Fully Vector-Quantized Neural Network-Based Code-excited Nonlinear Predictive Speech Coding [1]

Lizhong Wu Mahesan Niranjan Frank Fallside
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, U.K.
Tel: 44-223-332800.

Abstract

Recent studies have shown that nonlinear predictors can achieve about $2-3\ dB$ improvement in speech prediction over conventional linear predictors. In this paper, we take advantage of the nonlinear prediction capability of neural networks and apply it to the design of improved predictive speech coders. Our studies concentrate on the following three aspects: (a) the development of short-term (formant) and long-term (pitch) nonlinear predictive vector quantisers, (b) the analysis of the output variance of the nonlinear predictive filter to an input disturbance, and (c) the design of nonlinear predictive speech coders. The above studies have resulted in a fully vector-quantised, code-excited, nonlinear predictive speech coder. Performance evaluations and comparisons with linear predictive speech coding are presented. These tests have shown the applicability of nonlinear prediction to speech coding and the improvement in coding performance.

[1]Technical Report CUED/F-INFENG/TR.94, March 1992; Submitted to **IEEE Transactions on Speech and Audio Processing** in March 1992; Revised and resubmitted in May 1993.

1 Introduction

Most previous speech coding techniques are based on linear prediction. Speech coding systems with linear prediction can be described by a set of linear algebraic equations, and can be designed with the optimal solution of these equations (Atal, 1986; Kroon and Deprettere, 1988).

Linear prediction analysis is based on the assumption that the vocal tract can be approximated by a large number of short cylindrical segments with lossless transmission (Rabiner and Schafer, 1978). Without an alternative, this linear representation of the speech signal has been an important choice when nonlinear techniques have not been applicable. Linear simplification will doubtless lead to representational inaccuracy, processing inefficiency and poor coding performance (Wang et al., 1990).

Recent studies have shown that nonlinear prediction can be implemented with neural networks and various simulation results have demonstrated their applicability. In speech applications, it has been reported that neural network-based nonlinear prediction will achieve an improvement over conventional linear prediction of about $2-3\ dB$ in prediction gain (Tishby, 1990; Townshend, 1991).

In this paper, we take advantage of the nonlinear prediction capability of neural networks and apply it to the design of improved predictive speech coder. Our studies focus on:

1. the development of a neural network-based nonlinear predictive model

   In Section 2.1, we develop a nonlinear predictive model using a recurrent neural network. In this network, the outputs of the hidden-units are time-delayed and fed back to their input terminals. Linear predictors and many nonlinear predictors reported recently, e.g. (Lapedes and Farber, 1987), are special cases of this model. We describe a training algorithm and present simulation results and performance comparisons conducted using speech waveforms from the TIMIT database (Seneff and Zue, 1988).

2. the design of a nonlinear predictive vector quantiser

   When applied to speech coding, the nonlinear predictive parameters should be quantised. Usually, a low transmitted rate cannot be achieved by directly quantising the weight parameters of neural networks (Xie and Jabri, 1992). Instead, we train a finite set of nonlinear predictors to form a nonlinear predictive vector quantiser (VQ). The nonlinear prediction of each speech frame is thus encoded by the index of the selected predictor with the least predictive error. This is described in Section 2.2 of this paper.

3. the nonlinear prediction of long-term speech information

   Voiced speech consists of two types of redundant information, one between successive samples and another around adjacent pitch-periods. The former redundancy filtering is referred to as short-term (formant) prediction, and the latter as long-term (pitch) prediction. We demonstrate that not only can long-term prediction be carried out with a nonlinear model, but also that it outperforms a linear model. A long-term nonlinear predictive model and the integration of short-term and long-term predictors are studied in Section 2.3 of this paper.

4. the tolerance of the nonlinear predictive filter to an excitation disturbance

   A fully vector-quantised predictive speech coder consists of two main parts, a predictive VQ and an excitation VQ. For each frame of speech, an excitation codevector is chosen and applied to the selected predictive filter to reconstruct the speech. With a linear predictive VQ, the excitation codebook can be formed by stochastic samples. However, we find that the stochastic codebook is not suitable for nonlinear predictive coding due to the poor tolerance of the nonlinear predictive filter to an input disturbance. This is discussed in Section 3.1 in this paper.

   Two approaches have been considered to cope with this problem. One is to modify the training cost function so that the nonlinear predictive filter becomes less sensitive to its input disturbance. Another approach is to directly train an excitation codebook with the analysis-by-synthesis method, instead of using a stochastic codebook. We will concentrate on the second approach in this paper and discuss the first one elsewhere.

