EIN ZEITVARIABLER LINEARER PRÄDIKTIONSALGORITHMUS FÜR DIE SPRACHVERARBEITUNG

Karl Schnell und Arild Lacroix

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

Abstract: In diesem Beitrag wird eine zeitvariable lineare Prädiktion vorgestellt, die für die Analyse von instationären Prozessen der Sprachproduktion Verwendung findet. Die Instationarität des Sprachsignals wird maßgeblich durch zwei Effekte bei der Sprachproduktion hervorgerufen. Der erste Effekt wird durch die artikulationsbedingten Vokaltraktbewegungen verursacht, während der zweite Effekt durch die Stimmbandschwingungen der stimmhaften Anregung gegeben ist. Letztere stellen im Vergleich zu den Vokaltraktbewegungen wesentlich schnellere Änderungen dar, die zusätzlich als periodisch angenommen werden können. Eine Analyse mit der zeitvariablen Prädiktion kann daher beide oder nur einen der genanten Effekte erfassen. Der hier vorgestellte Ansatz der zeitvariablen Prädiktion zerlegt das Sprachsignal ähnlich der zeitinvarianten Prädiktion in Signalabschnitte. Im Gegensatz zum zeitinvarianten Ansatz wird die Instationarität durch eine lineare Zeitentwicklung der Prädiktorkoeffizienten innerhalb eines Abschnittes berücksichtigt. Um einer kontinuierlichen Bewegung der Artikulatoren Rechnung tragen zu können, wird eine stetige Entwicklung der Koeffizienten auch über die Abschnittsgrenzen hinaus gefordert, die durch einen stetigen Koeffizientenübergang zwischen den Abschnitten erfüllt wird. Das auf diese Weise gestellte Prädiktionsproblem lässt sich auf die Lösung eines linearen Gleichungssystems zurückführen. Damit können wie für die zeitinvariante lineare Prädiktion auch für die zeitvariable Prädiktion die optimalen Koeffizienten bezüglich eines minimalen Prädiktionsfehlers analytisch bestimmt werden.

1 Einleitung

Die lineare Prädiktion stellt für die Signalverarbeitung einen Basisalgorithmus zur Verfügung, mit deren Hilfe die Koeffizienten eines Nur-Pole Modells analytisch geschätzt werden können. Dadurch ergeben sich in der Sprachverarbeitung bekanntlich viele Anwendungsfelder [1, 2]. Insbesondere kann damit eine Schätzung der Sprechtraktresonanzen aus dem Sprachsignal effizient realisiert werden, wodurch auch eine Quelle-Filter Trennung erzielt werden kann. Für die Sprachanalyse wird das instationäre Sprachsignal in der Regel in kleinere Segmente zerlegt, da die gewöhnliche zeitinvariante Prädiktion ein stationäres Signal voraussetzt. Für die Analyse sollten die Segmente nicht zu klein gewählt werden, da sonst nicht genügend Abtastwerte für die Schätzung zur Verfügung stehen. Umgekehrt sollten die Segmente allerdings auch nicht zu groß gewählt werden, da sonst die Stationarität immer weniger erfüllt werden kann. Um die Instationarität bei der Analyse explizit zu berücksichtigen, muss ein Schätzverfahren verwendet werden, dass auch ein zeitvariables System zu Grunde legt. Für die zeitvariable Analyse existieren mehrere Ansätze. Die Klasse der adaptiven Filter stellt z.B. einen allgemein Ansatz für die zeitvariable Analyse dar [3]. Weiterhin kann auch ein statistischer Ansatz gewählt werden [4]. Eine an das Problem angepasste Methode ist durch eine lineare Prädiktion mit zeitvariablen Koeffizienten auf Grundlage vordefinierter Funktionen gegeben [5, 6]. Dabei wird die zeitliche Entwicklung der Koeffizienten auf den Raum der definierten Zeitfunktionen eingeschränkt, wodurch das

Problem mathematisch handhabbar wird. Der in diesem Beitrag vorgestellte Prädiktionsalgorithmus basiert auf diesen Ansatz. Im Vergleich zu den bisherigen Lösungen dieses Ansatzes werden hier die Prädiktionskoeffizienten segmentweise geschätzt mit der zusätzlichen Nebenbedingung eines kontinuierlichen Koeffizientenverlaufs zwischen den Segmentgrenzen [7].

