MODEL-BASED ANALYSIS OF SPEECH AND AUDIO SIGNALS FOR REAL-TIME PROCESSING BASED ON TIME-VARYING LATTICE FILTERS

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 paper, a time-varying analysis procedure based on lattice filters is proposed for real-time processing. The analysis procedure estimates the reflection coefficients from the signal frames successively in a way that the current frame is analyzed with respect to previous frames. For that purpose, a linear coefficient trajectory covering the current and previous frames is estimated under condition that the left-sided starting value of the trajectory is prescribed by analysis results of previous frames. The coefficients of the current frame can be determined from the trajectory at a particular point. Analyses of speech signals show that the algorithm yields good spectral modeling simultaneously with a smooth time development.

Index Terms**—** Real-time processing, Signal processing, Speech processing.

1. INTRODUCTION

Analysis and synthesis of speech and audio signals are often performed by a block-wise processing. The frames of the audio signals are usually assumed to be stationary and are analyzed via DFT or by model-based estimations such as linear prediction [1]. Two issues of this approach are not optimum for analysis and can be disadvantageous. Firstly, the frames are analyzed independently and, secondly, stationary statistics within the frames is assumed. For short frames the stationarity can be fulfilled sufficiently; however, the small amount of signal values of the frame increases the uncertainty of the estimation. This can lead to distorted analyses of the frames and/or to fluctuating estimation results from frame to frame, both of then decrease the audio quality of a succeeding synthesis based on the analyzed frames. This is especially interesting when very short frame lengths are desired for a high time resolution, additionally, a short frame length is usually favorable to meet the demands of real-time low-delay processing for applications like speech transmission and/or enhancement. This problem can be tackled by using explicitly time-varying models for the estimation. One group of time-varying estimation algorithms is the class of adaptive filters like LMS or Kalman filtering [2], [3]. A more analytical estimation approach is the expansion of the coefficient trajectory by basis functions [4], [5]. The time-varying trajectory in terms of direct-form coefficients can be estimated for a multi-frame analysis in an analytical way [6]; however, the trajectories of the direct-form coefficients don't produce everywhere usable coefficient configurations, besides the problem of unstable system configurations. The reflection coefficients of the lattice filter are more appropriate for modeling time-varying trajectories. In [7] an estimation algorithm of coefficient trajectories in terms of reflection coefficients is proposed; however, this approach needs for usable results a relatively long overlapping to the right side of the analyzed frame. To allow also low-delay processing, in this paper a procedure is proposed which is suitable for analyses with small frame lengths and needs, additionally, no right-sided overlapping. Furthermore, the proposed estimation procedure yields a better time resolution, too.

2. TIME-VARYING ANALYSIS

The FIR lattice filter shown in Fig.1 is used for the analysis by inverse filtering. Due to the time-varying estimation, the reflection coefficients *r* are time-varying within in each frame.

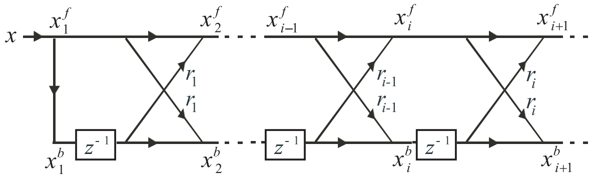

Figure 1: FIR lattice filter for inverse filtering.

2.1. Estimation procedure

For the estimation, the audio signal *s* is analyzed frame by frame from left to right. Each frame s_k is analyzed on the basis of the current frame s_k and previous frames s_{k-j}. The frames $s_k = s((k-1)L+1,\dots,kL)$ have the length L and border on each other directly without overlapping. For the estimation of the current frame s_k, longer segments $\overline{s}_k$ with length $L' > L$ are used for the analysis. The segments $\overline{s}_k = (s_{k-M},\dots,s_{k-1},s_k)$ used for the analysis include each the current frame s_k and are expanded with *M* previous frames to enlarge the amount of data for the estimation. The basic idea of the estimation procedure is to determine a coefficient vector for the current frame based on an estimated time-varying trajectory which covers the current frame and previous frames. In this way, previous data of last frames can be integrated consistently into the estimation by considering the time evolution of the underlying model. The segments $\overline{s}_k$ are pre-emphasized by

