Vokaltraktmodellbasierte Schätzung von Steuerparametern eines Moduls zur Sprechernormalisierung

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
Heiko Freienstein
aus Göttingen

Göttingen 2000

D 7
Referent: Prof. Dr. M. R. Schroeder
Korreferent: Prof. Dr. D. Ronneberger
Tag der mündlichen Prüfung: 27.04.2000

Inhaltsverzeichnis

1. Einleitung

Die vorliegende Arbeit ist Teil eines Projektes, das als Zielsetzung die Sprechernormalisierung zur automatischen Spracherkennung hat. Angestrebt ist die Entwicklung eines der Spracherkennung vorgeschalteten Moduls, das Sprecherunterschiede weitgehend kompensiert, sodass das aufwendige Training auf einen speziellen Sprecher entfällt oder wenigstens reduziert werden kann. Als Steuerparameter des Moduls kommen neben den geeignet aufbereiteten akustischen Daten (z. B. Barkspektrogramme) auch andere sprecherspezifische Parameter in Frage.

Diese Arbeit beschäftigt sich mit der Extraktion von specherspezifischen Parametern des Stimmkanals, die später als zusätzliche Steuerparameter getestet werden sollen. Dazu wurde ein neuer Ansatz verfolgt, der die anatomischen Besonderheiten des Stimmkanals eines Testsprechers als kleine Störungen des Stimmkanals eines Referenzsprechers auffasst. Dieser Ausgangspunkt führte zu einem Verfahren zur Schätzung von sprecherspezifischen Parametern aus Formantfrequenzmustern. Das entwickelte Verfahren bestimmt simultan nichtlineare Längenskalierungen und Skalierungen der Querschnittsfunktion, die die Abweichung des Testsprechers vom Referenzsprecher charakterisieren.

Die Arbeit gliedert sich in drei Teile. Nach einem ersten allgemeinen Teil über berührte Themengebiete wie Sprachproduktion, Sprachsynthese und das inverse Problem wird das neue Verfahren zur Extraktion von sprecherspezifischen Parametern vorgestellt. Im dritten Teil folgen Experimente mit synthetischen Vokalen, gehaltenen, isoliert gesprochenen Vokalen und schließlich Vokalen aus natürlicher Sprache.

2. Die Physik der Sprachproduktion

Der menschliche Sprachapparat ist in Abbildung 2.1 vereinfacht dargestellt. Bei der Sprachproduktion strömt Luft aus der Lunge durch die Stimmritze (Glottis) weiter durch den Rachenraum (Pharynx) in die Mundhöhle. Durch Heben und Senken des Gaumensegels (Velum) kann in unterschiedlichem Maße die Nasenhöhle angekoppelt werden. Die Schallabstrahlung erfolgt über die Lippen und bei angekoppelter Nasenhöhle auch durch die Nasenlöcher.

Die Stimmlippen, die die Glottis begrenzen, sind beim normalen Ausatmen weit geöffnet. Bei stimmhafter Sprache werden die nah zusammenliegenden Stimmlippen durch den Luftstrom zu Schwingungen angeregt und dienen so als Schallquelle. Der Stimmkanal selbst funktioniert beim Sprechen als Resonator und beeinflusst an den Lippen die Abstrahlung. Stimmlose Anregung des Stimmkanals geschieht durch Rauschen, das beim Durchströmen der verengten Glottis sowie weiteren Verengungen des Stimmkanals durch turbulenten Fluss entsteht. Ein weiterer Anregungsmechanismus ist der plötzliche Druckabbau im Stimmkanal, durch den Verschlusslaute gebildet werden.

Man unterscheidet Phonation, d. h. die Stimmbildung in der Stimmritze, und Artikulation, die Veränderung des Stimmkanals durch Bewegung von Unterkiefer, Zunge, Lippen und Velum, den sogenannten Artikulatoren. Phonation und Artikulation können weitgehend unabhängig voneinander geschehen. Die Phonation steuert die Tonhöhe und die Art des Lautes wird durch die Artikulation festgelegt [43].

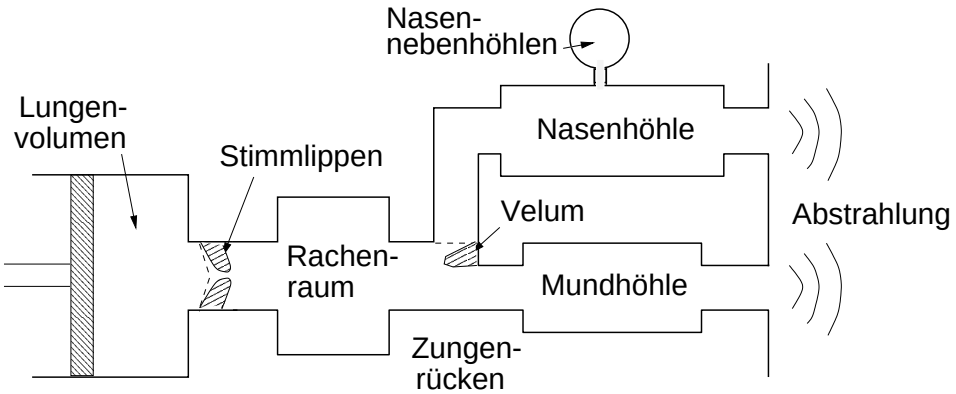


Abbildung 2.1.: Schema der funktionalen Bestandteile des Stimmkanals.

Analog zur Trennung zwischen Phonation und Artikulation werden bei der physikalischen Modellbildung im Allgemeinen die Anregung des Stimmkanals und der Stimmkanal

als Resonator in guter Näherung als wechselwirkungsfrei angenommen. Die physikalische Beschreibung wird in Quelle (Glottis) und Filter (Stimmkanal) aufgespalten [13].

2.1. Die Glottis

Der Mechanismus der Anregung der Stimmlippen durch den Luftstrom wird anhand eines Schwingungszyklus (der eingeschwungenen Schwingung) der Stimmlippen beleuchtet: Die aus der Lunge kommende Luft durchströmt die Stimmritze. Da diese eine Verengung darstellt, ist die Strömungsgeschwindigkeit erhöht. Auf Grund dieser Erhöhung kommt es zu einem Druckabfall in der Glottis. Dieser Druckabfall wird nach Daniel Bernoulli (1700 - 1782) als negativer Bernoulli-Druck bezeichnet. Die Bernoulli-Gleichung

$$p + \frac{1}{2}\varrho_0 v^2 = \text{const.} \tag{2.1}$$

ist eine dem Energiesatz äquivalente Beziehung für stationäre, laminare, reibungslose Flüssigkeiten und Gase (in Gleichung 2.1 ist die Beziehung in vereinfachter Form für inkompressible Medien angegeben). Der Bernoulli-Druck zieht die Stimmlippen bis zum Verschluss der Glottis zusammen. Die Elastizität der Stimmlippen und der steigende subglottale Druck öffnet die Glottis und der Zyklus beginnt erneut.

Titze [57] weist darauf hin, dass der beschriebene Prozess nicht ausreicht, um eine Schwingung dauerhaft aufrechtzuerhalten, da der Bernoulli-Druck auch der Öffnung der Stimmlippen entgegenwirkt. Um einen Nettoenergieübertrag aus der kinetischen Energie der Strömung auf die Schwingung zu gewährleisten, muss der intraglottische zeitliche Druckverlauf zwischen zwei Verschlüssen eine Asymmetrie aufweisen. Diese Asymmetrie kann durch zwei verschiedene Mechanismen hervorgerufen werden: Die Trägheit der bewegten Luftsäule und nichtgleichmäßige Schwingungsmoden. Die nichtgleichmäßige Schwingungsmode entsteht duch eine Überlagerung der Schwingung des Körpers der Stimmlippe (body) mit einer Welle, die in der Schleimhaut läuft, welche die Stimmlippe überzieht (cover). Titze legt die Physik der Schwingung der Stimmlippen mit kleinen Amplituden theoretisch und im Experiment [57–59] dar. Er kommt zu dem Schluss, dass die nichtgleichmäßige Schwingung bei normaler Phonation der dominante Effekt ist.

In Abbildung 2.2 ist der Zusammenhang der Ableitung des Flusses durch die Glottis während stimmhafter Phonation mit den spektralen Effekten in einem Simulationsexperiment dargestellt: Der Vokaltrakt wird durch einen sogenannten Pulszug angeregt. Im Fernfeld misst man ein Schalldrucksignal, das gegenüber dem Schallflusssignal im Nahfeld etwa einen spektralen Anstieg von 6 dB/Oktave hat. Durch die Differenzierung des Schallflusses wird diese Anhebung bereits bei der Anregung des Vokaltraktes durchgeführt, was den Vokaltrakt als lineares System voraussetzt. Die Faltung des abgeleiteten Glottispulses mit der Folge von Delta-Impulsen, die die zeitliche Wiederholung des Glottispulses bewirkt, wird nach dem *Faltungssatz* zu einer Multiplikation des fouriertransformierten Glottispulses mit der Fouriertransformierten der Pulsfolge, die wieder eine Pulsfolge ist.

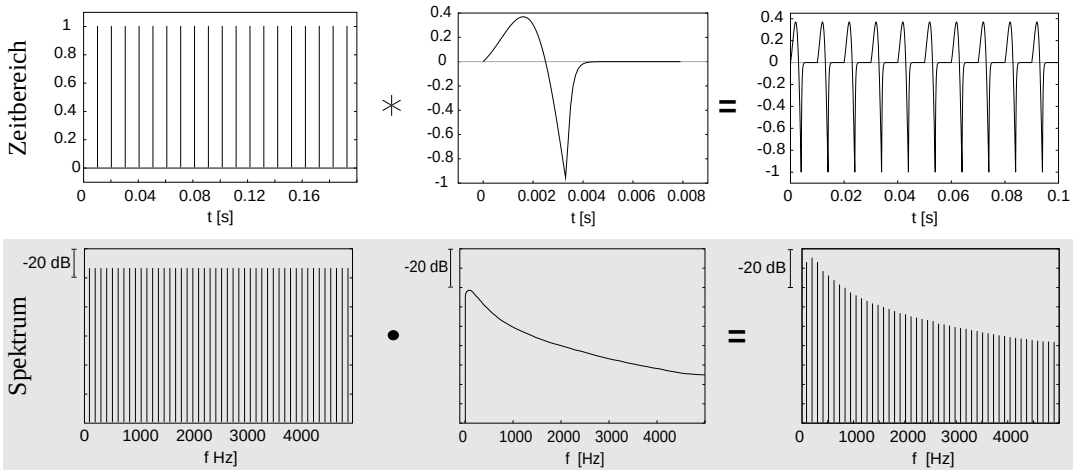


Abbildung 2.2.: Beschreibung des Schallflusses durch die Glottis im Zeit- und im Frequenzbereich in einem Simulationsexperiment.

Prominente Auswirkungen im Spektrum sind der spektrale Abfall, der von der Form des einzelnen Glottispulses abhängt, und die sich ergebenden spektralen Linien, die auf die periodische Anregung zurückgehen.

2.2. Physik des Stimmkanals

Abbildung 2.3 zeigt eine detaillierte, computertomographische Aufnahme des menschlichen Stimmkanals ohne angekoppelte Nasenhöhle. Die typische gekrümmte Form ist gut zu erkennen. Die Aufnahme zeigt im Rachenraum kompliziert geformte Strukturen. Die Anatomie des Stimmkanals weicht deutlich von einem runden oder elliptischen Querschnitt ab. Mit der heute zur Verfügung stehenden Rechenkapazität entstehen zur Zeit komplizierte Modelle zur Sprachsynthese, die die natürliche Form des Stimmkanals im Detail annähern und die dreidimensionale Wellenausbreitung berücksichtigen [64]. Doch gute Ergebnisse lassen sich bereits unter Annahme von einigen starken Vereinfachungen erzielen. Im Allgemeinen werden für die Schallausbreitung im Stimmkanal folgende Näherungen angenommen. Zur Beschreibung der Akustik wird die lineare Wellengleichung angesetzt. Nur bei den Stimmlippen und in der Nähe von starken Einschnürungen spielen nichtlineare Effekte eine Rolle.

Der Stimmkanal wird gestreckt. Die Ausbreitung des Schalls wird durch ebene Wellen beschrieben. Man kann zeigen, dass bei Vernachlässigung der Krümmung z. B. die Resonanzfrequenzen nahezu unverändert bleiben [49]. Höhere Moden sind bei einem maximalen Durchmesser von ca. 5 cm erst ab 4 kHz ausbreitungsfähig. Der Großteil der Energie des Sprachsignals liegt im Frequenzbereich bis 4 kHz.

Die Geometrie des Vokaltraktes reduziert sich somit auf die eines entsprechenden inhomogenen Rohres und wird durch die Querschnittsfunktion $A(x,t)$ vollständig bestimmt.

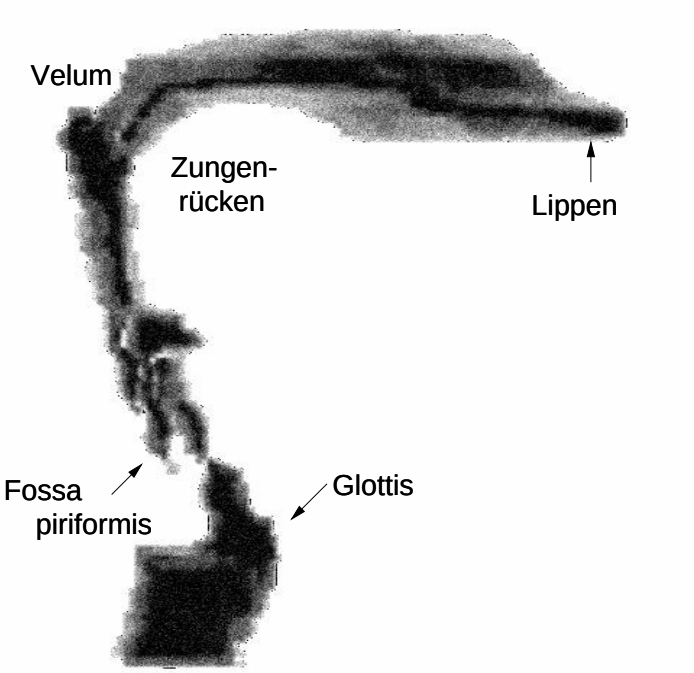


Abbildung 2.3.: Detaillierte Ansicht des Stimmkanals ohne Nasenhöhle nach einer Computertomographie, aufgenommen von Story et al. [51].

2.2.1. Das inhomogene, verlustlose, hartwandige Rohr

Für den Fall einer inhomogenen Röhre mit schallharten Wänden werden im Folgenden die akustischen Feldgleichungen hergeleitet. Die eindimensionale Schallausbreitung an der Stelle x wird durch den über den Querschnitt $A(x,t)$ gemittelten Schalldruck $p(x,t)$ und die über $A(x,t)$ integrierte Schallschnelle, den Schallfluss $u(x,t)$, beschrieben. Bei Vernachlässigung einer Gleichströmung Die Bewegung wird durch Newtons Gesetz (linearisierte Gleichung der Impulserhaltung)

$$\varrho_0 (u/A)^{\cdot} = -p', \tag{2.2}$$

der eindimensionalen linearisierten Gleichung der Massenerhaltung

$$(\varrho A)^{\cdot} + \varrho_0 \dot{A} = -\varrho_0 u' \tag{2.3}$$

(Gleichung 2.3) und der Zustandsgleichung

$$p = c^2 \varrho \tag{2.4}$$

beschrieben. Nur lineare Abhängigkeiten von den kleinen fluktuierenden Größen Dichte ϱ, Schallfluss u und Schalldruck p werden berücksichtigt. ϱ_0 ist die konstante durchschnittliche Dichte und c die Schallgeschwindigkeit. Punkte bezeichnen die partielle Ableitung nach der Zeit t, gestrichene Größen werden nach der Ortskoordinate x abgeleitet.

Gleichungen 2.3 und 2.4 können zu

$$(pA)^{\cdot}/\varrho_0 c^2 = -u' \tag{2.5}$$

kombiniert werden. Leitet man nun Gleichung 2.2 nach dem Ort x und Gleichung 2.5 nach der Zeit ab und setzt ineinander ein, so erhält man für eine zeitlich konstante Querschnittsfunktion $A(x)$ die Webstersche Horngleichung

$$p'' + (A'/A)p' - c^{-2}\ddot{p} = 0. \tag{2.6}$$

Sie gleicht bis auf den Term $A'/A = \frac{\partial \log A}{\partial x}$ der eindimensionalen Wellengleichung für homogene Querschnittsfunktionen. Die Darstellung nach zeitlicher Fouriertransformation (P ist die Fouriertransformierte des Schalldruckes)

$$P'' + (A'/A)P' + \lambda^2 P = 0 \tag{2.7}$$

lässt erahnen, dass die Eigenwerte $\lambda = (\omega/c)^2$ dieser Differentialgleichung, aus denen sich die Resonanzfrequenzen

$$f_m = c\sqrt{\lambda_m}/2\pi \tag{2.8}$$

des Stimmkanals, die sogenannten Formantfrequenzen, ergeben, nicht von der Querschnittsfunktion selbst, sondern von der logarithmierten Querschnittsfunktion abhängen.

Die Gleichungen 2.2 und 2.5 stehen in Analogie zu den Feldgleichungen einer verlustlosen elektrischen Leitung. Hierbei wird p mit der Spannung identifiziert und u mit dem Strom. Die Variablen

$$L' = \varrho_0/A \qquad \text{und} \qquad C' = A/\varrho_0 c^2 \tag{2.9}$$

entsprechen der Induktivitätsdichte beziehungsweise der Kapazitätsdichte. Das entsprechende infinitesimale Leitungssegment ist in Abb. 2.4 dargestellt.

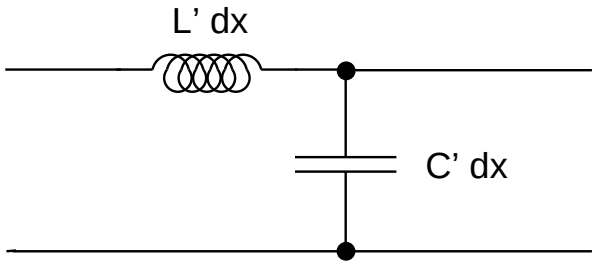


Abbildung 2.4.: Ersatzschaltbild eines infinitesimalen verlustlosen Rohrelementes.

2.2.2. Modellierung des verlustbehafteten Stimmkanals

Im Stimmkanal gibt es eine Reihe von Verlustmechanismen. Die wichtigsten sind viskose Reibung an den Wänden und Wärmeleitung. In dem im letzten Abschnitt eingeführten elektrischen Ersatzschaltbild (Abb. 2.4) können Verluste durch zusätzliche Widerstände realisiert werden (vgl. Abb. 2.5). Der Impedanzdichte R sind viskose Verluste an den Wänden des Stimmkanals zugeordnet. Der Admittanzdichte G wird der Verlust auf Grund von Wärmeleitung zugeordnet [17]. Ein weiterer Verlustmechanismus, die nicht schallharten Wände, wird durch Einfügen eines LRC-Serienresonanzkreises als Parallelschaltung zu C' und G' berücksichtigt. Dieser Effekt wirkt sich vor allem auf den ersten Formanten aus, der zu höheren Frequenzen verschoben wird. In diesem Ab-

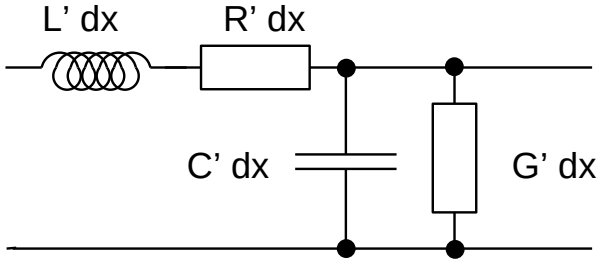


Abbildung 2.5.: Ersatzschaltbild eines infinitesimalen Rohrelementes unter Berücksichtigung von Verlusten auf Grund viskoser Reibung und Wärmeleitung.

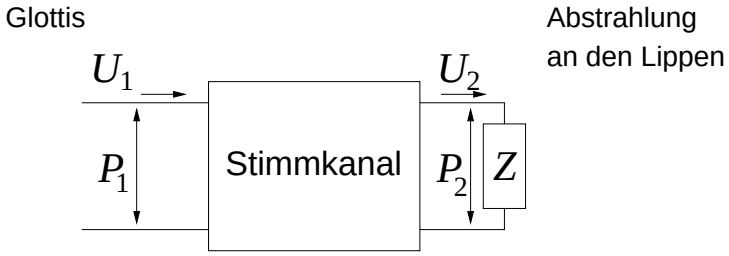


Abbildung 2.6.: Vierpolmodell des Stimmkanals. Die komplexen Eingangsgrößen, der Schalldruck P_1 und des Schallfluss U_1 können über eine Kettenmatrix, die den Stimmkanal beschreibt, mit den entsprechenden Ausgangsgrößen P_2 und U_2 verknüpft werden.

schnitt wird eine mögliche vierpoltheoretische Beschreibung (Abb. 2.6) des Vokaltraktes im Frequenzbereich dargestellt. Von der Darstellung des infinitesimalen Rohrleitungssegmentes gelangt man zu einer oft gebrauchten Modellierung des Stimmkanals (z. B. [2, 5, 17, 31, 63, 65]): Die Zerlegung in Segmente mit homogenem Querschnitt, die sich mit Hilfe sogenannter Kettenmatrizen beschreiben lassen.

Für ein infinitesimales Rohrleitungssegment gilt für den komplexen Schalldruck P und den komplexen Schallfluss U nach Flanagan [17]

$$\mathrm{d}\,U = -P\,y\,\mathrm{d}\,x \quad \text{und} \quad \mathrm{d}\,P = -U\,z\,\mathrm{d}\,x \tag{2.10}$$

mit $y = (G' + i\omega C')$ und $z = (R' + i\omega L')$. Der Schalldruck P und Schallfluss U erfüllen die Gleichungen

$$\frac{\mathrm{d}^2 P}{\mathrm{d}x^2} - zyP = 0 \quad \text{und} \quad \frac{\mathrm{d}^2 U}{\mathrm{d}x^2} - zyU = 0. \tag{2.11}$$

Die Lösungen der Differentialgleichungen sind

$$P = A_1 \exp(\gamma x) + B_1 \exp(-\gamma x) \tag{2.12}$$
$$U = A_2 \exp(\gamma x) + B_2 \exp(-\gamma x) \tag{2.13}$$

mit dem Ausbreitungskoeffizienten

$$\gamma = (z\,y)^{\frac{1}{2}} = (\alpha + \mathrm{i}\beta) \tag{2.14}$$

und den Integrationskonstanten A_1, A_2, B_1, B_2, die durch Randbedingungen bestimmt werden. Liegen als Eingangsgrößen die Spannung P_1 und der Strom U_1 und als Ausgangsgrößen P_2 und U_2 an, so ergibt sich das Gleichungssystem [17]

$$\begin{pmatrix} P_1 \\ U_1 \end{pmatrix} = \mathbf{T} \begin{pmatrix} P_2 \\ U_2 \end{pmatrix} \qquad \text{mit} \qquad \mathbf{T} = \begin{pmatrix} \cosh(\gamma l) & Z_0 \sinh(\gamma l) \\ Z_0^{-1} \sinh(\gamma l) & \cosh(\gamma l) \end{pmatrix}. \tag{2.15}$$

Dabei ist

$$Z_0 = (z/y)^{\frac{1}{2}} \tag{2.16}$$

die charakteristische Impedanz der Leitung. Die Matrix $\mathbf{T}$ beschreibt ein homogenes Rohrsegment. Ein inhomogenes Rohr wird aus diesen diskreten homogenen Rohrsegmenten unterschiedlichen Querschnittes zusammengesetzt: Für das Segment n gelte

$$\begin{pmatrix} P_n \\ U_n \end{pmatrix} = \mathbf{T}_n \begin{pmatrix} P_{n+1} \\ U_{n+1} \end{pmatrix}. \tag{2.17}$$

Die Gesamtmatrix gewinnt man durch Aufmultiplizieren der Teilmatrizen T_n:

$$\mathbf{A} = \mathbf{T}_1 \cdots \mathbf{T}_n = \mathbf{A}_{n-1} \mathbf{T}_n. \tag{2.18}$$

2×2 Matrizen der angegebenen Art werden als Kettenmatrizen bezeichnet.

Bei bekannter Abschlussimpedanz $Z = \frac{P_2}{U_2}$ ergibt sich bei bekannter Gesamtkettenmatrix $\mathbf{A}$ des Stimmkanals die Schallflussübertragungsfunktion H_U oder die Eingangsimpedanz Z_1 des Systems zu

$$H_U = H_U(Z, \mathbf{A}) = \frac{U_2}{U_1} = \frac{1}{a_{21}\, Z + a_{22}} \tag{2.19}$$

$$\text{bzw.} \qquad Z_1 = Z_1(Z, \mathbf{A}) = \frac{P_1}{U_1} = \frac{a_{11}\, Z + a_{12}}{a_{21}\, Z + a_{22}}. \tag{2.20}$$

Komplizierte Strukturen mit Seitenarmen, wie in Abbildung 2.1 dargestellt, werden mit Hilfe einer Kopplungsmatrix

$$\mathbf{Q}(Y) = \begin{pmatrix} 1 & 0 \\ Y & 1 \end{pmatrix}, \tag{2.21}$$

die die Admittanz Y des Seitenarmes in das Röhrenmodell einkoppelt, realisiert. Beispielhaft sei die Beschreibung des Systems Rachen A, Mundhöhle B (mit Eingangsadmittanz Y_B) und Nasenhöhle D gegeben:

$$\mathbf{M} = \mathbf{A}\mathbf{Q}_D\mathbf{B}. \tag{2.22}$$

Die Admittanz des Seitenarmes berechnet sich nach Gleichung 2.20. Da der Schallfluss additiv ist, ergibt sich die Gesamtübertragungsfunktion aus der Summe der Übertragungsfunktionen der möglichen Ausbreitungswege. Für einen nasalierten Vokal ist die schrittweise Berechnung der Übertragungsfunktion in Abbildung 2.7 dargestellt. Für

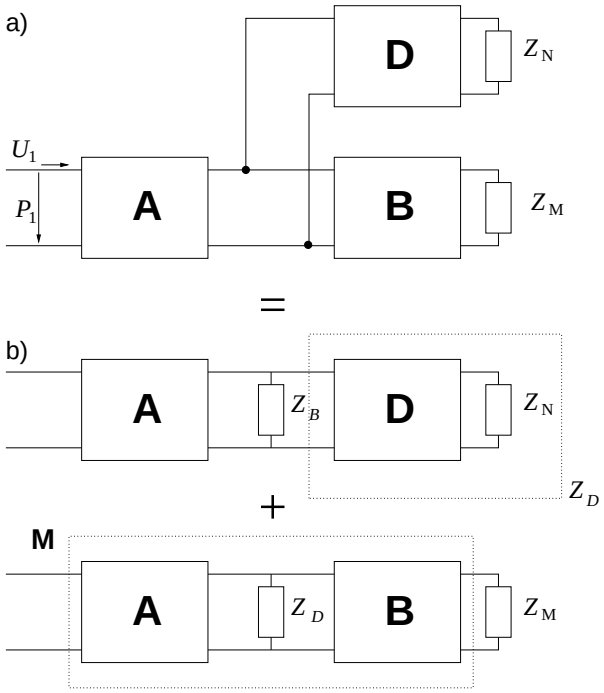


Abbildung 2.7.: Schrittweise Zusammensetzung eines Vierpolnetzwerkes (a), das einen verzweigten Stimmkanal darstellt. In Abbildung (b) ist die Einkopplung der Impedanzen der jeweiligen Seitenarme abgebildet. Die Übertragungsfunktionen beider Äste werden addiert, da der Schallfluss additiv ist.

die vorliegenden Arbeit wurde zusammen mit Kaufmann [25] ein Synthetisator im Frequenzbereich nach Sondhi und Schroeter implementiert [31]. Er verbindet realistische Dämpfung, die viskose Reibung, Wärmeleitung und mitschwingende Wände berücksichtigt, mit einer freien Längenskalierbarkeit der einzelnen Segmente. Bei der Berechnung der Gesamtkettenmatrix wird von einem schnellen Iterationsverfahren nach Strube [53]

Gebrauch gemacht. Der Ausdruck für die Gesamtkettenmatrix **A** des Systems reduziert sich auf das Ausmultiplizieren von zwei Polynomen in ω.

Die Abstrahlung an den Lippen stellt in guter Näherung die Impedanz eines Kolbenstrahlers in einer Kugel dar, wofür es keinen geschlossenen Ausdruck gibt [17]. Im Vokaltraktmodell dieser Arbeit wird die Näherung

$$\varrho c\, Z(\omega) = \frac{\varrho_0 \omega^2}{4\pi c}\, K_s(\omega) + \mathrm{i}\, \frac{8\,\varrho_0 \omega}{3\pi\sqrt{\pi A}} \tag{2.23}$$

von Wakita und Fant [61] benutzt. $A = S_n$ entspricht der Fläche des Kolbens und ist identisch mit der Querschnittsfläche S_n des letzten Traktabschnittes. Für $K_s(\omega)$ gilt die Näherung

$$K_s(\omega) = \begin{cases} \dfrac{0,6\,\omega}{\omega_0} + 1 & \text{für} \quad 0 \le \omega < \omega_0, \\ 1,6 & \text{für} \quad \omega \ge \omega_0 \end{cases} \tag{2.24}$$

mit $\omega_0 = 2\pi \cdot 1600\,\mathrm{Hz}$.

2.3. Das inverse Problem zur Sprachproduktion

Das direkte Problem der Sprachproduktion wurde im letzten Abschnitt behandelt und eine mögliche Implementation beschrieben. In diesem Abschnitt wird die Natur der umgekehrten Fragestellung, nämlich der akustisch–artikulatorischen Abbildung beschrieben, die kurz als das inverse Problem bezeichnet wird.

Im letzten Abschnitt wurden Formanten als Resonanzen des Stimmkanals (Gleichung 2.8) eingeführt. Die Betrachtungen von Schroeder [44] illustrieren die Beziehung zwischen Formanten und Querschnittsfunktion. Schroeder übertrug das Ehrenfestsche Adiabatentheorem[1] auf die kleinen Störungen eines homogenen Stimmkanals

$$\frac{\delta f}{f} = \frac{\delta W}{W}. \tag{2.25}$$

$\delta f/f$ ist die relative Abweichung der Formantfrequenz, die Arbeit δW wird gegen den Schallstrahlungsdruck auf die Wände verrichtet, W ist die Gesamtenergie in der Resonanz. Das Ergebnis der Störungstheorie erster Ordnung ist, dass im verlustlosen Vokaltrakt der m-te Formant direkt und ausschließlich mit dem m-ten ungeraden Koeffizienten der Kosinusreihenentwicklung der logarithmierten Querschnittsfunktion zusammenhängt:

$$\ln A(x)/A_0 = \sum_m^\infty a_m \cos(\pi m x/L), \tag{2.26}$$

$$\frac{a_{2n-1}}{2} = -\frac{\delta f_n}{f_n},\ n = 1, 2, 3, \ldots\,. \tag{2.27}$$

[1] Bei dem adiabatischen Prinzip $\Delta(W/f) = 0$ handelt es sich um ein sehr allgemeines physikalisches Prinzip, denn es beschreibt das Verhalten beliebiger Oszillatoren unter langsamer (adiabatischer) Änderung der Parameter.

Obwohl diese Näherung streng betrachtet nur für kleine Abweichungen von der homogenen Querschnittsfunktion gültig ist, geht aus [44] hervor, dass sie auch für große Störungen gute Ergebnisse liefert. Daher hat sich die Methode als Schätzverfahren für die grobe Querschnittsfunktion ohne die geraden Anteile der Fourierterme etabliert. Als Beispiel aus letzter Zeit sei [9] angeführt. Gleichung 2.26 zeigt auf eindrucksvolle Weise, dass die geraden Fourierterme von $\ln(A(x)/A_0)$ in erster Näherung keine Auswirkungen auf die Formantmuster haben. Es folgt, dass selbst bei Kenntnis aller (unendlich vielen) Formantfrequenzen in gewisser Weise die Hälfte der Querschnittsinformation fehlt.

Zur Konstruktion der Querschnittsfunktion werden zwei Datensätze benötigt. Mögliche Paare sind im Folgenden aufgeführt.

Zwei Sätze von Eigenwerten: Borg [6] zeigte für eine idealisierte Form des Problems, dass zwei komplette Sätze von Eigenwerten vorliegen müssen, um $A(x)$ aus der Gleichung der Art 2.7 zu bestimmen. Konkret könnten dies die Formantfrequenzen zu den Randbedingungen a) Abschluss des Vokaltraktes mit bekannter Strahlungsimpedanz Z_L und b) schallharter Abschluss des Vokaltraktes (geschlossene Lippen) sein.

Pole und Nullstellen der Eingangsimpedanz des Stimmkanals: Schroeder [44] nutzte die Pole und Nullstellen der von außen gemessenen Eingangsimpedanz an den Lippen. Diese entsprechen den zwei Formantensätzen für geschlossene Lippen und geöffnete Lippen.

Formanten und Bandbreiten: Im Fall eines dämpfungsbehafteten Abschlusses bestehen die Datensätze aus Formanten und den dazugehörigen Bandbreiten. Allerdings können nur drei bis vier Formanten und ihre Bandbreiten mit akzeptabler Genauigkeit aus gesprochenen Vokalen bestimmt werden, so dass die Genauigkeit dieser Methode sehr beschränkt ist: Atal et al. [2] gibt vier sehr verschiedene, durch Simulationen ermittelte Querschnittsfunktionen an, die in den ersten drei Resonanzfrequenzen und Bandbreiten fast exakt übereinstimmen.

Alle genannten Datensätze sind im wesentlichen äquivalent (Strube [54]). Sie haben gemeinsam, dass die Größen schwierig aus dem Sprachsignal zu bestimmen sind (Formanten und Bandbreiten) oder sich gar nicht ermitteln lassen (zwei Formantsätze oder Lippeneingangsimpedanz). Einen Ausweg aus dem Dilemma suchen vielfältige empirische Methoden. Gemein ist diesen Ansätzen die Einführung sinnvoller Einschränkungen, wie z. B. Parametrisierung der Querschnittsfunktionen mit Hilfe eines artikulatorischen Modells und *Codebooks* von Querschnitten. Eine weitere wichtige Information liefert der zeitliche Kontext der Querschnittsfunktion. Eine Auswahl an Verfahren dieser Art bietet der Überblicksartikel von Schroeter und Sondhi [46].

3. Ein neuer Ansatz zur Bestimmung von sprecherspezifischen Parametern

Die Problematik des inversen Problems zur Sprachproduktion wurde in Abschnitt 2.3 kurz dargestellt. Eine eindeutige Rekonstruktion der Querschnittsfunktion ist dem Sprachsignal nicht ohne weiteres zu entnehmen. In dieser Arbeit wird ein neuer Ansatz entwickelt und erprobt: die akustischen Abweichungen eines gegebenen Lautes, der von einem Testsprecher mit unbekannter Querschnittsfunktion gesprochen wurde, werden als kleine Störungen der bekannten Querschnittsfunktion des Referenzsprechers aufgefasst. Dieser Ansatz beschränkt sich nur auf die Analyse von bestimmten Anteilen des Sprachsignals, für die erstens verlässliche akustische Daten existieren und es zweitens bekannte Querschnittsfunktionen eines Referenzsprechers gibt. Unter diesen Voraussetzungen bieten sich Vokale als aussagekräftige Sprachdaten an, denn Vokale geben mit den charakteristischen Einschnürungen und Volumina der Querschnittsfunktion gute Einblicke in die anatomischen Dimensionen des Stimmkanals und bieten durch die dazugehörigen Formanten eine einfache, gut interpretierbare akustische Repräsentation. Hinzu kommt, dass detaillierte Messungen von Querschnittsfunktionen von Vokalen vorliegen.

3.1. Störungen der Querschnittsfunktion in Querschnitt und Länge

Dieser Abschnitt ist der Frage gewidmet, wie sich kleine Störungen einer beliebigen Querschnittsfunktion auf das Sprachsignal auswirken. Gesucht wird eine Verallgemeinerung der von Schroeder [44] gefundenen linearen Beziehungen zwischen $\delta\omega/\omega$ und kleinen Querschnittsstörungen $\delta A/A$ der homogenen Querschnittsfunktion. Eine Darstellung für den Fall diskretisierter Querschnittsfunktionen beliebiger Gestalt findet man bei Fant [15]. Fant untersuchte ein Vokaltraktmodell, das aus einem LC-Leiternetzwerk bestand, unter dem Einfluss von kleinen Querschnitts- und Längenstörungen. Das Ergebnis der Untersuchung sind zwei lineare Beziehungen zwischen den Größen $\tilde{A}_i = (\Delta A/A)_i$ und auch k_i (Gleichung 3.3) und den relativen Formantabweichungen $\tilde{\omega} = \Delta\omega/\omega$:

$$\tilde{\omega} = \frac{\sum_i \tilde{A}_i L_i}{\sum_i H_i} \quad \text{und} \tag{3.1}$$

$$\tilde{\omega} = \frac{\sum_i k_i H_i}{\sum_i H_i}\,. \tag{3.2}$$

$H_i = T_i + V_i$ und $L_i = T_i - V_i$ sind Summe und Differenz der potentiellen Energiedichte V_i und der kinetischen Energiedichte T_i in der Resonanz. Die Variable k_i verbindet eine monotone, nichtlineare Beziehung mit der relativen Längenstörung $\tilde{l}_i = \Delta l_i / l_i$ der Segmentlänge l_i. Es gilt

$$k_i = -\frac{\tilde{l}_i}{1+\tilde{l}_i}, \quad \text{bzw.} \tag{3.3}$$

$$\tilde{l}_i = -\frac{k_i}{1+k_i}\,. \tag{3.4}$$

Strube [54] erhielt aus der Websterschen Horngleichung (Gleichung 2.7) durch Variationsrechnung eine äquivalente Formulierung für die Querschnittsskalierung

$$\frac{\delta\omega}{\omega} = \frac{\int_0^l L(x)\frac{\delta A}{A}\,\mathrm{d}x}{\int_0^l H(x)\,\mathrm{d}x}\,. \tag{3.5}$$

Eine entsprechende Gleichungen für monotone Störungen ($x \to x + \delta x(x)$, $\delta x(0) = 0$, $\delta x' > -x'$) der x-Achse erhält Strube durch die Substitution $\delta A = -A'\delta x$ in Gleichung 3.5 und partielle Integration:

$$\frac{\delta\omega}{\omega} = -\frac{\int_0^l H(x)\,\delta x'(x)\,\mathrm{d}x}{\int_0^l H(x)\,\mathrm{d}x}\,. \tag{3.6}$$

Die Grössen $H(x)$ und $L(x)$ sind die kontinuierliche totale Energiedichte bzw. die Lagrangedichte. Eine lineare Variation der Gesamtlänge ergibt $\delta\omega/\omega = -\delta l/l$. Die Gleichung 3.6 von Strube stellt eine konsequente Näherung erster Ordnung dar, die eine Längenentzerrung der in Gleichung 3.4 dargestellten Art nicht berücksichtigt. Der Unterschied zwischen k bzw $\delta x'$ macht sich erst bei größeren Störungen bemerkbar.

