MODELLIERUNG VON LAUTÜBERGÄNGEN MITTELS NICHTLINEARER TRAJEKTORIEN DER VOKALTRAKTFLÄCHEN

K. Schnell, A. Lacroix

Institut für Angewandte Physik, J.W. Goethe Universität Frankfurt am Main
schnell@iap.uni-frankfurt.de

Abstract: Für die Spracherzeugung werden neben den Sprachsignalen auch Modelle der Sprachproduktion verwendet, welche die Akustik des Sprechtraktes nachbilden. Die Parameter der hier verwendeten Sprechtraktmodelle sind die Reflexionskoeffizienten bzw. Querschnittsflächen, die aus Sprachsignalen geschätzt werden. Durch Interpolieren der Modellparameter von einem Parametersatz zu dem folgenden können Lautübergänge gebildet werden, wobei die beiden Parametersätze die jeweils beteiligten Laute repräsentieren. Gegenstand der Untersuchungen sind die Auswirkungen von unterschiedlichen Rohrmodellen und Flächenübergängen auf die generierten Übergänge. Neben linearen Flächenübergängen werden insbesondere nichtlineare Übergänge behandelt.

1 Einleitung

Sprechtraktmodelle sind für die Sprachverarbeitung sowohl von praktischem wie auch von theoretischem Interesse. Die existierenden Modelle des Sprechtraktes unterscheiden sich bezüglich ihrer Komplexität und Modellierungsgüte. Einfache Modelle, wie das LPC-Modell, haben sich infolge ihrer mathematisch leicht handhabbaren Struktur für Anwendungen als nützlich erwiesen, besitzen allerdings ihre Grenzen in der Flexibilität und Generalisierbarkeit. Diese einfachen Modelle können durch eine genauere Modellierung der akustischen Gegebenheiten des Sprechtraktes erweitert werden. Die einfachen wie auch die erweiterten Rohrmodelle besitzen als Parameter die Querschnittsflächen. Artikulatorische Sprechtraktmodelle, wie z.B. [1, 2], modellieren die Artikulatoren direkt und besitzen als Modellparameter deren Einstellungen. Dadurch ist eine sehr hohe Flexibilität des Systems gegeben. Bei diesen Modellen besteht allerdings das Problem der Parametersteuerung und -einstellung. Im Vergleich dazu erweist sich die Parameterbestimmung für die in diesem Beitrag verwendeten Rohrmodelle, welche Querschnittsflächen als Modellparameter aufweisen, als weniger problematisch.

2 Sprechtraktmodelle

2.1 Vokaltrakt

Die Akustik des Vokaltrakts kann bekanntlich durch ein Rohr mit veränderlichen Querschnittsflächen approximiert werden. Die Geometrie des Vokaltraktes wird dann durch die Querschnittsflächen des Rohres beschrieben, die hauptsächlich durch die Zunge, sowie durch die Kiefer- und die Lippenstellung beeinflußt werden. Die Bewegungen der Artikulatoren bewirken Veränderungen der mediosagittalen Distanz *d* des Vokaltraktes (im folgenden abgekürzt mit Distanz bezeichnet), welche im mittleren Bereich den Abstand zwischen Zunge und Gaumen beschreibt. Für die Beziehung zwischen der Querschnittsfläche *A* und der Distanz *d* existieren einfache funktionale Modelle, wie das $\alpha-\beta-$Modell mit $A(t)=\alpha\cdot d(t)^{\beta}$; *t* stellt die Abhängigkeit von der Zeit dar. Hierbei muß erwähnt werden, daß die Beziehung zwischen *d* und *A* im Vokaltrakt auch ortsabhängig ist [3]. Bei einem linearen Übergang der Distanz in der Zeit *t* resultiert ein Flächenübergang der proportional zu $A(t)^{1/\beta}$

ist. Mit $\beta > 1$ werden kleinere Flächenwerte *A* stärker bewertet als größere. Dasselbe Verhalten ergibt sich auch bei der Verwendung von logarithmierten Flächen. Werden die Veränderungen der Vokaltraktgeometrie im Sagittalschnitt betrachtet, so sind die Distanzen *d* oder logarithmierten Flächen als Modellparameter in der Regel besser geeignet als die Flächen *A*. Die Bewegungen der Zunge erzeugen allerdings in der Regel keinen exakt linearen Übergang in *d* oder log(*A*), da sich die Zunge z.B. nicht nur senkrecht sondern auch längs zur Vokaltrakt-Mittellinie bewegen kann. Damit sind neben vertikalen Flächenbewegungen auch horizontale Bewegungen möglich.

