Effect of Adaptive Nonlinearity in Speech Coding

JARI TURUNEN, PEKKA LOULA, JUHA T. TANTTU
Tampere University of Technology, Pori
Pohjoisranta 11, P.O.BOX 300, FIN-28101, Pori
FINLAND

Abstract: - This paper is focused in the effect of adaptive weighted non-linearity in speech coding. The results showed that the effect of adaptive non-linearity improves the spectral distance measure when compared to the system without non-linearity.

Key-Words: - Coding, Non-linearity, Hammerstein Model, Adaptation

1 Introduction

Speech is a product of a non-stationary process. The speech production in human body is quite well known and it is studied over the decades. The different parts in speech production and propagation can be explained with partial differential equations. However, when thinking the general model combined from these partial differential equations, the efforts will lead to very difficult equation formulation and solving.

The most successful vocal tract estimator in speech processing is autoregressive (AR) filter, also known as Linear Predictive Coding (LPC). It has served successfully in speech coding over the years. Unfortunately, it has its limitations in speech modeling, for example a tendency to follow spectral peaks more than the spectral valleys and the nasal sound modeling accuracy. The proper nasal sound model filter needs at least one zero in the filter while the autoregressive model is all-pole filter and does not have any zeroes in its structure.

The LPC filter coefficients are easy to calculate and the redundancy removal filtering is easy to do, but the remaining residual after filtering process contains so much information that at least the most significant features from the residual must be modeled and transferred to the decoder somehow. This residual modeling is very intensive process and will take much more processing power than the LPC redundancy removal. These facts amongst others have partially guided the research work in nonlinear fields in order to reduce the residual handling process.

There have been analyses in several papers that the speech may contain different types of nonlinear components [1] – [5]. In the same time, nonlinear techniques have been tested with time series over several decades in order to improve modeling and estimation when compared to linear methods. For example, the logarithmic a-law/μ-law lossy compression in Pulse Code Modulation (PCM) coding [6] has worked successfully over the years and the information is possible to reduce 13-bit to 8-bit per sample without annoying disturbances.

In [7]-[9], the Radial Basis Functions (RBF) has been used in the parameter estimation process as a replacement or supporting part for the LPC estimator with good results. Multilayer perceptrons and time delay neural networks have been successfully used for speech analysis and synthesis in [10]. The Volterra kernel has been used as a short time predictor for speech in [2] as well as the chaotic processes in [1], [11][12].

The Hammerstein model has been used in many different types of time series analysis, for example in [13] for modeling biological systems, and echo cancellation in speech processing in [14]. In these papers the type of nonlinearity is known and thus results are good.

In speech processing the form of nonlinearity is not known precisely. It may be possible that speech consists of different types of nonlinearities. The vocal tract model structure should include different types of nonlinearities and emphasize the suitable ones for the current speech frame by weighting.

In this paper, we have proposed a structure for the weighted frame-by-frame adaptive non-linear structure and have studied the effect of weighted non-linearity to improve linear analysis method in vocal tract modeling considering speech coding.

2 Proposed system

The system described in this paper is based on two blocks, the non-linear and linear blocks. This type of system is also known as Hammerstein model. The proposed system is presented in Fig. 1.

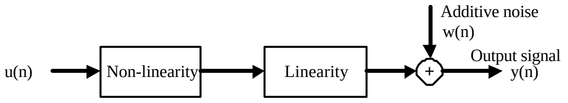

Fig. 1. The schematic diagram of the structure

In the proposed system the input signal, u(n), is first fed in the non-linear system where the non-linearities of the input signal is subtracted and rest of the signal are sent for linear processing. The additive noise, w(n), can be also a modeling error in this case, is cumulated in the output signal, y(n).

With the structure presented in Fig. 1, it is assumed that the static nonlinearity is able to remove the components in speech that the linear part cannot process. In the decoder, the inverse non-linear and linear processes must be able to accomplish. The non-linear and linear structures are static, but the weighting coefficients are adaptive and we used method described in [15] to calculate the adaptive coefficients for the speech frame in question.

The structure of the model is simple and the non-linear and linear coefficient calculation is an easy operation, but there are still several aspects to be concerned.

One problem is the shape of non-linearity in speech. In [1], it is suggested that the speech might be chaotic in nature, but in [2], only the consonants show chaotic behavior, and in [3] it is suggested that unvoiced fricatives may have quadratic non-linearities in speech signal.