3.2. Schätzung von Längen- und Querschnittsstörungen aus dem Formantmuster

In diesem Abschnitt wird das Verfahren zur Schätzung von Querschnitts- und Längenstörungen entwickelt. Es besteht aus einer Umkehrung der linearen Störungstheorie, die im vorhergehenden Abschnitt zusammengefassend dargestellt wurde. Es wird im Folgenden vorausgesetzt, dass die Störungen der einzelnen Segmente i der Querschnittsfunktion des Referenzsprechers in der Länge k_i und im Querschnitt $\tilde{A}_i$ nicht unabhängig voneinander sind, sondern in einer Anzahl von Störungsvektoren $\mathbf{v}_j$ mit $j = 1, \ldots, N$ zusammengefasst werden können. Ein triviales Beispiel für einen solchen Vektor $\mathbf{v}$ ist folgendes: $\mathbf{v}_1 = (1\,1\,1\,\ldots\,1)^T$. Dieser Störungsvektor bedingt eine lineare Längenskalierung der Querschnittsfunktion. Beliebige lineare Längenskalierungen werden durch die Wichtung mit einem Faktor a erreicht:

$$\tilde{\omega}_m = \frac{\sum_i H_{mi} a\, v_{1i}}{\sum_i H_{mi}}. \tag{3.7}$$

Interessanter ist der Fall der nichtgleichmässigen Längenstörung mit mehreren Vektoren in Verbindung mit einer gleichzeitigen Querschnittsskalierung durch den Vektor $\mathbf{v}_1$. Dabei ist $\mathbf{v}_2 = (1\,1\,\ldots\,0\,0)^T$ so konstruiert, dass nur der Rachenraum skaliert wird, und $\mathbf{v}_3 = (0\,0\,\ldots\,1\,1)^T$ beeinflusst nur die Mundhöhle. Es ergibt sich ein lineares Gleichungssystem

$$\begin{pmatrix} \tilde{\omega}_1 \\ \tilde{\omega}_2 \\ \tilde{\omega}_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{H_{1\cdot}}\sum_i L_{1i}v_{1i} & \frac{1}{H_{1\cdot}}\sum_i H_{1i}v_{2i} & \frac{1}{H_{1\cdot}}\sum_i H_{1i}v_{3i} \\ \frac{1}{H_{2\cdot}}\sum_i L_{2i}v_{1i} & \frac{1}{H_{2\cdot}}\sum_i H_{2i}v_{2i} & \frac{1}{H_{2\cdot}}\sum_i H_{2i}v_{3i} \\ \frac{1}{H_{3\cdot}}\sum_i L_{3i}v_{1i} & \frac{1}{H_{3\cdot}}\sum_i H_{3i}v_{2i} & \frac{1}{H_{3\cdot}}\sum_i H_{3i}v_{3i} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}. \tag{3.8}$$

Die totale Energiedichte der Resonanz m wurde mit $H_{m\cdot} = \sum_i H_{mi}$ bezeichnet. Im Folgenden wird die Aufstellung eines allgemeinen linearen Gleichungssystems für eine beliebige Anzahl von Formanten und Störungsvektoren beschrieben. Ausgehend von Gleichungen 3.2 und 3.1 werden die Störungen der Längen bzw. Querschnitte der einzelnen Segmente, k_i und $\tilde{A}_i$, in eine Basis von Störungsvektoren $\mathbf{v}_j$ entwickelt. Die Position im Vokaltrakt wird durch den Index i bezeichnet, und $j = 1, \ldots, N$ nummeriert die Basisvektoren der Störung.

$$\tilde{A}_i = a_1 v_{1i} + a_2 v_{2i} + \ldots + a_q v_{q_i} \tag{3.9}$$

$$k_i = a_{q+1} v_{q+1_i} + a_{q+2} v_{q+2_i} + \ldots + a_N v_{Ni}\,. \tag{3.10}$$

Die Basisvektoren $\mathbf{v}_1$ bis $\mathbf{v}_q$ bewirken eine Querschnittsstörung, $\mathbf{v}_{q+1}$ bis $\mathbf{v}_N$ gehören zur Längenstörung. Fasst man die Gleichungen der Längen- und Querschnittsskalierung für M verschiedene Resonanzfrequenzen zusammen, so erhält man ein lineares Gleichungssystem der Art

$$\mathbf{w} = \mathbf{E}\mathbf{a} \tag{3.11}$$

$$\text{mit} \quad \mathbf{w} = \begin{pmatrix} \tilde{\omega}_1 \\ \vdots \\ \tilde{\omega}_M \end{pmatrix}, \quad (\mathbf{E})_{mn} = \begin{cases} \sum_i L_{mi} v_{ni} / \sum_i H_{mi}, & n \le q, \\ \sum_i H_{mi} v_{ni} / \sum_i H_{mi}, & n > q, \end{cases} \tag{3.12}$$

$$\text{und} \quad \mathbf{a} = \left(a_1, \cdots a_q,\, a_{q+1}, \cdots a_M \right)^T. \tag{3.13}$$

Bei gegebenen relativen Abweichungen $\mathbf{w}$ kommt man über die Invertierung der Matrix $\mathbf{E}$ zu den Skalierungsparametern $\mathbf{a}$, die die Störungsvektoren $\mathbf{v}_j$ parametrisieren. Die Parameter $\mathbf{a}$ werden als sprecherspezifische Parameter interpretiert, da sie die Abweichung vom Referenzsprecher beschreiben und somit den Testsprecher charakterisieren.

3.2.1. Invertierung des Gleichungssystems

Die Matrix $\mathbf{E}$ muss nicht notwendigerweise eine nichtsinguläre, quadratische Matrix sein. Wählt man die Anzahl der Störungsvektoren N größer als die Anzahl M der Resonanzfrequenzen, so erhalt ein Unterbestimmtes Gleichungssystem und im Falle $M > N$ ein überbestimmtes Gleichungssystem. Im Falle $M = N$ ist das Gleichungssystem zwar eindeutig bestimmt, $\mathbf{E}$ kann aber durch eine ungünstige Wahl der Störungsvektoren $\mathbf{v}_j$

singulär werden, sodass keine Lösung für $\mathbf{a}$ existiert. Eine elegante Lösung bietet die Pseudoinverse. Die Pseudoinverse ist die Verallgemeinerung der Inversen von quadratischen, nichtsingulären Matrizen auf beliebige rechteckige Matrizen. Die Pseudoinverse kann auf verschiedene Weisen eingeführt werden [26]. An dieser Stelle wird die Definition über die Singulärwertzerlegung (SVD – singular value decomposition) gewählt.

Zu der Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ mit $r := \mathrm{Rang}(\mathbf{A})$ sei

$$\mathbf{U}^T \mathbf{A} \mathbf{V} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min(m,n)}) =: \Sigma \tag{3.14}$$

eine existierende Singulärwertzerlegung von $\mathbf{A}$ mit orthogonalen Matrizen $\mathbf{U} \in \mathbb{R}^{m \times m}$, $\mathbf{V} \in \mathbb{R}^{n \times n}$ und

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_{\min(m,n)} = 0 \tag{3.15}$$

Mit der $n \times n$-Matrix

$$\Sigma^+ := \begin{pmatrix} 1/\sigma_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & 1/\sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} \tag{3.16}$$

heißt

$$\mathbf{A}^+ := \mathbf{V} \Sigma^+ \mathbf{U}^T \in \mathbb{R}^{n \times m} \tag{3.17}$$

Pseudoinverse von $\mathbf{A}$.

Die Schätzung der Skalierungsparameter $\mathbf{a}$ wird durch (Pseudo-)Invertierung der Matrix $\mathbf{E}$ aus 3.11 mittels einer Singulärwertzerlegung (SVD) durchgeführt. Die Pseudoinverse hat bei Über- bzw. Unterbestimmung des Gleichungssytems $\mathbf{E}$ folgende Eigenschaften: Ist das Gleichungssystem nicht eindeutig lösbar, so findet die Pseudoinverse den Vektor $\mathbf{a}$ mit minimaler Norm. Ist das Gleichungssystem nicht lösbar, so findet man mit Hilfe der Pseudoinversen die Lösung des linearen Ausgleichsproblems mit $|\mathbf{E}\mathbf{a} - \mathbf{w}| \stackrel{!}{=} \min$ [26, 62].

Der Vorteil bei der Berechnung der Pseudoinversen über die SVD ist der, dass durch Manipulationen der σ_i gezielt die Genauigkeit der Berechnung des Vektors $\mathbf{a}$ beeinflußt werden kann: Ersetzt man $1/\sigma_i$ aus Gleichung 3.16 für kleine σ_i mit Null, so wird die dazugehörende Linearkombination des zu lösenden Satzes von Gleichungen bei der Berechnung von $\mathbf{a}$ nicht berücksichtigt [41]. Es hat sich in diesem Zusammenhang herausgestellt, dass es von Vorteil ist, sehr große $1/\sigma_i$ mit Null zu ersetzen, da der Lösungsvektor $\mathbf{a}$ so im Bereich der Gültigkeit der in den obigen Abschnitten gemachten linearen Näherung der Störungstheorie bleibt. Statt sehr große $1/\sigma_i$ ganz auf Null zu setzen wurde in den Experimenten dieser Arbeit $1/\sigma_i$ aus Gleichung 3.16 durch den Term

$$\frac{\sigma_i}{\sigma_i^2 + \sigma_{\mathrm{n}}^2} \tag{3.18}$$

ersetzt. Dies stellt eine Interpolation zwischen den Extremen dar: Für kleine σ_i nähert dieser Term sich Null, für große σ_i ergibt sich $1/\sigma_i$. Dieser Term heißt Wiener-Korrektur und ist formal analog zum Wiener Filter, das z. B. für die Korrektur verrauschter Bilder [48] genutzt wird. Der Wert σ_n^2 wird empirisch ermittelt. In dieser Arbeit wurde der Parameter folgendermaßen in Beziehung zu den Singulärwerten σ_i gesetzt:

$$\sigma_\mathrm{n}^2 = \epsilon_\mathrm{SVD} \sum_i {\sigma_i}^2 \,. \tag{3.19}$$

Bei einer Basis aus sechs Störungsvektoren wurde ϵ_SVD in den präsentierten Experimenten zwischen 10^{-9} und 0,015 gewählt.

3.2.2. Wahl der Störungsbasis

Eine erfolgreiche Schätzung der Parameter **a** hängt stark von einer geeigneten Wahl der Störungsbasis ab. Bei einem homogenen Querschnitt des Referenzsprechers macht es z. B. keinen Sinn, eine nichtlineare Längenskalierung einzuführen oder die gerade indizierten Glieder der Kosinusfunktion als Basisvektoren der Querschnittsstörung zu verwenden, die in erster Näherung keine Auswirkung auf die Formanten haben [44]. Ein einfaches Beispiel der Störungsbasis sind die ungerade indizierten Glieder der Kosinusfunktionen

$$\mathbf{v}_n(x) = \cos\left[(2n-1)\pi x / L_\mathrm{VT}\right] \tag{3.20}$$

mit $n = 1, 2, \ldots, q$. In Gleichung 3.20 ist x die Position im Vokaltrakt und L_VT die Vokaltraktgesamtlänge. Im Fall eines homogenen Vokaltraktes enspricht ein Vektor der angegebenen Art der Verschiebung genau eines Formanten (Schroeder [44]). Bei der Störung einer vom homogenen Querschnitt unterschiedlichen Geometrie ist diese Näherung nicht mehr gültig. Die Berechnung von Eigenfunktionen des Vokaltraktes mit komplizierter Geometrie findet sich bei Heinz [23]. Stehen genügend Querschnittsfunktionen von unterschiedlichen Sprechern zur Verfügung, so liegt es nahe, als Störungsbasis die Hauptkomponenten der Querschnittsfunktionen zu benutzen.

Die Längenstörung wird in dieser Arbeit entweder linear angesetzt oder als separate Skalierung von zwei Hälften vorgenommen.

In Kapitel 4 sind die aufgezählten Möglichkeiten der Wahl der Störungsbasen experimentell umgesetzt worden. Auf Grund der Flexibilität des Ansatzes sind darüber hinaus beliebig viele andere Konstruktionen von Störungsbasen möglich.

3.2.3. Realisation des Verfahrens

In Abbildung 3.1 ist das gesamte Verfahren zur Schätzung des Skalierungsparameter im Überblick dargestellt. Die Eingangsdaten bestehen aus einem Satz Formanten und die dazugehörigen Information, um welchen Vokal es sich handelt. Die Formantfrequenzen des Referenzsprechers ergeben sich über eine Nullstellensuche der Kehrwertes der nach Gleichung 2.19 berechneten Schallflussübertragungsfunktion $1/H_U$. Eine effektive Methode der Berechnung findet sich bei Fant ([13], S. 41 f.). Die Energie- und Lagrangedichtefunktionen, die zu diesen Resonanzfrequenzen gehören, werden mit Hilfe des

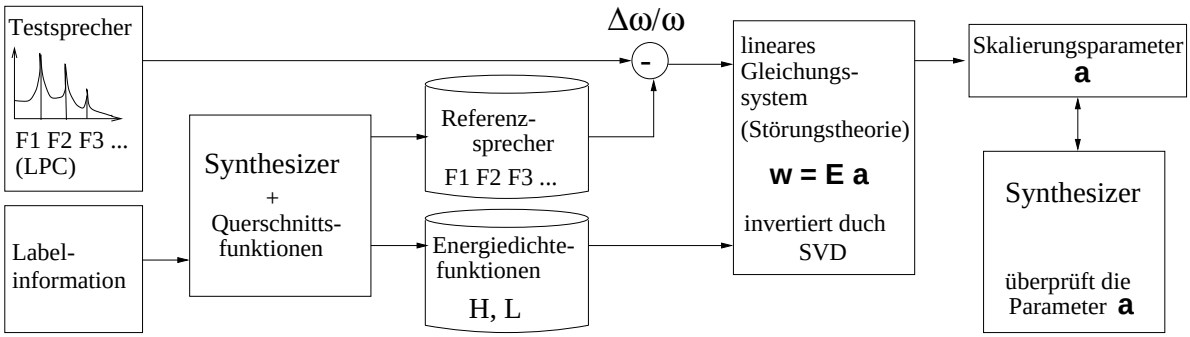


Abbildung 3.1.: Das neue Verfahren zur Schätzung von sprecherspezifischen Parametern aus Formantmustern.

Synthetisators [31] bestimmt. Es gilt $V_i = C_i'|P|^2/2$ und $T_i = L_i'|U|^2/2$. P und U sind der komplexe Schallfluss und Schalldruck. C_i' und L_i' sind Kapazitätsdichte bzw. Induktivitätsdichte des Rohrsegmentes i. Kapazitätsdichte und Induktivitätsdichte können entweder als verlustlose Näherung über die Gleichungen 2.9 aus den Querschnitten des Stimmkanalmodells errechnet werden oder ergeben sich über die charakteristische Impedanz Z_0 und die Ausbreitungskonstante γ des benutzten Stimmkanalmodells (nach Gleichungen 2.14 und 2.16). Die Schallfluss- und Schalldruckverläufe können mit dem Synthetisator berechnet werden. Sie werden an den Kopplungsstellen der Röhrensegmente abgegriffen. Während der Laufzeit wird praktisch nur die Formantextraktion und die Multiplikation mit der verallgemeinerten Inversen der Matrix **E** durchgeführt. Die Daten des Referenzsprechers (Formantfrequenzen, Energiedichtefunktionen und die Pseudoinverse) werden vorher berechnet und abgespeichert. Allerdings überprüft der Synthetisator die Anpassung der Formantmuster, indem die Formantmuster zu den geschätzten Querschnitten des Testsprechers bestimmt werden. Die Querschnittsstörung $\Delta A/A$ wird bei der Konstruktion der geschätzten Querschnittsfunktion des Testsprechers durch Addition zur logarithmierten Querschnittsfunktion des Referenzsprechers berechnet, denn es gilt

$$\Delta A/A \approx \mathrm{d}A/A = \mathrm{d}\ln A\,. \tag{3.21}$$

3.2.4. Einordnung des Verfahrens in die Literatur

Die vorliegende Arbeit „Vokaltraktmodellbasierte Schätzung von Steuerparametern eines Moduls zur Sprechernormalisierung“ liegt in der Schnittmenge dreier großer Themenkomplexe:

1. Sprecherindividualität,
2. Inverses Problem zur Sprachproduktion,
3. Sprechernormalisierung.

In diesem Abschnitt wird dieses Umfeld der Arbeit dargestellt und so eine Standortbestimmung vorgenommen.

Untersuchungen zum Thema Sprecherindividualität sind für Stimmenkonvertierung [8, 37, 50], Sprecheridentifikation [21, 22, 55] und schließlich auch Sprechernormalisierung [14, 33] von Interesse. Dem entsprechend lang ist die Liste von untersuchten akustischen Quellen der Individualität von Sprache (die Zitate sind Beispiele und ohne Anspruch auf Vollständigkeit zusammengestellt):

1. Quelle
   - Zeit-Frequenz-Muster der Tonhöhe (*pitch contour*) [1, 8]
   - Gestalt des glottalen Schallflusspulses (*glottal wave shape*)[8]
2. Filter
   - die spektrale Einhüllende (*spectral envelope*) und der Abfall des Spektrums (*spectral tilt*) [8, 27, 28, 37]
   - die Werte der Formantfrequenzen [8]
   - das Langzeitspektrum [21]
   - Bandbreiten der Formanten [37].

Das im Rahmen dieser Arbeit entwickelte Verfahren analysiert Formantmuster, die unter die zweite Kategorie fallen – Effekte, die dem Filter, dem Stimmkanal zugeordnet werden.

Zum Thema der Formantfrequenznormierung findet man in der Literatur häufig das Stichwort *Vocal Tract Length Normalization - VTLN*. Dies ist eine Methode der Sprechernormalisierung, die eine geeignete Verzerrung der Frequenzachse durchführt. Im einfachsten Fall handelt es sich um eine lineare Verzerrung der Frequenzachse, was im verlustlosen Fall einer Normalisierung der Länge des Stimmkanals entspricht. VTLN gibt es in vielen Variationen. Uebel und Woodland stellen einige Verfahren experimentell gegenüber [60]. Auch nichtlineare Verzerrungen sind möglich. Diese werden z. B. mit Hilfe eines DP (*Dynamik Programming*) basierten Verfahrens ermittelt [33]. Den VTLN-Verfahren gemeinsam ist, dass es Ansätze sind, denen kein physikalisches Modell des Stimmkanals zu Grunde liegt. Letzteres haben die meisten Methoden der Sprechernormalisierung gemeinsam.

Das Thema dieser Arbeit ist die Schätzung von anatomischen Parametern aus den akustischen Größen. Die umgekehrte Fragestellung, nämlich das Problem der Auswirkungen von Skalierungen der Querschnittsfunktion auf das Formantmuster wurde von Fant bearbeitet. Fant [14] diskutiert nichtgleichmäßige Skalierungen von Formantmustern und führt diese auf unterschiedliche Verhältnisse der Längen von Mundhöhle zu Rachenraum zurück. Fant [15] untersucht mit der Hilfe einer störungstheoretischen Beschreibung von Längenstörungen die Abhängigkeit von Längenverhältnissen und Formantmustern. Högberg [24] passt ein stilisiertes Stimmkanalmodell an gemessene Querschnittsfunktionen eines Mannes und einer Frau an und erhält so Hinweise auf die von Fant vermuteten Skalierungsunterschiede.

Die Umkehrung der von Fant berarbeiteten Fragestellung ist eng mit dem inversen Problem zur Sprachproduktion verknüpft. Die Schätzung von Querschnittsfunktionen

aus dem Sprachsignal ist wie bereits in Abschnitt 2.3 beschrieben ein *schlecht gestelltes* Problem. Die Anstrengungen bisheriger Arbeiten gehen daher in die Richtung, Mehrdeutigkeiten zu vermeiden, zum Beispiel durch Einführung eines artikulatorischen Modells [34] zur Einschränkung der Freiheitsgrade. Die Einführung weiterer Parameter, die Sprecherunterschiede modellieren, wird daher im Allgemeinen nicht vorgesehen. Ein wichtiger Parameter, der zwangsläufig festgelegt werden muss, ist die Gesamtlänge des Vokaltraktes. Diese wird meist einmalig festgelegt und bis auf mit Artikulation verbundenen kleinen Längenvariationen wie Lippenvorstülpung und Position des Kehlkopfes nicht mehr verändert. Ein verbreitetes Verfahren zur formantbasierten Vokaltraktlängenschätzung geht auf Paige und Zue zurück [40]. Für die Länge des Stimmkanals wird folgende Näherung angegeben:

$$L = \frac{c \sum_{n=1}^{M} [F_n/(2n-1)]}{\sum_{n=1}^{M} [4F_n/(2n-1)]^2}. \tag{3.22}$$

Um Eindeutigkeit zu erzwingen, wird als zusätzliches Kriterium die minimale Abweichung der Gestalt des Vokaltraktes vom homogenen Rohr angenommen. Erstaunlicherweise führt diese künstliche Einschränkung zu guten Schätzungen der Längen.

Ein vom Gesamtkonzept zur vorliegenden Arbeit sehr nahes Verfahren wurde parallel und unabhängig von Naito et al. [38] entwickelt. Naito et al. stellten eine Formantkarte zu den Querschnittsfunktionen der Vokale /aː/ und /iː/ zusammen, die auf einem artikulatorischen Sprachsynthetisator beruhte, der einer nichlinearen Längenskalierung unterworfen wurde. Einem Testsprecher wird anhand dieser Karte je ein Skalierungsfaktor des Mundraumes und des Rachenraumes zugeordnet. Diese Skalierungen werden von der nachgeschalteten Spracherkennung zur Einteilung des Sprecherraumes genutzt. In einer späteren Veröffenlichung [39] wurde zu den ermittelten Skalierungsparametern des Testsprechers eine Schar von Frequenzverzerrungen errechnet. Diese wurde aus den Abweichungen der ersten sieben Formanten des skalierten Vokaltraktmodells zum unskalierten Modell errechnet.

Die Methode nach Naito et al. und die in dieser Arbeit vorgestellte Methode gleichen sich in zwei wichtigen Punkten. Beide Verfahren sind vokaltraktmodellgestützt und ermitteln aus Formantmustern nichtlineare Skalierungen der Vokaltraktlänge. Allerdings unterscheiden sich beide Methoden grundlegend in der Berechnung dieser Skalierungen. Während Naito et al. die verschiedenen Formantfrequenzen zu unterschiedlichen Skalierungen abspeichert und die Datenbank dann bei der Schätzung durchsucht, ist das in Kapitel 3 beschriebene Verfahren in der Lage die Skalierungen direkt aus Formantfrequenzabweichungen zu bestimmen. Ferner lässt der in dieser Arbeit verfolgte Ansatz auch Skalierungen der Querschnittsfunktionen zu.

4. Versuche

Ebenso wie das Verfahren von Paige und Zue [40] ist das im letzten Kapitel vorgestellte Verfahren empirischer Art. Die Gültigkeit der Annahme, dass Sprecherspezifika durch Störungen einer mittleren Querschnittsfunktion des betreffenden Vokals erfasst werden, muss daher experimentell belegt werden. Die durchgeführten Versuche können in zwei Teile gegliedert werden. Zunächst werden synthetische Vokale mit bekannter Querschnittsfunktion untersucht. Im Mittelpunkt dieser Untersuchungen steht die Frage, ob die geschätzte Abweichung der Querschnittsfunktion des Referenzsprechers der wirklich vorliegenden Abweichung entspricht.

Der zweite Teil besteht aus der Analyse gesprochener Vokale. Ein Vergleich der angepassten mit der tatsächlichen Querschnittsfunktion des Testsprechers ist hier nicht mehr möglich. Die untersuchte Fragestellung ist, ob im Raum der geschätzten Parameter die verschiedenen Vokale eines Sprechers ähnliche Werte besitzen. Schließlich werden mit Parametersätzen aus Vokalen, die aus fließender Sprache gewonnen wurden, Sprechererkennungsexperimente durchgeführt.

4.1. Die Querschnittsfunktionen

Um festzustellen, ob das Verfahren individuelle Merkmale des Stimmkanals eines Testsprechers modelliert, sind realistische Querschnittsfunktionen des Stimmkanals, idealerweise von mehreren Sprechern, notwendig. In Tabelle 4.1 sind eine Reihe von Arbeiten aufgeführt, die Sätze von Querschnittsdaten von mehreren Sprechern enthalten. Diese aus kernspintomographischen Aufnahmen gewonnenen Daten sind Grundlage der durchgeführten Untersuchungen. Tiede und Yehia arbeiten an der Schätzung von Querschnittsfunktionen aus mediosagittalen Profilen [56]. Die noch unveröffentlichten Daten wurden uns für unsere Untersuchungen freundlicherweise zur Verfügung gestellt. Von Story et al. werden demnächst Daten von weiteren Sprechern veröffentlicht [50]. Die in diesem Kapitel präsentierten Ergebnisse basieren ausnahmslos auf den Daten von Tiede. Die Datenbasis von Tiede enthält Datensätze zu einer im Vergleich relativ großen Anzahl von Sprechern. Die Querschnittsfunktionen der Vokale in Nihongo ähneln zudem den deutschen (mit Ausnahme des /uː/), was für die Experimente an deutschen gehaltenen Vokalen von Bedeutung ist.

Autoren	Personen	Auflösung (in cm)	Äußerungen
Baer et al. [4]	2	0,875	4 englische Vokale
Story et al. [51] und [52]	2	0,396825	10 englische Vokale, 8 bzw. 2 Konsonanten
Tiede et al. (unveröffentlicht)	6/6	variabel (35 Segmente)	5 Vokale in Nihongo (Japanisch), neun englische Vokale

Tabelle 4.1.: Einige Studien zur Ermittlung von Querschnittsfunktionen des menschlichen Stimmkanals durch kernspintomographische Aufnahmen.

4.2. Die Wahl des Referenzsprechers

Der Referenzsprecher soll einen „mittleren“ Sprecher darstellen. Das ist in dem Sinne zu verstehen, dass der Referenzsprecher den jeweiligen Vokal optimal repräsentiert, sodass die Abweichung zu den Testsprechern durch eine kleine Störung der Querschnittsfunktion erfasst wird. Verschiedene Methoden der Berechnung des Referenzsprechers aus gemessenen Querschnittsfunktionen wurden verglichen:

1. Mittelung der linearen Querschnittsfunktionen der jeweiligen Vokale,
2. Mittelung der logarithmierten Querschnittsfunktionen der jeweiligen Vokale,
3. Synthese des Referenzsprechers aus mittleren Koeffizienten der Kosinusreihenentwicklung der logarithmischen Querschnittsfunktionen der jeweiligen Vokale.

Wenn im Folgenden von der Querschnittsfunktion des Referenzsprechers die Rede ist, bezieht sich das auf die nach 3. berechnete Querschnittsfunktion. Die Ordnung der Entwicklung wurde dabei auf 13 Koeffizienten begrenzt. Es wurden so geglättete Querschnittsfunktionen erhalten, die den Charakter des Vokals im Querschnittsverlauf beschreiben und auch (nach Gleichung 2.27) eine mittlere Lage der Formanten gewährleisten. Die logarithmierte Querschnittsfunktion des Vokals /aː/ des Testsprechers ist als durchgezogenen Linie in den Abbildungen 4.4 aufgetragen. Die übrigen Vokale können den entsprechenden Simulationsergebnissen entnommen werden, die im Anhang B zusammengetragen wurden.

4.3. Simulationen zur Längenskalierung

Bevor die eigentlichen Anpassungen der Querschnittsfunktionen vorgenommen werden, wird die für die meisten folgenden Experimente benutzte nichtlineare Längenskalierung näher beschrieben. Untersuchungen von Fant [14] zeigen, dass eine lineare Skalierung der Formanten, die etwa einer linearen Längenskalierung des Stimmkanals entspricht, für die Normalisierung von Formantfrequenzen nicht ausreicht. In den im Folgenden

vorgestellten Experimenten werden die erste und die zweite Hälfte des Vokaltraktes mit unterschiedlichen Faktoren in der Länge skaliert.

In Kapitel 3 wurden die Arbeiten von Strube [54] und Fant [15] angeführt, die Effekte von kleinen Längenstörungen der Querschnittsfunktion auf die relative Verschiebung der Formantfrequenzen $\tilde{\omega}$ beschreiben. Für den Fall eines Röhrenmodells, das sich aus diskreten Segmenten homogenen Querschnittes zusammensetzt, lassen sich beide Näherungen in der Form

$$\tilde{\omega} = \frac{\sum_i k_i H_i}{\sum_i H_i} \tag{4.1}$$

darstellen. k_i beschreibt die Längenskalierung des Segmentes i und wird je nach Herleitung unterschiedlich interpretiert, nämlich

$$k_i = -\Delta l_i / l_i \tag{4.2}$$

nach Strube (in diskrete Form umgeschrieben) gemäß Gleichung (3.6) und

$$k_i = -\frac{\Delta l_i / l}{1 + \Delta l_i / l_i} \tag{4.3}$$

nach Fant gemäß der entsprechenden Gleichung (3.2). Für kleine Störungen der Vokaltraktlänge sind Gleichung (4.2), die nur Effekte 1. Ordnung berücksichtigt, und (4.3) äquivalent. Die individuelle Länge des menschlichen Stimmkanals variiert allerdings sehr stark – Högberg [24] ermittelte bei der Umskalierung einer männlichen Querschnittsfunktion zu einer weiblichen eine Verkürzung um 17 %. In den folgenden Simulationen wird Fants Näherung (Gleichung (4.3)) verwendet, die auch noch für große lineare Längenskalierungen Gültigkeit besitzt, was aus einer einfachen Überlegung folgt: Für eine lineare Skalierung einer ungedämpften inhomogenen Röhre gilt die einfache Beziehung $\omega \propto 1/l$. ω ist die Resonanzfrequenz und l die Länge der Röhre. Multipliziert man die Länge l mit dem Faktor $1 + \Delta l/l$, so erhält man die relative Formantverschiebung

$$\Delta\omega/\omega = \frac{1/\left[l(1 + \Delta l/l)\right] - 1/l}{1/l} = -\frac{\Delta l/l}{1 + \Delta l/l}\,. \tag{4.4}$$

Es wurden eine Reihe von Simulationen durchgeführt, welche die Gültigkeit des gewählten Ansatzes zur Schätzung von nichtlinearen Längenskalierungen abstecken soll. Dabei wird als Beispiel die geglättete Querschittsfunktion des Vokals /u:/ (Abbildung 4.1 a) einer nichtlinearen Längenskalierung unterworfen. Mit unterschiedlicher Längenskalierung von linker und rechter Hälfte der Querschnittsfunktion werden unterschiedliche Längen der Mundhöhle und des Rachenraums realisiert. Aus den mit dem Synthetisator berechneten relativen Formantabweichungen wurden die Längenskalierungen zurückgerechnet und mit den wahren Werten verglichen. Als direktes Ergebnis der Schätzung ergeben sich nach Lösung des Gleichungsystems 3.11 die Parameter k_o und k_p, die in die Skalierung der Mundhöhle $1+\Delta l_o/l_o$ und des Rachenraumes $1+\Delta l_p/l_p$ umgerechnet werden. In den Darstellungen 4.1 b – e wurden die Längenverhältnisse von Mundhöhle zu Rachenraum als x-Achse und die Gesamtlängenskalierung als y-Achse gewählt.

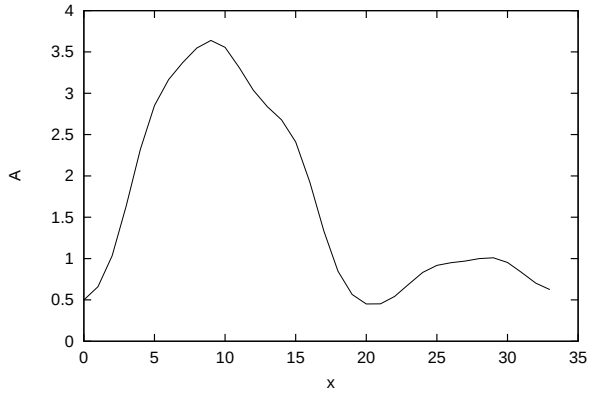
Die Abbildungen 4.1 b, c und d zeigen die Ergebnisse für den ungedämpften Synthetisator. $\Delta l/l$ wurde sowohl nach Gleichung 4.2 (Abbildung 4.1 b), als auch nach Gleichung 4.3 (Abbildung 4.1 c) berechnet. Die Gesamtlängenskalierung wurde zunächst aus dem arithmetischen Mittel der Teilskalierungen berechnet. Es stellt sich heraus, dass bei der Substitution $k = -\Delta l/l$ eine starke Verzerrung in der Längenschätzung auftritt, die bei der Annahme von $k = -\frac{\Delta l/l}{1+\Delta l/l}$ nicht beobachtet wird.

Die Abweichung der Gesamtlängenschätzung, die bei stark unterschiedlicher Skalierung von Rachen und Mundhöhle (Abbildung 4.1 b) beobachtet wird, verschwindet fast vollständig, wenn die Gesamtskalierung nicht als Mittel über die Teilskalierungen des Rachens und der Mundhöhle berechnet wird, sondern nach Gleichung 4.3 direkt aus $\bar{k} = (k_o + k_p)/2$ bestimmt wird. Die Gesamtlängenskalierung wird nach Gleichung 4.3 aus $\bar{k}$ ermittelt. Das Ergebnis ist in Abbildung 4.1 c dargestellt.

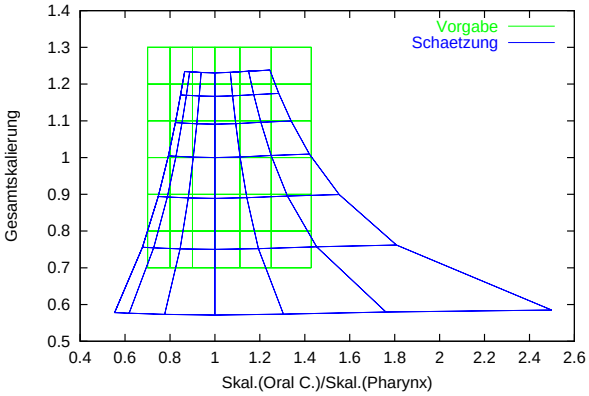
Bei der Schätzung von Skalierungsfaktoren eines gedämpften Systems liegen der Energiedichte H und der Lagrangedichte L die Schalldruck- und Schallflussverhältnisse des gedämpften Synthetisators zugrunde. Die Induktivitätsdichte und Kapazitätsdichte der einzelnen Segmente wurden aus der Kettenmatrix des jeweiligen Segmentes errechnet. Die Längenskalierung wurde gegenüber der vorherigen Simulation beibehalten.

Die Übereinstimmung mit den vorgegebenen Skalierungen ist gegenüber dem dämpfungsfreien Fall zwar schlechter geworden, bietet aber immer noch eine akzeptable Approximation. Die leichte Scherung des Gitters wird kompensiert, indem die Strahlungsimpedanz an den Lippen als Induktivität mit berücksichtigt wird. Bei der Berechnung der Summe der Energiedichtefunktion im Nenner von 4.1 ergibt sich ein Korrekturterm von $H_Z = |U|^2 L_Z$. Die Induktitivität wurde nach einer Näherung von Fant [15] berechnet, die einer Verlängerung der Röhre entspricht: $L_Z = \varrho_0/A \cdot 0.8\sqrt{A/\pi}$. Das Ergebnis ist in Abbildung 4.1 f) dargestellt.

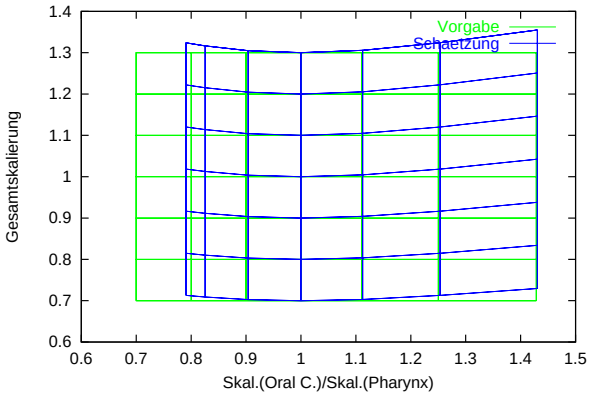
Das Schätzverfahren deckt einen weiten Bereich möglicher Längenverhältnisse des Stimmkanals ab, der etwa Högbergs [24] ermittelte Umskalierung der Längen von Mundhöhle und Rachen eines Mannes zu den entsprechenden Längen einer Frau einschließt: Eine Verkürzung der Gesamtlänge um 17% und das Verhältnis der Längen von Mundhöhle zum Rachenraum beträgt beim Mann 1,06, dagegen bei der Frau 1,24.



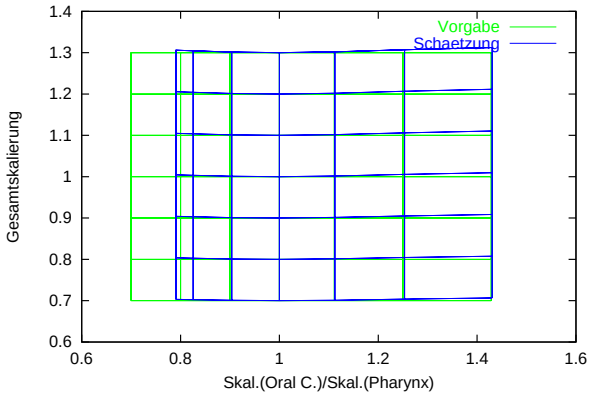
(a) Vokal /u:/, Querschnittsfunktion



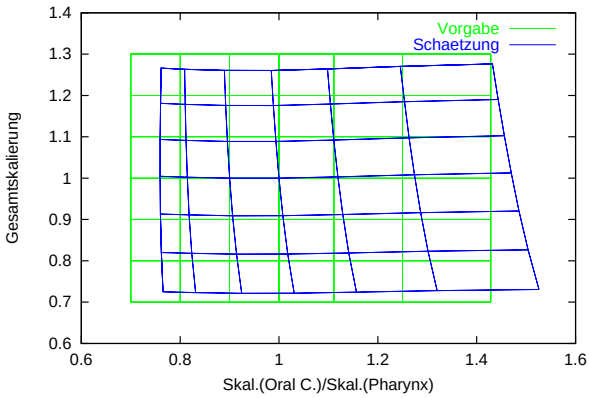
(b) Keine Dämpfung, Längenskalierung nur 1. Ordnung



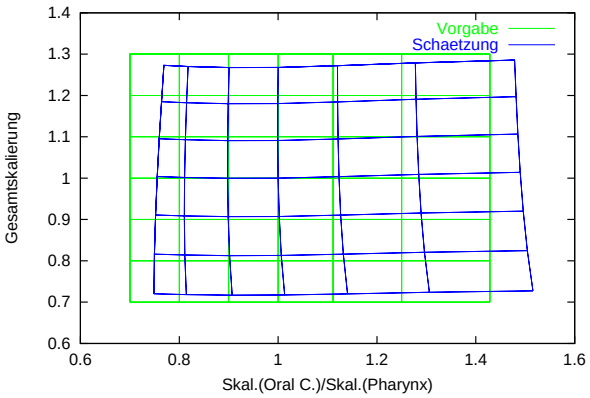
(c) Keine Dämpfung, Längenschätzung nach Fant



(d) Keine Dämpfung, Längenschätzung nach Fant, Mittelung über k



(e) Dämpfung, keine Korrektur der Abstrahlungsimpedanz



(f) Dämpfung, Korrektur der Abstrahlungsimpedanz

Abbildung 4.1.: Die Formantmuster des Vokals /u:/ wurden für verschiedene Skalierungen der Querschnittsfunktion berechnet und die Skalierungsfaktoren zurückgeschätzt.

4.4. Längen- und Querschnittsskalierung

In diesem Abschnitt werden Simulationsexperimente vorgestellt, in denen die Querschnittsfunktion des Testsprechers aus der Querschnittsfunktion des Referenzsprechers und der relativen Formantverschiebung geschätzt wird. Die Abweichungen des Formantmusters werden als Resultat kleiner Störungen der Querschnittsfunktion aufgefasst und das in Kapitel 3 entwickelte Verfahren zur Schätzung angewandt. Die Störungsbasen sehen eine simultane Skalierung von Querschnitten und Längen vor. Die einzelnen Simulationsexperimente unterscheiden sich nur in der Wahl der Störungsbasis. Eine Ausnahme ist das erste Experiment (Abschnitt 4.4.1), das die Längenschätzung nach Paige und Zue [40] als Sonderfall des neu entwickelten Verfahrens behandelt und den Referenzsprecher mit homogener Querschnittsfunktion ansetzt. Die in diesem Abschnitt vorgestellten logarithmischen Querschnittsfunktionen mit N diskreten Segmenten sind zur besseren Vergleichbarkeit normiert, sodass gilt

$$\sum_{i=1}^{N} \log(A_i) = 0\,. \tag{4.5}$$

4.4.1. Die Längenschätzung nach Paige und Zue als Sonderfall des Verfahrens

Eine eindeutige Längenschätzung des Stimmkanals nach Paige und Zue wird erreicht, indem die Abweichung vom homogenen Querschnitt minimal angesetzt wird. Mit der folgenden Wahl des Referenzsprechers und der Störungsbasis wird dieses Kriterium erfüllt: Die Querschnittsfunktion des „Referenzsprechers“ wird für alle Vokale homogen gewählt. Die Störungsbasen entsprechen den ungeraden indizierten Termen der Kosinusreihe (Gleichung 3.20, Abbildung 4.2). Die vorrangige Längenskalierung wird durch die

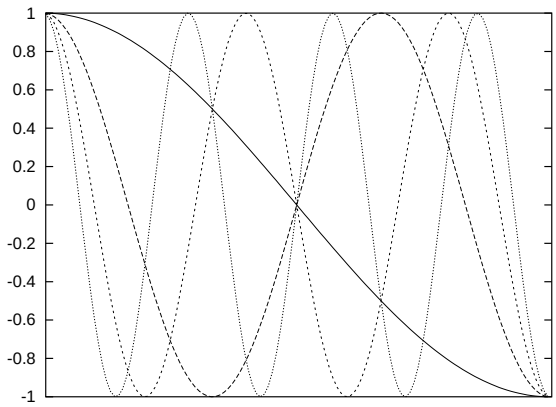


Abbildung 4.2.: Die ungerade indizierten Terme der Kosinusreihe (Gleichung 3.20).

schwächere Wichtung der Querschnittsstörungsbasis gegenüber der Längenstörungsbasis bewirkt. Die Amplitude der Kosinusfunktionen wird auf den Wert G_A festgelegt und der Längenstörungsvektor enthält die Einträge $-G_L/(1+G_L)$. Die Wichtungsparameter

sowie Wahl der Störungsbasen des Referenzsprechers und der freie Parameter der SVD (Gleichung 3.19) sind in der folgenden Tabelle zusammengestellt:

Referenz	homogen	
Störungsbasis		
Querschnitt	Kosinusbasis	G_A:0,3
Länge	linear	G_L:0,5
SVD-Einstellung	ϵ_{SVD}: 0,000000001	

In Tabelle 4.4.1 sind Werte angegeben, die die mittlere Formant- und Längenanpassung für die verschiedenen Testsprecher beschreiben. Die Größe $\hat{\omega}_{\mathrm{n/t}}$ fasst die relative Abweichung der Formanten des Normsprechers ω_{n} vom Testsprecher ω_{t} zusammen:

$$\hat{\omega}_{\mathrm{n/t}} = \sqrt{\sum_{i=1}^{N_{\mathrm{Form}}} \left(\frac{\omega_{\mathrm{n}i} - \omega_{\mathrm{t}i}}{\omega_{\mathrm{t}i}} \right)^2}. \tag{4.6}$$

Entsprechend ist $\hat{\omega}_{\mathrm{e/t}}$ die Zusammenfassung der relativen Formantabweichungen des geschätzten Testsprechers ω_{e} vom Testsprecher ω_{t}. Der über die Vokale eines Sprechers gemittelte Wert ist mit $\overline{\hat{\omega}_{\mathrm{n/t}}}$ bzw. $\overline{\hat{\omega}_{\mathrm{e/t}}}$ bezeichnet.

