ÜBER DIE RELEVANZ VON ALTERNATIVEN LP-METHODEN FÜR DIE SPRACHSYNTHESE

Erhard Rank

Institut für Nachrichtentechnik und Hochfrequenztechnik, TU Wien
Gusshausstrasse 25/E389, A–1040 Wien
Email: erank@nt.tuwien.ac.at

Für Sprachsynthese werden in zunehmenden Maße wieder Algorithmen verwendet, die das Quelle-Filter Modell der menschlichen Sprachentstehung berücksichtigen. Dabei wird mit einem Filter der Einfluss des Vokaltrakts auf das Sprachsignal angenähert. I.a. wird für die Realisierung dieses Filter lineare Prädiktion (LP) verwendet. Zur Synthese kann das LP-Filter mit dem Restsignal des inversen Filters angeregt werden (engl.: *residual excited linear prediction*, RELP).

In diesem Beitrag werden nun verschiedene Realisierungen des LP-Filters und unterschiedliche Methoden zur Ermittlung der Filterkoeffizienten untersucht, mit besonderem Augenmerk auf die Auswirkungen bei Prosodieveränderungen in der Sprachsynthese.

1 EINLEITUNG

In den letzten Jahren hat die Sprachsynthese durch die Verwendung von großen Sprachdatenmengen und Synthese durch Konkatenation im Zeitsignalbereich vermehrte Akzeptanz gefunden. Trotzdem wird heute zusätzlich oft das schon von den sog. Formant-Synthesizern bekannte Quelle-Filter-Modell der menschlichen Sprachentstehung berücksichtigt – meist in Form der linearen Prädiktion (*linear prediction*, LP) [1]. Dabei wird mittels eines digitalen Filters niedriger Ordnung die spektrale Einhüllende des Sprachsignales angenähert, und so der Einfluß des Vokaltraktes auf das Signal der Stimmbandschwingungen abgeschätzt.

Die dadurch erzielte Auftrennung in Quellsignal (Stimmbandschwingungen) und Filter (Vokaltrakt) bietet die Möglichkeit einer höheren Qualität der Synthese bei prosodischen Manipulationen und bei der Glättung von Konkatenationsstellen durch gezielte Beeinflussung der Parameter. So kann z.B. der unterschiedliche Frequenzinhalt an einer Konkatenationsstelle durch Interpolation des LP-Filters geglättet werden [2].

Im folgenden wird das Verhalten verschiedener Realisierungen des LP-Filters beim Umschalten der Filterkoeffizienten (Kapitel 2) sowie die Anwendung von unterschiedlichen Methoden zur Ermittlung der Filterkoeffizienten (Kapitel 3) insbesondere bei Prosodiemanipulation untersucht. Da die betrachteten Methoden zur Prosodiemodifikation auf grundfrequenzsynchroner Segmentierung basieren, sind die im folgenden beschriebenen Experimente auf grundfrequenzsynchrone Segmentierung beschränkt.

2 LP-FILTER-REALISIERUNGEN

Für die Realisierung des LP-Synthese-Filters gibt es unterschiedliche Möglichkeiten: Es können sowohl rekursive Transversalfilter als auch sogenannte „Lattice"-Filter (von engl. *lattice* = Gitter) verschiedener Bauart – die aus dem physikalischen Modell der Wellenausbreitung im Vokaltrakt abgeleitet werden – zum Einsatz kommen [1]. Mit beiden Filtertypen werden die Resonanzen des Vokaltrakts modelliert (*all-pole* Modell) und es können die selben Frequenzgänge erzielt werden. Folgende Realisierungen wurden betrachtet: Transversalfilter, Lattice-Filter mit zwei Multiplizierern, Kelly-Lochbaum Lattice-Filter bzw. das vom Verhalten her identische Lattice-Filter mit einem Multiplizierer und das normalisierte Lattice-Filter (vier Multiplizierer) [1].

Mit jedem dieser Filtertypen kann das Sprachsignal ohne Prosodiemanipulationen unter Verwendung des passenden inversen Analyse-Filters fehlerfrei resynthetisiert werden. Bei Prosodiemanipulationen allerdings, bei denen notwendigerweise einzelne Segmente des Signales ausgelassen oder wiederholt werden müssen, ergibt sich durch das Ausschwingen des LP-Filters ein Fehler, da i.a. nicht das gleiche Ausschwingverhalten wie beim jeweils vorhergehenden Segment im Orginalsignal erzielt wird. Dieser Fehler tritt insbesondere dann auf, wenn die LP-Filterkoeffizienten der Einfachheit halber nur an den Segmentgrenzen umgeschaltet und nicht bei jedem Signalabtastwert interpoliert werden.

Zur Untersuchung dieses Fehlers wurde hier das Anregungssignal des LP-Filters jeweils am Ende eines Segments abgeschaltet und das Auschwingverhalten ohne Filterkoeffizientenänderung und mit Umschaltung der Filterkoeffizienten betrachtet. In Abbildung 1 ist ein Beispiel für diesen Ausschwingvorgang für die unterschiedlichen Filtertypen dargestellt. Die verschiedenen Filtertypen erzeugen bei Umschalten der Koeffizienten unterschiedlich starke, abklingende Schwingungen bei den durch die neuen Koeffizienten gegebenen Resonanzfrequenzen.

