NONLINEAR FORMANT-PITCH PREDICTION USING RECURRENT NEURAL NETWORKS

Ekrem VAROGLU **Kadri HACIOGLU**
Department of Electrical and Electronics Engineering
Eastern Mediterranean University, Gazi Magosa, Mersin-10, Turkey
Tel: +90 (392) 366 65 88; Fax: +90 (392) 366 44 79; e-mail: evaroglu@salamis.emu.edu.tr

ABSTRACT

In this study, a parallel structure is proposed for the nonlinear formant and pitch prediction of speech signals using Recurrent Neural Networks (RNN) The well known Real Time Recurrent Learning (RTRL) algorithm is used as the learning algorithm. Its performance is evaluated in terms of the mean-square error and sensitivity to pitch errors through extensive computer simulations and compared to the combined formant-pitch RNN predictor and to the linear predictor.

1 INTRODUCTION

The most common model used in the speech production mechanism is the source-filter model. In this model, a linear all-pole filter is assumed and the filter coefficients are found by using linear predictive methods. Even though non-linearities exist in the speech production mechanism, in most of the practical applications, analysis and coding of speech signals, till recently were based on the above mentioned linear predictive methods due to their relatively good performance and computational efficiency. However, the success of the linear predictive methods are limited by the degree of linear relation between the speech samples. Recently non-linear techniques have become popular due to significant advances in the field of neural networks. A Time-Delay Neural Network (TDNN) [1], a Radial Basis Function (RBF) network [2], and a Recurrent Neural Network (RNN) [3] have been used as nonlinear predictors to cope with nonlinearities in speech waveforms and, hence, achieve a better prediction gain compared to that of linear predictors.

The prediction of a speech sample by using a fixed number of consecutive samples preceding the predicted sample is known as the short-term (formant) prediction, and the prediction of a sample using samples which are at a distance equal to the pitch period of the speech is known as the long-term (pitch) prediction. Even though Pitch-Formant (P-F) and Formant-Pitch (F-P) cascade predictors have been widely employed in the linear prediction, to the best of our knowledge, only the short term prediction has been studied in the case of nonlinear networks [1-3], except a brief study reported in [4]. On the other hand, it has been stated in [1] that, a nonlinear short term predictor is capable of removing most of the pitch information leaving little room for the long term prediction for further improvement.

In this study, an RNN is considered for the joint (but not cascade) short and long-term predictions where the RTRL algorithm [5] is used as the learning algorithm. In addition, we suggested a parallel network structure in which two RNNs are used- one for short-term prediction and the other for long-term prediction. We have tested the performance of the two structures with respect to the mean-square error (MSE) as well as sensitivity to pitch errors which may be due to an inaccurate estimate of the pitch period over a certain frame. Advantages of the proposed structure are clearly presented.

This paper is organized as follows. Section 2 presents the well known RNN predictor for joint long-term and short-term prediction. The proposed nonlinear predictor is introduced in section 3. Computer simulation results are discussed in section 4 and the conclusions are made in section 5

2 AN RNN PREDICTOR

A fully connected RNN consists of an input layer and a processing layer of neurons that can be classified as hidden and output neurons. The input layer is the concatenation of L external inputs, a bias input and N signals fed back from all available units. Thus, the network has a total of $N^2+(L+1)N$ weighted connections from the input layer to the corresponding processing layer. An RNN predictor with three neurons, two short-term inputs and one long-term input is as shown in Figure 1. Note that the bias term which is always one is not shown. Each neuron (or processing unit) consists of a linear combiner and a nonlinear function. Here, the bipolar sigmoid function is used as the nonlinear function. In this structure the predicted sample $\hat{x}(n)$ is the linear combination of the output of each neuron and is given by;

$$\hat{x}(n) = f(x(n-1), \cdots, x(n-M)\cdots) \qquad (1)$$

where f(·) is the overall function which is expected to realize a highly nonlinear relation and M is the pitch period. Thus, the total number of predictor coefficients becomes $N^2+(L+2)N$ where L is the summation of the number of short and long-term inputs. The network weights are adjusted by using the “backpropagated” RTRL algorithm. Being a nonlinear IIR filter, one expects the memory of the RNN to span a time interval long enough to include the pitch information. Thus, the short term prediction may be sufficient. However, the infinite memory is of fading nature and the degree of fading depends on the network size. A larger network may be capable of keeping the context information better than a smaller a network, but in the former, the major limitation is the computational complexity which is $O(N^4)$, where N is the total number of neurons. Hereafter, the network will be referred to as the combined RNN predictor.