2 Zeitvariable Prädiktion

Die lineare Prädiktion schätzt einen Signalwert $x(n)$ durch eine Linearkombination vergangener Werte $x(n-k)$. Der Prädiktionsfehler $e(n) = x(n) - \hat{x}(n)$ wird durch die Differenz zwischen dem tatsächlichen Wert $x(n)$ und dem geschätzten Wert $\hat{x}(n)$ repräsentiert. Bei der zeitvariablen Prädiktion sind die Koeffizienten $a_i(n)$ zeitabhängig gewählt mit

$$\hat{x}(n) = \sum_{i=1}^{N} a_i(n) \cdot x(n-i) \,. \tag{1}$$

Die Zeitabhängigkeit der Prädiktorkoeffizienten innerhalb eines Segments wird als linear angenommen, wodurch die Koeffizientenentwicklung einer Geradengleichung entspricht. Das *k*-te Segment verläuft von Index $m[k]$ bis $m[k+1]-1$ mit der Segmentlänge $l[k] = L[k]+1 = m[k+1]-m[k]$. Damit können die zeitvariablen Prädiktorkoeffizienten des *k*-ten Segments durch einen konstanten Anteil c_i^k und einer linearen Zeitentwicklung $u^k(n)$ zu

$$a_i(n) = c_i^k + d_i^k \cdot u^k(n) \qquad \text{mit} \;\; u^k(n) = (n - m[k]) / L[k] \tag{2}$$

dargestellt werden. Der Koeffizient d_i^k beschreibt dabei den Anteil des zeitvariablen linearen Anstiegs $u^k(n)$, der von Null bis Eins verläuft. Der Prädiktionsfehler resultiert damit zu

$$\begin{aligned} e(n) &= x(n) - \sum_{i=1}^{N} a_i(n) \cdot x(n-i) \\ e(n) &= x(n) - \sum_{i=1}^{N} \left(c_i^k \cdot x(n-i) + d_i^k \cdot u^k(n) \cdot x(n-i) \right). \end{aligned} \tag{3}$$

Werden die Segmentabschnitte als Vektoren interpretiert, so kann Gleichung (3) als Vektorgleichung

$$\boldsymbol{e}^k = \boldsymbol{x}_0^k - \sum_{i=1}^{N} \left(c_i^k \cdot \boldsymbol{x}_i^k + d_i^k \cdot \boldsymbol{w}_i^k \right) \tag{4}$$

geschrieben werden mit den Vektoren

$$\boldsymbol{x}_i^k = \left(x(m[k]-i),\, x(m[k]+1-i),\, \ldots,\, x(m[k]+L[k]-i) \right)^{\mathsf{T}},$$

$$\boldsymbol{e}^k = \left(e(m[k]),\, e(m[k]+1),\, \ldots,\, e(m[k]+L[k]) \right)^{\mathsf{T}},$$

$$\boldsymbol{w}_i^k = \boldsymbol{u}^k \otimes \boldsymbol{x}_i^k \;\; \text{und} \;\; \boldsymbol{u}^k = \left(0, \tfrac{1}{L[k]}, \tfrac{2}{L[k]}, \ldots \tfrac{L[k]-1}{L[k]}, 1 \right)^{\mathsf{T}}.$$

Der Index *k* stellt die Zugehörigkeit zum jeweiligen Segment dar und die Operation $\otimes$ führt eine elementweise Multiplikation mit $\boldsymbol{w} = \boldsymbol{u} \otimes \boldsymbol{x} \rightarrow w(n) = u(n) \cdot x(n)$ aus. Wird Gleichung (4) nach $\boldsymbol{x}_0^k$ aufgelöst, so können die Vektoren $\boldsymbol{x}_i^k$ und $\boldsymbol{w}_i^k$ als Basisvektoren interpretiert

werden, nach denen der Vektor $\boldsymbol{x}_0^k$ entwickelt wird. Für P Segmente ergeben sich dann die Gleichungen

$$\boldsymbol{x}_0^k = \sum_{i=1}^{N}\left(c_i^k \cdot \boldsymbol{x}_i^k + d_i^k \cdot \boldsymbol{w}_i^k\right) + \boldsymbol{e}^k \qquad \text{mit } k = 1\ldots P\,. \tag{6}$$

