Sprecher. Das Korpus umfasst ingesamt 3019 einzelne Diphone bei einer Abtastfrequenz von 16 kHz. Die Signale beginnen bei Diphonen in der Mitte des ersten Phons und enden in der Mitte des zweiten Phons.

6.1.1 Vokale

Auf den Abbildungen der folgenden Seiten sind die Ergebnisse der akustisch-artikulatorischen Inversion für Vokale zu sehen. Aus dem Diphon-Korpus wurden dazu jeweils die Einträge rausgesucht, die mit einem vorgegebenen Laut anfangen und mit einem beliebigen Laut enden. Je nach Anfangslaut sind das zwischen 10 und 20 Diphone. Für jedes dieser Diphone wurde ein Analysefenster von 25 ms verwendet, beginnend in der Mitte des ersten Lautes.

Aus den Analyseabschnitten wurden die artikulatorischen Parameter für das verlustlose und das verlustbehaftete akustische Modell ermittelt. Die Parameterschätzung für das verlustlose Modell erfolgte wie in Kapitel 5.1 beschrieben. Für das verlustbehaftete Modell wurde direkt das am besten passende Vektorpaar (kleinste euklidische Distanz zwischen den MFCC-Merkmalsvektoren) aus dem Codebuch gesucht.

Von allen geschätzten artikulatorischen Parametervektoren wurden je Anfangslaut die Vokaltraktkonturen und die Mittelwerte der einzelnen Parameter berechnet. Die gestrichelten Linien zeigen die Extremalwerte bezüglich der mediosagitallen Distanzen, zwischen denen die Vokaltraktkonturen liegen. Die mittlere durchgezogene Linie entspricht der Vokaltraktkontur des Mittelwertvektors.

Für die Laute [a],[i] und [u] sind zusätzlich noch Mediosagittal-Ansichten von MRT-Aufnahmen abgebildet. Die Aufnahmen wurden während der Artikulation der entsprechenden Laute gemacht. Sie stammen allerdings nicht von der gleichen Person wie die Sprachaufnahmen des Diphon-Korpus (Quelle der MRT-Aufnahmen: [16]).

Auf den ersten Blick erscheint die Varianz der Schätzungen relativ gross. Wie verschiedene Hörproben bestätigt haben, sind diese Variationen jedoch auch in den Lautcharakteristiken deutlich feststellbar, wohl hauptsächlich durch Koartikulation bedingt.

Die Ergebnisse der Schätzungen für das verlustlose und das verlustbehaftete akustische Mo-

dell sind bei den Vokalen insgesamt recht ähnlich. Einige Unterschiede sind jedoch gut sichtbar.

Im hinteren Mundraum bzw. oberer Rachenraum fallen beim verlustlosen Modell die Schätzung der Querschnittsflächen der vorderen Vokalen unverhältnismässig gross aus (erkennbar an der Einbuchtung des Zungenrückens). Da durch die LPC-Analyse nur die Flächenverhältnisse und nicht die absoluten Werte der Querschnittsflächen bestimmt werden, wurde die Ursache dafür zuerst im Skalierungsfaktor der Flächen in Verbindung mit der Schätzung der artikulatorischen Parameter vermutet. Dies konnte jedoch nicht bestätigt werden. Eine Änderung des Skalierungsfaktors führt zwar im gesamten Vokaltrakt zu kleineren oder grösseren mediosagittalen Distanzen, die Proportionen bleiben allerdings bestehen.

Deutlich unterschiedliche Vokaltraktformen kann man bei den Vokalen [a],[i] und [ɪ] erkennen. Die Positionen des Zungenrückens beim verlustlosen Modell liegen hier weiter vorne im Mundraum. Beim [i] liegt die engste Stelle im Vokaltrakt etwa im Bereich der Vorderzähne, wo fast ein Verschluss gebildet wird, während sie beim verlustlosen Modell ungefähr in der Mitte des harten Gaumens liegt. Nimmt man die vorhandenen MRT-Aufnahmen als Referenz, so sind die Schätzungen des verlustbehafteten Modells klar zu bevorzugen. Die Vokaltraktform und die Lage der Engstellen zeigen eine gute Übereinstimmung.

Die Unterschiede bei den vorderen Vokalen sind zu einem grossen Teil auf die Seitenzweige der Sinus piriformis zurückzuführen. Durch die verursachten Antiresonanzen verschieben sich die Formantfrequenzen merklich, wobei besonders der 2. Formant der vorderen Vokale davon betroffen ist (vgl. Abbildung 6.2).

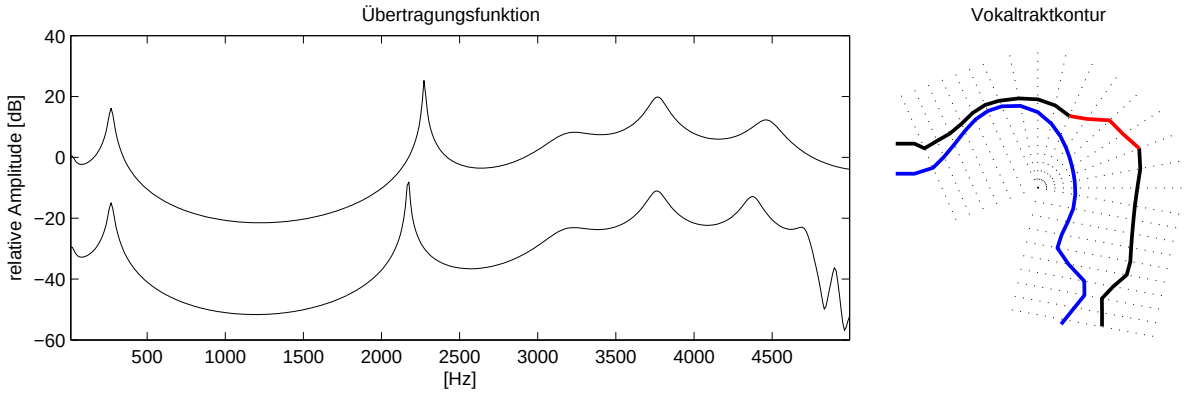


Abbildung 6.2: *Verschiebung der Formantfrequenzen durch die Sinus piriformis: Die Vokaltraktform entspricht der eines [i]. Die Übertragungsfunktion oben ergibt sich ohne Sinus piriformis, die Übertragungsfunktion unten mit den angekoppelten Seitenzweigen der Sinus piriformis.*

Die Abbildung 6.3 zeigt, wie sich die verschiedenen Verluste auf die Übertragungsfunktion des Systems auswirken. Die Verluste durch Wärmeleitung an den Rohrwänden hat nur einen sehr geringen Effekt auf die Formantbandbreiten. Für die Schätzung könnten diese Verluste vernachlässigt werden. Die Verluste durch die elastischen Rohrwände beeinflussen vor allem die niedrigen Frequenzen. Erkennbar ist eine Änderung der Frequenz und Bandbreite des ersten Formanten. Die Reibungsverluste und die Verluste durch die Schallabstrahlung nehmen gegen hohe Frequenzen zu. Während die Reibungsverluste vor allem zu einer Dämpfung der Formanten führen, haben die Verluste durch Schallabstrahlung auch eine Verschiebung der For-

mantfrequenzen zur Folge. Einen Eindruck, wie sich die Berücksichtigung der Verluste auf die Schätzungen auswirken, kann man durch den Vergleich der Ergebnisse der beiden Modelle gewinnen. Aufgrund des nichtlinearen Zusammenhangs sind die Auswirkungen auf die einzelnen Laute sehr unterschiedlich.

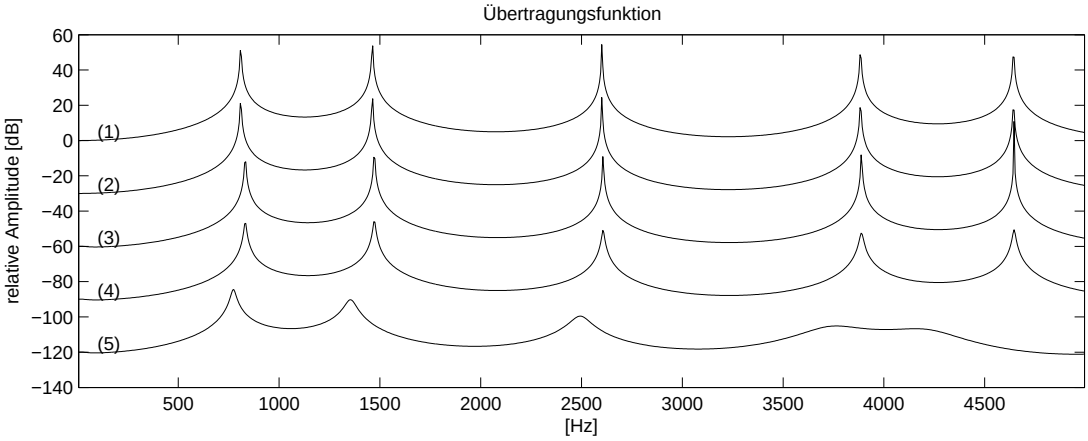


Abbildung 6.3: *Einfluss der Verluste auf die Übertragungsfunktion ([a]): Bei (1) sind alle Verluste vernachlässigt. Danach werden nacheinander berücksichtigt: (2) Verluste durch Wärmeleitung, (3) Verluste durch elastische Rohrwände, (4) Verluste durch Reibung an den Rohrwänden und (5) Verluste durch Schallabstrahlung.*

Im Vokalviereck klassifiziert man die Vokale nach der Lippenrundung und der Lage des Zungenrückens, wobei man zwischen vorderen und hinteren und zwischen tiefen und hohen Vokalen unterscheidet, je nachdem wo sich der höchste Zungenpunkt befindet. Entsprechend der Entfernung dieses Punktes vom Gaumen wird ein Vokal als offen oder geschlossen bezeichnet.

Die Lage der Zunge in den geschätzten Vokaltraktkonturen stimmt gut mit den zu erwartenden Positionen gemäss Vokalviereck überein. Abweichungen in den Schätzungen des verlustbehafteten Modells sind nur beim [u] und [ʊ] festzustellen. Beim [u] und [ʊ] wäre nach dem Vokalviereck eine höhere Zungenposition zu erwarten (im Vergleich zum [o]) und beim [ʊ] eine Lage des Zungenrückens etwas weiter vorne. Die Unterschiede zwischen den geschlossenen und offenen Varianten der Vokale (z.B. zwischen [e] und [ɛ]) sind überall erkennbar.

