VERKETTUNG VON SPRACHEINHEITEN FÜR DIE SPRACHERZEUGUNG MITTELS VERLUSTBEHAFTETER ROHRMODELLE

Karl Schnell, Arild Lacroix

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

Abstract: Die akustische Synthese einer konkatenativen Sprachsynthese wird häufig mit Hilfe von Zeitbereichsverfahren realisiert. In diesem Beitrag wird ein Ansatz diskutiert, der modellbasiert ist und einer parametrischen Synthese entspricht. Als Modell des Sprechtraktes wird das verlustbehaftete Rohrmodell verwendet, das in früheren Beiträgen schon vorgestellt wurde. Die zu verkettenden Einheiten für die akustische Synthese basieren auf analysierten Lautübergängen, die aus dem Sprachkorpus einer Diphondatenbank stammen.

1 Einleitung

Die akustische Synthese von beliebigen Sprachäußerungen wird in der Regel durch eine Verkettung von Spracheinheiten realisiert [1]. Die Spracheinheiten können dabei aus Diphonen, Triphonen, Silben oder ganzen Wörtern bestehen. Da diese Einheiten aus Sprachäußerungen stammen, die in der Regel nicht mit den zu synthetisierenden Äußerungen phonetisch identisch sind, müssen die Spracheinheiten an die neue phonetische Umgebung angepaßt werden. Dies beinhaltet die Anpassung der prosodischen Parameter, wie Grundfrequenz, Lautdauer und Lautstärke sowie eine spektrale Anpassung an den Konkatenationsstellen. Ein besonderes Problem stellt die Grundfrequenzänderung dar. Hierfür existieren Zeitbereichsverfahren beispielsweise wie TD-PSOLA, die auch durch einfache Modellbeschreibungen ergänzt werden können, entsprechend LPC-PSOLA. In diesem Beitrag wird eine ausschließlich modellhafte Beschreibung behandelt. Als Modell wird das verlustbehaftete Rohrmodell verwendet [2, 3]. Da das Residualsignal für die Spracherzeugung nicht verwendet wird, ist der vorgestellte Ansatz der einer parametrischen Synthese.

2 Analyse und Synthese mit dem verlustbehafteten Rohrmodell

2.1 Verlustbehaftetes Rohrmodell

Als Modellbeschreibung des Sprechtraktes wird das verlustbehaftete Rohrmodell verwendet, welches in [2] vorgestellt worden ist und das für die Resynthese in [3] schon benutzt wurde. Das verlustbehaftete Rohrmodell ist zeitdiskret in Kreuzgliedstruktur realisiert. Im Gegensatz zum verlustlosen Standard-Rohrmodell weist es einen frequenzabhängigen Lippenabschluß $\alpha L(z)$ nach [4] am Systemausgang auf. Zusätzlich werden verteilte frequenzabhängige Verluste, die bei der Wellenausbreitung innerhalb des Sprechtraktes auftreten, durch rekursive Systeme $V(z)$ berücksichtigt. Dafür werden die Laufzeiten der verlustlosen Rohrelemente durch verlustbehaftete Laufzeiten $\vartheta = V \cdot z^{-1}$ ersetzt, wie in [2] beschrieben. Die verlustbehafteten Laufzeiten sind, wie in Bild 1 dargestellt, abwechselnd im oberen und unteren Pfad des Signalflußgraphen des Rohrmodells plaziert. Damit können Verluste durch Wärmeleitung und viskose Reibung berücksichtigt werden, die sich überwiegend auf den höheren Frequenzbereich auswirken, sowie die Wandvibrationen, die sich insbesondere auf die niedrigen Frequenzen auswirken. Das Rohrmodell wird für die Synthese in der

Darstellung mit Leistungswellen betrieben, während für die Analyse eine Wellendarstellung in der Grundform verwendet wird [5], die Betriebskettenmatrizen mit dem Vorfaktor 1 benutzt statt des Vorfaktors $(1+r^2)^{-1/2}$ für Leistungswellen. In Bild 1 ist der Signalflußgraph des Rohrmodells mit den verlustbehafteten Laufzeiten ϑ gezeigt; r_i stellt den i-ten Reflexionsfaktor dar. Zwischen den Reflexionsfaktoren und den Flächen A_i existiert die Beziehung $r_i = (A_i - A_{i+1})/(A_i + A_{i+1})$.

