



Time-Varying Pre-emphasis and Inverse Filtering of Speech

Karl Schnell, 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 method is applied to speech processing. In contrast to the commonly used linear prediction approach, the proposed time-varying method considers the continuous time evolution of the vocal tract and, additionally, avoids block-wise processing. On the assumption that the linear predictor coefficients evolve linearly in sections and continuously over the whole signal, the optimum time-varying coefficients can be determined quasi-analytically by a least mean square approach. The investigations show that the method fits very well the realization of a time-varying pre-emphasis. Furthermore, the results show that the method is suitable for time-varying inverse filtering.

Index Terms: time-varying prediction, pre-emphasis, inverse filtering

1. Introduction

Linear prediction is a basic technique in the field of speech processing [1, 2]. The products of the linear prediction are, firstly, the predictor coefficients and, secondly, the residual or error signal. Both the coefficients and the residual can be used for several applications such as speech analysis, coding, and synthesis. The residual signal itself, for example, can be used for voice analysis or pitch modification [3-6]. The common linear prediction approach implies a block-wise analysis of speech signals using the autocorrelation or covariance method; due to the block-wise analysis, windowing of the frames is appropriate. Prior to the actual prediction analysis, usually a pre-emphasis is performed. The pre-emphasis compensates the spectral decrease eliminating the influence of the lip radiation and voiced excitation on the spectral envelope; this is relevant, for example, for the estimation of formants or vocal tract areas [7]. An adaptive pre-emphasis can be realized by a linear prediction of order one [1]. In contrast to the common linear prediction approach assuming a stationary signal, also time-varying analysis approaches exist considering that the vocal tract is changing more or less during the speech segment. A general approach is to use adaptive algorithms like LMS or RLS [8]. The autoregressive modeling technique (TVAR) estimates an AR model with time-varying coefficients; the coefficients can be described by basis functions, e.g. in [9, 10]. In [11] an approach using time-frequency shifts is given. Besides all-pole modeling, also approaches using ARMA modeling of non-stationary signals exist [12]. In this contribution, a time-varying prediction based on AR model is proposed allowing the joint estimation of adjacent segments. For that purpose, the predictor coefficients are constricted in a way that the predictor coefficients are connected continuously between the frames and that the coefficients evolve linearly within the frames; this allows a quasi-analytical estimation procedure analyzing a sequence of adjacent speech segments.

2. Time-varying linear prediction

The predictor equation of a time-varying linear prediction can be generally described by

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

$\hat{x}(n)$ is the estimation of $x(n)$. In comparison to the common time-invariant prediction, the predictor coefficients depend on the time index n . For the proposed approach, the time dependence of the coefficients $a_i(n)$ is prescribed linearly within the frames. Therefore, for one frame the evolution of each coefficient is defined by a straight line with the constant term c and slope d . The speech signals are segmented into adjoining frames with the markings $m[k]$; $m[k]$ is the starting index of the k -th frame with the length $l[k] = L[k] + 1$. Since the frames are joined without overlapping, the relationship between the markings of adjacent frames can be described by $m[k+1] = m[k] + L[k] + 1$. The predictor coefficients, corresponding to the k -th frame, can be described by

$$a_i(n) = c_i^k + d_i^k \cdot u^k(n) \quad (2)$$

with $u^k(n) = (n - m[k]) / L[k]$ for $n = m[k] \dots m[k] + L[k]$. The superscript k of the coefficients indicates the corresponding frame. The coefficients c_i^k describe the constant components whereas the d_i^k describe the time-varying components of the predictor coefficients $a_i(n)$ for the k -th frame. $u^k(n)$ represents a straight line from 0 to 1 with the weight d_i^k implying the linear time evolution; $u^k(n)$ can be described by the vector \mathbf{u}^k . The prediction error $e(n) = x(n) - \hat{x}(n)$ of the k -th frame results in

$$e(n) = x(n) - \sum_{i=1}^N a_i(n) \cdot x(n-i) \quad (3)$$

$$e(n) = x(n) - \sum_{i=1}^N (c_i^k \cdot x(n-i) + d_i^k \cdot u^k(n) \cdot x(n-i))$$

for $n = m[k] \dots m[k] + L[k]$. Equation (3) can be expressed for each frame k by a vector-based description

