TIME-VARYING LINEAR PREDICTION FOR SPEECH ANALYSIS AND SYNTHESIS

Karl Schnell and Arild Lacroix

Institute of Applied Physics, Goethe-University Frankfurt
Max-von-Laue-Str. 1, 60438 Frankfurt am Main, Germany
schnell@iap.uni-frankfurt.de

ABSTRACT

In this contribution, a time-varying linear prediction is proposed for speech analysis and synthesis. In comparison to the time-invariant prediction, the predictor coefficients are time-varying within the frames. For that purpose, the coefficient trajectories can be described by basis functions. This approach leads to discontinuities between the frames if the frames are analyzed independently. Therefore, continuous conditions are defined which force continuous trajectories also between the frames. The estimation of the optimum coefficients of the basis functions is solved analytically by a least mean square approach. The analysis results show that the estimation algorithm achieves smooth trajectories of the vocal-tract resonances together with a high time resolution, which is interesting for a variety of application.

Index Terms— Time-varying filters, Prediction methods, Speech analysis, Speech synthesis

1. INTRODUCTION

Linear prediction is a commonly used technique in the field of speech processing. Since speech utterances are not time-invariant, the speech signals are usually segmented into frames, which are analyzed independently by one of the time-invariant prediction algorithms such as the autocorrelation, covariance, or Burg method assuming stationary statistics within the frames [1]. For nonstationary estimation of audio and speech signals, some approaches exist allowing also time-varying coefficients of the underlying model. A general approach is to use adaptive algorithms like LMS, RLS or Kalman filter [2]. One more specialized approach is the time-varying autoregressive modelling technique (TVAR). These methods estimate an AR model with time-varying coefficients; the coefficients can be described by basis functions, e.g. in [3]-[5]; in [5] also ARMA is treated. For abrupt changes of coefficients the consideration of smoothness prior (SP) is suitable [6]. A more statistical approach of estimating time-varying AR models for example is described in [7]. In [8]-[9], a time-varying linear prediction is proposed for speech analysis and pre-emphasis, which assumes piece-wise linearly time-varying coefficients. In contrast to that, in this contribution a time-varying prediction algorithm is proposed based on basis functions and, additionally, on continuous conditions, which ensure continuous connections of the basis functions between the frames. Furthermore, the time-varying prediction is discussed for speech analysis and synthesis techniques regarding which benefits can be expected from time-varying approaches.

2. TIME-VARYING LINEAR PREDICTION

In the case of time-varying prediction, the predictor coefficients are time-dependent. For a segment-wise analysis, the signal $x(n)$ to be analyzed is segmented into *P* non-overlapping frames

$$x(n) = \ldots x^{k-1}(L^{k-1}),\ \underbrace{x^k(1), x^k(2), \ldots x^k(L^k),}_{k\text{-th frame}}\ x^{k+1}(1), \ldots ;$$

with $k = 1 \ldots P$. The superscript denote the corresponding frame of length L. The time-varying linear prediction for the *k*-frame can be described in generally by

$$\hat{x}^k(n) = \sum\nolimits_{i=1}^{N} a_i^k(n) \cdot x^k(n-i)\,. \qquad (1)$$

$\hat{x}^k(n)$ is the estimation of $x^k(n)$ and the superscript *k* denotes the frame. If $i \geq n$ is valid, the values of $x^k(n-i)$ in Eq. (1) are defined by values of the previous frame with

$$x^k(-m) = x^{k-1}(L^{k-1} + 1 - m) \quad \text{for} \quad m \geq 0\,.$$

To ensure continuous functions of the predictor coefficients and to reduce the number of parameters, the predictor coefficients can be described by a superposition of basis functions $\phi^{j,k}(n)$ leading to

$$a_i^k(n) = \sum\nolimits_{j=0}^{M} d_i^{j,k} \cdot \phi^{j,k}(n)\,. \qquad (2)$$

Hence, the coefficient trajectories are determined by the parameters $d_i^{j,k}$. Here, the first basis function $\phi^{0,k}(n) = 1$ is defined constant to one describing the stationary component of $a_i^k(n)$, whereas the functions $\phi^{j,k}(n)$ with $j > 0$ are time-varying and describe the nonstationary components. The signals and basis functions of each frame can be described vector-based by the definitions:

$\boldsymbol{x}_i^k = \left(x^k(1-i),\ \ldots x^k(L^k - i)\right)^{\mathsf{T}}$, $\boldsymbol{e}^k = \left(e^k(1), e^k(2), \ldots, e^k(L^k)\right)^{\mathsf{T}}$,

$\boldsymbol{\varphi}^{j,k} = \left(\phi^{j,k}(1), \phi^{j,k}(2), \ldots, \phi^{j,k}(L^k)\right)^{\mathsf{T}}$, and $\boldsymbol{w}_i^{j,k} = \boldsymbol{\varphi}^{j,k} \otimes \boldsymbol{x}_i^k$;

the operation $\otimes$ describes the element-by-element multiplication with $\boldsymbol{w}_i^{j,k} = \boldsymbol{\varphi}^{j,k} \otimes \boldsymbol{x}_i^k \rightarrow w_i^{j,k}(n) = \phi^{j,k}(n) \cdot x_i^k(n)$. The vector $\boldsymbol{x}_0^k$ represents the *k*-th frame whereas $\boldsymbol{x}_i^k$ for $i > 0$ represents the shifted *k*-th frame with a shifting of *i* samples. The prediction error vector $\boldsymbol{e}^k$ of the *k*-th frame is defined by

$$\boldsymbol{e}^k = \boldsymbol{x}_0^k - \sum_{i=1}^{N} \sum_{j=0}^{M} \left(d_i^{j,k} \cdot \boldsymbol{w}_i^{j,k}\right) \qquad (3)$$

with values $e^k(n) = x^k(n) - \hat{x}^k(n)$. Considering $\boldsymbol{w}_i^{0,k} = \boldsymbol{x}_i^k$ due to $\phi^{0,k}(n) = 1$ and solving Eq. (3) for the vector $\boldsymbol{x}_0^k$ lead to

$$\boldsymbol{x}_0^k = d_i^{0,k} \cdot \boldsymbol{x}_i^{0,k} + \sum_{i=1}^{N} \sum_{j=1}^{M} \left(d_i^{j,k} \cdot \boldsymbol{w}_i^{j,k}\right) + \boldsymbol{e}^k \quad \text{for } k = 1 \ldots P\,. \qquad (4)$$

Eq. (4) represents a vector expansion of $\boldsymbol{x}_0^k$ by the basis vectors $\boldsymbol{w}_i^{j,k} = \boldsymbol{\varphi}^{j,k} \otimes \boldsymbol{x}_i^k$, which are the basis functions multiplied by the signal values. The error of approximation is $\boldsymbol{e}^k$ representing the prediction error. The coefficients $d_i^{j,k}$ can be determined by regression minimizing the norm of the error vector $|\boldsymbol{e}^k|$ for each frame k separately. However, in this way the segments would be analyzed independently, which causes discontinuities. To consider the continuous movements of the vocal tract, the coefficients should evolve continuously in time, also across frames. Therefore, a continuity condition is defined by

$$a_i^k(L^k]) = a_i^{k+1}(1) \tag{5}$$

$$\sum\nolimits_{j=0}^{M} d_i^{j,k} \cdot \phi^{j,k}(L^k) = \sum\nolimits_{j=0}^{M} d_i^{j,k+1} \cdot \phi^{j,k+1}(1)$$

ensuring that the last coefficient value $a_i^k(L^k])$ of each frame k is connected continuously to the first value $a_i^{k+1}(1)$ of the next frame. Eq. (5) implies a coupling of Eq. (4) for $k = 1 \ldots P$. Solving Eq. (5) for the coefficient $d_i^{0,k}$ results in

$$d_i^{0,k+1} = \left(\sum\nolimits_{j=0}^{M} d_i^{j,k} \cdot \phi^{j,k}(L^k) - \sum\nolimits_{j=1}^{M} d_i^{j,k+1} \cdot \phi^{j,k+1}(1)\right) / \phi^{0,k+1}(1) . \tag{6}$$

