NONLINEAR PREDICTION FOR SPEECH CODING USING RADIAL BASIS FUNCTIONS

Fernando Díaz-de-María

Dpto. Electrónica, EPSI–Ingeniería de Telecomunicación–Universidad de Cantabria, Avda. Los Castros s/n, 39005 Santander, Spain

Aníbal R. Figueiras-Vidal

DSSR, ETSI Telecomunicación-Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain

ABSTRACT

Radial Basis Functions (RBF) networks constitute an interesting option to deal with nonlinear prediction of speech because they provide a regularized solution and, then, they can guarantee the stability of the corresponding synthesis scheme; consequently, they are adequate to be used in Code Excited Nonlinear Prediction (CENP) coders.

In this paper this approach is presented, and some simulations examples show its advantage in prediction performance. After this, the main points to arrive to practical implementations of CENP coders are addressed.

1. INTRODUCTION

Predictive coding of speech signals serves to get efficiency in representing digitally the human voice because redundancies are removed before digitization. The practical coders which have been developed till now make use of linear predictors for reasons of simplicity: they provide a reasonable redundancy reduction by using a reduced number of parameters which are obtained with a moderate computational effort.

When taking advantage of this, the price is to accept an intrinsically limited quality, due to the lack of capability of linear models to cope with the nonlinear phenomena which appear in the speech production mechanism [1]. Thus, there is a room to explore the application of nonlinear predictors in speech coding in order to obtain a higher quality; obviously, these approaches will introduce more complex schemes and will require more computational effort, as clearly perceived when reading [2,3,4,5].

When looking for practical results, some aspects have to be carefully considered: first at all, the universal approximation characteristic of the prediction scheme, which is not provided by all them [2,3]; but RBF are universal approximators [6]. Second, the kind of approximation is being proposed: RBF offer a regularized one [7], and this is useful to keep the number of modelling parameters relatively low (although their quantization remains as a problem to be studied, the parallelism of the networks opens a way for an efficient quantization, and backward adaptive schemes are also possible) (the fact of having a regularized solution would also serve to control the error amplification in the synthesis side by means of the regularization parameter); this important subject is not considered in [4], and indirect solutions are suggested in other works: [3] includes a trained codebook, reducing robustness, and [5] allows to switch between the nonlinear predictor and a linear one to cope with instability situations. Third, developing good training algorithms is essential to reach acceptable computational efforts.

The regularization character of RBF prediction is discussed in Section 2; Section 3 shows, by means of applying a correlation based model validity test [8], that the nonlinear correlations between residuals from a short-term linear prediction and the speech signal do not remain when a RBF prediction is applied. Some simulations show the improvement which RBF prediction obtains with respect to linear prediction; while other aspects, as well as some additional practical issues, are addressed at the Conclusions section, along with some suggestions to progress in solving the operational difficulties to make feasible the corresponding class of CENP speech coders.

2. THE PROPOSED RBF PREDICTOR

The prediction problem in speech coding can be stated as follows: let $S = \{(\mathbf{x}_i, x_{i+1}) \in R^p \times R \,|\, i = 1, \cdots, N\}$ a set of data pairs to approximate by means of a function f, where each pair consist on a sample to predict x_{i+1}, and a vector of samples $\mathbf{x}_i$ corresponding to

0-7803-2431-5/95 $4.00

prevoius samples of x_{i+1}. The problem of learning a mapping from examples is ill-posed, in the sense that information in the data is not sufficient to reconstruct uniquely the mapping in the region where data are not available; in addition, the data are usually noisy (the speech signal is not completely predictable). Then, some underlying idea of what the mapping should look like is needed to make the problem well-posed: smoothness is one of the most general and weakest constraints that renders the approximation possible.

The regularization approach that exploits this type of constraints leads to the RBF as a solution to the stated problem [7]. Specifically, this network computes the formula

$$f(\mathbf{x}) = \sum_{i=1}^{M} \lambda_i h_i(||\mathbf{x} - \mathbf{c}_i||) \tag{1}$$

where $\{h_i(\cdot)\}$ are the RBF, $||\cdot||$ is a norm in R^p, $\{\mathbf{c}_i\}$ are the RBF centers, $\{\lambda_i\}$ are the weights of the linear combination, and M is the number of RBF used in the approximation.