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

an adaptive pre-emphasis $P_k(z) = (1-\rho_1 z^{-1})(1-\rho_2 z^{-1})$ resulting in the pre-emphasized segments

$$\bar{x}_k = \bar{s}_k * p_k \,. \tag{1}$$

The two pre-emphasis coefficients ρ_l with $l = 1,2$ are determined for each segment by a repeated linear prediction of first order. In the next paragraph, the estimation of each section of the FIR lattice filter is treated.

2.1.1. Estimation of each section

The reflection coefficients r_i are estimated successively by minimizing the output powers of each section of the FIR lattice filter. Fig. 2 shows the *i*-th section of the FIR lattice filter. The first section is initialized by $x^f_{1,k} = \bar{x}_k$ and $x^b_{1,k} = \bar{x}_k$. Since now a single section is treated exemplarily, the index *i* of the section is left out with $r = r_i$ for example. Furthermore, the input and output signals are denoted by $o(n) = x^f_{i,k}(n)$, $u(n) = x^b_{i,k}(n-1)$, $v(n) = x^f_{i+1,k}(n)$, and $w(n) = x^b_{i+1,k}(n)$ for a readable description (compare Fig. 2).

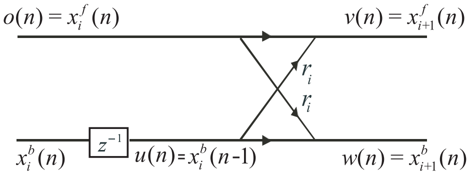

Figure 2: *i*-th section of FIR lattice filter.

The trajectory $\bar{r}(n)$ of a time-varying reflection coefficient is assumed to be linear for each analyzed segment $\bar{x}_k$. Hence, the trajectory can be expressed by a linear combination of a constant or starting value r^{st} and a linear basis function $\phi(n)$ resulting in

$$\bar{r}(n) = r^{\text{st}} + d \cdot \phi(n) \tag{2}$$

with $\phi(n) = (n-1)/(L'-1)$ for $n = 1 \ldots L'$.

The parameters r^{st} and d describe the constant and time-varying component respecting the linear trajectory. Since the basis function $\phi(n)$ has values reaching from zero to one with $\phi(1) = 0$ and $\phi(L') = 1$, the coefficient trajectory starts with $\bar{r}(1) = r^{\text{st}}$ and ends with $\bar{r}(L') = r^{\text{st}} + d$. If analysis results of left-sided frames are available, the starting value r^{st} is determined from the coefficients of previous frames. This can be motivated by the demand of a continuous trajectory which is generally favorable for applications of analysis and synthesis. Hence, for the estimation of the trajectory the starting value r^{st} can be assumed to be fixed so that only the coefficient d is to be estimated. d describes the slope of the trajectory. The output signals

$$\begin{aligned} v(n) &= o(n) + \bar{r}(n)u(n) = o(n) + (r^{\text{st}} + d \cdot \phi(n))u(n) \\ w(n) &= u(n) + \bar{r}(n)o(n) = u(n) + (r^{\text{st}} + d \cdot \phi(n))o(n) \end{aligned} \tag{3}$$

of a section can be expressed by

$$\begin{aligned} v(n) &= o(n) + r^{\text{st}} \cdot u(n) + d \cdot \tilde{u}(n) \\ w(n) &= u(n) + r^{\text{st}} \cdot o(n) + d \cdot \tilde{o}(n) \end{aligned} \tag{4}$$

with the definitions

$\tilde{o}(n) = \phi(n) \cdot o(n)$ and $\tilde{u}(n) = \phi(n) \cdot u(n)$.