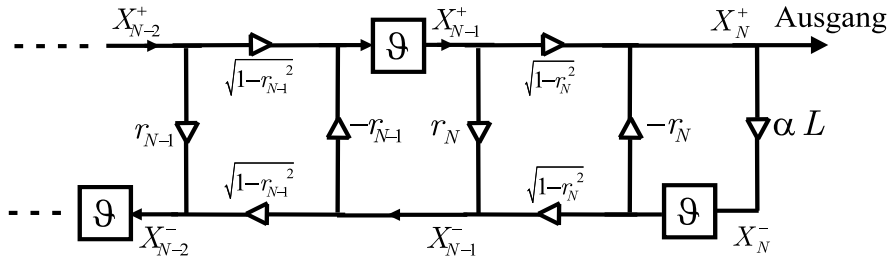

Bild 1 - Signalflußgraph des verlustbehafteten Rohrmodells.

2.2 Parameterbestimmung des Rohrmodells

Für die Spracherzeugung werden die Parameter des Rohrmodells aus Sprachsignalen geschätzt. Die zu schätzenden Parameter sind die Reflexionskoeffizienten. Die Parameter des Rohrabschlusses $\alpha L(z)$ und der Verlustsysteme $V(z)$ sind hingegen für die Analyse vorbestimmt. Das zu analysierende Sprachsignal s wird durch eine adaptive Präemphase vorgefiltert, die aus einer mehrfach ausgeführten Burg-Analyse erster Ordnung besteht. Danach werden aus dem Betragsspektrum des gefilterten Sprachsignals die Reflexionskoeffizienten geschätzt. Dies wird durch einen Optimierungsalgorithmus erreicht, der eine inverse Filterung im Spektralbereich durchführt [2]. Die theoretischen Grundlagen des Fehlerkriteriums sind in [5] diskutiert.

2.3 Analyse von Diphonen

Als Spracheinheiten werden Diphone der Diphondatenbank de1 verwendet [6]. Diese Diphondatenbank wurde von einer Sprecherin mit einer Abtastrate von 16 kHz aufgenommen. Für eine Synthese müssen die Diphone zuerst analysiert werden. Die stimmhaften Laute und stimmlosen Frikative der Diphone werden dafür in überlappende Abschnitte aufgeteilt, aus denen jeweils die Modellparameter mittels des Optimierungsalgorithmus' bestimmt werden. Um einen zeitlich glatten Verlauf der Parametersätze zu erreichen, werden bei der Durchführung des Optimierungsalgorithmus' nach jeweils einigen Iterationen die resultierenden Parameter von aufeinanderfolgenden Abschnitten gemittelt. Die Analyse der Diphone kann auf unterschiedliche Weise durchgeführt werden. Die zu analysierenden Abschnitte des Sprachsignals werden entweder mit fester Blocklänge oder grundperiodensynchron segmentiert. Die Grundperioden werden für die grundperiodensynchrone Segmentierung an den Nulldurchgängen markiert; die Blocklänge ist dann durch die Länge einer bestimmten Anzahl von Perioden bestimmt. Da bei stimmlosen Lauten keine periodische Struktur vorliegt, kommt dort nur die feste Blocklänge zur Anwendung. Neben einer unterschiedlichen Segmentierung kann sich die Analyse der Sprachabschnitte auch durch Anwendung von verschiedenen Fensterfunktionen auf die Signalabschnitte unterscheiden. Im Falle einer festen Blocklänge ist die Gewichtung mit einer an den Rändern verschwindenden Fensterfunktion erforderlich, um störende

Blockgrenzeneffekte zu vermeiden. Bei der grundperiodensynchronen Segmentierung ist hingegen auch das Rechteckfenster sinnvoll. Neben den Reflexionskoeffizienten werden auch die Präemphase-Koeffizienten für die Diphone adaptiv bestimmt. Diese werden im Gegensatz zu den Reflexionskoeffizienten aus dem gesamten zu analysierenden Sprachabschnitt geschätzt. Damit besitzen sie für alle stimmhaften bzw. stimmlosen Parametersätze eines Diphons dieselben Werte.

