ITERATIVE INVERSE FILTERING BY LATTICE FILTERS FOR TIME-VARYING ANALYSIS AND SYNTHESIS OF SPEECH

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, an analysis procedure is proposed for time-varying analysis and synthesis of speech based on lattice filters. The estimation is performed by an iterative inverse filtering approach exploiting analytical suboptimal solutions. Starting from an initial configuration of coefficients, the procedure estimates a continuous piece-wise linear trajectory in terms of reflection coefficients. In this way, smooth trajectories can be estimated which have additionally a high time resolution. One advantage of this analysis technique is that the coefficient trajectories are estimated in the same way as they are used for the synthesis. Examples of synthesized speech signals show that the proposed algorithm suppresses artifacts which are caused by the use of time-invariant estimation procedures.

Index Terms— Time-varying filters, Speech analysis, Speech synthesis.

1. INTRODUCTION

Model-based analysis and synthesis of speech are often based on all-pole models. The model parameters are commonly estimated from speech frames by conventional time-invariant linear prediction techniques. Since the speech production process is a non-stationary process, also time-varying estimation algorithms exist [1]-[7]. One important aim of estimating time-varying coefficient trajectories is to yield a smooth trajectory simultaneously with a high time resolution and a good approximation of the spectral envelope. A general category of time-varying analysis techniques are adaptive filtering algorithms like LMS or Kalman filtering [1], [2]. One practical solution to determine the coefficients is to develop the coefficient trajectory by basis functions, which is proposed for individual frames in [3], [4] for direct-form and reflection coefficients. A joint analysis of adjacent frames is favorable, to yield a continuous trajectory also between frames. This is shown in [5] by presenting an analytical multi-frame analysis based on time-varying basis functions in terms of direct-form coefficients. The direct-form coefficients are rather not suitable for interpolation and, therefore, not optimum for time-varying analysis and synthesis of speech. In [6] an algorithm is proposed which estimates a continuous time-varying trajectory in terms of reflection coefficients frame by frame. For that purpose, the coefficients of one frame are estimated with respect to the analysis results of previous frames. Although this procedure is simple and efficient, the optimization is not optimum since the estimation is not consistent with respect to the whole analyzed speech signal. In comparison to that, in this contribution an algorithm is proposed which estimates the continuous trajectory with respect to the whole speech signal by an iterative procedure.

2. TIME-VARYING ANALYSIS

The IIR lattice filter is related to a simple vocal tract model and is used for synthesis. The FIR lattice filter, which is shown in Fig. 1, is used for the estimation by inverse filtering. Due to the time-varying estimation, the reflection coefficients are time-varying within each frame. The trajectory $r(n)$ of the reflection coefficients is assumed to be continuous and piece-wise linear.

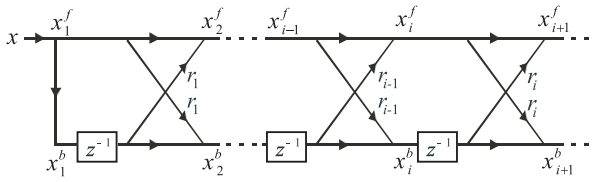

Figure 1: FIR lattice filter for inverse filtering.

2.1. Estimation procedure

Prior to the analysis, the speech signal *s* is pre-emphasized by a repeated adaptive pre-emphasis resulting in the signal *x*. According to the piece-wise linear trajectory, the pre-emphasized speech signal *x* is segmented into adjacent frames x_k of length *L* for the frames $k = 1 \ldots P$. The trajectory of the *i*-th reflection coefficient and the *k*-th frame is denoted by $r_{i,k}$. The segmentation is performed in a way that the trajectories within each frame are linear with

$$r_{i,k}(n) = c_i + d_i \cdot (n-1)/(L-1) \quad \text{for} \quad n = 1 \ldots L \,. \tag{1}$$

The coefficients which are located at the left and right boundary of the frame are denoted by

$$r_{i,k}^l = r_{i,k}(1) \quad \text{and} \quad r_{i,k}^r = r_{i,k}(L) \,, \tag{2}$$

respectively. Since the whole trajectories are continuous, the linear trajectories $r_{i,k}(n)$ of the frames are connected continuously by $r_{i,k}(L) = r_{i,k+1}(1)$. The frames x_k are analyzed by an iterative algorithm which needs a starting configuration of the coefficients. Then, the coefficients are updated iteratively with the aid of analytical suboptimal solutions for one coefficient. The time-

varying reflection coefficients of one frame are updated each after the other by minimizing the output powers of each section. In Fig. 2 one section is depicted.