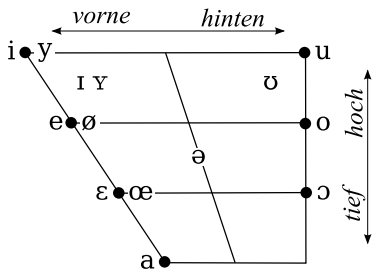
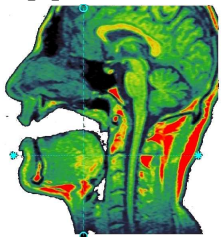
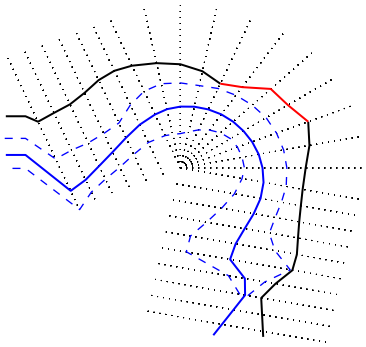
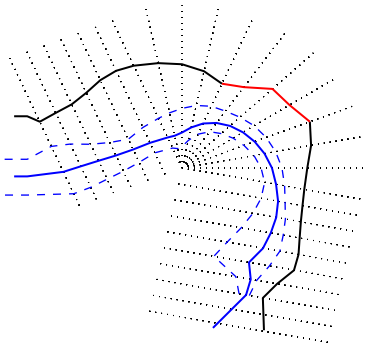
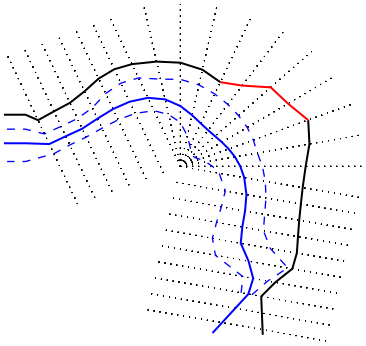
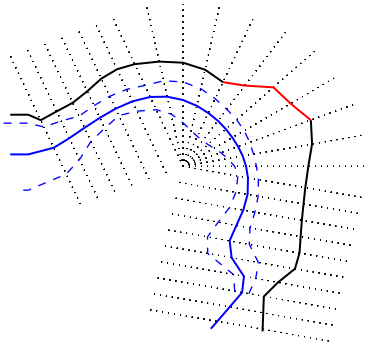
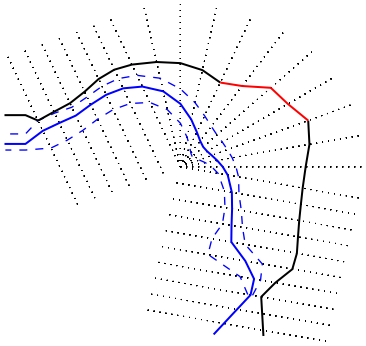
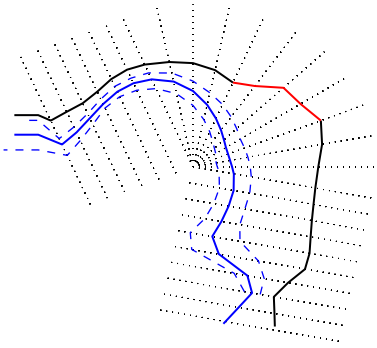
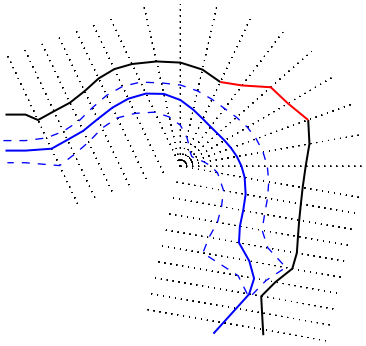
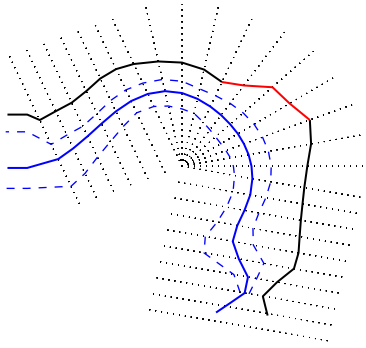
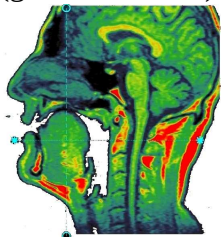
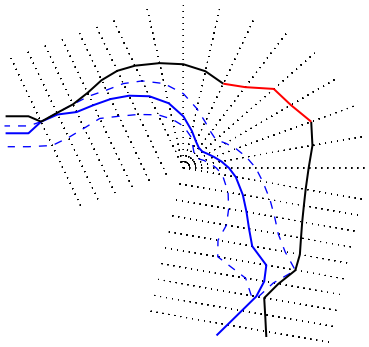
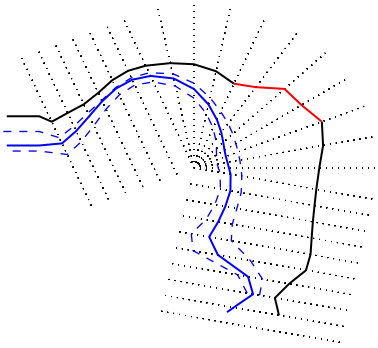
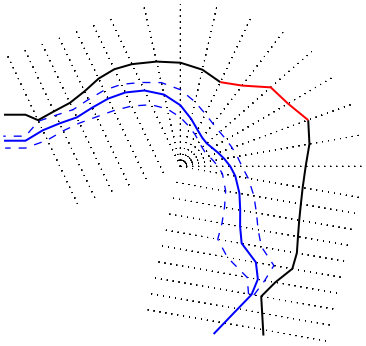
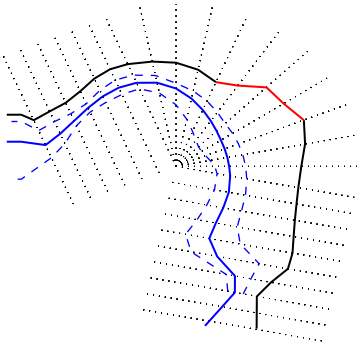
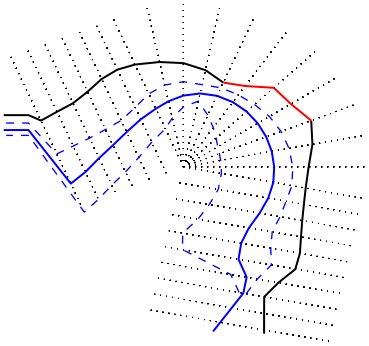
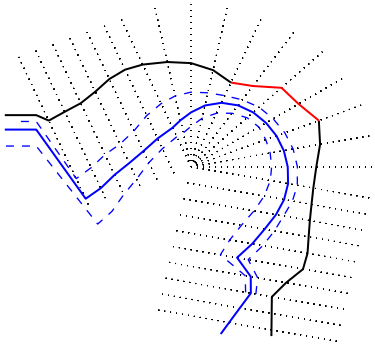
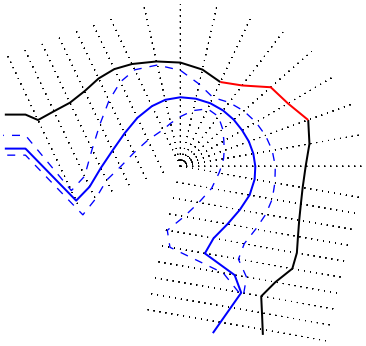
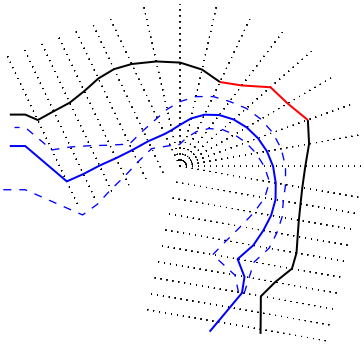


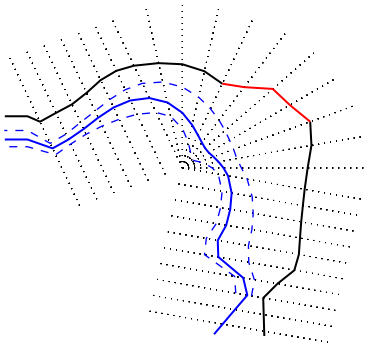
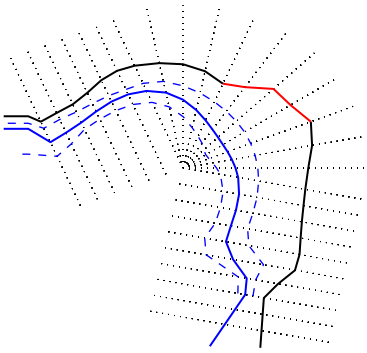
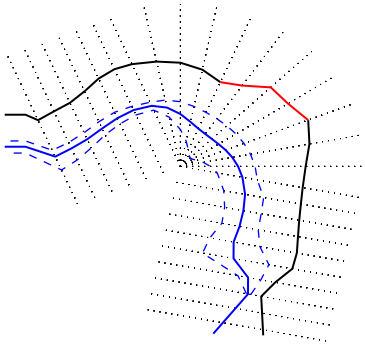
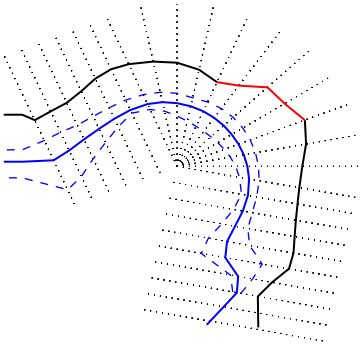
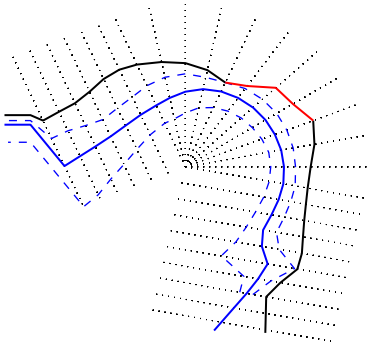
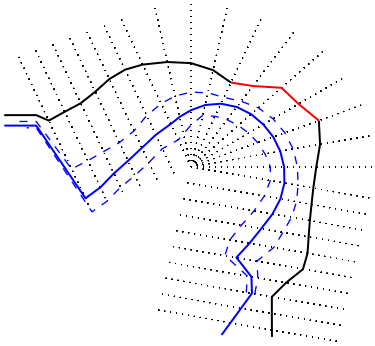
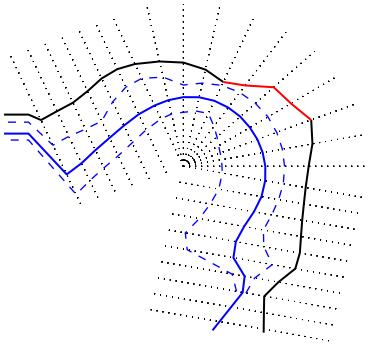
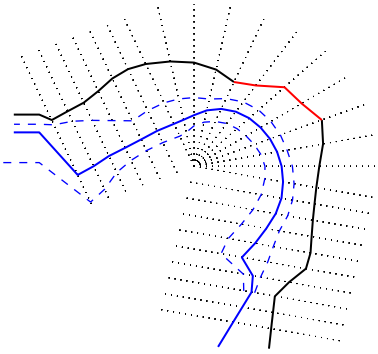
Abbildung 6.4: *Vokalviereck: Einteilung der Vokale nach Lage des Zungenrückens. Die Lage der Vokale links oder rechts der Punkte gibt an, ob die Lippen ungerundet oder gerundet sind (nach [1]).*