3 Synthese mit Diphonen

Nach der Analyse stehen für jedes Diphon jeweils eine bestimmte Anzahl von Parametersätzen zu Verfügung, die für eine Synthese mit dem Rohrmodell verwendet werden. Die Parametersätze beinhalten die logarithmierten Flächensätze, die Präemphasekoeffizienten und eine Leistungsangabe. Für die akustische Synthese mit dem Rohrmodell wird neben den Modellparametern noch eine entsprechende Anregung des Systems benötigt, die abhängig von dem zu synthetisierenden Laut ist. Für die stimmhafte Anregung wird eine Impulsfolge verwendet, der ein hochpaß-gefiltertes Rauschen additiv überlagert ist. Die stimmlose Anregung für Frikative wird durch weißes Rauschen realisiert. Für die Synthese werden die Parametersätze nacheinander als Modellparameter verwendet. Der Wechsel der Parametersätze wird nach jeder synthetisierten Periode bzw. Block vollzogen, wobei wegen der Lautdaueranpassung nicht alle Parametersätze berücksichtigt werden. Diese Parametersätze werden dabei linear von einem in den nächsten Parametersatz überführt. Durch die Verwendung der Parametersätze in der beschriebenen Weise können die Diphone resynthetisiert werden. Für eine Synthese von Lautfolgen werden die Diphone, die durch ihre Parametersätze repräsentiert werden, miteinander verkettet. Die Anpassung der Diphone an die Lautfolge erfordert eine Änderung ihrer ursprünglichen prosodischen Parameter. Weiterhin sollten die Flächenverläufe an den Diphongrenzen stetig überführt werden, um Unstetigkeiten zu vermeiden.

3.1 Prosodische Parameter

Die Änderung des Grundfrequenzverlaufs an die neue Lautfolge kann sich bei Zeitbereichsverfahren für größere Grundfrequenzänderungen als problematisch erweisen. Die Grundfrequenzänderung kann hingegen bei der hier vorliegenden parametrischen Synthese unmittelbar durchgeführt werden, da die Systemanregung unabhängig von dem analysierten Signal ist. Neben der Grundfrequenzänderung muß auch eine Anpassung der Lautdauer durchgeführt werden, die durch Auslassen oder Verdoppeln von Parametersätzen erreicht werden kann. Um einen möglichst glatten Verlauf der Parameter zu erhalten, kann bei einer Lautdauerverlängerung statt einer Verdopplung auch ein neuer Parametersatz eingefügt werden, der aus einer Mittelung der beiden Nachbarsätze resultiert. Bei den untersuchten synthetisierten Beispielen ist allerdings im Vergleich zur Parametersatzverdopplung auditiv kaum ein Unterschied feststellbar.

3.2 Verkettung

Werden die Diphone bzw. Parametersätze für die Synthese hart aneinander gesetzt, so entstehen Unstetigkeiten an den Konkatenationsstellen, die sich negativ auf die Sprachqualität auswirken. Da hier der Ansatz einer parametrischen Synthese gegeben ist, wird das Residualsignal für die Synthese nicht verwendet, so daß durch die Anregung keine Unstetigkeiten auftreten. Unstetigkeiten entstehen jedoch in den Flächensätzen und damit auch in den Modellfrequenzgängen. Ein Flächensatz $\boldsymbol{a} = \big(a(0), a(1), a(2), \ldots, a(N)\big)$ besteht aus den logarithmierten Flächen $a(k) = \log(A(k))$. Um die Unstetigkeiten an den