The coefficient d can be estimated by minimizing the powers of $v(n)$ and $w(n)$. To increase the robustness of the estimation, also a time-invariant estimation is considered for the estimation. To integrate the time-invariant estimation consistently with the time-varying approach, the time-invariant estimation contributes only to the estimation of the trajectory at one point. The relationship between the time-varying trajectory and the reflection coefficient r^{ti} corresponding to the time-invariant estimation is defined by $r^{\text{ti}} = r^{\text{st}} + \beta d$. β is a fixed parameter and determines the position $n = \beta L'$ of the coefficient $r^{\text{ti}} = \bar{r}(\beta L')$ related to the time-varying trajectory. Hence, it is sensible to chose the segment for the time-invariant estimation in a way that the segment is centered at the position $n = \beta L'$. This is valid for segments with indices $[n'] = L' - 2(1-\beta)L' \ldots L'$. A sketch of the trajectory is shown in Fig. 3. For the time-invariant estimation, a weighting of the input signals is performed by a Hamming window υ of length $2(1-\beta)L'+1$ with $\hat{o} = \upsilon o$ and $\hat{u} = \upsilon u$. Then, the output signals of the time-invariant case are given by

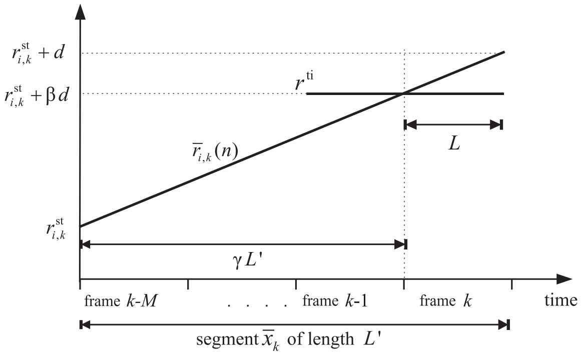

Figure 3: Trajectory $\bar{r}(n)$ for the analysis of frame *k*.

$$\begin{aligned} \hat{v}[n'] &= \hat{o}[n'] + r^{\text{ti}}\hat{u}[n'] = \hat{o}[n'] + (r^{\text{st}} + \beta d)\hat{u}[n'] \\ \hat{w}[n'] &= \hat{u}[n'] + r^{\text{ti}}\hat{o}[n'] = \hat{u}[n'] + (r^{\text{st}} + \beta d)\hat{o}[n']. \end{aligned} \tag{5}$$

For the estimation of the parameter *d*, the error to be minimized is defined by

$$e(d) = \mathsf{E}\left[\alpha(v^2 + w^2) + (1-\alpha)(\hat{v}^2 + \hat{w}^2)\right] \to \min \,. \tag{6}$$

The error *e* is a linear combination of the powers of the output signals corresponding to the time-varying and time-invariant case. By analogy with the Burg method, the arithmetic mean of the two output signals is chosen for both cases. To yield a minimum error *e*, the derivate of the error with respect to the coefficient *d* is set to be equal to zero

$$\frac{\partial e(d)}{\partial d} = 0 \,. \tag{7}$$

Solving (7) for *d* leads to the formula of the optimum coefficient:

$$d = -\mathsf{E}\left[\frac{\alpha(\tilde{u}o + u\tilde{o} + r^{\text{st}}(\tilde{u}u + o\tilde{o})) + \lambda(2\hat{u}\hat{o} + r^{\text{st}}(\hat{u}^2 + \hat{o}^2))}{\alpha(\tilde{u}^2 + \tilde{o}^2) + \lambda\beta(\hat{u}^2 + \hat{o}^2)}\right] \tag{8}$$

with $\lambda = (1-\alpha)\beta$. The coefficient of formula (8) minimizes the error under the condition of the prescribed parameter r^{st}. The parameter α influences the ratio of the time-invariant and time-varying component; e.g. $\alpha = 1$ leads to a vanishing contribution of the time-invariant estimation. The expected value E is calculated by the means of the signal values. To force stable solutions, the value of the parameter d is bounded with respect to $|r^{\text{st}}+d| \leq 0.99$. It should be noted that the modification concerning stable solutions is very scarce. After the calculation of the parameter d, the output signals of the time-varying processing determined by (3) are used as the input signals of the next section with $o(n) := v(n)$ and $u(n) := w(n-1)$.

