Modellbeschreibung des stimmhaften Anregungssignals für die Spracherzeugung

Karl Schnell, Arild Lacroix

Institut für Angewandte Physik, Goethe-Universität Frankfurt,
Max-von-Laue-Straße 1, D-60438 Frankfurt am Main
Email: schnell@iap.uni-frankfurt.de

Einleitung

Für die modellbasierte Spracherzeugung wird die stimmhafte Anregung in der Regel durch ein periodisches Anregungssignal beschrieben. Neben der Impulsfolge als einfachstes stimmhaftes Anregungssignals werden auch parametrisierte Modelle des glottalen Flusses bzw. dessen Ableitung verwendet, wie z.B. das Rosenberg oder LF-Modell [1]. Diese Modelle stellen idealisierte glottale Signalverläufe dar, wodurch synthetisierte Sprachsignale etwas künstlich klingen können. Eine Erweiterung des LF-Modells durch Rauschanteile ist in [2] zu finden. Im Gegensatz zu den vereinfachten Glottissignalverläufen des Rosenberg oder LF-Modells kann ein weniger eingeschränktes Anregungsmodell, dessen Parameter aus dem Sprachsignal geschätzt werden, Vorteile bringen durch eine genauere Modellierung. In [3], [4] wird eine Periodenmodellierung des tiefpassgefilterten Residualsignals durch eine Polynomapproximation diskutiert. Durch diesen Ansatz wird allerdings nur der glatte Verlauf der Anregungsperioden berücksichtigt. Dies führt insbesondere für den Bereich der Glottisschließung zu einer ungenauen Modellierung. Daher wird in dem hier vorgestellten Ansatz die Polynommodellierung nur auf Periodenabschnitte angewendet [5]. Weiterhin wird auch der Approximationsfehler der Modellierung berücksichtigt, da dieser auch die rauschartigen Komponenten der stimmhaften Sprache enthält.

Anregungsmodell

Der Einfluss des Sprechtraktes kann bekanntlich durch eine inverse Filterung mittels linearer Prädiktion aus dem Sprachsignal weitgehend beseitigt werden. Das resultierende Residualsignal beschreibt ein Anregungssignal im Sinne eines Quelle-Filter Modells und ist Träger der Grundfrequenz. Für die Modellierung des Anregungssignals verwendet der vorgestellte Ansatz, statt dem Residualsignal selbst, eine Tiefpass gefilterte Darstellung g des Residualsignals, die als Schätzung des glottalen Flusses interpretiert werden kann. Das Verfahren zerlegt das tiefpassgefilterte Residualsignal g in abschnittsweise glatte Verläufe, die durch ein Polynommodell approximiert werden. Der Approximationsfehler des Polynommodells sowie der Zeitabschnitt der Glottisschließung werden hingegen durch Zeitsignale repräsentiert. Diese Aufspaltung bietet für eine Modifikation der Grundfrequenz Vorteile, da der glatte Verlauf und die Fluktuationen bzw. Unstetigkeiten auf unterschiedliche Weise an die neue Grundperiodenlänge angepasst werden können. Bei einer geeigneten Längenanpassung der einzelnen Komponenten und anschließender Zusammensetzung ergeben sich auch bei größeren Grundfrequenzänderungen Anregungssignale, die zu vergleichsweise natürlich klingender Sprache führen. In Abbildung 1 ist die Zerlegung des Signals g dargestellt. Das Signal $g = h_\mathrm{T} * r$ wird durch eine Tiefpassfilterung mittels der Impulsantwort h_T des Tiefpasses H_T gewonnen. H_T besitzt zwei reelle Polstellen nahe Eins. Der Bereich der Glottisschließung wird für die i-te Periode durch den Signalabschnitt bzw. Signalvektor $\boldsymbol{y}^i = \left(y^i(1), y^i(2), \ldots y^i(L_y)\right)$ dargestellt, der übrige Abschnitt der Periode wird durch den Signalvektor $\boldsymbol{x}^i = \left(x^i(1), x^i(2), \ldots x^i(L_x^i)\right)$ dargestellt. Die i-te Periode wird damit durch den zusammengesetzten Signalvektor $\boldsymbol{g}^i = \left[\boldsymbol{y}^i, \boldsymbol{x}^i\right]$ repräsentiert wie in Abb. 1