Die Gleichungen (6) stehen jeweils für P Segmente und werden durch die Stetigkeitsbedingung zwischen den Segmenten miteinander gekoppelt. Die Forderung eines stetigen Koeffizientenübergangs zwischen den Segmenten werden durch die Bedingungsgleichungen

$$a_i(m[k+1]) \;=\; a_i(m[k]+L[k]) \qquad \text{bzw.} \qquad c_i^{k+1} = \; c_i^k + d_i^k \tag{7}$$

erfüllt. An den Stetigkeitsbedingungen (7) ist zu erkennen, dass die Koeffizienten c_i^k mit $k > 0$ von den Koeffizienten d_i^k und c_i^1 abhängen, so dass nur letztere zu bestimmen sind. Werden die Kopplungen (7) für die Gleichungen (6) berücksichtigt, so folgt das Gleichungssystem

$$\boldsymbol{q}_0^0 = \sum_{i=1}^{N} c_i^1 \cdot \boldsymbol{q}_i^0 + \sum_{k=1}^{P}\sum_{i=1}^{N} d_i^k \cdot \boldsymbol{q}_i^k + \tilde{\boldsymbol{e}}^P \tag{8}$$

mit den Vektordefinitionen

$$\boldsymbol{q}_i^0 = \begin{pmatrix} \boldsymbol{x}_i^1 \\ \boldsymbol{x}_i^2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{x}_i^P \end{pmatrix},\; \boldsymbol{q}_i^1 = \begin{pmatrix} \boldsymbol{w}_i^1 \\ \boldsymbol{x}_i^2 \\ \boldsymbol{x}_i^3 \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{x}_i^P \end{pmatrix} \ldots\; \boldsymbol{q}_i^k = \begin{pmatrix} \boldsymbol{0}^1 \\ \vdots \\ \boldsymbol{0}^{k-1} \\ \boldsymbol{w}_i^k \\ \boldsymbol{x}_i^{k+1} \\ \vdots \\ \boldsymbol{x}_i^P \end{pmatrix} \ldots\; \boldsymbol{q}_i^P = \begin{pmatrix} \boldsymbol{0}^1 \\ \boldsymbol{0}^2 \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{0}^{P-1} \\ \boldsymbol{w}_i^P \end{pmatrix} \quad \text{und} \quad \tilde{\boldsymbol{e}}^P = \begin{pmatrix} \boldsymbol{e}^1 \\ \boldsymbol{e}^2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{e}^P \end{pmatrix}.$$