Tabelle 4.4.1 enthält auch die über die Vokale eines Sprechers gemittelten relativen Fehler der Längenschätzung. $\tilde{L}_{\mathrm{n/t}}$ ist die relative Abweichung der Länge von Referenzsprecher zum Testsprecher, $\tilde{L}_{\mathrm{e/t}}$ ist die relative Abweichung der geschätzten von der tatsächlichen Länge des Stimmkanals des Testsprechers und $\tilde{L}_{\mathrm{pz/t}}$ ist der relative Fehler der direkt nach der Formel von Paige und Zue [40] berechneten Längenschätzung

$$L = \frac{c \sum_{n=1}^{M} [F_n/(2n-1)]}{\sum_{n=1}^{M} [4F_n/(2n-1)]^2}. \tag{4.7}$$

Die Werte der Vokale im einzelnen können der Tabelle im Anhang B.1 entnommen werden.

In Abbildung 4.3 sind die logarithmierten Querschnittsfunktionen des Vokals /a:/ des Referenzsprechers und aller Testsprecher aus der Nihongo-Datenbasis von Tiede und die Schätzungen der logarithmierten Querschnittsfunktionen der Testsprecher aufgetragen.

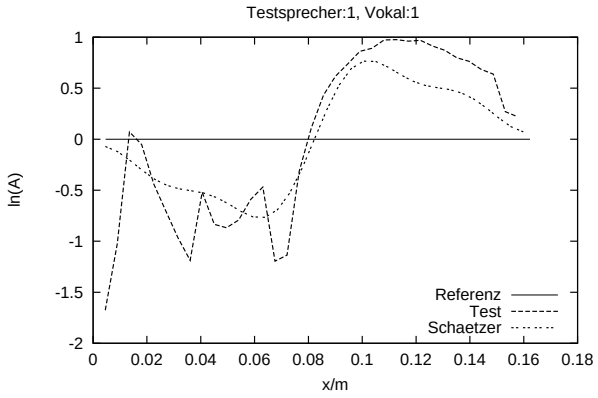
Die Schätzung approximiert den groben Verlauf der logarithmierten Querschnittsfunktion des Testsprechers. Auffällig ist der Abfall auf Null an der Glottis und Lippenende. Dies ist die Folge davon, dass die gleichmäßige Verschiebung aller Formanten zu höheren oder tieferen Frequenzen durch die Längenskalierung vorgenommen wird. Die Querschnittsskalierung mit der Kosinusbasis verschiebt die Formanten nur gegeneinander. Alle Vektoren der Kosinusbasis (Abbildung 4.2) haben an der Glottis und an den Lippen identische Werte, sodass sich bei paarweiser Verschiebung der Formanten die Querschnittsänderungen an den Enden der Querschnittsfunktion aufheben.

Das Verfahren ergibt eine gute Längenschätzung des Testsprechers. $\tilde{L}_{\mathrm{e/t}}$ und $\tilde{L}_{\mathrm{pz/t}}$ sollten nach dem Ansatz dieses Versuches übereinstimmen und tatsächlich zeigen sie

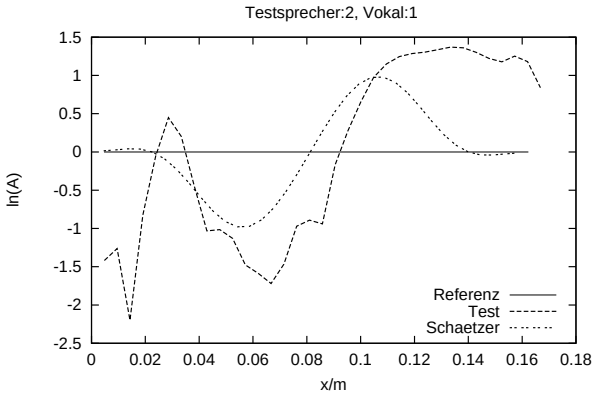
Sprecher	$\overline{\hat{\omega}_{\mathrm{n/t}}}$	$\overline{\hat{\omega}_{\mathrm{e/t}}}$	$\overline{\tilde{L}_{\mathrm{n/t}}}$	$\overline{\tilde{L}_{\mathrm{e/t}}}$	$\overline{\tilde{L}_{\mathrm{pz/t}}}$
1	0,594	0,089	0,022	0,068	0,080
2	0,704	0,144	0,042	0,062	0,075
3	0,591	0,092	0,105	0,066	0,087
4	0,782	0,133	0,087	0,095	0,113
5	0,847	0,154	0,095	0,055	0,064
6	0,707	0,134	0,134	0,071	0,084

Tabelle 4.2.: Die Abweichungen der Formantmuster und der Längen der Querschnittsfunktion des Referenzsprechers und der geschätzten Querschnittsfunktion des Testsprechers zur Querschnittsfunktion des Testsprechers, gemittelt über die fünf untersuchten Vokale. Die Schätzung wird mit einem homogenen Querschnittsverlauf als Referenzsprecher und der Kosinusbasis durchgeführt.

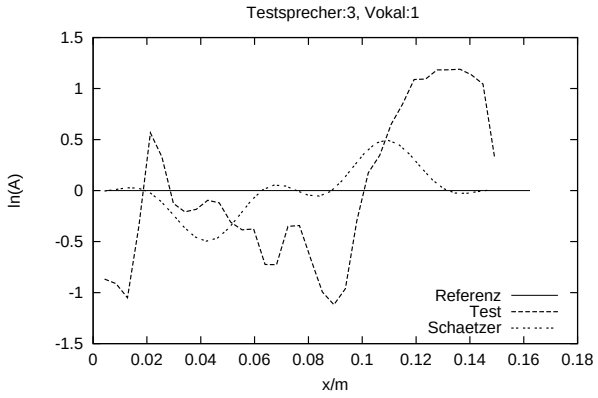
denselben Trend. Allerdings ist die Schätzung aus dem Referenzsprecher nach dem neuen Verfahren etwas genauer als die direkte Schätzung nach Gleichung 4.7. Dies kann auf die Berücksichtigung der Dämpfung des Synthetisators zurückgeführt werden.



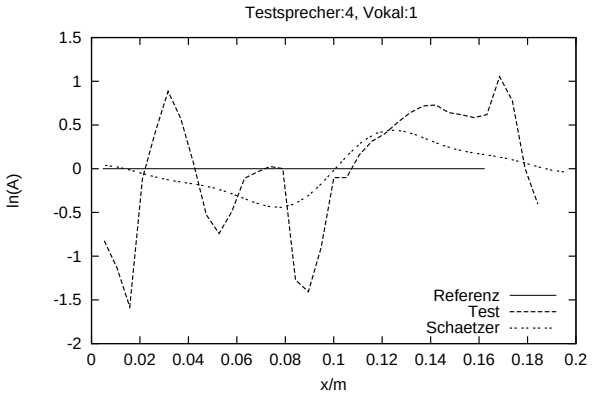
(a) Testsprecher 1



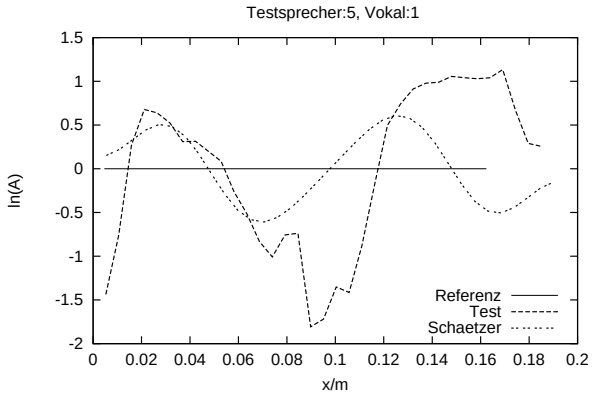
(b) Testsprecher 2



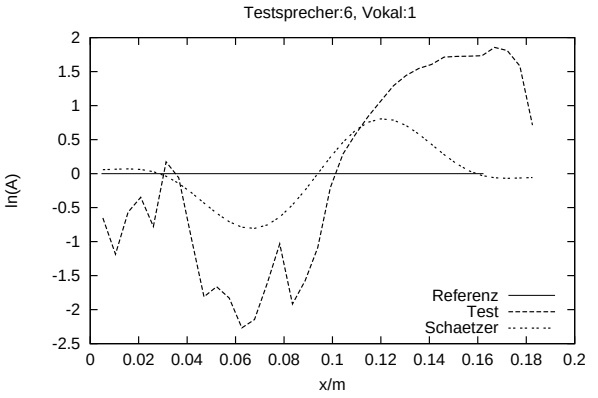
(c) Testsprecher 3



(d) Testsprecher 4



(e) Testsprecher 5



(f) Testsprecher 6

Abbildung 4.3.: Schätzung der Querschnittsfunktion (/aː/) des Testsprechers aus der homogenen Querschnittsfunktion. Die Abweichungen werden durch lineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis nachgebildet.

4.4.2. Die Querschnitts- und Längenschätzung mit nichtlinearer Längenskalierung und der Kosinusbasis

In diesem Abschnitt wird das Verfahren aus Kapitel 3 zur simultanen Schätzung von nichtlinearen Längenskalierungen und Querschnittsskalierungen angewandt.

Referenz	mittlerer jeweiliger Vokal	
Störungsbasis		
Querschnitt	Kosinusbasis	G_A:0,3
Länge	nichtlinear	G_L:0,1
SVD-Einstellung	ϵ_{SVD}:0,0025	

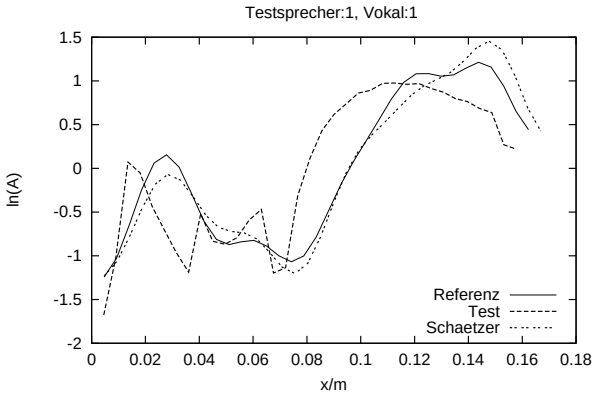
Die Abweichungen der Formantmuster und Längenschätzungen sind in Tabelle 4.3 für alle Vokale zusammengefasst. In Abbildung 4.4 sind die logarithmierten Querschnittsfunk-

Sprecher	$\overline{\hat{\omega}_{\mathrm{n/t}}}$	$\overline{\hat{\omega}_{\mathrm{e/t}}}$	$\overline{\tilde{L}_{\mathrm{n/t}}}$	$\overline{\tilde{L}_{\mathrm{e/t}}}$	$\overline{\tilde{L}_{\mathrm{pz/t}}}$
1	0,200	0,099	0,022	0,022	0,080
2	0,212	0,037	0,042	0,030	0,075
3	0,201	0,058	0,105	0,045	0,087
4	0,529	0,043	0,087	0,079	0,113
5	0,433	0,104	0,095	0,016	0,064
6	0,414	0,056	0,134	0,015	0,084

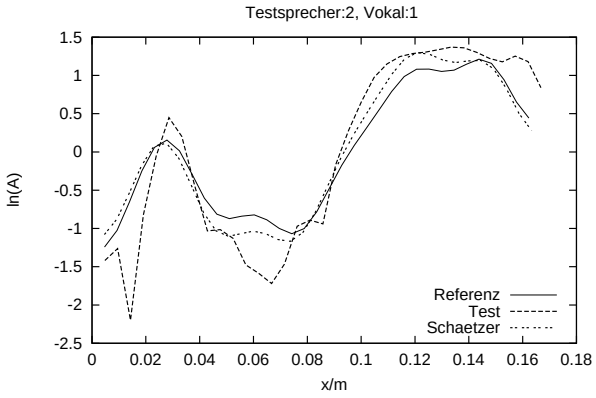
Tabelle 4.3.: Die Abweichungen der Formantmuster und der Längen der Querschnittsfunktion des Referenzsprechers und der geschätzten Querschnittsfunktion des Testsprechers zur Querschnittsfunktion des Testsprechers, gemittelt über die fünf untersuchten Vokale. Die Schätzung wird mit einem mittleren Querschnittsverlauf des jeweiligen Vokals als Referenzsprecher und der Kosinusbasis durchgeführt.

tionen des Vokals /a:/ des Referenzsprechers und aller Testsprecher sowie die Schätzungen der logarithmierten Querschnittsfunktionen der Testsprecher aufgetragen. Der grobe Verlauf des Vokals ist durch die Querschnittsfunktion des Referenzsprechers bereits vorgegeben. Ziel der Anpassung ist die Modellierung von Details der Querschnittsfunktion. Die Auswahl an Testsprechern ist zwar sehr begrenzt, jedoch kann man aus Abbildung 4.4 Erkenntnisse über verschiedene Fälle gewinnen. Allgemein gilt, dass, über alle Vokale gemittelt, eine deutliche Verbesserung der Anpassung der Gesamtlänge gegenüber dem letzten Experiment gelungen ist (Tabelle 4.3). Die geschätzten logarithmierten Querschnittsfunktionen der Testsprecher drei bis sechs haben gegenüber denen des Referenzsprechers eine Annäherung im Detail an die Querschnittsverläufe des Testsprechers erfahren (zu vergleichen sind die beiden gestrichelt dargestellten logarithmischen Querschnittsverläufe). Man erkennt eine bessere qualitative Übereinstimmung des Rachenraumvolumens und der Mundhöhle sowie eine gute Übereinstimmung der Einschnürung

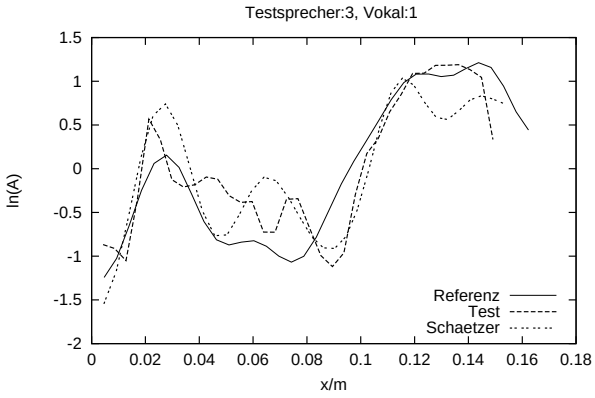
der Querschnittsfunktion. Testsprecher zwei stimmt bereits gut mit dem Referenzsprecher überein, sodass keine wesentliche Verbesserung beobachtet werden kann. Testsprecher eins führt das Verfahren an seine Grenzen. Die Annahme, dass der Testsprecher durch eine kleine Abweichung vom Referenzsprecher angenähert werden kann, ist hier nicht gerechtfertigt. Statt die charakteristische, stark vom Referenzsprecher abweichende Beschaffenheit der Mundhöhle anzunähern, werden die Frequenzmuster durch eine kleine Korrektur der Querschnittsfunktion angenähert. Die Anpassung der Formantmuster von Testsprecher vier kann sowohl durch eine Längenskalierung als auch durch eine Querschnittsskalierung erreicht werden. Das Ergebnis der Anpassung ist sehr von der Abstimmung der Störungsvektoren des Querschnittes und der Länge abhängig. Die Abstimmung wurde so vorgenommen, dass Testsprecher drei (mit dem kürzesten Stimmkanal aller Testsprecher) eine akzeptable Längenanpassung aufweist. Eine vorrangige Längenskalierung wie im letzten Experiment hatte eine nahezu perfekte Anpassung in Querschnitt und Länge des Testsprechers drei zur Folge, führte aber zu einer erheblichen Längenüberschätzung von Testsprecher vier.



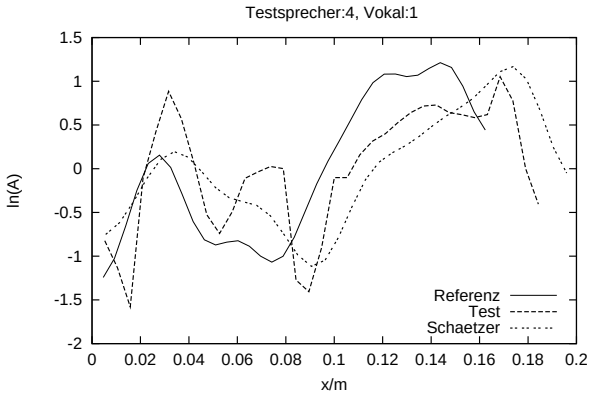
(a) Testsprecher 1



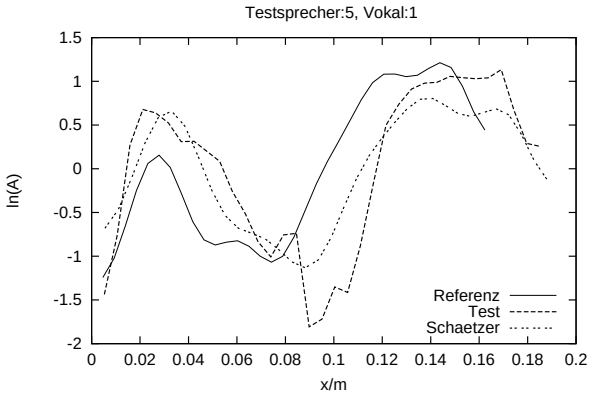
(b) Testsprecher 2



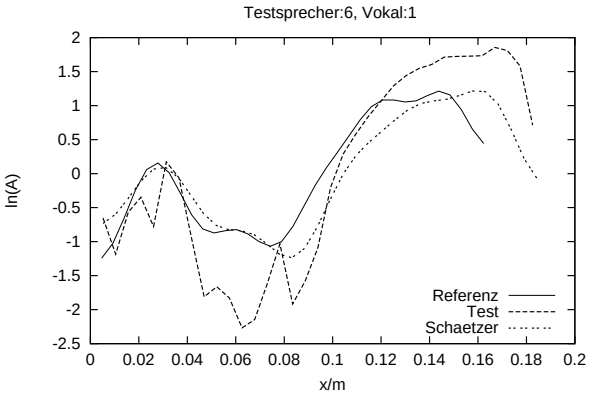
(c) Testsprecher 3



(d) Testsprecher 4



(e) Testsprecher 5



(f) Testsprecher 6

Abbildung 4.4.: Schätzung der Querschnittsfunktion (/aː/) des Testsprechers aus der Querschnittsfunktion des Referenzsprechers. Die Abweichungen werden durch nichtlineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis nachgebildet.

4.4.3. Entwicklung einer Störungsbasis durch eine Hauptkomponentenanalyse der Querschnittsdaten

In den vorherigen Versuchen wurde die Umskalierung der Querschnittsfunktion des Referenzsprechers auf die geschätzte Querschnittsfunktion des Testsprechers mit Hilfe von Kosinusgliedern, die zu ungeradzahligen Koeffizienten der Kosinusreihe gehören, durchgeführt. Dieses Vorgehen hat zwei Nachteile: Erstens ist eine vollständige Anpassung der Querschnittsfunktionen prinzipiell nicht möglich, da die Kosinusglieder zu geradzahligen Koeffizienten einfach fehlen. Zweitens enthalten diese künstlichen Skalierungsfunktionen keinerlei Beschränkungen auf realistische Querschnittsskalierungen. Es wird nun untersucht, ob die auf Messdaten beruhenden Skalierungsfunktionen die Mehrdeutigkeiten reduzieren und besser geeignet sind, die Unterschiede zwischen den Sprechern nachzubilden.

Die ideale Störungsbasis enthält ausschließlich Sprecherspezifika. Um die Daten von vokalspezifischen Einflüssen zu befreien, wurde von jeder logarithmierten Querschnittsfunktion die über alle Sprecher gemittelte logarithmierte Querschnittsfunktion des entsprechenden Vokals abgezogen:

$$x^v_{ijk} = x_{ijk} - \frac{1}{N_{\mathrm{sp}}} \sum_{n=1}^{N_{\mathrm{sp}}} x_{ijn}. \tag{4.8}$$

Dabei sei x_{ijk} der logarithmierte Querschnitt des Vokals j des Sprechers k im Segment i und N_{sp} die Anzahl der Sprecher. Die auf diese Art mittelwertbefreite Datenbasis wird mit X^v bezeichnet. Die Querschnittsfunktionen wurden der Datenbasis von Tiede entnommen.[1] Die auf diese Art mittelwertbefreiten Daten X^v wurden einer Hauptkomponentenanalyse unterzogen. Die Analysen wurden auf den logarithmierten Querschnittsfunktionen durchgeführt, da sich die ergebenden Hauptkomponenten direkt als Störungsbasis $\mathbf{v}_j$ interpretieren lassen. Dies zeigt die folgende Überlegung: Die Störungsbasis $\mathbf{v}_j$ wird nach Gleichung 3.10 zu relativen Störungen $\Delta A/A$ der Querschnittsfunktion zusammengesetzt. Die auf die oben beschriebene Art gewonnenen Hauptkomponenten stellen die Differenz $\Delta \ln A$ der logarithmischen Querschnittsfunktion zum mittleren Vokal als eine Linearkombination dar. Da für kleine ΔA

$$\Delta A/A \approx \mathrm{d}A/A = \mathrm{d} \ln A \tag{4.9}$$

gilt, können die Hauptkomponenten als Störungsbasis $\mathbf{v}_j$ betrachtet werden.

Die Anteile der Varianzen der einzelnen Hauptkomponenten von X^v sind in Tabelle 4.4 aufgeführt. Die ersten vier Hauptkomponenten, die zusammen 84,4 % der Varianz erklären, sind in Abbildung 4.5 dargestellt. Sie werden im nachfolgenden Experiment als Störungsbasis eingesetzt.

[1] Die Vokale von Sprecher eins wurden aus der Analyse ausgenommen, da er die Vokale gegenüber den anderen Sprechern auf *sehr* unterschiedliche Weise bildet und daher die Voraussetzung der Methode, dass die Querschnittsfunktion des Testsprechers als kleine Störung der Querschnittsfunktion des Referenzsprechers betrachtet werden kann, offensichtlich nicht erfüllt ist.

HK.	Varianz
1	0,366
2	0,654
3	0,798
4	0,844
5	0,885

Tabelle 4.4.: Anteile der Varianz der Hauptkomponentenanalyse.

Referenz	mittlerer jeweiliger Vokal	
Störungsbasis		
Querschnitt	Hauptkomponentenbasis	G_A:0,3
Länge	nichtlinear	G_L:0,05
SVD-Einstellung	ϵ_{SVD}:0,015	

Sprecher	$\overline{\hat{\omega}_{\mathrm{n/t}}}$	$\overline{\hat{\omega}_{\mathrm{e/t}}}$	$\overline{\tilde{L}_{\mathrm{n/t}}}$	$\overline{\tilde{L}_{\mathrm{e/t}}}$	$\overline{\tilde{L}_{\mathrm{pz/t}}}$
1	0,200	0,091	0,022	0,037	0,080
2	0,212	0,057	0,042	0,021	0,075
3	0,201	0,071	0,105	0,041	0,087
4	0,529	0,135	0,087	0,077	0,113
5	0,433	0,111	0,095	0,017	0,064
6	0,414	0,109	0,134	0,041	0,084

Tabelle 4.5.: Die Abweichungen der Formantmuster und der Längen der Querschnittsfunktion des Referenzsprechers und der geschätzten Querschnittsfunktion des Testsprechers zur Querschnittsfunktion des Testsprechers, gemittelt über die fünf untersuchten Vokale. Die Schätzung wurde mit einer aus der Hauptkomponentenanalyse gewonnenen Störungsbasis durchgeführt.

Tabelle 4.5 enthält Informationen über Formant- und Längenanpassung der geschätzten Querschnittsfunktion an den Testsprecher. Die Anpassung der Querschnittsfunktionen ist gegenüber der letzten Versuchsreihe mit Kosinusfunktionen nicht verbessert worden. Die mittleren Längenschätzungen und Formantanpassungen (Tabelle 4.5) sind sogar schlechter als in der vorherigen Versuchsreihe. Eine mögliche Ursache ist, dass die Störungsbasis die Querschnittsfunktionen zwar gut anpasst, akustisch aber die Formanten nicht wirkungsvoll beeinflusst. Die erste Hauptkomponente beispielsweise hat einen ausgeprägten Gleichanteil und die dritte Hauptkomponente zeigt eine Symmetrie bezüglich der Traktmitte. Diese Anteile sind zumindest auf eine homogene Röhre angewandt in erster Ordnung wirkungslos.

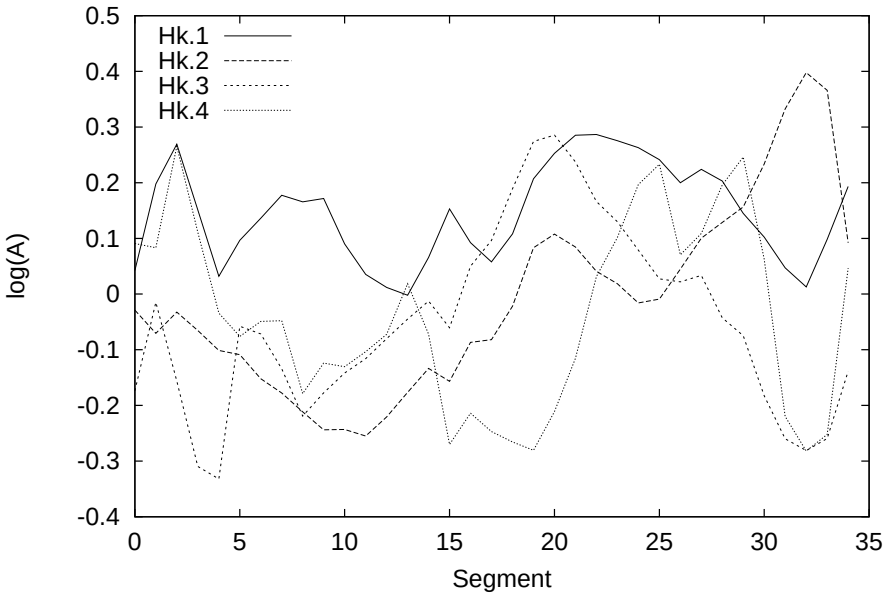
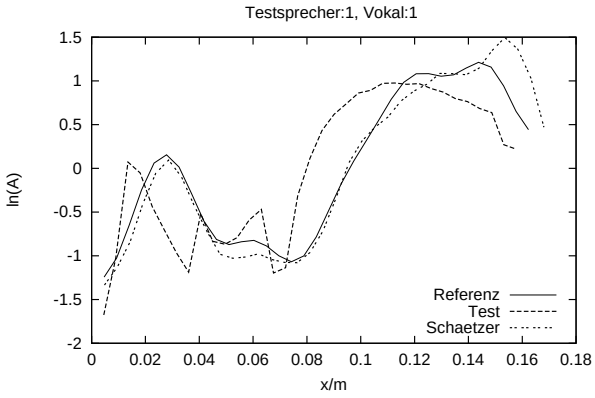
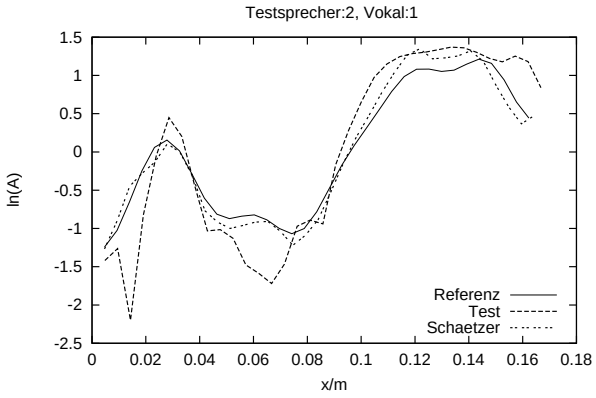


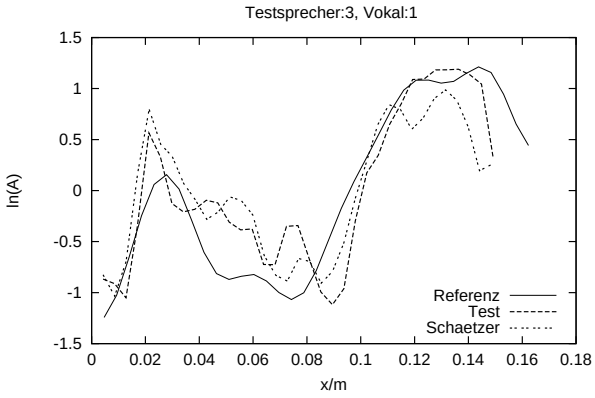
Abbildung 4.5.: Hauptkomponenten aus den zuvor nach Gleichung 4.8 mittelwertbefreiten logarithmierten Querschnittsfunktionen.



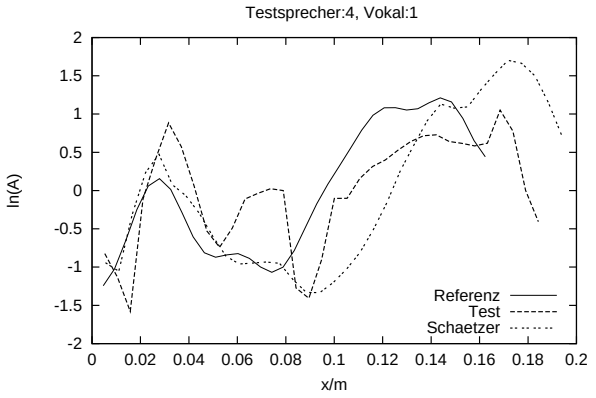
(a) Testsprecher 1



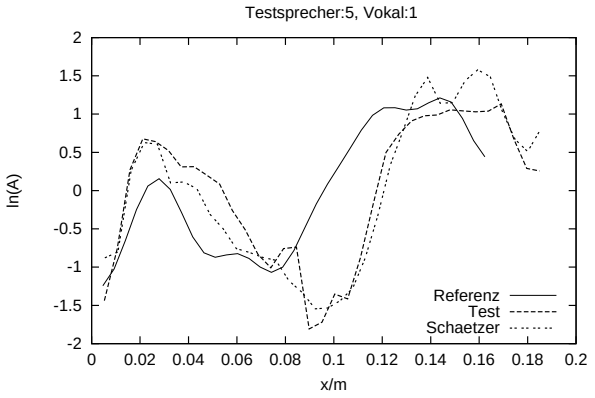
(b) Testsprecher 2



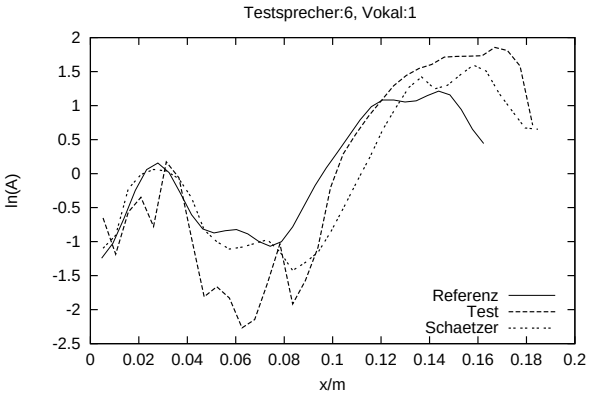
(c) Testsprecher 3



(d) Testsprecher 4



(e) Testsprecher 5



(f) Testsprecher 6

Abbildung 4.6.: Anpassung des Referenzsprechers an den Testsprecher durch unterschiedliche Längenskalierung beider Vokaltrakt-Hälften. Die Querschnittsskalierung wird mit einer aus der Hauptkomponentenanalyse der Querschnittsfunktionen gewonnenen Basis durchgeführt.

4.4.4. Entwicklung einer Störungsbasis durch Analyse der Lagrangedichtefunktionen

Das Konzept des letzten Abschnitts war, die Störungsfunktionen an die Querschnittsfunktionen anzupassen. In diesem Abschnitt hingegen wird das entgegengesetzte Konzept getestet: Die Störungsbasen werden so gewählt, dass sie das Formantmuster kontrolliert und eindeutig beeinflussen, und zwar so, dass jeder Störungsvektor so angepasst wird, dass er in erster Näherung genau einen Formanten beeinflusst. Das ergibt für jeden Vokal des Referenzsprechers eine eigene Störungsfunktion. Heinz [23] berechnete solche Störungsfunktionen aus der Websterschen Horngleichung. In diesem Absatz wird ein pragmatischer Ansatz verfolgt. Die bereits berechneten Lagrangedichtefunktionen des Referenzsprechers werden genutzt, um die Vektoren der angegebenen Art zu erhalten. Gegeben seien die Lagrangedichtefunktionen L für M Formantfrequenzen. Die diskreten Störungsvektoren $\mathbf{v}_j$ der Länge N müssen, um obigen Forderungen zu genügen, nach Gleichung 3.1 folgende Bedingung erfüllen:

$$\delta_{mj} = \sum_i v_{j_i} L_{mi}. \tag{4.10}$$

Für $m = 1, \ldots, M$ und $j = 1, \ldots, M$ ergibt sich ein in aller Regel unterbestimmtes (da $M < N$) lineares Gleichungssystem

$$\mathbf{e}_j = \begin{pmatrix} L_{11} & L_{12} & \cdots & L_{1N} \\ L_{12} & L_{22} & \cdots & L_{2N} \\ \vdots & \vdots & & \vdots \\ L_{1M} & L_{2M} & & L_{MN} \end{pmatrix} \mathbf{v}_j \,. \tag{4.11}$$

Dieses Gleichungssystem wurde mit Hilfe der Pseudoinversen gelöst, sodass als Störungsbasis diejenigen Vektoren $\mathbf{v}_j$ aus der Lösungsmenge von Gleichung 4.10 ausgewählt wurden, die die kleinste Norm besitzen. Die Störungsbasis wurde anschließend normiert. Die Störungsbasen des Vokals /a/ des Referenzsprechers sind in Abbildung 4.7 aufgetragen.

Referenz	mittlerer jeweiliger Vokal	
Störungsbasis		
Querschnitt	Kosinusbasis	G_A:0,3
Länge	nichtlinear	G_L:0,1
SVD-Einstellung	ϵ_{SVD}:0,0025	

Die errechneten Störungsbasen ergeben eine gute Annäherung der Formantmuster und Schätzungen der Längen des Testsprechers (Abbildung 4.6). Die angenäherten Querschnittsfunktionen des Testsprechers sind in Abbildung 4.7 aufgetragen. Bei dieser Störungsbasis handelt es sich sicherlich um ein brauchbares Instrument zur Anpassung von Querschnittsfunktionen und Schätzungen von Längen des Stimmkanals. Insbesondere fällt die gute Approximation des Sprechers vier auf. Insgesamt lässt sich aber keine bedeutende Veränderung gegenüber der Kosinusbasis erkennen.

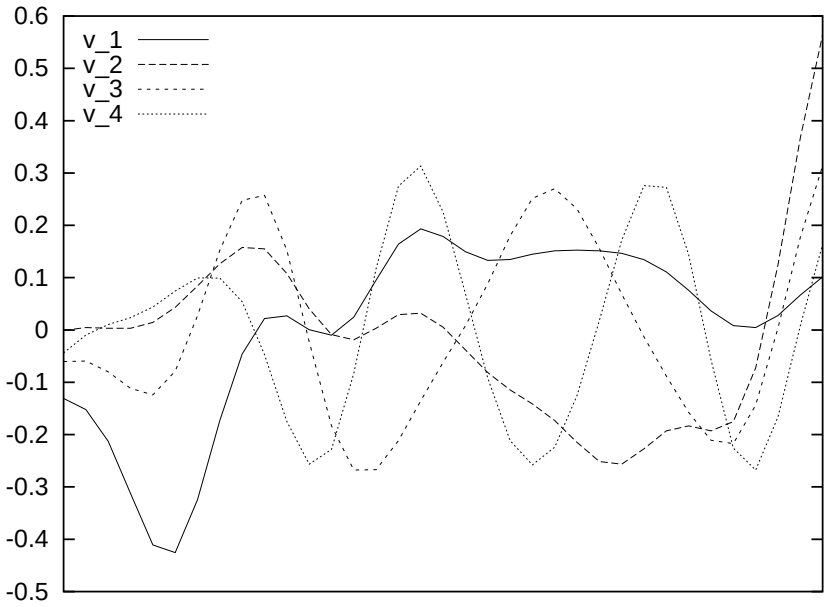


Abbildung 4.7.: Störungsbasis aus der Analyse der Lagrangedichtefunktionen des Stimmkanalmodells des Vokals /a/ des Referenzsprechers (Gleichung 4.10).

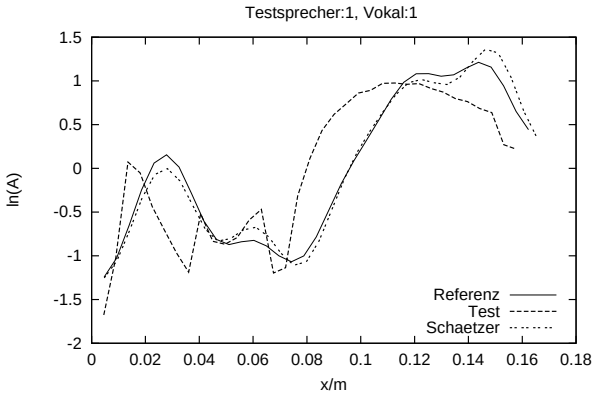
4.5. Beurteilung der Simulationen

Die Simulationen haben gezeigt, dass mit dem vorgestellten Verfahren gute Längenschätzungen gemacht werden können, die im Mittel den Schätzungen von Paige und Zue überlegen sind. Im Allgemeinen liefert das Verfahren auch eine brauchbare Approximation des Testsprechers. Die verschiedenen Störungsbasen liefern insgesamt sehr ähnliche Ergebnisse. Am erfolgreichsten ist aber die Kosinusbasis mit nichtlinearer Längenskalierung, da sie in ihrer Konzeption einfach ist, keine zusätzlichen Informationen über den Testsprecher benötigt und darüber hinaus für alle Vokale geeignet zu sein scheint. Die Simulationen haben aber auch die Grenzen des Verfahrens aufgezeigt: Bei zu unterschiedlichen Querschnittsfunktionen von Referenz- und Testsprecher wie bei Sprecher eins versagt das Verfahren. Außerdem hat das Verfahren gezeigt, dass es zu Problemen kommen kann, wenn die zu schätzenden Querschnittsfunktionen in der Länge zu weit auseinander liegen. Testsprecher drei und Testsprecher vier zeigten sich relativ unvereinbar. Eine Verbesserung der Approximation des einen Sprechers führte zu einer Verschlechterung der Schätzung der Querschnittsfunktion des anderen.

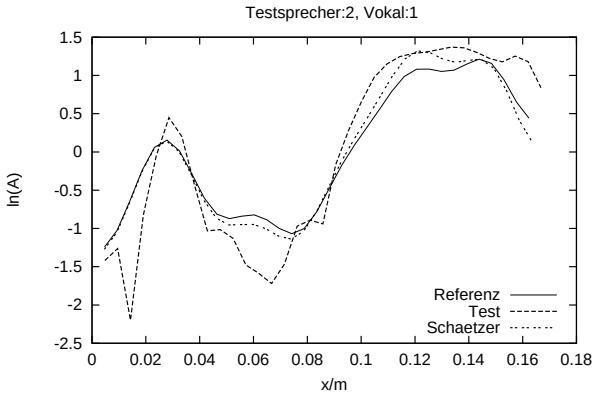
Für die Experimente an natürlicher Sprache wird auf die Kosinusstörungsbasis mit nichtlinearer Längenskalierung zurückgegriffen.