Eq. (6) shows that the coefficients $d_i^{0,k}$ with $k > 1$ are determined by the other coefficients, which reduces the number of parameters to be estimated. The coefficients $d_i^{0,1}$ remain which represent the stationary components. Eqs. (4) and (5) can be combined in one vector expansion or one single system of equations covering the frames $k = 1 \ldots P$. For that purpose, the vectors of each frame are arranged on top of each other, which leads to the combined vector equation

$$\boldsymbol{q}_0^0 = \sum_{i=1}^{N} d_i^{0,1} \cdot \boldsymbol{q}_i^0 + \sum_{k=1}^{P} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(d_i^{j,k} \cdot \boldsymbol{q}_i^{j,k} \right) + \boldsymbol{e} \tag{7}$$

with the vectors $\boldsymbol{q}_i^0$ and $\boldsymbol{q}_i^{j,k}$ for $k = 0 \ldots P$ defined by

$$\boldsymbol{q}_i^0 = \begin{pmatrix} \boldsymbol{x}_i^1 \\ \boldsymbol{x}_i^2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{x}_i^P \end{pmatrix}, \boldsymbol{q}_i^{j,1} = \begin{pmatrix} \boldsymbol{w}_i^{j,1} \\ \boldsymbol{v}_{i,1}^{j,2} \\ \boldsymbol{v}_{i,1}^{j,3} \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{v}_{i,1}^{j,P} \end{pmatrix} \cdots \ \boldsymbol{q}_i^{j,k} = \begin{pmatrix} \boldsymbol{u}_{i,k}^{j,1} \\ \vdots \\ \boldsymbol{u}_{i,k}^{j,k-1} \\ \boldsymbol{w}_i^{j,k} \\ \boldsymbol{v}_{i,k}^{j,k+1} \\ \vdots \\ \boldsymbol{v}_{i,k}^{j,P} \end{pmatrix}, \cdots \ \boldsymbol{q}_i^{j,P} = \begin{pmatrix} \boldsymbol{u}_{i,P}^{j,1} \\ \boldsymbol{u}_{i,P}^{j,2} \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{u}_{i,P}^{j,P-1} \\ \boldsymbol{w}_i^{j,P} \end{pmatrix}$$

and with the vectors $\boldsymbol{u}_{i,m}^{j,k} = \phi^{j,m}(1) \cdot \boldsymbol{x}_i^k$ and $\boldsymbol{v}_{i,m}^{j,k} = \phi^{j,m}(L^m) \cdot \boldsymbol{x}_i^k$. The vector $\boldsymbol{e}$ contains the vectors $\boldsymbol{e}^k$ one after the other for $k = 1 \ldots P$, analogously to the vectors $\boldsymbol{q}_i^0$ and $\boldsymbol{q}_i^{j,k}$. The vectors $\boldsymbol{u}_{i,m}^{j,k}$ and $\boldsymbol{v}_{i,m}^{j,k}$ considers the continuous condition of Eq. (5). This is explained in the following by regarding the vector $\boldsymbol{q}_i^{j,k}$. The vector $\boldsymbol{q}_i^{j,k}$ includes the sub-vectors $\boldsymbol{u}_{i,m}^{j,k}$, $\boldsymbol{v}_{i,m}^{j,k}$, and $\boldsymbol{w}_i^{j,k}$. The vectors $\boldsymbol{u}_{i,m}^{j,k}$ and $\boldsymbol{v}_{i,m}^{j,k}$ imply the values $\phi^{j,k}(1)$ and $\phi^{j,k}(L^k)$, respectively, and are arranged before and after the vector $\boldsymbol{w}_i^{j,k}$. Since the vector $\boldsymbol{w}_i^{j,k}$ contains function values ranging from $\phi^{j,k}(1)$ to $\phi^{j,k}(L^k)$, the arrangement of the sub-vectors ensures a continuous trajectory of the predictor coefficients for all vectors $\boldsymbol{q}_i^{j,k}$. This is achieved by the extension of the boundary coefficient values $\phi^{j,k}(1)$ and $\phi^{j,k}(L^k)$ of $\boldsymbol{w}_i^{j,k} = \boldsymbol{\varphi}^{j,k} \otimes \boldsymbol{x}_i^k$ to the other frames by introducing $\boldsymbol{u}_{i,m}^{j,k-l} = \phi^{j,m}(1) \cdot \boldsymbol{x}_i^{k-l}$ and $\boldsymbol{v}_{i,m}^{j,k+l} = \phi^{j,m}(L^m) \cdot \boldsymbol{x}_i^{k+l}$ with $l > 0$. The optimum solution of Eq. (7) is determined by regression minimizing the length of the error vector or the power of the prediction error, respectively.