2.2 Rohrmodell

Die Schallausbreitung im Vokaltrakt kann näherungsweise durch die eindimensionale Ausbreitung ebener Wellen beschrieben werden. Daraus läßt sich das Rohrmodell für den Vokaltrakt ableiten. Die zeitdiskrete Realisierung sieht quantisierte Laufzeiten für die Rohrelemente sowie Adaptoren für die Querschnittssprünge vor. In der Regel wird im Vokaltrakt eine verlustlose Wellenausbreitung angenommen; daraus resultiert das verlustlose Standardrohrmodell, ein LPC-Modell mit Reflexionskoeffizienten als Modellparametern. Im Vokaltrakt treten real allerdings unterschiedliche Verlustmechanismen auf, so daß auch Rohrmodelle verwendet werden [4], welche die Wandvibrationen, die viskose Reibung und die thermischen Verluste berücksichtigen. Dies wird durch ein frequenzabhängiges Verlustsystem $V(z)$ realisiert, zusätzlich zu den Laufzeitgliedern z^{-1} der Rohrelemente. Neben den inneren verteilten Verlusten, wird der Lippenabschluß ebenfalls frequenzabhängig realisiert. Die Parameter dieses verlustbehafteten Rohrmodells können nicht durch Standardverfahren, wie den Burg-Algorithmus, adäquat geschätzt werden. Daher wird für die Schätzung ein Optimierungsverfahren verwendet, welches ein spektrales Abstandsmaß zwischen dem Modellbetragsgang und dem Sprachspektrum minimiert [4]. Vor der eigentlichen Schätzung der Rohrmodellparameter wird das zu analysierende Sprachsignal mit einer adaptiven Präemphase gefiltert, um den Einfluß der Anregung und der Abstrahlung zu eliminieren.

3 Modellierung von Lautübergängen

Die Modellierungen von Lautübergängen werden mit dem Rohrmodell in der Flächendarstellung vorgenommen. Für den Übergang stehen eine Flächenkonfiguration $\boldsymbol{a}_1$ des ersten Lautes und eine Flächenkonfiguration $\boldsymbol{a}_2$ des folgenden Lautes zur Verfügung. Die Generierung des Lautüberganges wird durch eine Interpolation zwischen den Konfigurationen $\boldsymbol{a}_1$ und $\boldsymbol{a}_2$ erzielt. Eine Konfiguration $\boldsymbol{a} = (a(0), a(1), ..., a(N-1), a(N))^{\mathrm{T}}$ stellt dabei einen Vektor mit den einzelnen Komponenten $a(i)$ dar. Die Komponenten a können die Flächen *A* selbst oder von ihnen abgeleitete Größen wie Potenzen oder Logarithmen von *A* sein.

3.1 Linearer Flächenübergang

Bei dem linearen Ansatz ergibt sich der resultierende Flächenvektor aus einer Linearkombination der beteiligten Flächenvektoren $\boldsymbol{a}_1$ und $\boldsymbol{a}_2$:

$$\boldsymbol{a}(t) = (1 - f(t)) \cdot \boldsymbol{a}_1 + f(t) \cdot \boldsymbol{a}_2 \quad , \; t = t_1 \ldots t_2 .$$

$f(t)$ beschreibt dabei die Dynamik des Übergangs und ist eine zwischen Null und Eins monoton steigende Funktion, die im einfachsten Fall linear anwächst. Für die Spracherzeugung sind sigmoide Zeitverläufe realistischer, da die Bewegungen der

massebehafteten Artikulatoren den mechanischen Gesetzen unterliegen. Unabhängig von der zeitlichen Dynamik wird die Bahn bei dem linearen Ansatz durch eine Gerade beschrieben. Die Güte der Modellierung der Lautübergänge mit dem linearen Ansatz ist von den jeweiligen Lautkombinationen abhängig. Dies begründet den Ansatz, auch nichtlineare Trajektorien zu betrachten.