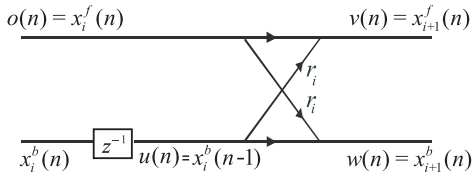

Figure 2: *i*-th section of FIR lattice filter.

2.1.1. Suboptimal solution

In the following the suboptimal solution of each section of the FIR lattice filter is treated. For readability the index *i* of the reflection coefficients is left out, which leads to $r_k = r_{i,k}$ for example. The suboptimal solution estimates the right-sided coefficient $r_k^r = r_k(L)$ of one frame on condition that the coefficients of the other frames are known. For the estimation of the coefficients of the *k*–th frame, only the coefficients r_{k-1}^r and r_{k+1}^r of the left-sided and right-sided frame are needed. Since each coefficient $r_k^r = r_k(L)$ to be estimated is located between the frames *k* and $k+1$, the two frames *k* and $k+1$ are united to one segment $\overline{x}_k$ by

$$\overline{x}_k = (x_k(1), x_k(2), \ldots, x_k(L), x_{k+1}(1), \ldots, x_{k+1}(L)) . \quad (3)$$

Analogously, the coefficient trajectory $\overline{r}_k$ of the segment $\overline{x}_k$ is defined by

$$\overline{r}_k = (r_k(1), r_k(2), \ldots, r_k(L), r_{k+1}(1), \ldots, r_{k+1}(L)) . \quad (4)$$

The trajectory $\overline{r}_k$ is linear from $\overline{r}_k(1)$ to $\overline{r}_k(L)$ and from $\overline{r}_k(L+1)$ to $\overline{r}_k(2L)$ considering the linear trajectories of the embedded two frames; furthermore, $\overline{r}_k$ is continuous with $\overline{r}_k(L) = \overline{r}_k(L+1)$. The coefficient to be estimated lies in the center of the segment $\overline{x}_k$. To describe the coefficient trajectory $\overline{r}_k$ of the segment $\overline{x}_k$ by basis functions, the two basis functions

$$\phi_1(n) = \begin{cases} (n-1)/(L-1) & \text{for} \quad n = 1 \ldots L \\ (2L-n)/(L-1) & \text{for} \quad n = L+1 \ldots 2L \end{cases} \quad (5)$$

$$\phi_2(n) = \quad (n-1)/(2L-1) \quad \text{for} \quad n = 1 \ldots 2L$$

are defined. Then, the trajectory $\overline{r}_k$ can be described by

$$\overline{r}_k(n) = c + d_1\phi_1(n) + d_2\phi_2(n) \quad \text{for} \quad n = 1 \ldots 2L . \quad (6)$$

Important values of the basis functions are located at the segment boundaries with $\phi_1(1) = \phi_2(1) = 0$, $\phi_1(2L) = 0$ and $\phi_2(2L) = 1$ as well as at the center with $\phi_1(L) = 1$, $\phi_2(L) \simeq 0.5$. Hence, the coefficient values at the segment boundaries are $\overline{r}_k(1) = c$ and $\overline{r}_k(2L) = c + d_2$. Considering the prescribed coefficients r_{k-1}^r and r_{k+1}^r, the coefficients c and d_2 are fixed with

$$c = \overline{r}_k(1) = r_{k-1}^r \quad \text{and} \quad d_2 = \overline{r}_k(2L) - c = r_{k+1}^r - r_{k-1}^r . \quad (7)$$