5. the design of a fully vector-quantised, code-excited, nonlinear predictive speech coder

   Unlike linear predictive coding, a complete form of excitation vectors no longer exists in nonlinear predictive coding. Therefore, we train the excitation codebook with a gradient descent approach. The excitation codebook is initialised by Gaussian samples and trained directly to minimise the error between the synthesised speech and the original one. This is described in Section 3.2 in this paper.

The above studies result in a fully vector-quantised, trained coded-excited nonlinear predictive (TCENLP) speech coder. Performance evaluations and comparisons are made of predictive gains, distortion-rate curves, mean opinion scores of reconstructed speech.

2 Neural Network-Based Nonlinear Prediction

Time series prediction can be defined as: Given p previous observations of the signal $s(t)$, $X = (s(t-1), \ldots, s(t-p))^T$, find out a function $g(\cdot)$ which minimises the predictive residual

$$D = \int\int \parallel s - g(X) \parallel^2 P(X,s) dX ds. \tag{1}$$

where $P(X,s)$ is the density function of the joint probability of X and s. The theoretical solution of Eqn(1) is a posterior mean estimator:

$$g(X) = \int s P_{s|X}(X,s) ds, \tag{2}$$

where $P_{s|X}(X,s)$ is the density function of the conditional probability of s given X.

With a multi-layer neural network, if the number of input-units is p and there is only one output-unit, the network can be trained as a p^{th}-order predictor. Assume that $F(\Phi, X)$ is the transfer function of network. The aim of training is to determine the architecture of the hidden-layers of the network and to adjust its weights, Φ, so that $F(\Phi, X)$ approaches $g(X)$. This is a problem of non-parametric estimation with neural networks. Several studies, e.g. (White, 1989), have demonstrated that single hidden-layer, feedforward, networks are capable of learning an arbitrarily accurate approximation to an unknown function, provided that they increase in complexity at a rate approximately proportional to the size of the training data.

Neural network-based predictors can be used for modelling data without any specific prior assumption about the form of nonlinearity. Their advantages have been reported by a number of researchers, e.g. (Lapedes and Farber, 1987). In this section, we study a general predictive model with a recurrent neural network and apply it to nonlinear predictive vector quantisation and long-term (pitch) nonlinear speech prediction.

2.1 A Neural Model of Nonlinear Prediction

A general structure of neural network-based nonlinear predictor is shown in Figure 1. It consists of three layers, which contain N_i, N_h and 1 units in the input, hidden and output layers respectively, with N_i set to the given predictive order. To predict observations with any scale of amplitude, no nonlinear activation function is imposed on the output-unit. The output of the hidden-units is delayed by τ and fed back to the inputs of the hidden-units via a weighting matrix W. This predictor

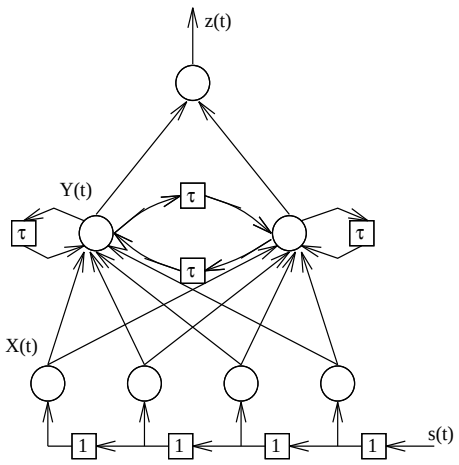

Figure 1: Neural network-based nonlinear predictor. The small squares stand for time-delay and the numbers inside the squares represent their time-delay units.

can be described by the equations:

$$\begin{aligned} Y(t) &= f\left(WY(t-\tau) + VX(t)\right) \\ z(t) &= UY(t) \end{aligned} \tag{3}$$

where U is the weight vector between the output-unit and the hidden-layer, and V is the weight matrix between the hidden-layer and the input-layer. $X(t) = (s(t-1), \ldots, s(t-N_i))^T$ is the input vector, $Y(t)$ the hidden vector, and $z(t)$ the output variable. $f(\cdot)$ is a differentiable nonlinear function. In this paper, we define $f(\cdot)$ as the commonly used sigmoid function and set the time delay τ to one sample-period.

In Eqn(3), if $W = 0$, the network is of feedforward type. This form of predictors has been widely studied, e.g. (Lapedes and Farber, 1987). Moreover, if $N_h = N_i$, $V = I$, where I is the identity matrix, $f(X) = X$, $W = 0$ and $U = \alpha$, then $z(t) = \alpha X(t)$, and the predictor becomes the linear one.