Der Aufbau eines Vektors $\boldsymbol{q}_i^k$ ist in der Weise gestaltet, dass in der k-ten Zeile der innere Vektor $\boldsymbol{w}_i^k$ steht, und weiterhin innere Nullvektoren $\boldsymbol{0}^k = (0,0,\ldots 0)^{\mathsf{T}}$ darüber und Vektorelemente $\boldsymbol{x}_i^k$ darunter stehen. Die dem Vektor $\boldsymbol{w}_i^k$ folgenden Elemente $\boldsymbol{x}_i^{k+j}$ mit $j = k+1\ldots P$ sind eine Folge des linearen Anstiegs $\boldsymbol{u}^k$, der von Null bis Eins verläuft. Da der lineare Anstieg mit Eins endet, wird der Endwert Eins den Prädiktorkoeffizienten der nachfolgenden Segmente dazuaddiert. Dadurch wird ein stetiger Segmentübergang der Koeffizienten realisiert, der durch die nachfolgenden Elemente $\boldsymbol{x}_i^k$ gewährleistet ist. Die Koeffizienten c_i^1 und d_i^k sind optimal gewählt, falls die Norm bzw. die Länge des Fehlervektors $|\tilde{\boldsymbol{e}}^P|$ minimal wird. Dies ist gerade bei einer Entwicklung des Vektors $\boldsymbol{q}_0^0$ nach den Basisvektoren $\boldsymbol{q}_i^k$ mit $i > 0$ gegeben, vergleichbar einem Regressionsproblem. Die Minimierung des Fehlervektors kann durchgeführt werden, indem das Basissystem $\boldsymbol{q}_i^k$ in ein orthogonales Basissystem $\boldsymbol{v}_m$ mittels des Gram-Schmidt Verfahrens überführt wird. Die optimalen Koeffizienten in Darstellung der orthogonalen Basis können mit Hilfe des Skalarprodukts $\langle\,|\,\rangle$ durch die verallgemeinerten Fourierkoeffizienten $b_m = \langle \boldsymbol{q}_0^0 \,|\, \boldsymbol{v}_m \rangle / |\,\boldsymbol{v}_m\,|^2$ bestimmt werden. Anschließend müssen die Koeffizienten b_m noch in die Koeffizienten der ursprünglichen Basis $\boldsymbol{q}_i^k$ zurück transformiert werden.

Die Rechenzeit für das Lösen der Gleichung (8) wächst stärker als linear mit der Anzahl der Segmente. Für die Berechnung vieler Segmente kann die Rechenzeit verkürzt werden, indem die Sequenz der Segmente in überlappende Subsequenzen zerlegt wird. Die Koeffizienten der ersten Subsequenz werden dann entsprechend (8) gelöst. Für die weiteren Subsequenzen wird der erste Anfangskoeffizientensatz c_i^1 nicht geschätzt, da der entsprechende Koeffizientensatz der letzten Subsequenz für c_i^1 verwendet wird. Bei einer Anzahl von 7 Segmenten pro Subsequenz und einem Überlapp von 3 Segmenten sind die Analyseergebnisse praktisch identisch mit denen von Gl. (8) für alle Segmente.

3 Analyse von Sprachsignalen

3.1 Segmentlänge für die zeitvariable Analyse

Bei der Wahl der Segmentlänge der zeitvariablen Prädiktion ist zu beachten, dass die Zeitvariabilität der Stimmbandschwingungen innerhalb einer Grundperiode vollständig zur Geltung kommt, während sich der Einfluss der Vokaltraktbewegungen fließend über die Periodengrenzen hinaus auswirkt. Daher kann durch Segmentlängen, die wesentlich größer sind als die Grundperiodenlänge, der Einfluss der Stimmbandschwingungen weitgehend unterdrückt werden.

3.2 Zeitvariable Präemphase

Um den Einfluss des Sprechtraktes von dem der Anregung und Abstrahlung zu separieren, wird gewöhnlich eine Präemphase verwendet. Diese wirkt sich insbesondere auf die Schätzung der Vokaltraktflächen aus [8] und kann adaptiv durch die lineare Prädiktion erster Ordnung realisiert werden [2]. Da der spektrale Abfall von der Stimmqualität und der Mundöffnungsfläche abhängig ist, welche sich während der Artikulation ändern können, ist die Verwendung einer zeitvariablen Präemphase sinnvoll. Dafür wird eine zeitvariable Prädiktion erster Ordnung mit $N = 1$ durchgeführt, womit der spektrale Abfall des Sprachspektrums durch eine reelle Polstelle modelliert wird [9]. Für eine genauere Modellierung des spektralen Abfalls kann die adaptive Präemphase wiederholt angewendet werden, so dass der spektrale Abfall durch mehrere reelle Polstellen approximiert werden kann. Das mit der Präemphase gefilterte Sprachsignal s' ist durch den Fehlervektor $\tilde{\boldsymbol{e}}^P$ der Prädiktion gegeben.

