Improved Non-Linear Long-Term Predictors based on Volterra Filters

Vladimir Despotović [1], Norbert Görtz [2], Zoran Perić [3]
[1] University of Belgrade, Technical Faculty in Bor, Vojske Jugoslavije 12, 19210 Bor, Serbia
[2] Vienna University of Technology, Institute of Telecommunications, Gußhausstr. 25-29, 1040 Vienna, Austria
[3] University of Nis, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Nis, Serbia
vdespotovic@tf.bor.ac.rs

Abstract - Speech prediction is extensively based on linear models. However, components generated by nonlinear effects are also contained in speech signals, which is neglected using linear techniques. This paper presents long-term nonlinear predictors based on second-order Volterra filters that are shown to be superior to linear long-term predictors with only a minimal increase in complexity and the number of coefficients.

Keywords – Volterra filters; Speech prediction; Pitch; Nonlinear signal processing

I. INTRODUCTION

Models based on linear prediction have been used for several decades in different areas of speech signal processing, such as coding, synthesis, speech and speaker recognition. While the linear approach has led to great advances in the last 40 years, it neglects structure known to be present in the speech signal [1].

Linear models are based on the assumption that the vocal cords and the vocal tract are completely independent in the speech production process, which enables separation of the source and filter in the model (i.e. source-filter model). Furthermore, it is assumed that the airflow through the vocal tract is laminar (without turbulences) and that the vocal cords vibrations are ideally periodical for voiced speech. However, numerous results confirm that these assumptions are not always justified. Teager [2] presents several physical measures that show turbulences in the airflow. Richard [3] shows that even for sustained vowels exactly periodic vibration of vocal cords is not a reasonable assumption; hence the excitation source can be decomposed into a quasi-periodic and an aperiodic component. All these results lead to the conclusion that linear models neglect nonlinear effects during the speech production process. Hence, nonlinear models should enable a more accurate description of speech and lead to better performance of practical speech processing applications [1].

A large variety of techniques dealing with nonlinear speech prediction are reported in the literature. An excellent overview of these techniques is given in [4]. Given such a huge variety of options, it is clear that nonlinear prediction is not a trivial issue [5]. The use of the Volterra series model for short-term nonlinear speech prediction is reported in works of Thyssen [6], Mumolo [7] and Alipoor [8]. This paper deals with long-term prediction (LTP) of speech based on truncated Volterra series. Having in mind the exponential increase in the number of coefficients and the resulting complexity of the algorithm for Volterra series [9], the model proposed here will be limited to the second order.

II. NONLINEAR PREDICTION BASED ON THE SECOND ORDER VOLTERRA FILTERS

A nonlinear predictor based on Volterra filters estimates a current signal value by a linear combination of past signal values, and additionally by linear combinations of products of past signal values [10]. Hence, the predictor is nonlinear in the signal values, but linear in the filter coefficients. As a consequence, adaptation algorithms valid in the linear case can be extended to Volterra filters. Without loss of generality the system will be treated with the first and second-order kernels only (Figure 1). Then, the predicted signal is given by

$$\hat{x}(n) = \sum_{k=1}^{p} h_1(k) \cdot x(n-k) + \sum_{i=1}^{p}\sum_{j=1}^{p} h_2(i,j) \cdot x(n-i) \cdot x(n-j)\,, \qquad (1)$$

where $\hat{x}(n)$ is the estimation of $x(n)$ and p is the prediction order. The coefficients $h_1(k)$ and $h_2(i,j)$ represent linear and nonlinear components, respectively. The symmetry of the coefficients is assumed, so $h_2(i,j) = h_2(j,i)$. In this case, the overall number of coefficients for the second order Voltera predictor equals

$$n_c = p + \frac{p \cdot (p+1)}{2}\,. \qquad (2)$$

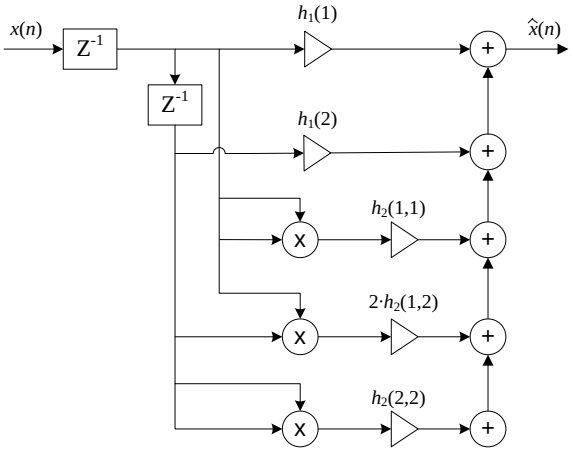

Figure 1. Model of the second order Volterra filter for prediction order p=2

The estimation of the coefficients can be achieved using adaptive algorithms such as Recursive Least Squares (RLS) or Least Mean Squares (LMS) [7].

