Speech Analysis and Synthesis by Time-Varying Lattice Filters

K. Schnell, A. Lacroix

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

Introduction

Model-based analysis of speech is often performed by time-invariant linear prediction [1]. However, the articulatory speech production process is a continuous non-stationary process where a time-varying analysis is adequate. In comparison to the time-invariant estimation approach, a time-varying analysis considers explicitly time-varying model parameters within the frames. A general approach to perform time-varying analyses is given by adaptive filtering procedures [2]. An analytical approach is given if the time-varying coefficients of a segment are modelled by basis functions, which can be treated in terms of direct-form [3] or reflection coefficients [4]. In [5] a time-varying analysis procedure based on basis functions is proposed which estimates a continuous piece-wise linear trajectory in terms of reflection coefficients under the constraints of a continuous time evolution of the model parameters over frames. In this contribution, an extended procedure of [5] is used, which is proposed in [6] and improves the estimation results by an iterative re-estimation of the coefficients. Depending on the aim of the estimation, different adjustments can be favourable such as the segment length.

Time-Varying Analysis

The time-varying analysis considers a continuous piece-wise linear trajectory in terms of reflection coefficients. Prior to the analysis, the speech signal s is pre-emphasized leading to the signal x and segmented into frames x_k of uniform length L. The trajectory of the i-th reflection coefficient of the k-th frame is linear and can be expressed by

$$r_{i,k}(n) = c_{i,k} + d_{i,k} \cdot (n-1)/(L-1) \qquad (1)$$

for $n = 1 \ldots L$. The coefficients to be estimated are the boundary coefficients $r_{i,k}(1)$ and $r_{i,k}(L)$ of each frame. Since the trajectory is assumed to be continuous, the values of the corresponding right-sided and left-sided boundary coefficients of adjacent frames are equal

$$r_{i,k}^r = r_{i,k}(L) = r_{i,k+1}(1) \, . \qquad (2)$$

The right-sided boundary coefficients $r_{i,k}^r = c_{i,k} + d_{i,k}$ and the trajectories of the frames can be described by the parameters $c_{i,k}$ and $d_{i,k}$. For these parameters eq. (2) results in

$$c_{i,k+1} = c_{i,k} + d_{i,k} \, . \qquad (3)$$

Since the continuity conditions of eqs. (2)-(3) are valid, only the coefficients $d_{i,k}$ have to be estimated, except for the first frame. The estimation of the coefficients is based on time-varying inverse filtering. The inverse filter in terms of reflection coefficients is depicted in fig. 1. Analogously to the time-invariant estimation of the Burg method, the coefficients for one frame are estimated one after the other. The estimation of each coefficient is performed by minimizing the output powers of the corresponding section which is depicted in fig. 2. For a compact notation, also the declarations of the input and output signals of fig. 2 are valid with $o(n) = x_i^f(n)$, $u(n) = x_i^b(n-1)$, $v(n) = x_{i+1}^f(n)$, and $w(n) = x_{i+1}^b(n)$.

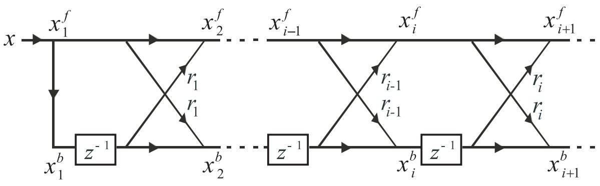

Figure 1: FIR lattice filter for inverse filtering.

In [5] an estimation algorithm performing a frame-by-frame analysis from left to right is proposed, which estimates a piece-wise continuous trajectory. However, due the consideration of only left-sided frames for the estimation of each frame, the estimation is not optimum. Hence, in the next section an iterative updating procedure of the coefficients is explained.

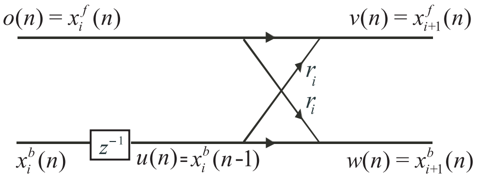

Figure 2: Section of FIR lattice filter for the estimation.