3.3 Analyse von Sprachsignalen

Nach der zeitvariabeln Präemphase wird das gefilterte Sprachsignal s' mit einer zeitvariablen Prädiktion höherer Ordnung analysiert, um den Einfluss des Sprechtraktes zu erfassen. Durch die zeitvariable Prädiktion wird jedem analysierten Segment k ein Koeffizientensatz zugeordnet, der aus den Koeffizienten c_i^k bestimmt wird. Die Analyseergebnisse zeigen, dass die Koeffizienten c_i^k selbst oftmals fluktuierende Modellschätzungen von Segment zu Segment erzeugen und daher inadäquat sind. Im Gegensatz dazu, führen die zeitlichen Mittelwerte der Koeffizienten

$$\overline{a}_i^k = c_i^k + c_i^{k+1}$$

innerhalb eines Segmentes k zu verwertbaren Analyseergebnisse. Die Koeffizienten $\overline{a}_i^k$ führen meistens zu zeitlich glatteren Modellverläufen als die Ergebnisse mit der herkömmlichen zeitinvarianten Prädiktion. Abbildung 1 zeigt die geschätzten Modellbetragsgänge der Äußerung [ta] ab dem Stimmeinsatz. Da der Lautübergang relativ

schnell artikuliert wird, wurden verhältnismäßig kurze Segmentlängen gewählt. Der analysierte Abschnitt beinhaltet die Formantabbiegungen infolge des Explosivs. Für die zeitinvariante Prädiktion wird die Kovarianzmethode verwendet, da sie insbesondere für kürzere Segmente bessere Ergebnisse liefert als die Autokorrelationsmethode. Es ist zu sehen, dass die zeitvariable Prädiktion einen kontinuierlicheren Verlauf der Modellbetragsgänge aufweist. Dieser glattere Verlauf ist eine Auswirkung der Stetigkeitsbedingungen von Gl (7), die einen kontinuierlichen Koeffizientenverlauf zwischen den Segmentgrenzen fordern. Durch

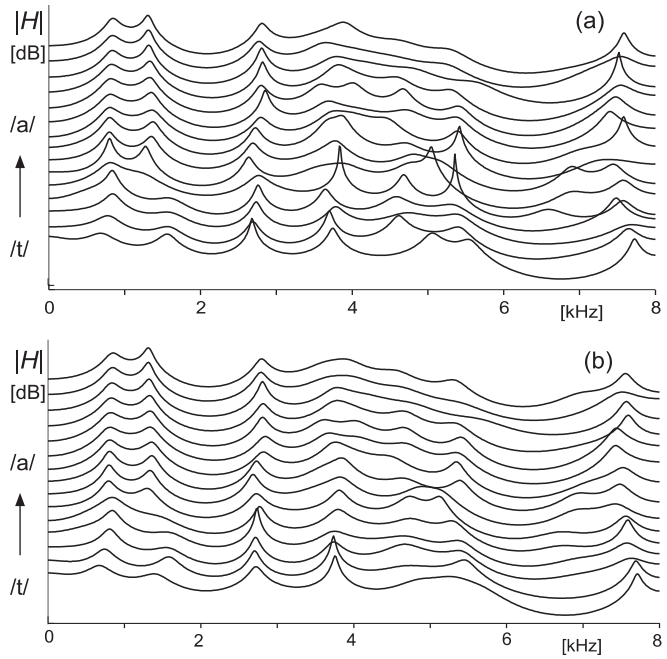

Abbildung 1 – Geschätzte Betragsgänge des Lautübergangs [ta]: (a) segmentweise Analyse mit zeitinvarianter Prädiktion, (b) Analyse mit zeitvariabler Prädiktion.