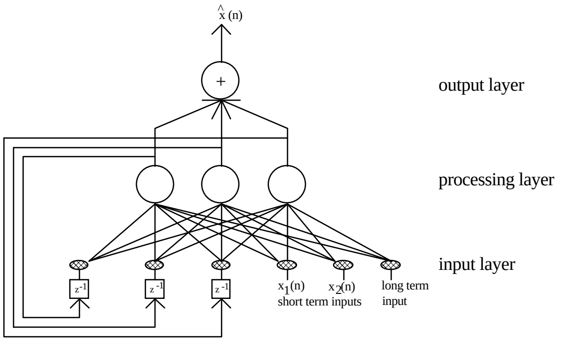

Figure 1. RNN Predictor

3 PROPOSED NONLINEAR PREDICTOR

In order to minimize the computational complexity, make learning easier, and have more freedom to keep the context information, we adopted the principle of divide and conquer, and proposed a network structure in which two separate RNNs are employed. One of the RNNs is for the short-term prediction and the other is for the long-term prediction. The network structure is shown in Fig.2 In this case the predicted sample $\hat{x}(n)$ is given by;

$$\hat{x}(n) = w_f\, f_f\,(x(n-1)\cdots) + w_p\, f_p(x(n-M+k)\cdot\cdot) \qquad (2)$$

where 2k+1 is the pitch prediction order.

This network structure decreases the computational complexity allowing relatively smaller size networks to be used for each task. The sensitivity to pitch errors is expected to decrease since the pitch prediction is separated from the formant prediction. In contrast to the combined RNN predictor, in this structure, only one neuron from each RNN block is used for the prediction.

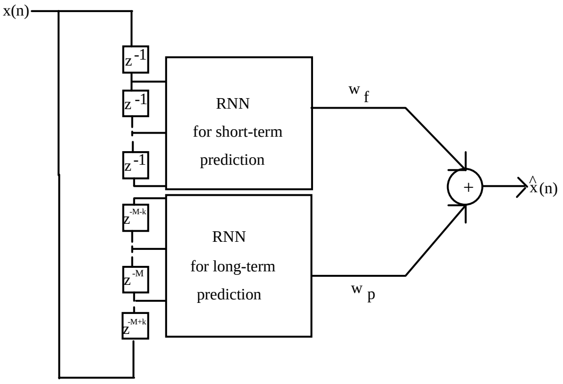

Figure 2. Proposed Nonlinear Predictor

4 SIMULATIONS AND RESULTS

In this study all simulations have been performed by using a female speech which has a very low pitch period (~3ms). The speech waveform has been lowpass filtered at 3.4 kHz cut-off frequency, sampled at 8 kHz and stored at 16-bits. All results presented are the average of 5 different trials. The analysis was performed over a frame of length 256 samples. The short-term prediction order, p, and the long-term prediction order, 2k+1, were fixed to 8 and 1, respectively. The mean-squared prediction error is used as the performance measure. The learning rate, was fixed to 0.1 for the nonlinear part and to 0.05 for the linear part throughout the simulations. Number of epochs used in each case was 2000.

The effect of the pitch prediction on the combined RNN predictor with respect to the number of neurons is shown in Figure 3. It can be seen that for small number of neurons there is a gain of 7-10 dB when the pitch prediction is included and that this gain decreases approximately to 1 dB as the number of neurons are increased. The performance of the linear predictor (with and without pitch prediction) is also included for reference. The results clearly show that the non-linear processing of speech signals using the RNN is promising and the pitch prediction should be included in small size networks.

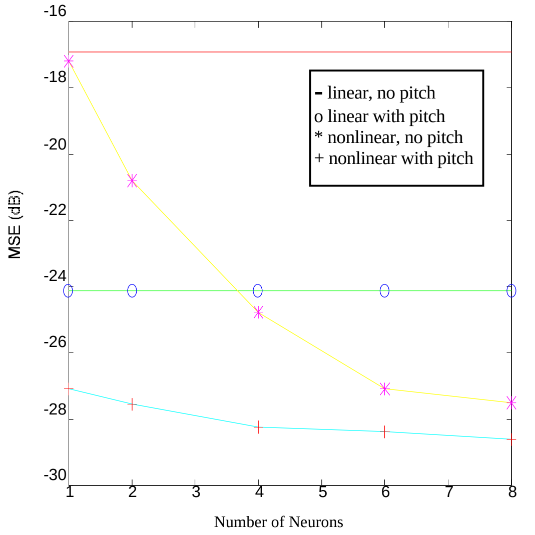

Figure 3. *Performance of linear and Nonlinear networks with and without pitch prediction.*

Table 1 compares the combined predictor and the proposed predictor in terms of MSE, order of computational complexity and the number of coefficients used. For the proposed predictor, (N_f,Np) indicates that N_f neurons are used for the formant prediction and N_p neurons are used for the pitch prediction. It can be easily seen that the proposed structure suggests alternative networks with comparable or better performance, smaller number of coefficients and reduced computational complexity.