In [14]-[15] the series of types of $\{x, x^2, x^3\ldots\}$ were tested as non-linearities with Hammerstein model. Several other series also exists as candidates to be a static non-linearity in Hammerstein model, but finding a suitable combination from several non-linearities, which will shape the speech signal for linear filter, is very difficult. The same non-linearity must work in the decoder and form the estimate of the original speech signal from inverse linear filtered residual signal, which itself is in the decoder side only a good estimate of the original residual. The stability must also be preserved at least between values –1 and 1, because in the decoder the output signal is limited between these values.

There are several good candidates to be used as non-linearity of the Hammerstein model, but the non-linearity should act in that sense that it will leave enough information in the signal that the inversion is possible in the decoder. If the weighted non-linearity will process the signal too close to the zero-level, the linear process will do more harm than good to the final filtered signal. This can also be an indication of the false identification of the parameter definition process. In order to get the combination work, the non-linearity and linear part should work together, and the identification process will model true dynamics of the speech signal.

2.1 Parameter identification

The non-linear and linear coefficients are calculated with the method described in [15]. This method is based on the well-known idea of minimizing the prediction error by using the least square estimate for the output signal and combined non-linearity and linearity (without coefficients) matrix. The output vector of the least square estimate method is organized in to matrix form, which is processed by "economy size" singular value decomposition. The nonlinear and linear weighting coefficients are possible to extract from the remaining three matrices with simple row operations.

The proposed system attached to parameter identification process is presented in Fig. 2.

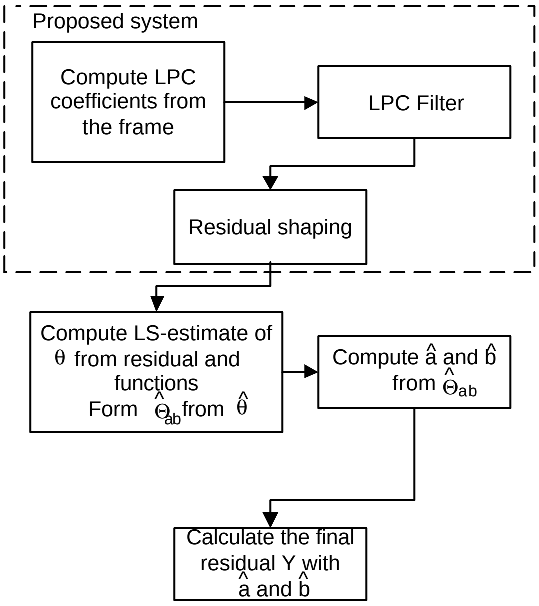
Proposed system
Compute LPC coefficients from the frame
LPC Filter
Residual shaping
Compute LS-estimate of θ from residual and functions Form $\hat{\Theta}_{ab}$ from $\hat{\theta}$
Compute $\hat{a}$ and $\hat{b}$ from $\hat{\Theta}_{ab}$
Calculate the final residual Y with $\hat{a}$ and $\hat{b}$

Fig. 2. The schematic structure of the parameter identification algorithm.

The Hammerstein system can be described as:

$$y(n)=\sum_{k=0}^{p}\sum_{i=0}^{r} b_k a_i g_i(u(n-k)) + w(n) \qquad (1)$$

where the b_k, k=0..p, are the linear coefficients, u(n-k) are the past k-delayed values in input, a_i, i=0..r, are the non-linear coefficients, g_i are the non-linear functions and w(n) is the additive noise. The y(n) can be formed as follows:

$$y(n) = q^T f + w(n) \qquad (2)$$

where the parameter vectors θ and ϕ are formed as follows:

$$q = [b_0 a_0, \ldots b_0 a_r, \ldots, b_p a_0, \ldots, b_p a_r]^T \qquad (3)$$

$$f = [g_0(u(n)), \ldots, g_r(u(n)), \ldots, \\ g_0(u(n-p)), \ldots, g_r(u(n-p))]^T \qquad (4)$$

The output signal vector Y is formed from N-samples of y:

$$Y_N = \Phi_N^T q + W_N \qquad (5)$$

where the Y_N, Φ_N and W_N vectors are formed as:

$$\begin{aligned} Y_N &= [y_1, y_2, \ldots, y_N]^T \\ \Phi_N &= [f_1, f_2, \ldots, f_N] \\ W_N &= [w_1, w_2, \ldots, w_N]^T \end{aligned} \qquad (6)$$

the θ vector has the following property:

$$\hat{q} = (\Phi_N \Phi_N^T)^{-1} \Phi_N Y_N \qquad (7)$$

In other words, the estimate of θ vector is least square estimate formed from the output signal Y_N and vector Φ_N, which is formed from non-linear and linear effects without any weighting operations.