After the architecture and size of the neural network have been decided, U, V and W can be trained. Let Φ be the set $\{U, V, W\}$. Then,

$$\Delta\Phi = \eta e(t) \nabla_\Phi z(t) \tag{4}$$

where η is a learning rate and $e(t) = s(t) - z(t)$. Back-propagating $\nabla_\Phi z(t)$ to $\nabla_\Phi Y(t)$, we find

$$\begin{aligned} \nabla_U z(t) &= Y(t) \\ \nabla_W z(t) &= U \nabla_W Y(t) \\ \nabla_V z(t) &= U \nabla_V Y(t). \end{aligned} \tag{5}$$

Defining a matrix $A = W\nabla_\Phi Y(t-\tau)$ and a column vector $G = f'(WY(t-\tau) + VX(t))$, $\nabla_W Y(t)$ and $\nabla_V Y(t)$ can recursively be computed from

$$\begin{aligned} \frac{\partial y_j(t)}{\partial w_{rs}} &= g_j[a_{rs} + \delta_{jr} y_s(t-\tau)] \\ \frac{\partial y_j(t)}{\partial v_{mn}} &= g_j[a_{mn} + \delta_{jm} s(t-n)], \end{aligned} \tag{6}$$

for $1 \le j, r, s, m \le N_h$ and $1 \le n \le N_i$.

2.2 Nonlinear Predictive Vector quantisation

Like the linear predictive case (Wu and Fallside, 1991), a nonlinear predictive VQ consists of a set of predictors $\{F(\Phi_k, X),\ k = 1, \ldots, K\}$. The number of predictors K is equal to 2^r, r (in bits) being the size of quantiser. The performance of predictive quantisers is evaluated by their distortion-rate functions. In the nonlinear predictive case, each predictor is trained to cover certain regions of the nonlinear predictive parameter space. The nonlinear predictive quantiser is therefore expected to cope with the nonstationarity of the predicted signal and further improve the predictive performance over that of an individual predictor. During quantisation, each frame of speech is successively applied to all the predictors in the quantiser. The predictor with the least predictive error, averaged over the whole frame, is then selected to quantise the current frame. The process can be described by the equation:

$$c = \arg \min_{1 \le k \le K} \sum_t^N \| s(t) - F(\Phi_k, X(t)) \|^2, \tag{7}$$

where N is the frame length. The nonlinear prediction of the speech frame is then represented by a code symbol of r-bit length instead of the weight parameters of the selected predictor.

The training process of the nonlinear predictive quantiser is the same as that for a single predictor, except that the cost function becomes

$$D = \sum_k^K \sum_{\mathbf{S(t)} \in k} \sum_t^N \| s(t) - F(\Phi_k, X(t)) \|^2, \tag{8}$$

and that the updating equation becomes

$$\Delta\Phi_k = \eta \sum_{\mathbf{S(t)} \in k} \sum_t^N \| s(t) - F(\Phi_k, X(t)) \| \nabla_{\Phi_k} z_k(t), \quad \text{for} \quad k = 1, \ldots, K. \tag{9}$$

where $\mathbf{S(t)} = \{s(t), t = 1, 2, \ldots, N.\}$.

Unlike linear prediction, however, the output of a nonlinear predictor cannot be simply expressed as the sum of both the zero-state response and the zero-input response. The memory effect of the

last selected predictor cannot be directly removed by subtracting the current frame speech with the zero-input response. The memory of the last frame should be taken into account during the predictor selection for the current frame.

2.3 Long-term (Pitch) Nonlinear Speech Prediction

Speech signal consists of two types of redundancy or correlation: a short-term one between successive speech samples and a long-term one between adjacent pitch-periods. Consequently, the linear prediction speech production model contains two time-varying linear predictive filters, a short-term one and a long-term one. The former gives the spectral envelope of the speech signal and the latter reproduces the spectral fine structure. The long-term linear prediction of speech has been studied, for example, by (Ramachandran and Kabal, 1989).

We studied the implementation of the long-term nonlinear prediction of speech using a neural network. The structure chosen for the long-term nonlinear predictive model is the same as that for the short-term one shown in Figure 1, but the input is delayed by one pitch-period before being applied to the input-layer. The current sample is thus predicted from the samples around its last pitch-period. The long-term nonlinear predictor is expressed by the following equation [2]:

$$\begin{aligned} Y_l(t) &= f(W_l Y_l(t-\tau) + V_l X_l(t-T)) \\ z_l(t) &= U_l Y_l(t), \end{aligned} \tag{10}$$

where T is the pitch-period, which is found from the short-time autocorrelation function of the short-term residual (Rabiner and Schafer, 1978):

$$r_s(t) = s(t) - z_s(t). \tag{11}$$

Long-term prediction is always accompanied by short-term prediction in speech coding. The long-term predictor can be connected to the short-term one in cascade or in parallel. In the cascade form, the short-term residual is delayed by one pitch-period and fed to the long-term predictor. In the parallel form, the original speech is delayed by one pitch-period and applied to the long-term predictor. So, in Eqn(10),

$$X_l(t) = (r_s(t - N_{li}/2), \ldots, r_s(t + N_{li}/2))^T \tag{12}$$

for the cascade form and

$$X_l(t) = (s(t - N_{li}/2), \ldots, s(t + N_{li}/2))^T \tag{13}$$

[2] We use the subscript l to identify the long-term predictor and s the short-term one.