Laut und Beispiel	*verlustloses Modell*	*verlustbehaftetes Modell*
[a] Sch**a**tten 		
[ə] bitt**e**		
[e] **e**gal (geschlossenes e)		
[ɛ] **E**nde (offenes e)		

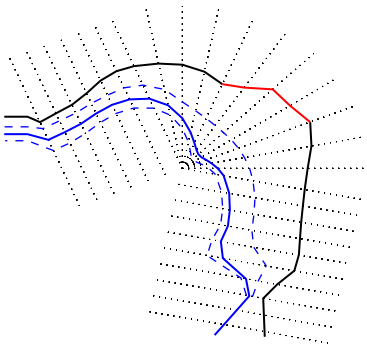
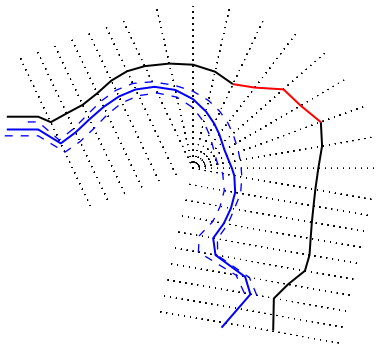
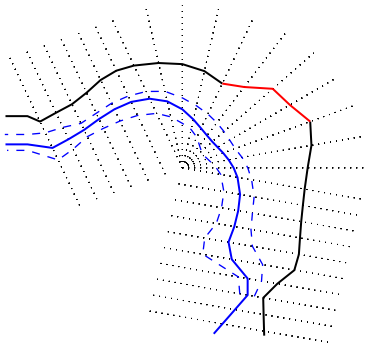
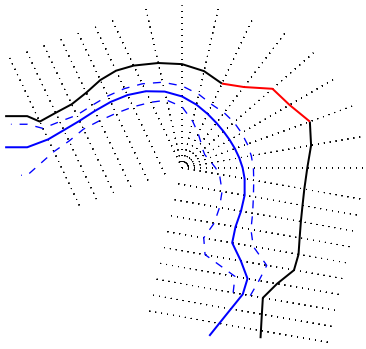
akustisch-artikulatorische Inversion: Vokale

Laut und Beispiel	*verlustloses Modell*	*verlustbehaftetes Modell*
[i] v**i**tal (geschlossenes i) 		
[ɪ] T**i**sch (offenes i)		
[o] M**o**tiv (geschlossenes o)		
[ɔ] G**o**tt (offenes o)		

akustisch-artikulatorische Inversion: Vokale

Laut und Beispiel	*verlustloses Modell*	*verlustbehaftetes Modell*
[ø] sch**ö**n (geschlossenes ö)		
[œ] **ö**ffentlich (offenes ö)		
[u] br**u**tal (geschlossenes u) 		
[ʊ] L**u**ft (offenes u)		

akustisch-artikulatorische Inversion: Vokale

Laut und Beispiel	*verlustloses Modell*	*verlustbehaftetes Modell*
[y] Ph**y**sik (geschlossenes ü)		
[ʏ] d**ü**nn (offenes ü)		

akustisch-artikulatorische Inversion: Vokale

6.1.2 Konsonanten

Die Inversionsmethoden wurden ebenso an den Konsonanten des Diphon-Korpus erprobt. Das Vorgehen bei den Schätzungen und die Berechnungen für die abgebildeten Vokaltraktkonturen sind wie bei den Vokalen durchgeführt worden.

Frikative, Vibranten und Approximanten

Das verlustlose Modell sieht grundsätzlich nur eine Anregung an einem Rohrende vor. Bei den Frikativen findet die Anregung jedoch grösstenteils innerhalb des Vokaltraktes statt. Die Schätzung für diese Frikative sind daher wenig aussagekräftig. Die LPC-Analyse führt typischerweise zu einer Flächenfunktion, die Richtung Lippen immer weiter wächst und anatomisch unmögliche Grössen erreicht. Die zugehörigen Vokaltraktformen zeigen scheinbar eine Engstelle im Bereich des harten Gaumens. Ursache dieser Täuschung sind allerdings nur die unrealistischen Flächenfunktionen.

Die geschätzten Vokaltraktkonturen des verlustbehafteten Modells weisen bei den Zischlauten [s],[z],[ʃ] und [ç] Konstriktionen im Bereich des Zahndammes auf. Vereinzelt liegen die Engstellen auch im palatalen Bereich, was eher unwahrscheinlich ist. Die Konturen der Laute unterscheiden sich nur wenig voneinander. Insbesondere beim [ç] würde man die Verengung generell etwas weiter hinten im Mundraum vermuten (vgl. Abbildung 2.6).

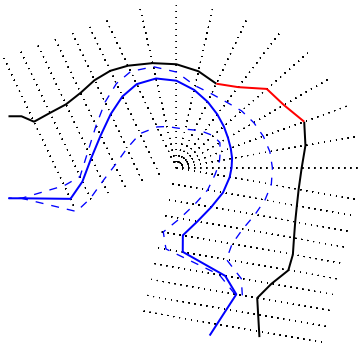
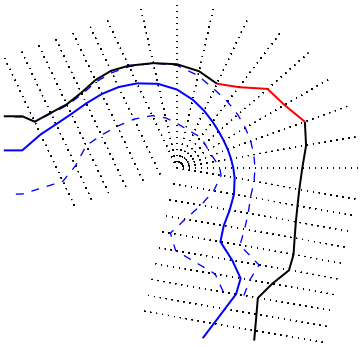
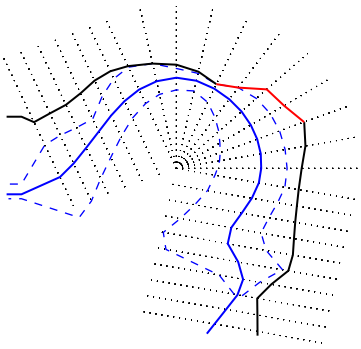
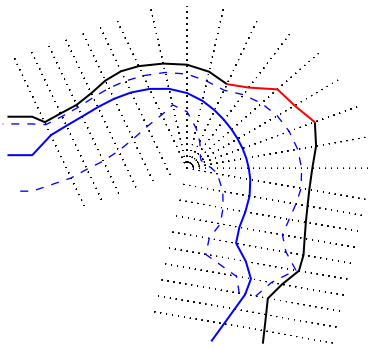
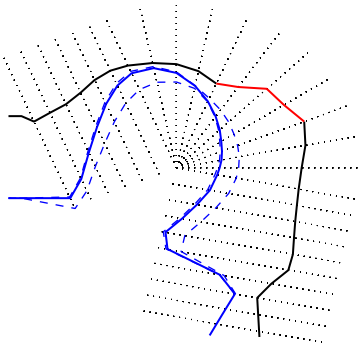
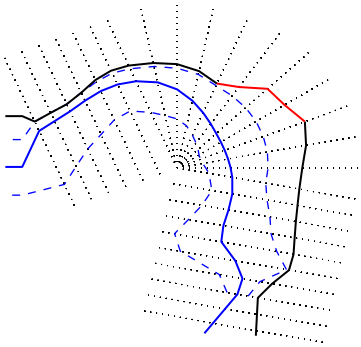
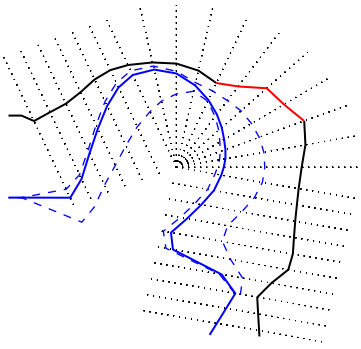
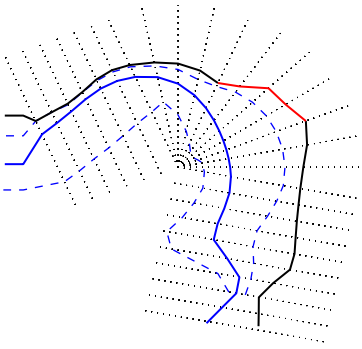
Für das [f] gelingt keine sinnvolle Schätzung der Vokaltraktkonturen. Die Konstriktionsorte reichen vom bilabialen bis zum velaren Bereich. Die Hauptursache dafür ist die hohe spektrale Variabilität der f-Laute. Diese Beobachtung haben auch NARAYANAN und ALWAN [35] gemacht. Sie erklären die Variabilität durch unterschiedliche Positionen der Zunge. Für die Erzeugung der labiodentalen Laute sind in erster Linie die Unterlippe und die Vorderzähne entscheidend, während die Zunge im Mundraum ziemlich frei bewegt werden kann. Dadurch ändern sich die akustischen Eigenschaften des Resonanzraumes hinter der Konstriktion.

Beim [v] ist die Engstelle hingegen klar an den Lippen erkennbar. Die im Vergleich zum [f] bessere Schätzung ist vermutlich auf die zusätzlichen Informationen über die Vokaltraktkonfiguration durch die stimmhafte Anregung an der Glottis zurückzuführen. Eine genauere Unterscheidung zwischen bilabialer und labiodentaler Konstriktion ist mit dem verwendeten artikulatorischen Modell nicht möglich.