The coefficient to be estimated at $n = L$ can be described by

$$r_k^r = \overline{r}_k(L) = c + d_1 + 0.5d_2 . \quad (8)$$

Since c and d_2 are prescribed by the left-sided and right-sided frames, for the estimation of r_k^r only the coefficient d_1 has to be estimated. The estimation is performed by a minimization of the output powers. The outputs of one section corresponding to the designations of Fig. 2 are

$$\begin{aligned} v(n) &= o(n) + \overline{r}_k(n)u(n) \\ &= o(n) + (c + d_1\phi_1(n) + d_2\phi_2(n))u(n) \\ w(n) &= u(n) + \overline{r}_k(n)o(n) \\ &= u(n) + (c + d_1\phi_1(n) + d_2\phi_2(n))o(n) \end{aligned} \quad (9)$$

and can be described by the definitions $\tilde{o}_l(n) = \phi_l(n) \cdot o(n)$ and $\tilde{u}_l(n) = \phi_l(n) \cdot u(n)$ for $l = 1, 2$ with

$$\begin{aligned} v &= o + c\tilde{u}_1 + d_1\tilde{u}_1 + d_2\tilde{u}_2 \\ w &= u + c\tilde{o}_1 + d_1\tilde{o}_1 + d_2\tilde{o}_2 . \end{aligned} \quad (10)$$

The basis function ϕ_1 can produce a sharp bend of the trajectory which can be affected by the voiced excitation. To exclude this effect, also a time-invariant estimation is integrated into the estimation, whose impact can be adjusted by the parameter α. The time-invariant estimation implies the coefficient $r_k^r = c + d_1 + 0.5d_2$ as constant coefficient located in the center at $n = L$. Therefore, segments with indices $[n'] = L - M \ldots L + M$ are used for the time-invariant component. M is here chosen corresponding to the frame length by $M = L/2$. The output signals in the time-invariant case are described by

$$\begin{aligned} v^{\text{ti}}[n'] &= o[n'] + r_k^r u[n'] = o[n'] + (c + d_1 + 0.5 \cdot d_2)u[n'] \\ w^{\text{ti}}[n'] &= u[n'] + r_k^r o[n'] = u[n'] + (c + d_1 + 0.5 \cdot d_2)o[n']. \end{aligned} \quad (11)$$

For the time-invariant estimation, the input signals $o[n']$ and $u[n']$ are weighted by a Hamming window of length $2M+1$ resulting in the signals $\hat{o}(n)$ and $\hat{u}(n)$. The output signals incorporating the windowing are defined by

$$\begin{aligned} \hat{v}^{ti}(n) &= \hat{o}(n) + (c + d_1 + 0.5 \cdot d_2)\hat{u}(n) \\ \hat{w}^{ti}(n) &= \hat{u}(n) + (c + d_1 + 0.5 \cdot d_2)\hat{o}(n). \end{aligned} \quad (12)$$

For the estimation of the parameter d_1, the error *e* to be minimized is a linear combination of the time-varying and time-invariant case. By analogy with the Burg method, the arithmetic mean of the output powers are chosen for both cases resulting in

$$e(d_1) = \mathsf{E}\left[\alpha((v)^2 + (w)^2) + (1-\alpha)((\hat{v}^{\text{ti}})^2 + (\hat{w}^{\text{ti}})^2)\right] \rightarrow \min . \quad (13)$$

To minimize the error *e*, the derivate of the error with respect to the coefficient d_1 is set to zero

$$\frac{\partial e}{\partial d_1} = 0 . \quad (14)$$

Solving (14) for d_1 leads to formula

$$d_1 = -\mathsf{E}\left[\frac{\alpha(\varepsilon_{\text{n}}^{\text{tv}}) + (1-\alpha)(\varepsilon_{\text{n}}^{\text{ti}})}{\alpha(\varepsilon_{\text{d}}^{\text{tv}}) + (1-\alpha)(\varepsilon_{\text{d}}^{\text{ti}})}\right] \quad (15)$$

of the suboptimal coefficient with the definitions

$$\begin{aligned} \varepsilon_{\text{n}}^{\text{tv}} &= \tilde{u}_1 o + \tilde{o}_1 u + c(\tilde{u}_1 u + \tilde{o}_1 o) + d_2(\tilde{u}_1\tilde{u}_2 + \tilde{o}_1\tilde{o}_2) \\ \varepsilon_{\text{d}}^{\text{tv}} &= (\tilde{u}_2)^2 + (\tilde{o}_2)^2 \\ \varepsilon_{\text{n}}^{\text{ti}} &= \hat{u}^{\text{ti}}\hat{o}^{\text{ti}} + \hat{o}^{\text{ti}}\hat{u}^{\text{ti}} + (c + 0.5 \cdot d_2)((\hat{u}^{\text{ti}})^2 + (\hat{o}^{\text{ti}})^2) \\ \varepsilon_{\text{d}}^{\text{ti}} &= (\hat{u}^{\text{ti}})^2 + (\hat{o}^{\text{ti}})^2. \end{aligned}$$