Sprecher	$\overline{\hat{\omega}_{\mathrm{n/t}}}$	$\overline{\hat{\omega}_{\mathrm{e/t}}}$	$\overline{\tilde{L}_{\mathrm{n/t}}}$	$\overline{\tilde{L}_{\mathrm{e/t}}}$	$\overline{\tilde{L}_{\mathrm{pz/t}}}$
1	0,200	0,062	0,022	0,022	0,080
2	0,212	0,034	0,042	0,017	0,075
3	0,201	0,027	0,105	0,057	0,087
4	0,529	0,054	0,087	0,038	0,113
5	0,433	0,100	0,095	0,018	0,064
6	0,414	0,052	0,134	0,048	0,084

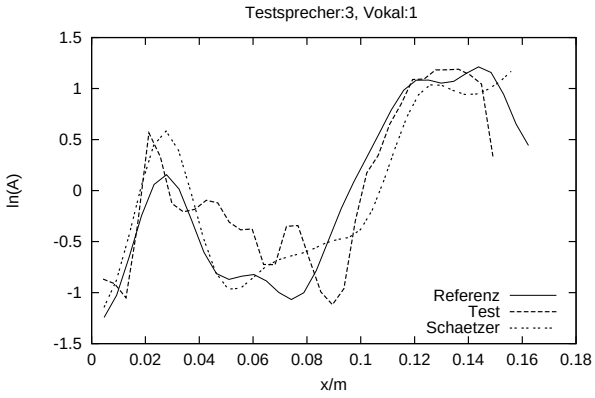
Tabelle 4.6.: Die Abweichungen der Formantmuster und der Längen der Querschnittsfunktion des Referenzsprechers und der geschätzten Querschnittsfunktion des Testsprechers zur Querschnittsfunktion des Testsprechers, gemittelt über die fünf untersuchten Vokale. Die Schätzung wird mit einem mittleren Querschnittsverlauf des jeweiligen Vokals als Referenzsprecher und der aus den Lagrangedichten berechneten Störungsbasen durchgeführt.



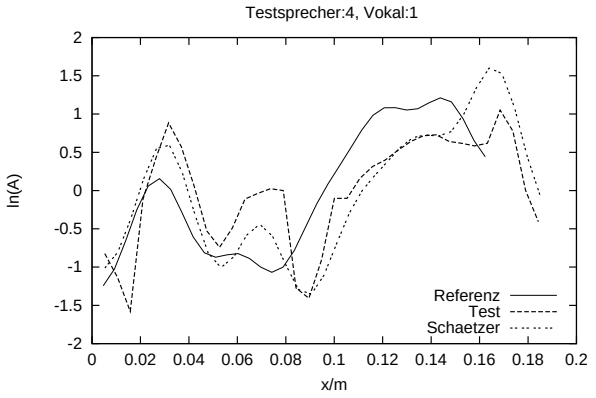
(a) Testsprecher 1



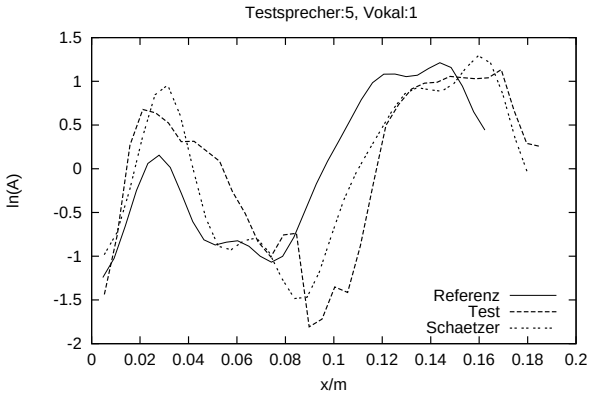
(b) Testsprecher 2



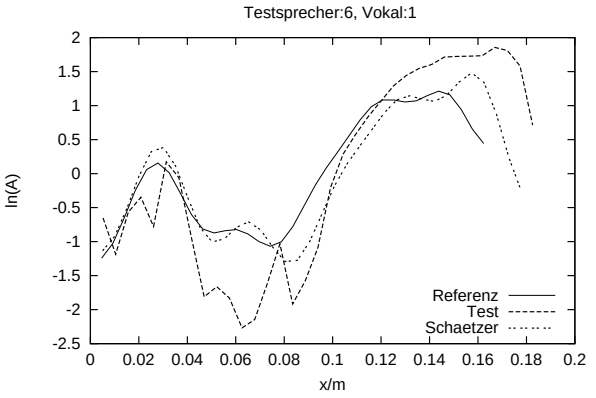
(c) Testsprecher 3



(d) Testsprecher 4



(e) Testsprecher 5



(f) Testsprecher 6

Abbildung 4.8.: Anpassung des Referenzsprechers an den Testsprecher durch unterschiedliche Längenskalierung beider Vokaltrakthälften und Querschnittsskalierung mit einer aus der Lagrangedichtefunktion abgeleiteten Störungsbasis.

4.6. Natürliche Sprache

In natürlicher Sprache sind Vokale der Koartikulation und Vokalreduktion unterworfen [47]. Koartikulation ist die kontextabhängige Artikulation von Vokalen, der Effekt der Vokalreduktion besteht darin, dass die Vokale in kontinuierlicher Sprache nicht mehr voll ausartikuliert werden. Sie nähern sich dem neutralen "schwa"-Laut /ə/ an. Das Vokaltrapez, das die Formanten F_1 und F_2 der verschiedenen Vokale aufspannt, zeigt eine entsprechende Tendenz zu schrumpfen.

Bei den folgenden Experimenten wird zunächst auf eine Datenbasis [36] zurückgegriffen, die isoliert gesprochene, gehaltene Vokale enthält. Damit können Koartikulation und Vokalreduktion ausgeschlossen werden. Im zweiten Teil dieses Abschnittes wird das Verfahren auf Vokale ausgedehnt, die aus fließender Sprache gewonnen wurden.

4.6.1. Vorverarbeitung der Daten

Die zu verarbeitenden Signale sind gehaltene, isoliert gesprochene Vokale. Die beschriebene Methode zur Extraktion sprecherspezifischer Parameter basiert auf der Verarbeitung von Formantmustern. Zur Bestimmung dieser Formantmuster wurde eine LPC-Analyse [3, 30, 32] des Sprachsignals durchgeführt, die zusammen mit Kaufmann [25] auf das spezielle Problem der stationären Vokale angepasst wurde. Die Parameter der LPC-Analyse sind in der folgenden Tabelle zusammengestellt.

LPC-Methode	Kovarianzmethode
Anzahl d. Formanten	4
Abtastfrequenz	10 kHz
Prädiktorordnung	12
Präemphase μ	15/16
Fensterlänge	40 ms
Fenstervorschub	10 ms

Aus dem LPC-Spektrum werden die Maxima als Startwerte für eine genauere Suche der Formanten extrahiert (*Peak-Picking*). Eine numerische iterative Nullstellensuche im Nennerpolynom des Prädiktions-Synthese-Filters, das sich aus den Prädiktorkoeffizienten ergibt, liefert eine Schätzung der Formantfrequenzen. Als problematisch erwies sich eine Nullstelle des Nennerpolynomes, die bei niedrigen Frequenzen (in der Nähe des ersten Formanten) gelegen ist und den Abfall des Spektrums modelliert. Bei der Nullstellensuche wird statt des ersten Formanten gelegentlich diese Nullstelle gefunden. Das geschätzte Formantmuster dieses Fensters verschiebt sich: Die Nullstelle wird als erster Formant, der erste Formant als zweiter Formant, usw. aufgefasst. Dieser Verschiebung wurde vorgebeugt, indem die Nullstellensuche mit dem Mittelwert[2] der aktuellen Formantfrequenzschätzung über alle Fenster als Startwert erneut durchgeführt wurde. Der

[2] Als geeigneter Mittelwert hat sich der Mid-Mean-Wert erwiesen – eine Kombination aus Median und arithmetischem Mittel. Das obere und untere Viertel der Werte wird verworfen und anschließend das arithmetische Mittel gebildet.

Vokal	ΔF_1	ΔF_2	ΔF_3	ΔF_4
i	0,444 %	0,142 %	1,010 %	0,331 %
ɪ	2,552 %	0,146 %	0,176 %	1,412 %
ɛ	4,093 %	0,411 %	0,445 %	1,625 %
æ	2,287 %	0,392 %	0,355 %	1,339 %
ʌ	0,020 %	1,162 %	0,020 %	0,690 %
ɑ	0,197 %	0,022 %	0,536 %	0,736 %
ɔ	2,213 %	0,892 %	0,333 %	0,093 %
o	3,636 %	0,512 %	0,227 %	2,237 %
ʊ	3,330 %	1,056 %	0,415 %	0,582 %
u	4,277 %	0,812 %	0,140 %	0,237 %

Tabelle 4.7.: Relativer Fehler der Formantschätzung aus synthetischen Vokalen mit der Grundfrequenz 100 Hz

letzte Schritt wird wiederholt, bis eine vorgegebene Anzahl von Iterationen durchlaufen wurde.

Die Methode wurde mit Hilfe von synthetischen Vokalen mit bekannten Formantfrequenzen getestet. Die relativen Fehler der Formantschätzung für verschiedene Vokale und Grundfrequenzen kann den Tabellen 4.7 und 4.8 entnommen werden.

4.6.2. Gehaltene, isoliert gesprochene Vokale

Das Ziel dieses Abschnittes ist, Einblicke in den Parameterraum zu gewinnen, der von der in den Simulationen bereits erprobten Störungsbasis aufgespannt wird. Es sollen folgende Fragen erläutert werden:

- Wie konsistent sind die einzelnen Parameter für die Vokale eines Sprechers?
- Wie verhalten sich die Parameter verschiedener Sprecher zueinander?

In den folgenden Versuchen wurde der freie Parameter der SVD ϵ_{SVD} zu 0,0025 gewählt. In Abbildung 4.9 sind die Längenskalierungsparameter der vier Vokale /aː/, /eː/, /iː/, /oː/ eines Sprechers aufgetragen[3]. Die y-Achse stellt die Gesamtskalierung gegenüber dem Referenzsprecher dar, auf der x-Achse ist das Skalierungsverhältnis der Mundhöhle im Verhältnis zur Gesamtskalierung aufgetragen. Da es sich um ein Verfahren handelt, das die einzelnen Zeitfenster, die sich aus der LPC-Analyse ergeben, unabhängig bearbeitet, ist die Standardabweichung der Werte für den jeweiligen Vokal eingetragen. Es ist zu erkennen, dass die einzelnen Vokale durchaus unterschiedliche Längenschätzungen ergeben. Diese liegen zwischen 7 % und etwa 17 % über der Vokaltraktlänge des Referenzsprechers. Die nichtlineare Längenskalierung ergibt ebenfalls

[3]Die Analyse des Vokals /uː/ wurde nicht mit den natürlichen Lauten durchgeführt, da die Querschnittsfunktion aus der Nihongo-Datenbasis sich nicht auf das Deutsche übertragen lässt.

Vokal	ΔF_1	ΔF_2	ΔF_3	ΔF_4
i	31,549 %	0,103 %	1,203 %	0,721 %
ɪ	2,279 %	0,013 %	0,760 %	1,028 %
ɛ	6,248 %	1,174 %	0,832 %	1,823 %
æ	2,241 %	0,410 %	0,791 %	1,062 %
ʌ	7,569 %	5,883 %	0,805 %	0,743 %
ɑ	0,015 %	1,511 %	0,083 %	0,679 %
ɔ	3,279 %	2,825 %	1,453 %	0,157 %
o	0,110 %	5,626 %	0,806 %	0,748 %
ʊ	0,150 %	1,344 %	0,795 %	0,131 %
u	27,652 %	0,491 %	0,807 %	0,426 %

Tabelle 4.8.: Relativer Fehler der Formantschätzung aus synthetischen Vokalen mit der Grundfrequenz 200 Hz

unterschiedliche Werte für die vier Vokale. Die Vokale /aː/, /iː/, /oː/ liegen etwa beim Verhältnis 0,54 der Längenskalierung Mundhöhle zur Gesamtlänge, während das /eː/ mit 0,49 fast beim neutralen Wert liegt.

Abbildung 4.10 zeigt für einen Sprecher die Koeffizienten zur Skalierung mit der Kosinusbasis (Abbildung 4.2). Der Koeffizient, der die vierte Komponente der Störungsbasis beschreibt, ist bei drei von vier Vokalen in Übereinstimmung. Bei den niedrigeren Koeffizienten der einzelnen Vokale ist kein eindeutiger Trend zu erkennen. Für denselben Sprecher sind die Ergebnisse dreier Messreihen, die an unterschiedlichen Tagen aufgenommen wurden, in Abbildung 4.11 aufgetragen. Die Parameter liefern für die drei Messungen konsistente Werte. Die Längenskalierungen sechs verschiedener Sprecher sind in Abbildung 4.12 dargestellt. Man erkennt, dass – wie zu erwarten – eine gute Unterscheidung zwischen männlichen und weiblichen Testsprechern auf Grund der Gesamtlängenskalierung getroffen werden kann. Die nichtlineare Längenskalierung der verschiedenen Vokale eines Sprechers ist, wie bereits beim ersten Sprecher festgestellt wurde, für die verschiedenen Vokale nicht einheitlich.

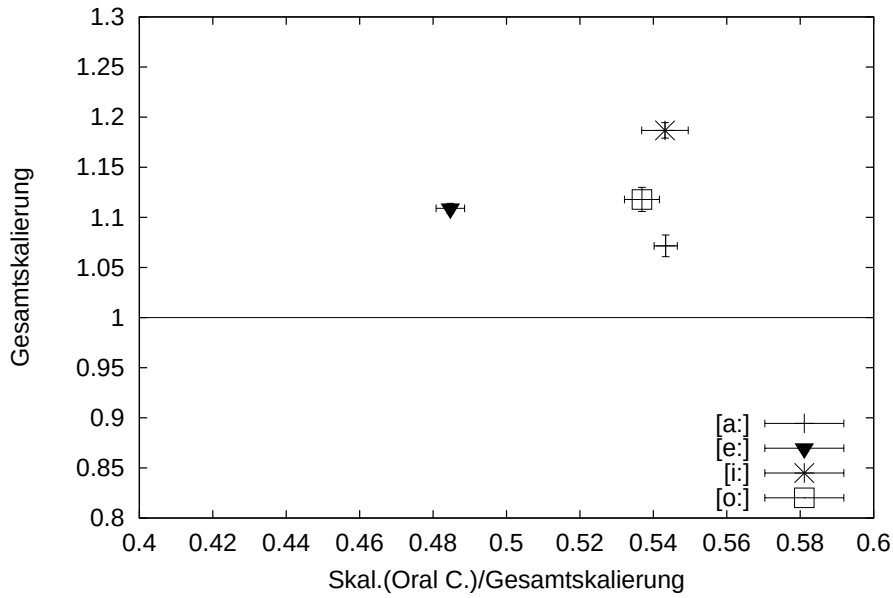


Abbildung 4.9.: Analyse isoliert gesprochener, gehaltener Vokale eines Sprechers.

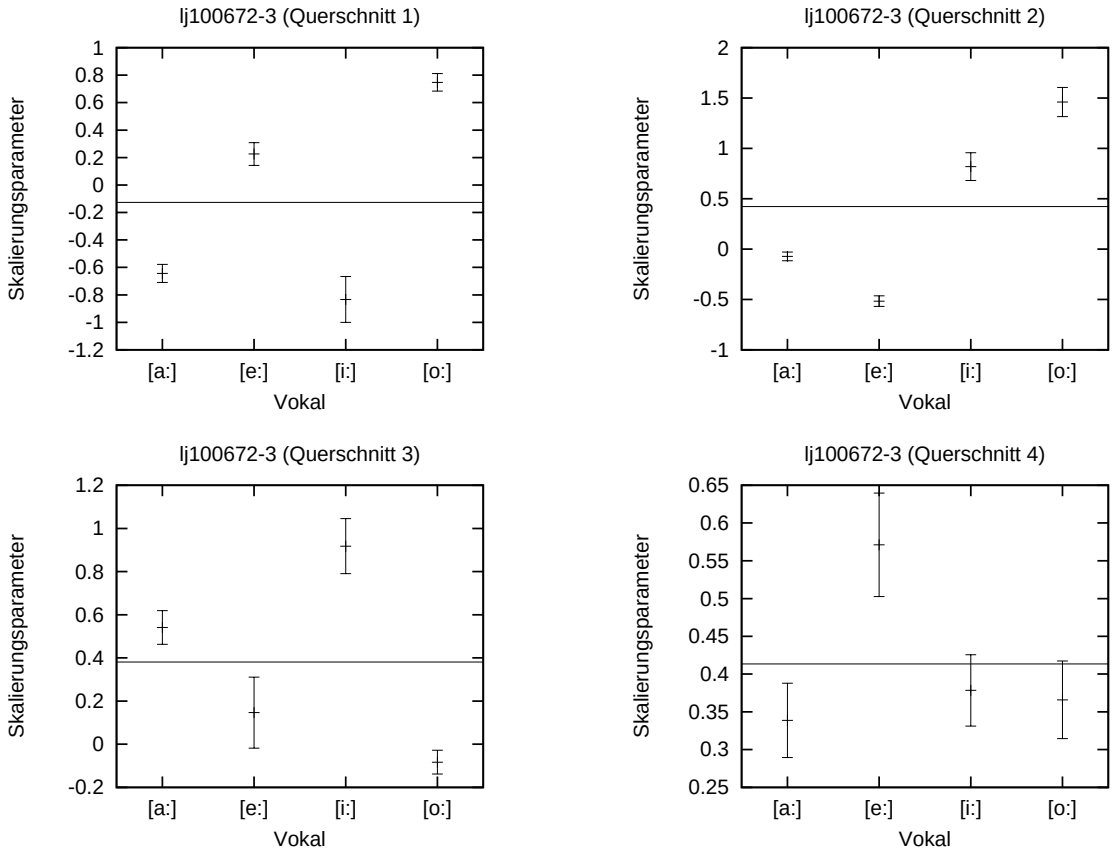


Abbildung 4.10.: Analyse isoliert gesprochener, gehaltener Vokale eines Sprechers. Aufgetragen sind die Koeffizienten der Kosinusstörungsbasis.

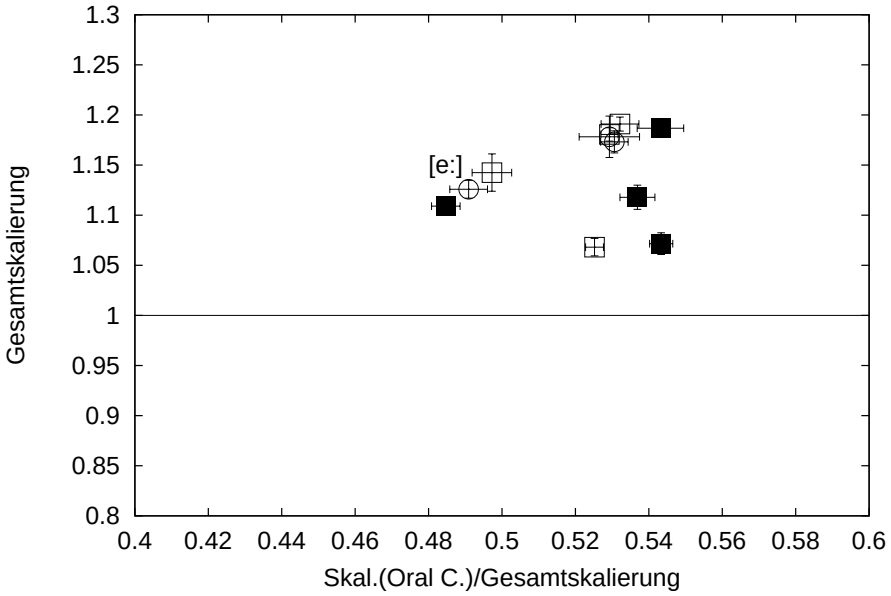


Abbildung 4.11.: Analyse isoliert gesprochener, gehaltener Vokale eines Sprechers. Aufgetragen sind die Ergebnisse verschiedener Aufnahmesitzungen, die mit verschiedenen Symbolen gekennzeichnet sind.

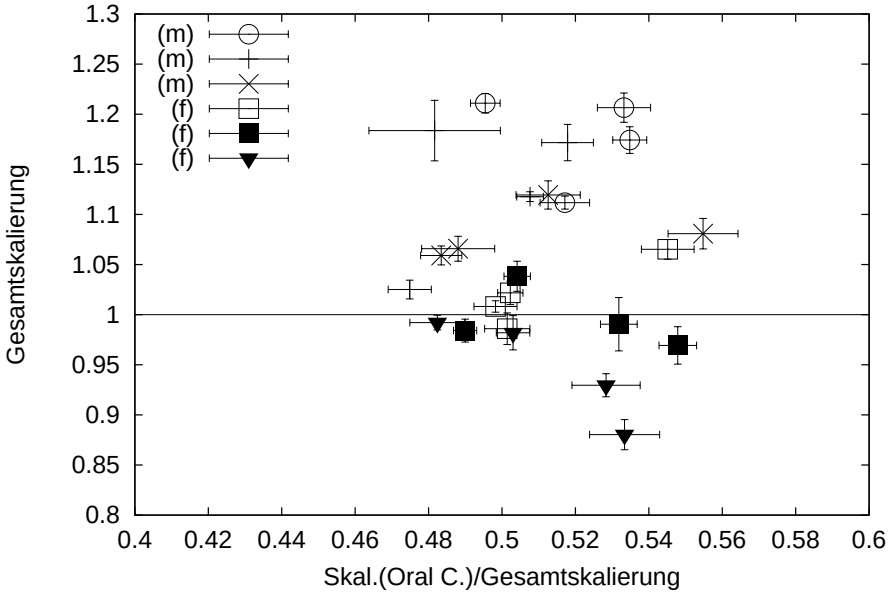


Abbildung 4.12.: Analyse isoliert gesprochener, gehaltener Vokale verschiedener Sprecher.

Um einen vollständigen Überblick der Skalierungsparameter zu erhalten, ist in den Abbildungen 4.13 bis 4.20 der volle Satz der Parameter dargestellt. Dieser unterteilt sich in die männlichen (Abbildung 4.13 bis 4.16) und weiblichen Sprecher (Abbildung 4.17 bis 4.20). Auf der x-Achse sind die verschiedenen Sprecher aufgetragen.

Teil (a) der Abbildungen zeigt jeweils die Längenskalierung der Mundhöhle im Verhältnis zur Gesamtlängenskalierung ($Skal_{\text{oral}}/Skal_{\text{total}}$) und die Gesamtlängenskalierung ($Skal_{\text{total}}$). Teil (b) der Abbildungen enthält die Koeffizienten der vier Komponenten der Kosinusstörungsbasis.

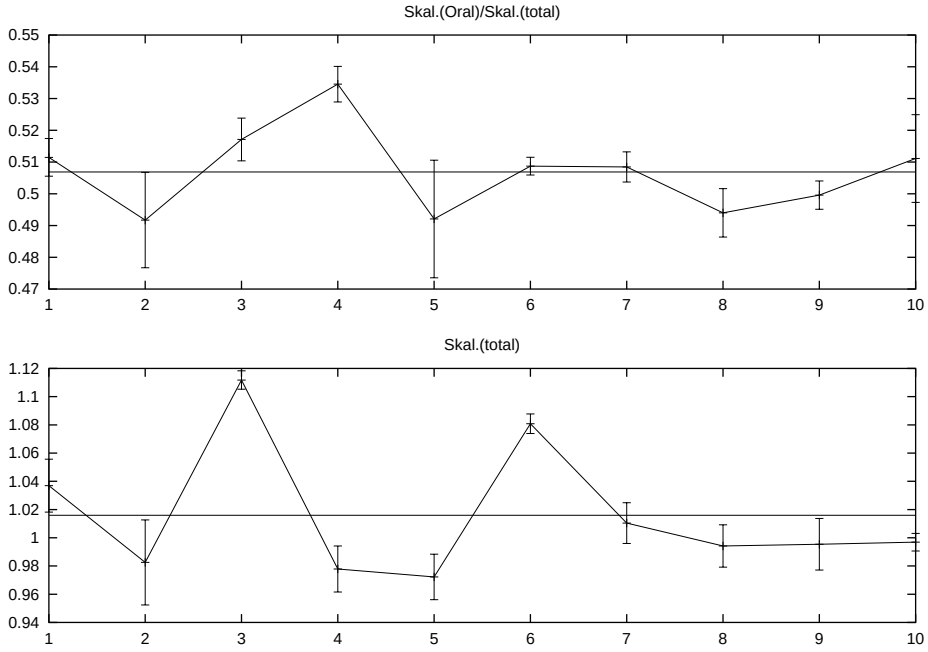
Bei den männlichen Sprechern ist die mittlere Gesamtlängenskalierung 1,02 für den Vokal /aː/, 1,11 für /eː/und 1,16 für /iː/ und /oː/. Das mittlere Verhältnis der Längenskalierung der Mundhöhle zur Gesamtlängenskalierung ($Skal_{\text{oral}}/Skal_{\text{total}}$) ist für /aː/ und /eː/ nahezu neutral, während /oː/ eine schwache und /iː/ mit 0,54 eine ausgeprägtere Verlängerung der Mundhöhle im Vergleich zur Gesamtlänge des Stimmkanals aufweisen.

Bei den weiblichen Sprechern liegt die mittlere Gesamtlängenskalierung zu den unterschiedlichen Vokalen mit etwa 0,9 – 1,0 deutlich unter der der männlichen Testsprecher. Hier ist das Verhältnis der Längenskalierung der Mundhöhle zur Gesamtlängenskalierung ebenfalls beim /iː/ mit 0,55 besonders ausgeprägt. Bei /aː/ liegt eine leichte Verschiebung des Verhältnisses zu einer verstärkten Skalierung der Mundhöhle im Vergleich zur Gesamtlänge vor. /oː/ und /eː/ sind nahezu neutral. Es wurde also kein geschlechtsspezifischer Trend bezüglich des Verhältnisses von Mundhöhle zur Gesamtlänge festgestellt.

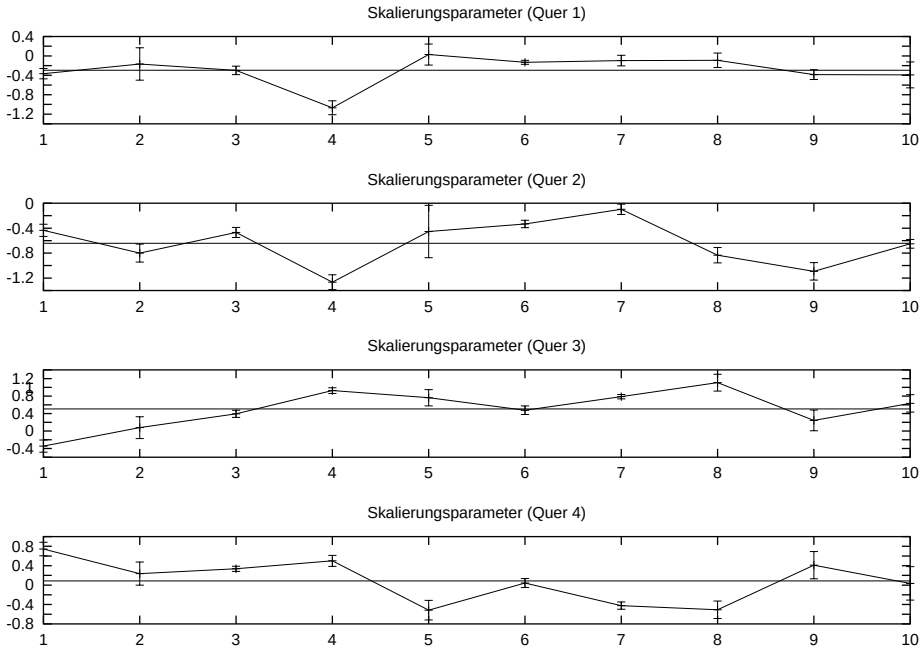
Die Gesamtlängenskalierung der männlichen Testsprecher ergibt Werte zwischen 0,95 und 1,35. Aus den Abbildungen 4.13 bis 4.16 kann folgende Zuordnung vorgenommen werden: Die männlichen Sprecher 3 und 6 erfahren gegenüber der mittleren Skalierung eine größere Gesamtlängenskalierung, während diese bei Sprecher 2, 5, 8 und 10 relativ klein ist. Sprecher 1, 4, 7 und 9 liegen im Mittelfeld.

Das Verhältnis der Längenskalierung der Mundhöhle zur Gesamtlängenskalierung ($Skal_{\text{oral}}/Skal_{\text{total}}$) ergibt sowohl für die männlichen als auch für die weiblichen Sprecher keinen ausgeprägten Trend. Die Koeffizienten zur Kosinusstörungsbasis der männlichen Sprecher lassen ebenfalls keine Tendenz erkennen.

Bei den weiblichen Sprechern zeigen eine relativ kleine Gesamtlängenskalierung Sprecher 1, 6 (mit einem unglaubwürdigen Ausreißer zu sehr großer Skalierung) und Sprecher 10. Eine mittlere Skalierung ergibt sich für die Sprecher 3, 4, 7 und 11. Größere Gesamtlängenskalierungen weisen Sprecher 2, 5, 8 und 9 auf. Interessanterweise gleichen sich bei den weiblichen Testsprechern die Koeffizienten der Komponenten der Kosinusstörungsbasis der Vokale /oː/ und /aː/, insbesondere die vierte Komponente.

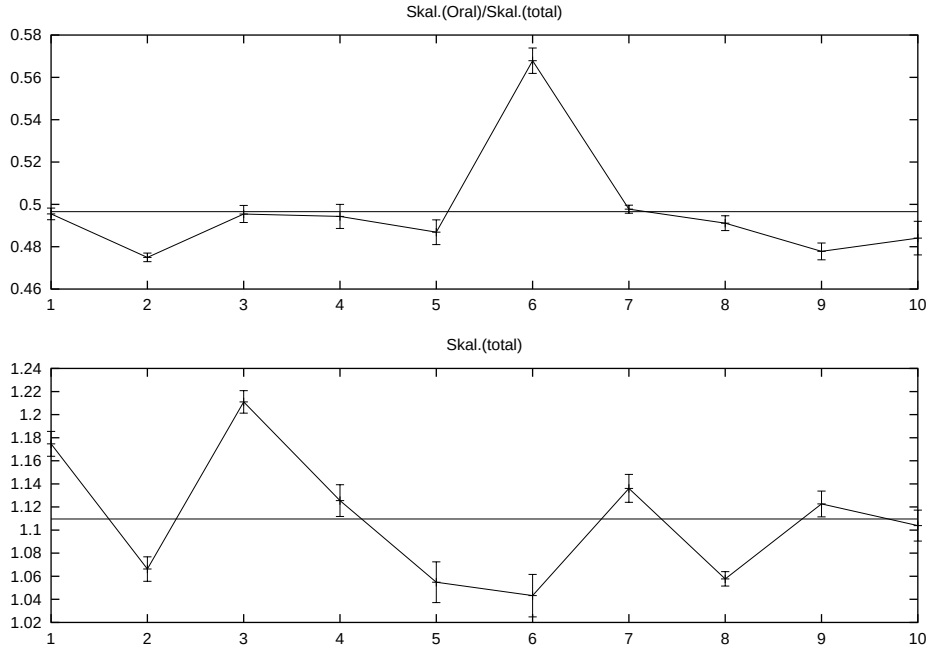


(a) Längenskalierung der männl. Sprecher 1 – 10

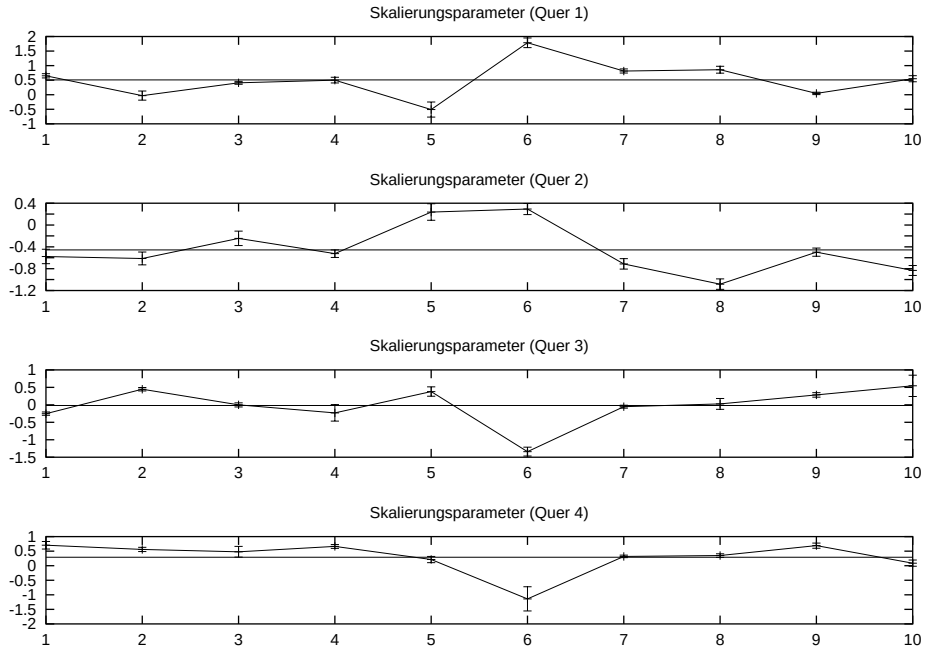


(b) Querschnittsskalierung der männl. Sprecher 1 – 10

Abbildung 4.13.: Überblick über den Vokal /aː/ aller männlichen Sprecher.

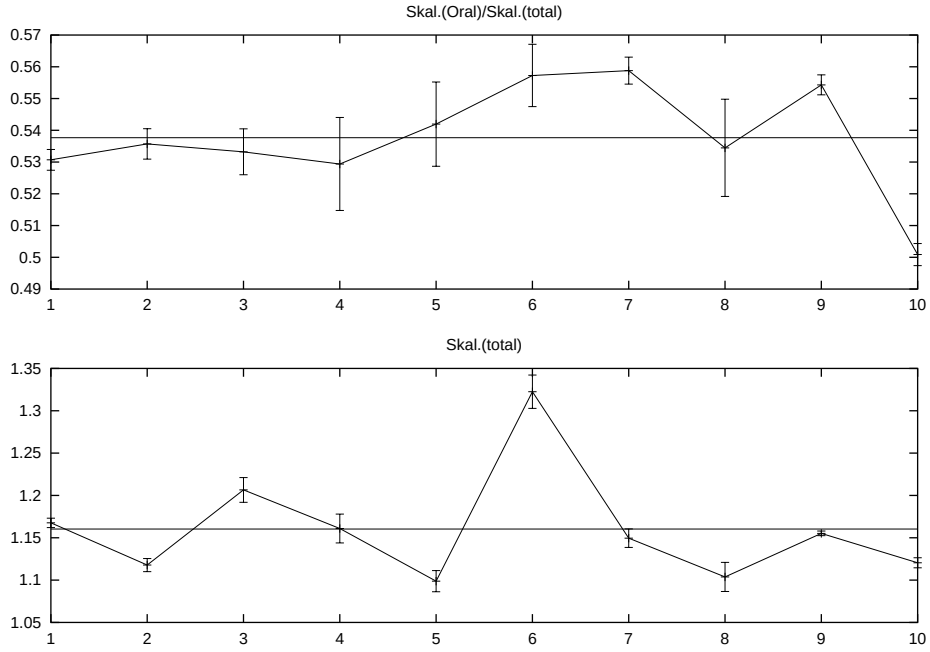


(a) Längenskalierung der männl. Sprecher 1 – 10

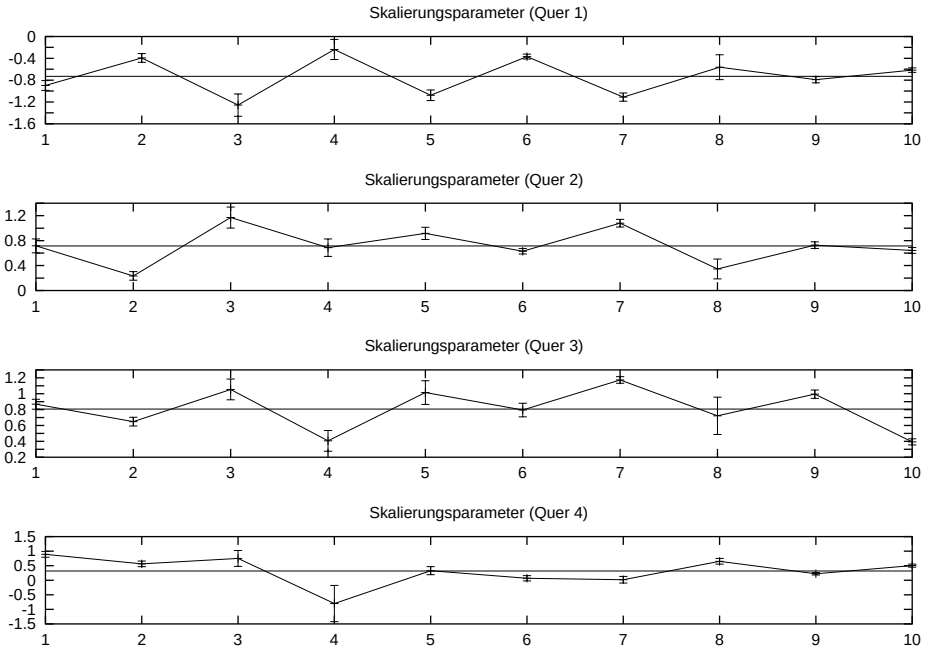


(b) Querschnittsskalierung der männl. Sprecher 1 – 10

Abbildung 4.14.: Überblick über den Vokal /eː/ aller männlichen Sprecher.

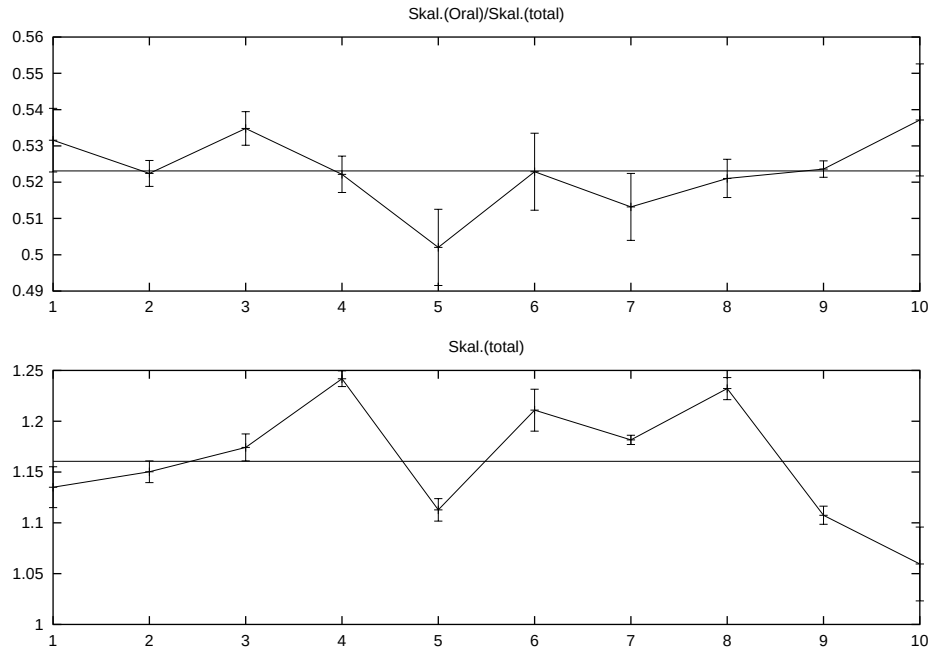


(a) Längenskalierung der männl. Sprecher 1 – 10

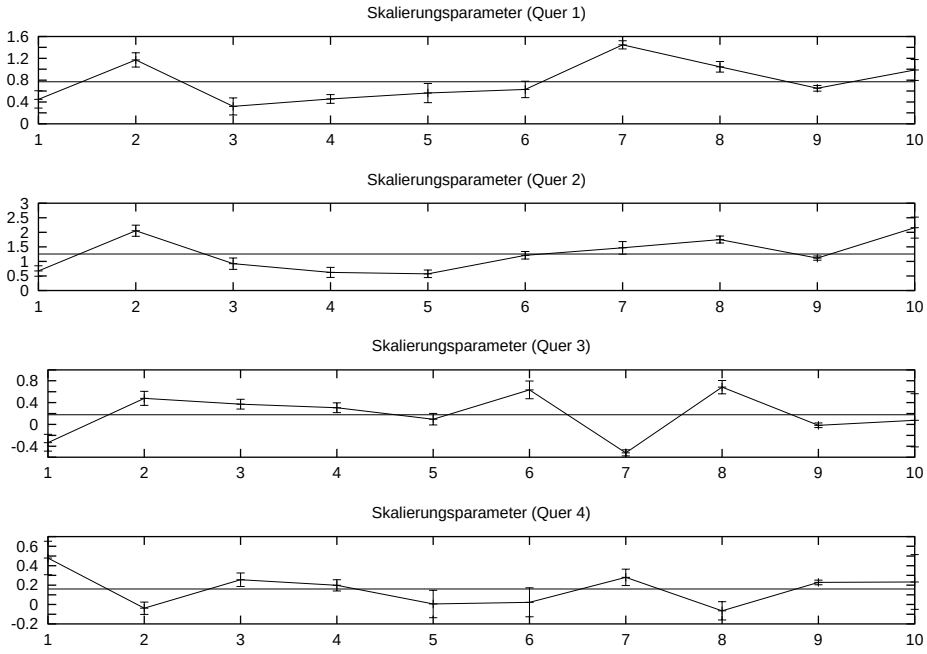


(b) Querschnittsskalierung der männl. Sprecher 1 – 10

Abbildung 4.15.: Überblick über den Vokal /iː/ aller männlichen Sprecher.