die Stetigkeitsbedingungen werden die Segmente nicht mehr unabhängig voneinander geschätzt, wodurch die Analyse weniger empfindlich gegenüber Variationen der Segmentgrenzen ist. Abbildung 2 zeigt Analysen der Äußerung „blaue“ [blaU@] von periodengenauen Segmentierungen mit den Markierungen $m[k]$, wobei sich diese an Nulldurchgängen befinden. Um die Auswirkungen von Segmentierungsfehlern zu untersuchen, werden die Markierungen $m[k]$ zusätzlich mittels zufälliger Fehler modifiziert. Dazu wird den ursprünglichen Markierungen ein Rauschen r aufgeprägt, wodurch sich die modifizierten Segmentgrenzen mit $m'[k] = m[k] + r[k]$ ergeben. In Abbildung 2 sind die geschätzten Betragsgänge der Segmente mit den originalen und fehlerbehafteten Markierungen überlagert gezeichnet. Abweichungen zwischen den beiden Betragsgängen mit und ohne Markierungsfehler zeigen, wie empfindlich der Schätzalgorithmus bezüglich kleiner Änderungen der Segmentgrenzen ist. Abbildung 2(a) zeigt die Auswirkungen der Markierungsfehler für die zeitinvariante lineare Prädiktion, während Abbildung 2(b) die Auswirkungen für die zeitvariable lineare Prädiktion zeigt. Es ist zu erkennen, dass sich die Markierungsfehler für die zeitvariable Prädiktion weniger auswirken als im zeitinvarianten Fall. Dieses Verhalten ist allerdings keine direkte Folge der expliziten Berücksichtigung der Zeitvariabilität, sondern eine Folge der Stetigkeitsbedingungen, da damit bis zu einem gewissen Grad kontinuierliche Zeitentwicklungen der Modellfunktionen erzwungen werden.

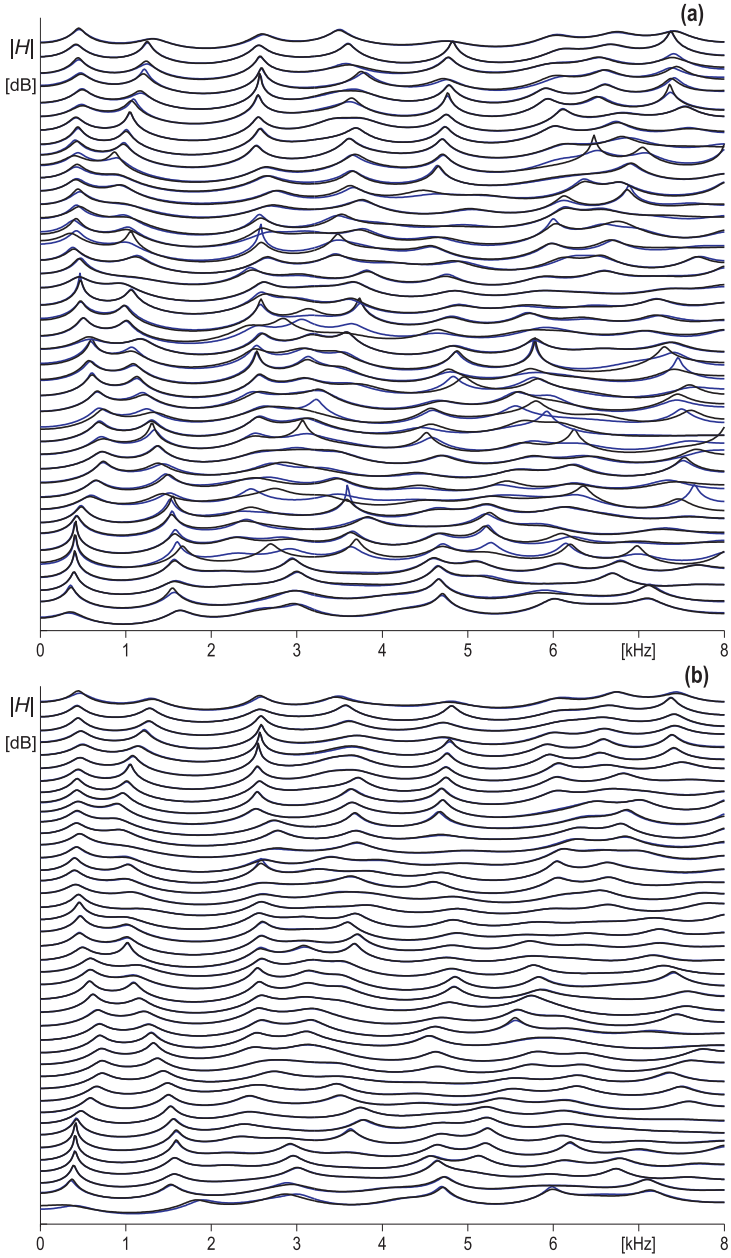

Abbildung 2 – Analyse der Äußerung „blaue“ [blaU@]. Geschätzte Betragsgänge periodensynchroner Segmente mit und ohne Markiergunsfehler überlagert dargestellt: (a) für Analyse mit zeitinvarianter Prädiktion, (b) für Analyse mit zeitvariabler Prädiktion.