for the parallel form. The final residual of the combined short-term and long-term prediction for both the cascade and the parallel forms is,

$$r(t) = s(t) - z_s(t) - z_l(t). \tag{14}$$

The short-term and long-term predictors are trained in sequence. First, the short-term predictor is trained to minimise the short-term residual, $r_s(t)$. With the short-time autocorrelation function of $r_s(t)$, the pitch-period is then estimated (Rabiner and Schafer, 1978). Finally, the long-term predictor is trained to reduce the combined short-term and long-term residual $r(t)$.

2.4 Nonlinear Predictive Quantisation Performance

Besides network architecture, two parameters, the predictive order and the total number of weights, also affect the property of predictors. We compared the predictive quantisation performance between different architectures under either of following two conditions: (a) same predictive order, or (b) same number of weights.

We have used the following notations, $LP(p)$ stands for a p^{th}-order linear predictor, and $NLP(N_i - N_h)$ stands for a nonlinear predictor with N_i input-units and N_h hidden-units with recurrent neural network architecture. In all of our studies, N_h is set to two.

Since linear prediction is an all-pole model, good results will only be obtained when the overall speech spectrum is like the response of an all-pole filter. For vowel sounds, the all-pole model can make good approximations to the spectral shapes of formants. But for the nasals and some voiced consonants, all-pole spectral modelling is poor and its performance usually saturates at a very small predictive order. A general nonlinear model may be more suitable.

As an example, Figure 2 compares the spectra of linear and nonlinear predictive residuals for a frame of nasal speech. We see that when the predictive order is increased from 10 to 26, only a small improvement is obtained for the linear predictor. The $NLP(10-2)$, with a predictive order equal to 10 and a total number of weights of 26, outperforms both the $LP(10)$ and the $LP(26)$. Its predictive residual is closer to white noise.

The nonlinear predictive quantiser is now evaluated with continuous speech data. The speech samples were from the TIMIT database (Seneff and Zue, 1988) and were pre-filtered to 8 kHz. The training set consists of ten different spoken sentences by ten different speakers (five females). The test set consists of another four different spoken sentences by another four different speakers (two females).