2.1.2. Estimation of adjacent frames

Since now the successive analysis of adjacent frames is treated, the signals and coefficients have two lower indices i and k for the section number and the frame number, respectively. For example, $r_{i,k}$ is the i-th reflection coefficient corresponding to the k-th frame. The frames are analyzed successively from left to right. Since for the first frame no analysis results of left-sided frames are available, the first frame is analyzed by the conventional time-invariant Burg method. The next frames are analyzed with respect to the results of the left-sided frames. For the following description, the existence of left-sided frames is taken for granted. Since the multi-frame segments $\bar{x}_k$ are used for the analysis, the first left-sided reflection coefficient $\bar{r}_{i,k}(1)$ of the trajectory corresponds to the frame $k-M$. Therefore, the coefficients of the frame $k-M$ and its adjoining frames can be exploited to determine the starting values $r^{\text{st}}_{i,k}$. The values $r^{\text{st}}_{i,k}$ are determined by a weighted mean around the interesting frame $k-M$ by

$$r^{\text{st}}_{i,k} = \sum_{m=k-M-W}^{k-M+W} w_m r_{i,m} \quad \text{with} \quad \sum_m w_m = 1\,. \qquad (9)$$

The weights w_m have the maximum value in the central position and are decreasing towards both ends. With the aid of the value $r^{\text{st}}_{i,k}$ the coefficient $d_{i,k}$ can be estimated by equation (8). The resulting reflection coefficients corresponding to frame k are

$$r_{i,k} = r^{\text{st}}_{i,k} + \gamma \cdot d_{i,k}\,. \qquad (10)$$

The parameter γ determines at which position the values of the coefficient trajectories are chosen as coefficients for the frame. $\gamma = M/(M+1)$ implies as position the left side of the k-th frame as illustrated in Fig. 3, whereas the value $\gamma = 1$ implies the right side of the k-th frame. The analyses have shown that values of γ which tend more to 0.5 produce smoother coefficient trajectories than those tending to one. For the frames $k = 2 \ldots M$, the procedure is modified considering the fact that less than M left-sided frames are available. In this way the frames can be analyzed successively from left to right. The computational costs can be derived from (8) and depend, additionally, on the parameter M.

3. ANALYSIS OF SPEECH SIGNALS

In the following, analysis results of the time-varying approach and of the conventional linear prediction are shown. Other analysis results which are not depicted show similar effects. The analyzed signals are speech signals of 16 kHz sampling rate and the order of the system of the time-varying and time-invariant analysis is 24. For the time-varying analysis, the value $\alpha = 0.75$ is chosen in (8); it should be mentioned that this value is not optimized. Fig. 4 shows analysis results of the utterance [julI] and Fig. 5 shows analysis results of the same utterance, however, corrupted by added babble noise. The analyses of Fig. 4(b)-(d) and Fig. 5 imply a processing delay of 5 ms due to the block length, whereas the analysis of Fig. 4(a) implies a processing delay of 2.5 ms. It can be seen that the time-varying analyses yield smooth trajectories, which is mainly caused by the consideration of previous frames with the starting value of the linear trajectory. In this way a joint analysis of the frames is realized. In contrast to that, the time-invariant estimation analyzes the frames independently which causes fluctuations of the trajectories. The time-invariant covariance method yields better spectral estimations than the autocorrelation method; however, the covariance method