Iterative re-estimation of the coefficients

Based on an initial coefficient configuration, each coefficient is re-estimated iteratively by the analysis results of the left-sided and right-sided frames. In the following, the task is to re-estimate the coefficient $r_{i,k}^r$ by the fixed coefficients $r_{i,k-1}^r$ and $r_{i,k+1}^r$. For that purpose, the segment $\overline{x}_k$ for the estimation is defined which covers the k-th and k+1-th segment

$$\overline{x}_k = (x_k(1), \ldots, x_k(L), x_{k+1}(1), \ldots, x_{k+1}(L))^{\mathsf{T}} \, . \qquad (4)$$

Analogously, the coefficient trajectory

$$\overline{r}_{i,k} = (r_{i,k}(1),\ldots,r_{i,k}(L),r_{i,k+1}(1),\ldots,r_{i,k+1}(L))^{\mathrm{T}} \qquad (5)$$

of the segment $\overline{x}_k$ is defined. In the following, the estimation of the i-th right-sided coefficient $r_{i,k}^r = \overline{r}_{i,k}(L)$ of the k-th frame is treated. For abbreviation, the index i and k if possible is left out leading to e.g. $r_k^r := r_{i,k}^r$. The trajectory $\overline{r}_k(n)$ can be described by a linear combination of two basis functions ϕ_1 and ϕ_2 by

$$\overline{r}_k(n) = c + d_1\phi_1(n) + d_2\phi_2(n) \qquad (6)$$

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

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

Due to the definition of the segment $\overline{x}_k$, the relationships between the coefficients and the trajectory are $r_{k-1}^r = \overline{r}_k(1) = c$, $r_k^r = \overline{r}_k(L) = c + d_1 + 0.5d_2$ and $r_{k+1}^r = \overline{r}_k(2L) = c + d_2$. The segment and the basis functions are illustrated in fig. 3. The basis function ϕ_2 determines the right-sided boundary value $\overline{r}_k(2L)$ alone since $\phi_1(2L) = 0$ is valid. The basis function ϕ_1 determines the central value $\overline{r}_k(L)$. Hence, for estimating the coefficient $r_k^r = \overline{r}_k(L) = c + d_1 + 0.5d_2$ the coefficient d_1 only has to be estimated. The other coefficients are prescribed by the fixed coefficients by $c = r_{k-1}^r$ and $d_2 = r_{k+1}^r - r_{k-1}^r$.

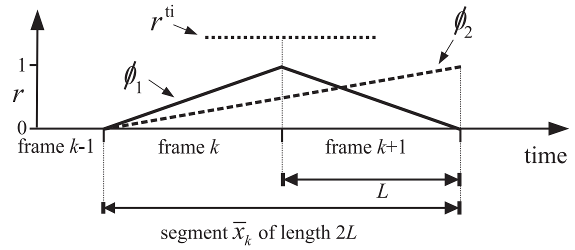

Figure 3: Basis functions ϕ_1 and ϕ_2 of time-varying part and segment for r^{ti} of time-invariant part for the estimation of the trajectory $\overline{r}_k(n)$.

To stabilize the estimation of the coefficients $r_k^r = \overline{r}_k(L)$, also a time-invariant estimation component represented by the coefficient r_k^{ti} is included. For r_k^{ti} a segment is considered which encloses the position L of the segment $\overline{x}_k$ centrically as illustrated by the dotted line in fig. 3. Since the coefficient $r_k^r = \overline{r}_k(L)$ is at position L, the relationship between the coefficient r_k^{ti} of the time-invariant component and the time-varying trajectory is

$$r_k^{\text{ti}} = r_k^r = c + d_1 + 0.5d_2\,. \qquad (7)$$

The coefficient d_1 of the basis function ϕ_1 is estimated by minimizing the error

$$e(d_1) = \mathrm{E}\left[\alpha((v)^2 + (w)^2) + (1-\alpha)((\widehat{v}^{\text{ti}})^2 + (\widehat{w}^{\text{ti}})^2)\right] \qquad (8)$$

which includes the signals

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

of the time-varying inverse filtering and the signals

$$\begin{aligned} \widehat{v}^{\text{ti}} &= \widehat{o} + r_k^{\text{ti}} \cdot \widehat{u} = \widehat{o} + (c + d_1 + 0.5\cdot d_2)\widehat{u} \\ \widehat{w}^{\text{ti}} &= \widehat{u} + r_k^{\text{ti}} \cdot \widehat{o} = \widehat{u} + (c + d_1 + 0.5\cdot d_2)\widehat{o} \end{aligned} \qquad (10)$$