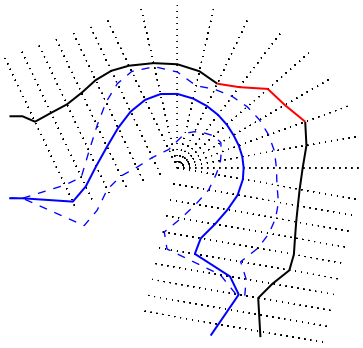
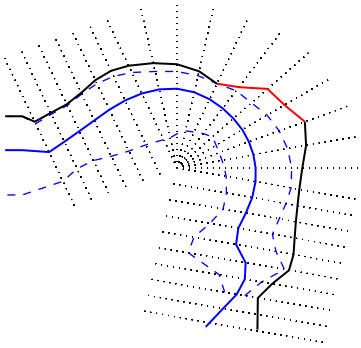
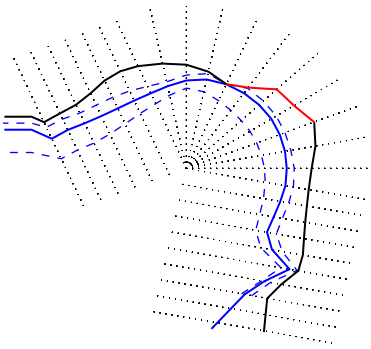
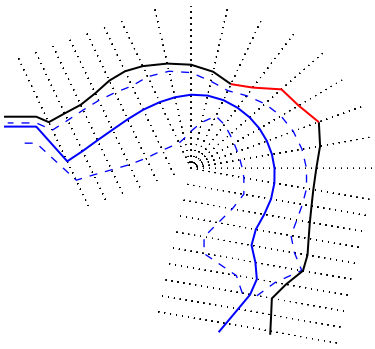
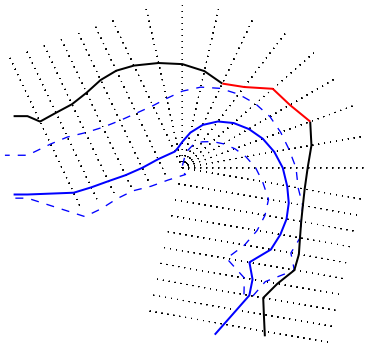
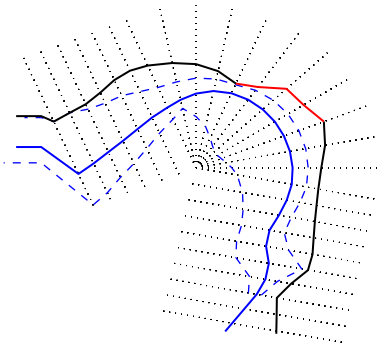
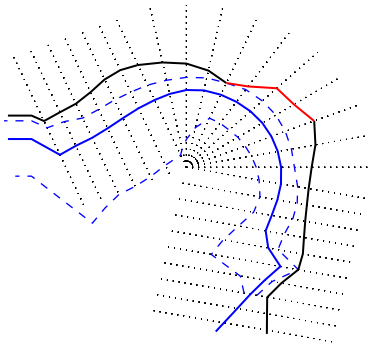
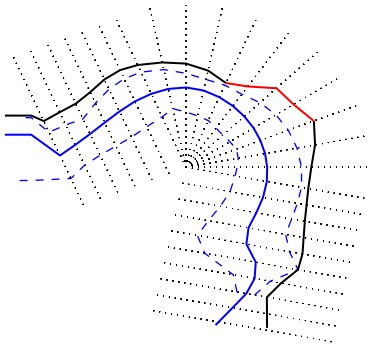
Beim [x] wird die Engstelle mehrheitlich richtig am weichen Gaumen geschätzt.

Für das [h] und das [r] ist eine Beurteilung der Schätzungen schwierig. Die Friktion vom [h] wird an der Glottis erzeugt. Das akustische Modell sieht jedoch nur eine stimmhafte Anregung an der Glottis vor. Bei der berechneten Übertragungsfunktion fehlt daher der Einfluss des Quellenspektrums. Zudem kann der Vokaltrakt bei der Produktion des Lautes unterschiedliche Konfigurationen einnehmen, weshalb das Spektrum stark von den benachbarten Lauten abhängt. Für das [r] erfolgt die Schätzung des Sprechtraktes wie bei den Vokalen. Eine deutliche Artikulationsstelle ist hier nicht erkennbar.

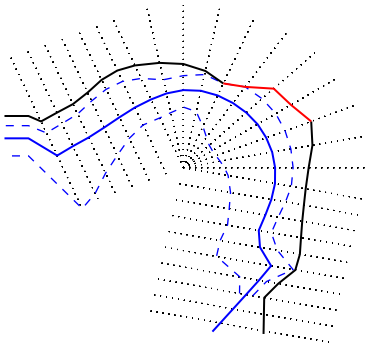
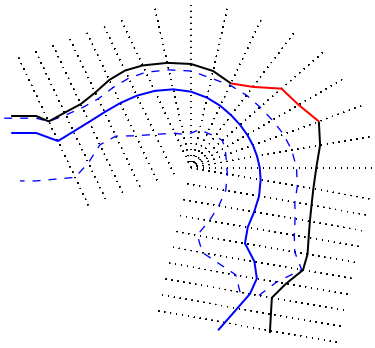
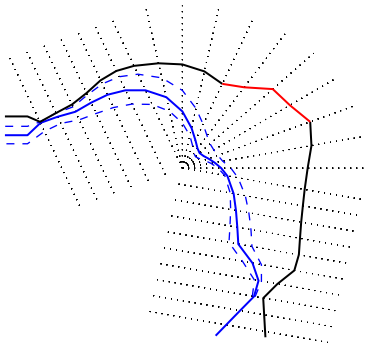
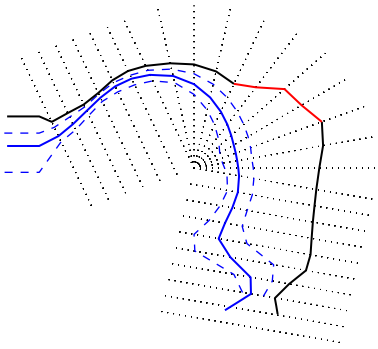
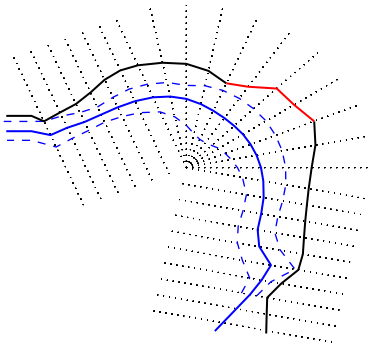
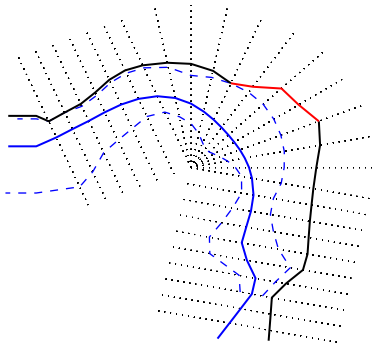
Der Halbvokal [j] zeigt erwartungsgemäss eine engere Konstriktion als die Vokale. Beim [l] berührt die Zunge nicht den Zahndamm. Dies würde beim zweidimensionalen Modell einen vollständigen Verschluss bedeuten. Stattdessen ist eine leichte Verengung im vorderen Mundbereich sichtbar.

Laut und Beispiel	*verlustloses Modell*	*verlustbehaftetes Modell*
[s] Hau**s** (stimmloses s)		
[z] Ha**s**e (stimmhaftes s)		
[ʃ] Fi**sch** (stimmloses sch)		
[ç] mö**ch**te		

akustisch-artikulatorische Inversion: Konsonanten

Laut und Beispiel	*verlustloses Modell*	*verlustbehaftetes Modell*
[f] Ha**f**t		
[v] **W**elt		
[x] Bu**ch**		
[h] **H**all		

akustisch-artikulatorische Inversion: Konsonanten

Laut und Beispiel	*verlustloses Modell*	*verlustbehaftetes Modell*
[r] **R**atte		
[j] **j**äh		
[l] **L**ast		

akustisch-artikulatorische Inversion: Konsonanten

Nasale und Plosive Bei Nasalen gelingt keine realistische Schätzung des Vokaltraktes. Im Falle des unverzweigten Rohrsystems ist dies nicht erstaunlich. Allerdings bringt auch der Einbezug der verbundenen Resonanzräume (geschlossener Mundraum, Sinus piriformis und Nasennebenhöhlen) keine Verbesserung, obwohl sich damit Nasale synthetisieren lassen, die man durchaus auch als solche wahrnimmt. Die berechneten Übertragungsfunktionen unterscheiden sich aber stark von den Spektren der natürlichen Laute.

In Abbildung 6.5 ist oben das Betragsspektrum eines [m] dargestellt. Aus einem kleinen Codebuch mit 500 Einträgen (nur Nasale), mit dem Mundraum als einziger angekoppelter Resonanzraum, wurde das am besten passende Vektorpaar ausgesucht. Der Einfluss der verschlossenen Mundhöhle als Resonanzraum ist sowohl im Betragsspektrum des natürlichen Signals als auch in der Übertragungsfunktion des akustischen Modells erkennbar und ebenso der Einfluss der Sinus piriformis. Die Antiresonanzen der Nasennebenhöhlen hingegen lassen sich im Betragsspektrum des [m] kaum ausmachen. Obwohl die Übertragungsfunktion ohne Nasennebenhöhlen besser zum beobachteten Spektrum passt, sind die damit synthetisierten Laute nicht mit einem [m] vergleichbar. Erst durch das Ankoppeln der Nasennebenhöhlen gelingt dies.

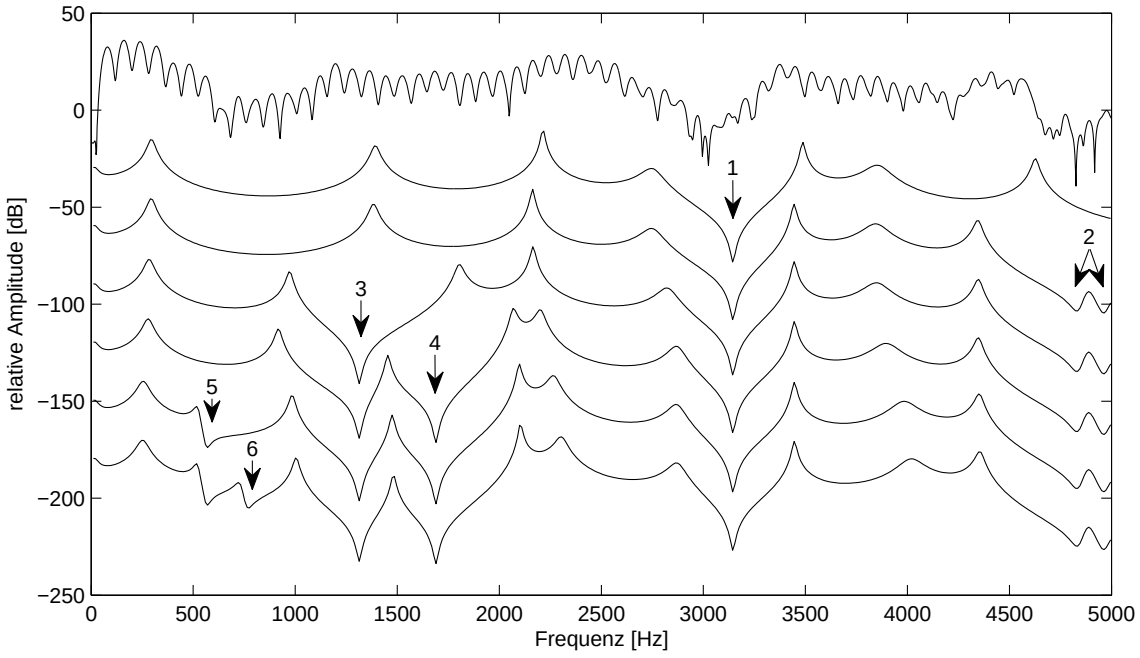


Abbildung 6.5: *Betragsspektrum eines [m] (oben) und Übertragungsfunktionen des akustischen Modells. Markiert sind die Antiresonanzen erzeugt durch Ankopplung von: 1. Mundraum, 2. Sinus piriformis, 3. grosse Keilbeinhöhle, 4. kleine Keilbeinhöhle, 5. Kieferhöhle, 6. Stirnhöhle.*

Plosivlaute lassen sich aufgrund ihrer Instationarität nicht unmittelbar mit den verwendeten akustischen Modellen beschreiben. Die Artikulationsorte lassen sich aber oftmals aus den geschätzten Vokaltraktkonturen vor und nach den Plosivlauten ableiten. Beispiele dazu folgen im nächsten Abschnitt.