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

using the definitions

$$\mathbf{x}_i^k = (x(m[k] - i), x(m[k] + 1 - i), \dots, x(m[k] + L[k] - i))^T,$$

$$\mathbf{e}^k = (e(m[k]), e(m[k] + 1), \dots, e(m[k] + L[k]))^T,$$

$$\mathbf{w}_i^k = \mathbf{u}^k \otimes \mathbf{x}_i^k, \text{ and } \mathbf{u}^k = \left(0, \frac{1}{L[k]}, \frac{2}{L[k]}, \dots, \frac{L[k]-1}{L[k]}, 1\right)^T;$$

the operation \otimes describes the element-by-element multiplication with $\mathbf{w} = \mathbf{u} \otimes \mathbf{x} \rightarrow w(n) = u(n) \cdot x(n)$. The vector \mathbf{x}_0^k in eq. (4) represents the k -th frame whereas \mathbf{x}_i^k for

$i > 0$ represents the shifted k -th frame with a shifting of i samples. Eq. (4) can be solved for the vector \mathbf{x}_0^k resulting in

$$\mathbf{x}_0^k = \sum_{i=1}^N (c_i^k \cdot \mathbf{x}_i^k + d_i^k \cdot \mathbf{w}_i^k) + \mathbf{e}^k \quad \text{with } k=1 \dots P. \quad (5)$$

P is the number of frames. Eq. (5) represents a vector expansion of \mathbf{x}_0^k by the vectors \mathbf{x}_i^k and \mathbf{w}_i^k ; the error of approximation is \mathbf{e}^k . The weights of the vectors \mathbf{x}_i^k and \mathbf{w}_i^k are the coefficients c_i^k and d_i^k . These coefficients can be determined by a minimization of the norm of the error vector $|\mathbf{e}^k|$ for each frame k separately. However, in this way the segments would be analyzed independently, which can cause 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

$$a_i(m[k+1]) = a_i(m[k] + L[k]) \quad (6)$$

ensuring that the last coefficient $a_i(m[k] + L[k])$ of each frame k is connected continuously to the first value $a_i(m[k+1])$ of the next frame. Considering eq. (2), eq. (6) can be expressed by the coefficients c and d with

$$c_i^{k+1} = c_i^k + d_i^k. \quad (7)$$

c_i^{k+1} and $c_i^k + d_i^k$ are the first and last value of the frame $k+1$ and k , respectively. Eq. (7) implies a coupling of the eqs. of (5) for $k=1 \dots P$. The equations (5) and (7) can be combined in one vector expansion or one single system of equations covering the frames $k=1 \dots P$. For that purpose, the vectors of eq. (5) are arranged on top of each other

$$\begin{aligned} \mathbf{x}_0^1 &= \sum_{i=1}^N (c_i^1 \cdot \mathbf{x}_i^1 + d_i^1 \cdot \mathbf{w}_i^1) + \mathbf{e}^1 \\ &\vdots \\ \mathbf{x}_0^P &= \sum_{i=1}^N (c_i^P \cdot \mathbf{x}_i^P + d_i^P \cdot \mathbf{w}_i^P) + \mathbf{e}^P. \end{aligned}$$

Since the coefficients c_i^k for $k > 1$ depend on the other coefficients, these coefficients are eliminated applying eq. (7) iteratively. After sorting the vectors, the matching vectors of each column are combined to single extended vectors leading to one vector equation

$$\mathbf{q}_0^0 = \sum_{i=1}^N c_i^1 \cdot \mathbf{q}_i^0 + \sum_{k=1}^P \sum_{i=1}^N d_i^k \cdot \mathbf{q}_i^k + \mathbf{e}^P \quad (8)$$

using the vectors \mathbf{q}_i^k for $k=0 \dots P$ defined by

$$\mathbf{q}_i^0 = \begin{pmatrix} \mathbf{x}_i^1 \\ \mathbf{x}_i^2 \\ \vdots \\ \mathbf{x}_i^P \end{pmatrix}, \mathbf{q}_i^1 = \begin{pmatrix} \mathbf{w}_i^1 \\ \mathbf{x}_i^3 \\ \vdots \\ \mathbf{x}_i^P \end{pmatrix}, \dots, \mathbf{q}_i^k = \begin{pmatrix} \mathbf{0}^1 \\ \vdots \\ \mathbf{0}^{k-1} \\ \mathbf{w}_i^k \\ \mathbf{x}_i^{k+1} \\ \vdots \\ \mathbf{x}_i^P \end{pmatrix}, \dots, \mathbf{q}_i^P = \begin{pmatrix} \mathbf{0}^1 \\ \mathbf{0}^2 \\ \vdots \\ \vdots \\ \mathbf{0}^{P-1} \\ \mathbf{w}_i^P \end{pmatrix}.$$