The estimate of θ vector is then arranged to a block column matrix Θ_{ab}

$$\hat{\Theta}_{ab} = Blockvec(\hat{q}) \qquad (8)$$

The coefficients a_i and b_k can be solved from the estimate of Θ_{ab} matrix with “economy size” singular value decomposition (SVD) operation:

$$\hat{\Theta}_{ab} = U_j \Sigma_j V_j^T = \sum_{i=1}^{j} u_i s_i v_i^T \qquad (9)$$

where Σ_j is a diagonal matrix containing the j nonzero singular values, U_j and V_j contain only first j columns of unitary matrices U and V provided by full SVD of estimate of Θ_{ab}. The estimates of a and b coefficients are obtained as follows:

$$\hat{a} = U_1 \qquad (10)$$

$$\hat{b} = V_1 \Sigma_1 \qquad (11)$$

2.2 Simulations

In the simulations the LPC-filtered speech residual, from the current speech frame, was used as a base signal for construction of the output signal y(n) to parameter identification process, as shown in Fig. 2. This residual was then shaped by removing the periodical spikes that exceeds the statistical “3*√(var(y))” threshold rule. Also the trend was removed and finally, the residual signal was low pass filtered with four sample sliding window.

This modified signal is now used as an output signal y(n) and the current speech frame is used as input signal u(n) in the parameter identification process.

The selected static non-linearity component in the examples is derived from the A-law, as presented in Fig. 3, so that the final non-linearity will be:

$$g(x) = a_0 x + a_1 sign(x) * \frac{\log(9*|x| + 1.2)}{\log(10.2)} \qquad (12)$$

The selected nonlinearity was chosen due to its simplicity and ability to compress the speech signal. The parameter identification process itself allows more complex structures to be used as non-linearities.

Every item is multiplied with the respective coefficient a_i, i=0..1. If the speech input signal u(n) is normalized between {-1,1} and a_1=1, the effect of the logarithm component is presented in Fig. 3.

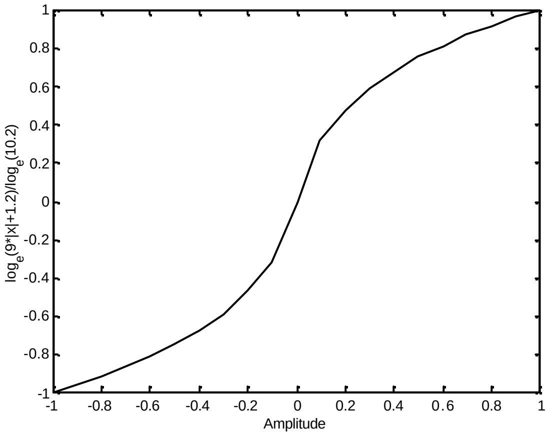

Fig. 3. The non-linear logarithmic function

The proposed model testing was done in Mathworks Matlab environment. The data was taken from TIDIGITS discrete number database. The database consist of 16-bit discrete speech samples, numbers, sampled with 8 kHz sampling frequency. The

Hammerstein system structure was selected to be the logarithm function as in equation 12, and the linear model combined with 10th order filter. As a reference 10th order filter without non-linearities were included in the test.

The proposed system (LOG-linear) and reference (linear) system were tested with following coding scheme: both residual signals were down-sampled that only every 8th residual sample was collected and transmitted in to the decoder. This method was selected for finding out the modeling differences and stability with coarse residual and both systems' ability to reconstruct the estimate of the original signal from this kind of excitation.

When processing samples, corresponding approximately 10 minutes of speech of both sexes (5 minutes each), the RMS signal error was approximately the same with all models. However, the spectral distance measures the similarity better than RMS error. In the Table 1, Itakura distances for decoded signals for all filtering methods are presented.