III. Nonlinear Long-Term Prediction of Speech

The predictor presented in the previous section is essentially a short-term predictor, which eliminates correlation between nearby samples. It is well known that the prediction order must be high enough to include at least one pitch period, in order to model a voiced signal adequately [11]. However, this is not acceptable for most practical implementations due to large delay and increased complexity. The standard solution to this problem in linear prediction is the use of a model with two predictors, short-term and long-term, connected in cascade. A long-term predictor in such a realization targets correlation between samples one or multiple pitch periods apart. This solution is used in a number of speech coders (CELP, RPE-LTP etc.).

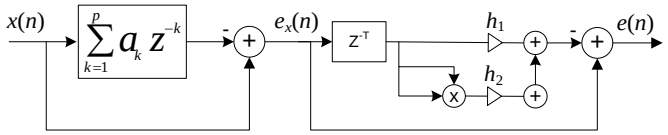

Figure 2. Short-term linear predictor connected in cascade with a long-term second-order Volterra predictor

In this paper we substitute the linear long-term predictor by a nonlinear one, based on the second order Volterra filter and connect it in cascade with the short-term linear predictor, as shown in Figure 2. It predicts the current signal sample from a past sample that is one or more pitch periods apart. The method can be defined as the one-tap predictor; that is, prediction is based on one single sample from the distant past. The corresponding predicted sample equals

$$\hat{e}_x(n) = h_1 \cdot e_x(n-T) + h_2 \cdot e_x^{\,2}(n-T)\,, \qquad (3)$$

where T is the pitch period and h_1 and h_2 are LTP coefficients. Compared to linear LTP, the number of coefficients is increased only by one (h_2). Within a given time interval of interest, we seek to find h_1 and h_2 such that the sum $J = \sum_n \left(e_x(n) - \hat{e}_x(n)\right)^2$ of squared errors is minimized. Substituting (3), differentiating with respect to h_1 and h_2 and equating to zero, the following LTP coefficients are obtained [12]

$$h_1 = \frac{\sum_n e_x^{\,4}(n-T) \cdot \sum_n e_x(n) \cdot e_x(n-T) - \sum_n e_x^{\,3}(n-T) \cdot \sum_n e_x(n) \cdot e_x^{\,2}(n-T)}{\sum_n e_x^{\,4}(n-T) \cdot \sum_n e_x^{\,2}(n-T) - \left(\sum_n e_x^{\,3}(n-T)\right)^2}\,, \qquad (4)$$

$$h_2 = \frac{\sum_n e_x^{\,2}(n-T) \cdot \sum_n e_x(n) \cdot e_x^{\,2}(n-T) - \sum_n e_x^{\,3}(n-T) \cdot \sum_n e_x(n) \cdot e_x(n-T)}{\sum_n e_x^{\,4}(n-T) \cdot \sum_n e_x^{\,2}(n-T) - \left(\sum_n e_x^{\,3}(n-T)\right)^2}\,. \qquad (5)$$

It is necessary to determine a pitch period T in order to locate and reduce periodic components that are exactly one pitch period apart. A search procedure within the range $\mathrm{Tmin} \le T \le \mathrm{Tmax}$ has to be applied to find the optimal T, where typically $T_{\min} = 20$ and $T_{\max} = 140$.

A. The Frame/Subframe Structure

It can be shown that the effectiveness of the presented long-term predictor on removing long-term correlation is limited. The parameters of the long-term predictor need to be updated more frequently than the parameters of the short-term predictor. That is, it loses its effectiveness when the time interval used for estimation becomes too long, which is due to the dynamic nature of the pitch period [11]. Experiments have shown that shortening the time interval for estimation of linear LTP coefficients from 160 to 40 samples results in an increase of the LTP gain of up to 2.2 dB [13]. We are going to show that similar conclusions are also true for nonlinear LTP based on the second order Volterra filter.

The frame/subframe structure is used, where short-term linear prediction is applied to frames with the length of 240 samples. The frame is divided into four intervals of equal length, known as subframes. Second order Volterra LTP analysis is then applied to each subframe separately.

Note that a more frequent update of the long-term predictor obviously requires a higher bitrate. However, this could be justified for high enough prediction gain.

IV. Results and Discussion

Let us assume a short-term linear predictor with the order p=10 connected in cascade with the one-tap long-term nonlinear predictor based on the second order Volterra filter, as shown in Figure 2. The predictor operates on frames of 240 samples, divided into four subframes of 60 samples.