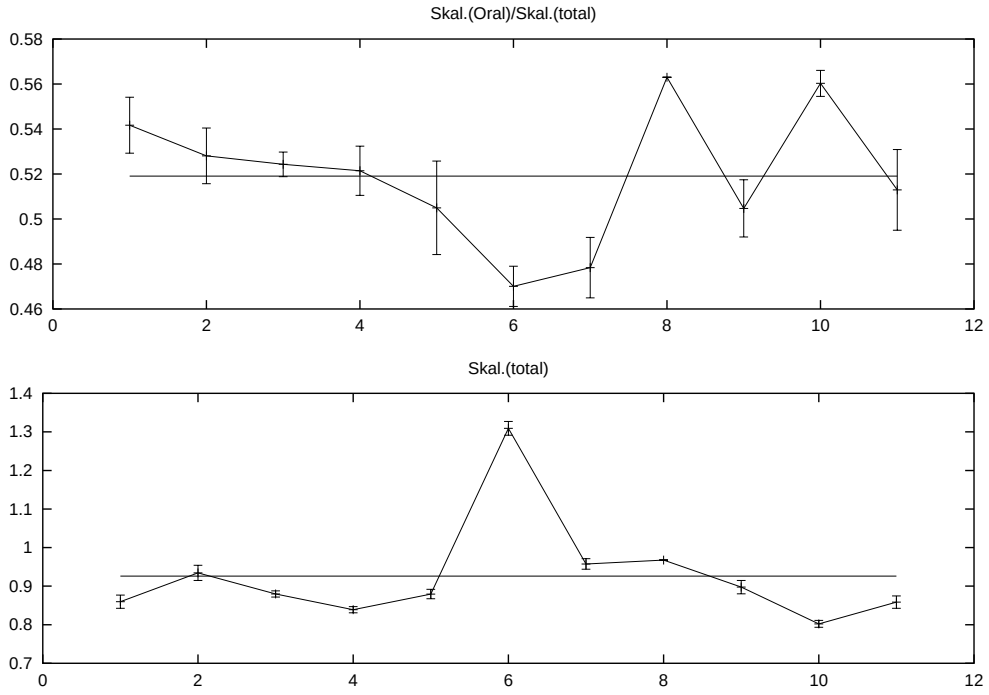


(a) Längenskalierung der männl. Sprecher 1 – 10

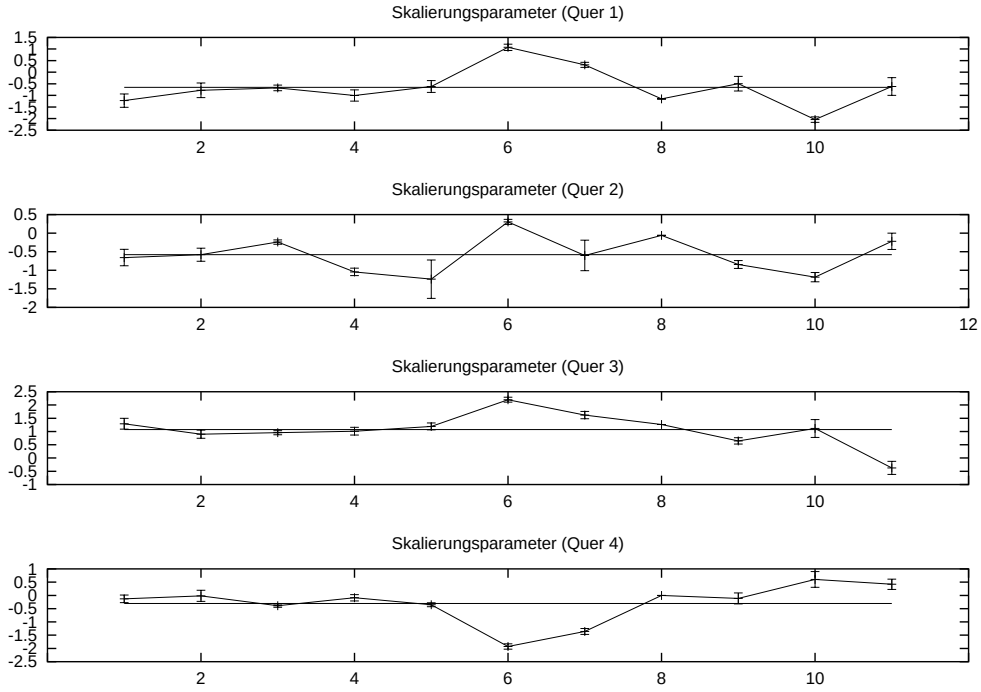


(b) Querschnittsskalierung der männl. Sprecher 1 – 10

Abbildung 4.16.: Überblick über den Vokal /oː/ aller männlichen Sprecher.

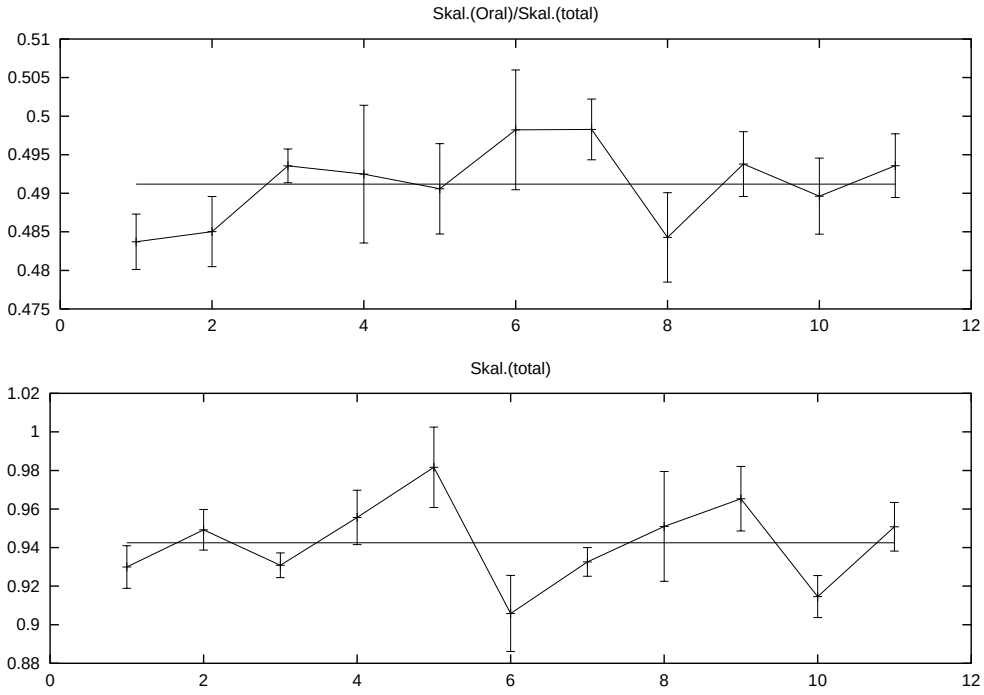


(a) Längenskalierung der weibl. Sprecher 1 – 11

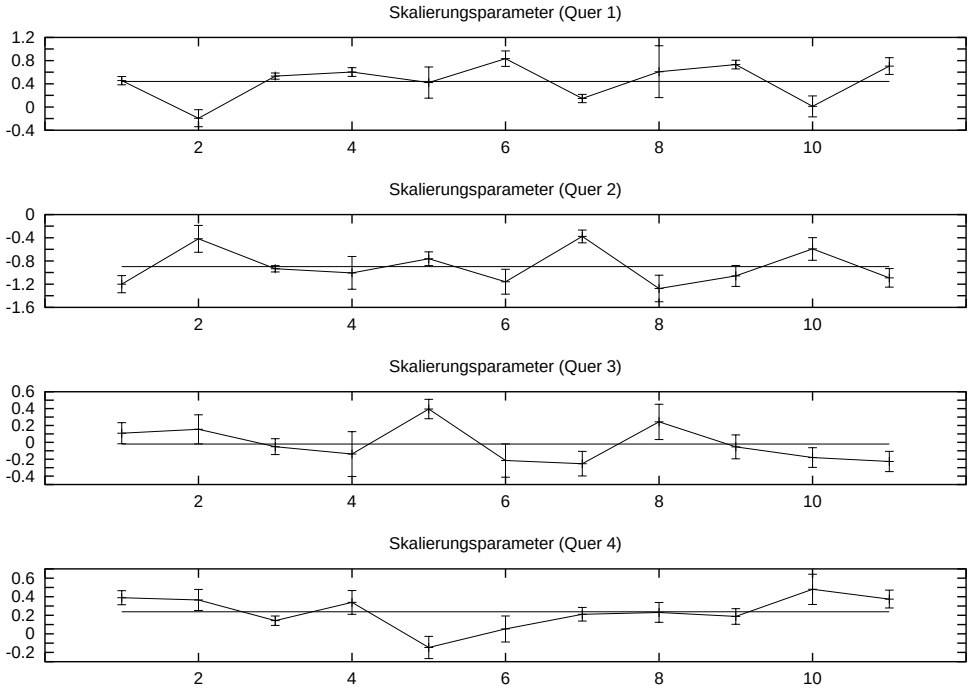


(b) Querschnittsskalierung der weibl. Sprecher 1 – 11

Abbildung 4.17.: Überblick über den Vokal /aː/ aller weiblichen Sprecher.

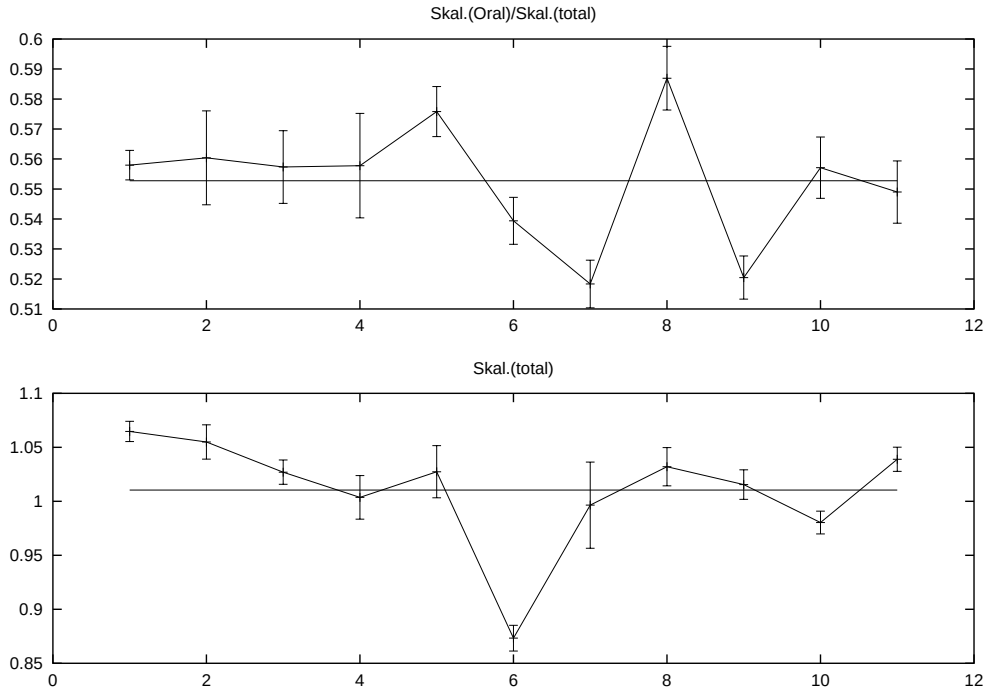


(a) Längenskalierung der weibl. Sprecher 1 – 11

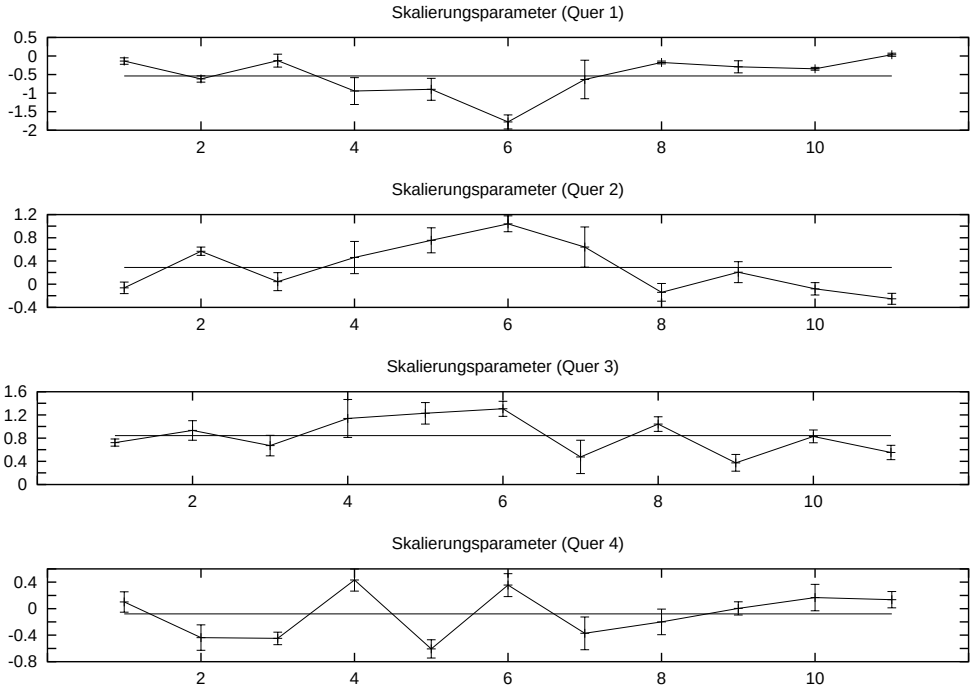


(b) Querschnittsskalierung der weibl. Sprecher 1 – 11

Abbildung 4.18.: Überblick über den Vokal /eː/ aller weiblichen Sprecher.

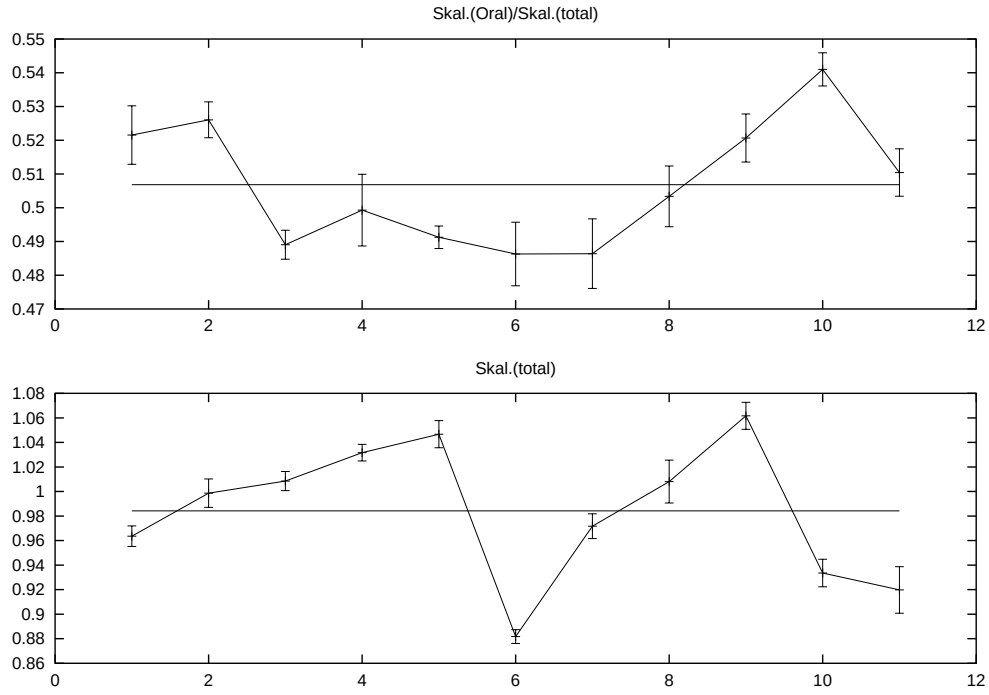


(a) Längenskalierung der weibl. Sprecher 1 – 11

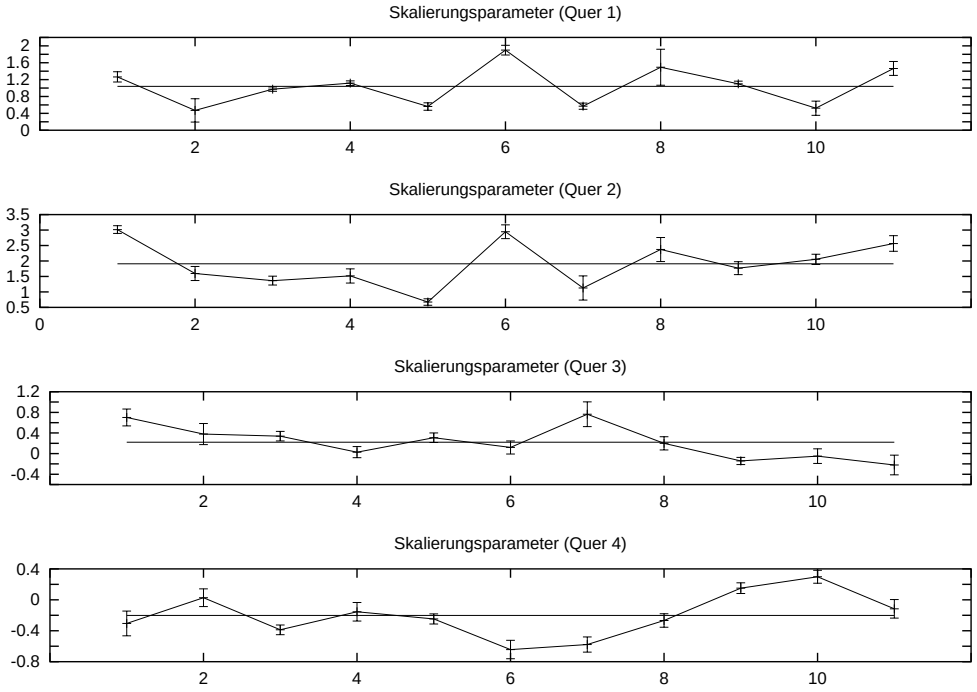


(b) Querschnittsskalierung der weibl. Sprecher 1 – 11

Abbildung 4.19.: Überblick über den Vokal /iː/ aller weiblichen Sprecher.



(a) Längenskalierung der weibl. Sprecher 1 – 11



(b) Querschnittsskalierung der weibl. Sprecher 1 – 11

Abbildung 4.20.: Überblick über den Vokal /oː/ aller weiblichen Sprecher.

4.6.3. Analyse von Vokalen aus fließender Sprache

Das Verfahren zur Schätzung von sprecherspezifischen Parametern ist eine Methode zur Analyse von gehaltenen Vokalen. In diesem Abschnitt wird das Verfahren dennoch auf aus fließender Sprache extrahierte Vokale angewandt. Es wurden gegenüber der Analyse von synthetischen Vokalen zwei Auswahlkriterien für gültige Schätzungen eingeführt:

- Nur Fenster, deren Signalenergie größer als 75 % der maximalen Energie ist, die in einem Fenster des betreffenden Vokals vorkommt, werden zur Analyse zugelassen.
- Überschreitet das Maß $\hat{\omega}_{\mathrm{e/t}}$ der relativen Formantabweichung des geschätzten Sprechers zum Testsprecher (vgl. Gleichung 4.6) den Wert 0,25, so wird die Schätzung für ungültig erklärt.

Die Abbildungen 4.21 und 4.22 zeigen die geschätzten Skalierungsparameter für den Vokal /a/ von sieben Sprechern (wobei *gvnn*, *rtda* und *wind* weibliche und *heid*, *kkoa* und *rorn* männliche Sprecher sind). Die in Abbildung 4.21 und 4.22 aufgetragenen Parame-

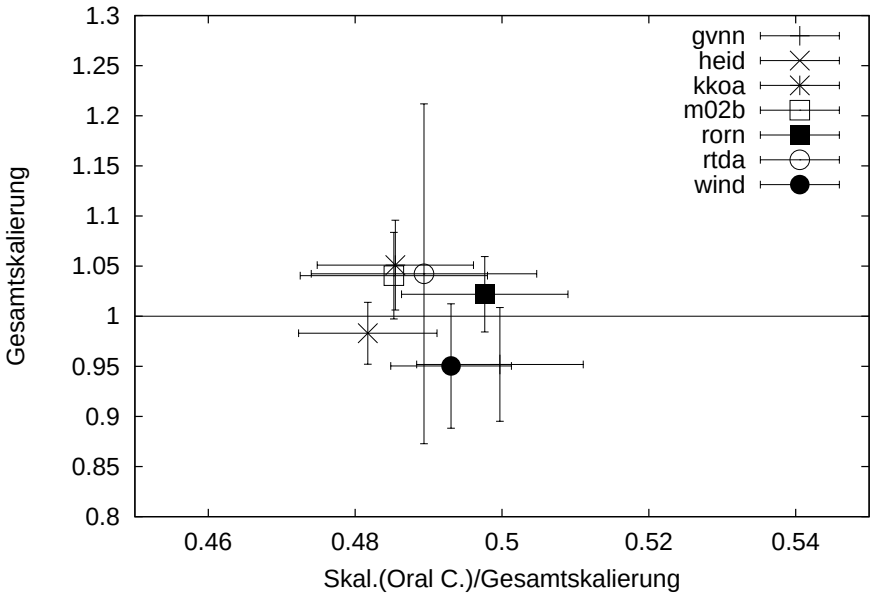


Abbildung 4.21.: Schätzung der Längenskalierung der Vokale /a/ von sieben Sprechern einer Datenbank mit fließender Sprache, PHONDAT 1.

ter beziehen sich auf ca. je 40 s Sprachsignal. Allgemein kann man beobachten, dass die Standardabweichung der Parameter stark zugenommen hat. Insbesondere Testsprecher *rtda* streut über den gesamten sinnvollen Parameterbereich. Unter den übrigen Sprechern kann dennoch in Abbildung 4.22 unterschieden werden. Sowohl die Gesamtskalierung als auch das Verhältnis der Skalierung der Mundhöhle zur Gesamtskalierung sind zur Unterscheidung der Sprecher geeignet. Besonders die Koeffizienten, die zum ersten und vierten Kosinusterm der Querschnittsstörungsbasis gehören (*Quer 1* und *Quer 4* in Abb. 4.22), unterscheiden sich für die unterschiedlichen Sprecher. Für den Vokal /e/ (Abbildungen

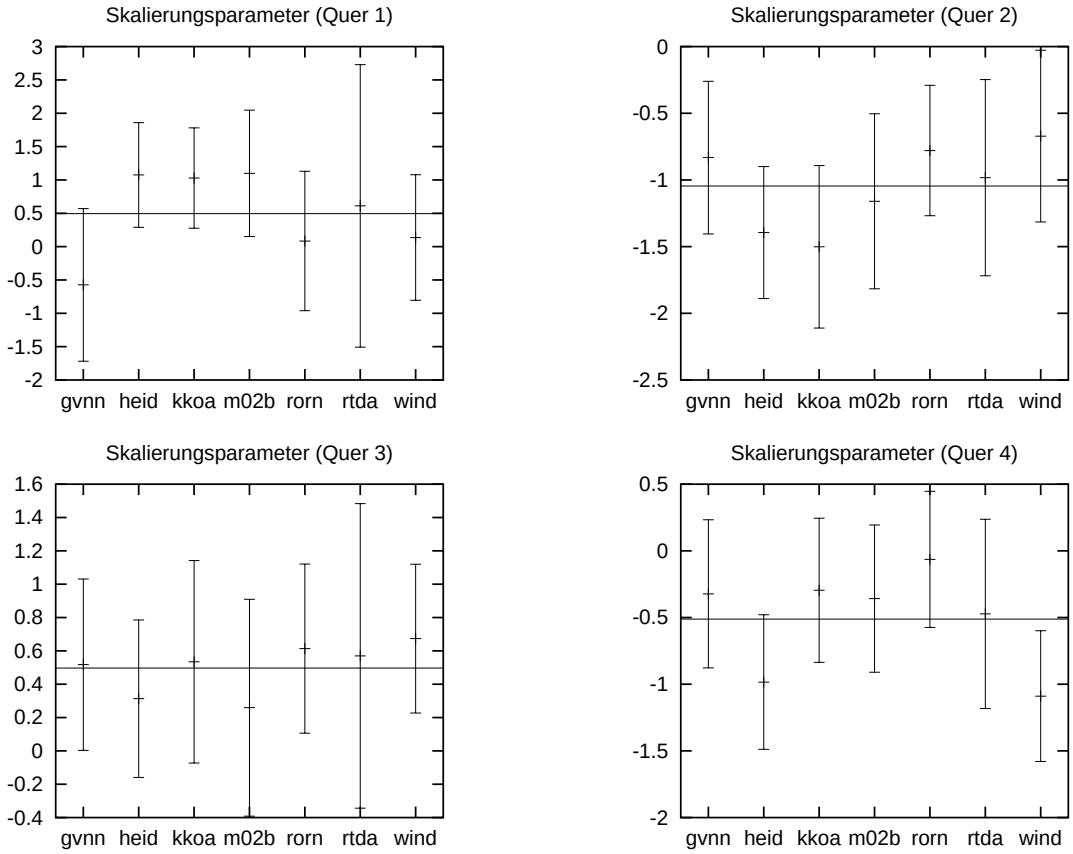


Abbildung 4.22.: Schätzung der Querschnittsskalierung der Vokale /a/ von sieben Sprechern einer Datenbank mit fließender Sprache, PHONDAT 1.

4.23 und 4.24) gilt Ähnliches. Das zugrunde liegende Sprachsignal besteht aus ca. 14 s geschnittener Vokale /e/. Für diesen Vokal liefert *rtda* auch brauchbare Schätzungsparameter. Das Verhältnis der Skalierung der Mundhöhle zum Rachenraum ist für alle Sprecher nahe 0,5, dem neutralen Wert.

In einem letzten Experiment werden die Parametersätze benutzt, um ein Sprechererkennungsexperiment durchzuführen.

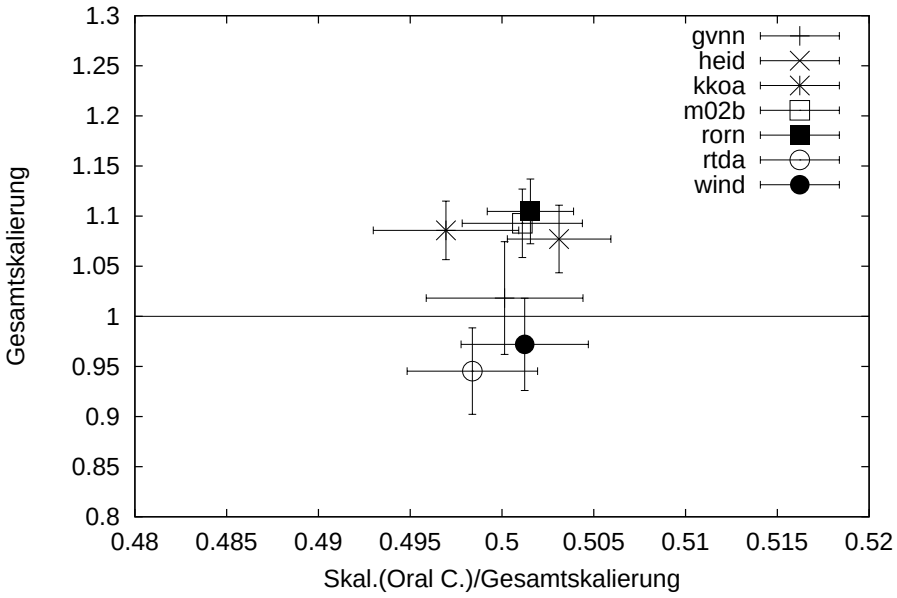


Abbildung 4.23.: Schätzung der Längenskalierung der Vokale /e/ von sieben Sprechern einer Datenbank mit fließender Sprache, PHONDAT 1.

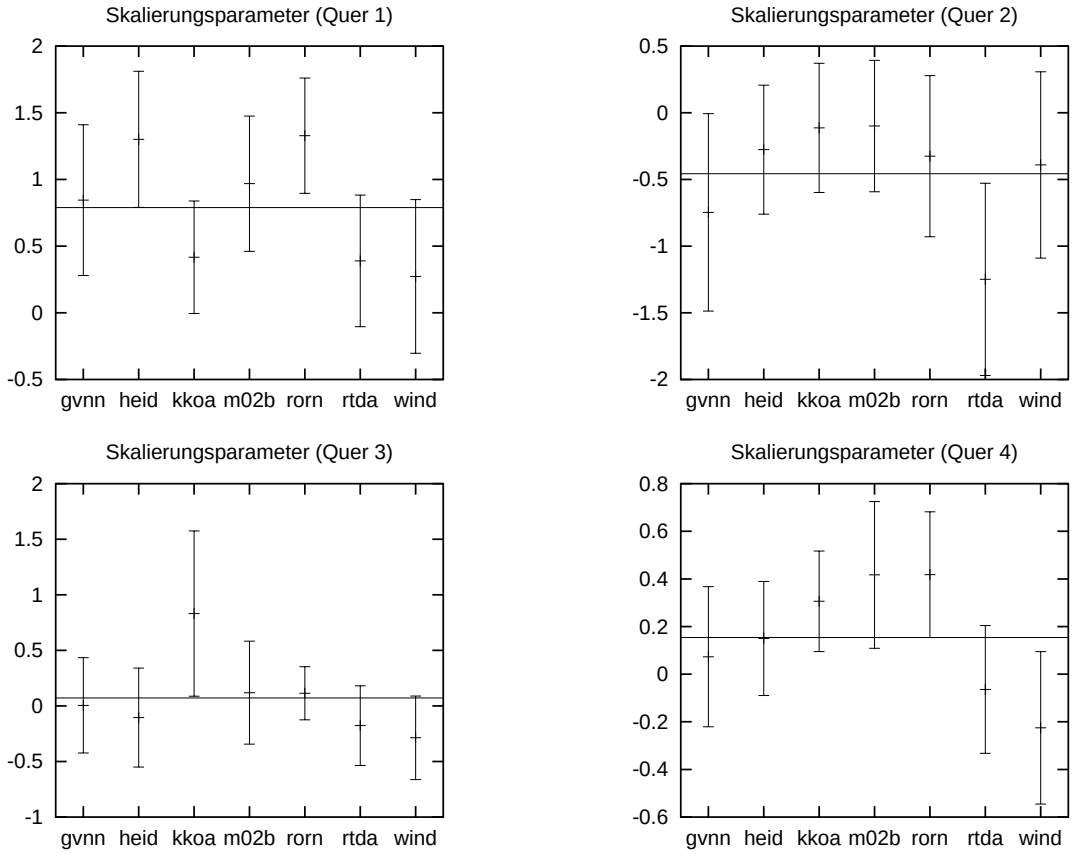


Abbildung 4.24.: Schätzung der Längenskalierung der Vokale /e/ von sieben Sprechern einer Datenbank mit fließender Sprache, PHONDAT 1.

4.6.4. Das Kontrollnetzwerk

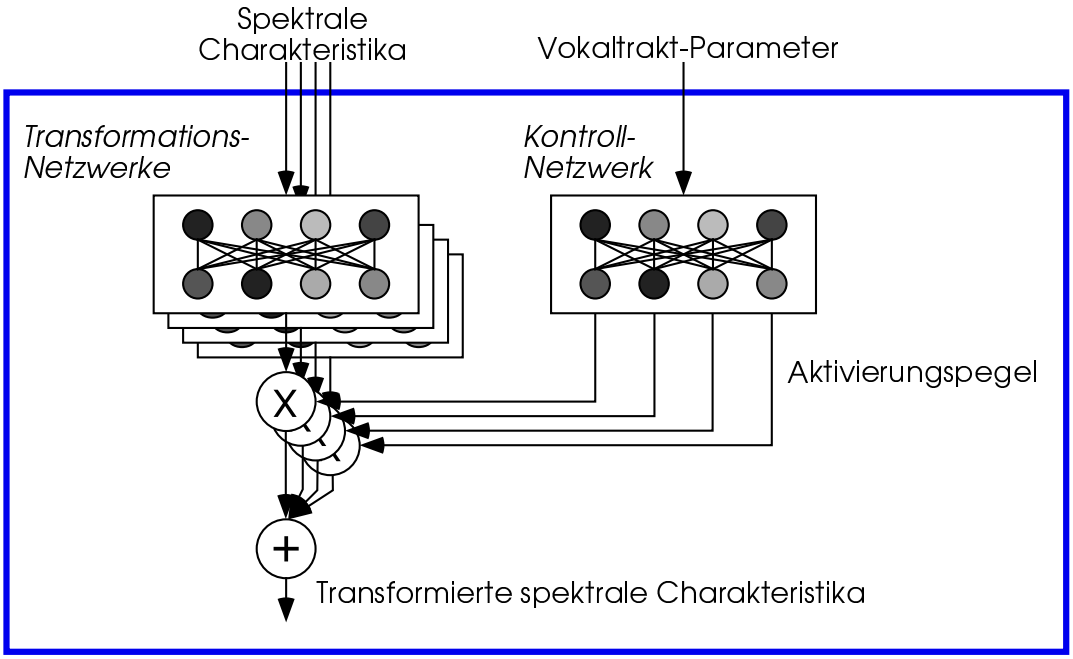


Abbildung 4.25.: Architektur des Normalisierungsmoduls zur automatischen Spracherkennung.

Dieser Abschnitt schließt den Bogen: Die extrahierten sprecherspezifischen Parameter werden in einem ersten Erkennungsexperiment getestet. Die Architektur des Normalisierungsmoduls ist in Abb. 4.25 dargesellt. Man erkennt ein Kontrollnetzwerk und eine Schar von Transformationsabbildungen. Die Aufgabe des Kontollnetzwerkes ist im einfachsten Fall die Bestimmung derjenigen Abbildung, die den Testsprecher optimal auf den Referenzsprecher transformiert[4]. Diese Aufgabe ist verwandt mit der der Sprecherklassifizierung. Ist das Netzwerk in der Lage, einem gegebenen Satz sprecherspezifischer Parameter einen bestimmten Sprecher zuzuordnen, kann zur Normalisierung die dem Sprecher zugeordnete normalisierende Abbildung eingesetzt werden. Eine Verbesserung der Zuordnung durch die geschätzten sprecherspezifischen Vokaltraktparameter gibt Einblicke in ihre Bedeutung.

Das konkrete Testszenario besteht aus einem *Backpropagation*-Netzwerk [42] als Implementation des Kontrollnetzwerks. Eingänge sind die sechs Kanäle für sprecherspezifische Parameter. Zum Test wurden sechs Sprecher aus der deutschen PhonDat 1-Sprachdatenbank ausgewählt. Es wurden die Vokale /a/ extrahiert. Die Zahl der Ausgangskanäle entspricht der Anzahl der Sprecher. Der Anteil der korrekten Klassifizierung auf Grund der Parameter des Vokals /a/, die in den Abbildungen 4.21 und 4.22 abgebildet sind, ergab sich in dem Vorexperiment zu 66,7 %. Dieses Ergebnis ist noch

[4] Es ist ferner vorgesehen, dass das Kontrollnetz eine auf den Testsprecher zugeschnittene optimale Interpolation der Transformationsabbildungen erstellt.

nicht befriedigend, zeigt aber, dass trotz der großen Standardabweichungen der einzelnen Parameter und der nahezu unbrauchbaren Parameter der Sprecherin *rtda* ein Ergebnis erzielt wird, das im Vergleich zur Ratewahrscheinlichkeit von 16,67 % sehr hoch ist. Weitere Untersuchungen mit unterschiedlichen Parametersätzen, Vorverarbeitungen und Netzwerkarchitekturen werden in der Dissertation von Knut Müller zu finden sein.

5. Diskussion und Ausblick

Das vorgestellte Verfahren zur Extraktion sprecherspezifischer Parameter aus Formantfrequenzen wurde zunächst an synthetischen Vokalen mit bekannten Querschnittsfunktionen getestet. Für eine erfolgreiche Schätzung der Skalierungsparameter des jeweiligen Testsprechers ist als Ausgangspunkt eine zum untersuchten Vokal gehörende Querschnittsfunktion eines Referenzsprechers Voraussetzung. Den Untersuchungen an synthetischen Vokalen, aber auch an natürlichen, stationären Vokalen und an Vokalen aus fließender Sprache lag als Querschnittsfunktion des Referenzsprechers die über den entsprechenden Vokal von sechs japanischen Sprechern gemittelte und geglättete Querschnittsfunktion zugrunde. Als beste Wahl der Störungsbasis, in der die Abweichungen vom Referenzsprecher beschrieben werden, haben sich die ungerade indizierten Terme der Kosinusreihe herausgestellt. Die Wahl dieser Störungsbasis in Kombination mit dem oben beschriebenen Referenzsprecher führte in den Experimenten an synthetischen Vokalen zu guten Ergebnissen. In den durchgeführten Simulationen mit synthetischen Vokalen wurden für fünf von sechs Sprechern eine gute Anpassung der Länge und des Querschnittes erreicht. Der mittlere relative Fehler der Längenschätzung von sechs Sprechern und fünf Vokalen liegt bei 3,5 %. Das Verfahren nach Paige und Zue [40] hingegen erreicht als relativen Fehler im Mittel 8,4 % (bzw. 7,0 % als Sonderfall des vorgestellten Verfahrens).

Die Experimente an gehaltenen, isoliert gesprochenen Vokalen führten zu dem Ergebnis, dass das Verhältnis der Skalierung der Mundhöhle zur Gesamtlängenskalierung durch das Verfahren nicht als reines Sprecherspezifikum extrahiert werden konnte. Die einzelnen Vokale eines jeden Sprechers glichen sich nicht in diesem Parameter. Es wurde auch kein Hinweis darauf gefunden, dass ein geschlechtsspezifischer Trend dieses Quotienten existiert, denn die Mittelwerte dieses Parameters waren sowohl für männliche als auch weibliche Sprecher vergleichbar. Die Gesamtlängenschätzung dagegen ließ einen Trend bezüglich der einzelnen Sprecher erkennen. Die Simulationen zeigten, dass neben der Anpassung der Längenverhältnisse der Skalierung des Querschnitts bei der Schätzung der Querschnittsfunktion aus dem Referenzsprecher eine große Bedeutung zukommt. Im Gegensatz zu den Betrachtungen von Naito et al. [38, 39] wurden auch bei der Analyse von gesprochenen Vokalen diese Größen in der vorliegenden Arbeit berücksichtigt. Für die Vokale /a:/ und /o:/ der weiblichen Testsprecher ergaben sich ausgeprägte Übereinstimmungen, was den Schluss nahelegt, dass diese Parameter sprecherspezifische Information unabhängig von artikulatorischen Eigenheiten tragen. Bei den männlichen Testsprechern hingegen konnte kein Trend über verschiedene Vokale beobachtet werden.

Insgesamt sprechen die erzielten Ergebnisse dafür, dass die individuellen Sprachgewohnheiten der Artikulation verschiedener Vokale die sprecherspezifischen anatomischen

Eigenheiten zum Teil stark überdecken, was zur Konsequenz hat, dass die extrahierten Parameter als Beschreibung der sprecherspezifischen Realisation des jeweiligen Lautes betrachtet werden müssen.

Die Schätzung der Skalierungsparameter aus Vokalen, die aus fließender Sprache gewonnen wurden, wurde ebenfalls untersucht. Die große Standardabweichung der einzelnen Parameter ist auf die Effekte von Koartikulation und Vokalreduktion zurückzuführen. Vermutlich lassen sich Fehlerquellen dieser Art reduzieren, indem die analysierten Vokale einer strengeren Auswahl unterworfen werden. Zum einen können die in Abschnitt 4.6.3 eingeführten Kontrollmechanismen der Schwellwerte der Signalenergie und der spektralen Approximation verschärft werden. Zum anderen sind weitere Auswahlkriterien denkbar, wie z. B. die Elimination von Signalfenstern, deren Formantmuster im Vokaltrapez nicht den ausartikulierten Vokalen entsprechen, sondern dem neutralen Laut /ə/ nahe kommen. Von diesen Mechanismen wurde bisher abgesehen, da die genutzte Datenbasis zum Training der neuronalen Netze möglichst viele Werte enthalten sollte. Ein erstes Sprechererkennungsexperiment lässt vermuten, dass die extrahierten Parameter eines Vokals trotz der großen Schwankungen sprecherspezifische Information (diesmal im Sinne von individuellen artikulatorischen und anatomischen Eigenheiten) enthalten, die die Steuerung des Kontrollnetzes (gegebenenfalls unter Berücksichtigung der spektralen Parameter) ermöglicht. Diese Vermutung wird im zweiten Teil dieses Projektes, der Sprechernormalisierung, im Detail überprüft werden. Dazu muss das Problem der Vokalreduktion und Koartikulation eingehender untersucht und die bereits erwähnten Maßnahmen bezüglich einer strengeren Vokalauswahl durchgeführt werden.