Für den Gebrauch der hier vorgestellten zeitvariablen Prädiktion sollten für die Erzielung adäquater Ergebnisse zwei Dinge beachtet werden. Erstens sollte der zeitvariablen Prädiktion eine Filterung des Sprachsignals mit einer Präemphase vorausgehen und zweitens sollte sich die segmentverbundene Analyse nur auf Sprachabschnitte erstrecken, die entweder eine stimmhafte Anregung oder eine Rauschanregung aufweisen. Die Präemphase ist notwendig, da sonst die Prädiktionskoeffizienten ungünstige Trajektorien aufweisen und der stetige Übergang von stimmhaft zu stimmlos zu vermeiden ist, da dieser keinen Übergang der Modellfunktion des Sprechtraktes darstellt, sondern die Anregung betrifft.

4 Zusammenfassung

Für eine zeitvariable Sprachanalyse wurde ein zeitvariabler Prädiktionsalgorithmus vorgestellt, der eine stetige und segmentweise lineare Entwicklung der Prädiktorkoeffizienten annimmt. Die optimalen Koeffizienten können dafür analytisch bestimmt werden. Der Prädiktionsalgorithmus kann als zeitvariable Präemphase und als Prädiktion für die Vokaltraktschätzung eingesetzt werden. Die Analysen zeigen, dass bei Verwendung der zeitvariablen Prädiktion eine Präemphase für die Erzielung adäquater Ergebnisse erforderlich ist. Die Schätzergenisse der zeitvariablen Prädiktion weisen glattere Zeitverläufe der Modellbetragsgänge auf als die der gewöhnlichen linearen Prädiktion. Weiterhin zeigen die Untersuchungen, dass die zeitvariable Prädiktion im Vergleich zur zeitinvarianten Kovarianzmethode weniger empfindlich gegenüber Variationen der Segmentgrenzen ist, wodurch sich Schätzungen ergeben, die verhältnismäßig robust gegenüber Segmentierungsfehler ausfallen.

Literatur

[1] Makhoul, J.: "Linear prediction: A tutorial review", in *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[2] Markel, J. and Gray, A.: *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[3] Haykin, S.: *Adaptive Filter Theory*. New Jersey: Prentice-Hall, Inc., 3 ed., 1996.

[4] Jachan, M., Matz, G., and Hlawatsch, F.: "Time-Frequency-Autoregressive Random Processes: Modeling and Fast Parameter Estimation", in *Proc. ICASSP'03*, Hong Kong 2003, pp. 125–128.

[5] Subba Rao, T., "The Fitting of Non-stationary Time-series Models with Time-dependent Parameters," in *J. Roy. Statist. Soc. Series B*, vol. 32, no. 2, pp. 312-322, 1970.

[6] Grenier, Y.: "Time-Dependent ARMA Modeling of Non-stationary Signals,', in *IEEE Trans. ASSP-31*, no. 4, pp. 899–911, August 1983.

[7] Schnell, K. and Lacroix, A.: "Time-Varying Linear Prediction for Speech Analysis", in *Proc. EUSIPCO'07*, Poznan 2007 (in print).

[8] Wakita, H.: "Estimation of Vocal-Tract Shapes from Acoustical Analysis of the Speech Wave: The State of the Art", in *IEEE Trans. ASSP-27*, no. 3, pp. 281–285, June 1979.

[9] Schnell, K. and Lacroix, A.: "Time-Varying Pre-emphasis and Inverse Filtering of Speech", in *Proc. INTERSPEECH'07*, Antwerp 2007 (in print).