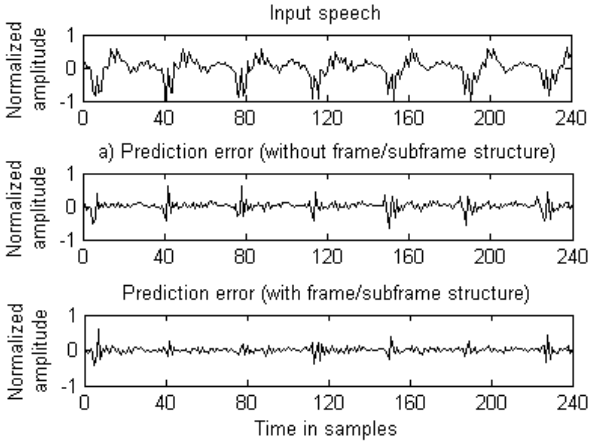

Figure 3. Prediction error using the short-term linear predictor connected in cascade with the long-term second-order Volterra predictor; a) without frame/subframe structure, b) with frame/subframe structure

Figure 3 shows the prediction error on one characteristic frame of speech, with and without use of the frame/subframe structure. Without dividing frames into subframes nonlinear

long-term prediction gain is modest and equals only 0.56 dB, which is indicated by the strong remaining periodic pitch component in the error signal. On the other hand, when the frame/subframe structure is introduced the nonlinear long-term gain raises up to 3.50 dB, with pitch components substantially suppressed.

Experiments using an extensive amount of speech samples performed on nearly 3 minutes of speech extracted from the TIMIT database [14] (56 sentences, American English speakers) confirmed this result. Using the frame/subframe structure an increase of nonlinear long-term prediction of 2 dB was achieved (Table I). The obtained gain is also approximately 1 dB higher compared to a case when linear LTP with the same frame/subframe structure was used.

TABLE I. PREDICTION GAIN FOR NONLINEAR PREDICTORS BASED ON THE 2ND ORDER VOLTERRA FILTERS

Long-term prediction gain [dB]	
Without frame/subframe structure	With frame/subframe structure
1.84	3.86

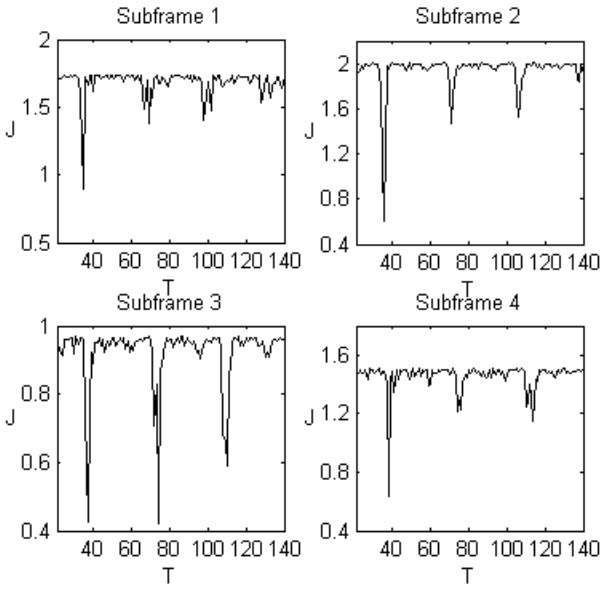

Figure 4. Sum of squared error *J* as a function of pitch period *T* on different subframes

An example of the sum of squared errors *J* as a function of pitch period *T* for the four subframes of one characteristic frame is given in Figure 4. The pitch period is searched in the range $20 \le T \le 140$. The minimum value of *J* on the graph indicates the optimal pitch period. Nonlinear long-term predictor coefficients and optimal pitch periods found on a given frame are listed in Table II. Note that both the pitch period and coefficients change substantially in different subframes. Without the frame/subframe structure the optimal pitch period was found to be $T = 35$. Those significant changes are due to integer multiples of the pitch-period that may provide smaller values for *J*, because the pitch period may not be an integer multiple of the sampling period used. Using a subframe structure allows to pick a "good" value for the pitch period more frequently, which results in higher prediction gain.

Finally, we have analyzed the distribution of nonlinear long-term prediction gain over segments of speech signal. As expected, the highest additional gains of up to 5 dB are achieved in voiced speech frames; however, even in unvoiced speech frames the additional prediction gain hardly drops below 1dB. A typical situation for one word containing voiced to unvoiced transitions (the word *Dishes*) is shown in Figure 5.