6.2 Natürliche Sprachsignale

Neben der Inversion einzelner Signalabschnitte wurden die Schätzmethoden auch an kurzen natürlichen Sprachsignalen erprobt. Für das verlustbehaftete Modell kann hierbei die Suche mittels dynamischer Programmierung verwendet werden. Die Schätzung des verlustlosen Modells profitiert nicht von den zusätzlichen zeitlichen Informationen. Die Ergebnisse entsprechen daher der Analyse einzelner unabhängiger Abschnitte wie beim Diphon-Korpus. Auf die Darstellung dieser Ergebnisse wird daher verzichtet.

Die nachfolgenden Analysen wurde immer mit einem 25 ms langen Analysefenster durchgeführt, welches jeweils um 10 ms verschoben wurde. Das verwendete Codebuch umfasst knapp 16’000 Einträge.

Für längere Sprachsignale kann der zeitliche Verlauf der geschätzten Vokaltraktformen mit Hilfe einer dreidimensionalen Darstellung ähnlich eines Spektrogramms visualisiert werden [41]. Zuerst werden aus den Vokaltraktkonturen die Flächenfunktionen als Funktion der Zeit ermittelt. Die Grössen der Querschnittsflächen (Amplituden) werden als Schwärzung dargestellt. Je kleiner die Fläche desto dunkler der Abschnitt. Für eine bessere Unterscheidbarkeit kleiner Werte wird eine logarithmische Skalierung der Graustufenskala verwendet. Die horizontale Achse repräsentiert die Zeitachse und die vertikale Achse den Abstand von der Glottis. In Abbildung 6.6 ist die Transformation an einem Beispiel veranschaulicht.

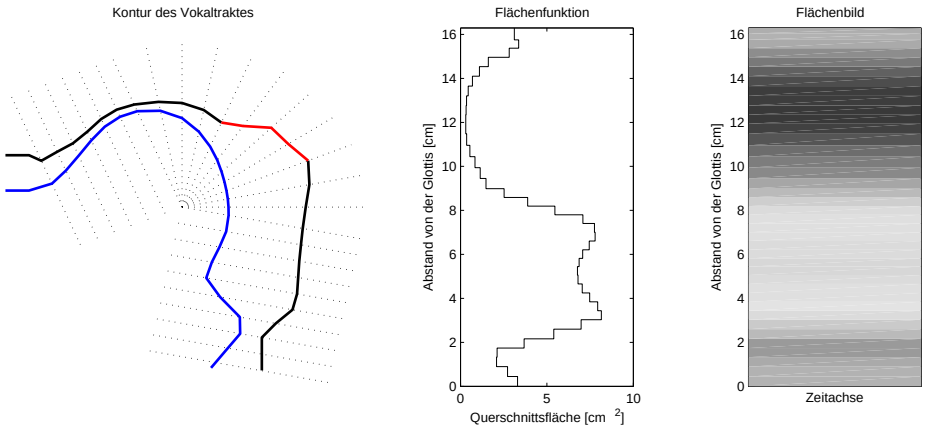


Abbildung 6.6: *Kompakte Darstellung der Querschnittsflächen als Flächenbild.*

Abbildung 6.7 zeigt das Flächenbild für das Sprachsignal “Quand j’ai du temps libre”, gesprochen von einer Frau. Die Flächenwerte wurden für die Darstellung zusätzlich zwischen den einzelnen Analysezeitpunkten interpoliert. Die geschätzten Konstriktionen sind gut an den dunklen Bereichen erkennbar. Für die Plosivlaute [d], [t] und [b] wird korrekt ein vollständiger Verschluss geschätzt. Die Position des Verschlusses und die Form des Vokaltraktes in diesem Bereich werden durch die vorangehenden und nachfolgenden Vokaltraktkonturen bestimmt. Das [k] am Anfang weist einen Verschluss im Bereich des vorderen Gaumens auf. Das nachfolgende [ã] genügt hier nicht für eine korrekte Schätzung des Verschlussortes. Das Gaumensegel bleibt bei den zwei nasalierten Lauten geschlossen.

Das Flächenbild des Sprachsignals “Should we chase those cowboys?” ist in Abbilung 6.8 dargestellt. Die Sprachaufnahme stammt in diesem Fall von einem Mann. Bei den Frikativen

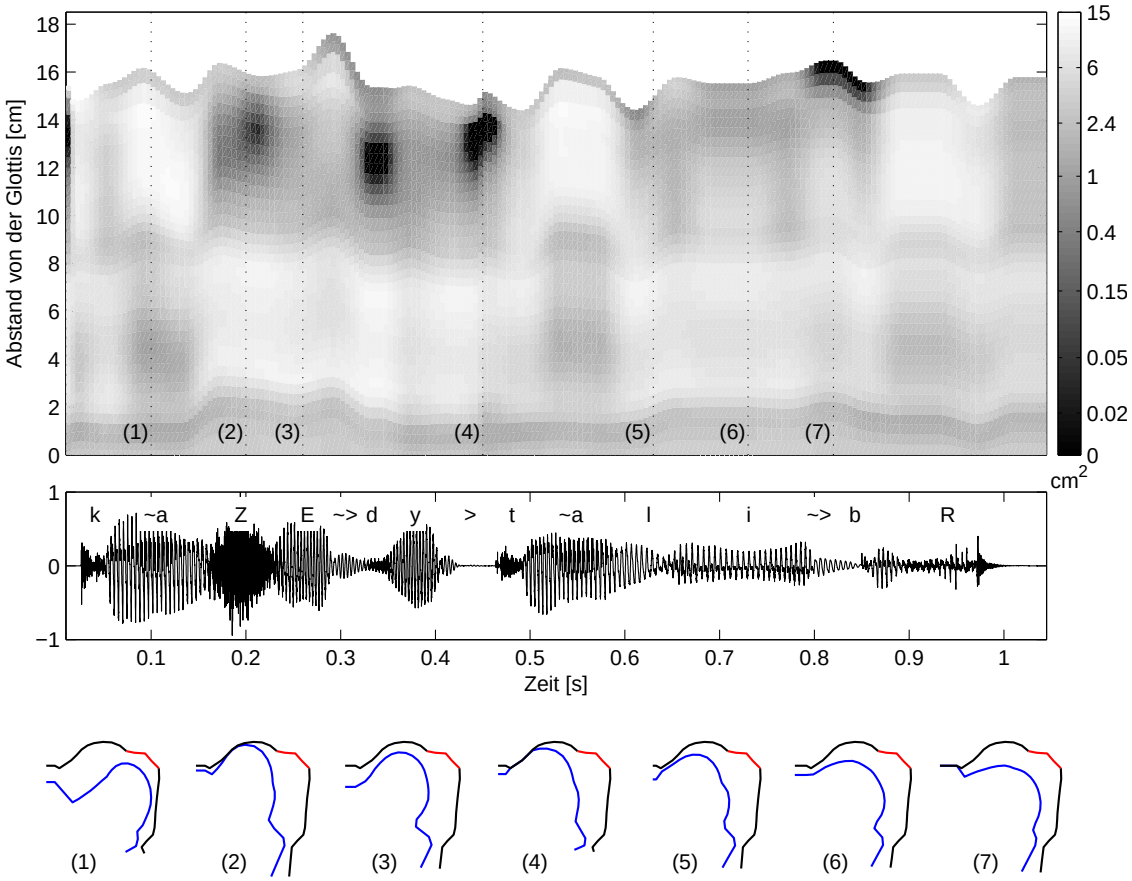


Abbildung 6.7: *Flächenbild des Sprachsignals "Quand j'ai du temps libre". Für die gestrichelten Linien sind die zugehörigen Vokaltraktkonturen angegeben.*

liegen die Konstriktionen an den zu erwartenden Stellen. Mit Ausnahme des [d] werden auch die Verschlüsse der Plosivlaute gut geschätzt.

Allgemein werden die Verschlussorte gut aus den benachbarten Vokaltraktkonturen geschätzt sofern es sich dabei um Vokale handelt. Folgen mehrere Konsonanten aufeinander, treten häufiger Fehler auf. Die Fehlschätzungen sind vermehrt bei den stimmhaften Plosivlauten festzustellen, wo anstatt eines vollständigen Verschlusses nur eine Engstelle in den Ergebnissen zu sehen ist. Bei den Vokalen sind keine offensichtlichen Fehler auszumachen. Auffällig sind die teilweise recht grossen Änderungen zwischen zwei aufeinanderfolgenden Vokaltraktkonturen. Mit wachsendem Codebuch werden die Verläufe glatter.

Auf der beiliegenden CD finden sich zahlreiche weitere Ergebnisse der Inversionsexperimente (im Verzeichnis `Beispiele_Inversion`). Es sind auch kurze Videosequenzen aus den Schätzungen erstellt worden, welche die zeitlichen Veränderungen des Vokaltraktes in der mediosagittalen Ebene zeigen. Die Animationen sind mit den synthetisierten Sprachsignalen unterlegt.