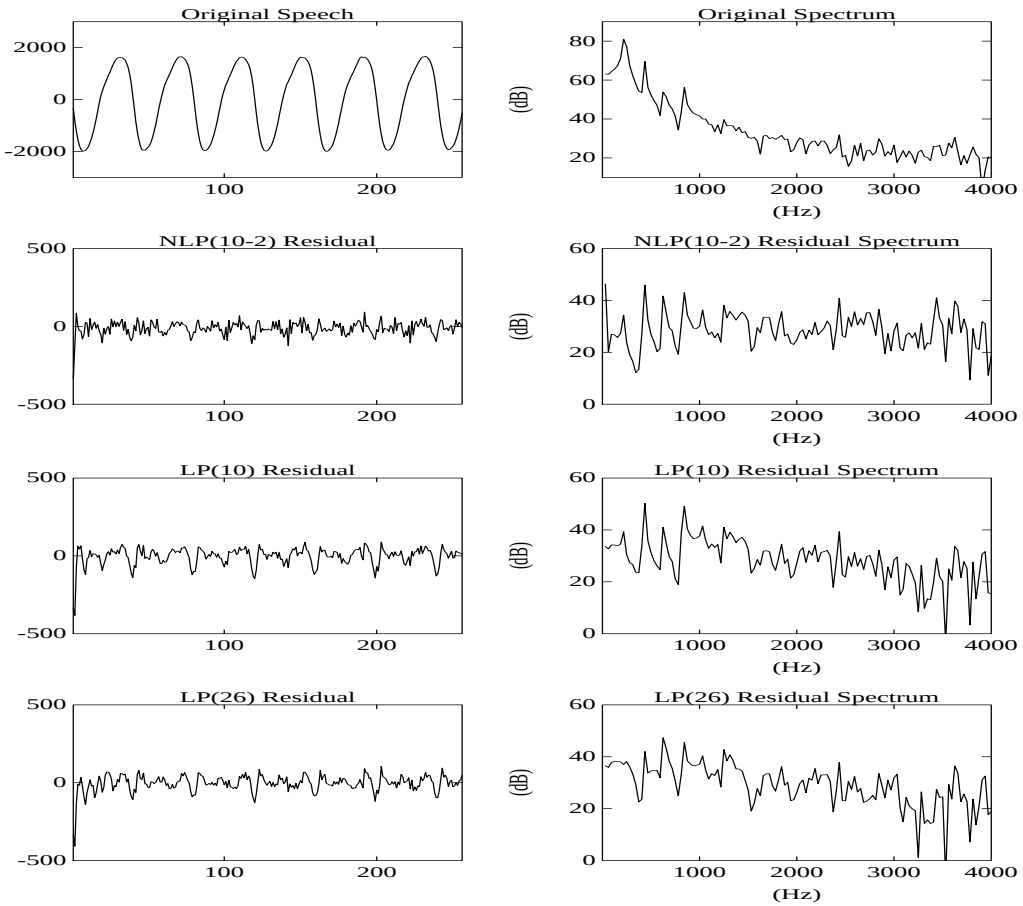

Figure 2: Spectra of linear and nonlinear predictive residual of nasal speech.

Figure 3 compares the predictive gain-rate functions of the linear and nonlinear predictive VQs. Their sizes vary from one-bit to six-bits. The nonlinear predictive VQ is formed by 2^{r_s} $NLP(10-2)$ recurrent neural networks, where r_s (in bits) is the size of the short-term predictive VQ. The frame length was set to 256 samples without overlapping between frames. Therefore, the transmitted rate is $\frac{r_s}{256}$ bits/sample for the nonlinear predictive information.

The linear predictive VQ was designed using the approach developed in (Buzo et al., 1980). Each frame consists of 256 or 384 Hamming-windowed samples and is overlapped by 128 samples. Therefore, the transmitted rate is $\frac{r_s}{128}$ bits/sample for the frame size of 256 and $\frac{r_s}{256}$ bits/sample for the frame size of 384.

Figure 3 also shows the predictive gain-rate functions of the combined short-term and long-term VQs. The sizes of long-term predictive VQs are also from one-bit to six-bits. The long-term predictive VQ is cascaded with a six-bit, short-term, predictive VQ. Each predictor in the long-term nonlinear predictive VQ is a $NLP(3-2)$ recurrent neural network, and that in the linear predictive VQ is a $LP(3)$ linear predictor.

The pitch information is estimated using the short-time autocorrelation function of the short-term predictive residuals (Rabiner and Schafer, 1978). It is updated every 64 samples. The long-term predictor is switched on only if the pitch-period was not equal to zero.

The design of the long-term linear predictive VQ was based on (Davidson and Gersho, 1987; Ramachandran and Kabal, 1989). The predictor parameters were vector-quantised using a one-step identification/compression technique. A codebook of long-term predictor parameters is exhaustively searched to identify which predictor minimises the short-term residual.

The frame length of long-term nonlinear prediction is 64 samples as is the processing step-size of long-term linear prediction. Therefore, their transmitted rates are the same and equal to $\frac{r_l}{64}$ bits/sample, where r_l is the size of long-term predictive VQ in bits.

From inspection of Figure 3, we conclude that all nonlinear predictive quantisers outperform linear ones either with the same predictive order or with equal numbers of weights. The test-performance of linear predictive VQs is $0.5-1$ dB worse than the training-performance. In contrast, with the same training and test data sets, Figure 3 shows that the test-performance of nonlinear predictive VQ is close to, or even better than, the training-performance. Therefore, we can also conclude that the nonlinear predictive VQ generalises better than the linear one.

As an example, Figure 4 plots a segment of short-term nonlinear predictive residual and the combined short-term and long-term nonlinear predictive residual obtained from the test set.