6. Zusammenfassung

Sprecherspezifische Parameter des Stimmkanals, wie z. B. dessen mittlere Länge, der mittlere Quotient der Querschnitte von Rachenraum und der Mundhöhle oder der Quotient von deren Längen sind für Sprecherunterschiede mitverantwortlich und somit für die Sprechernormalisierung von Bedeutung.

Es wurde ein neues Verfahren zur Extraktion dieser sprecherspezifischer Parameter aus Formantfrequenzen vorgestellt. Die zugrunde liegende Idee ist, dass sprecherspezifische Eigenheiten der Querschnittsfunktion eines Testsprechers durch kleine Störungen der bekannten Querschnittsfunktion desselben Lautes eines Referenzsprechers beschrieben werden können. Es ist bekannt, dass Auswirkungen von kleinen Störungen der Geometrie in Länge und Querschnitt des Stimmkanals auf die Formantfrequenzen sich mit Hilfe einfacher linearer Gleichungen ausdrücken lassen.

Dieser Ausgangspunkt führte zu einem neuen Verfahren zur Schätzung von sprecherspezifischen Parametern aus Formantfrequenzmustern. Nachdem wenige relevante Größen des Referenzspechers mit Hilfe eines Vokaltraktmodells einmalig berechnet worden ist, bestimmt das entwickelte Verfahren simultan nichtlineare Längenskalierungen und Skalierungen der Querschnittsfunktion, die die Abweichung des Testsprechers vom Referenzsprecher charakterisieren. Das Verfahren zeichnet sich durch seine Einfachheit aus, denn die Berechnung der Störungen aus den Formantfrequenzen stellt im Wesentlichen eine Matrixmultiplikation dar. Da es sich um einen empirischen Ansatz handelt, wurde das Verfahren in Simulationen mit synthetischen Vokalen, mit natürlichen, stationären Vokalen und schließlich mit Vokalen aus fließender Sprache getestet. Die Simulationen ergeben in den meisten Fällen gute Schätzungen der Querschnittsfunktionen des Testsprechers aus den Formantmustern und liefern eine zuverlässige Schätzung der Gesamtlängenskalierung des Testsprechers gegenüber dem Referenzsprecher.

Die Ergebnisse der Analyse statischer Vokale deutet darauf hin, dass die individuellen Sprachgewohnheiten die sprecherspezifischen anatomischen Eigenheiten zum Teil stark überdecken. Die extrahierten Parameter müssen daher als Beschreibung der sprecherspezifischen Realisation des jeweiligen Lautes interpretiert werden.

Die ermittelten Parameter aus Vokalen, die aus natürlicher Sprache gewonnen wurden, sind starken Schwankungen unterworfen. Diese sind der Koartikulation und Vokalreduktion zuzuordnen. Ein erstes Sprechererkennungsexperiment gibt Grund zur Annahme, dass diese Parameter trotz starker Streuung sprecherspezifische Information enthalten.

A. Seitenarme des Stimmkanals

Der Fokus dieser Arbeit liegt auf der Extraktion sprecherspezifischer Parameter aus Formantmustern des unverzweigten Stimmkanals. In diesem Kapitel wird speziell auf die Bedeutung von angekoppelten Seitenarmen eingegangen. Da der Querschnitt des Nasaltraktes nicht von Artikulation betroffen ist, wird im Folgenden auf die Auswirkungen der Nasenhöhle auf das Spektrum eingegangen. Hörtests belegen, dass insbesondere das Frequenzband über 20 ERB (1740 Hz) Träger sprecherspezifischer Information ist [27]. Das stützt die Vermutung, dass kleine Strukturen wie z. B. die Schleimhautausbuchtungen im Bereich zwischen Kehle und Rachen, die sogenannte Fossa piriformis, eine gewisse Rolle spielen. Sowohl der Nasaltrakt als auch die Fossa piriformis können als an den Hauptast des Stimmkanals angekoppelte Seitenarme betrachtet werden. Durch Ankopplung von Seitenarmen verändert sich die Übertragungsfunktion eines Vokales deutlich. Auffällig sind Nullstellen in der Übertragungsfunktion, sogenannte Antiresonanzen, aber auch zusätzliche Pole. Nullstellen in der Übertragungsfunktion enstehen, wenn die Impedanz des Seitenarmes gegen Null geht. Diese Situation des Kurzschlusses tritt ein, wenn die Resonanzfrequenz des Seitenarmes getroffen wird.

Eine einfache Überlegung verhilft zu qualitativem Verständnis der zusätzlichen Pole. Es ist aus der allgemeinen Netzwerktheorie bekannt, dass jede mögliche Übertragungsfunktion des Systems (die einen beobachteten Strom oder eine Spannung mit dem Strom oder der Spannung einer beliebigen Quelle, die die Impedanzstruktur nicht stört, in Verbindung bringt) dieselben Pole hat [15], da alle Übertragungsfunktionen dieselbe Systemdeterminante besitzen (siehe z. B. [29]). Also kann man den Ort der Anregung auch in den Verzweigungspunkt der verschiedenen Äste legen ([13], S. 145), wie in Abb. A.1 a durch einen Pfeil angedeutet ist. Das System wird nun als Parallelschwingkreis aufgefasst (Abb. A.1 b). Die Resonanzbedingung des Kreises ist, dass sich die Blindströme der einzelnen Zweige aufheben [35], sodass nur der ohmsche Widerstand übrig bleibt. Bei Fant [13] findet sich ein aus dieser Bedingung (unter Vernachlässigung der Dämpfung) resultierender Ansatz zur grafischen Bestimmmung der Resonanzen aus den Admittanzen:

$$Y_P + Y_0 + Y_N = 0. \tag{A.1}$$

Bei der grafischen Bestimmung wird die Frequenz des Schnittpunktes der beiden frequenzabhängigen Terme $Y_P + Y_0$ und $-Y_N$ bestimmt. Unter Berücksichtigung der Dämpfung gilt der Merksatz: Die Summe der Suszeptanzen am Ankopplungspunkt ist in der Resonanz Null [7]. Der Einfluss der Seitenarme wird in den nächsten Abschnitten am Beispiel des Nasaltraktes und der Fossa piriformis diskutiert.

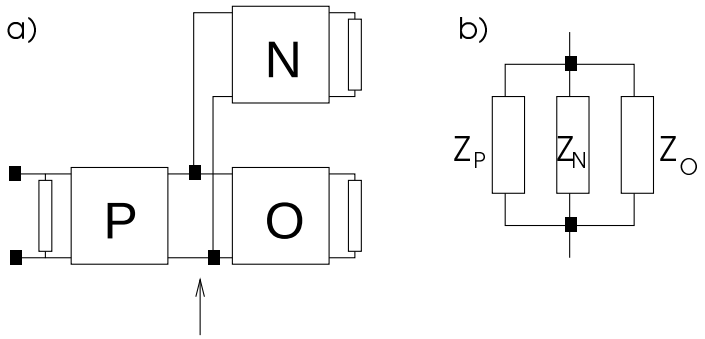


Abbildung A.1.: Betrachtung zur Abschätzung von zusätzlichen Polen der Übertragungsfunktion bei Ankopplung eines Seitenarmes an den Stimmkanal.

A.1. Nasaltrakt

Der Nasaltrakt kann als Seitenarm des Stimmkanals betrachtet werden. Die Artikulation beschränkt auf sich die Steuerung des Velums, das in unterschiedlichem Maße die Nasenhöhle an den Rachenraum angekoppelt.

Vokale: Es besteht keine Ankopplung der Nasenhöhle durch das Velum an den Stimmkanal (z. B. [a]).

nasalierte Vokale: Es liegt nur eine geringe Ankopplung der Nasenhöhle an den Stimmkanal (z. B. [ã]) vor.

nasale Konsonanten: Die Nasenhöhle ist an den Rachenraum angekoppelt. Die Mundhöhle ist ein Stichkanal ([m], [n]).

Pharyngonasaler Konsonant: Die Mundhöhle ist vollständig abgekoppelt. Die Schallausbreitung erfolgt über Rachen- und Nasenhöhle ([ŋ]).

Die Nasenhöhle mit ihren Nebenhöhlen bildet ein kompliziertes akustisches System, das *in vivo* schwer zugänglich ist. Die Pneumatisation (Ausbildung luftgefüllter Volumina) der Nasennebenhöhlen erfolgt in den ersten zehn Lebensjahren. Der Mensch besitzt vier Paare von Nasennebenhöhlen (die Kieferhöhle, die Stirnhöhle, die Keilbeinhöhle und die Siebbeinzellen) mit mehr oder weniger großem akustischem Einfluss.

Erst in jüngerer Zeit sind kernspintomographische Messungen möglich [12], die – im Gegensatz zu den aus Sektionen gewonnenen Querschnittsinformationen – die Beschaffenheit der Schleimhäute realistisch wiedergeben. Die Bestimmung der sogenannten Ostia, der Verbindungskanäle der Nebenhöhlen zur Nasenhöhle, stellt auch mit modernen Untersuchungsverfahren ein Problem dar. Die Nasennebenhöhlen weichen stark von einem elliptischen Querschnitt ab und sind mit einer dicken Schleimhaut ausgekleidet, die etwa 50% des Querschnittes im abgeschwollenen Zustand ausmacht [12]. Die Untersuchungen von Dang und Honda ergaben Asymmetrien der Nasenhöhle und der Nebenhöhlen in Länge und Volumen, das führt dazu, dass die angekoppelten Resonatoren

nicht mehr zusammengefasst werden können und mehrere Pole und Nullstellen hervorrufen. Die akustischen Auswirkungen der Nasenhöhle und Nebenhöhle lassen sich schlecht vereinheitlichen. Feng und Castelli [16] sprechen von „komplexen und simplen Mustern“. Dang und Honda [10] erreichten eine qualitativ gute Übereinstimmung von gemessenen und simulierten Übertragungsfunktionen, indem sie die Nasenhöhle des Modells mit nur vier angekoppelten Helmholtzresonatoren versahen. Studien zur Sprechererkennung [22, 55] ohne zugrunde liegendes Modell des Nasaltraktes belegen, dass nasale Phonation sprecherspezifische Information enthält. Der Nasaltrakt ist in das beschriebene Konzept von Anpassung eines Modells nicht ohne weiteres einzufügen, da sich komplexe Effekte ergeben, nämlich eine individuell unterschiedliche Anzahl zusätzlicher Pol-Nullstellen-Paare, die den anatomischen Größen eines Modells nicht zugeordnet werden können.

A.2. Fossa piriformis

Als Fossa piriformis werden zwei konisch zulaufende Schleimhautausbuchtungen am Epilarynx bezeichnet. Die Seitenarme haben eine ungefähre Länge von 2 cm. Die Rolle der

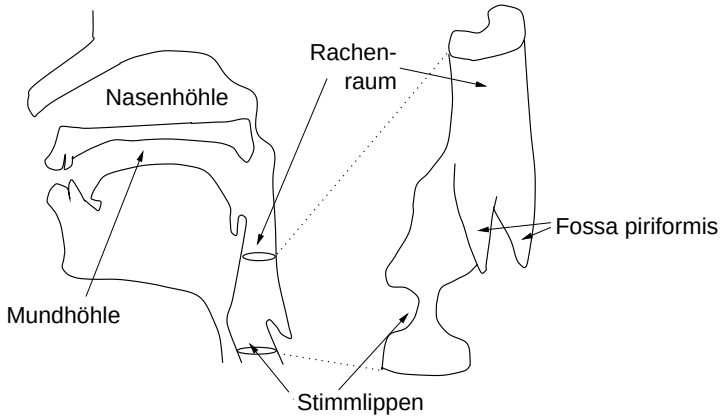


Abbildung A.2.: Die Fossa piriformis nach Dang und Honda [11].

Fossa piriformis wurde in der Literatur bereits verschiedentlich diskutiert: In Simulationen stellte Fant [13] eine Antiresonanz bei 5 kHz sowie eine leichte Verschiebung der Formantfrequenzen unter 3,5 kHz fest. Neueren Untersuchungen von Dang und Honda [11] liegen kernspintomographische Aufnahmen zugrunde. Experimente dieser Autoren an plastischen Modellen sowie Simulationen lassen den Schluss zu, dass der globale Effekt der Fossa piriformis unterschätzt wurde. Die Auswirkungen auf das in Kapitel 3 vorgestellte Verfahren lassen sich wie folgt zusammenfassen: Die Fossa piriformis hat auf Grund ihrer kleinen Dimensionen eine Antiresonanz in der Region von 4–5 kHz. Bei tiefen Frequenzen wirkt sie in etwa wie ein Zusatzvolumen und muss nicht extra als Seitenarm modelliert werden. Auf eine Schätzung der Fossa piriformis wurde bei dem in dieser Arbeit entwickelten Verfahren verzichtet.

B. Simulationsergebnisse

Die Ergebnisse der Simulationen mit verschiedenen Störungsbasen werden in diesem Anhang in Tabellenform und als Abbildung zusammengestellt.

Sprecher	Vokal	$\hat{\omega}_{\mathrm{n/t}}$	$\hat{\omega}_{\mathrm{e/t}}$	$\tilde{L}_{\mathrm{n/t}}$	$\tilde{L}_{\mathrm{e/t}}$	$\tilde{L}_{\mathrm{pz/t}}$
1	/aː/	0,375	0,090	0,030	0,014	−0,009
1	/eː/	0,330	0,043	−0,008	−0,037	−0,034
1	/iː/	1,269	0,138	0,004	−0,202	−0,283
1	/oː/	0,604	0,115	−0,025	0,032	0,012
1	/uː/	0,393	0,057	−0,042	0,053	0,063
2	/aː/	0,495	0,137	−0,027	−0,053	−0,075
2	/eː/	0,299	0,042	−0,051	−0,027	−0,010
2	/iː/	1,386	0,262	−0,028	−0,103	−0,178
2	/oː/	0,614	0,089	−0,051	0,067	0,055
2	/uː/	0,727	0,191	−0,055	0,060	0,054
3	/aː/	0,233	0,037	0,089	−0,014	−0,005
3	/eː/	0,375	0,095	0,081	−0,065	−0,066
3	/iː/	1,197	0,161	0,096	−0,117	−0,200
3	/oː/	0,451	0,117	0,125	0,111	0,094
3	/uː/	0,700	0,050	0,136	−0,025	−0,070
4	/aː/	0,537	0,062	−0,119	0,063	0,071
4	/eː/	0,429	0,063	−0,094	0,026	0,041
4	/iː/	1,331	0,296	−0,081	−0,045	−0,115
4	/oː/	1,106	0,151	−0,076	0,218	0,193
4	/uː/	0,505	0,094	−0,065	0,121	0,143
5	/aː/	0,595	0,040	−0,122	0,029	0,032
5	/eː/	0,487	0,082	−0,092	−0,009	−0,002
5	/iː/	1,532	0,320	−0,100	−0,138	−0,219
5	/oː/	0,663	0,062	−0,098	0,073	0,067
5	/uː/	0,960	0,268	−0,065	0,028	−0,001
6	/aː/	0,636	0,127	−0,111	0,000	−0,015
6	/eː/	0,322	0,041	−0,110	−0,076	−0,069
6	/iː/	1,228	0,282	−0,152	−0,162	−0,209
6	/oː/	0,760	0,120	−0,135	0,060	0,048
6	/uː/	0,589	0,100	−0,163	0,055	0,080

Tabelle B.1.: Schätzung der Querschnittsfunktion des Testsprechers aus der homogenen Querschnittsfunktion. Die Abweichungen werden durch nichtlineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis (4.4.1) nachgebildet.

Sprecher	Vokal	$\hat{\omega}_{\mathrm{n/t}}$	$\hat{\omega}_{\mathrm{e/t}}$	$\tilde{L}_{\mathrm{n/t}}$	$\tilde{L}_{\mathrm{e/t}}$	$\tilde{L}_{\mathrm{pz/t}}$
1	/aː/	0,096	0,033	0,030	0,058	−0,009
1	/eː/	0,166	0,054	−0,008	0,033	−0,034
1	/iː/	0,218	0,297	0,004	−0,001	−0,283
1	/oː/	0,164	0,043	−0,025	−0,017	0,012
1	/uː/	0,354	0,068	−0,042	0,002	0,063
2	/aː/	0,131	0,022	−0,027	−0,020	−0,075
2	/eː/	0,230	0,024	−0,051	0,031	−0,010
2	/iː/	0,173	0,044	−0,028	0,048	−0,178
2	/oː/	0,211	0,013	−0,051	0,018	0,055
2	/uː/	0,316	0,081	−0,055	0,034	0,054
3	/aː/	0,246	0,106	0,089	0,027	−0,005
3	/eː/	0,139	0,030	0,081	0,019	−0,066
3	/iː/	0,114	0,038	0,096	0,055	−0,200
3	/oː/	0,158	0,090	0,125	0,066	0,094
3	/uː/	0,350	0,027	0,136	0,057	−0,070
4	/aː/	0,530	0,043	−0,119	0,063	0,071
4	/eː/	0,472	0,034	−0,094	0,055	0,041
4	/iː/	0,469	0,067	−0,081	0,106	−0,115
4	/oː/	0,693	0,039	−0,076	0,097	0,193
4	/uː/	0,482	0,030	−0,065	0,072	0,143
5	/aː/	0,603	0,105	−0,122	0,016	0,032
5	/eː/	0,380	0,056	−0,092	0,041	−0,002
5	/iː/	0,270	0,091	−0,100	0,008	−0,219
5	/oː/	0,475	0,111	−0,098	0,003	0,067
5	/uː/	0,438	0,155	−0,065	−0,013	−0,001
6	/aː/	0,386	0,029	−0,111	0,009	−0,015
6	/eː/	0,264	0,032	−0,110	−0,030	−0,069
6	/iː/	0,399	0,125	−0,152	−0,025	−0,209
6	/oː/	0,419	0,055	−0,135	0,006	0,048
6	/uː/	0,601	0,036	−0,163	0,007	0,080

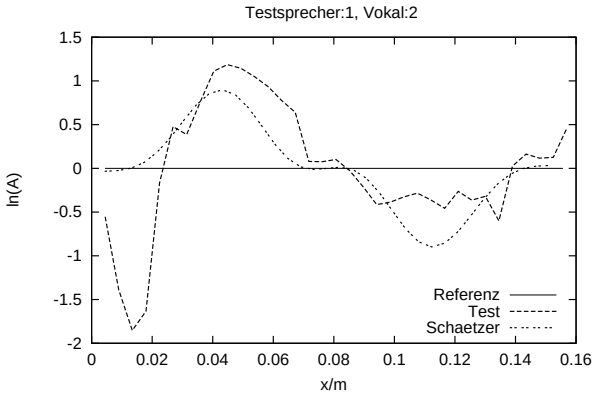
Tabelle B.2.: Schätzung der Querschnittsfunktion des Testsprechers aus der Querschnittsfunktion des Referenzsprechers. Die Abweichungen werden durch nichtlineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis (4.4.2) nachgebildet.

Sprecher	Vokal	$\hat{\omega}_{\mathrm{n/t}}$	$\hat{\omega}_{\mathrm{e/t}}$	$\tilde{L}_{\mathrm{n/t}}$	$\tilde{L}_{\mathrm{e/t}}$	$\tilde{L}_{\mathrm{pz/t}}$
1	/aː/	0,096	0,042	0,030	0,059	−0,009
1	/eː/	0,166	0,123	−0,008	0,019	−0,034
1	/iː/	0,218	0,143	0,004	−0,035	−0,283
1	/oː/	0,164	0,088	−0,025	0,023	0,012
1	/uː/	0,354	0,057	−0,042	0,051	0,063
2	/aː/	0,131	0,026	−0,027	−0,011	−0,075
2	/eː/	0,230	0,037	−0,051	0,011	−0,010
2	/iː/	0,173	0,087	−0,028	0,009	−0,178
2	/oː/	0,211	0,039	−0,051	0,029	0,055
2	/uː/	0,316	0,095	−0,055	0,045	0,054
3	/aː/	0,246	0,097	0,089	0,015	−0,005
3	/eː/	0,139	0,053	0,081	0,035	−0,066
3	/iː/	0,114	0,055	0,096	0,066	−0,200
3	/oː/	0,158	0,067	0,125	0,088	0,094
3	/uː/	0,350	0,084	0,136	−0,004	−0,070
4	/aː/	0,530	0,092	−0,119	0,060	0,071
4	/eː/	0,472	0,132	−0,094	0,045	0,041
4	/iː/	0,469	0,180	−0,081	0,013	−0,115
4	/oː/	0,693	0,160	−0,076	0,174	0,193
4	/uː/	0,482	0,110	−0,065	0,093	0,143
5	/aː/	0,603	0,091	−0,122	0,008	0,032
5	/eː/	0,380	0,087	−0,092	0,010	−0,002
5	/iː/	0,270	0,126	−0,100	−0,042	−0,219
5	/oː/	0,475	0,098	−0,098	−0,020	0,067
5	/uː/	0,438	0,153	−0,065	0,005	−0,001
6	/aː/	0,386	0,056	−0,111	0,022	−0,015
6	/eː/	0,264	0,069	−0,110	−0,033	−0,069
6	/iː/	0,399	0,233	−0,152	−0,109	−0,209
6	/oː/	0,419	0,061	−0,135	0,016	0,048
6	/uː/	0,601	0,124	−0,163	0,024	0,080

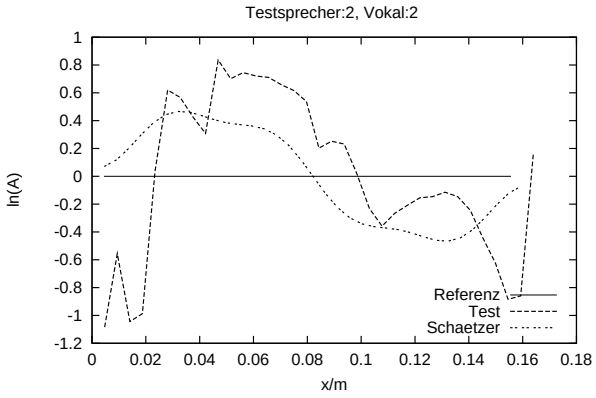
Tabelle B.3.: Schätzung der Querschnittsfunktion des Testsprechers aus der Querschnittsfunktion des Referenzsprechers. Die Abweichungen werden durch nichtlineare Längenskalierung und Querschnittsskalierung mit der Hauptkomponentenbasis (4.4.3) nachgebildet.

Sprecher	Vokal	$\hat{\omega}_{\mathrm{n/t}}$	$\hat{\omega}_{\mathrm{e/t}}$	$\tilde{L}_{\mathrm{n/t}}$	$\tilde{L}_{\mathrm{e/t}}$	$\tilde{L}_{\mathrm{pz/t}}$
1	/aː/	0,096	0,012	0,030	0,049	−0,009
1	/eː/	0,166	0,029	−0,008	0,017	−0,034
1	/iː/	0,218	0,210	0,004	−0,027	−0,283
1	/oː/	0,164	0,014	−0,025	−0,012	0,012
1	/uː/	0,354	0,046	−0,042	0,004	0,063
2	/aː/	0,131	0,016	−0,027	−0,020	−0,075
2	/eː/	0,230	0,013	−0,051	0,015	−0,010
2	/iː/	0,173	0,046	−0,028	0,031	−0,178
2	/oː/	0,211	0,009	−0,051	−0,007	0,055
2	/uː/	0,316	0,085	−0,055	0,010	0,054
3	/aː/	0,246	0,048	0,089	0,046	−0,005
3	/eː/	0,139	0,016	0,081	0,031	−0,066
3	/iː/	0,114	0,032	0,096	0,057	−0,200
3	/oː/	0,158	0,030	0,125	0,097	0,094
3	/uː/	0,350	0,011	0,136	0,057	−0,070
4	/aː/	0,530	0,054	−0,119	0,004	0,071
4	/eː/	0,472	0,062	−0,094	0,026	0,041
4	/iː/	0,469	0,050	−0,081	0,086	−0,115
4	/oː/	0,693	0,048	−0,076	0,048	0,193
4	/uː/	0,482	0,056	−0,065	0,028	0,143
5	/aː/	0,603	0,088	−0,122	−0,027	0,032
5	/eː/	0,380	0,060	−0,092	0,005	−0,002
5	/iː/	0,270	0,094	−0,100	−0,012	−0,219
5	/oː/	0,475	0,107	−0,098	−0,030	0,067
5	/uː/	0,438	0,152	−0,065	−0,015	−0,001
6	/aː/	0,386	0,016	−0,111	−0,025	−0,015
6	/eː/	0,264	0,024	−0,110	−0,048	−0,069
6	/iː/	0,399	0,093	−0,152	−0,061	−0,209
6	/oː/	0,419	0,038	−0,135	−0,051	0,048
6	/uː/	0,601	0,088	−0,163	−0,054	0,080

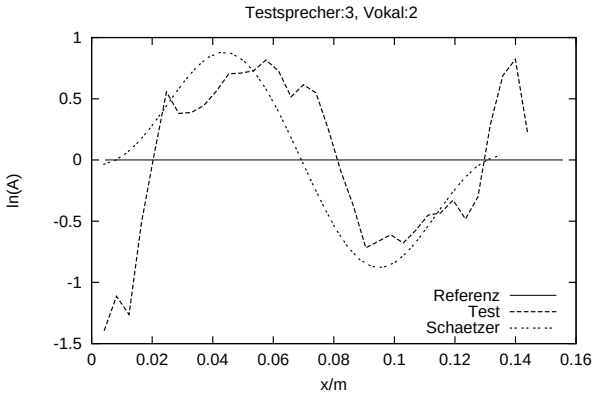
Tabelle B.4.: Schätzung der Querschnittsfunktion des Testsprechers aus der Querschnittsfunktion des Referenzsprechers. Die Abweichungen werden durch nichtlineare Längenskalierung und Querschnittsskalierung mit der aus der Lagrangedichte errechneten Störungsbasis (4.4.4) nachgebildet.



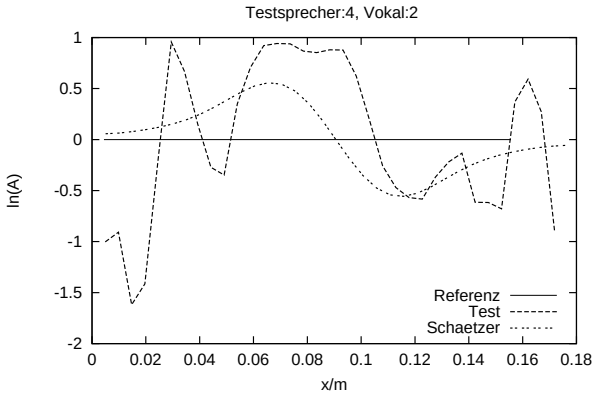
(a) Testsprecher 1



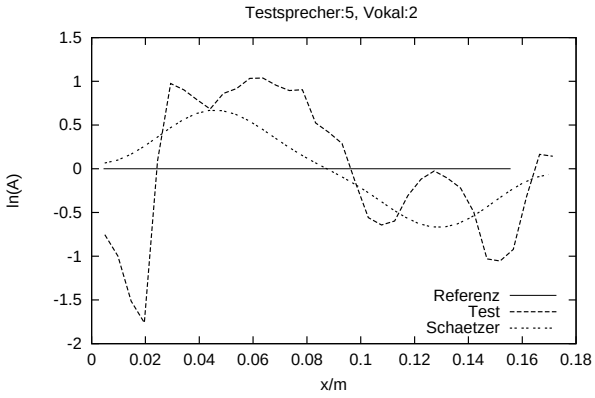
(b) Testsprecher 2



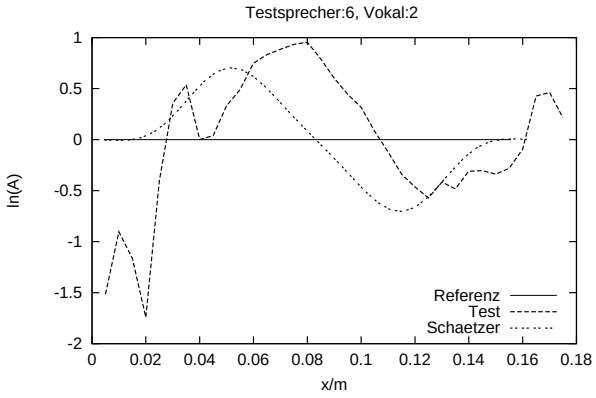
(c) Testsprecher 3



(d) Testsprecher 4

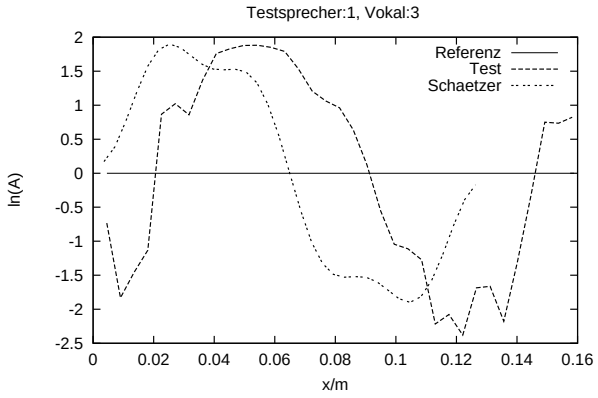


(e) Testsprecher 5

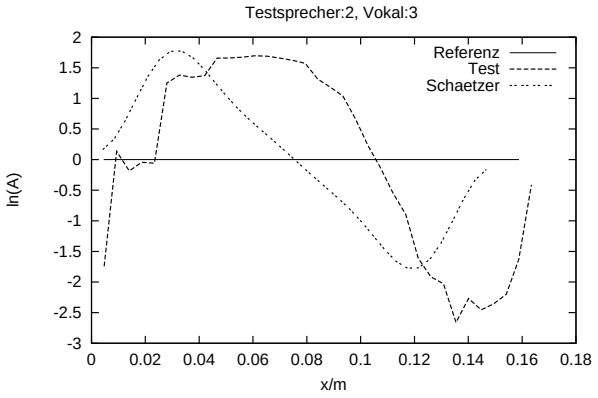


(f) Testsprecher 6

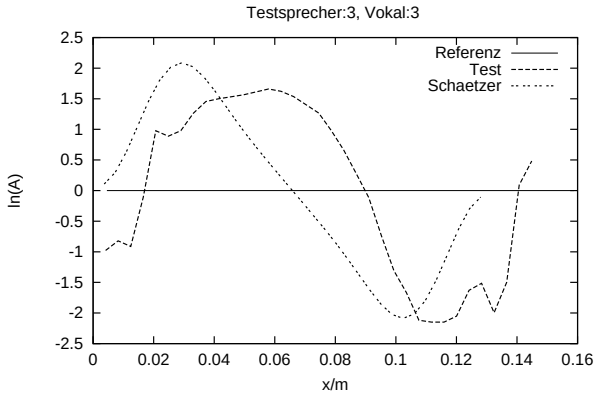
Abbildung B.1.: Schätzung der Querschnittsfunktion (/eː/) des Testsprechers aus der homogenen Querschnittsfunktion. Die Abweichungen werden durch lineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis nachgebildet.



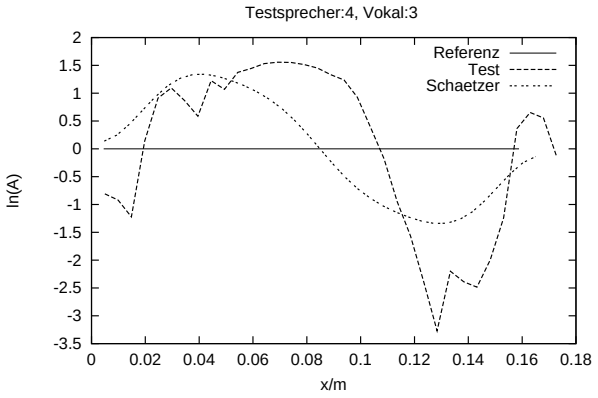
(a) Testsprecher 1



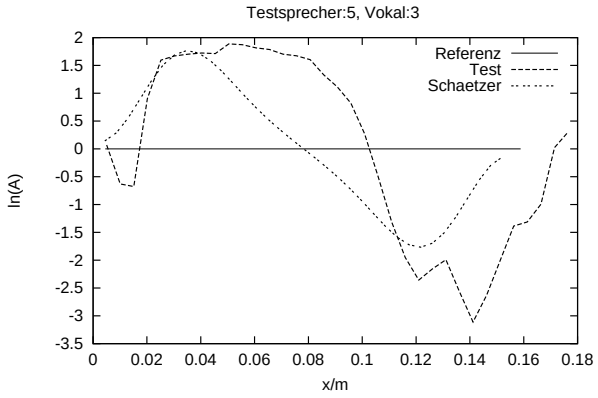
(b) Testsprecher 2



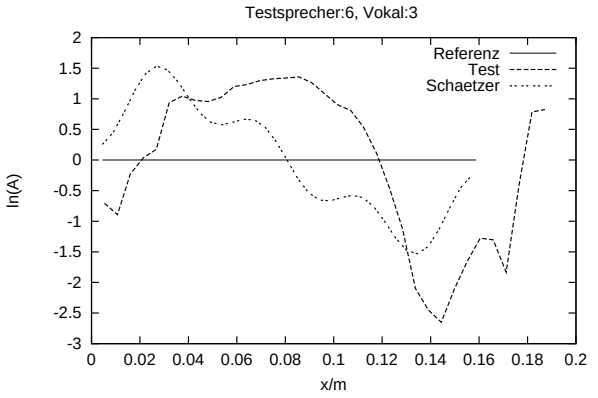
(c) Testsprecher 3



(d) Testsprecher 4

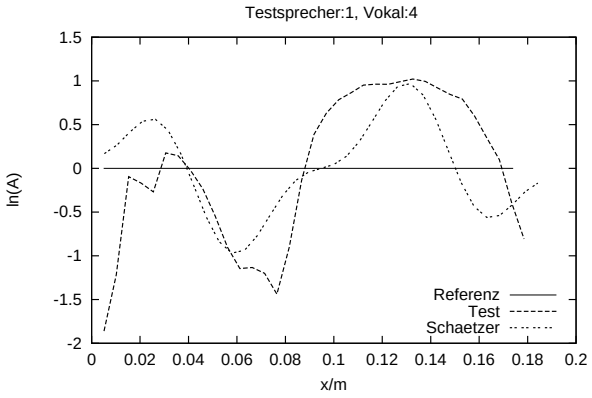


(e) Testsprecher 5

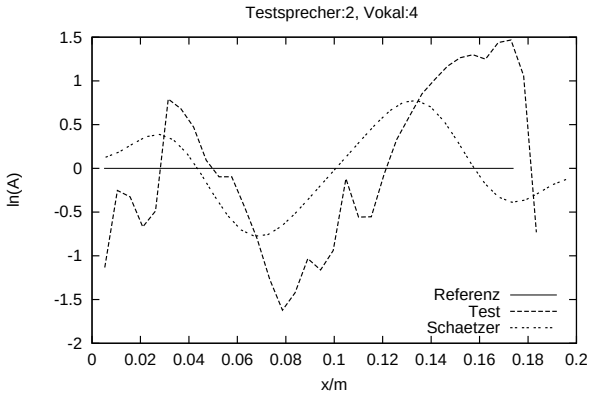


(f) Testsprecher 6

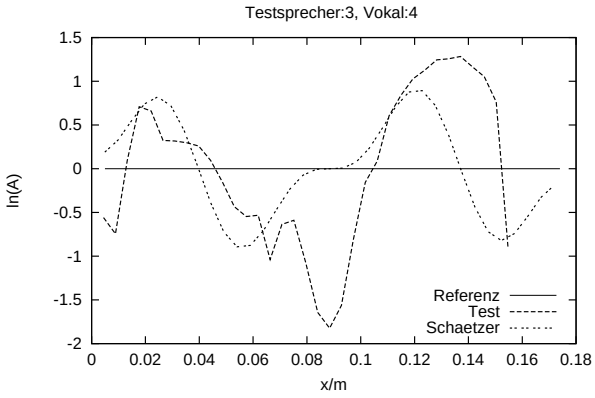
Abbildung B.2.: Schätzung der Querschnittsfunktion (/i:/) des Testsprechers aus der homogenen Querschnittsfunktion. Die Abweichungen werden durch lineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis nachgebildet.



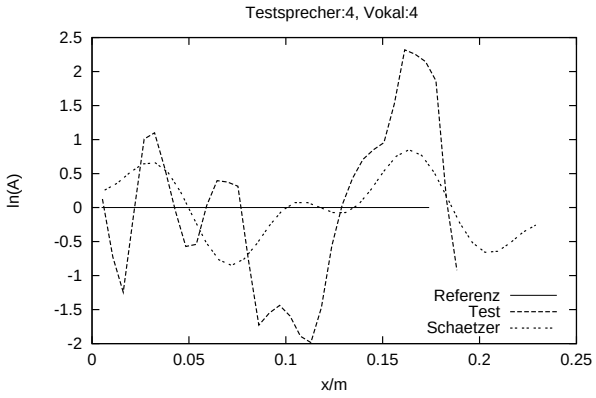
(a) Testsprecher 1



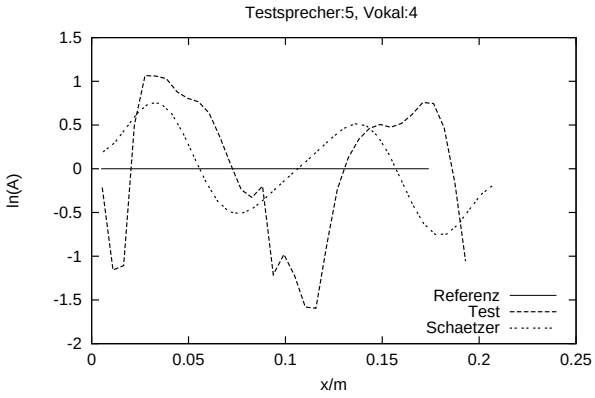
(b) Testsprecher 2



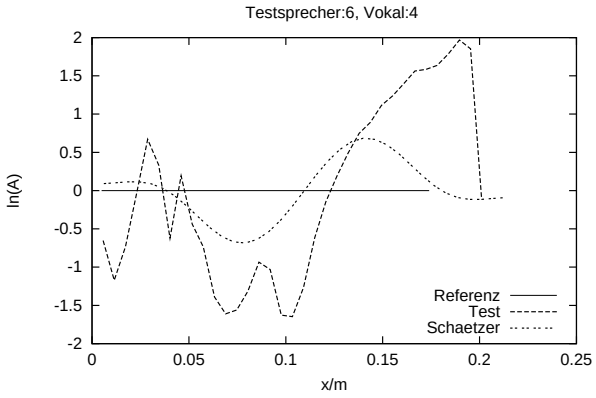
(c) Testsprecher 3



(d) Testsprecher 4

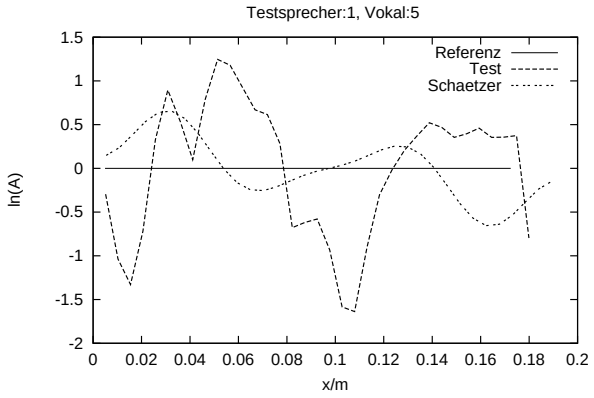


(e) Testsprecher 5

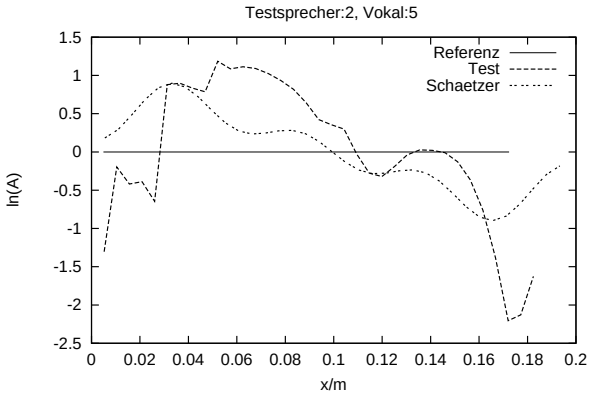


(f) Testsprecher 6

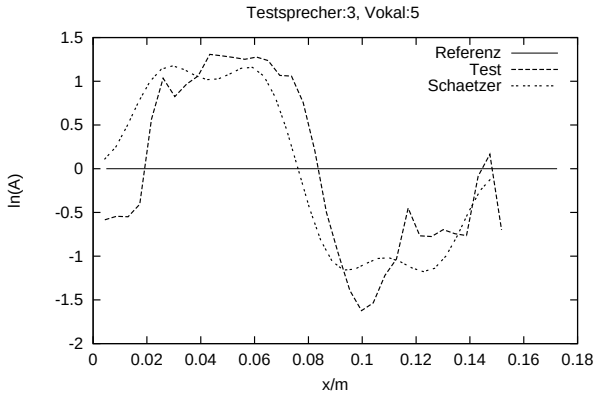
Abbildung B.3.: Schätzung der Querschnittsfunktion (/oː/) des Testsprechers aus der homogenen Querschnittsfunktion. Die Abweichungen werden durch lineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis nachgebildet.