The expected value E is calculated by the means of the signal values. To ensure stable solutions, the coefficient d_1 is bounded by $|r_k^r| \leq 0.99$. After the determination of the parameter d_1, the output signals of the time-varying processing determined by (9) or

(10) are used as input signals for the next section with $o(n) := v(n)$ and $u(n) := w(n-1)$. Due to the delays in the FIR lattice filters also time-shifted values $w(n-1)$ are involved, which imply values from the previous frame, too. To yield an estimation which is analogous with the covariance method, longer signals $o(n), u(n), v(n)$ and $w(n)$ with indices $n = -N + i \ldots 2L$ are used for the inverse filtering in each section *i*. For the initialization of the FIR lattice filter, the segments

$$\vec{x}'_k = (x_{k-1}(1+L-N), \ldots x_{k-1}(L), x_k(1), \ldots x_k(L), x_{k+1}(1), \ldots x_{k+1}(L))$$

are used. In comparison to the filtering, the estimation by formula (15) uses segments with constant length of $2L$.

2.1.2. Iterative procedure

The estimation algorithm starts with an initial configuration of reflection coefficients. The initial coefficients can be chosen by coefficients $r_{i,k}^{\text{aut}}$ or $r_{i,k}^{\text{cov}}$ which are determined by a conventional time-invariant linear prediction of the autocorrelation or covariance method, respectively. Then, several iterations are performed for an updating of the coefficients $r_{i,k}^{r}$ by the suboptimal solutions. In each iteration, the coefficients of all frames $k = 1 \ldots P$ are estimated once by formula (15). Since the order of the frames for the updating can have an effect, two orders are used. The first order updates the coefficients of the frames $k = 1 \ldots P$ one after the other, and the second order updates firstly the coefficients of the frames with even indices $k = 2, 4 \ldots$ and then with odd indices. The analyses of speech signals have shown that the convergence by using the order in succession $k = 1 \ldots P$ is faster.

3. ANALYSIS AND SYNTHESIS OF SPEECH

In the following examples of analyzed and synthesized speech signals are shown based on time-invariant and iterative time-varying estimation. For the iterative analysis, the value $\alpha = 0.75$ in (15) is chosen. The analyzed speech signals are pre-emphasized and have a sampling rate of 16 kHz; the order of the lattice filter is 24. In Fig. 3, the analysis results of the German word “weile” [vaIl@] by time-invariant and time-varying estimation are shown. The frame length is $L = 160$ corresponding to 10 ms. In Fig. 3(a)-(d), one magnitude response per frame is depicted. In the case of the time-varying estimation, the magnitude responses which correspond to the coefficients $r_{i,k}^{r}$ are shown. Since the frame length is relatively short, the covariance method yields predominantly better estimation results than the autocorrelation method, which can be seen in Fig. 1(b) and (d). However, the magnitude responses and, especially, the resonance bandwidths are estimated worse by the covariance method in the region of the transition between /v/ and /aI/. It should be noted that overlapping frames can produce smoother trajectories, but the time resolution is worse. In Fig. 1(a) and (c) the estimation results by the time-varying estimation procedure after one and three iterations, respectively, are shown. The starting configurations $r_{i,k}^{r}$ of the iterative procedure are the coefficients $r_{i,k}^{\text{cov}}$ of a time-invariant linear prediction of the covariance method, which is used for the results of Fig. 3(d). The convergence of the iterative procedure is rather fast; therefore, the computational costs are in an acceptable range. The first iteration yields usually already a smoother trajectory, which can be seen from Fig. 3(a) and (d) as well. The use of more iterations can improve the trajectory further, especially, in non-stationary regions. This can be seen for the sound transition [v-aI] in Fig. 3(a) and (c), where the trajectories of the first formants are more appropriate after the third iteration; the interesting magnitude response is indicated by a marker ‘►’ at the left side of Fig. 3.