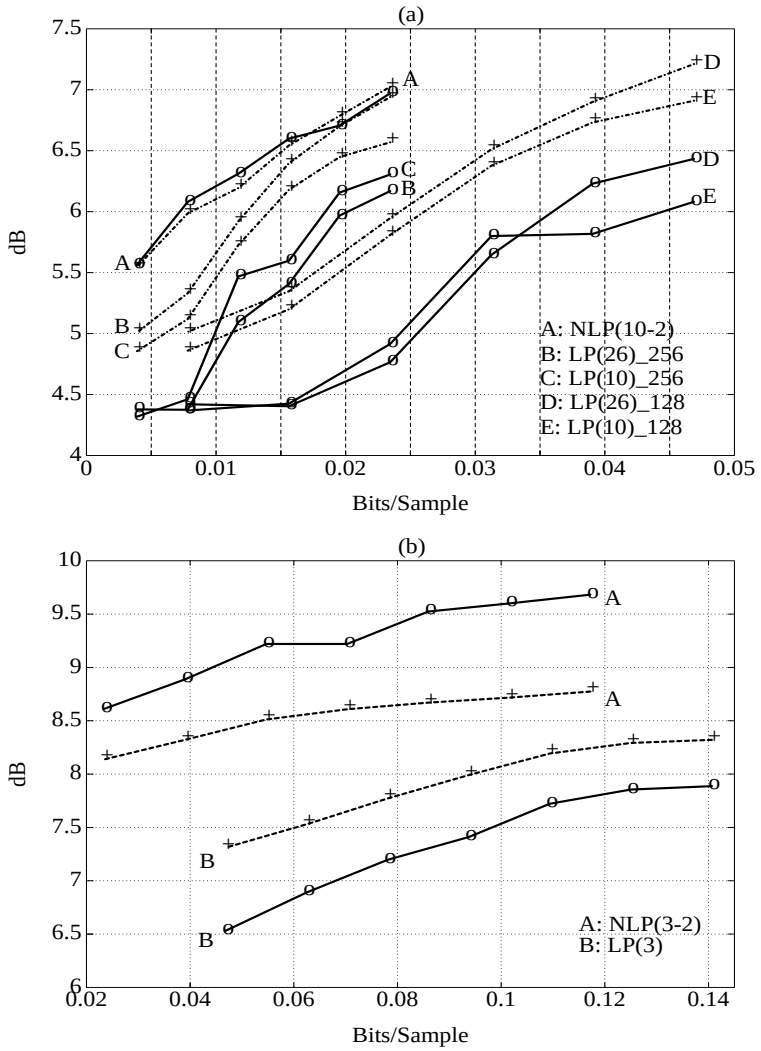

Figure 3: Comparison of the predictive gain-rate functions of linear and nonlinear predictive VQs. Panel (a) is for the VQs with short-term prediction only and panel (b) for those with combined short-term and long-term prediction in the cascade form. The training-performance is connected by dashed curves, and the test-performance is connected by solid ones. In panel (a), "-256" and "-128" represent the frame advance. The transmitted rate of short-term VQs is $\frac{r_s}{256}$ bits/sample for nonlinear prediction and $\frac{r_s}{128}$ or $\frac{r_s}{256}$ for linear prediction. Because the long-term VQs are cascaded with a 6-bit, short-term VQ, their transmitted rate is $(\frac{6}{256} + \frac{r_l}{64})$ bits/sample for nonlinear prediction and $(\frac{6}{128} + \frac{r_s}{64})$ for linear prediction (with the frame advance of 128 in short-term prediction). r_s and r_l are respectively the sizes of short-term and long-term VQs in bits. The transmitted rate for coding the pitch information has not been included in this figure.

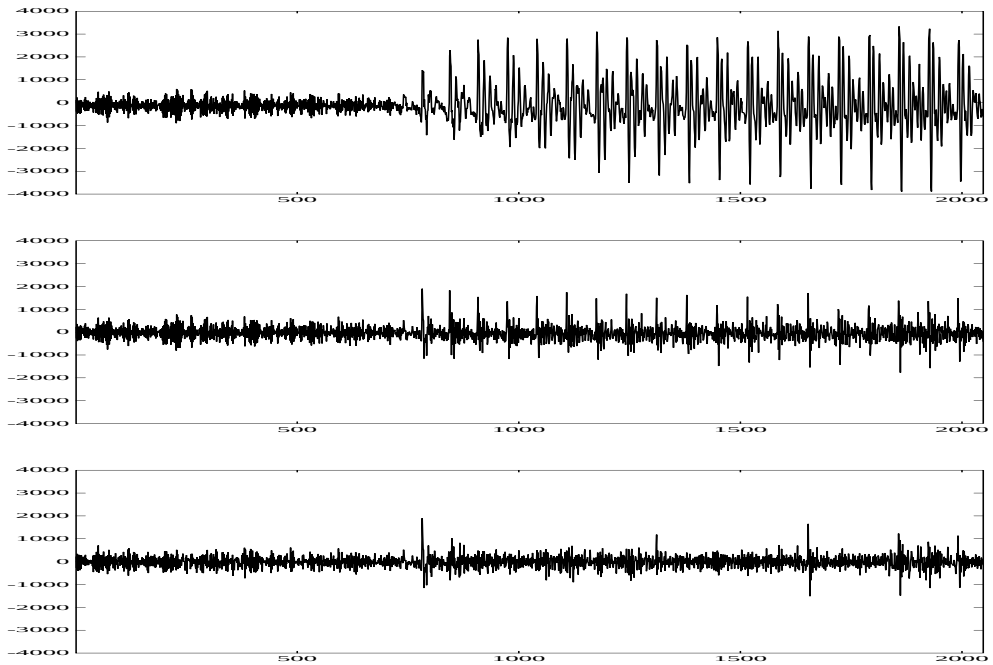

Figure 4: Nonlinear predictive residuals of speech. From the top downwards, they are respectively the original speech, the short-term nonlinear predictive residual, and the combined short-term and long-term nonlinear predictive residual in the cascade form.