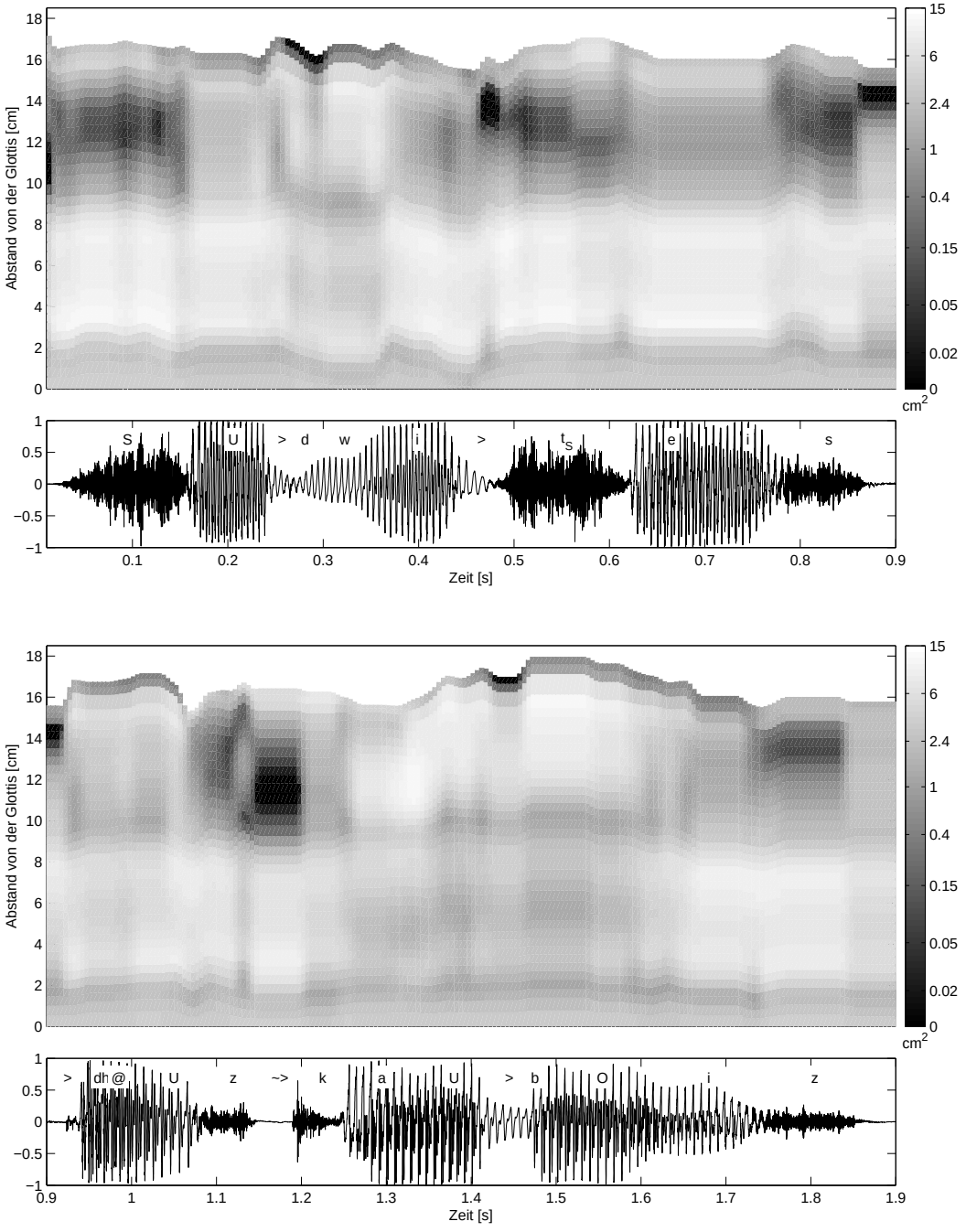


Abbildung 6.8: *Flächenbild des Sprachsignals "Should we chase those cowboys?".*

7 Diskussion und Ausblick

In der vorliegenden Arbeit werden zwei Methoden zur Schätzung der Vokaltraktform aus dem Sprachsignal vorgestellt. Die Schätzmethoden beruhen auf einer akustischen und einer artikulatorischen Modellierung des Sprechapparates.

Während für beide Methoden das gleiche artikulatorische Modell von MAEDA [26] verwendet wird, unterscheiden sich die akustischen Modelle durch den Grad der vorgenommenen Vereinfachungen.

Die Untersuchungen anhand von Sprachproben haben gezeigt, dass beide Schätzmethoden für Vokale insgesamt brauchbare Ergebnisse liefern. Gerade bei den vorderen Vokalen wird aber klar, dass der akustische Effekt der Sinus piriformis nicht vernachlässigt werden darf. Ebenso führen die verschiedenen Verluste zu deutlichen Änderungen der Frequenzen und Bandbreiten der Formanten. Während die Formantbandbreiten für die Schätzungen eher eine untergeordnete Rolle spielen (nicht aber für die Synthese), folgen aus den Verschiebungen der Formantfrequenzen unterschiedliche Vokaltraktkonturen. Es ist daher anzunehmen, dass eine realistische Modellierung der Verluste die Ergebnisse der Schätzungen verbessert.

Für eine sinnvolle Schätzung des Vokaltraktes bei Frikativlauten müssen die zusätzlich verteilten Geräuschquellen berücksichtigt werden. Der physikalische Mechanismus, der zur Entstehung dieser Laute beiträgt, wird hierfür durch drei konzentrierte Quellen nachgeahmt. Trotz dieser starken Vereinfachung können im Allgemeinen gute Ergebnisse erzielt werden. Wichtig für weitere Verbesserungen in diesem Zusammenhang wäre vor allem eine genauere Beschreibung des vorderen Mundraumes. Die Flächenfunktion in diesem Bereich hängt bei Frikativen von zahlreichen Details ab, die im artikulatorischen Modell von MAEDA nur sehr grob wiedergegeben werden. Für aspirierte Laute sollte das Frikativmodell noch um eine zusätzliche Geräuschquelle unmittelbar nach der Glottis erweitert werden.

Die Schätzung des Vokaltraktes bei Nasalen hat sich als sehr schwierig erwiesen. Obwohl mit dem akustischen Modell perzeptiv befriedigende Nasale synthetisiert werden können, unterscheidet sich die spektrale Zusammensetzung der synthetisierten Laute stark von den Spektren natürlicher Laute. Schwierigkeiten bereiten vor allem die grossen individuellen Unterschiede des Nasenraumes, die sich auch in den Spektren der Laute wiederfinden. Die feste Vorgabe der geometrischen Grössen des Nasenraumes, wie im Rahmen dieser Arbeit getan, hat sich nicht bewährt. Eine Möglichkeit, die Problematik zu umgehen, wäre, durch zusätzliche Hilfsmittel die Grenzen der Nasallaute zu bestimmen. Diese Signalpassagen könnten bei der Schätzung mittels dynamischer Programmierung wie kurze Pausen behandelt werden und der Ort des Verschlusses allenfalls aus den benachbarten Vokaltraktkonturen geschätzt werden. Eine wichtige Rolle dürfte auch, gerade bei den Nasalen, die Schallabstrahlung über die Haut spielen. Gemäss FANT *et al.* [12] ist diese im Bereich der Lippen und des Kehlkopfes bei geschlossenem Mund am stärksten und beeinflusst hauptsächlich den ersten Formaten, welcher das Spektrum bei Nasalen dominiert.

Die Ergebnisse zeigen, dass die Inversion insbesondere von einer genaueren akustischen Modellierung des Sprechapparates profitiert. Die Schätzmethode basierend auf dem verlustlosen Rohrmodell bietet hierfür keine Möglichkeiten, zeichnet sich aber durch eine hohe Effizienz und akzeptable Ergebnisse bei Vokalen aus. Bei der akustisch-artikulatorischen Inversion auf der Grundlage eines im Voraus erstellten Codebuches sind diese Einschränkungen nicht vor-

handen und die Schätzung des Vokaltraktes unabhängig von der Komplexität der verwendeten Modelle durchführbar. Die Methode der dynamischen Programmierung scheint eine geeignete Strategie zu sein, um dem Problem der Nichteindeutigkeit zu begegnen und kontinuierliche Bewegungsabläufe zu schätzen. Die Gewichtsfunktion kann problemlos erweitert werden, wenn zusätzliche Kriterien bei der Suche berücksichtigt werden sollen.

Die Simulation des Sprechtraktes im Frequenzbereich eignet sich gut für die Analyse (quasi-) stationärer Laute. Für Laute, die durch schnelle Bewegungen gekennzeichnet sind (z.B. [R]), oder für transientes Verhalten, wie es bei Plosivlauten der Fall ist, wäre auch eine Zeitbereichssimulation denkbar ([27]). Bis auf die frequenzabhängigen Elemente (für Verluste durch Reibung und Wärmeleitung) könnte das gleiche akustische Netzwerk verwendet werden. Eine solche Simulation setzt allerdings Vorwissen über die Bewegungen der Artikulatoren voraus.

Anhang A: Technische Realisierung

Dieser Abschnitt gibt einen kurzen Überblick über das Computerprogramm, in welchem die in der Arbeit beschriebenen Modelle und Methoden umgesetzt wurden. Um eine einfache Bedienung zu gewährleisten, wurde eine grafische Benutzeroberfläche erstellt. Die Möglichkeit, interaktiv die Parameter und Konfigurationen der Modelle zu verändern und ihre Auswirkung auf die akustischen Merkmale beobachten zu können, hat sich als sehr nützlich erwiesen.

Die Programmdateien befinden sich auf der beiliegenden CD im Verzeichnis `matlab/demo_inv`. Das Hauptprogramm wird mit dem Skript `demo_inv.m` gestartet (MATLAB v7.9).

Das Programm bietet folgende Funktionalität:

- Die akustisch-artikulatorische Inversion eines Sprachsignals mit den vorgestellten Schätzverfahren.
- Die Visualisierung der Schätzungen in diversen Formen.
- Die Generierung synthetischer Sprachsignale.
- Die interaktive Konfiguration der Modellparameter.

In Abbilung A1 ist die Hauptansicht der graphischen Oberfläche zu sehen. Die Ansicht gibt einen Überblick über das geladene Sprachsignal und die aktuelle Vokaltraktform.

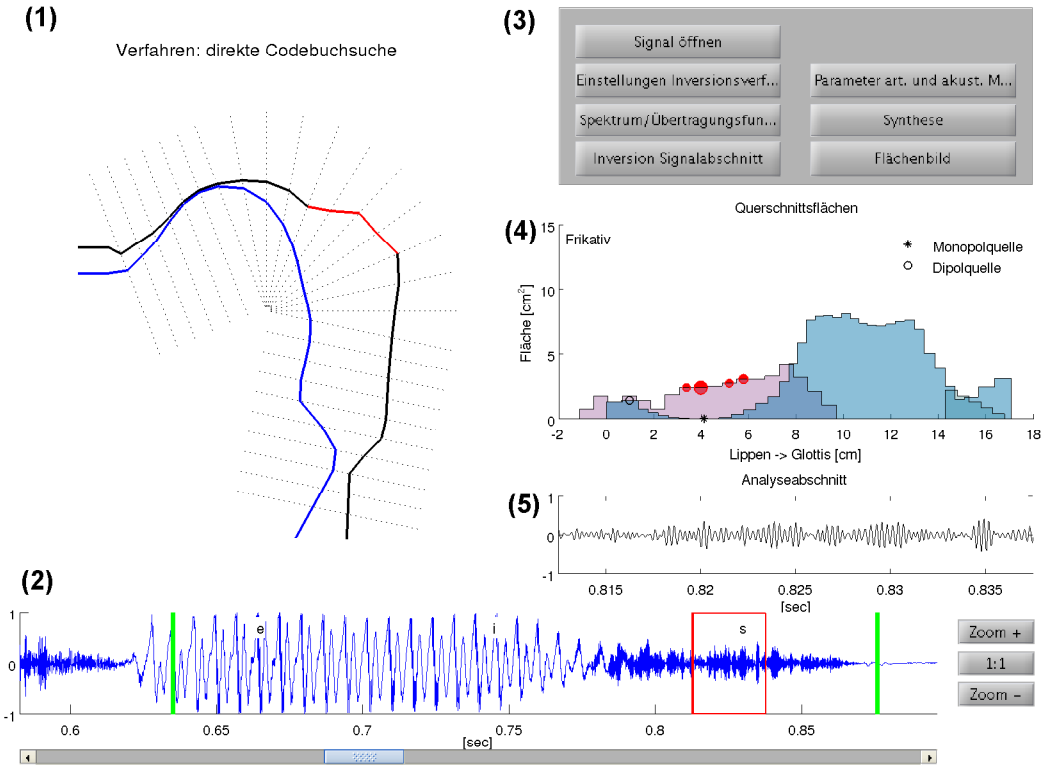


Abbildung A1: *Benutzeroberfläche des Hauptprogramms.*

In **(1)** wird die Vokaltraktkontur des artikulatorischen Modelles dargestellt, wobei im Titel die gewählte Schätzmethode angezeigt wird.