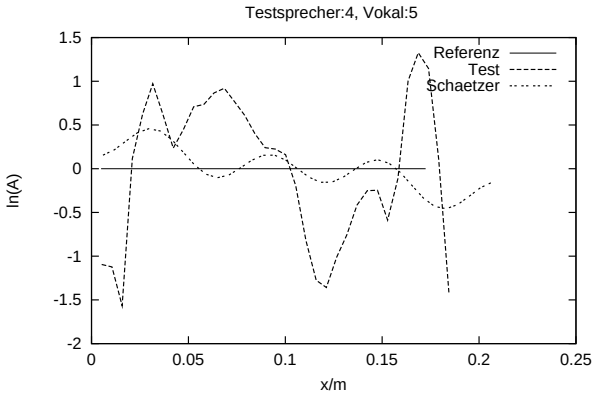
(a) Testsprecher 1



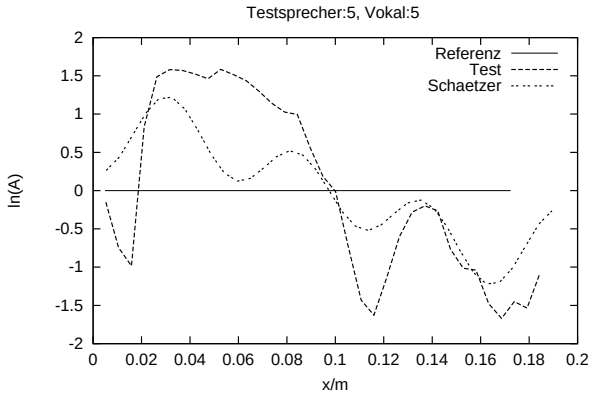
(b) Testsprecher 2



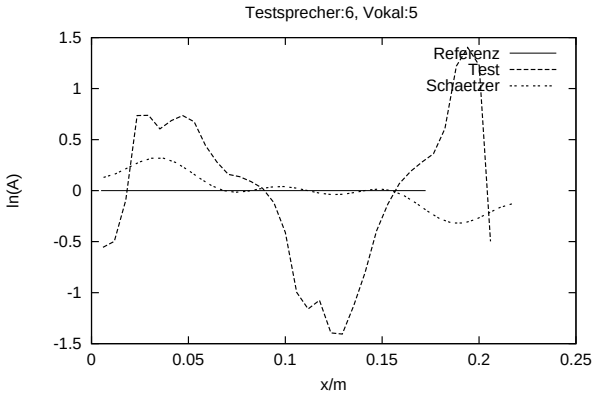
(c) Testsprecher 3



(d) Testsprecher 4

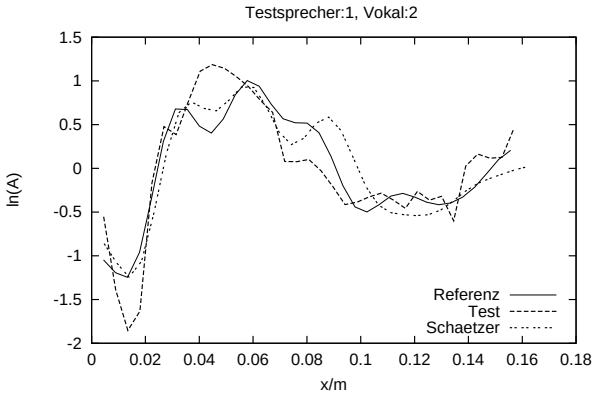


(e) Testsprecher 5

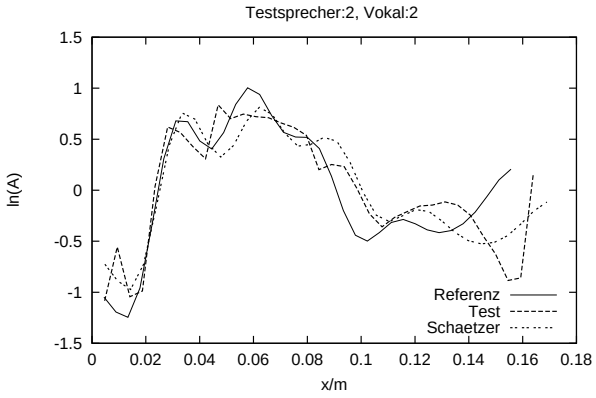


(f) Testsprecher 6

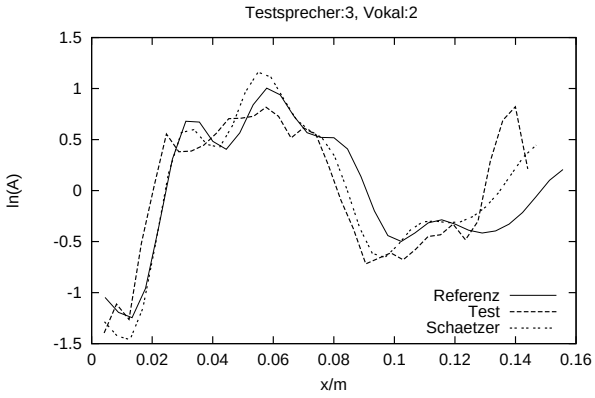
Abbildung B.4.: Schätzung der Querschnittsfunktion (/u:/) des Testsprechers aus der homogenen Querschnittsfunktion. Die Abweichungen werden durch lineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis nachgebildet.



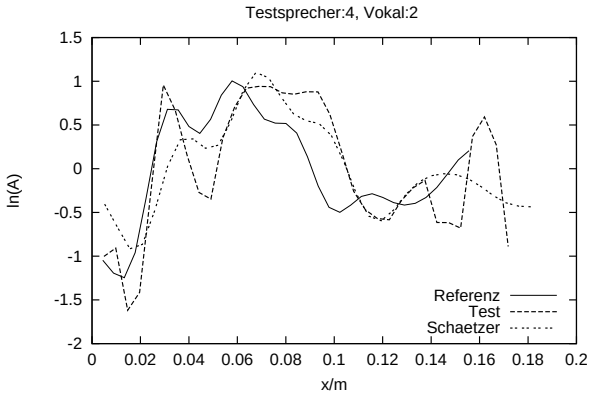
(a) Testsprecher 1



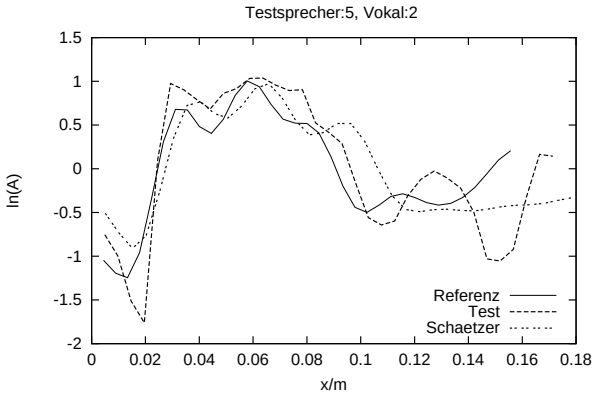
(b) Testsprecher 2



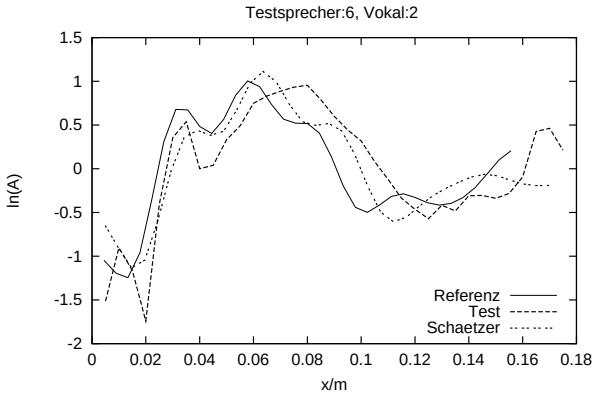
(c) Testsprecher 3



(d) Testsprecher 4

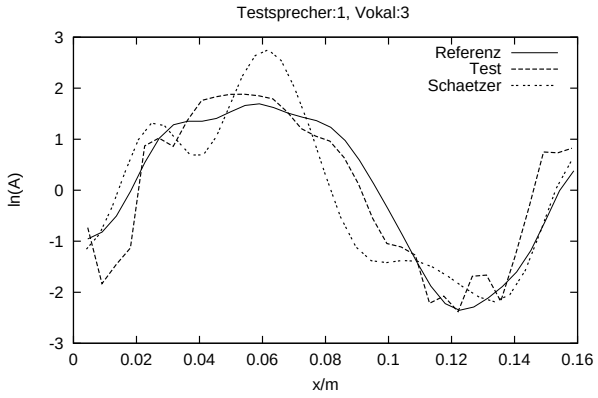


(e) Testsprecher 5

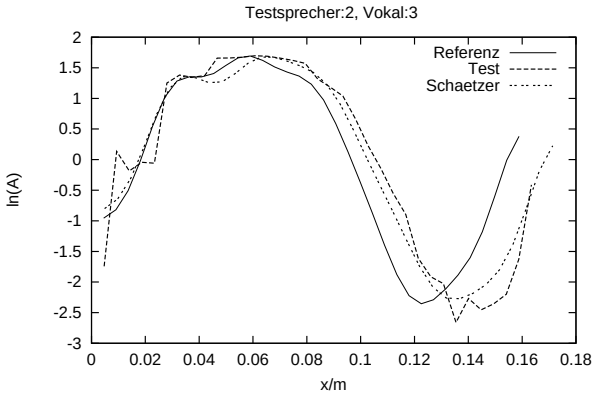


(f) Testsprecher 6

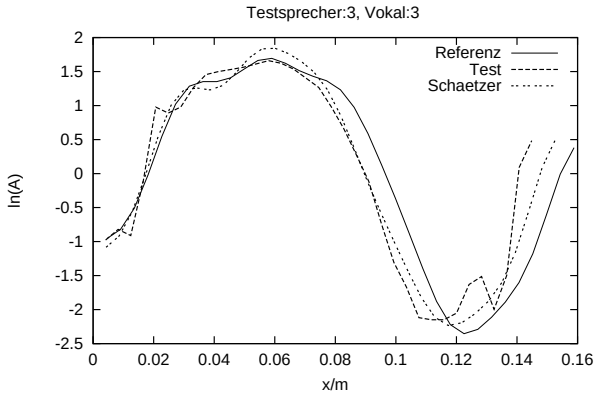
Abbildung B.5.: Schätzung der Querschnittsfunktion (/eː/) des Testsprechers aus der Querschnittsfunktion des Referenzsprechers. Die Abweichungen werden durch nichtlineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis nachgebildet.



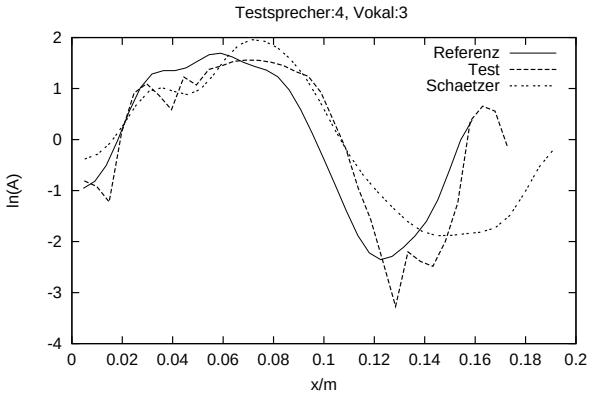
(a) Testsprecher 1



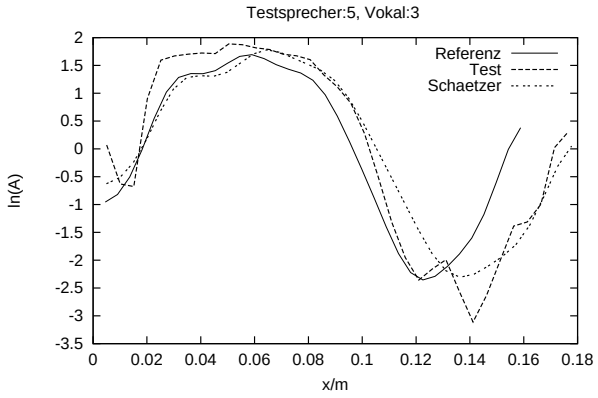
(b) Testsprecher 2



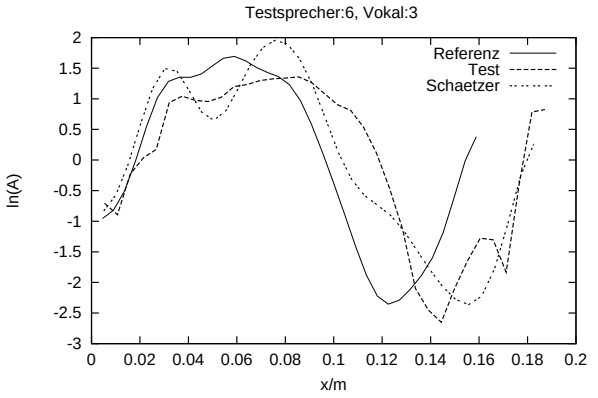
(c) Testsprecher 3



(d) Testsprecher 4

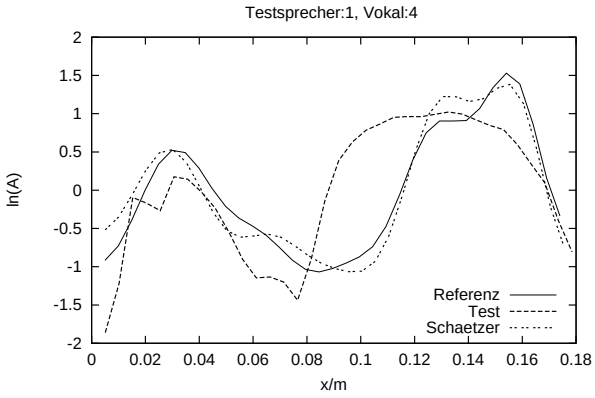


(e) Testsprecher 5

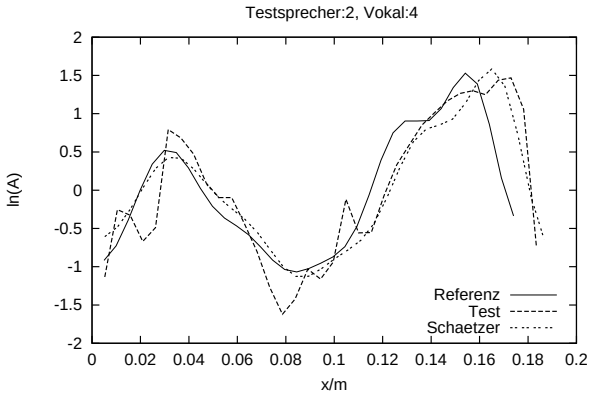


(f) Testsprecher 6

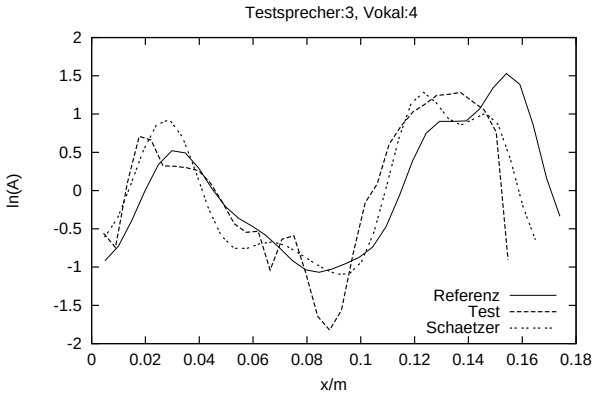
Abbildung B.6.: Schätzung der Querschnittsfunktion (/i:/) des Testsprechers aus der Querschnittsfunktion des Referenzsprechers. Die Abweichungen werden durch nichtlineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis nachgebildet.



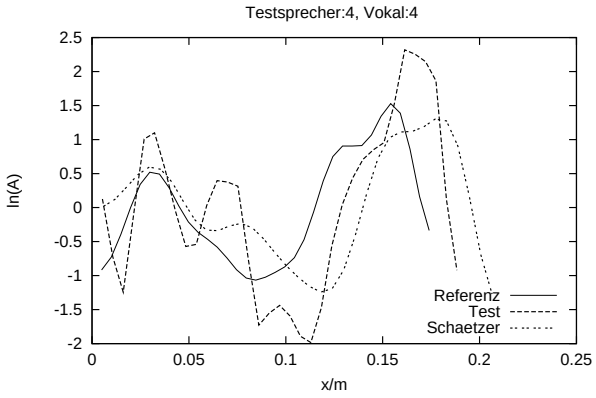
(a) Testsprecher 1



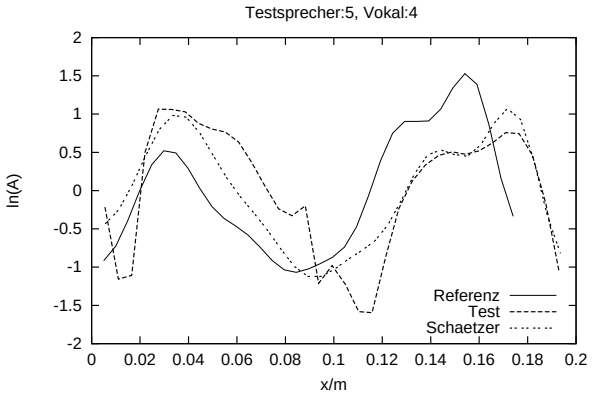
(b) Testsprecher 2



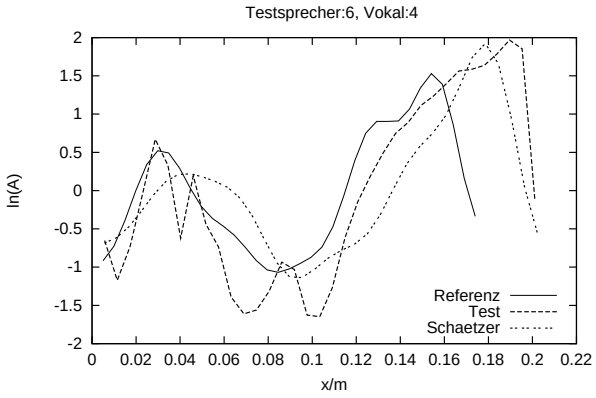
(c) Testsprecher 3



(d) Testsprecher 4

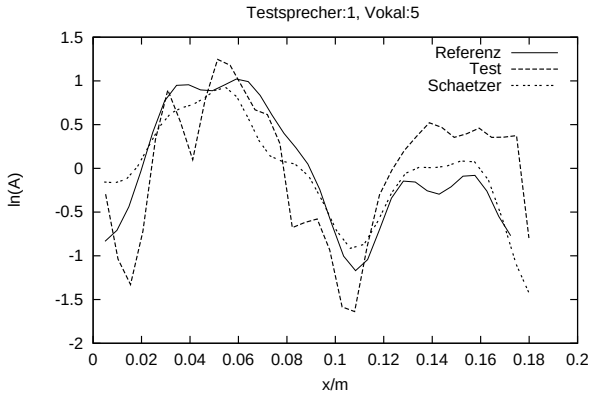


(e) Testsprecher 5

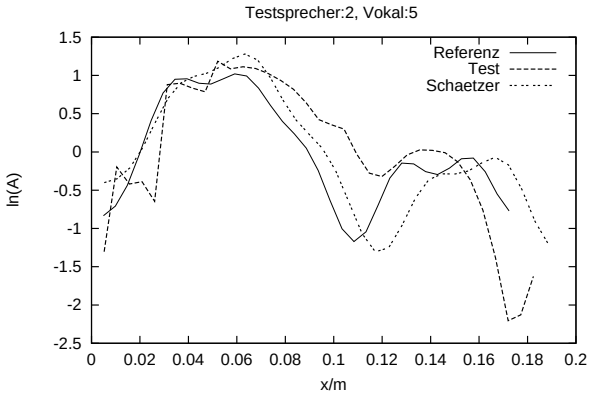


(f) Testsprecher 6

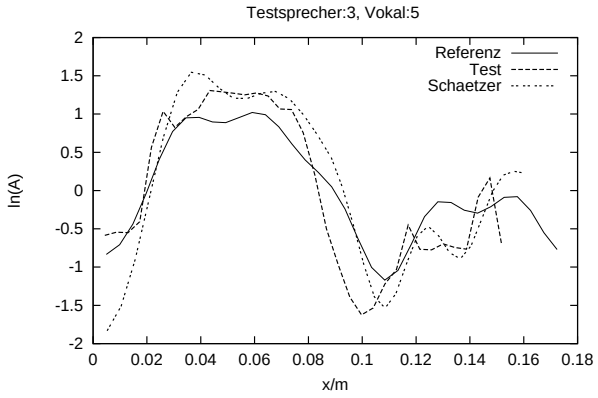
Abbildung B.7.: Schätzung der Querschnittsfunktion (/oː/) des Testsprechers aus der Querschnittsfunktion des Referenzsprechers. Die Abweichungen werden durch nichtlineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis nachgebildet.



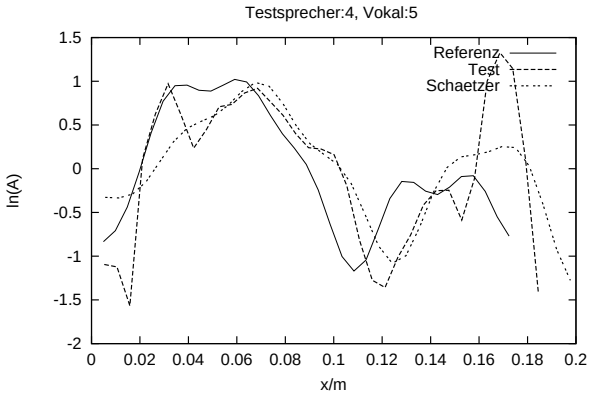
(a) Testsprecher 1



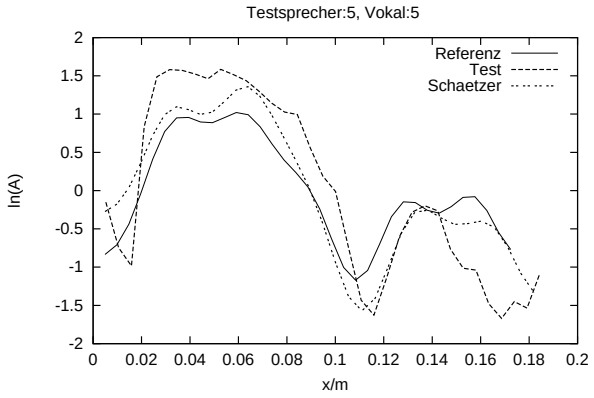
(b) Testsprecher 2



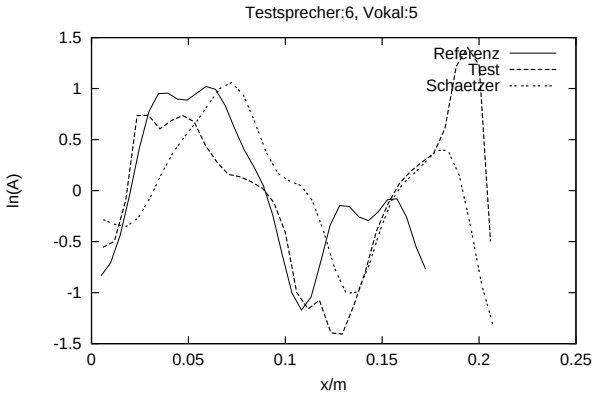
(c) Testsprecher 3



(d) Testsprecher 4

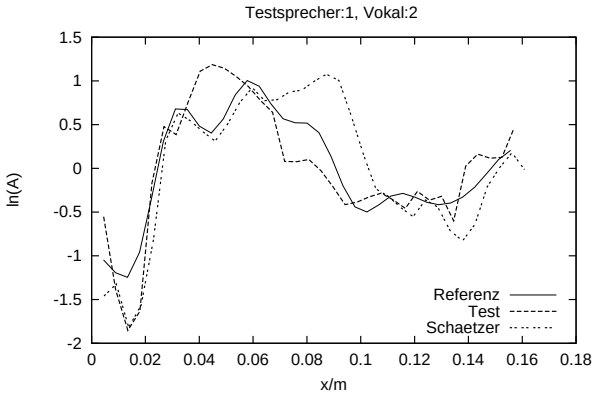


(e) Testsprecher 5

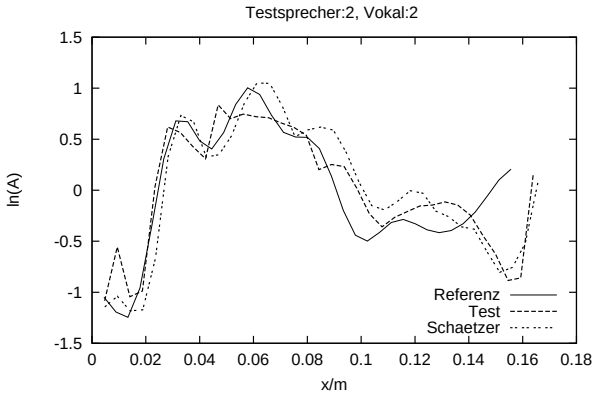


(f) Testsprecher 6

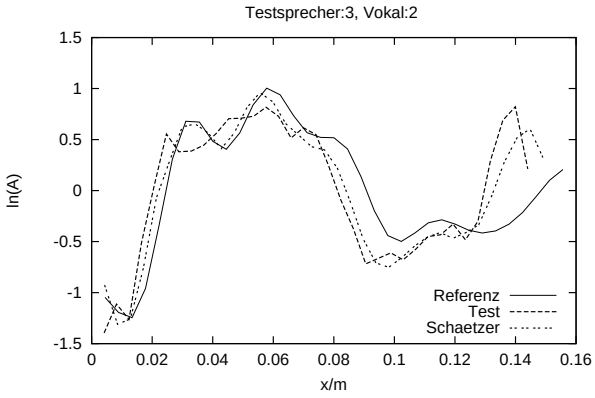
Abbildung B.8.: Schätzung der Querschnittsfunktion (/u:/) des Testsprechers aus der Querschnittsfunktion des Referenzsprechers. Die Abweichungen werden durch nichtlineare Längenskalierung und Querschnittsskalierung mit der Kosinus-Störungsbasis nachgebildet.



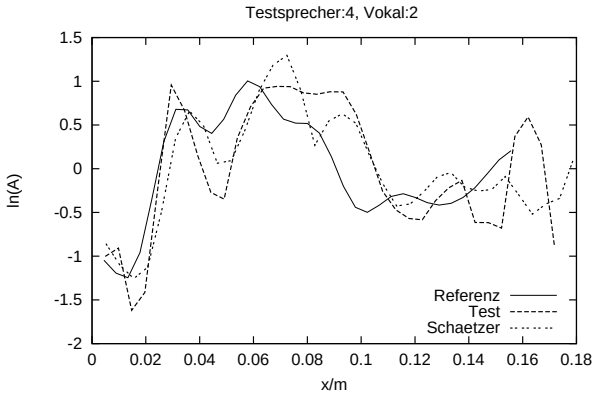
(a) Testsprecher 1



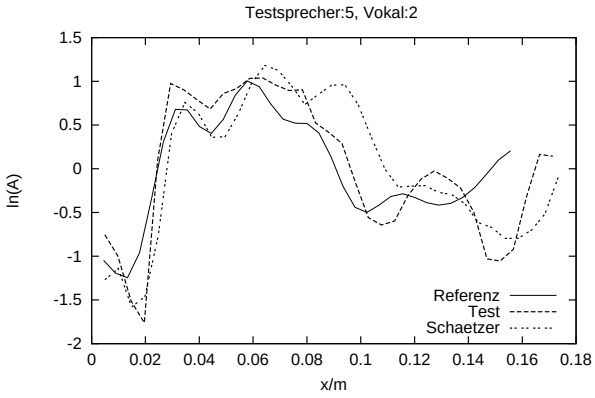
(b) Testsprecher 2



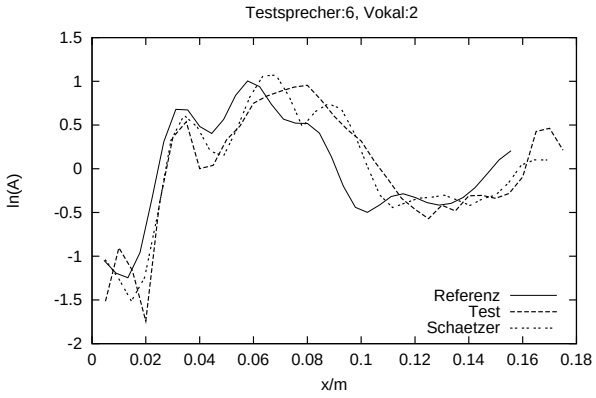
(c) Testsprecher 3



(d) Testsprecher 4

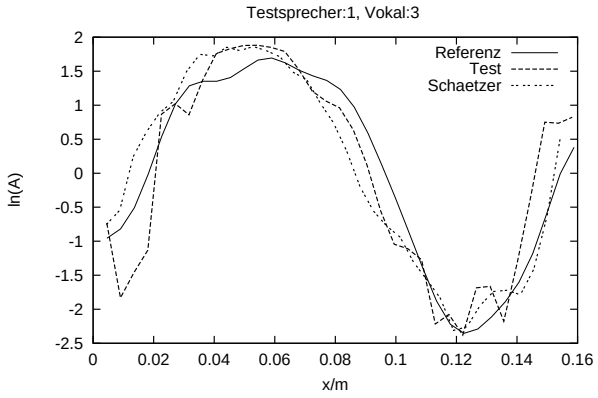


(e) Testsprecher 5

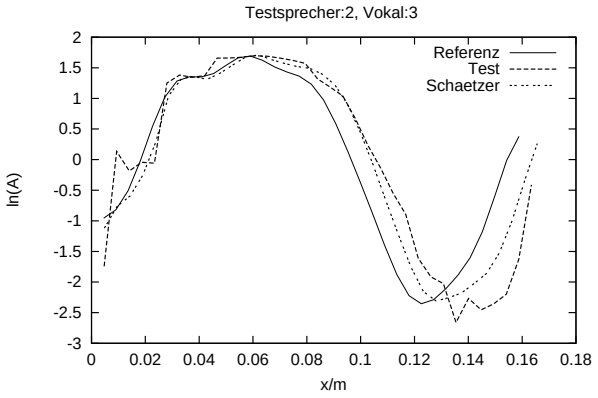


(f) Testsprecher 6

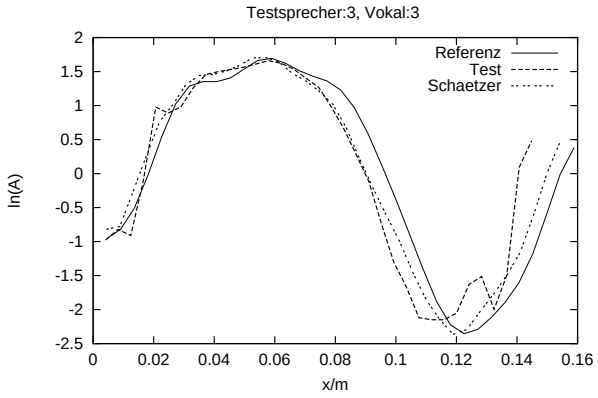
Abbildung B.9.: /eː/ :Anpassung des Referenzsprechers an den Testsprecher durch unterschiedliche Längenskalierung beider Vokaltrakthälften. Die Querschnittsskalierung wird mit einer aus der Hauptkomponentenanalyse der Querschnittsfunktionen gewonnenen Basis durchgeführt.



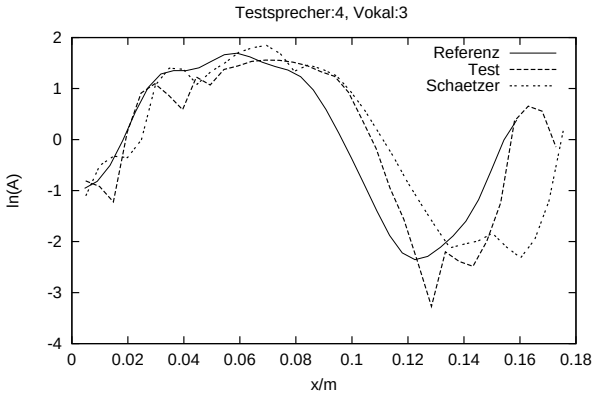
(a) Testsprecher 1



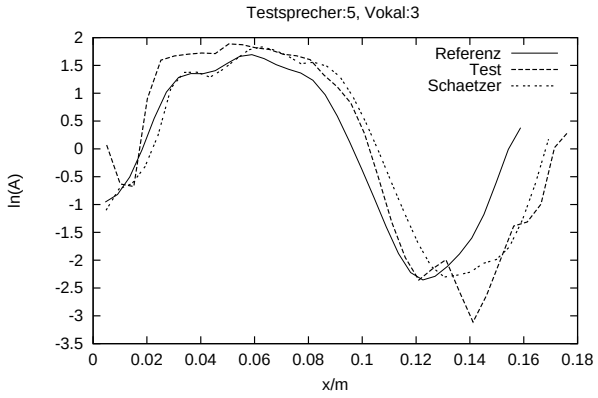
(b) Testsprecher 2



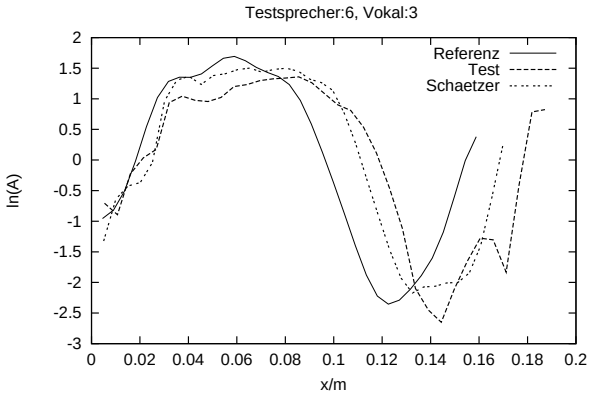
(c) Testsprecher 3



(d) Testsprecher 4

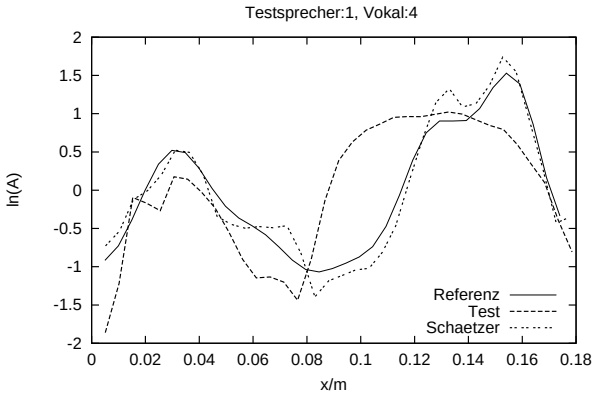


(e) Testsprecher 5

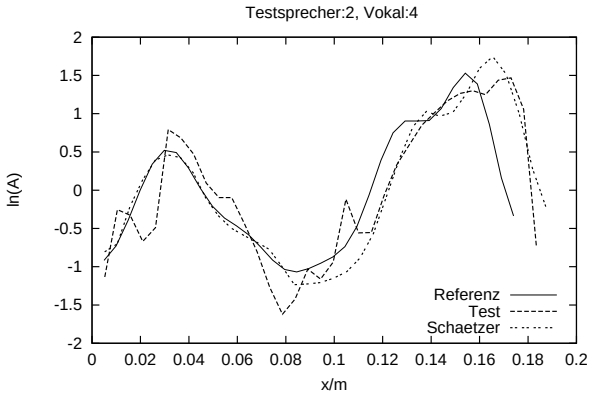


(f) Testsprecher 6

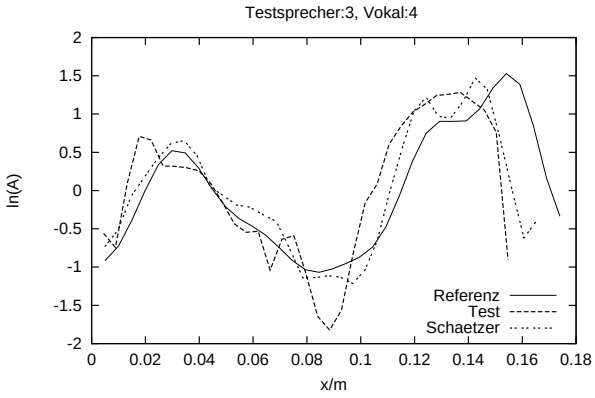
Abbildung B.10.: /iː/: Anpassung des Referenzsprechers an den Testsprecher durch unterschiedliche Längenskalierung beider Vokaltrakthälften. Die Querschnittsskalierung wird mit einer aus der Hauptkomponentenanalyse der Querschnittsfunktionen gewonnenen Basis durchgeführt.



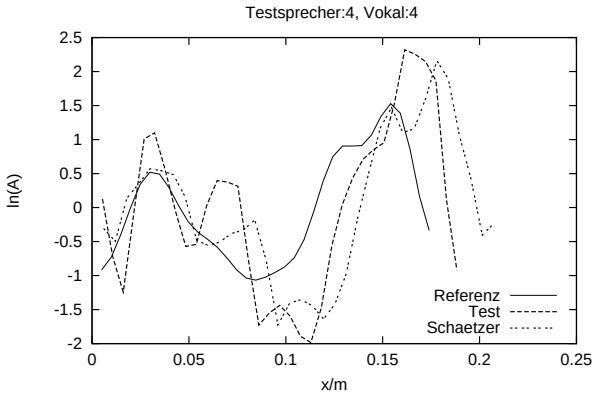
(a) Testsprecher 1



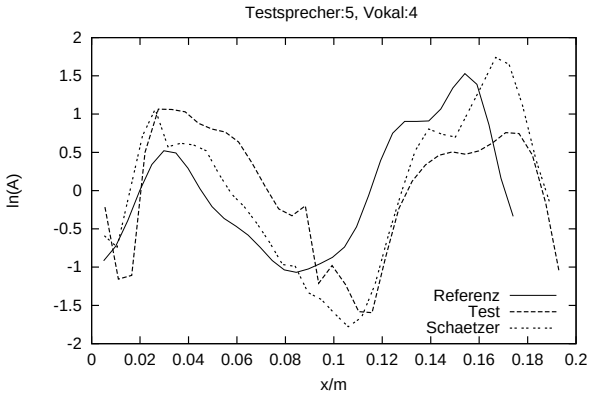
(b) Testsprecher 2



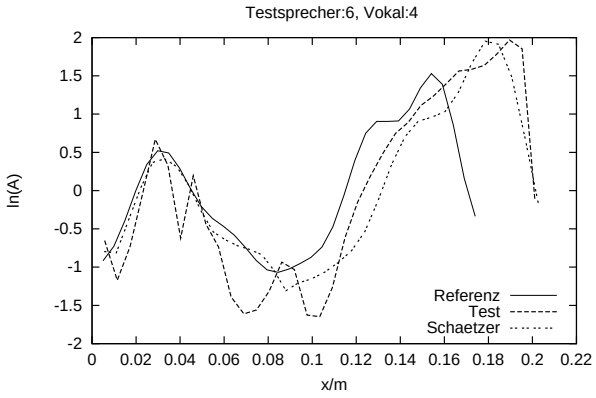
(c) Testsprecher 3



(d) Testsprecher 4

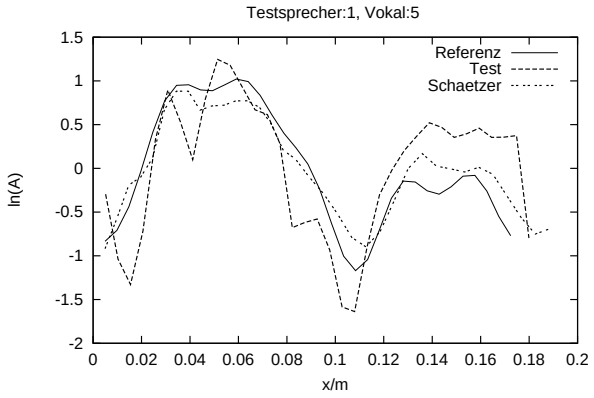


(e) Testsprecher 5

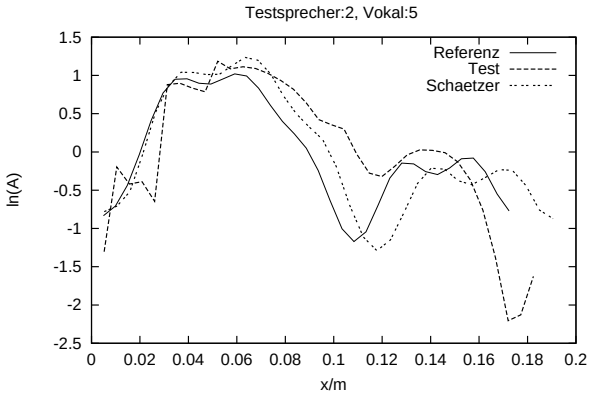


(f) Testsprecher 6

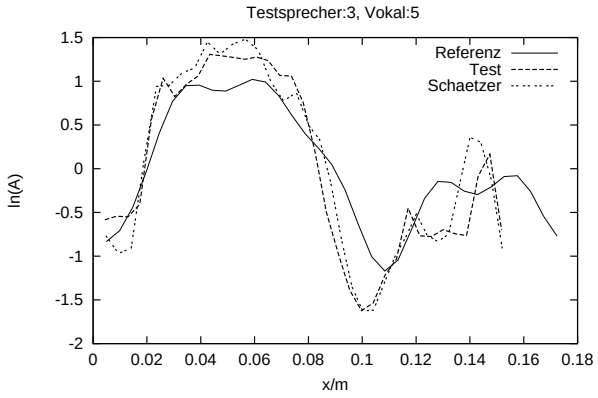
Abbildung B.11.: /oː/: Anpassung des Referenzsprechers an den Testsprecher durch unterschiedliche Längenskalierung beider Vokaltrakthälften. Die Querschnittsskalierung wird mit einer aus der Hauptkomponentenanalyse der Querschnittsfunktionen gewonnenen Basis durchgeführt.