On the other hand, an important constraint in predictive coding is that the synthesis system must be stable. Therefore, the nonlinear autoregressive model based on this network

$$x_{i+1} = f(\mathbf{x}_i) + e_{i+1} \tag{2}$$

must be stable. To comply with this, the chosen RBFs $\{h_i(\cdot)\}$ have to verify [9]

$$\lim_{||\mathbf{x}||\to\infty} \frac{|f(\mathbf{x})|}{||\mathbf{x}||} = 0 \tag{3}$$

It can be easily verified that Gaussian RBF

$$h_i(x) = exp(-\frac{x^2}{\sigma_i^2}), \tag{4}$$

σ_i^2 being the variance associated to each RBF, vale it.

Consequently, the proposed nonlinear predictor is defined by three sets of parameters; specifically:

- $\{\ \sigma_i^2\ \}$: variances of the RBF;
- $\{\ \mathbf{c}_i\ \}$: centers of the RBF; and
- $\{\ \lambda_i\ \}$: weights of the linear combination;

with $i = 1, ..., M$. Then, the total number of parameters of a model with M RBFs is $(p+2)M$.

As previously said, training is an inportant practical aspect; here, to combine a moderate amount of calculations with a reasonable performance, a first solution is obtained by the Moody and Darken method [10], and it is refined using a standard gradient algorithm.

3. MODEL VALIDATION

In the case of nonlinear systems, like speech production, it is well known that the traditional validation test for linear models, which consist on computing the autocorrelation of the residual and the cross-correlation between the input signal and the residual, is useless. Billings and Voon proposed in [8] a high order correlation based validity test for nonlinear models: a nonlinear model, estimated using suboptimal least-squares procedures will be unbiased if:

$$\begin{cases} \phi_{x^{2\prime}e^2}(m) = 0, & \forall m \\ \phi_{x^{2\prime}e}(m) = 0, & \forall m \\ \phi_{xe}(m) = 0, & \forall m \end{cases} \tag{5}$$

where, in our case, x is the original speech, e is the prediction residual, ϕ is the biased estimate of the cross-correlation function, and $x^{2'}$ denotes $x^2 - \overline{x^2}$.

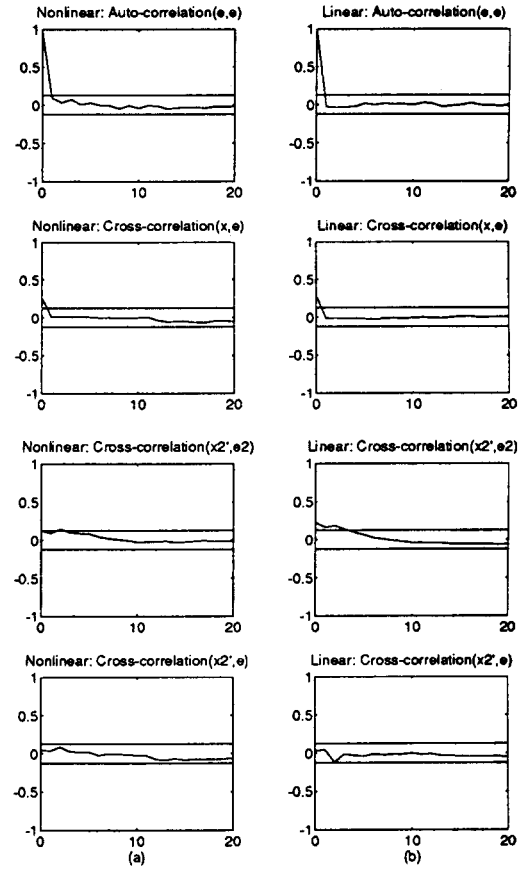

Figure 1: Model validation: $\phi_{ee}(m)$, $\phi_{xe}(m)$, $\phi_{x^{2\prime}e^2}(m)$ and $\phi_{x^{2\prime}e}(m)$. (a) Nonlinear predictor; (b) Linear predictor.