3.2 Nichtlineare Flächenübergange

Für die Realisierung eines nichtlinearen Flächenüberganges wird der Übergang in zwei lineare Übergänge aufgeteilt. Die zwei linearen Übergänge verbinden die Anfangskonfiguration $\boldsymbol{a}_1$ mit einer Zwischenkonfiguration $\boldsymbol{b}$ und diese mit der Endkonfiguration $\boldsymbol{a}_2$

$$\boldsymbol{a}_1 \;\rightarrow\; \boldsymbol{b} \;\rightarrow\; \boldsymbol{a}_2 .$$

Die Aufgabe besteht nun darin, die Zwischenkonfiguration $\boldsymbol{b}(\boldsymbol{a}_1,\boldsymbol{a}_2)$, die den zeitlichen Mittelpunkt des Überganges darstellt, aus den beiden Konfigurationen $\boldsymbol{a}_1$ und $\boldsymbol{a}_2$ zu ermitteln. Im linearen Übergang ist $\boldsymbol{b}$ gleich $\boldsymbol{a}_{12} = (\boldsymbol{a}_1 + \boldsymbol{a}_2)/2$. Im nichtlinearen Fall erhält $\boldsymbol{b}$ eine Komponente von $\boldsymbol{a}_{12}$, sowie auch nichtlineare Komponenten. Diese nichtlinearen Anteile werden aus zwei Ansätzen ermittelt:

3.2.1 Flächenübergang mit dem Anteil einer neutralen Konfiguration

Der erste nichtlineare Ansatz besteht aus der Berücksichtigung eines Anteils der neutralen Konfiguration des Schwa-Lautes in der zeitlichen Mitte des Übergangs. Dies kann durch die besondere Eigenschaft des Schwa-Lautes motiviert werden, der sich in der Mitte des Vokalviereckes befindet; um den Schwa-Laut herum sind die restlichen Vokale gruppiert. Bei Lautübergängen von einem Vokal zu einem anderen ist zu erkennen, daß sich die Trajektorie im Vokalviereck der Position des Schwa-Lautes annähert. Dieses Verhalten wird auch für die Übergänge zwischen Konsonanten und Vokalen angenommen. Aus dieser Betrachtung heraus erhält die Trajektorie der Vokaltraktkonfigurationen für einen Lautübergang im Mittelpunkt eine bestimmte Komponente der neutralen Vokaltraktkonfiguration. Die neutrale Flächenkonfiguration $\boldsymbol{\eta}$ ergibt sich aus der Analyse des Schwa-Lautes und wird mit dem Vektor $\boldsymbol{a}_{12}$ nichtlinear verknüpft. Abhängig von einem Gewichtsvektor $\boldsymbol{g}$ werden die Komponenten von $\boldsymbol{\eta}$ und $\boldsymbol{a}_{12}$ zu einem neuen Vektor $\boldsymbol{b}'$ zusammengefügt

$$\boldsymbol{b}' = \left(g(0)\cdot\eta(0) + (1-g(0))\cdot a_{12}(0) \;,\; g(1)\cdot\eta(1) + (1-g(1))\cdot a_{12}(1) \;,\; \ldots \right)^{\mathrm{T}} .$$

Der Gewichtsvektor $\boldsymbol{g}$ wird so gewählt, daß am Systemausgang (Lippenabschluß) und am Systemeingang des Rohrmodells der Einfluß des Vektors $\boldsymbol{a}_{12}$ stärker gewichtet wird als in der Mitte des Rohrmodells. Dies liegt der Annahme zugrunde, daß die Lippen und der Rachenabschluß stärker von den beiden beteiligten Lauten beeinflußt werden, während im mittleren Bereich, der von der Zunge dominiert wird, die neutrale Konfiguration mehr zur Geltung kommt.