The vector $\mathbf{0}^k = (0, 0, \dots, 0)^T$ contains zero values and its length is equal to the length of \mathbf{x}_i^k . Since eq. (8) is linear with the coefficients c_i^1 and d_i^k , the optimum coefficients can be determined by least mean square estimation minimizing $|\mathbf{e}^P|$. This can be performed by the vector expansion of \mathbf{q}_0^0 with the vectors \mathbf{q}_i^k represented by eq. (8). The vectors \mathbf{q}_i^k can be interpreted as a basis which is usually not orthogonal. Therefore, the vectors \mathbf{q}_i^k are transformed into an orthogonal

basis with the vectors \mathbf{v}_i using the Gram-Schmidt orthogonalization. Then, the optimum coefficients $\lambda_i = \langle \mathbf{q}_0^0, \mathbf{v}_i \rangle / |\mathbf{v}_i|^2$ in the description of the orthogonal basis can be obtained by the dot product $\langle \cdot, \cdot \rangle$. Finally, these coefficients are converted back $\lambda_i \rightarrow c_i, d_i$ into the original basis of \mathbf{q}_i^k yielding the optimum coefficients.

2.1.1. Efficient analysis of many frames

The computational cost of the estimation procedure increases linearly with the length of the analyzed signal. Unfortunately, the computational cost increases also with the number of frames P , due to the increasing number of vectors in eq. (8). Fortunately, this problem can be solved by analyses of overlapping ranges of frames; a range is defined by several adjoining frames. In the following description, the overlapping of the ranges is W frames and each range contains B frames. The first range contains the segments with index $k=1 \dots B$, which is analyzed as described in the last section by equation (8) with $P=B$. The next range with index $\tilde{k}=k-B+W$ contains the segments $k=B-W+1 \dots B-W$. Now, the index and coefficients of the next frame are denoted with a tilde ' \sim '. This range is analyzed utilizing the results of the previous range. The coefficients c_i^{B-W+1} of the analysis of the previous range can be used as the starting coefficients $\tilde{c}_i^1 = c_i^{B-W+1}$ for the analysis of the next range leading to

$$\tilde{\mathbf{q}}_0^0 = \tilde{\mathbf{q}}_0^0 - \sum_{i=1}^N \tilde{c}_i^1 \cdot \tilde{\mathbf{q}}_i^0 = \sum_{k=1}^B \sum_{i=1}^N \tilde{d}_i^k \cdot \tilde{\mathbf{q}}_i^k + \tilde{\mathbf{e}}^P. \quad (9)$$

Eq. (9) is the vector expansion for the next range and can be solved analogously to eq. (8) yielding the \tilde{d}_i^k . The following ranges are treated in the same way. By this procedure the computational cost of the time-varying prediction can be significantly reduced for analyses with many frames. The differences of the results between the analyses with and without overlapping ranges can be neglected; for example, if each range has 7 segments with an overlapping chosen by 3, the analysis results are almost the same.

3. Time-varying speech processing

In the following, the time-varying linear prediction procedure is applied to speech signals; the sampling rate of the speech signals is 16 kHz. For the analysis, the speech signal is divided into segments determining the markings $m[k]$. The optimum coefficients are determined by the vector expansion, as described in the previous section. The error vector represents the residual signal of the time-varying prediction. The requirements for the time-varying prediction are the choice of the prediction order N and the markings $m[k]$.

3.1. Time-varying pre-emphasis

The pre-emphasis compensates the spectral decrease of speech, which eliminates the influence of the voiced excitation and the lip radiation. This effect can be explained by a simplified model of the speech production process implying a series connection of excitation, vocal tract, and radiation; the lip radiation is modeled by a real zero and the voiced excitation by two real poles. Since both the zero and the poles are assumed close to one, a single pole remains for the combined excitation-radiation model; however, this is only a simplification of the actual conditions. In addition, the spectral decrease of voiced speech is time-varying. This can

be caused, for example, by alterations of the glottal flow waveform. Furthermore, the lip radiation depends on the area of lip opening. An adaptive pre-emphasis, which is constant for the analyzed speech segment, can be estimated by the common linear prediction of order one [1]; the resulting prediction error system $e(n) = x(n) - a_1 \cdot x(n-1)$ represents a pre-emphasis filtering. For the realization of a time-varying pre-emphasis, the time-varying prediction of the previous section with order one fits perfectly for that task. The error signal for order $N=1$ corresponding to eq. (3) results in