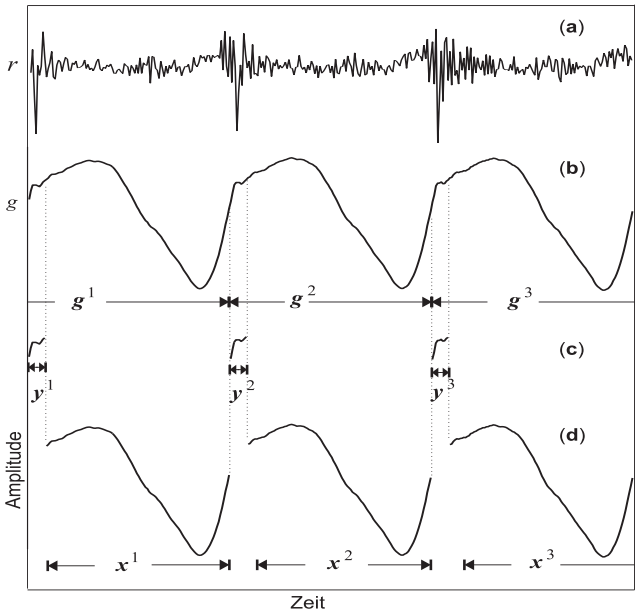

Abbildung 1: Zerlegung der tiefpassgefilterten Residualperioden: (a) Residualsignal r; (b) tiefpassgefiltertes Residualsignal g; (c) Abschnitte $\boldsymbol{y}^i$ der Glottisschließung; (d) übrige Abschnitte $\boldsymbol{x}^i$.

gezeigt. Der glatte Verlauf der Signalabschnitte $\boldsymbol{x}^i$ wird durch eine Polynomapproximation $\hat{\boldsymbol{x}}^i$ bestimmt. Der Approximationsfehler $\boldsymbol{e}^i = \boldsymbol{x}^i - \hat{\boldsymbol{x}}^i$ enthält die Fluktuationen des zugehörigen Zeitabschnitts, die auch die Rauschkomponente der Anregung beinhalten. Für eine Grundfrequenzänderung wird der Signalvektor $\boldsymbol{y}^i$ der Glottisschließung unverändert übernommen, während der Bereich des Vektors $\boldsymbol{x}^i$ gestaucht oder gestreckt wird entsprechend der neuen Grundperiodenlänge. Das Polynommodell kann für diesen Zweck interpoliert werden, so dass der Signalvektor $\hat{x}^i(1 \ldots L_x^i)$ mit der Länge L_x^i in den Vektor $\hat{x}_\mathrm{neu}^i(1 \ldots L_{x\mathrm{neu}}^i)$ der neuen Länge $L_{x\mathrm{neu}}^i$ konvertiert

wird. Die Längenanpassung des Fehlervektors $\boldsymbol{e}^i$ wird durch eine OLA-Technik vorgenommen. Dafür wird der Vektor in drei Bereiche zerlegt, die überlappend ineinander oder auseinander verschoben werden. Der mittlere Bereich sollte dabei möglicht die Glottisöffnung beinhalten. Der dadurch in der Länge veränderte Fehlervektor $\boldsymbol{e}^i_{\text{neu}}$ der Länge $L^i_{x\text{neu}}$ wird dem Vektor $\hat{x}^i_{\text{neu}}$ überlagert, der wiederum mit dem unveränderten Abschnitt $\boldsymbol{y}^i$ die neue Grundperiode $\boldsymbol{g}^{\text{i}}_{\text{neu}}$ bildet, wie in (1) aufgezeigt:

$$\boldsymbol{g}^i = [\boldsymbol{y}^i\ ,\ \hat{\boldsymbol{x}}^i + \boldsymbol{e}^i]$$

$$\text{mit}\quad \hat{x}^i(n) = \sum_{k=0}^{N} p^i_k \cdot \left(\frac{n}{L^i_x}\right)^k \quad n = \left(1 \ldots L^i_x\right)$$

$$\downarrow \qquad (f_0\text{-Änderung}) \qquad (1)$$

$$\boldsymbol{g}^{\text{i}}_{\text{neu}} = \left[\boldsymbol{y}^i, \hat{\boldsymbol{x}}^i_{\text{neu}} + \lambda \cdot \boldsymbol{e}^i_{\text{neu}}\right]$$

$$\text{mit}\quad \hat{x}^i_{\text{neu}}(n) = \sum_{k=0}^{N} p^i_k \cdot \left(\frac{n}{L^i_{x\text{neu}}}\right)^k \quad n = \left(1 \ldots L^i_{x\text{neu}}\right).$$