Eine statistische Auswertung über einen kleinen Korpus[1] von Einzelvokalen und Sätzen von drei verschiedenen männlichen Sprechern ist in Tabelle 1 dargestellt. Die Zahlen geben das Verhältnis zwischen Energie des Fehlersignals und Energie des Auschwingvorganges bei nicht umgeschaltetem LP-Filter an. Die Werte in der ersten Zeile gelten für einfache LP-Analyse über ein Segment, die in der zweiten Zeile für LP-Analyse mit Korrektur des Einflusses auf die jeweils nachfolgenden Segmente nach [3]. Je kleiner der angebene Wert ist, desto geringer sind die Auswirkungen der Koeffizientenveränderung.

Bei einfacher LP-Analyse ergibt sich der beste Wert für die Transversalfilterstruktur, bei den Lattice-Filtern liegt das normalisierte Filter am besten. Das normalisierte Lattice-Filter gibt auch den mit Abstand besten Wert, wenn bei der Analyse der Einfluss auf die Folgesegmente berücksichtigt wird. Das kann dadurch erklärt werden, dass die Variablen im normalisierten Lattice-Filter Leistungswellen entsprechen, und die Gesamtenergie im Filter daher beim Umschalten der Koeffizienten erhalten bleibt [4].

3 ALTERNATIVE LP-METHODEN

Die Ermittlung der LP-Filterkoeffizienten erfolgt durch Schätzung der Autokorrelation bzw. der Kovarianz des Sprachsignals und nachfolgender Minimierung der Energie

[1] Es wurden 2534 stimmhafte Grundfrequenzperioden analysiert.

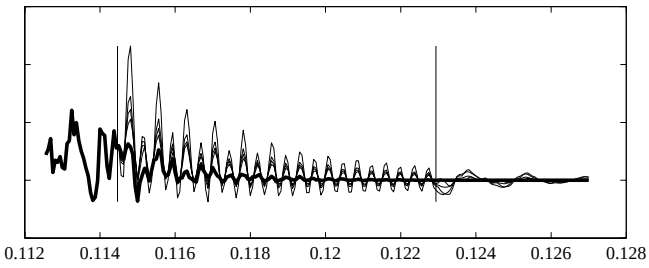

Abbildung 1: Transienten beim Umschalten der LP-Filterkoeffizienten. Die dick ausgezogene Linie stellt das Ausschwingen bei nicht umgeschalteten LP-Koeffizienten (für alle Filtertypen) dar; die nicht durchgezogenen Linien zeigen Transienten der verschiedenen Filtertypen beim Umschalten an den Segmentgrenzen (vertikale Linien).

2-Multi.	Normal.	KL (1-Multi.)	Transvers.
-4,249 dB	-4,537 dB	-4,102 dB	-4,980 dB
-3,608 dB	-6,073 dB	-4,360 dB	-4,292 dB

Tabelle 1: Mittelwerte der Fehler im Ausschwingvorgang beim Umschalten der LP-Filterkoeffizienten für die verschiedenen Filtertypen. Die Zahlen geben die Energie des Fehlersignales im Verhältnis zur Energie des Ausschwingvorganges bei nicht umgeschalteten LP-Filter an. Erste Zeile: einfache LP-Analyse über ein Segment; zweite Zeile: LP-Analyse mit Korrektur des Einflusses auf die Folgesegmente.

des Fehlersignals (Restsignal) entweder direkt als Matrixoperation oder implizit in einer Lattice-Filter Struktur[1, 5]. Dabei kann der betrachtete Signalabschnitt mit einer Fensterfunktion multipliziert werden, was eine Betonung des zentralen Signalbereichs bedeutet.

Wenn man näher an das Quelle-Filter Modell herankommen möchte, kann es vorteilhaft sein, die LP-Analyse auf jene Signalbereiche zu beschränken, in denen die Glottis geschlossen ist: Hier kann der Vokaltrakt als abgeschlossenes System (ohne Anregung und Einfluss des subglottalen Bereichs) angesehen werden, und das resultierende LP-Filter sollte damit dem Resonanzverhalten des Vokaltrakts alleine entsprechen.

Im folgenden wird ein Vergleich zwischen LP-Analyse mit und ohne Fensterung über eine Grundfrequenzperiode oder nur im Bereich der Glottisverschlussphase bei der Anwendung von Prosodiemanipulationen angestellt. Hierfür wird das Setting zur Evaluation von Sprachsynthesecodern verwendet, das im Rahmen von Cost258[2] entwickelt wird [6]. Die Aufgabenstellung ist, von einem mit monotoner Prosodie aufgenommenen Sprachsignal durch Manipulation von Lautdauer und Grundfrequenz möglichst gut an natürliche Sprachsignale mit unterschiedlicher Prosodie heranzukommen. Die Glottisverschlussphasen wurden für ausgewählte Samples im Korpus nach [7] ermittelt und in der Folge die LP-Analyse auf den typischen Bereich – nämlich 3 ms nach dem Glottisverschluss – beschränkt. Zur Prosodiemanipulation wurden SRELP [8, 9] und LP-PSOLA [10] verwendet.