Konkatenationsstellen zu beseitigen, werden die letzten Flächensätze des ersten Diphons ***a*** linear in die ersten Flächensätze des nächsten Diphons ***b*** überführt:

$$\begin{aligned} &\ldots, \boldsymbol{a}_{N-2}, \boldsymbol{a}_{N-1}, \boldsymbol{a}_N, \boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3, \ldots \\ &\qquad\qquad \downarrow \\ &\ldots, \tilde{\boldsymbol{a}}_{N-2}, \tilde{\boldsymbol{a}}_{N-1}, \tilde{\boldsymbol{a}}_N, \tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, \tilde{\boldsymbol{b}}_3, \ldots \end{aligned} \tag{1}$$

$$\text{mit} \quad \tilde{\boldsymbol{a}}_{N-k+1} = (1-f)\cdot \boldsymbol{a}_{N-k+1} + f \cdot \boldsymbol{b}_1 \quad , \quad f = \frac{k}{2M}$$

$$\text{und} \quad \tilde{\boldsymbol{b}}_k = (1-f)\cdot \boldsymbol{a}_N + f \cdot \boldsymbol{b}_k \quad , \quad f = \frac{k}{2M} + \frac{1}{2} \quad , \quad k = 1 \ldots M.$$

Mit der Konstanten M in (1) kann die Länge des Übergangs festgelegt werden; die Gesamtlänge des Übergangs ist durch $2M$ gegeben. Die Flächensätze $\tilde{\boldsymbol{a}}$ und $\tilde{\boldsymbol{b}}$ stellen die an der Konkatenationsstelle stetig übergehenden geglätteten Flächenvektoren dar. In der gleichen Weise wie (1) wird auch ein Übergang für die Präemphasekoeffizienten und den Leistungswert vorgenommen. Der resultierende Verlauf der Modellparameter besitzt infolge der Glättung keine Unstetigkeitsstellen mehr an den Verkettungsstellen; dies konnte in [3] an einigen Beispielen mit einer Verkettungsstelle schon gezeigt werden. Bei der hier diskutierten Verkettung von Diphonen entstehen allerdings verhältnismäßig viele Verkettungsstellen. Die synthetisierten Sprachäußerungen mit Diphonen belegen, daß auch hier der verwendete Flächenübergang geeignet ist. Die untersuchten Beispiele zeigen weiterhin, daß die Artefakte, hervorgerufen durch hartes Aneinandersetzen der Parametersätze, unterschiedlich stark ausgeprägt sind. Dies kann in erster Linie damit erklärt werden, daß die Diphone aus phonetisch unterschiedlichen Umgebungen stammen.

3.3 Synthese

Die mit dem Rohrmodell analysierten stimmhaften Laute und stimmlosen Frikative werden den stationären Lauten zugeordnet. Die übrigen Laute weisen einen stark instationären Charakter auf. Für die Synthese dieser Laute werden deren Zeitsignale verwendet. Bei den stimmhaften Explosiven bezieht sich das instationäre Signal auf das Explosionsgeräusch (burst) mit nachfolgendem Rauschen. Die synthetisierten Beispiele der Explosivlaute legen nahe, daß die Zeitsignale gut mit den durch das Rohrmodell synthetisierten Signalen kombiniert werden können. Dabei ist auf einen angemessenen Abstand des Zeitsignals zu dem modellgenerierten Signal und deren Amplitudenrelation zu achten. Es muß angemerkt werden, daß nicht alle instationären Laute der Diphondatenbank untersucht wurden. Welche instationären Sprachlaute für die Synthese unter Umständen auch darüber hinaus modellgestützt behandelt werden können, müssen weitere Untersuchungen zeigen.

3.4 Variation der Verkettung und der Analyseblöcke

Im folgenden werden die Auswirkungen verschiedener Segmentierungen und Fensterfunktionen auf die Analyse diskutiert. Da die Analyseergebnisse durch die geschätzten Parametersätze repräsentiert werden, haben die unterschiedlichen Analysetechniken unmittelbaren Einfluß auf die Synthese. In Bezug auf die Konkatenationen werden darüber hinaus auch die Verbesserungen durch den Flächenübergang deutlich. In Bild 2 werden Ergebnisse mit unterschiedlichen Analysetechniken anhand eines Ausschnitts der synthetisierten Äußerung „Lawine“ gezeigt. Es sind jeweils die Betragsgänge des Rohrmodells in zeitlicher Entwicklung dargestellt. Da für die synthetisierte Äußerung eine Lautdaueranpassung vorgenommen wird, werden nicht immer alle Parametersätze der

analysierten Diphone verwendet. Die Bilder 2a-d zeigen jeweils einen Ausschnitt mit drei Konkatenationsstellen innerhalb der Laute /a/, /v/, /i:/. Für die Betragsgänge der Bilder 2a-d wurden dieselben Diphone verwendet, allerdings unterschiedlich analysiert.