$$e(n) = x(n) - (c_1^k + d_1^k \cdot u^k(n)) \cdot x(n-1) \quad (10)$$

for $n = m[k] \dots m[k] + L[k]$; this can be expressed as

$$e(n) = x(n) - p(n) \cdot x(n-1) \quad (11)$$

with the time-varying pre-emphasis coefficient $p(n) = c_1^k + d_1^k \cdot u^k(n)$. The time-varying pre-emphasis can be used repeatedly applying the procedure on the resulting error signal again. If the error signal of the first pre-emphasis is

$$e_1(n) = x(n) - p_1(n) \cdot x(n-1), \quad (12)$$

the error signal of the $(j-1)$ -th repetition is defined by

$$e_j(n) = e_{j-1}(n) - p_j(n) \cdot e_{j-1}(n-1). \quad (13)$$

The repeated pre-emphasis describes a pre-emphasis modelled by more than one real zero achieving a better compensation of the spectral decrease. Typically, the first two pre-emphasis coefficients cause the major effect; however, the first coefficient is dominant. Fig. 1 shows the estimated time-varying pre-emphasis coefficients for the utterance [jUlja] of the German word “Julia”. The segment length is $L=320$ according to 20 ms. Since the pre-emphasis coefficients are more sensitive for

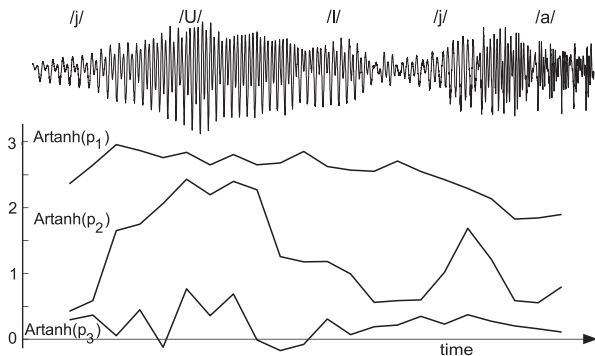


Figure 1: Analysis of word “Julia” [jUlja]: (top) speech, (bottom) first three time-varying pre-emphasis coefficients in description of Artanh.

absolute values close to one, the nonlinear function of the inverse hyperbolic tangent (Artanh) is applied to the pre-emphasis coefficients. For the example shown in fig. 1, the first pre-emphasis coefficient $p_1(n)$ is between 0.95 and 0.99, $p_2(n)$ is between 0.43 and 0.98, and $p_3(n)$ is between -0.1 and 0.49. It can be seen that the first two coefficients depend on the phoneme. The first coefficient is relatively low for the vowel /a/ due to the open mouth. In contrast to that, the vowel /U/ has a high first pre-emphasis coefficient and, in addition, a high second coefficient; the vowel /U/ causes a strong pre-emphasis due to lip rounding. The third pre-emphasis coefficient is fluctuating nearby zero with minor effect.

3.2. Time-varying inverse filtering

The time-varying prediction can be used for time-varying inverse filtering, with the additional benefit of avoiding block-wise processing. For that purpose, at first a time-

varying pre-emphasis is carried out as described in the previous section. The resulting pre-emphasized speech signal $e_j(n)$ with $j=3$ is used for the time-varying prediction with $N>1$, for vocal tract modeling. This means substituting $x(n) = e_j(n)$ for eq. (8) or (9), respectively. The segmentation is chosen with a uniform frame length $L=320$ for each frame. To obtain one area configuration from each segment k , the time averaged predictor coefficients \bar{a}_i^k of each segment k is calculated by $\bar{a}_i^k = (c_i^k + c_i^{k+1})/2$. Then, the polynomials are converted into reflection coefficients and, finally, into logarithmic areas. The estimated logarithmic areas of the analysis with prediction order $N=21$ of the word “Julia” (same utterance as in fig. 1) is shown in fig. 2. The lips are at position 21 and the glottis is approximately in region 2-6. By comparison with areas from literature obtained from NMR or X-ray investigations, it can be seen that the areas in fig. 2(b) obtained with a repeated pre-emphasis lead to more realistic vocal tract configurations than those in fig. 2(a) without repetition. Other analyses show similar effects. From fig. 2(b), the constriction in the mouth region for the sound /j/ and the open mouth of vowel /a/ can be recognized.