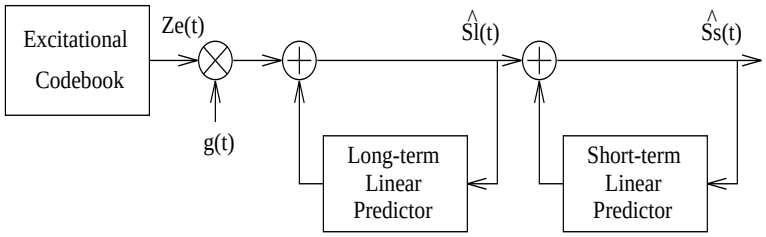

Figure 5: Code-excited linear predictive (CELP) speech coder (Schroeder and Atal, 1985).

3 Code-excited Nonlinear Predictive Speech Coding

In a code-excited linear predictive speech coding (CELP) system, as shown in Figure 5, speech is synthesized by passing an excitation vector through a cascade of short-term and long-term linear predictive filters. The excitation codebook usually performs the codevector search in an analysis-by-synthesis manner (Kroon and Deprettere, 1988). The excitation codebook may either be stochastic, i.e. pseudo-randomly populated, or pre-designed over some training data for minimum global distortion.

In a code-excited nonlinear predictive speech coding system, the linear predictive filters in Figure 5 are replaced by nonlinear ones. In this section, we investigate the problems caused by such a replacement and study how to achieve better coding performance based on the capability of nonlinear predictive vector quantisation.

3.1 Tolerance of Nonlinear Predictive Filter to an Excitation Disturbance

In the prediction mode, the inputs are the original speech samples. In the synthesis mode, however, the predictive filter receives only the estimations of speech. The difference between the estimation and the original speech signal is referred to as an excitation or input disturbance.

In this sub-section, we will analyse the tolerance of nonlinear predictive filter to an excitation disturbance. We show that the stochastic excitation codebook does not apply to nonlinear predictive coding due to its poor tolerance.

Assume that the input of a given predictive filter is changed from X_0 to X with $D_x = \| X - X_0 \|^2$, and their corresponding outputs are respectively z and z_0 with $D_z = \| z - z_0 \|^2$.

In the linear prediction case, $z = \alpha^T X$, where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ is the p^{th}-order linear

predictive coefficient vector. We easily find

$$D_z = \| z - z_0 \|^2 \leq \| \alpha \|^2 D_x. \tag{15}$$

In the nonlinear prediction case, we have (see Appendix)

$$D_z \leq \rho^2 D_x, \tag{16}$$

where

$$\rho = \frac{\gamma \| U \| \| V \|}{\gamma \| W \| - 1}, \tag{17}$$

$\| U \|$, $\| V \|$ and $\| W \|$ are the Euclidean norms of U, V and W respectively, $\gamma = \max | f'(O^*(t)) |$, $O^*(t) \in (O(t), O_0(t))$, and $O(t) = WY(t-r) + VX(t)$.

Assume that two predictive filters have sensitivity parameters of ρ_1 and ρ_2. If they are connected in cascade, the sensitivity parameter of the combined filter is $\rho = \rho_1 \rho_2$, and if they are connected in parallel, the sensitivity parameter of the combined filter is $\rho = \rho_1 + \rho_2$.

Eqns (15) and (16) give the upper bounds of the output variations of linear and nonlinear predictive filters, respectively, due to an input disturbance.

As shown in the Appendix

$$0 \leq \gamma \| W \| < 1,$$

so

$$(\gamma \| W \| - 1)^2 < 1.$$

If $(\gamma \| W \| - 1)^2$ is close to zero, then ρ becomes very large. Therefore, the output will greatly deviate from its original value, even for a small input disturbance.

Table 1 compares the sensitivity parameter, $c = \frac{D_z}{D_x}$, of the linear and nonlinear predictive filters for a VQ of four-bit size. All nonlinear units have a sigmoid function $f(x) = \frac{2}{1+e^{-2x}} - 1$; therefore $\gamma = 1$. We see that the sensitivity parameter of nonlinear predictive filters is much greater than the linear one, in particular when a predictive order is large.

To deal with the poor tolerance of the nonlinear predictive filter to an input disturbance in a nonlinear predictive coding system, two approaches can be applied:

1. as an accompanying training criterion, to minimise the sensitive parameters at the same time as minimising the predictive error during the design of nonlinear VQs;

2. instead of using the stochastic codebook, to use the trained excitation codevectors in the code-excited nonlinear predictive coder.