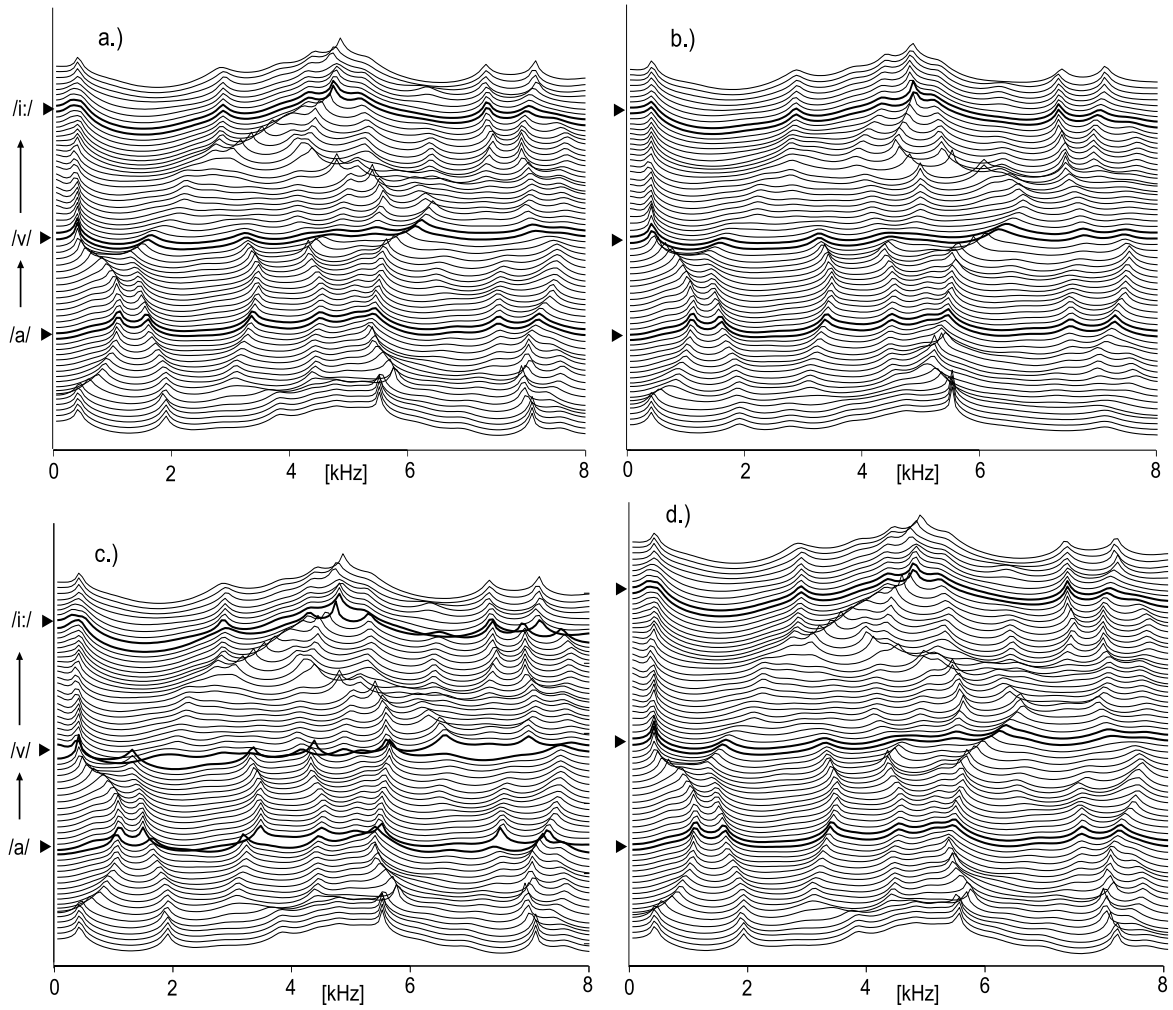

Bild 2 – Betragsgänge der verketteten Flächensätze der Diphone für die Synthese eines Ausschnitts der Äußerung „Lawine“: (a) mittels Flächenübergang an den Konkatenationsstellen und einer Analyse der Diphone mit Hann-Fenster und grundperiodensynchrone Blocklänge; (b) mittels Flächenübergang an den Konkatenationsstellen und einer Analyse der Diphone mit Rechteck-Fenster und grundperiodensynchrone Blocklänge; (c) ohne Flächenübergang an den Konkatenationsstellen und einer Analyse der Diphone mit Hann-Fenster und grundperiodensynchrone Blocklänge; (d) mittels Flächenübergang an den Konkatenationsstellen und einer Analyse der Diphone mit Hann-Fenster und fester Blocklänge.

Die Segmentierung erfolgt in den Bildern 2a, b, und c grundperiodensynchron mit 4 Perioden, in Bild 2d fest mit einer Blocklänge von 320 Abtastwerten, entsprechend der mittleren Länge von 4 Grundperioden. In Bild 2b werden die zu analysierenden Signalblöcke mit einer Rechteck-Fensterfunktion gewichtet, während in den übrigen Fällen ein Hann-Fenster zur Anwendung kommt. Die Betragsgänge der Konkatenationsstellen sind kräftiger gezeichnet und mit einem Dreieck markiert. In Bild 2a und 2c sind Ergebnisse gezeigt, die mit der selben Analysetechnik erzeugt wurden, allerdings mit dem Unterschied daß für Bild 2c keine Flächenübergange an den Verkettungsstellen der Diphone vorgenommen wurden. Es ist in Bild 2c zu sehen, daß ohne diese Übergange Unstetigkeiten im Betragsgang an den Verkettungsstellen entstehen. In dem dargestellten Abschnitt sind die Unstetigkeiten durch die