Table 1. *Comparison of the combined predictor with the proposed predictor*

Method	# of Neurons	MSE (dB)	# of Coeff.	Computational Complexity
Combined	1	-27.10	11	O(1)
	2	-27.57	24	O(16)
	4	-28.24	54	O(256)
Proposed	(1,1)	-27.17	13	2×(O(1))
	(1,2)	-27.18	18	O(1)+O(16)
	(2,1)	-28.54	25	O(16)+O(1)
	(2,2)	-28.56	30	2×(O(16))

Figure 4 shows the performance of the combined RNN predictor with respect to pitch errors. It can be seen that when a small number of neurons are used, the network is incapable of keeping the context information and as a result the performance of the RNN drops significantly with pitch errors. As the number of neurons is increased the performance gets better (as also shown in Figure 3) and the capability of keeping the context information increases. As a result, the sensitivity to pitch errors drops at the expense of increased computational complexity.

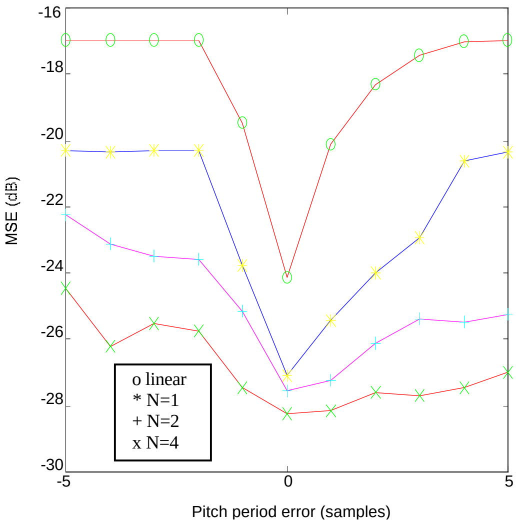

Figure 4. *Sensitivity to pitch errors of the combined RNN predictor.*

Figure 5 compares the sensitivity to pitch period errors of the two network structures when equal number of neurons are used in each case.

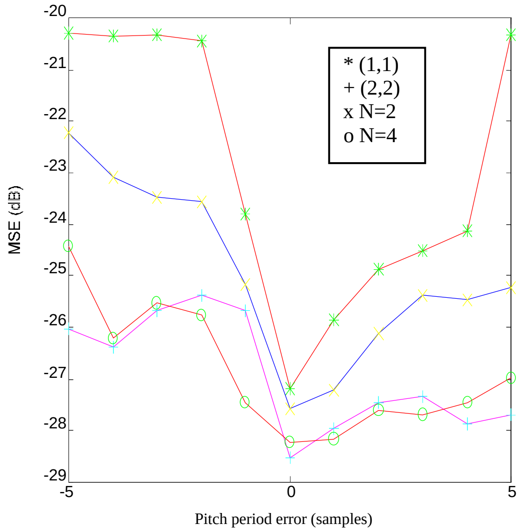

Figure 5. *Sensitivity to pitch period errors of the combined RNN predictor and the proposed predictor.*

It can be seen that when a total number of 2 neurons are used the proposed structure may not be a very good alternative for the combined predictor if compared in the sense of MSE performance. However, the former has lower computational complexity and smaller number of coefficients. Thus the proposed predictor can be chosen when these issues are a concern. On the other hand, when 4 neurons are used the proposed structure becomes a good alternative to the combined predictor in all aspects.

5 CONCLUSIONS

In this paper, a parallel structure for the nonlinear speech prediction has been proposed. This structure reduces interaction between short and long term samples. As a result, the network has been found to produce better or comparable results even with reduced network complexity. It has been demonstrated that a large network is required in the combined approach to keep the context information successfully. The proposed approach has been shown to offer a more efficient network for the same task in terms of the number of coefficients and order of the computational complexity. Currently Formant-Pitch and Pitch-Formant cascade configurations are under consideration.

REFERENCES

[1] J.Thyssen, H.Nielsen, and S.D.Hansen, "Non-linear Short-term Prediction in Speech Coding ," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, ICASSP-94, pp. 185-188.

[2] D de Maria and A. R. F. Vidal, "Nonlinear Prediction for Speech Coding using Radial Basis Functions,"Proc.IEEE Int. Conf. Acoust., Speech, Signal Processing, ICASSP-95, pp. 788-791.

[3] Wu and M. Niranjan, " On the Design of Nonlinear Speech Predictors with Recurrent Nets," Proc.IEEE Int. Conf. Acoust., Speech, Signal Processing, ICASSP-94, pp. 529-532.

[4] Wu, M. Niranjan, and F. Fallside, "Fully-Vector Quantized Neural Network Based Code-Excited Nonlinear Predictive Speech Coding," IEEE Transactions on Speech and Audio Processing, Vol.2, No.4, pp. 432-439, October 94.

[5] Williams, R. J., and Zipser, D., "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks," Neural Computation,pp. 270-280, 1989.