Table 1. The spectral distance measure

Method	Itakura distance
LOG-linear	0.20
Linear	0.32

In Figure 4, the LOG-linear and linear based vocal tracts are compared to the speech frame spectrum. The effect of the logarithmic transform is also presented in Fig. 4. It can be seen from Fig. 4 that the Log-linear model is tracking the vocal tract model better than the linear-model. The weighted logarithmic function has emphasized certain spectral features from the speech that the linear model matches better with the log-modified speech spectra.

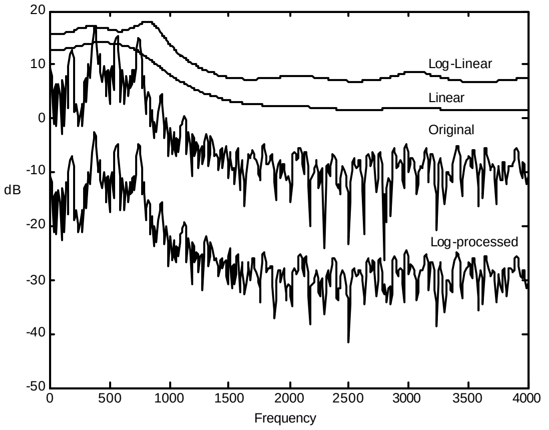

Fig. 4. Two vocal tract models compared to the original signal spectrum. The LOG-linear vocal tract is shifted +10 dB, linear +5 dB, original signal is in zero level and log processed signal is shifted down with –20 dB.

In Fig. 5, the results of decoding process are presented. It can be seen that the non-linearity will enhance the peaks of the signal when compared to the reference linear model. It can be seen that the original signal features can be recovered with less information when compared to linear method. Both of the signals are constructed from the residual excitation signal that consists of only every 8th sample transmitted to the decoder.

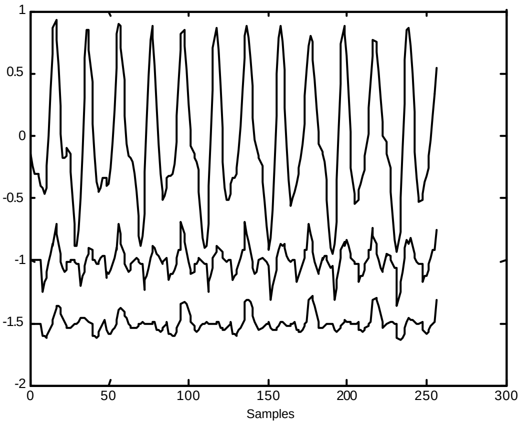

Fig. 5. The original, "Log-linear" decoded (shifted down by –1 units) and "linear" decoded (shifted down by –1.5 units) signals.

3 Conclusion

The non-linear methods can improve significantly the speech coding mechanism. The method in [15], further extended in [16] that gives the mathematical background and the algorithm extension also for other topologies, is suitable for static non-linearity and linearity weighting identification and thus frame based adaptive filter coefficient calculation. The method is easy to use but the model calculation is slightly more complex when compared to linear predictive coding. But with suitable non-linear and linear combination the residual modeling can be reduced significantly when compared to LPC residual modeling methods.

The experiment in this paper shows the suitability of the method for speech coding. The identification method estimates the combined system parameters from input and output signals. This approach is very close to black box model identification, but the inner structure must be known in this case. The output signal that will be fed in the coefficient estimation process is not known in our case, but this allows an opportunity to feed almost any signal into the system. If the output signal can be regarded as residual signal and the input signal is speech signal then in coefficient determination process, it is possible to force the determination process to analyze very simple and low-complex residual signal that is wanted as the true output of the 'black-box' system.

What form should the (simulated) output signal be? This signal can be formed to almost any shape, but when thinking the practical considerations, the final, calculated output from the system should be either zero vector or as easy as possible to model with, for example, a simple codebook. In the decoding side the inversion transform from this signal and other parameters to estimate of the original speech signal to the estimate of the original must be possible. The form of the output signal y(n) or the shaped residual in the experiment, which the method uses to estimate the parameters, is also dependable of the selected linear/non-linear combination. The property that the parameters are estimated from the output and input signals, allows countless possibilities to model the signal by forming the output signal estimate to selected form before fed in the identification algorithm. On the other hand, the output signal that will be fed in the estimation process must be chosen carefully that the selected linear and non-linear combination is able to shape the signal in to the wanted form.