of the time-invariant filtering. The signals $\widehat{o}(n)$ and $\widehat{u}(n)$ are the hamming-windowed versions of the output signals for the segment with indices $n = L - M \ldots L + M$ corresponding to r_k^{ti} in fig. 3. The parameter α in eq. (8) adjusts the impact of the time-invariant and time-varying estimation components. The optimum coefficient can be derived by the derivative of eq. (8) with $\partial e/\partial d_1 = 0$ resulting in the formula

$$d_1 = -\,\mathrm{E}\left[\frac{\alpha(\varepsilon_{\text{n}}^{\text{tv}}) + (1-\alpha)(\varepsilon_{\text{n}}^{\text{ti}})}{\alpha(\varepsilon_{\text{d}}^{\text{tv}}) + (1-\alpha)(\varepsilon_{\text{d}}^{\text{ti}})}\right] \quad \text{with} \qquad (11)$$

$$\begin{aligned} \varepsilon_{\text{n}}^{\text{tv}} &= \tilde{u}_1 o + \tilde{o}_1 u + r(\tilde{u}_1 u + \tilde{o}_1 o) + d_2(\tilde{u}_1\tilde{u}_2 + \tilde{o}_1\tilde{o}_2) \\ \varepsilon_{\text{d}}^{\text{tv}} &= (\tilde{u}_2)^2 + (\tilde{o}_2)^2 \end{aligned}$$

$$\begin{aligned} \varepsilon_{\text{n}}^{\text{ti}} &= \widehat{u}^{\text{ti}}\widehat{o}^{\text{ti}} + \widehat{o}^{\text{ti}}\tilde{u}^{\text{ti}} + (r + 0.5\cdot d_2)((\widehat{u}^{\text{ti}})^2 + (\widehat{o}^{\text{ti}})^2) \\ \varepsilon_{\text{d}}^{\text{ti}} &= (\widehat{u}^{\text{ti}})^2 + (\widehat{o}^{\text{ti}})^2. \end{aligned}$$

The expectation value E can be calculated by the means of the signal values; the signals $\tilde{o}_l(n) = \phi_l(n)\cdot o(n)$ and $\tilde{u}_l(n) = \phi_l(n)\cdot u(n)$ for $l = 1,2$ are the input signals which are weighted by the basis functions. Eq. (11) represents an optimum solution overall only if the prescribed coefficients r_{k-1}^r and r_{k+1}^r are optimum. Formula (11) is applied for the coefficients from $i = 1$ to $i = N$. After the estimation of r_k^r the output signals are calculated by time-varying filtering of the estimated section. For the re-estimation eq. (11) is applied to all coefficients of all frames, which describes one iteration. The analyses show that only few iterations are necessary for convergence.

Time-Varying Analysis

Fig. 4 shows estimation results obtained by the iterative time-varying analysis from the German utterance "Julia". The sampling rate of the analyzed speech signal is 16 kHz. The prediction order is 30 and the segment length is 200 samples. The analysis results are robust against moderate changes of the parameter α. Here, the parameter is chosen by $\alpha = 0.66$. The magnitude responses of the trajectory are shown which correspond to the coefficients r_k^r and to the centrally interpolated configurations $(r_k^r + r_{k+1}^r)/2$. It can be seen that the trajectory is rather smooth and shows the movements of the formants.