To asses the impact of the estimation algorithms on synthesis, speech signals are synthesized using the estimated coefficients $r_{i,k}^{r}$ of the time-varying estimation as well as using $r_{i,k}^{\text{aut}}$ and $r_{i,k}^{\text{cov}}$ of the time-invariant estimation. If coefficients $r_{i,k}^{\text{cov}}$ imply unstable systems, the coefficients $r_{i,k}^{\text{cov}}$ are modified to meet stability. For synthesis the IIR lattice filter based on power waves is used. The voiced excitation of the lattice filter is independent from the analyzed speech signal and is based on repeated pitch-modified residual segments of the schwa-sound. For that purpose, the pitch modification algorithm of [8] is used. Since an abrupt changing of the coefficients introduces discontinuities degrading the speech quality, during the synthesis the reflection coefficients are linearly interpolated between the coefficient configurations. The synthesis of several utterances shows that the use of the covariance method can achieve better results than the use of the autocorrelation method due to the more precise estimation. However, the use of the coefficients $r_{i,k}^{\text{cov}}$ can cause artifacts or glitches. For examples, in Fig. 4(a)-(b) a segment of the synthesized speech signal based on the coefficients $r_{i,k}^{\text{cov}}$ and $r_{i,k}^{r}$ corresponding to the magnitude responses of Fig. 3(d) and (c) are shown. The synthesized signal of Fig. 4(a) by using the coefficients of the time-invariant covariance method has an artifact, which decreases the speech quality significantly. A comparison with Fig. 4(b) shows that this artifact is removed by using the updated coefficients with the time-varying estimation. This is valid for all investigated speech utterances. It should be noted that for the coefficients $r_{i,k}^{\text{cov}}$ of this example no modifications due to stability were necessary. To determine this artifact in the model-based domain, in Fig. 3(e) and (f) the magnitude responses of the coefficients $r_{i,k}^{\text{cov}}$ and $r_{i,k}^{r}$ together with those of the interpolated coefficients $(r_{i,k}^{\text{cov}} + r_{i,k+1}^{\text{cov}})/2$ and $(r_{i,k}^{r} + r_{i,k+1}^{r})/2$ in between are shown. From Fig. 3(f), it can be seen that the interpolated configurations of $r_{i,k}^{\text{cov}}$ have regionally strongly fluctuating bandwidths in comparison to those of the time-varying estimation. This is caused by the fact that the coefficients $r_{i,k}^{r}$ are estimated jointly based on a time-varying trajectory, whereas the coefficients of the time-invariant estimation are determined independently.

The example in Fig. 4(c)-(d) demonstrates that the time-varying analysis can be advantageous for the synthesis of plosives. The analysis and synthesis is performed in the same manner as the example of the utterance [vaIl@]. It can be seen that the abrupt starting of the phonation caused by the voiced plosive /b/ can be modeled more accurately by the use of the time-varying estimation than by the time-invariant estimation. Additionally, it can be seen that the artifact at the right side caused by the covariance method can be removed by the time-varying estimation.