Verkettung des Lautes /v/ am stärksten ausgeprägt. Bei den anderen Bildern Bild 2a,-b und 2d sind Flächenübergänge an den Konkatenationsstellen mit $M = 4$ entsprechend (1) durchgeführt worden. Es ist deutlich zu erkennen, wie die Konkatenationsstellen spektral geglättet und die Unstetigkeiten beseitigt sind. Dieser positive Effekt ist auch auditiv wahrnehmbar. Neben den Parameterübergängen an den Konkatenationsstellen wirken sich auch die verschiedenen Analysetechniken der Diphone aus, wie in den Bildern 2a, b und d zu erkennen ist. Die Betragsgänge von Bild 2a sind ähnlicher den Betragsgängen von Bild 2d als mit denen von Bild 2b. Dies läßt sich beispielsweise an dem Abschnitt kurz hinter der Verkettungsstelle des Lautes /v/ bei 5,5 kHz erkennen. Die Resultate lassen vermuten, daß die Fensterfunktion einen stärkeren Einfluß hat als die Segmentierung. Der auditive Eindruck der synthetisierten Beispiele legt nahe, daß die Gewichtung mit dem Hann-Fenster bei der Analyse vorteilhaft ist im Gegensatz zur Gewichtung mit der Rechteckfunktion. Bei Verwendung eines Hann-Fensters bei der Analyse kann eine grundperiodensynchrone oder eine Segmentierung mit fester Blocklänge verwendet werden. Für die grundperiodensynchrone Segmentierung müssen die Perrioden bereits markiert vorliegen, was zusätzlichen Aufwand bedeutet. Deshalb könnte es für die Anwendung interessant sein, daß sich auch mit fester Blocklänge eine verhältnismäßig gute Sprachqualität erzielen läßt.