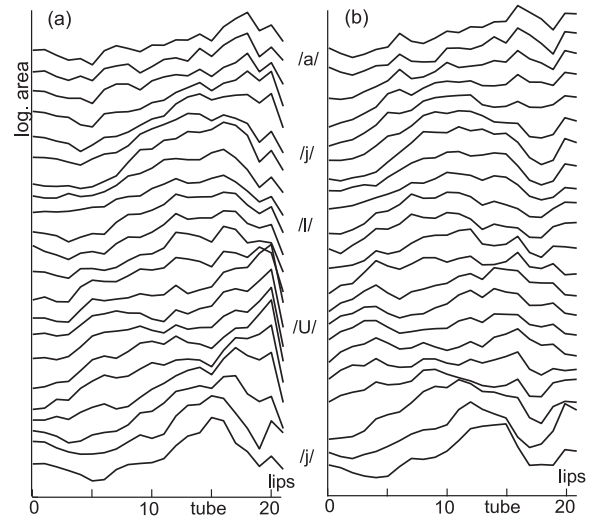


Figure 2: Logarithmic areas of [jUlja] estimated by time-varying prediction and pre-emphasis (a) without repetition $j=1$ and (b) with repetition $j=3$.

3.2.1. Comparison of inverse filtering methods

In the following, the effect of different inverse filtering methods on the residual is discussed. For comparison, the common block-wise time-invariant linear prediction is used; the residual is calculated with constant coefficients for each block. Additionally, to achieve a quasi-continuous processing comparable to the time-varying approach, the coefficients of the time-invariant analysis are interpolated, by single-sample processing during the inverse filtering; at the frame boundaries the coefficients contain half-sharing the two coefficient configurations of the two frames. The interpolation of the predictor coefficients is performed in the description of reflection coefficients; this leads to slightly better results than interpolating the predictor coefficients themselves. The analyses show that the interpolated inverse-filtering with time-invariant analysis increases the residual power considerably. Therefore, interpolation is not adequate to achieve continuous processing. Furthermore, the analyses show that the time-varying prediction leads to usually five percent less residual power than the block-wise time-invariant approach. Fig. 3 gives an example of

residual signals representing the beginning of the diphthong /aI/ from [vall]. The residuals of block-wise time-invariant analysis 3(c) and time-varying analysis 3(d) are similar. However, it can be seen that the interpolation, used for the residual of 3(b), leads to artefacts, especially in non-stationary regions. The differences

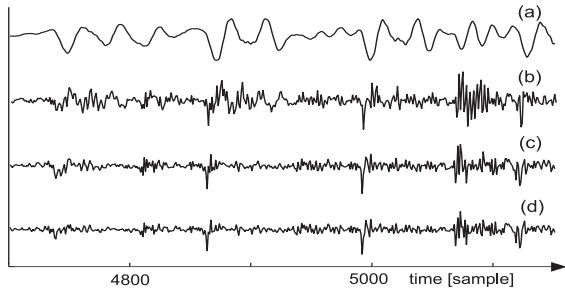


Figure 3: Analysis of /aI/: (a) speech; residual (b) by time-invariant analysis with interpolation and (c) without interpolation, (d) by time-varying analysis.

between the residuals from the time-varying and the block-wise (non-interpolated) time-invariant approach can be revealed by a low-pass filtering of the residuals. For that purpose, a low-pass with two real poles is used. The low-pass filtered residual is roughly related to the glottal flow, and is useful for speech analysis or synthesis; for example, in [5, 6] pitch modification is performed in a low-pass filtered description of the residual. It should be mentioned that also inverse filtering approaches exist for the purpose of glottal flow estimation such as [13, 14]; in [13] only the closed phase of a glottal period is used for the estimation of the inverse filter. Fig. 4 shows an example of low-pass filtered residuals obtained by the block-wise time-invariant approach 4(c) and the time-varying approach 4(b). The analyses show that the time-varying approach leads to residuals which are more regular in the low-passed description than those obtained by the time-invariant approach.