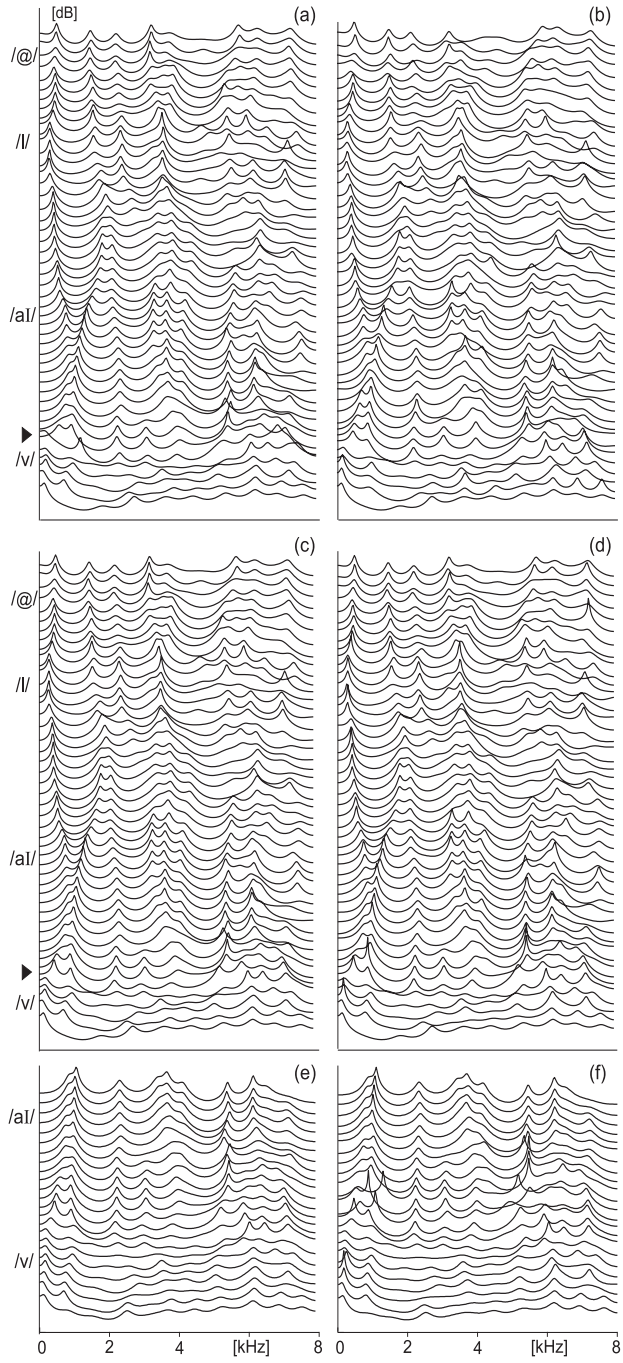

Figure 3: Estimated magnitude responses of utterance [vaIl@]: Iterative time-varying analysis, starting from time-invariant covariance method after one iteration (a) and after three iterations (c),(e); time-invariant analysis by autocorrelation method (b) and covariance method (d),(f); (e) and (f) represent regions from trajectories of (c) and (d), respectively, plus one interpolated configuration in terms of reflection coefficients in between.

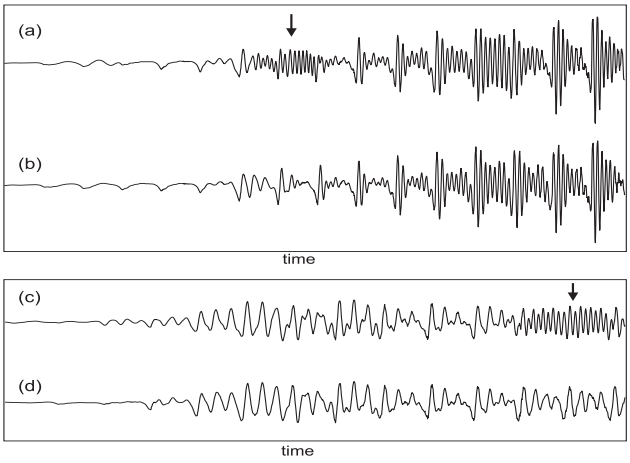

Figure 4: Segments of synthesized speech signals representing [v-aI] from [vaIl@] in (a)-(b) and representing [ba] for (c)-(d). Using analysis results of covariance method (a),(c) and using analysis results of iterative time-varying analysis (b),(d), which starts from analysis results of time-invariant covariance method. Glitches caused by synthesis are marked by arrows.

4. CONCLUSIONS

The proposed time-varying analysis procedure enables an analysis of speech signals which is consistent with the continuous vocal-tract movements and, therefore, also with synthesis by time-varying lattice filters. The time-varying analysis yields a smooth trajectory with an accurate time resolution and spectral modeling. Since the coefficients are estimated time-varyingly in the same manner as they are used for the synthesis, artifacts can be avoided for the synthesized speech.

5. REFERENCES

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

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

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

[5] K. Schnell and A. Lacroix, "Time-Varying Linear Prediction for Speech Analysis and Synthesis," in *Proc. ICASSP*, Las Vegas, pp. 3941-3944, 2008.

[6] K. Schnell, "Time-Varying Burg Method for Speech Analysis," in *Proc. EUSIPCO*, Lausanne Switzerland, pp. 2045-2049, 2008.

[7] J. P. Kaipio and M. Juntunen, "Deterministic Regression Smoothness Priors TVAR Modelling, " in *Proc. ICASSP*, Phoenix USA, 1999.

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