Voiced Excitation Models for Speech Production Based on Time Variable Volterra Systems

Karl Schnell, Arild Lacroix

Institute of Applied Physics, Goethe-University Frankfurt,
Frankfurt am Main, Germany
{Schnell, Lacroix}@iap.uni-frankfurt.de

Abstract. The speech production can be modeled by linear and nonlinear systems. In this contribution a time variable nonlinear Volterra system is used to model the fluctuations of the voiced excitation while a linear system models the resonances of the speech production system. The estimation of the Volterra system is performed by a prediction algorithm. This is enabled by a description of the prediction problem as an approximation by a series expansion. Speech examples show that the use of a time variable Volterra system improves the naturalness of the synthetic speech.

1 Introduction

Linear systems provide adequate modeling of the resonances of the vocal tract. The parameters of the linear system can be estimated by linear prediction or inverse filtering. The speech signal is described by the linear model only partially [1], therefore an estimation by a nonlinear system is performed with respect to the residual signal of speech. Nonlinear prediction based on Volterra systems is used for the estimation. Since in the residual signal the linear relations of the speech signal are eliminated mostly, the nonlinear predictor consists of nonlinear terms only.

2 Nonlinear Prediction

The prediction $\hat{x}(n)$ of a signal value $x(n)$ is performed by a combination of products of previous signal values $x(n-i)\cdot x(n-k)$ with $i,k>0$. For a signal x the prediction error e is defined as the difference between the estimated value $\hat{x}$ and the actual value x:

$$e(n) = x(n)-\hat{x}(n) = x(n)-\sum_{i=1}^{M}\sum_{k=1}^{i} h'_2(i,k)\cdot x(n-k)x(n-i)\,. \tag{1}$$

In (1) the second-order kernel h_2 is assumed symmetrically resulting $h_2'(i,k) = h_2(i,k)$ for i=k and $h_2'(i,k) = 2 \cdot h_2(i,k)$ for $i \neq k$. The coefficients of the predictor are optimal if the expected value $\mathrm{E}[e(n)^2]$ is minimized; this is approximated by $\sum e(n)^2 \rightarrow \min$.

2.1 Prediction based on Vectors

If the analyzed signal $x(n)$ is a finite signal of length L the prediction (1) can be described in a vector notation

$$\boldsymbol{x} = \hat{\boldsymbol{x}} + \boldsymbol{e} = \sum_{i=1}^{M} \sum_{k=1}^{i} h_2'(i,k) \cdot \boldsymbol{x}_{i,k} + \boldsymbol{e} \tag{2}$$

with the vectors:

$$\boldsymbol{x} = \big(x(0), x(1), \ldots, x(L-1)\big)^{\mathrm{T}}$$
$$\boldsymbol{x}_{i,k} = \big(x(-i)x(-k),\ x(-i+1)x(-k+1),\ \ldots\big)^{\mathrm{T}} .$$

Equation (2) represents a vector expansion of the vector $\boldsymbol{x}$ by the vectors $\boldsymbol{x}_{i,k}$. The error of the approximation $|\boldsymbol{e}| = |\boldsymbol{x} - \hat{\boldsymbol{x}}|$ is to be minimized. The estimation can be performed by a transformation of the vectors $\boldsymbol{x}_{i,k}$ into an orthogonal basis by the Gram-Schmidt algorithm. Then the coefficients can be easily determined with the aid of the dot product of the vector $\boldsymbol{x}$ and the orthogonal basis vectors. After this the basis is returned into the original vectors $\boldsymbol{x}_{i,k}$ yielding the estimated coefficients $h_2'(i,k)$.

3 Analysis and Synthesis of Speech

The prediction error filter (1) is suitable for the analysis of signals while the inverse system of the prediction error filter can be used for synthesis. The inverse system H_2 of (1) has a recursive structure:

$$y(n) = x(n) + \sum_{i=1}^{M} \sum_{k=1}^{i} h_2'(i,k) \cdot y(n-k) y(n-i) . \tag{3}$$

The recursive Volterra system is used for speech synthesis which can be seen in fig. 1. The parameters of the nonlinear system are time variable to model the fluctuations of the voiced excitation. The linear system $H(z)$ in fig. 1 models the resonance structure of the vocal tract and may include a real pole system $P(z)$ for deemphasis. The voiced excitation for speech synthesis is realized with the recursive Volterra

system H_2 which is excited by an impulse train corresponding to the fundamental frequency. The parameters for the nonlinear system are estimated from the residual of speech; to enable a time variability of the parameters the analysis is performed blockwise in short time intervals. For that purpose at first the speech signal is inverse-filtered by LPC-analysis representing the conventional linear prediction. The resulting residual signal is segmented into overlapping segments which are analyzed by the nonlinear Volterra prediction. The lengths of the segments are about two pitch periods and the overlapping is about one period. The nonlinear prediction described in section 2 yields for every segment a diagonal matrix $h'_{2\lambda}(i,k)$ of estimated coefficients; the index λ represents the analyzed λ-th segment. During the synthesis the parameters of the system H_2 are controlled by the parameter matrices $h'_{2\lambda}(i,k)$ consecutively, modeling the fluctuations of the voiced excitation.