TABLE II. OPTIMAL PITCH PERIOD AND LONG-TERM NONLINEAR PREDICTOR COEFFICIENTS

Subframe	h_1	h_2	T
1	0.86	-0.003	35
2	0.81	-0.04	36
3	0.64	-0.27	74
4	0.43	-0.76	38

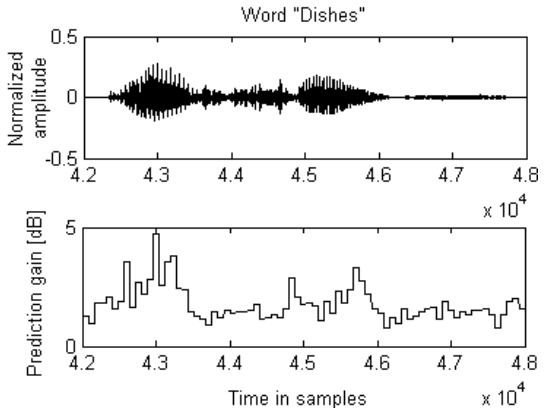

Figure 5. Nonlinear long-term prediction gain over different segments of speech in the word *Dishes*

V. CONCLUSION

A long-term second order Volterra predictor is proposed in this paper. The frame/subframe structure was used, where each frame was divided into four subframes of equal length, and second order Volterra LTP analysis was applied to each subframe separately. An increase in prediction gain of 2 dB was achieved this way. Compared to the linear LTP predictor with the frame/subframe structure, the performance was improved by 1 dB, whereas number of predictor coefficients was only increased by one. This also proves that the pitch complexes contain nonlinear components.

Analyzing the nonlinear long-term prediction gain over segments of speech we found that the highest gain was observed in voiced frames, as expected. However, even in unvoiced frames the prediction gain was not negligible.

ACKNOWLEDGMENT

Vladimir Despotovic wishes to thank the Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH) for research grant within the framework of WUS Austria scholarship program that enabled his stay at the Vienna University of Technology, Institute of Telecommunications.

REFERENCES

[1] M. Faúndez-Zanuy, G. Kubin, W. B. Kleijn, P. Maragos, S. McLaughlin, A. Esposito, A. Hussain and J. Schoentgen, "Nonlinear speech processing: overview and applications," Control and Intelligent Systems, vol. 30, no.1, 2002, pp. 1-10.

[2] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in Speech Production and Speech Modelling, W. J. Hardcastle and A. Marchal, Eds., vol. 55 of NATO Advanced Study Institute Series D, pp. 241–261, Bonas, France, July 1989.

[3] G. Richard and C. R. D'Alessandro, "Modification of the aperiodic component of speech signals for synthesis," in Progress in Speech Synthesis, R. Sproat, J. Olive, J. Hirschberg J. P.H. van Santen, Ed. New York: Springer-Verlag, 1997, pp. 41-56.

[4] M. Chetouani, A. Hussain, M. Faúndez-Zanuy and B. Gas, "Non-linear predictive models for speech processing," ICANN 2005, Lecture Notes in Computer Science, W. Duch et al. (Eds.), vol. 3697, pp. 779–784, 2005.

[5] G. Kubin, "Nonlinear processing of speech," in Speech coding and synthesis, Chapter 16, W. B. Kleijn & K. K. Paliwal, Ed.: Elsevier, 1995.

[6] J. Thyssen, H. Nielsen and S. Hansen, "Non-linear short-term prediction in speech coding," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, 1994, pp. 185-188.

[7] E. Mumolo, A. Carini and D. Francescato, "ADPCM with nonlinear predictors," in Signal Processing VII: Theories and applications.: Elsevier, 1994, pp. 387-390.

[8] Gh. Alipoor and M. H. Savoji, "Employing Volterra filters in the ADPCM technique for speech coding: a comprehensive investigation," European Transactions on Telecommunications, vol. 22, no. 2, 2011, pp. 81-92.

[9] T. Ogunfunmi, Adaptive nonlinear system identification: the Volterra and Wiener Model Approaches, Springer, 2007.

[10] K. Schnell and A. Lacroix, "Estimation of Speech Features of Glottal Excitation by Nonlinear Prediction," in Proceedings of the ISCA ITRW Non-Linear Speech Processing (NOLISP 2007), Paris, France, 2007, pp. 116-119.

[11] W. C. Chu, Speech coding algorithms: foundation and evolution of standardized coders, New Jersey: John Wiley & Sons, 2003.

[12] V. Despotovic, N. Goertz and Z. Peric, "Nonlinear long-term prediction of speech based on truncated Volterra series," IEEE Transactions on Audio, Speech & Language Processing, vol. 20, no. 3, 2012, pp. 1069-1073.

[13] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 4, pp. 467–478, 1989.

[14] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus Linguistic Data Consortium, " Philadelphia, USA, 1993.