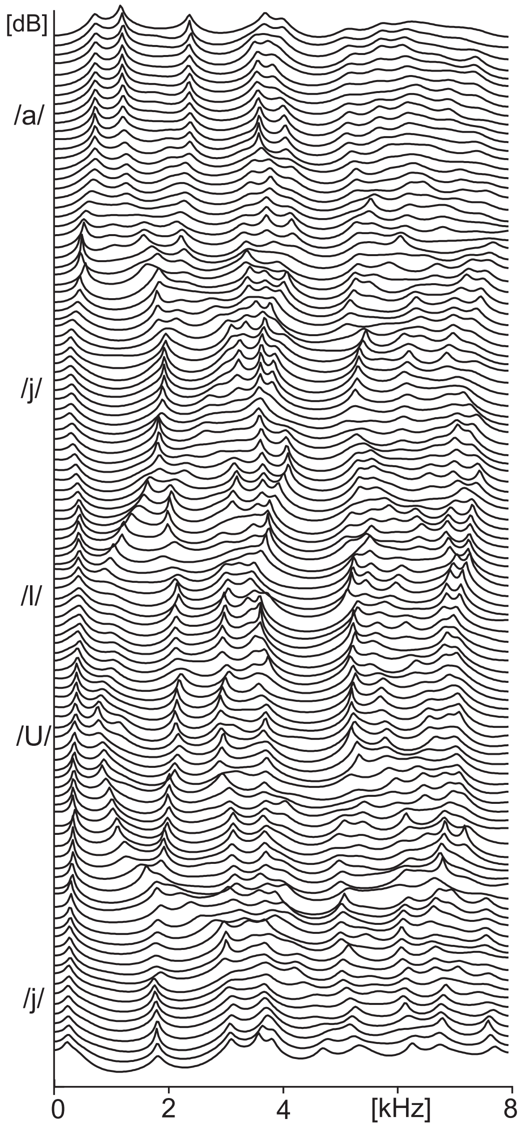

Figure 4: Estimated magnitude responses of r_k^r and $(r_k^r + r_{k+1}^r)/2$ from utterance "Julia" by iterative time-varying analysis with segment length $L = 200$.

Effect of time-varying glottis

In the following example the influence of the time-varying glottis is discussed. As a result of the vocal-fold vibrations, the vocal-tract termination at the vocal chords functions acoustically as a time-varying spatially concentrated loss. To achieve a time resolution which covers the vocal-fold vibrations, a small segment length is chosen. Fig. 5 shows estimation results by a segment length of $L = 40$ obtained from a stationary speech signal of vowel /a:/ of 16 kHz sampling rate. In addition to the boundary configurations $r_k^r = \overline{r}_k(L)$ also one interpolated configuration $(r_k^r + r_{k+1}^r)/2$ per segment is shown. Hence, the distance between the magnitude responses corresponds to 20 samples. Since the pitch period lengths of the speech signal /a:/ are between 100 and 120 samples, the open and closed glottis phases are about 50-60 samples long, which corresponds to three configurations. This is in coincidence with the estimation results of fig. 5. The formants can be observed clearly in the closed phase in contrast to the open phase of the glottis which widens the formant bandwidths and, additionally, alters the vocal tract transfer function by the coupled sub-glottal tract.

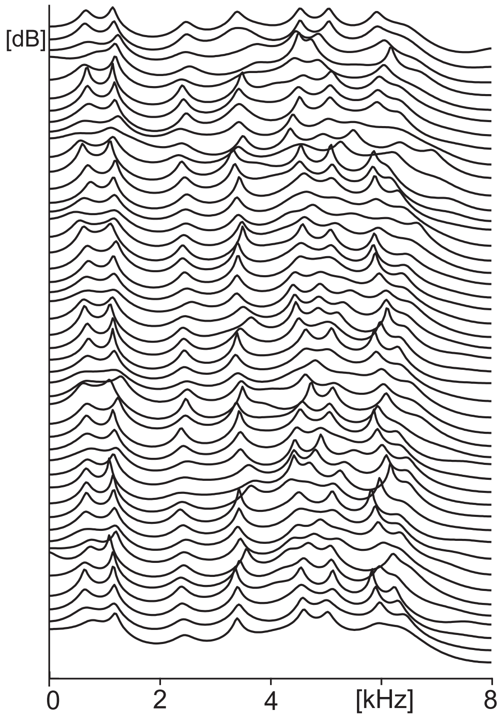

Figure 5: Estimated magnitude responses from vowel /a:/ by the iterative time-varying analysis with segment length $L = 40$, configurations of r_k^r plus one interpolated configuration in terms of reflection coefficients in between.

Time-Varying Analysis and Synthesis

The estimated coefficients can be used for synthesis. For that purpose, the time-varying IIR-lattice filter is used. The vocal-tract movements are usually comparatively slow; however, for particular sounds also fast articulatory movements occur. Especially, plosives include fast vocal-tract movements since the tongue or the lips are snapping back for the release of the closure. Therefore, a fine time resolution is favorable to cover such events [7]. In the following example the analysis and synthesis of voiced plosives are discussed. Fig. 6 shows synthesized speech signals based on an analysis of the German word "audio" [aUdIo:]. For re-synthesis, the coefficients are controlled by the estimated coefficients r_k^r from the analyzed speech. During the filtering, the reflection coefficients are interpolated linearly between the configurations r_k^r . The excitation of the time-varying lattice filter is independent of the analyzed speech signals and consists of a repeated segment of the schwa sound LPC-residual, which is pitch-modified for the prescribed fundamental frequency. The pitch modification is carried out for each period and is based on the algorithm proposed in [8]. For the results of fig. 6(a)-(c), the estimation results of a time-invariant linear prediction are used for r_k^r. In comparison to that, fig. 6(d)-(f) represents synthesized signals using the estimation results of the iterative time-varying analysis. To demonstrate the effect of the segmentation, the segment lengths $L = 320$, $L = 240$, and $L = 160$ are used for the fig. 6(a)/(d), (b)/(e), and (c)/(f), respectively. The graphs in fig. 6 show a part of the synthesized speech signal which represents [..UdI.]. This part includes the voiced plosive /d/. It can be seen that by using the time-varying estimation results, the discontinuity of the plosive, which is indicated by an arrow in fig. 6(e)-(f), can be observed for the segment lengths 240 and 160. In contrast to that, the time-invariant estimation results give no or weaker indications for the burst of the plosive. Perceptive tests show that the plosive of the synthesized signals of fig. 6(e)-(f) can be clearly perceived in comparison to the other shown cases.