Die harte Verkettung von Diphonen kann noch stärkere Unstetigkeiten hervorrufen, wie das folgende Beispiel der zu synthetisierenden Äußerung „vielleicht“ zeigt. Bild 3a zeigt die Betragsgänge der Verkettung ohne Flächenübergang im Diphthong [aI]. Diese Verkettung kommt durch die Diphone [l-aI] und [aI-C] zustande. Die Unstetigkeit in Bild 3a ist durch den

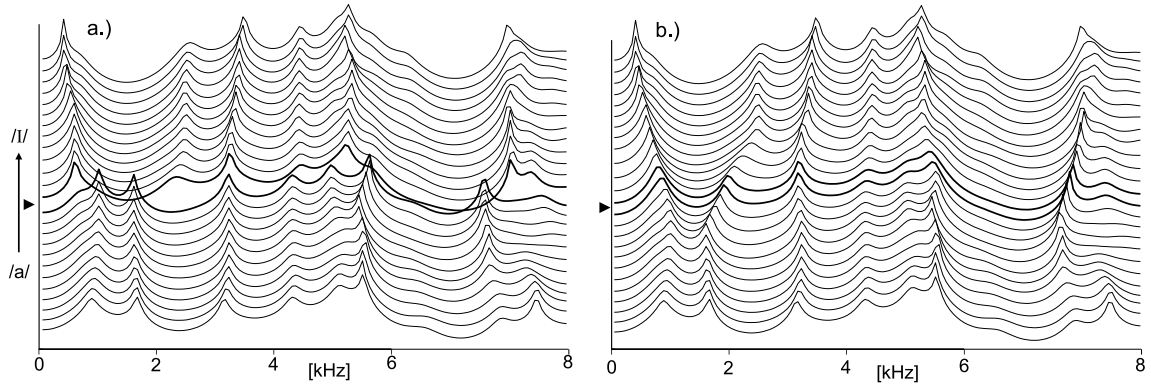

Bild 3 – Betragsgänge der Verkettung der Diphone [l-aI] und [aI-C] in /aI/ für die Äußerung „vielleicht“: (a) ohne Flächenübergang, (b) mit Flächenübergang an der Konkatenationsstelle.

vorliegenden Laut bedingt, da bei einem Diphthong die Lautmitte selbst einen Übergang darstellt und die Konkatenationsstelle somit in einem instationären Bereich liegt. Bild 3b zeigt die Betragsgänge nach der Verkettung derselben Diphone mit einem Flächenübergang von $M = 4$. Durch den Übergang der Modellparameter verlaufen die Formanten nun stetig von /a/ zu /I/, insbesondere zu sehen an den ersten beiden Formanten.

Neben den stimmhaften Lauten können auch stimmlose Frikative mit dem Rohrmodell analysiert und synthetisiert werden. Bei der Synthese der stimmlosen Frikative wird, wie schon erwähnt, weißes Rauschen für die Systemanregung verwendet. Für die Synthese der Äußerung „vielleicht“ kommen die Frikative /f/ und /C/ vor. Das Spektrogramm der synthetisierten Äußerung ist in Bild 4 zu sehen. Die Flächensätze an den Konkatenationsstellen im Laut /f/ und /C/ sind genauso wie die stimmhaften Laute durch einen Übergang mit $M = 4$ entsprechend (1) geglättet. Die Frikative weisen trotz der Verkettung einen ausgeprägten stationären Charakter auf. Die Verkettungsstellen im

stimmhaften Abschnitt sind fast nicht zu erkennen. Hinter dem Laut /C/ ist eine Pause eingefügt, an die das originale Sprachsignal des Explosivs /t/ anschließt.