For an efficient calculation of the prediction algorithm, the sequence of frames can be segmented into overlapping sub-sequences of frames which are processed analogous to Eq. (7) by regression one after the other. To consider the continuous condition between the sub-sequences, the constant coefficients $d_i^{0,1}$ are adopted from the analysis of the previous sub-sequence.

2.1. Basis functions

The basis functions $\phi^{j,k}(n)$ describe the space of the possible parameter trajectories. The constant term $\phi^{0,k}(n) = 1$ is used for all different types of basis functions. For defining the basis functions, the definition of the linear sequences

$$\theta_L = (0,1,\ldots L-1)^{\mathrm{T}} / (L-1) \quad \text{and} \quad \theta_L^r(n) = (L-1,\ldots 1,0)^{\mathrm{T}} / (L-1)$$

of length L ranging from zero to one and reversed are useful. Basis functions can be defined by applying a function f to θ. The use of θ_L^r can be suitable if $d \cdot f(\theta)$ with an arbitrary coefficient d cannot describe its time-reversed counterpart $f(\theta^r)$; for examples, this is the case for polynomials. The simplest time-varying basis functions is the linear function

$$\phi_{\text{lin}}^{1,k}(n) = \theta_{L^k}(n) ,$$

which describes a straight line. L^k is equal to the length of frame k. Polynomial basis functions are defined, here, by

$$\phi_{\text{poly}}^{j,k}(n) = (\theta_{L^k}(n))^{(j+1)/2} \quad \text{for even indices } j = 1,3,\ldots,M$$

and $\phi_{\text{poly}}^{j,k}(n) = (\theta_{L^k}^r(n))^{j/2+1}$ for odd indices $j = 2,4,\ldots M-1$.

Since $\phi_{\text{poly}}^{1,k}$ is the linear function θ_{L^k} and the following basis functions are the polynomials alternating with their reversed counterparts, even numbers M are appropriate.

Another type of basis functions is motivated by the periodic behavior of the glottis termination of the vocal tract. For that purpose, trigonometric basis functions are defined by

$$\phi_{\text{trig}}^{j,k}(n) = \sin\left(\theta_{L^k}(n) \cdot 2\pi(j+1)/2\right) \quad \text{for even indices } j = 1,3,..M-1$$

$$\phi_{\text{trig}}^{j,k}(n) = \cos\left(\theta_{L^k}(n) \cdot 2\pi j/2\right) \quad \text{for odd indices } j = 2,4,\ldots M .$$

3. ANALSYSIS AND SYNTHESIS OF SPEECH

For speech analysis, the speech signal is segmented into adjoining non-overlapping frames. Then, the segmentation together with the basis functions determine the vectors $\boldsymbol{q}_i^0$ and $\boldsymbol{q}_i^{j,k}$. The regression of the vector expansion (7) yields the optimum predictor coefficients $a_i(n) = \ldots a_i^{k-1}(L^{k-1}), \underbrace{a_i^k(1), a_i^k(2), \ldots a_i^k(L^k),}_{k-\text{th frame}} a_i^{k+1}(1), \ldots$ by the estimated coefficients $d_i^{j,k}$ for the whole analyzed signal $x(n)$. To obtain a reduced number of coefficient sets, the sequences of the coefficients $a_i(n)$ are down-sampled by selecting every L^{step}-th sample, which leads to the sequences

$\tilde{a}_i^m = a_i^k(m \cdot L^{\text{step}})$. If the transfer function $H(a_i, z)$ corresponds to the coefficients a_i, the transfer functions $H_m = H(\tilde{a}_i^m, z)$ represent a sequence of transfer functions implying a step size of L^{step} samples. Figure 1 shows the transfer functions estimated from the utterance [jUlIa] by time-invariant and time-varying prediction of order $N = 24$ with a step size L^{step} of 100 samples. The sampling rate of the analyzed speech is 16 kHz and the prediction order is $N = 24$.