3.2.2 Flächenübergang mit horizontalen Flächenbewegungen

Der zweite nichtlineare Ansatz berücksichtigt explizit neben vertikalen auch horizontale Bewegungen der Vokaltraktflächen. Das Problem hierbei besteht darin, die horizontalen Bewegungen aus den Informationen der beiden Lautkonfigurationen zu bestimmen. Dafür

werden horizontale Bewegungen gesucht, welche die Konfiguration $\boldsymbol{a}_1$ nach erfolgter Bewegung der Konfiguration $\boldsymbol{a}_2$ ähnlicher werden läßt und umgekehrt. Dieses Problem wird durch dynamische Programmierung bzw. den Viterbi-Algorithmus gelöst. Es wird ein dem Dynamic-Time-Warping vergleichbares Verfahren angewendet, bei dem zwei Flächensätze örtlich angepaßt werden und nicht zeitlich. Für die dynamische Programmierung stehen die Vektoren $\boldsymbol{a}_1$ und $\boldsymbol{a}_2$ zur Verfügung, wie in Bild 1 zu sehen ist. Die Elemente der Vektoren

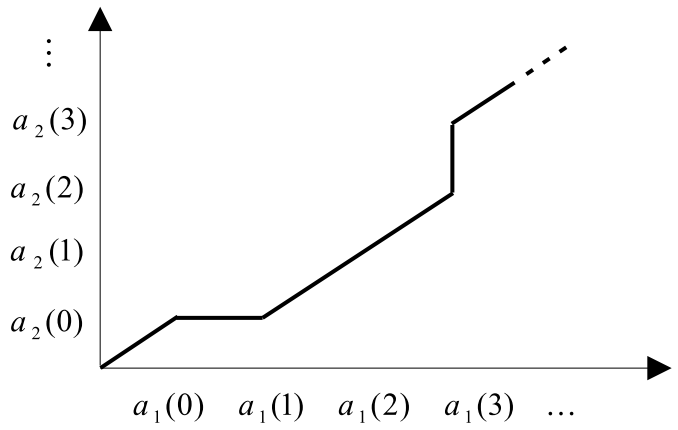

Bild 1 – Beispiel für das Ergebnis der Dynamischen Programmierung mit zwei Flächenvektoren $\boldsymbol{a}_1$ und $\boldsymbol{a}_2$.

sind an den zwei Koordinaten des Graphen angeordnet, durch den ein Pfad von der linken unteren Ecke in die rechte obere gelegt wird. Horizontale oder vertikale Schritte des Pfades verdoppeln Rohrstücke in einer Konfiguration. Dadurch können die Konfigurationen angeglichen werden. Für das zu minimierende Abstandsmaß wird der euklidische Abstand zwischen den beiden Ergebnisvektoren verwendet. Die Mengen λ_1 und λ_2 stellen die Positionen der Vektoren $\boldsymbol{a}_1$ bzw. $\boldsymbol{a}_2$ dar, an denen ein Rohrelement verdoppelt wird. Der optimale Pfad kann durch dynamische Programmierung gefunden werden. Als Resultat ergeben sich zwei Vektoren $\tilde{\boldsymbol{a}}_1$ und $\tilde{\boldsymbol{a}}_2$, die in der Regel länger sind als $\boldsymbol{a}_1$ und $\boldsymbol{a}_2$. Daher müssen die beiden Vektoren auf die Länge der ursprünglichen gekürzt werden. Um die Korrelation zwischen den Vektoren nicht zu verändern, werden in beiden Vektoren ausschließlich Rohrstücke an denselben Positionen eliminiert. Als Positionen werden abwechselnd Werte aus λ_1 und λ_2 verwendet. Dabei werden aus den Mengen λ_1 und λ_2 der Größe nach jede zweite Position berücksichtigt, solange bis die ursprüngliche Vektorlänge erlangt ist. Aus den resultierenden gekürzten Vektoren $\tilde{\boldsymbol{a}}_1'$ und $\tilde{\boldsymbol{a}}_2'$ wird durch Mittelung der Vektor $\tilde{\boldsymbol{a}}' = (\tilde{\boldsymbol{a}}_1' + \tilde{\boldsymbol{a}}_2')/2$ berechnet. $\tilde{\boldsymbol{a}}'$ wird als Komponente der Konfiguration $\boldsymbol{b}'$ zugefügt, so daß die Zwischenkonfiguration $\boldsymbol{b} = \gamma \cdot \boldsymbol{b}' + (1-\gamma) \cdot \tilde{\boldsymbol{a}}'$ resultiert, welche die Stützstelle in der zeitlichen Mitte des Lautübergangs darstellt. Der Gewichtsvektor $\boldsymbol{g}$ und der Faktor γ bestimmen den Grad der nichtlinearen Modellierungen und somit die Abweichungen zum linearen Flächenübergang.