This test has been employed to validate the two models (short-term predictors) which are being com-

pared, i.e., the RBF one and the linear one, with the same number of coefficients. The RBF model has 4 Gaussian RBF and the dimension of the input vector is $p = 10$; hence it is characterized by 48 parameters. In this paper we test the RBF predictor in the scope of the standard 16 kbps CELP-LD, which employs a linear one of 50 coeffients. The results, obtained by averaging cross-correlations over a sentence of a male speaker, are shown in Figure 1: the straight lines delimit the 95 % confidence intervals, which indicate if the correlation is significant or not. Inspection of part (b) of Figure 1, corresponding to the linear model, clearly shows that $\phi_{x^2 e^2}(m)$ is outside the confidence bands, indicating that there are unmodeled nonlinear correlations. Part (a) of Figure 1 shows, however, that these correlations have been correctly modeled by the proposed RBF scheme.

4. SIMULATIONS

As a consequence of the above facts, the use of this kind of predictor must offer better results than linear schemes. The following simulations support this idea.

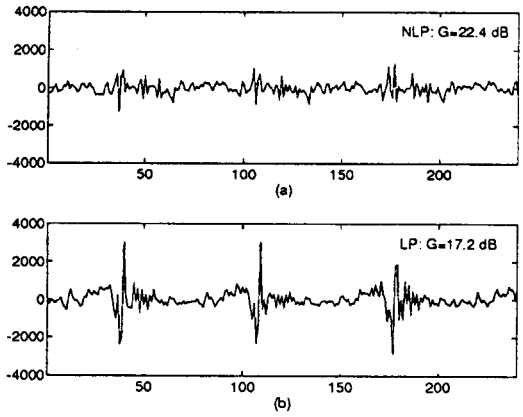

Figure 2: Residuals from a voiced speech frame: (a) RBF prediction; (b) Linear prediction.

Prediction Gain	frame type		
	voiced	transition	unvoiced
Nonlinear (dB)	15.1	13.2	7.1
Linear (dB)	13.6	11.3	5.9

Table I: Prediction gains

First, we compare the proposed short-term nonlinear predictor with a linear one with the same number of coefficients. The network which is used is a RBF model with 4 RBF and centers of dimension $p = 10$. The residuals and prediction gains obtained in both cases for one frame (30 msec.) of voiced speech (corresponding to a male speaker) are shown in Figure 2 as an illustrative example.

The averaged results over 4 sentences from 4 speakers are presented in Table I. We have distinguished among voiced, unvoiced and transition (from voiced to unvoiced and viceversa) frames. As it can be seen, the proposed model is found to perform significantly better, in particular over voiced and transition frames, than the linear procedure with the same number of parameters.

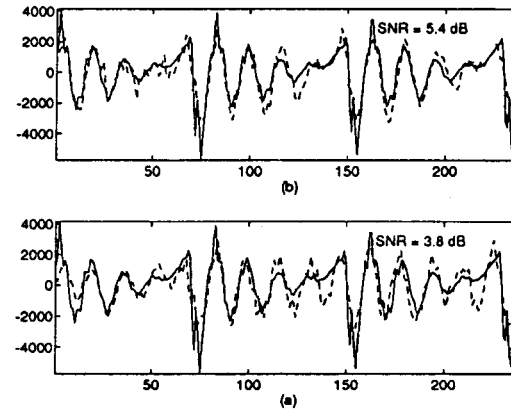

Figure 3: Original speech (solid line) and reconstructed speech (dashed line): (a) RBF predictor; (b) Linear predictor

These results suggest the possibility of applying our nonlinear predictor to the CCITT 16 kb/s speech coding standard [11]: this coder is a CELP-LD which does not use long-term predictor due to the pitch predictor adaptation sensitivity to channels errors and, to compensate for the loss in speech quality, the short-term predictor order is increased from 10 to 50. Figure 3 shows a preliminary comparison of the reconstruction performance of a CELP coder (without pitch predictor) using our predictor, with the same coder using a linear predictor with 48 coefficients (in both cases, the predictor has not been quantized, a 128 excitation codebook has been used, and the adaptation has been performed in a forward way). The results for the nonlinear and the linear schemes are presented in parts (a) and (b) of this Figure, respectively: the solid lines correspond to the original speech, and the dashed lines correspond to the nonlinear and linear reconstructed versions. These results clearly show that the reconstructed speech using

the proposed approach provides a better approximation to the actual speech signal than the linear procedure.