Für das akustische Modell wird die Distanzfunktion der Vokaltraktkontur in die entsprechende diskrete Flächenfunktion umgerechnet. Die geometrischen Grössen, welche für die akustische Simulation benutzt werden, sind im Bereich **(4)** zu sehen. Die blaue Fläche entspricht der Flächenfunktion des Rachen- und Mundraumes. Die Lippenöffnung wird hier als Referenzpunkt benutzt. Die rosa eingefärbe Fläche stellt die Flächenfunktion des Nasenraumes dar und die roten Kreise symbolisieren die Verzweigungen zu den Nasennebenhöhlen. Die grüne Fläche im Rachenraum entspricht der addierten Flächenfunktionen der Sinus piriformis. Ist eine Konstriktion im Vokaltrakt vorhanden, werden die Positionen der zwei Geräuschquellen mit einem Kreis bzw. einem Stern angezeigt. Ausser für die LPC-basierte Schätzung wird im Programm immer das verlustbehaftete Rohrmodell zur akustischen Simulation des Sprechapparates benutzt.

Im Signalfenster **(2)** werden zusätzlich zum Signal Textlabels angezeigt. Beim Öffnen eines Signals werden diese selbstständig geladen, falls im Verzeichnis eine Datei mit demselben Namen und der Endung ".lab" vorhanden ist. Die Zeilen der Label-Datei enthalten jeweils eine Spalte für den Beginn und das Ende der Segmentgrenze in 100 ns Einheiten und eine Spalte für den angezeigten Text (HTK label format).

Die 3 überlagerten Markierungen sind per Drag and Drop verschiebbar. Mit den grünen Markierungen kann ein bestimmter Abschnitt aus dem gesamten Signal ausgewählt werden. Das rote Rechteck entspricht dem aktuellen Analysefenster, welches nochmals in **(5)** vergrössert dargestellt wird. Beim Verschieben des Fensters werden die Darstellungen automatisch aktualisiert.

Über die Schaltfläche "Einstellungen Inversionsverfahren" **(3)** können die Parameter der Kurzzeitanalyse (Grösse und Verschiebung des Analyseabschnittes) und das Schätzverfahren gewählt werden. Die Schätzung der artikulatorischen Parameter kann unmittelbar für das ausgewählte Analysefenster erfolgen (LPC-Analyse bzw. direkte Suche im Codebuch) oder die Parameter werden im Voraus berechnet. Dies geschieht über die Schaltfläche "Inversion Signalabschnitt". Durch die Ausdehnung der Schätzung auf einen längeren Signalabschnitt können die Vorteile der optimierten Codebuchsuche (DP) genutzt werden. Die Ergebnisse der Schätzung werden gespeichert und können durch Verschieben des Analysefensters im Signalfenster betrachtet werden. Es besteht danach auch die Möglichkeit, die Ergebnisse in Form eines Flächenbildes, wie in Abschnitt 6.2 vorgestellt, darzustellen.

Die Schaltfläche "Spektrum/Übertragungsfunktion" öffnet ein Fenster (Abbildung A2), in welchem die entsprechende Übertragungsfunktion des akustischen Modells (schwarz) zusammen mit dem Betragsspektrum des Signals (blau) dargestellt wird. Anstatt der Übertragungsfunktion können auch die Koeffizienten des Mel-Cepstrums oder die berechnete Eingangsimpedanz des Vokaltraktes angezeigt werden.

Über die Schaltfläche "Parameter art. und akust. Modell" sind die wichtigsten Parametereinstellungen der Modelle zugänglich. Dazu gehören:

- die 8 Parameter des artikulatorischen Modells. Sie können mit Schiebereglern innerhalb ihrer Wertebereiche verstellt werden. Die Einstellungen können gespeichert und wieder geladen werden.

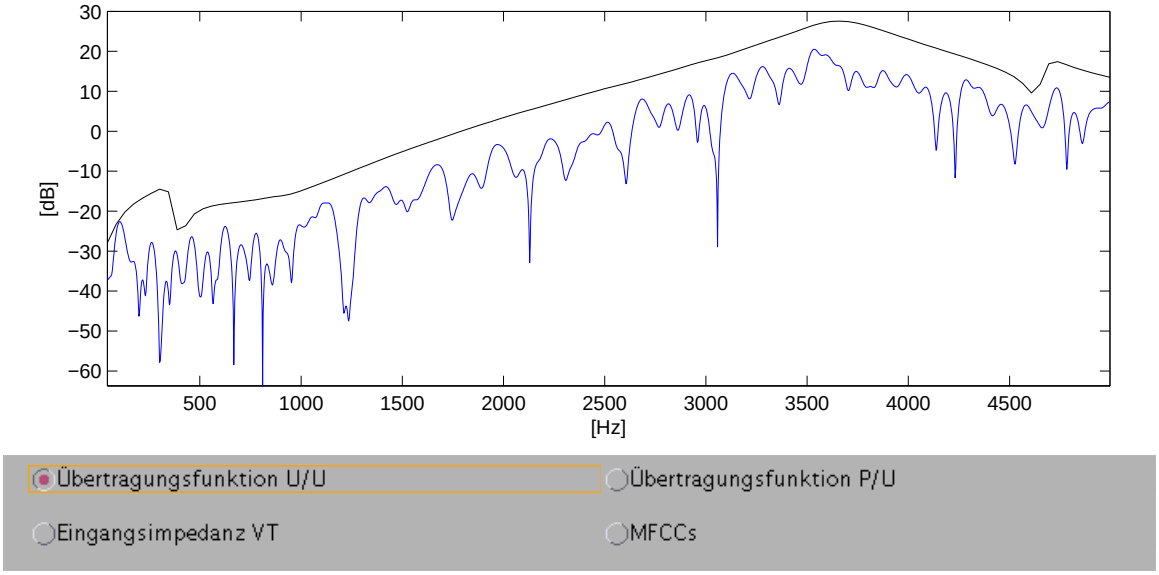


Abbildung A2: *Ansicht der Übertragungsfunktion des akustischen Modells und des Betragsspektrums des Analysefensters.*

- die Parameter der stimmhaften Anregungsfunktion für die Synthese.
- die Volumina der Nasennebenhöhlen.
- die allgemeine Konfiguration des akustischen Modells. Diese beinhaltet z.B. die Ab-/Ankopplung der verschiedenen Seitenzweige oder die Auswahl der Energieverluste, welche berücksichtigt werden sollen.

Bei einer Änderung der Parameter werden die Darstellungen der Hauptansicht und die Übertragungsfunktion automatisch aktualisiert, so dass sich die Auswirkungen direkt verfolgen lassen.

Einen akustischen Eindruck der Modelleinstellungen erhält man über die Schaltfläche “Synthese”. Für die aktuelle Vokaltraktkonfiguration kann ein kurzes synthetisches Sprachsignal erzeugt werden. Um einen natürlicheren Klang zu erhalten, werden Grundfrequenz F_0 und Amplitude A_0 über die Zeit variiert. Die verwendeten Werte sind in Tabelle A1 aufgeführt.

Zeit [ms]	0	250	500	600
Frequenz	$0.8 \cdot F_0$	$1 \cdot F_0$	$0.65 \cdot F_0$	$0.65 \cdot F_0$
Zeit [ms]	0	40	400	600
Amplitude	$0.0 \cdot A_0$	$1 \cdot A_0$	$0.95 \cdot A_0$	$0.0 \cdot A_0$

Tabelle A1: *Variation von Grundfrequenz und Amplitude für die Synthese kurzer Signale. Zwischen den Zeitpunkten werden die Werte linear interpoliert.*

Die Synthese kann ebenfalls für die geschätzten artikulatorischen Parameter eines ganzen Signalabschnittes erfolgen. Der zeitliche Verlauf der Grundfrequenz und der Intensität werden hierfür aus dem natürlichen Sprachsignal übernommen. In Kombination mit einer Animation der Vokaltraktkonturen lassen sich so auch Videosequenzen erzeugen. Um besser die Verände-

rungen verfolgen zu können, ist eine Zeitlupenfunktion integriert worden. Für die Erstellung der Videos wird die frei verfügbare FFmpeg-Bibliothek[1] verwendet.

Die Synthese kann wahlweise mit einer Frequenzbereichssimulation, wie in Abschnitt 6 beschrieben, erfolgen, oder die Synthese wird mit dem von MAEDA entwickelten Verfahren [27] im Zeitbereich durchgeführt. Bei der Simulation im Zeitbereich werden nur Rachen- und Mundraum berücksichtigt.

Die Skripte zum Erzeugen und Clustern des Codebuches befinden sich im Unterverzeichnis `codebook_generation`. Sie können über eine einfache grafische Oberfläche `codebook_gen_gui.m` bedient werden. Das verwendete Codebuch befindet sich im Unterverzeichnis `codebook`. Es setzt sich zusammen aus den akustischen und artikulatorischen Merkmalsvektoren, den Mittelwertvektoren der Partitionen und die Zuordnung der Vektorpaare zu den Partitionen. Ein zusätzliches Feld weist jedem Vektorpaar eine Lautklasse zu, welche bei der Erzeugung anhand der Geometrie des Vokaltraktes bestimmt wurde. Unterschieden wird zwischen Vokalen, Frikativen, nasalierten Lauten, Nasalen und Verschlusslauten.

[1] http://ffmpeg.org/

Anhang B: Aufgabenstellung

Institut für
Technische Informatik und
Kommunikationsnetze

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo

(MA-2009-14)

MASTERARBEIT

für

Herrn Simon Simonet

Betreuer: Dr. B. Pfister, ETZ D97.6
M. Gerber, ETZ D97.4

Ausgabe: 1. Oktober 2009
Abgabe: 31. März 2010

Schätzung des bewegten Vokaltraktes aus dem Sprachsignal

Einleitung

Um bei der Sprachcodierung eine hohe Datenreduktion zu erreichen, wird meistens ein an die menschliche Sprachproduktion angelehntes Modell eingesetzt. Ein gebräuchliches Modell ist das sogenannte Source-Filter-Modell, das nur die beiden Hauptfunktionen, nämlich die Signalgenerierung und die Klangformung unterscheidet. Der zugehörige mathematische Formalismus ist unter der Bezeichnung *linear predictive coding* (LPC) bekannt (siehe z.B. [1] oder Kapitel 4.5 in [2]).

In [3] ist gezeigt worden, dass aus dem Sprachsignal mit der LPC-Analyse (Methode A im Abschnitt VI) direkt die Form des Vokaltraktes geschätzt werden kann. Diese Schätzung ist jedoch nur dann gültig, wenn die Stimmlippen die einzige Schallquelle bilden, was hauptsächlich für die Vokale zutrifft.

Problemstellung

Das Ziel dieser Masterarbeit ist, zu untersuchen, ob und wie der Ansatz der Vokaltraktschätzung in [3] auf andere Lautgruppen ausgeweitet werden kann. Es sind dazu auch neuere Arbeiten wie [4] zu konsultieren. Zudem sollen verschiedene graphische Visua-

lisierungen des aus dem Sprachsignal geschätzten Bewegungsablaufes des Vokaltraktes realisiert werden.