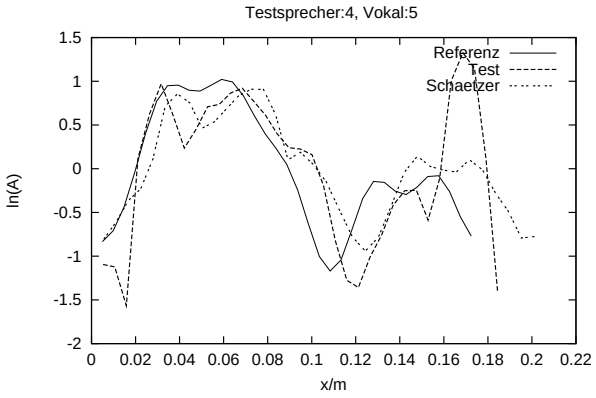
(a) Testsprecher 1



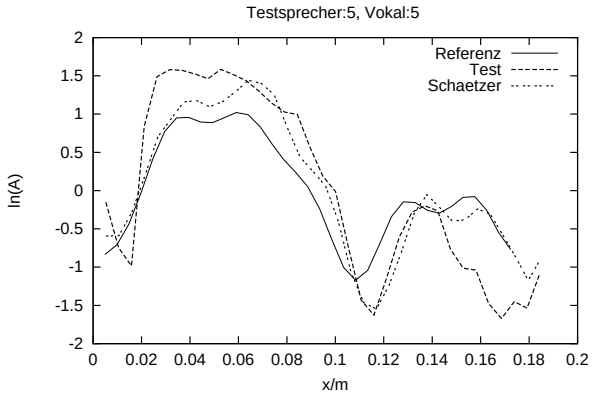
(b) Testsprecher 2



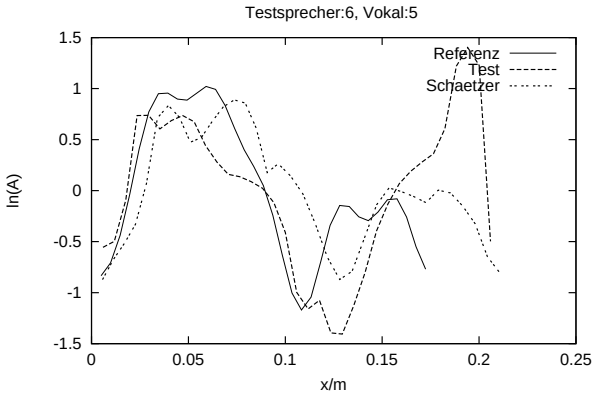
(c) Testsprecher 3



(d) Testsprecher 4

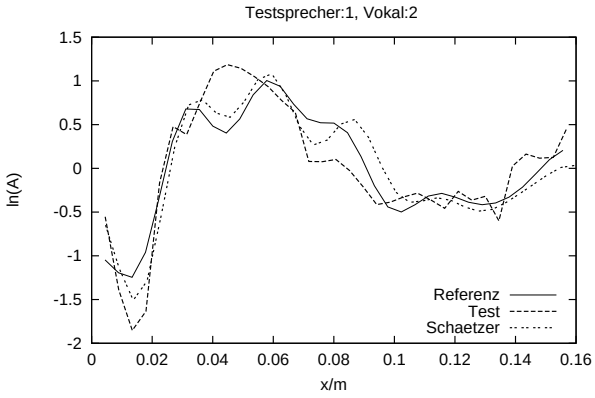


(e) Testsprecher 5

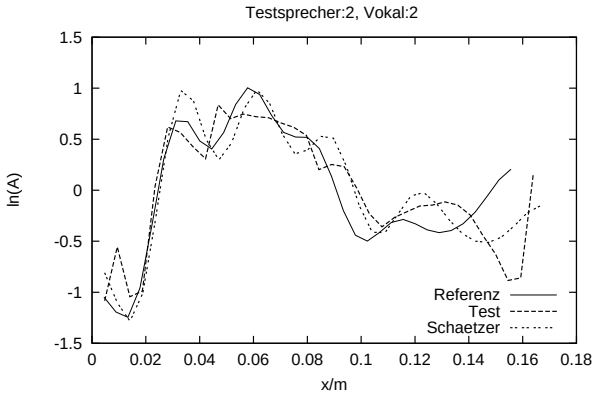


(f) Testsprecher 6

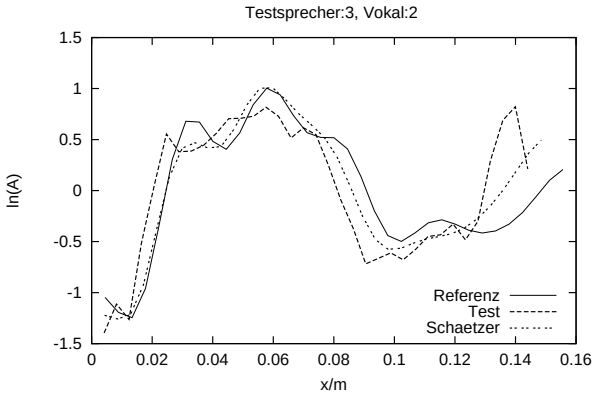
Abbildung B.12.: /uː/: Anpassung des Referenzsprechers an den Testsprecher durch unterschiedliche Längenskalierung beider Vokaltrakthälften. Die Querschnittsskalierung wird mit einer aus der Hauptkomponentenanalyse der Querschnittsfunktionen gewonnenen Basis durchgeführt.



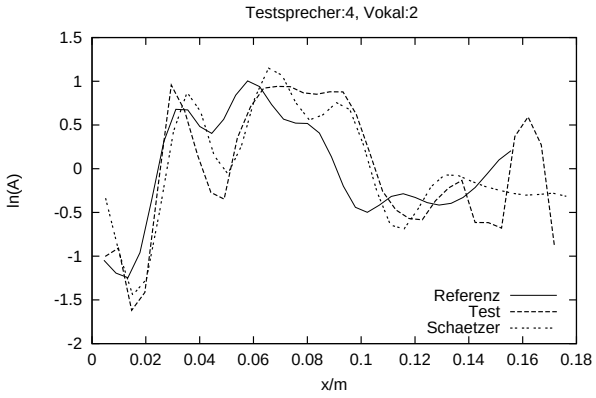
(a) Testsprecher 1



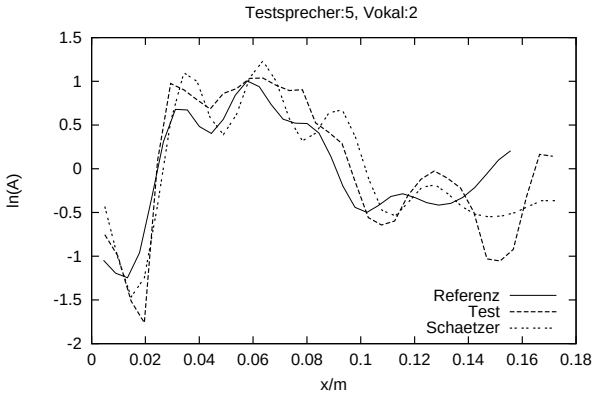
(b) Testsprecher 2



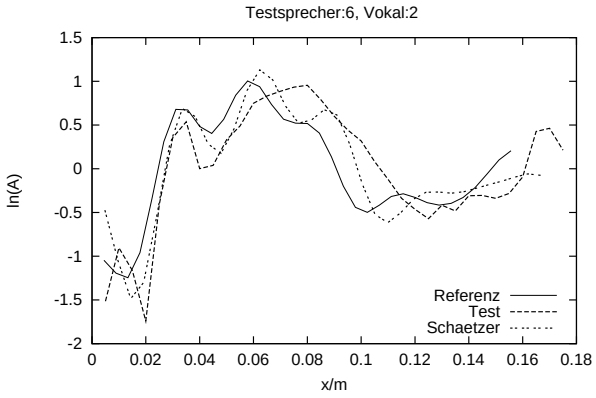
(c) Testsprecher 3



(d) Testsprecher 4

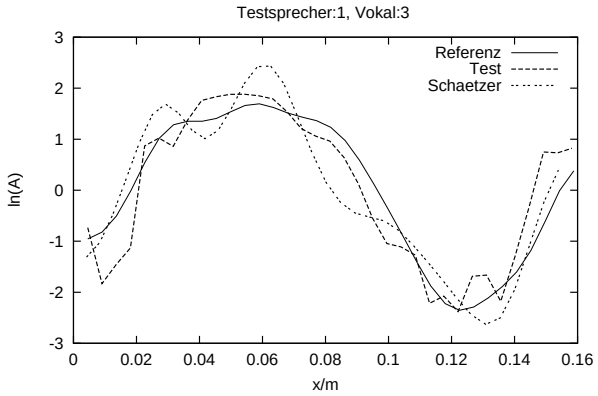


(e) Testsprecher 5

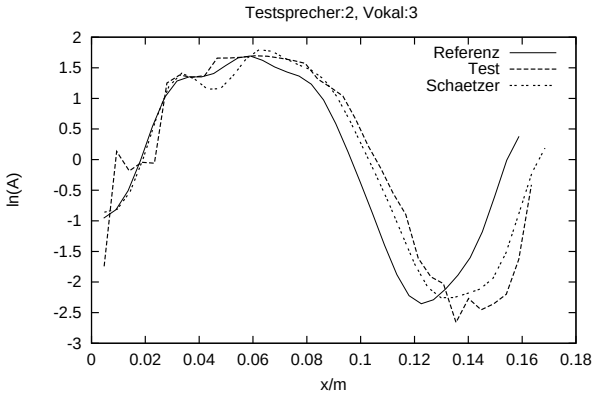


(f) Testsprecher 6

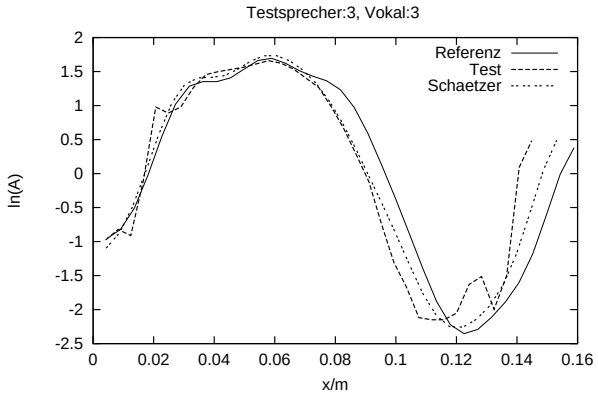
Abbildung B.13.: /eː/: Anpassung des Referenzsprechers an den Testsprecher durch unterschiedliche Längenskalierung beider Vokaltrakthälften und Querschnittsskalierung mit einer aus der Lagrangedichtefunktion abgeleiteten Störungsbasis.



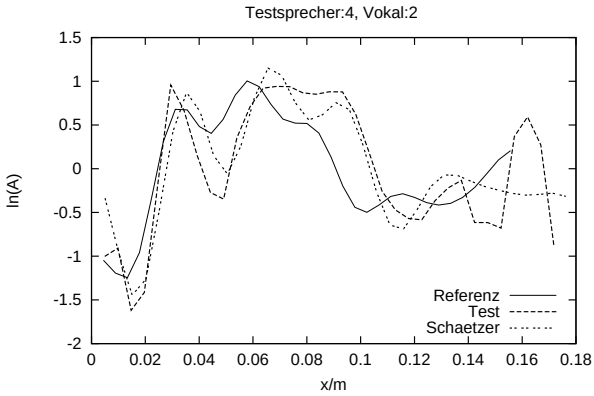
(a) Testsprecher 1



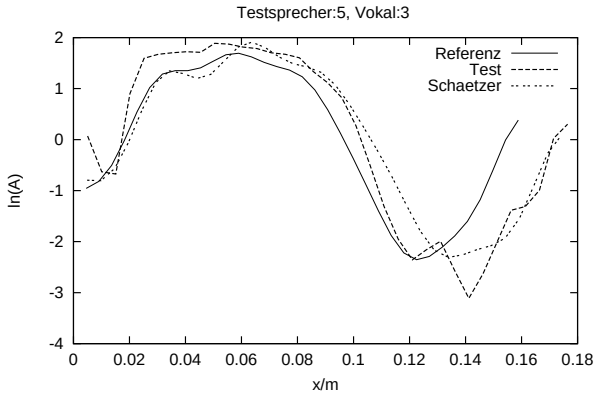
(b) Testsprecher 2



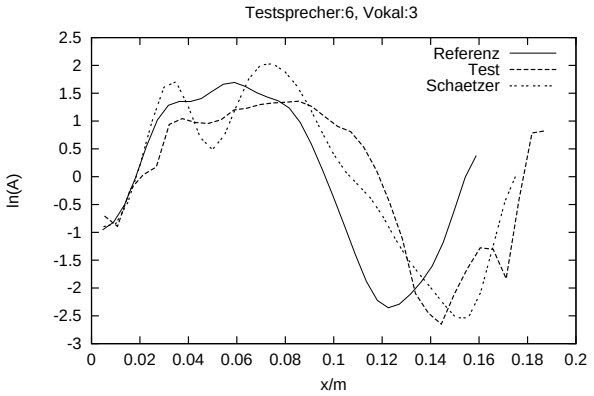
(c) Testsprecher 3



(d) Testsprecher 4

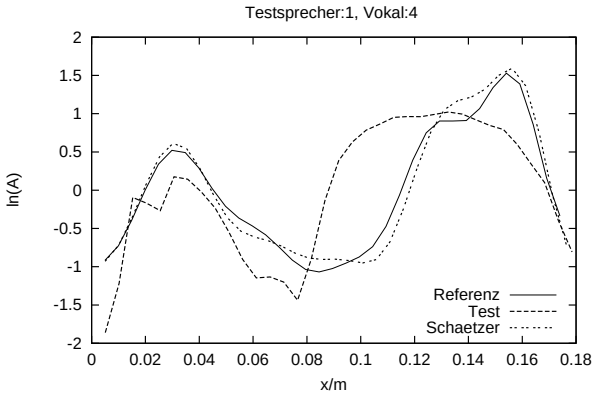


(e) Testsprecher 5

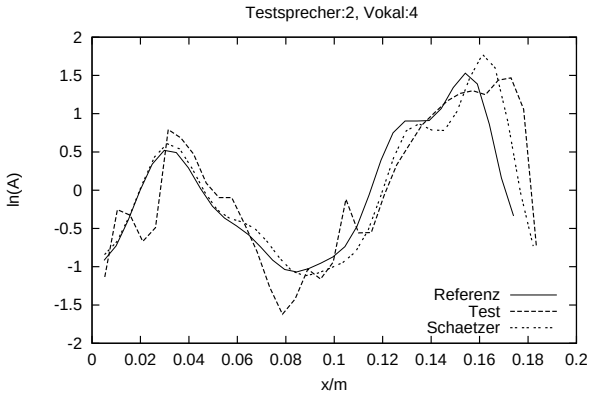


(f) Testsprecher 6

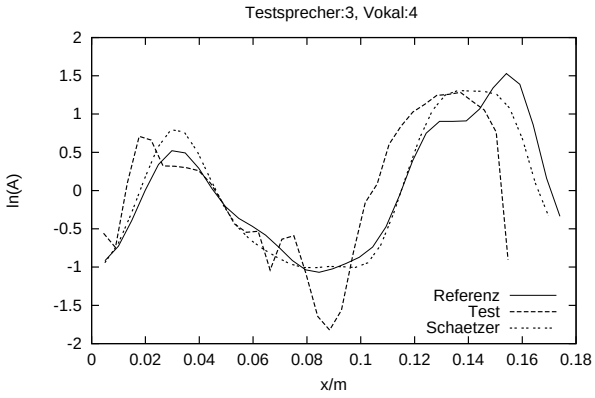
Abbildung B.14.: /iː/: Anpassung des Referenzsprechers an den Testsprecher durch unterschiedliche Längenskalierung beider Vokaltrakthälften und Querschnittsskalierung mit einer aus der Lagrangedichtefunktion abgeleiteten Störungsbasis.



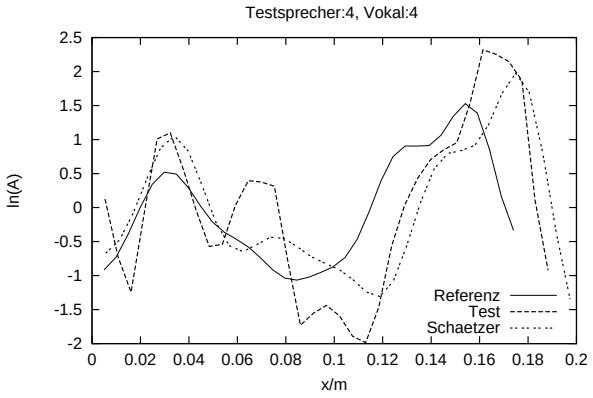
(a) Testsprecher 1



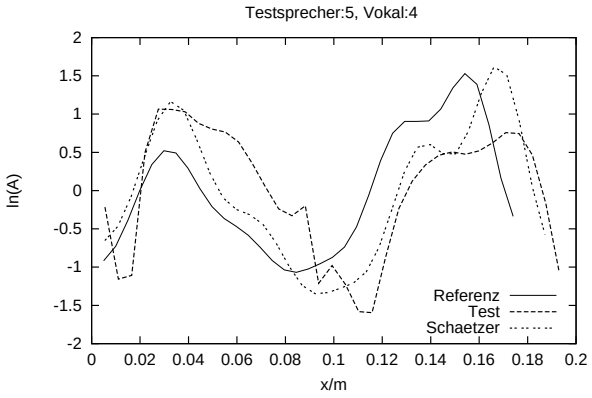
(b) Testsprecher 2



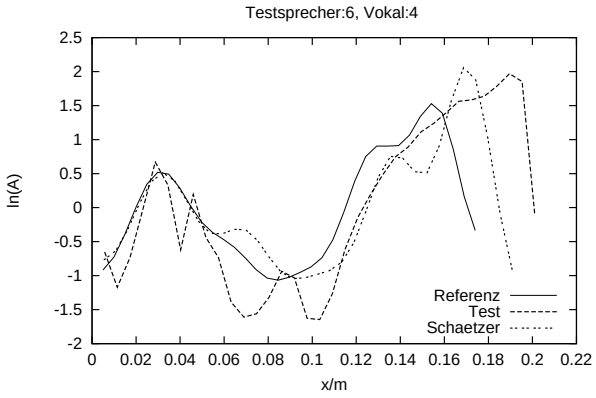
(c) Testsprecher 3



(d) Testsprecher 4

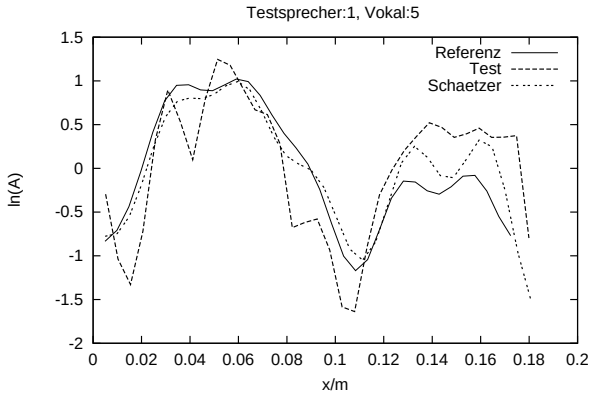


(e) Testsprecher 5

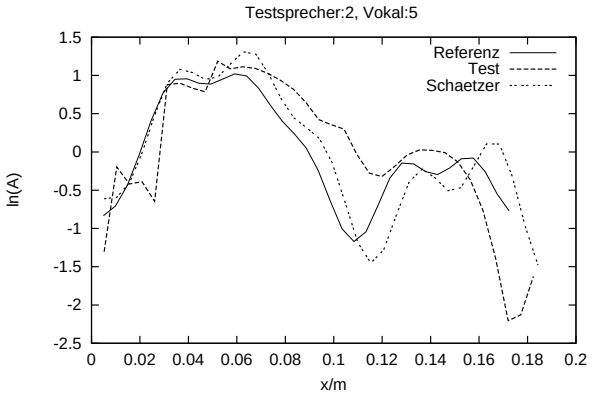


(f) Testsprecher 6

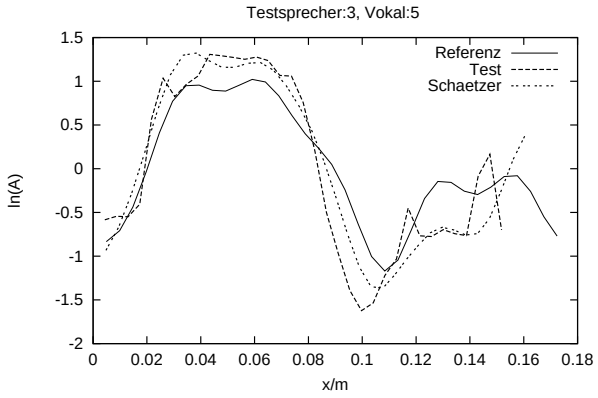
Abbildung B.15.: /oː/: Anpassung des Referenzsprechers an den Testsprecher durch unterschiedliche Längenskalierung beider Vokaltrakthälften und Querschnittsskalierung mit einer aus der Lagrangedichtefunktion abgeleiteten Störungsbasis.



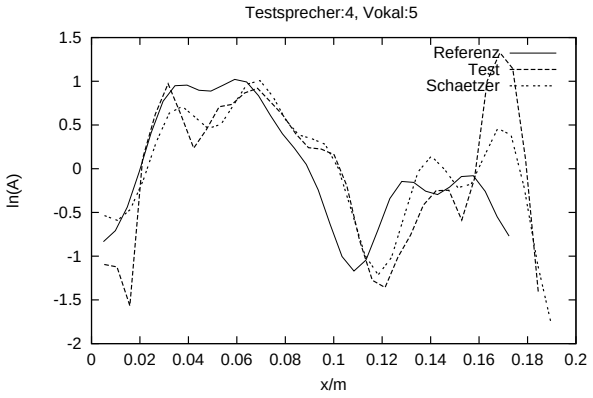
(a) Testsprecher 1



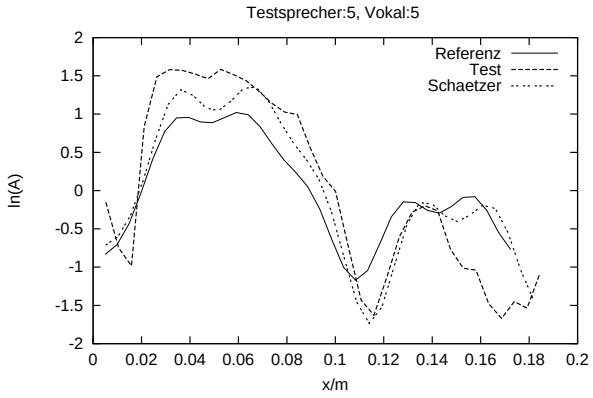
(b) Testsprecher 2



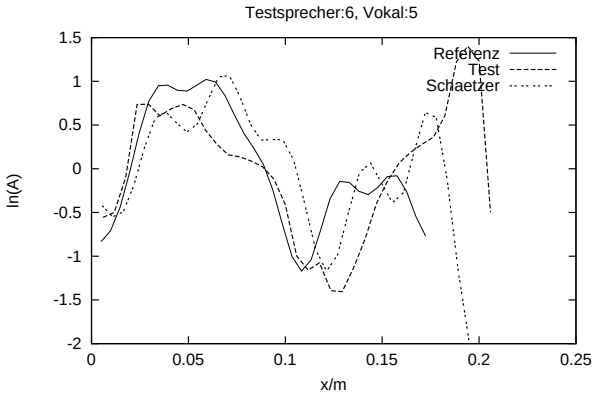
(c) Testsprecher 3



(d) Testsprecher 4



(e) Testsprecher 5



(f) Testsprecher 6

Abbildung B.16.: /uː/: Anpassung des Referenzsprechers an den Testsprecher durch unterschiedliche Längenskalierung beider Vokaltrakthälften und Querschnittsskalierung mit einer aus der Lagrangedichtefunktion abgeleiteten Störungsbasis.

Literaturverzeichnis

[1] Akagi, Masato und Ienaga Taro: *Speaker individuality in fundamental frequency contours and its control.* J. Acoust. Soc. Jpn. (E), 18(2):73–88, 1997.

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

[3] Atal, B.S. und M. R. Schroeder: *Adaptive predictive coding of speech signals.* Bell Sys. Tech. J., 49:1973–1986, 1970.

[4] Baer, T., J. C. Gore, L. C. Gracco und P. W. Nye: *Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels.* J. Acoust. Soc. Am., 90(2):799–828, 1991.

[5] Båvegård, Mats: *Towards an articulatory speech synthesiser: Model development and simulations.* TMH-QPSR, 1/1996:1–15, 1996.

[6] Borg, G.: *Eine Umkehrung der Sturm–Liouvilleschen Eigenwertaufgabe.* Acta Mathematica, 78:1–96, 1946.

[7] Chen, Marylin Y.: *Acoustic correlates of English and French nasalized vowels.* J. Acoust. Soc. Am., 102(4):2360–2370, 1997.

[8] Childers, D.G., Ke Wu, Hicks D. M. und B. Yegnanarayana: *Voice conversion.* Speech Communication, 8:147–158, 1989.

[9] Clermont, Frantz und Parham Mokhtari: *Acoustic-articulatory evaluation of the upper vowel-formant region and its presumed speaker-specific potency.* In: *Fifth International Conference on Spoken Language Processing*, Band 2, Seiten 527–530, 1998.

[10] Dang, Jianwu und Kiyoshi Honda: *Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation.* J. Acoust. Soc. Am., 100(5):3374–3382, 1996.

[11] Dang, Jianwu und Kiyoshi Honda: *Acoustic characteristics of the piriform fossa in models and humans.* J. Acoust. Soc. Am., 101(1):456–465, 1997.

[12] Dang, Jianwu, Kiyoshi Honda und Hisayoshi Suzuki: *Morphological and acoustical analysis of the nasal and the paranasal cavities.* J. Acoust. Soc. Am., 96(4):2088–2100, 1994.

[13] Fant, Gunnar: *Acoustic Theory of Speech Production.* Mouton, Den Haag, 2. Auflage, 1970.

[14] Fant, Gunnar: *Non-Uniform Vowel Normalization.* STL-QPSR, 2-3/1975:1–19, 1975.

[15] Fant, Gunnar: *Vocal-Tract Area and Length Perturbations.* STL-QPSR, 4/1975:1–14, 1975.

[16] Feng, Gang und Eric Castelli: *Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization.* J. Acoust. Soc. Am., 99(6):3694–3706, 1996.

[17] Flanagan, James L.: *Speech Analysis, Synthesis and Perception.* Springer-Verlag Berlin Heidelberg New York, 2. Auflage, 1972.

[18] Freienstein, H., K. Müller und H.W. Strube: *Vocal-tract parameter estimation from formant patterns.* In: *J. Acoust. Soc. Am.*, Band 105, Seite 978, 1999.

[19] Freienstein, Heiko, Knut Müller und Hans Werner Strube: *Extraktion sprecherspezifischer Parameter des Stimmkanals aus Vokalen.* In: Sill, Albert (Herausgeber): *Fortschritte der Akustik – DAGA98*, Seiten 336–337, Oldenburg: DEGA, 1998.

[20] Freienstein, Heiko, Knut Müller und Hans Werner Strube: *Schätzung von sprecherspezifischen Vokaltraktparametern.* In: Mehnert, D. (Herausgeber): *Elektronische Sprachsignalverarbeitung, Studientexte zur Sprachkommunikation*, Band 16, Seiten 208–215, Dresden: w.e.b. Universitätsverlag, 1999.

[21] Furui, Sadaoki: *Research on individuality features in speech waves and automatic speaker recognition techniques.* Speech Communication, 5:183–197, 1986.

[22] Glenn, James W. und Norbert Kleiner: *Speaker Identification Based on Nasal Phonation.* J. Acoust. Soc. Am., 43(2):368–372, 1968.

[23] Heinz, J. M.: *Perturbation functions for the determination of vocal-tract area functions from vocal-tract eigenvalues.* STL-QPSR, 1/1967:1–14, 1967.

[24] Högberg, Jesper: *From sagittal distance to area function and male to female scaling of the vocal tract.* STL-QPSR., 4/1995:11–53, 1995.

[25] Kaufmann, Joachim: *Ein Stimmkanalmodell zur Schätzung von sprecherspezifischen Vokaltraktparametern.* Diplomarbeit, Drittes Physikalisches Institut der Georg-August-Universität Göttingen, 1999.

[26] KIELBASINSKI, A. und H. SCHWETLICK: *Numerische lineare Algebra.* Verlag Harry Deutsch, 1988.

[27] KITAMURA, T. und M. AKAGI: *Relationship between physical characteristics and speaker individualities in speech spectral envelopes.* In: *Proceedings of the Third Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan*, Seiten 833–838, 1996.

[28] KUWABARA, HISAO und YOSHINORI SAGISAKA: *Acoustic characteristics of speaker individuality: Control and conversion.* Speech Comm., 16:165–173, 1995.

[29] LEPAGE, WILBUR R. und SAMUEL SEELY: *General Network Analysis.* McGraw-Hill, 1952.

[30] MAKHOUL, JOHN: *Linear Prediction: A Tutorial Review.* Proceedings of the IEEE, 63(4):561–580, 1975. Korr. 64:285, 1976.

[31] MAN MOHAN SONDHI und JÜRGEN SCHROETER: *A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer.* IEEE Trans. Acoust. Speech Signal Process., 35(7):955–967, 1987.

[32] MARKEL, J. D. und A. H. GRAY, JR.: *Linear Prediction of Speech.* Springer-Verlag Berlin, 1976.

[33] MATSUMOTO, HIROSHI und HISASHI WAKITA: *Vowel Normalization by frequency warped spectral matching.* Speech Communication, 5:239–251, 1986.

[34] MERMELSTEIN, P.: *Articulatory model for the study of speech production.* J. Acoust. Soc. Am., 53(4):1070–1082, 1973.

[35] MEYER, E. und D. GUICKING: *Schwingungslehre.* Friedr. Vieweg & Sohn, Braunschweig/Wiesbaden, 1. Auflage, 1974.

[36] MICHAELIS, DIRK, MATTHIAS FRÖHLICH und HANS WERNER STRUBE: *Selection and combination of acoustic features for the description of pathologic voices.* J. Acoust. Soc. Am., 103(3):1628–1639, 1998.

[37] MIZUNO, HIDEYUKI und MASANOBU ABE: *Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt.* Speech Comm., 16:153–164, 1995.

[38] NAITO, MASATI, LI DENG und YOSHINORI SAGISAKA: *Speaker clustering for speech recognition using the parameters characterizing vocal-tract dimensions.* In: *Proceedings of ICSSP '99*, Band 2, Seiten 981–984, Seattle, WA, 1998.

[39] NAITO, MASATI, LI DENG und YOSHINORI SAGISAKA: *Model-based Speaker Normalization Methods for Speech Recognition.* In: *Proceedings of Eurospeech '99*, Band 6, Seiten 2515–2518, Budapest, 1999.

[40] Paige, A. und V. W. Zue: *Calculation of Vocal Tract Length.* IEEE Trans. Audio Electroacoust., AU-18(3):268–270, 1970.

[41] Press, William H., Brian P. Flannery und Saul A. Teukolskyand William T. Vetterling: *Numerical recipes in C.* Cambridge University Press, 1989.

[42] Rummelhart, D. E., G. E. Hinton und R. J. Williams: *Learning representations by Back-propagating Errors.* Nature, 323:533, 1986.

[43] Schmidt, Robert F. und Gerhard Thews (Herausgeber): *Physiologie des Menschen.* Springer-Verlag Berlin Heidelberg New York, 1995.

[44] Schroeder, Manfred R.: *Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements.* J. Acoust. Soc. Am., 41(4):1002–1010, 1967.

[45] Schroeder, M. R.: *Computer Speech.* Springer, 1999.

[46] Schroeter, Jürgen und Man Mohan Sondhi: *Techniques for Estimating Vocal-Tract Shapes from the Speech Signal.* IEEE Trans. Speech Audio Process., 2(1):133–150, 1994.

[47] Sickert, Klaus (Herausgeber): *Automatische Spracheingabe und Sprachausgabe.* Markt & Technik, 1983.

[48] Sondhi, M. M.: *Image restoration: The removal of spatially invariant degradations.* Proc. IEEE, 69(7):842–853, 1972.

[49] Sondhi, M. M.: *Resonances of a bent vocal tract.* J. Acoust. Soc. Am., 79(4):1113–1116, 1986.

[50] Story, Brad H. und Ingo R. Titze: *Parameterization of vocal tract area functions by empirical orthogonal modes.* Journal of Phonetics, 26:223–260, 1998.

[51] Story, Brad H., Ingo R. Titze und Eric A. Hoffman: *Vocal tract area functions from magnetic resonance imaging.* J. Acoust. Soc. Am., 100(1):537–554, 1996.

[52] Story, Brad H., Ingo R. Titze und Eric A. Hoffman: *Vocal tract area functions for an adult female speaker based on volumetric imaging.* J. Acoust. Soc. Am., 104(1):471–487, 1998.

[53] Strube, Hans Werner: *Bestimmung der Querschnittsfunktion des menschlichen Stimmkanals aus dem Sprachsignal.* Doktorarbeit, Georg-August-Universität, Göttingen, 1974.

[54] Strube, H. W. in: M. R. Schroeder: *Computer Speech*, Kapitel Appendix A. Springer, Berlin, Heidelberg, New York, 1999.

[55] Su, Lo-Soun, K.-P. Li und K. S. Fu: *Identification of speakers by use of nasal coarticulation.* J. Acoust. Soc. Am., 56(6):1876–1882, 1974.

[56] Tiede, Mark K. und Hani Yehia: *A shape-based approach to vocal tract area function estimation.* In: *Acoust. Soc. Am. and Acoust. Soc. Jap., Third Joint Meeting*, Seiten 861–866, 1996.

[57] Titze, Ingo R.: *The physics of small-amplitude oscillation of the vocal folds.* J. Acoust. Soc. Am., 83(4):1536–1552, 1988.

[58] Titze, Ingo R.: *Principles of Voice Production.* Prentice Hall, Englewood Cliffs, N.J. 07632, 1994.

[59] Titze, Ingo R., Sheila S. Schmidt und Michael R. Titze: *Phonation threshold pressure in a physical model of the vocal fold mucosa.* J. Acoust. Soc. Am., 97(5):3080–3084, 1995.

[60] Uebel, L. F. und P. C. Woodland: *An Investigation into Vocal Tract Length Normalisation.* In: *Proceedings of Eurospeech '99*, Band 6, Seiten 2572–2530, Budapest, 1999.

[61] Wakita, H. und G. Fant: *Toward a better vocal tract model.* STL-QPSR, 1/1978:2–29, 1978.

[62] Werner, Jochen: *Numerische Mathematik 1.* vieweg studium, 1992.

[63] Wilhelms, Reiner: *Schätzung von artikulatorischen Bewegungen eines stilisierten Artikulatormodells aus dem Sprachsignal.* Doktorarbeit, Georg-August-Universität, Göttingen, 1987.

[64] Wilhelms-Tricarico, Reiner F.: *Blueprint of a biomechanical model of vocal tract structures.* In: *ACUSTICA - acta acustica*, Band 85 Suppl. 1, Seite 52, 99.

[65] Yehia, Hani und Fumitada Itakura: *A method to combine acoustic and morphological constraints in the speech production inverse problem.* Speech Communication, 18:151–174, 1996.

Danksagung

Ich möchte an erster Stelle Herrn Prof. Manfred Robert Schroeder für die Ermöglichung dieser Arbeit danken. Er verlieh dem Seminar seine unverwechselbare persönliche Note.

Herrn Dr. H. W. Strube gilt mein Dank für die stets hilfsbereite Betreuung der Arbeit und die Mühe, die er aufbrachte, wenn bei rasant nahenden Abgabeterminen die Zeit für Korrekturen knapp wurde.

Den Mitgliedern der Arbeitsgruppe Knut, Dirk, Matthias, Joachim, Jan, Tillman, Jannis, Olaf und natürlich den Ehemaligen Hans, Holger und Kyrill sei gedankt. Insbesondere möchte ich Joachim für die gute Zusammenarbeit und seine Freundschaft danken und nicht zu vergessen das Korrekturlesen dieser Arbeit. In diesem Zusammenhang danke ich auch Matthias für die Anregungen zur frühen Version und Olaf für tatkräftige Hilfe bei der späten Version dieser Arbeit. Viel Dank gebührt auch Knut, der immer Rat wusste - sei es in TEX-Fragen oder bei Herausforderungen von Si++/C++.

Sie alle haben die Atmosphäre der Arbeitsgruppe geprägt und mit jedem Einzelnen durfte ich schon mit Gewinn über wissenschaftliche und weniger wissenschaftliche Themen sprechen.

Im Dritten Physikalischen Institut habe ich mich immer zu Hause gefühlt, wofür ich allen Mitgliedern danken möchte. Besonders die Gespräche mit Herrn Dr. Guicking, Fabian Evert und Christian Merkwirth habe ich immer sehr genossen.

Ich möchte meiner Familie danken, in der ich mich sehr geborgen fühle. Besonders meinen Eltern, die mich in vielerlei Hinsicht unterstützt und gefördert haben. Besonderen Dank schulde ich meiner lieben Frau Anja. Besonders in der heißen Phase dieser Arbeit war sie oft mit vielen Problemen allein und hat mir den Rücken freigehalten. Anja und meine Tochter Antonia gaben mir Kraft für meine Arbeit.

Herrn Mark K. Tiede, ATR Human Information Processing Research Laboratories, Kyoto, Japan möchte ich für die Mitbenutzung der Datenbank der Querschnittfunktionen danken.

Der Deutschen Forschungsgemeinschaft danke ich schließlich für die Förderung (Str 225-7/1 und Str 225-7/2), die diese Arbeit erst ermöglicht hat.

Ich habe in meinem Leben nichts
gefunden was mein Herz so still und froh
gemacht hätte wie die Worte aus Psalm 23:
„Du bist bei mir“
Immanuel Kant

Lebenslauf

Ich wurde am 12.03.1970 in Göttingen als zweites von drei Kindern von Waltraud und Heinrich Freienstein in Göttingen geboren. Meine Eltern waren beide Lehrer in Witzenhausen, wo ich aufwuchs und bis zur zehnten Klasse die Schule besuchte. 1986 wechselte ich die Schule und besuchte bis zu meinem Abitur 1989 die Gymnasiale Oberstufe der Gesamtschule Bad Sooden-Allendorf. Es schloss sich ein fünfzehnmonatiger Wehrdienst in Hessisch-Lichtenau an.

Im Wintersemester 1990/91 begann ich mein Physikstudium an der Georg-August-Universität zu Göttingen und legte im Oktober 1992 die Diplomvorprüfung ab. Meine Diplomarbeit zum Thema „Breitbandige aktive Schallabsorption mit mehreren Kompensationslautsprechern in einem Modellkanal“ fertigte ich bei Herrn Dr. Guicking im Dritten Physikalischen Institut an und erlangte im Februar 1996 das Diplom.

Im Mai 1996 heiratete ich meine Frau Anja.

Im Sommersemester 1996 wechselte ich in die Arbeitsgruppe „Sprache, Neuronale Netze und Gehörmodelle“ von Herrn Dr. H. W. Strube und Professor Dr. M. R. Schroeder. Ich begann meine Arbeit als wissenschaftlicher Mitarbeiter im Projekt „Sprechernormalisierung unter Einbeziehung artikulatorischer Parameter“, das von der deutschen Forschungsgemeinschaft gefördert wurde und dessen Resultat die vorliegende Arbeit ist.

Es bleibt zu erwähnen, dass im Februar 1999 meine geliebte Tochter Antonia zur Welt kam.