These results encourage the design and test of CELP-LD schemes based on this kind of predictors: additional practical aspects are discussed in the following Section.

5. CONCLUSIONS AND FURTHER WORK

After discussing the theoretical possibilities which Gaussian RBF have to perform an efficient prediction of speech signals, these advantages have been verified by means of validiting the modelling advantage with respect to linear prediction and simulating the work of two coders, one with a linear and the other using an RBF predictor.

It exits the alternative of compensating the disadvantage of linear prediction with a better excitation if enough bits are available: however, it seems reasonable to dedicate excitation to model only the strictly unpredictable part of the signal.

Some practical issues have to be addressed before proposing practical CENP schemes:

- First at all, the error amplification phenomenon which appears when synthesizing speech using quantized excitation need to be carefully solved: the use of RBF models allows to establish a compromise between closeness to the data and smoothness of the solution by selecting the value of the regularization parameter, ensuring in this way a reasonable balance between a high prediction capability and the need of having small errors in the synthesized signal when small errors appear in the excitation.

- Second, when following the above line, the loss in prediction gain, which is equivalent to an increase in residual correlation, can be partially compensated by adding a linear predictor (an thus, in turn, will probably allow a reduction in the RBF model size). This will open a way to getting practical combined predictors under reasonable guidelines to establish their sizes; and, consequently, to develop other kinds of CENP coders.

- Third: in any case, increasing the efficiency of training algorithms is needed to make useful this kind of coding schemes.

All these aspects, as well as other appearing in designing CENP coders such as designing the excitation and selecting the gain for the analysis-by-synthesis procedure, are currently under investigation.

REFERENCES

[1] H. M. Teager: "Some Observations on Oral Air Flow Vocalization"; IEEE Trans. ASSP, vol. 28, pp. 559–601; Oct. 1980.

[2] S. Haykin and L. Li: "16 kb/s Adaptive Differential Pulse Code Modulation of Speech"; Proc. Int. Workshop on Applications of Neural Networks to Telecommunications, pp. 132–138; Princeton, NJ, 1993.

[3] L. Wu, M. Niranjan and F. Fallside: "Nonlinear Predictive Vector Quantization with Recurrent Neural Nets"; Proc. 1993 IEEE-SP Workshop on Neural Networks for Signal Processing, pp. 372–381; Baltimore, MA, 1993.

[4] J. Thyssen, H. Nielsen and S. D. Hansen: "Non-Linear Short-Term Prediction in Speech Coding"; Proc. ICASSP-94, vol. I, pp. 185–188; Adelaide, Australia, 1994.

[5] E. Mumolo, A. Carini, and D. Francescato: "ADPCM With Non Linear Predictors"; Proc. EUSIPCO-94, vol. I, pp. 387–390; Edinburgh, Scotland, U.K., 1994.

[6] F. Girosi and T. Poggio: "Networks and the Best Approximation Property"; Biological Cybernetics, vol. 2, pp. 169–176; 1990.

[7] T. Poggio and F. Girosi: "Networks for Approximation and Learning"; Proc. IEEE, vol. 78, pp. 1481–1497; Sept. 1990.

[8] S. A. Billings and W. S. F. Voon: "Correlation Based Model Validity Tests for Non-Linear Models"; Int. J. Control, vol. 44, pp. 235–244; 1986.

[9] J. M. Vesin: *On Some Aspects of Non-linear Signal Modeling and its Real World Applications*; PhD Thesis, Ecole Polytechnique Federale de Lausanne, Switzerland; 1992.

[10] J. Moody and C. Darken: "Fast-learning in Networks of Locally-tuned Processing Units"; Neural Computation, vol. 1, pp. 281–294; 1989.

[11] J.H. Chen, R.V. Cox, Y.C. Lin, N. Jayant and M.J. Melchner: "A Low-Delay CELP Coder for the CCITT 16 kb/s Speech Coding Standard"; IEEE Journal on Selected Areas in Communications, vol. 10, pp.830–849; 1992.