Das Ziel der Arbeit ist ein Tool, mit dem die zeitliche Veränderung des Vokaltraktes für ein beliebiges Sprachsignal ermittelt und auf verschiedene Arten (z.B. als eine Folge von graphischen Darstellungen oder als Video) visualisiert werden kann.

Vorgehen

Wie die Literatur zeigt, ist das Bestreben, aus dem Sprachsignal Rückschlüsse auf die Stellung der Artikulatoren zu ziehen nicht neu. Die meisten Arbeiten haben sich jedoch nur mit den Vokalen beschäftigt.

Für diese Masterarbeit wird das folgende Vorgehen empfohlen:

1. Anhand einer Literaturrecherche ist zu untersuchen, für welche Lautklassen welche Methoden zur Schätzung des Vokaltraktes entwickelt worden sind. Von Interesse ist dabei auch, wie die Schätzungen graphisch dargestellt werden.

2. Als erster Schritt zu einem Visualisierungs-Tool soll die Vokaltraktschätzung nach [3] implementiert werden. Es ist zu untersuchen, für welche Laute mit dieser Methode sinnvolle Resultate erzielt werden können.

3. Falls nötig sollen nun für die Lautkategorien Nasale und Frikative erweiterte Methoden zur Schätzung entwickelt werden und es ist zu überlegen, wie die Resultate zweckmässig visualisiert werden können.

4. Eine zweckmässige Methode zur dynamischen Visualisierung soll konzipiert und implementiert werden. Die Methode muss nicht echtzeitfähig sein; die Bildsequenz kann off-line berechnet werden.

5. Anhand von Sprachproben verschiedener Sprecher und Sprecherinnen ist zu untersuchen, wo die entwickelten Vokaltraktschätzmethoden brauchbare Ergebnisse liefern und wo noch Entwicklungsbedarf besteht.

Die ausgeführten Arbeiten und die erhaltenen Resultate sind in einem Bericht zu dokumentieren (siehe dazu [5]), der in gedruckter und in elektronischer Form (als PDF-Datei) abzugeben ist. Zusätzlich sind im Rahmen eines Kolloquiums zwei Präsentationen vorgesehen: etwa drei Wochen nach Beginn soll der Arbeitsplan und am Ende der Arbeit die Resultate vorgestellt werden. Die Termine werden später bekannt gegeben.

Literaturverzeichnis

[1] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

[2] B. Pfister and T. Kaufmann. *Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Springer Verlag (ISBN: 978-3-540-75909-6), 2008. http://www.springer.com/978-3-540-75909-6.

[3] H. Wakita. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Transactions on Audio and Electroacoustics*, 21(5):417–427, October 1973.

[4] S. Krstulovic. *Speech Analysis with Production Constraints*. PhD thesis, EPFL, 2001.

[5] B. Pfister. *Richtlinien für das Verfassen des Berichtes zu einer Semester- oder Diplomarbeit*. Institut TIK, ETH Zürich, Februar 2009. (http://www.tik.ee.ethz.ch/~spr/SADA/richtlinien_bericht.pdf).

[6] B. Pfister. *Hinweise für die Präsentation der Semester- oder Diplomarbeit*. Institut TIK, ETH Zürich, März 2004. (http://www.tik.ee.ethz.ch/~spr/SADA/hinweise_praesentation.pdf).

Zürich, den 1. Oktober 2009

Literatur

[1] International Phonetic Association. The International Phonetic Alphabet. *http://www.langsci.ucl.ac.uk/ipa/*, 2005.

[2] B.S. Atal, J.J. Chang, M.V. Mathews, and J.W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, 63(5):1535–1555, 1978.

[3] P. Birkholz. *3D-Artikulatorische Sprachsynthese*. PhD thesis, Universität Rostock, 2005.

[4] C.P. Browman and L. Goldstein. Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, 18(3):299–320, 1990.

[5] C.H. Coker. A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64(4):452–460, April 1976.

[6] J. Dang and K. Honda. Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation. *The Journal of the Acoustical Society of America*, 100(5):3374–3383, 1996.

[7] J. Dang and K. Honda. Acoustic characteristics of the piriform fossa in models and humans. *The Journal of the Acoustical Society of America*, 101(1):456–465, 1997.

[8] J. Dang and K. Honda. Construction and control of a physiological articulatory model. *The Journal of the Acoustical Society of America*, 115(2):853–870, 2004.

[9] J. Dang, K. Honda, and H. Suzuki. Morphological and acoustical analysis of the nasal and the paranasal cavities. *The Journal of the Acoustical Society of America*, 96(4):2088–2100, 1994.

[10] G. Fant. *Acustic Theory of Speech Production*. The Hague: Mouton, 1960.

[11] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. Quarterly Progress and Status Report 4, Royal Institute of Technology, Stockholm, 1985.

[12] G. Fant, L. Nord, and P. Branderud. A note on the vocal tract wall impedance. Quarterly Progress and Status Report 4, Royal Institute of Technology, Stockholm, 1976.

[13] J. Flanagan. *Speech Analysis, Synthesis and Perception*. Springer-Verlag Berlin Heidelberg, 1972.

[14] G. Habermann. *Stimme und Sprache: eine Einführung in ihre Physiologie und Hygiene*. Thieme Stuttgart, 1986.

[15] J.M. Heinz and K.N. Stevens. On the relations between lateral cineradiographs area functions, and acoustic spectra of speech. In *Proc. Fifth Int. Congr. Acoust. Liège*, page A44, 1965.

[16] P. Hoole. MRI of vowel articulation. *http://www.phonetik.uni-muenchen.de/∼hoole/kurse/movies/mrivowel_notes.pdf*, 2001.

[17] K. Ishizaka, J. French, and J. Flanagan. Direct determination of vocal tract wall impedance. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 23(4):370–373, Aug 1975.

[18] K. Ishizaka, M. Matsudaira, and T. Kaneko. Input acoustic-impedance measurement of the subglottal system. *The Journal of the Acoustical Society of America*, 60(1):190–197, 1976.

[19] J.L. Kelly and C.C. Lochbaum. Speech synthesis. In *Proceedings of the Fourth International Congress on Acoustics, Kopenhagen*, volume Bd. G42, pages 1–4, 1962.

[20] S. Krstulović. *Speech analysis with production constraints*. PhD thesis, École polytechnique fédérale de Lausanne, 2001.

[21] B. J. Kröger. *Ein phonetisches Modell der Sprachproduktion*. PhD thesis, Universität Köln, 1998.

[22] A. Lacroix. Speech production: Acoustics, models, and applications. In J. Blauert, editor, *Communication Acoustics*, pages 321–337. Springer-Verlag Berlin Heidelberg, 2005.

[23] J.N. Larar, J. Schroeter, and M.M. Sondhi. Vector quantization of the articulatory space. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(12):1812–1818, 1988.

[24] R. Lerch, G. Sessler, and D. Wolf. *Technische Akustik*. Springer-Verlag Berlin Heidelberg, 2009.

[25] P.J. Lynch and M. Komorniczak. Paranasal sinuses. *http://commons.wikimedia.org/wiki/File:Paranasal_sinuses_numbers.svg*, 2009.

[26] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, 1979.

[27] S. Maeda. A digital simulation method of the vocal tract system. *Speech Communication*, 1:199–229, 1984.

[28] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

[29] J. Makhoul. Stable and efficient lattice methods for linear prediction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 25(5):423–428, Oct 1977.

[30] P. Mermelstein. Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

[31] P. Mermelstein and M.R. Schroeder. Determination of smoothed cross-sectional-area functions of the vocal tract from formant frequencies. *The Journal of the Acoustical Society of America*, 37(6):1186–1186, 1965.

[32] P. Meyer, R. Wilhelms, and H.W. Strube. A quasiarticulatory speech synthesizer for german language running in real time. *The Journal of the Acoustical Society of America*, 86(2):523–539, 1989.

[33] F.D. Minifie. Speech acoustics. In F.D. Minifie, T.J. Hixon, and F. Williams, editors, *Normal Aspects of Speech, Hearing, and Language*, page 173. Englewood Cliffs: Prentice-Hall, 1973.

[34] M. Mrayati, R. Carre, and B. Guerin. Distinctive regions and modes: a new theory of speech production. *Speech Communication*, 7(3):257 – 286, 1988.

[35] S. Narayanan and A. Alwan. Noise source models for fricative consonants. *Speech and Audio Processing, IEEE Transactions on*, 8(3):328–344, May 2000.

[36] G. Papcun, J. Hochberg, T.R. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *The Journal of the Acoustical Society of America*, 92(2):688–700, 1992.

[37] B. Pfister and T. Kaufmann. *Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Springer-Verlag Berlin Heidelberg, 2008.

[38] G. Ramsay and L. Deng. Maximum likelihood estimation for articulatory speech recognition using a stochastic target model. In *Proceedings of the EUROSPEECH'95*, pages 1401–1404, 1995.

[39] K. Schnell. *Rohrmodelle des Sprachtraktes: Analyse, Parameterschätzung und Syntheseexperimente*. PhD thesis, Johann Wolfgang Goethe-Universität Frankfurt am Main, 2003.

[40] J. Schroeter and M.M. Sondhi. Dynamic programming search of articulatory codebooks. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, volume 1, pages 588–591, May 1989.

[41] M.S. Shah and P.C. Pandey. Areagram display for investigating the estimation of vocal tract shape for a speech training aid. In *Symposium on Frontiers of Research on Speech and Music (Kanpur,India)*, pages 121–124, 2003.

[42] K. Shirai and M. Honda. Estimation of articulatory motion. In *Dynamic Aspects of Speech Production*, pages 279–302. Tokyo University Press, 1976.

[43] K. Shirai and T. Kobayashi. Estimation of articulatory motion using neural networks. *Journal of Phonetics*, 19:379–385, 1991.

[44] M. Sondhi and J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(7):955–967, Jul 1987.

[45] M.M. Sondhi. Resonances of a bent vocal tract. *The Journal of the Acoustical Society of America*, 79(4):1113–1116, 1986.

[46] A. Soquet, V. Lecuit, T. Metens, and D. Demolin. Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with mri. *Speech Communication*, 36:169–180(12), March 2002.

[47] V.N. Sorokin and A.V. Trushkin. Articulatory-to-acoustic mapping for inverse problem. *Speech Communication*, 19(2):105 – 118, 1996.

[48] B. H. Story, I.R. Titze, and E. A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 100(1):537–554, 1996.

[49] M. Vojnovic and M. Mijic. An improved model for the acoustic radiation impedance of the mouth based on an equivalent electrical network. *Applied Acoustics*, 66(5):481 – 499, 2005.

[50] H. Wakita. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *Audio and Electroacoustics, IEEE Transactions on*, 21(5):417–427, Oct 1973.

[51] H. Wakita and G. Fant. Toward a better vocal tract model. Quarterly progress and status report, Royal Institute of Technology, Stockholm, 1978.

[52] E.R. Weibel. *Morphometry of the Human Lung*. Springer-Verlag Berlin Heidelberg, 1968.