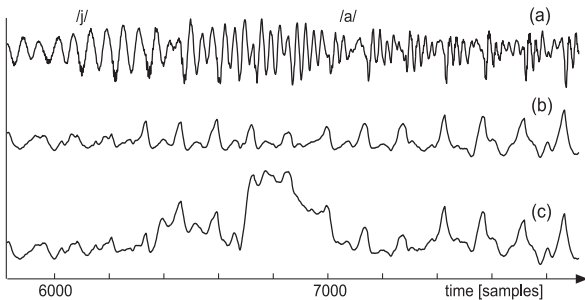


Figure 4: Analysis of transition [ja] from [jUlja]: (a) speech; low-pass filtered residual signal obtained from (b) time-varying analysis and (c) non-interpolated time-invariant analysis.

3.3. Synthesis based on time-varying prediction

The inverse system of the time-varying predictor error system of eq. (3) can be used for synthesis. For that purpose, the averaged predictor coefficients of each frame $\bar{a}_i^k = (c_i^k + c_i^{k+1})/2$ are used; in comparison to that, the coefficients c_i^k are less suitable. To ensure the stability of the synthesis system, the coefficients \bar{a}_i^k are converted into reflection coefficients \bar{r}_i^k . If one $|\bar{r}_i^k|$ is equal or greater than one, the absolute value can be decreased to a value smaller than one, typically 0.99; however, this case occurs scarcely. Re-synthesis examples based on the averaged time-varying coefficients indicate that the procedure is suitable for synthesis. The excitation of the synthesis system can be

residual-based or model-based. One favourable feature of the time-varying prediction is that the inclusion of the continuity condition by eq. (7) into the estimation procedure causes smooth trajectories of the coefficient vectors, which can help reducing the buzz in model-based synthesis.

4. Conclusions

A time-varying prediction is proposed for speech processing, which can be performed in a quasi-analytical way. The time-varying prediction can be used for a time-varying pre-emphasis and for time-varying prediction analysis or inverse filtering. Additionally, synthesis can be performed by the estimated coefficients. The investigations indicate that the differences of the results obtained by the time-varying approach compared to the time-invariant approach occur mostly in non-stationary speech regions such as sound transitions. Besides the modeling of the time-dependence itself, the time-varying approach enables a joint estimation of adjacent speech segments avoiding block-wise processing.

5. References

- [1] Markel, J. and Gray, A., Linear Prediction of Speech. New York: Springer-Verlag, 1976.
- [2] Makhoul, J., "Linear prediction: A tutorial review", in Proc. IEEE, vol. 63, pp. 561–580, Apr. 1975.
- [3] Childers, D. G. and Lee, C. K., "Vocal quality factors: Analysis, synthesis, and perception", J. Acoust. Soc. Am., Vol. 90, no. 5, pp. 2394–2410, 1991.
- [4] Ananthapadmanabha, T. V. and Yegnanarayana, B., "Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval", IEEE Trans. ASSP-27, no. 4, pp. 309–319, August 1979.
- [5] P.H. Milenkovic, "Voice source model for continuous control of pitch period", J. Acoust. Soc. Am. 93 (2), pp. 1087–1096, 1993.
- [6] Schnell, K., "Pitch Modification of Speech Residual Based on Parameterized Glottal Flow with Consideration of Approximation Error", Proc. ICASSP, Toulouse 2006, pp. 737–740.
- [7] Wakita, H., "Estimation of Vocal-Tract Shapes from Acoustical Analysis of the Speech Wave: The State of the Art", IEEE Trans. ASSP-27, no. 3, pp. 281–285, June 1979.
- [8] Haykin, S., Adaptive Filter Theory. New Jersey: Prentice-Hall, Inc., 3 ed., 1996.
- [9] 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.
- [10] Härmä, A., Juntunen, M., and Kaipio, J. P., "Time-Varying Autoregressive Modeling of Audio and Speech Signals", Proc. EUSIPCO, Tampere Finland 2000.
- [11] Jachan, M., Matz, G., and Hlawatsch, F., "Time-Frequency-Autoregressive Random Processes: Modeling and Fast Parameter Estimation", Proc. ICASSP, Hong Kong 2003, pp. 125–128.
- [12] Grenier, Y., "Time-Dependent ARMA Modeling of Non-stationary Signals", IEEE Trans. ASSP-31, no. 4, pp. 899–911, August 1983.
- [13] Wong, D. Y., Markel, J. D., and Gray, A.H., "Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform", IEEE Trans. ASSP-27, no. 4, pp. 350–355, August 1979.
- [14] Alku P., "Glottal Wave Analysis with Pitch Synchronous Iteratively Adaptive Inverse Filtering", Speech Communication, Vol. 11, nos. 2-3, pp. 109–118, 1992.