The non-linearity, in a combined model, should work together with linear model and also reduce or, shape the information to a suitable form for linear filtering. Another considerations are that the total number of the parameters in speech coding should not exceed the number of parameters of traditional coding methods, and the residual modeling should be much easier when compared to the traditional coders. The combined linear and nonlinear model based system should be also as insensitive as the LPC-based speech coding system to the residual modeling inaccuracies.

The question is, what combination will produce best output in the name of parameter reduction, signal reconstruction quality and simplicity of the coding process?

This question cannot be answered directly, because there are many factors, also mentioned in this paper that will affect to the solution. But as shown in this paper, in Figs. 4 and 5 and in Table 1, the nonlinearities can improve the signal modeling better than using linear modeling methods.

The results in this paper showed that the non-linearity with adaptive weighting coefficient could reduce the information in the encoder side. However, in coding process the stability of the filter is the most essential aspect. The algorithm presented in [15] produces finite solution for the both non-linear and linear filter coefficients, but the stability of the linear filter coefficients can be ensured by checking that the zeros of the filter lies within unit circle and, if necessary, doing the minimum phase correction.

References:

[1] T. Miyano, A. Nagami, I. Tokuda and K. Aihara, "Detecting Nonlinear Determinism in Voiced Sounds of Japanese Vowel /a/," *Intern. J. of Bifurcation and Chaos*, vol. 10, no. 8, 2000, pp. 1973-1979.

[2] J. Thyssen, H. Nielsen and S. Hansen, "Non-Linear Short Term Prediction in Speech Coding," in *Proc. ICASSP,* 1994, pp. 185-188.

[3] J. Fackrell, "Bispectral Analysis of Speech Signals," PhD. Thesis, University of Edinburgh, 1996.

[4] I. Mann, "An Investigation of Nonlinear Speech Synthesis and Pitch Modification Techniques," PhD. Thesis, University of Edinburgh, 1999.

[5] G. Kubin, "Nonlinear Processing of Speech", in *Speech Coding and Synthesis*, W. Klein and K. Paliwal (Eds.), Amsterdam, Elsevier Science, 1995, pp. 559-568.

[6] ITU-T Recommendation G.711, "Pulse Code Modulation of Voice Frequencies", ITU, 10 pages, 1993.

[7] N. Ma and G. Wei, "Speech Coding with Nonlinear Local Prediction Model," in *Proc. ICASSP*, 1998, pp. 1101-1104.

[8] F. Díaz-de-María and A. Figueiras-Vidal, "Radial Basis Functions for Nonlinear Prediction of Speech in Analysis-by-Synthesis Coders," in *Proc. ICASSP,* 1995, pp. 788-791.

[9] M. Birgmeier, H. Bernhard and G. Kubin, "Nonlinear Long_term Prediction of Speech Signals," in *Proc. ICASSP,* 1997, pp. 1283-1286.

[10] M. Faundez-Zanuy, F. Vallverdú and E. Monte, "Nonlinear Prediction with Neural Nets in ADPCM," in *Proc. ICASSP,* 1998, pp. 345-349.

[11] M Banbrook, S McLaughlin and I. Mann, "Speech Characterization and Synthesis by Nonlinear Methods," *IEEE Trans. Speech Audio Processing,* vol.7, 1999, pp. 1-17.

[12] B. Townshend, "Nonlinear Prediction of Speech," in *Proc. ICASSP,* 1991, pp. 425-428.

[13] D. Westwick and R. Kearney, "Identification of a Hammerstein Model of the Stretch Reflex EMG will be processed Using Separable Least Squares," in *Proc of 22nd annual conf. of the*

IEEE Eng. in Med. And Biology Society, 2000, pp. 1901-1904.

[14] L. Hngia and J. Sjöberg, “Nonlinear acoustic echo cancellation using a Hammerstein model,” in *Proc. ICASSP,* 1998, pp. 1229-1232.

[15] J. Gómez and E. Baeyens, ‘Identification of multivariable Hammerstein systems using rational orthonormal bases,” in *Proc. 39th IEEE Conf. on Decision and Control*, 2000, (3), pp. 2849 –2854.

[16] J. Gomez, “Analysis of Dynamic System Identification using Rational Orthonormal Bases,” PhD. Thesis, University of Newcastle, 1998.