Die zusammengesetzten Grundperioden $[\ldots \boldsymbol{g}^{\text{i}}_{\text{neu}}, \boldsymbol{g}^{\text{i+1}}_{\text{neu}}, \ldots]$ stellen das Signal g_{neu} dar, das eine grundfrequenzveränderte Version des Signals *g* ist. Durch Filterung mit dem inversen System H^{-1}_{T} des Tiefpasses gelangt man wieder zu einer Residualsignaldarstellung $r_{\text{neu}} = h^{-1}_{\text{T}} * g_{\text{neu}}$. Mit dem Parameter λ in (1) können die Einflüsse des Rauschanteils und der Glottisöffnung abgeschwächt oder verstärkt werden im Vergleich zur Normaleinstellung $\lambda = 1$. Damit kann neben der Grundfrequenz auch die Stimmqualität beeinflusst werden. Die Analysen zeigen, dass insbesondere für stimmhafte Frikative der Approximationsfehler $\boldsymbol{e}$ für die Anregung wichtig ist, da er das Friktionsrauschen enthält. Abbildung 2 zeigt beispielhaft grundfrequenzveränderte Signale die aus der Analyse des Schwa-Lautes gewonnen wurden. Abb. 2 (a)-(b) zeigt die Residualsignale r_{neu} für Parameterwerte λ von 1 und 0,25. Durch Filterung des Residualsignals r_{neu} mit dem Nur-Pole Modell *H* der linearen Prädiktion des Schwa-Lautes kann ein Sprachsignal mit der gewünschten Grundfrequenz generiert werden. Die Graphen 2 (c) und (d) sind die mit *H* gefilterten Signale aus (a) und (b). Die Auswirkung der Parametereinstellung für λ ist in der Residualdarstellung deutlicher zu sehen als im synthetisierten Sprachsignal. Für die synthetisierten Signale der Graphen (e) und (f) ist die Grundfrequenz tiefer eingestellt mit einem entsprechend höheren Wert für $L^i_{x\text{neu}}$. Betragspektren der Signale aus (c) und (e) sind in den Graphen (g) und (h) dargestellt. Die Spektren sind aus einem Signalabschnitt von exakt sechs Perioden mit Hamming Fenster ermittelt worden. Es ist zu erkennen, dass die Spektren wie bei natürlicher Sprache im unteren Frequenzbereich stark harmonisch sind, während sie im oberen Frequenzbereich zunehmend unharmonischer werden. Es ist zu beachten, dass die Grundfrequenz der synthetisierten Signale (a) bis (f) konstant eingestellt ist ohne irgendwelche Grundfrequenzschwankungen. Abweichungen zur exakten Periodizität tragen bekanntlich zur Natürlichkeit der Sprache bei. Die Untersuchungen zeigen, dass die mit Hilfe der grundfrequenzveränderten Residualsignale erzeugten Sprachsignale eine hohe Natürlichkeit aufweisen, so dass sich daraus eine gewinnbringende Anwendung für die modellbasierte Sprachsynthese ergibt.

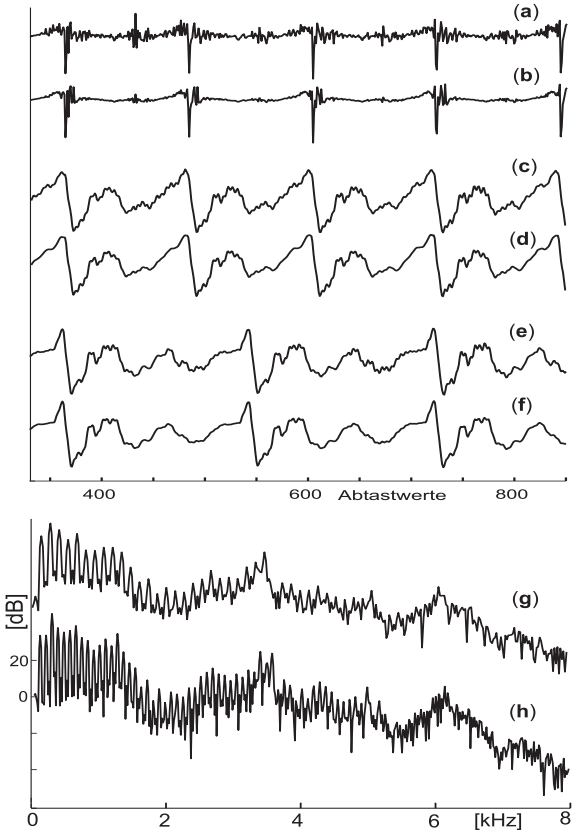

Abbildung 2: Grundfrequenzveränderte Signale des Schwa-Lautes: Residualsignal für Periodenlänge 120 mit (a) $\lambda = 1$ und (b) $\lambda = 0{,}25$; (c),(d) sind die mit einem Sprechtraktmodell *H* gefilterten Signale von (a),(b); (e),(f) wie (a),(b) jedoch für Periodenlänge 180; (g),(h) sind Spektren der Signale aus (c),(e).

Literatur

[1] G. Fant, J. Liljencrantz, Q. Lin: "A Four Parameter Model of Glottal Flow", STL-QPSR, 2-3, pp. 119-156, 1985.

[2] C. Gobl: "Modelling Aspiration Noise During Phonation Using the LF Voice Source Model", Proc. INTERSPEECH'06, Pittsburgh, pp. 965-968, 2006.

[3] D.G. Childers, T.H. Hu: "Speech Synthesis by Glottal Excited Linear Prediction", J. Acoust. Soc. Am. 96(4), pp. 2026-2036, 1994.

[4] P.H. Milenkovic: "Voice Source Model for Continuous Control of Pitch Period", J. Acoust. Soc. Am. 93(2), pp. 1087-1096, 1993.

[5] K. Schnell: "Pitch Modification of Speech Residual Based on Parameterized Glottal Flow with Consideration of Approximation Error", Proc. ICASSP'06, Toulouse, pp. 737-740, 2006.