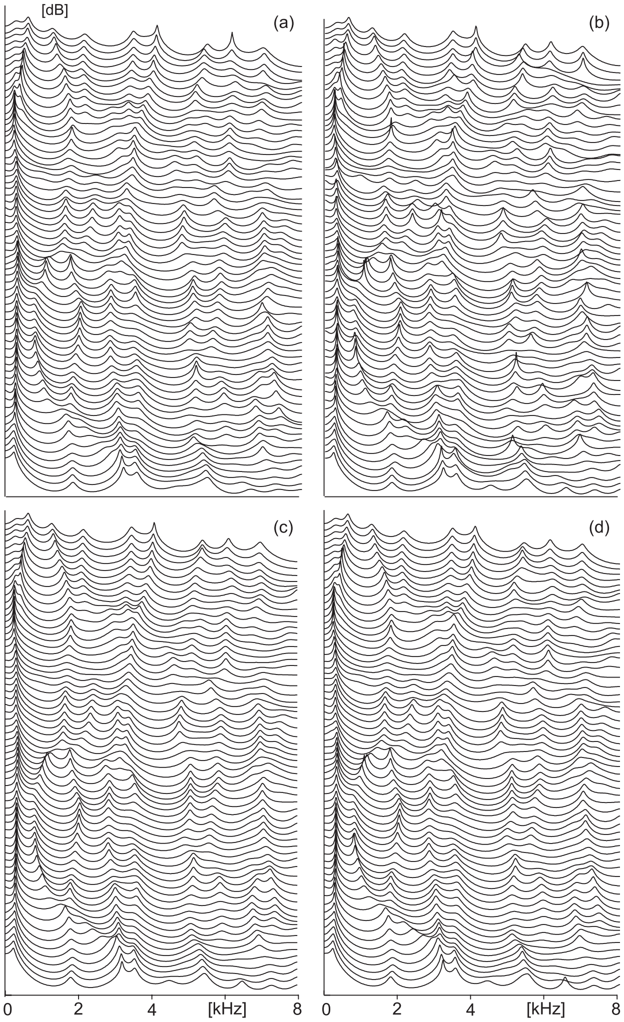

Figure 1: Estimated magnitude responses of utterance [jUlia]: (a) time-invariant prediction; (b) time-varying prediction with linear basis without continuous condition; (c)-(d) time-varying prediction with continuous condition, (c) linear basis and (d) polynomial basis with $M = 5$; frame length is 300 samples for (a)-(c) and 800 samples for (d).

For comparison, Fig. 1(a) represents the conventional time-invariant prediction which is performed by the covariance method with overlapping frames with a length of 300 samples. Figs. 1 (c) and (d) represent the results by the proposed time-varying prediction with linear and polynomial basis functions, respectively, and with the continuous condition. It can be seen that the spectral trajectories of the time-invariant (graph (a)) and time-varying approach (graphs (c)-(d)) show similarities; however, the time-varying approach leads to a smoother trajectory. With larger frame lengths also the time-invariant approach yields smoother trajectories, but the time resolution can be decreased. Figs. 1(c) and (d) demonstrate that similar results can be achieved with linear and polynomial basis functions. For that, the frame length for the prediction with polynomial functions should be larger than that for the prediction with the linear functions. Fig. 1(b) shows the time-varying analysis with same conditions as used for Fig. 1(c), but without using the continuous condition, which leads to fluctuating trajectories. The results of Fig. 1(b) are obtained by considering Eq. (4) only. This demonstrates that the introduction of the continuous condition by Eq. (5) is sensible.