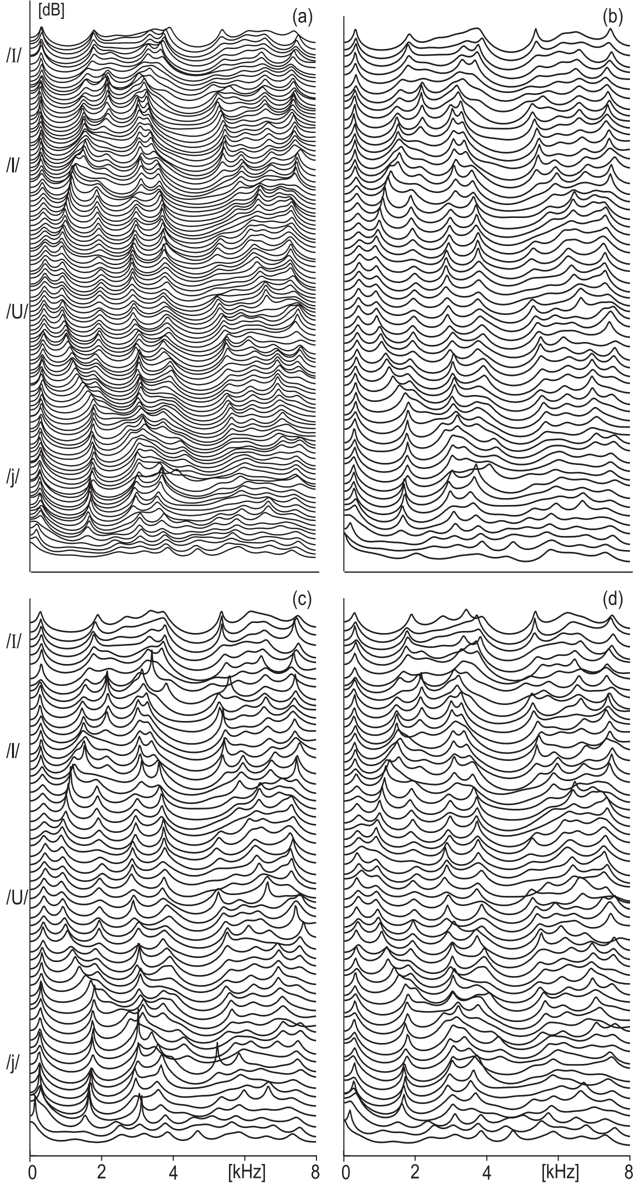

Figure 4: Estimated magnitude responses of utterance [julI]: (a)-(b) time-varying analysis, (a) with $L = 40$ samples (2.5 ms overlap) and (b) with $L = 80$ (5 ms overlap). (c)-(d) conventional time-invariant analysis by covariance method (c) and autocorrelation method (d) with block length of 10 ms and an overlap of 5 ms.

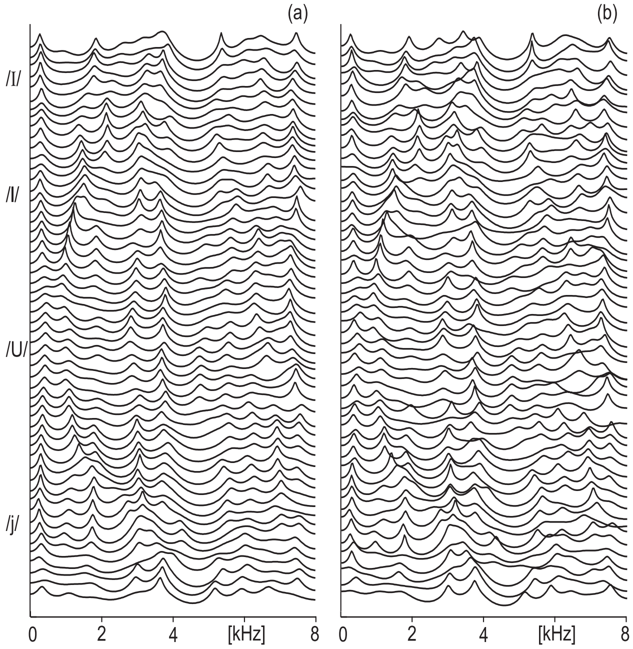

Figure 5: Estimated magnitude responses of utterance [julI] with added babble noise of 20 dB SNR (SNR for speech only): (a) time-varying analysis with $L = 80$ (5 ms overlap). (b) conventional time-invariant analysis by autocorrelation method (block length 10 ms and overlap of 5 ms).