3.3 Beispiele von Lautübergängen

Im folgenden werden Realisierungen des Lautübergangs [gi] mit den zuvor diskutierten Ansätzen behandelt. Für die zu generierenden Flächenverläufe werden die Anfangs- und die Endkonfiguration $\boldsymbol{a}_1$ und $\boldsymbol{a}_2$ benötigt, welche die beteiligten Laute des Übergangs repräsentieren sollen. Dafür werden die Konfigurationen $\boldsymbol{a}_1$ und $\boldsymbol{a}_2$ aus Sprachsignalabschnitten geschätzt, die von einem männlichen Sprecher mit einer Abtastrate von 22 kHz aufgenommen wurden. Die Signalabschnitte bestehen aus vier Grundperioden. Die geschätzten Flächenvektoren beinhalten 28 Flächenwerte entsprechend der Anzahl der Rohrstücke des verwendeten Rohrmodells. Für die zu generierenden Übergänge werden nicht die Flächenwerte A verwendet, sondern Potenzen $A^{2/3}$, entsprechend $\beta = 3/2$ für das $\alpha - \beta -$Modell. Die zwei Zeitabschnitte aus denen $\boldsymbol{a}_1$ und $\boldsymbol{a}_2$ ermittelt werden, stammen aus demselben Sprachsignal. Damit steht der natürliche Flächenverlauf für einen Vergleich mit den erzeugten Übergängen zur Verfügung. Für die Bestimmung des originalen Verlaufs des Lautübergangs werden überlappende Signalabschnitte analysiert, so daß eine dichte zeitliche Sequenz von Flächenvektoren resultiert. Die Betragsgänge des originalen Lautübergangs [gi] sind in Bild 2a zu sehen; das dabei verwendete Rohrmodell ist verlustbehaftet. Für die Generierung des Lautübergangs [gi] mit linearen und nichtlinearen Flächenübergängen werden jeweils die erste und letzte Konfigurationen aus Bild 2a für $\boldsymbol{a}_1$ und $\boldsymbol{a}_2$ herangezogen; die resultierenden Verläufe sind in den Bildern 2b-e zu sehen. Für die Darstellung der Übergänge ist $f(t)$ linear gewählt. Der Betragsgang der Zwischenkonfiguration ist jeweils durch eine dickere Strichstärke gekennzeichnet. Dieser repräsentiert bei einem linearen Flächenübergang die Konfiguration $\boldsymbol{a}_{12}$ und bei einem nichtlinearen Übergang die Konfiguration $\boldsymbol{b}$. Für einen Vergleich ist auch im originalen Übergang von Bild 2a die mittlere Kurve dicker gezeichnet. An den Bildern 2b-e ist zu sehen, daß sich die Zwischenkonfigurationen hinsichtlich des verwendeten Rohrmodells und der Art des Flächenübergangs unterscheiden. Für eine Beurteilung der Betragsgänge ist zu beachten, daß der niederfrequente Bereich für Sprache mehr Informationen enthält als der höherfrequente. Abweichungen zum originalen Übergang von Bild 2a ergeben sich in der Güte der zweiten Resonanz, knapp unter 2 kHz, und für die zwei dicht benachbarten Resonanzen um 3 kHz. Bei den linearen Übergängen der Bilder 2b und 2d ist die zweite Resonanz im Gegensatz zu Bild 2a zu schwach ausgeprägt. Weiterhin sind die beiden Resonanzen um 3 kHz nicht deutlich bzw. gar nicht getrennt. Die Konfiguration mit dem verlustbehafteten Rohrmodell ist diesbezüglich dem originalen Verlauf ähnlicher als die des verlustlosen LPC-Rohrmodells. Bei anderen Lautübergängen kann dieser Unterschied auch stärker hervortreten. In den Bildern 2c und 2e ist ein nichtlinearer Übergang mit einem verlustbehafteten Rohrmodell realisiert. Dieser verbessert die mittlere Konfiguration im Vergleich zum linearen Übergang von Bild 2d. Für das Ergebnis in Bild 2c wurde nur der nichtlineare Ansatz mit dem anteiligen Schwa-Laut berücksichtigt, während für den Übergang des Bildes 2e zusätzlich horizontale Flächenbewegungen berücksichtigt sind. Es ist zu sehen, das hier der Ansatz mit dem Anteil der neutralen Konfiguration die wesentliche Verbesserung erbringt. Dies ist auch bei anderen untersuchten Lautkombinationen erkennbar. Der Verlauf in Bild 2f wurde durch denselben Typ des nichtlinearen Flächenübergang wie in Bild 2e erzielt, berücksichtigt allerdings für $\boldsymbol{a}_2$ eine andere Konfiguration. Die Konfiguration $\boldsymbol{a}_2$ von Bild 2f stellt auch den Vokal /i/ dar, wurde allerdings aus einem Signalabschnitt der Äußerung [di] ermittelt. Die mittlere Konfigurationen der Bilder 2e und 2f sind insbesondere in der unteren Hälfte des Frequenzbereichs ähnlich.