3.1. Formant patterns

From the estimated sequence of predictor coefficients the corresponding roots can be calculated from which the resonance trajectories can be obtained. In Fig. 2 the frequencies of the roots are depicted representing the estimated trajectories of the formant or resonance frequencies obtained from the utterance [laN@waIl].

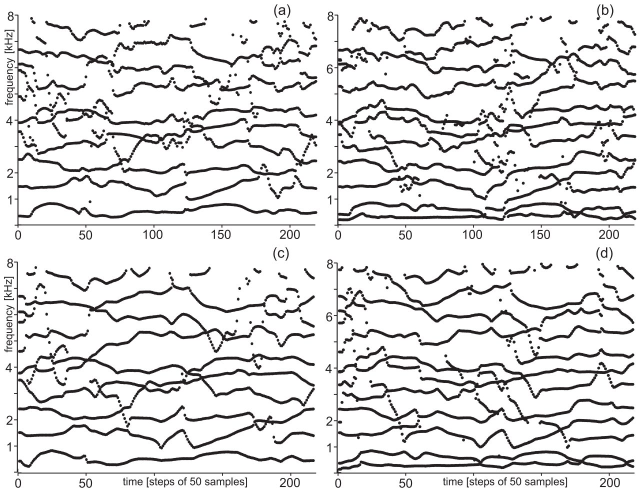

Figure 2: Estimated resonances from the utterance [laN@waIl]: (a)-(b) time-invariant prediction with frame length 600 samples; (c)-(d) time-varying prediction with polynomial basis, continuous condition and frame length 400 samples; (a) and (c) with prediction order $N = 20$ and with pre-emphasis; (b) and (d) with prediction order $N = 24$ and without pre-emphasis.

Figs. 2 (c) and (d) are obtained by time-varying prediction with a frame length of 400 samples, whereas Figs. 2 (a) and (b) is obtained by time-invariant prediction with a frame length of 600 samples. Although the time resolution of the time-varying estimation is better, the trajectories of the time-varying estimation are equal or more smooth than those of the time-invariant estimation. Figs. 2 (a) and (c) show analysis results with pre-emphasis and Figs. 2 (b) and (d) without pre-emphasis. Since the pre-emphasis compensates the so-called glottal formant, the prediction order without pre-emphasis is chosen higher than with pre-emphasis. The pre-emphasis is performed adaptively by a linear prediction of order one [1]. The pre-emphasis is repeated twice by using the prediction error as input signal for the next pre-emphasis. The error signal of the last prediction of order one represents the pre-emphasized speech. In the time-varying case, also the time-varying prediction of order one is used [8].

Analyses of several utterances indicate that for the time-varying prediction the resonance patterns are more invariant against variations of prediction order and pre-emphasis than the time-invariant prediction. Furthermore, the investigations show that in the case of noisy speech the resonance trajectories are relatively smooth, even in the disturbed spectral regions.

3.2. Periodic basis functions

The time-varying prediction with frame lengths which are small compared to the pitch period leads to fluctuating trajectories. These fluctuations are caused by statistical effects and by the voiced excitation. Therefore, to obtain smooth trajectories the frame length should be larger than the pitch period. The following example shows that if the basis functions allow dents in the coefficient trajectories within pitch periods, the excitation is affected by the linear prediction. For that purpose, trigonometric basis functions are used. The frame length is pitch-synchronous two or three pitch periods. The trigonometric functions are chosen in a way that the periodicity of the sine and cosine functions corresponds to the pitch periods. For that purpose, a selection j' of indices j of the trigonometric basis functions is chosen corresponding to the number R of pitch periods in each frame. This means that, for example, if the frames are exactly two periods long, the functions $\phi_{\text{trig}}^{j',k}(n)$ with the indices $j' = 3,4,7,8,11\ldots$ are chosen. From Fig. 3, it can be seen that the inclusion of the trigonometric functions decreases significantly the prediction error power and affects the peaks of the glottal excitation.