Abbildung 2 zeigt die Ergebnisse für drei verschiedene Bewertungsverfahren (*Weighted Spectral Slope, Log-Likelyhood Ratio* und *Log-Area Ratio*), die jeweils ein Abstandsmaß zwischen synthetischem und natürlichem Sprachsignal angeben. Interessanterweise ergibt sich der geringste Abstand (beste Qualität) durchgehend bei LP-Analyse über das gesamte Segment ohne Fensterung. Die Mittelwerte für LP-PSOLA liegen bei Bewertung durch den *Weighted Spectral Slope* etwas besser als für SRELP, sonst zeigt sich kein signifikanter Unterschied zwischen den beiden Verfahren.

[2] COST Aktion 258 der EU: „The Naturalness of Synthetic Speech“.

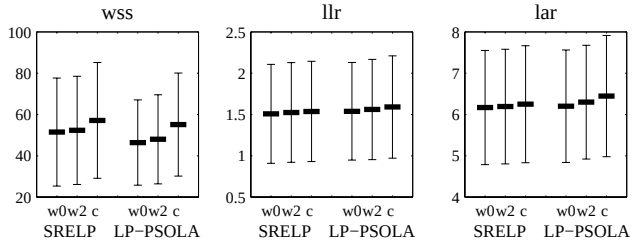

Abbildung 2: Bewertung (Mittelwerte und Bereich für Standardabweichung) von Prosodiemanipulationen mit SRELP und LP-PSOLA für LP-Analyse über ein ganzes Segment ohne (w0) und mit (w2) Fensterung (Hanning-Fenster) und über den Bereich des Glottisverschlusses (c). Die Bewertung enthält die Abstandmaße *Weighted Spectral Slope (wss)*, *Log-Likelyhood Ratio (llr)* und *Log-Area Ratio (lar)*.

4 ZUSAMMENFASSUNG

Untersucht wurde das Verhalten von LP-Filtern bei Prosodiemanipulationen, wie sie zur Sprachsynthese i.a. benötigt werden. Im speziellen wurden die transienten Fehler analysiert, die beim Umschalten der Filterkoeffizienten an Segmentgrenzen auftreten; hier zeigte sich, dass bei einfacher LP-Analyse über eine Grundfrequenzperiode die Transveralfilterstruktur, bei Korrektur des Ausschwingvorgangs bei der Analyse die normalisierte Lattice-Filter-Struktur die kleinsten Fehler ergibt. Weiters wurde mittels objektiver Abstandsmaße der Einfluss von Fensterung bzw. Beschränkung auf die Glottisverschlussphase bei der LP-Analyse beurteilt; die Ergebnisse deuten auf eine minimale Verschlechterung der Qualität der synthetischen Sprache bei Analyse in der Glottisverschlussphase, die allerdings im Bereich der Standardabweichung der Ergebnisse liegt.

Literatur

[1] J. D. Markel und A. H. Gray, Jr.: *Linear Prediction of Speech*. Berlin, Heidelberg, New York: Springer, 1976.

[2] E. Rank: Exploiting Improved Parameter Smoothing within a Hybrid Concatenative/LPC Speech Synthesizer, in *Proceedings of Eurospeech'99*, Budapest, S. 2339–2342, 1999.

[3] A. Ferencz, I. Nagy, T.-C. Kovács, T. Raţiu und M. Ferencz: On a Hybrid Time Domain-LPC Technique for Prosody Superimposing used for Speech Synthesis, in *Proceedings of Eurospeech'99*, Vol. 4, Budapest, S. 1831–1834, 1999.

[4] G. Kubin: Wave Digital Filters: Voltage, Current, or Power Waves?, in *Proceedings of ICASSP'86*, Tampa (FL), März 1986.

[5] M. L. Honig und D. G. Messerschmitt: *Adaptive Filters: Structures, Algorithms, and Applications*. Boston-The Hague-London-Lancaster: Kluwer Academic Publishers, 1984.

[6] E. Keller, Hrsg.: *Cost258 – Working Papers*. Springer, 2000. In Vorbereitung.

[7] D. Y. Wong, J. D. Markel und J. Augustine H. Gray: Least Squares Glottal Inverse Filtering from the Acoustic Waveform, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-27, S. 350–355, aug 1979.

[8] B. E. Caspers und B. S. Atal: Changing Pitch and Duration in LPC Synthesized Speech using Multipulse Excitation, *JASA*, Vol. S5(A), S. 73–, 1983.

[9] M. Macchi, M. J. Altom, D. Kahn, S. Singhal und M. Spiegel: Intelligibility as a Function of Speech Coding Method for Template-Based Speech Synthesis, in *Proceedings of Eurospeech'93*, Berlin, Germany, S. 893–896, 1993.

[10] E. Moulines und F. Charpentier: Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones, *Speech Communications*, Vol. 9, S. 452–467, 1990.