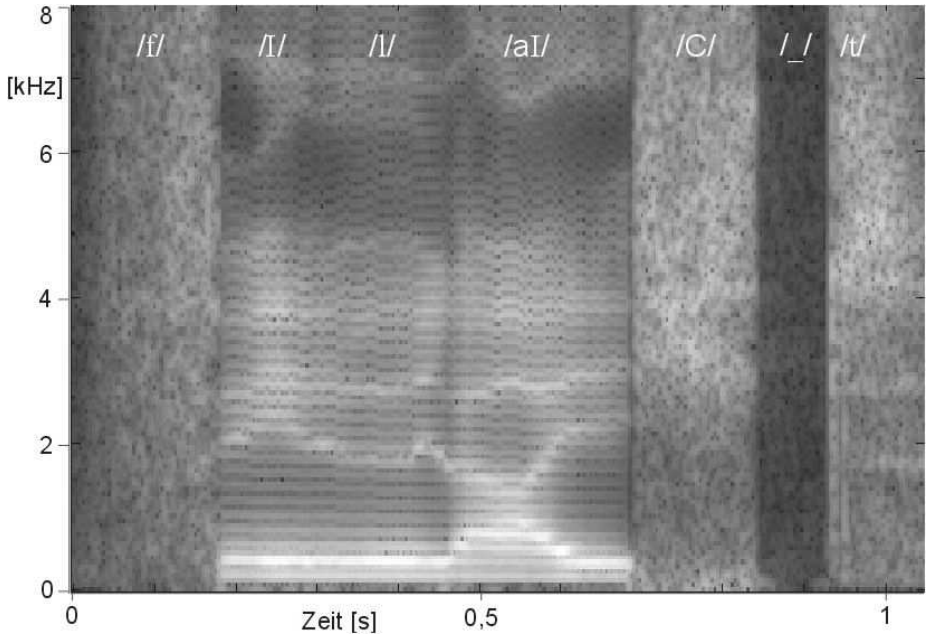

Bild 4 – Spektrogramm der synthetisierten Äußerung „vielleicht“.

4 Zusammenfassung

In diesem Beitrag konnte gezeigt werden, wie das verlustbehaftete Rohrmodell für eine akustische Synthese eingesetzt werden kann. Für die Realisierung der Synthese werden Modellparametersätze von analysierten Diphonen verkettet. Die Lautdaueranpassung kann durch Verdoppeln oder Auslassen von einzelnen Parametersätzen vorgenommen werden. Die Grundfrequenzänderung ist ebenfalls einfach zu realisieren, da eine Modellanregung verwendet wird, die unabhängig von den analysierten Sprachsignalen ist. Durch die Verkettung von Parametersätzen unterschiedlicher Diphone ergeben sich Unstetigkeiten an den Konkatenationsstellen, die durch einen linearen Flächenübergang beseitigt werden können. Weiterhin wurden verschiedene Analysetechniken der Diphone und deren Auswirkungen auf die Synthese diskutiert. Die erzielte Sprachqualität der synthetisierten Beispiele kann für eine parametrische Synthese als gut eingestuft werden. Dabei muß allerdings berücksichtigt werden, daß bisher nur ein kleiner Teil der Diphondatenbank herangezogen wurde, so daß noch keine vollständige Bewertung möglich ist. Eine Verbesserung der Synthesequalität kann unter Umständen durch ein anderes Anregungsmodell erreicht werden. Für die untersuchten Beispiele der Diphonsynthese stellen sich die vielen Konkatenationsstellen sowie die hohe Grundfrequenz der weiblichen Stimme als relativ unproblematisch dar.

Herr Dr. Fred Englert (früher Institut für Phonetik, J.W.Goethe-Universität Frankfurt am Main) hat uns die Diphondatenbank de1 für Untersuchungen zur Verfügung gestellt; dafür sei an dieser Stelle gedankt.

Literatur

[1] Dutoit, T.: “An Introduction to Text-to-Speech Synthesis”, Kluwer Academic Publishers, Dordrecht/Boston/London, 1997.

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

[3] Schnell, K.; Lacroix, A.: “Sprachanalyse und –erzeugung mit verlustbehafteten zeitdiskreten Rohrmodellen”, Tagungsband ESSV-2003, Karlsruhe, Web-Verlag Dresden pp. 196-202.

[4] Laine, U.K.: “Modeling of lip radiation impedance in the z-domain”, Proc. ICASSP'82, Paris 1982, pp. 1992-1995.

[5] Schnell, K.: “Rohrmodelle des Sprechtraktes – Analyse, Parameterschätzung und Syntheseexperimente”, Dissertation J.W.Goethe-Universität Frankfurt am Main 2003, http://dbib.uni-frankfurt.de .

[6] Englert, F.: “Acquisition of a Diphone Database for German” In: Wodarz, H.-W. (Ed.): Forum Phoneticum 63, Speech Processing, Hector-Verlag Frankfurt am Main, 1997, pp. 23–32.