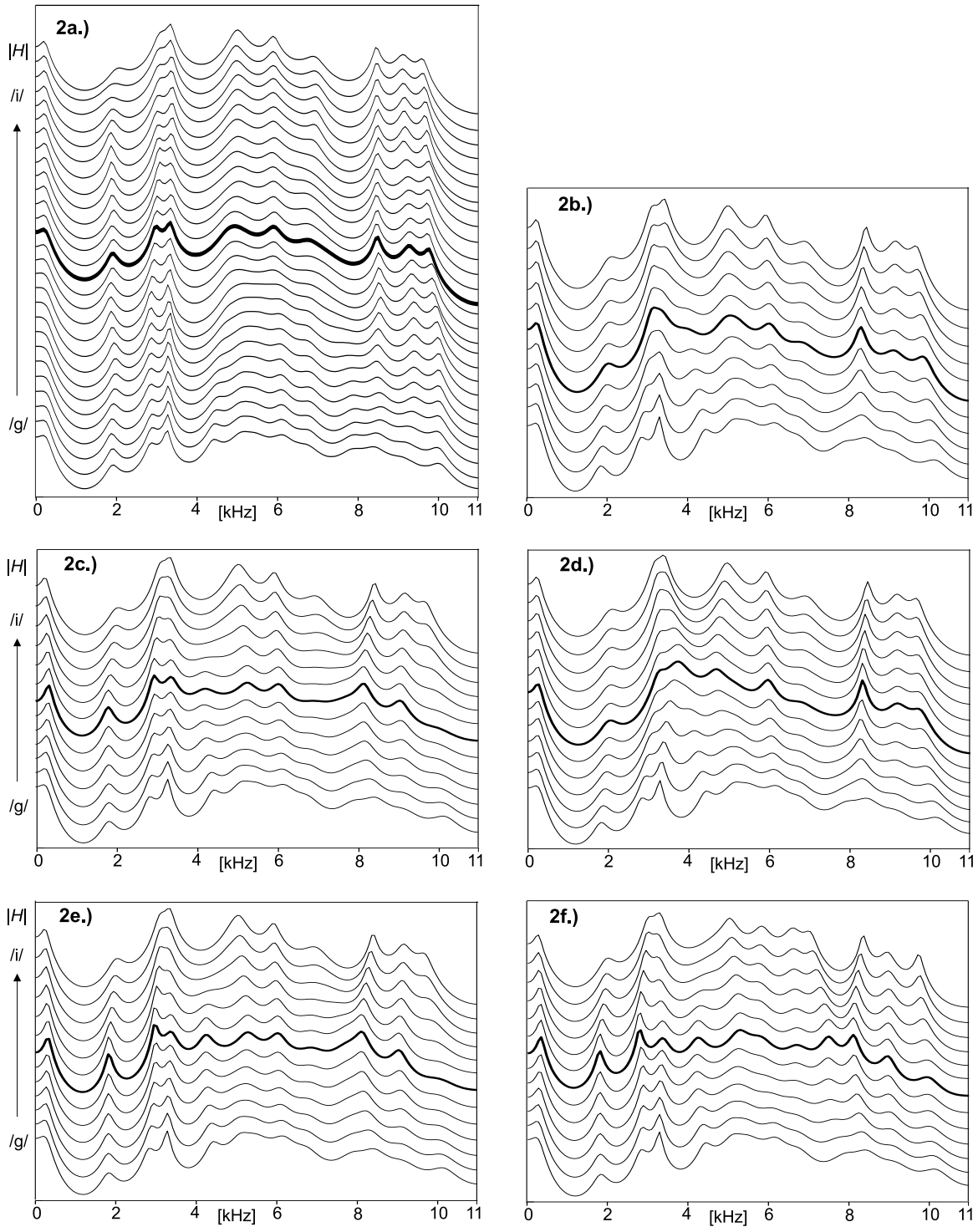

Bild 2 – Betragsgänge des Lautübergangs [gi]: (a) originaler Übergang; (b) linearer Flächenübergang mit LPC-Rohrmodell; (c) nichtlinearer Flächenübergang (ohne horizontale Bewegung) mit verlustbehaftetem Rohrmodell; (d) linearer Flächenübergang mit verlustbehaftetem Rohrmodell; (e) nichtlinearer Flächenübergang mit verlustbehaftetem Rohrmodell; (f) nichtlinearer Flächenübergang mit verlustbehaftetem Rohrmodell, die Konfiguration $\boldsymbol{a}_2$ wurde aus [di] ermittelt.

4 Zusammenfassung

Mit den vorgestellten Methoden wird der Ansatz verfolgt, die Flächendarstellung des Rohrmodells für die Beschreibung von Lautübergängen auszunutzen. Neben dem linearen Flächenübergang werden insbesondere nichtlineare Übergänge diskutiert. Diese weisen in der Übergangsmitte eine Tendenz zu einer neutralen Konfiguration auf und erlauben auch die Berücksichtigung horizontaler Flächenbewegungen. Anhand der Modellbetragsgänge wird nachgewiesen, daß durch solche nichtlinearen Übergänge Verbesserungen erzielt werden können. Weiterhin hat sich das verlustbehaftete Rohrmodell für die Flächenübergänge im Gegensatz zu dem verlustlosen Standardrohrmodell durchweg als vorteilhaft erwiesen.

Literatur

[1] Dang, J., Honda, K.: “Construction and control of a physiological articulatory model” J.A.S.A. Vol. 115 (2), February 2004, pp. 853-870.

[2] Kröger, B. J.: „Ein phonetisches Modell der Sprachproduktion“, Linguistische Arbeiten 387, Max Niemeyer Verlag Tübingen, 1998.

[3] Soquet, A. ;Lecuit, V; Metens, T. ;Demolin, D.: “Mid-sagittal cut to area function transformation: Direct measurements of mid-sagittal distance and area with MRI”, Speech Communication Vol. 36, 2002, pp. 169-180.

[4] Schnell, K.; Lacroix, A.: “Analysis of lossy vocal tract models for speech production”, Proc. EUROSPEECH-2003, Geneva Switzerland, pp. 2369-2372.