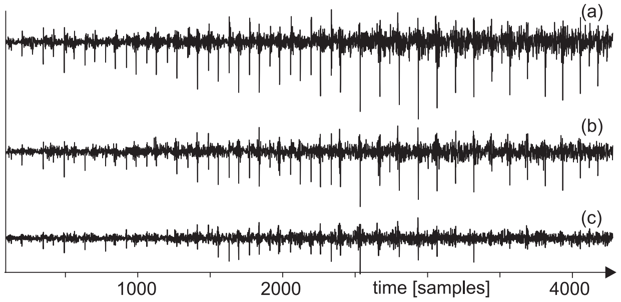

Figure 3: Prediction error signal of voiced speech by time-varying prediction of order $N = 24$, the frame length is pitch-synchronous 2 periods for (a),(c) and 3 periods long for (b); (a) linear basis only, (b) linear basis and 4 trigonometric basis functions, (c) linear basis and 6 trigonometric basis functions.

3.3. Synthesis by time-varying prediction

Besides considering spectral trajectories, also re-synthesis is useful to asses the estimation results. For examples, the re-synthesis reveals that the speech quality is not always correlated with the prediction gain and that the trajectories of the coefficients and resonances should be smooth, but not too smooth, since over-smoothing can also degrade the speech quality. To realize the re-synthesis, the estimated sequences of predictor coefficients $\tilde{a}_i^m$ are used for controlling an all-pole system. The excitation of the all-pole system is independently of the analyzed speech. Since the predictor coefficients can lead to unstable systems, the coefficients $\tilde{a}_i^m$ are checked for stability. In case of instability, the roots z_i^m of the polynomial with coefficients $\tilde{a}_i^m$ are calculated and the radii of the unstable poles z_λ^m with index $i = \lambda$ are pulled into the unit circle by changing the radius of each unstable pole by $|z'^m_\lambda| = 1/|z^m_\lambda|$. After exchanging the roots $z^m_\lambda := z'^m_\lambda$, the roots are converted back into coefficients. Fortunately, unstable configurations occur not often.

Overall, the re-synthesis indicates that the time-varying prediction algorithm is suitable for synthesis techniques and can lead to a moderately different speech quality compared to time-invariant prediction. It seems that fast sound transitions can be better produced by using the time-varying prediction. It should be considered, that not only the best possible speech quality of re-synthesis is important, since for applications such as parametric synthesis or speaker transformation also the trajectory of the spectral envelope has to be modified; for that purpose, the availability of suitable resonance trajectories can be important, too.

4. CONCLUSIONS

A segment-wise time-varying prediction algorithm is proposed for speech analysis and synthesis. To yield smooth coefficient trajectories of the basis functions between adjacent frames, constraints are defined ensuring continuous frame connections. On these constraints, the optimum coefficients of the basis functions can be estimated quasi-analytically by a least mean square approach. The investigations show that the proposed algorithm achieves smooth spectral trajectories together with a high time resolution which can be utilized for speech analysis and synthesis.

5. REFERENCES

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

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

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

[4] A. Härmä, M. Juntunen, and J. P. Kaipio, “Time-Varying Autoregressive Modeling of Audio and Speech Signals,” *Proc. EUSIPCO*, Tampere Finland, 2000.

[5] Y. Grenier, “Time-Dependent ARMA Modeling of Non-stationary Signals, ” *IEEE Trans.* ASSP-31, no. 4, pp. 899–911, August 1983.

[6] J. P. Kaipio and M. Juntunen, “Deterministic Regression Smoothness Priors TVAR Modelling, ” *Proc. ICASSP*, Phoenix USA, 1999.

[7] K. M. Malladi and R. V. Rajakumar, “Estimation of Time-Varying AR Models of Speech through Gauss-Markov Modeling,” *Proc. ICASSP*, Hong Kong, pp. 305–308, 2003.

[8] K. Schnell and A. Lacroix, “Time-Varying Pre-emphasis and Inverse Filtering of Speech,” *Proc. INTERSPEECH*, Antwerp Belgium, pp. 530-533, 2007.

[9] K. Schnell and A. Lacroix, “Time-Varying Linear Prediction for Speech Analysis,” *Proc. EUSIPCO*, Poznan Poland, pp. 2045-2049, 2007.