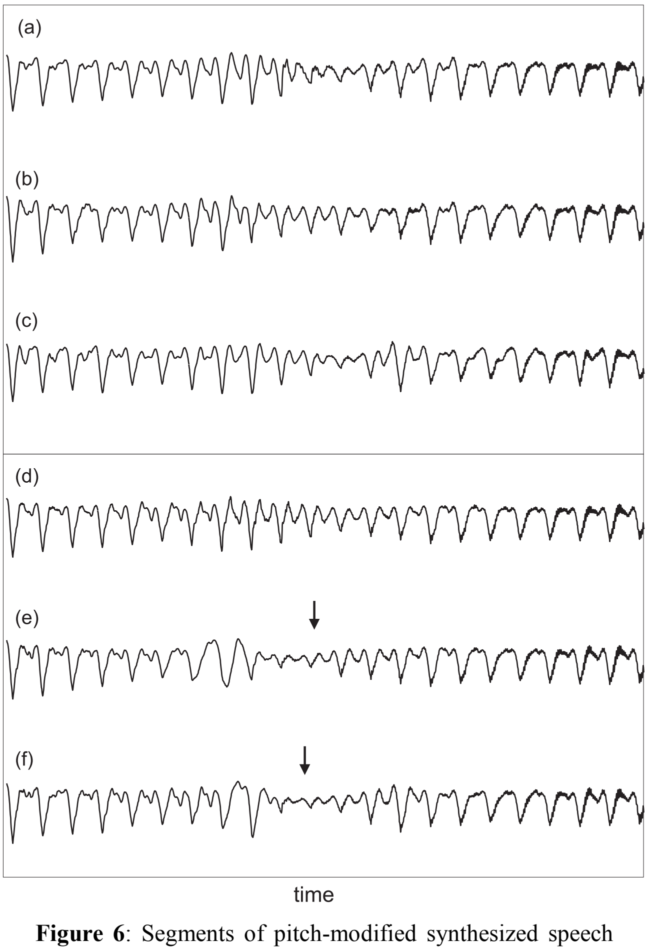

Figure 6: Segments of pitch-modified synthesized speech signals representing [..UdI..] from [aUdIo:]: (a)-(c) using analysis results of time-invariant autocorrelation method and (d)-(f) using those of the iterative time-varying analysis; frame length $L = 320$ for (a)/(d), $L = 240$ for (b)/(e), and $L = 160$ for (c)/ (f).

Conclusions

The time-varying analysis is discussed for the analysis and synthesis of speech signals. The analyses show that the estimation procedure is able to estimate piece-wise linear trajectories in terms of reflection coefficients. To cover the influence of the vocal tract transfer function alone or in combination with the time-varying glottis termination, the segment length should be chosen greater or smaller than the period length of the fundamental frequency, respectively. For covering the relatively fast vocal-tract movements of plosives, the segment length should be chosen rather small but excluding the vocal-fold vibrations.

References

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

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

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

[5] Schnell, K.: "Time-Varying Burg Method for Speech Analysis", in Proc. EUSIPCO'08, Lausanne Switzerland, 2008.

[6] Schnell, K.; Lacroix, A.: " Iterative Inverse Filtering by Lattice Filters for Time-Varying Analysis and Synthesis of Speech", in Proc. ICASSP'09, Taipei Taiwan, 2009.

[7] Kaipio, J. P.; Juntunen, M.: "Deterministic Regression Smoothness Priors TVAR Modelling,", in Proc. ICASSP'99, Phoenix USA, 1999.

[8] Schnell, K.: "Pitch Modification of Speech Residual Based on Parameterized Glottal Flow with Consideration of Approximation Error", in Proc. ICASSP'06, Toulouse France, pp. 737-740, 2006.