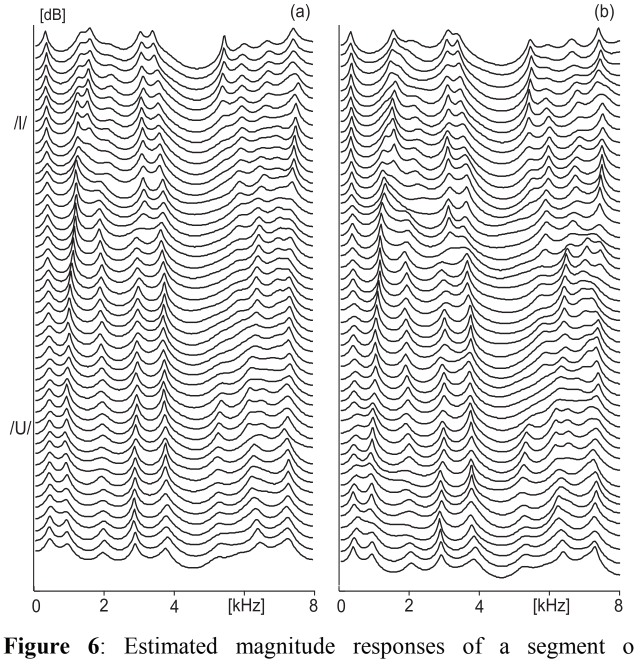

Figure 6: Estimated magnitude responses of a segment of utterance [julI]: (a) time-varying analysis with $L = 24$ (1.5 ms overlap). (b) conventional time-invariant analysis by autocorrelation method with block length of 10 ms and an overlap of 1.5 ms.

produces here and there unstable solutions or runaways. Concerning that, it is known that the time-invariant autocorrelation method provides a more robust estimation. The joint estimation of the segments by the time-varying approach yields both a good and robust spectral estimation generating a smooth trajectory. The time-varying analysis is more robust against noise, since the time-varying analysis maintains generally the smoothness of the trajectory, whereas the time-invariant estimation is affected stronger by noise, which can also be seen from a comparison of Figs. 4 and 5. Fig. 6 shows the results of the analysis of a segment of the utterance [julI] (clean speech) implying a processing delay of 1.5 ms due to the block length. Also in this case, the advantage of the time-varying estimation can be seen. For all analyses, the values of the parameters β and γ are chosen in a way that the temporally point of the spectral estimation is 5 ms before the last value of the analyzed segment at $n = L' - 80$. The length L' is about 160 samples which corresponds to 10 ms. The block length of the time-invariant estimation corresponds to 10 ms, too.

To asses the estimation results for an analysis-synthesis approach, synthesis is performed by the estimated coefficients. For that purpose, the IIR lattice filter is controlled by the estimated coefficients. The lattice filter is excited by a prescribed excitation, since any estimation error influences the residual signal. The excitation is noise or an impulse train depending of the type of sound. The analysis-synthesis results show that the smooth trajectory of the time-varying analysis is favorable concerning the yielded audio quality.

4. CONCLUSIONS

A model-based analysis procedure for real-time applications is proposed by exploiting time-varying analysis techniques. The estimation of a time-varying trajectory can be used for a joint analysis of segments. As a result of that, a smooth trajectory of the coefficients from frame to frame can be estimated with a good spectral estimation and high time resolution. The analysis results show that the time-varying approach is suitable for both processing of clean and noisy speech.

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] 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.

[4] 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.

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

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

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