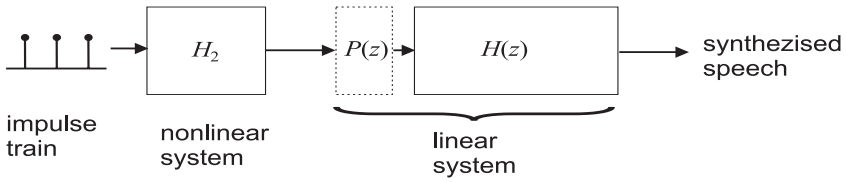

Fig. 1. Systems involved for the synthesis of voiced speech: Excitation of the time variable recursive Volterra system H_2 by an impulse train; filtering by a linear system consisting of $P(z)$ for deemphasis and $H(z)$ for the resonances of the vocal tract.

To demonstrate the effect of the time variable nonlinear system in the following example the vowel /a:/ is analyzed and resynthesized with constant fundamental frequency without any jitter. The speech signal of the vowel /a:/ is filtered by linear prediction and the residual is analyzed blockwise as described above. The linear system $H(z)$ is in this case the standard all-pole model obtained from the linear prediction. To show the impact of the nonlinear system, the spectra of the resynthesized vowel /a:/ are shown in fig. 2 with and without the recursive Volterra system. The use of the time variable Volterra system causes nonperiodicities, which can be seen in fig. 2(a) in contrast to 2(b). It is known, that a nonperiodic component is favorable for the naturalness of synthetic speech [2]. Additionally to the nonlinear system a noise component may be included into the excitation, to further increase the nonperiodicity.
Besides of resynthesis of stationary vowels, the excitation is used for a parametric synthesis which includes a lossy tube model as linear system $H(z)$ in fig. 1. The parameters of this lossy tube model are estimated from diphones by an optimization algorithm [3]. Figure 3 shows short-time spectra of the parametric synthesis representing the beginning of the diphthong /aI/. The time variable nonlinear system reduces the periodicity especially in the high frequency range which can be seen in fig. 3(a). Synthesized examples of words show that the inclusion of the time variable nonlinear system improves the naturalness of the synthetic speech.

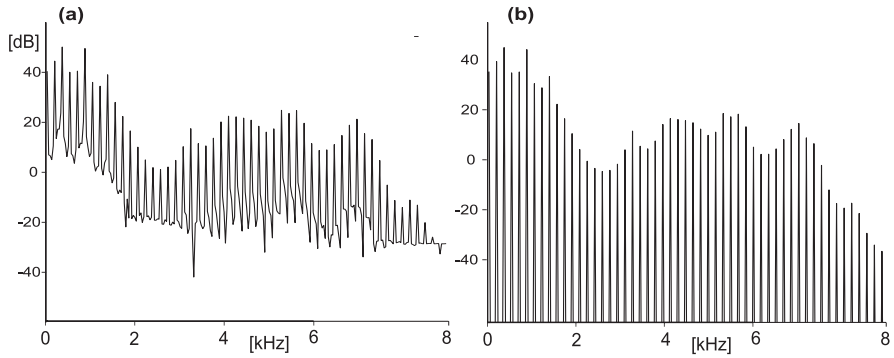

Fig. 2. Spectra of synthesized vowel /a:/: (a) excitation consists of an impulse train and the subsequent nonlinear system; (b) excitation consists of an impulse train without the nonlinear system.

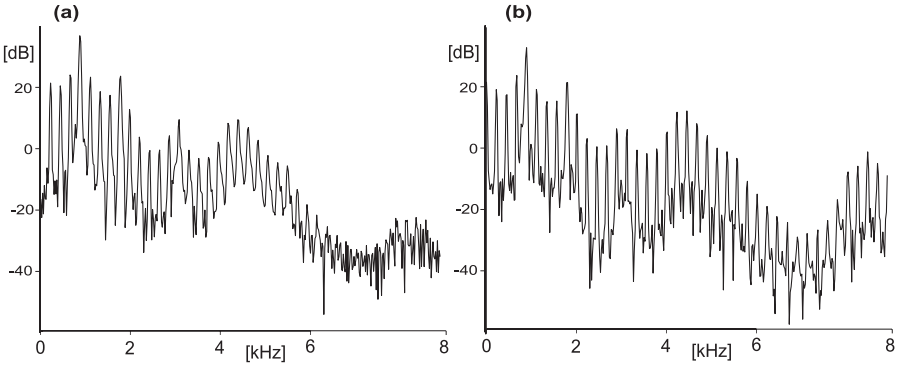

Fig. 3. Spectra of a segment of the synthesized diphthong /aI/ by parametric synthesis with lossy tube model: (a) excitation consists of an impulse train and the subsequent nonlinear system; (b) excitation consists of an impulse train without the nonlinear system.

References

1. McLaughlin, S.: Nonlinear Speech Synthesis, Proc. EUSIPCO-2002, Toulouse France, pp. 211–218 (2002).
2. Sambur M.R., Rosenberg A.E., Rabiner L.R., McGonegal C.A.: On reducing the buzz in LPC synthesis, J.Acoust.Soc.Am. (63), pp. 918-924, (1978).
3. Schnell, K., Lacroix, A.: Speech Production Based on Lossy Tube Models: Unit Concatenation and Sound Transitions, Proc. INTERSPEECH-2004 ICSLP, Jeju-Island Korea, Vol. I, pp. 505-508 (2004).