Predictor	*Predictive Order*			
	4		14	
	Nonlinear	*Linear*	*Nonlinear*	*Linear*
1	24.90	1.28	63.40	3.91
2	25.88	1.29	48.17	2.92
3	7.63	1.14	422.50	4.25
4	6.18	2.18	37.28	2.81
5	768.55	7.03	1004.43	2.65
6	470.50	1.48	530.01	2.11
7	70.82	5.20	19.76	2.04
8	47.15	1.37	13.93	1.84
9	23.73	0.35	66.55	1.19
10	65.67	0.35	58.79	1.03
11	25.75	0.21	27.03	0.92
12	12.76	0.25	10.37	0.80
13	37.72	0.35	15.09	0.57
14	7.23	0.24	6.42	0.44
15	2.50	0.18	11.50	0.23
16	1.23	0.14	6.59	0.58

Table 1: Comparison in the sensitivity parameters of predictive filters in linear and nonlinear predictive VQs.

We discuss the first approach elsewhere, and focus here on the second approach.

3.2 Trained Code-excited Nonlinear Predictive Coding

In a nonlinear predictive coder, the complete form of the excitation codevectors does not exist. Therefore, we obtain the excitation codebook with the gradient descent technique.

Assuming N_e is the dimension of the excitation vectors, $Z_e = (z_{e1}, \ldots, z_{eN_e})^T$, we define $\hat{S}(t)$, $\hat{S}_l(t)$, $Z_s(t)$ and $Z_l(t)$ as column vectors of size N_e, and $\mathbf{Y}_s(t)$, $\mathbf{Y}_l(t)$, $\mathbf{X}_s(t)$ and $\mathbf{X}_l(t)$ as $N_{sh} \times N_e$, $N_{lh} \times N_e$, $N_{si} \times N_e$ and $N_{li} \times N_e$ matrices respectively [3]. The predictive speech coder of a recurrent neural network-based, short-term and long-term cascade, nonlinear predictive VQ can be expressed as shown below:

$$\begin{aligned} \mathbf{Y}_l(t) &= f(W_l \mathbf{Y}_l(t-\tau) + V_l \mathbf{X}_l(t-T)), & Z_l(t) &= [U_l \mathbf{Y}_l(t)]^T, \\ \mathbf{Y}_s(t) &= f(W_s \mathbf{Y}_s(t-\tau) + V_s \mathbf{X}_s(t)), & Z_s(t) &= [U_s \mathbf{Y}_s(t)]^T, \end{aligned} \tag{18}$$

where

$$\begin{aligned} \mathbf{X}_l(t) &= (\hat{S}_l(t - N_{li}/2), \ldots, \hat{S}_l(t + N_{li}/2))^T \\ \mathbf{X}_s(t) &= (\hat{S}(t-1), \ldots, \hat{S}(t - N_{si}))^T, \end{aligned} \tag{19}$$

and

$$\begin{aligned} \hat{S}_l(t) &= Z_l(t) + Z_e(t) \\ \hat{S}(t) &= Z_s(t) + \hat{S}_l(t). \end{aligned} \tag{20}$$

The excitation vector Z_e is chosen from the excitation codebook by reducing the coding distortion

$$D(t) = \| E(t) \|^2 = \| S(t) - \hat{S}(t) \|^2. \tag{21}$$

The training process of Z_e is outlined below. For simplicity, we assume that the coder consists of a short-term predictive quantiser only. For the case of a combined short-term and long-term predictive quantiser, the training algorithm can be derived analogously.

From the gradient descent training algorithm, we have

$$\Delta Z_e = \eta E(t) \nabla_{Z_e} \hat{S}(t). \tag{22}$$

Since

$$\hat{S}(t) = Z_s(t) + Z_e(t), \tag{23}$$

[3] To avoid confusing $Y(t)$ and $X(t)$ in Eqn(3) and Eqn(10), the bold font is used here.

then

$$\nabla_{Z_e}\hat{S}(t) = \nabla_{Z_e}Z_s(t) + \nabla_{Z_e}Z_e(t). \tag{24}$$

The back-propagated $\nabla_{Z_e}Z_s(t)$ in the nonlinear predictor network is

$$\begin{aligned} \nabla_{Z_e}\mathbf{Y}_s(t) &= [W_s\nabla_{Z_e}\mathbf{Y}_s(t-\tau) + V_s\nabla_{Z_e}\mathbf{X}_s(t)]\,{G_s}^T \\ \nabla_{Z_e}Z_s(t) &= U_s\nabla_{Z_e}\mathbf{Y}_s(t), \end{aligned} \tag{25}$$

where

$$\begin{aligned} \nabla_{Z_e}\mathbf{X}_s(t) &= (\nabla_{Z_e}\hat{S}(t-1), \ldots, \nabla_{Z_e}\hat{S}(t-N_{si}))^T \\ G_s &= f'(W_s\mathbf{Y}_s(t-\tau) + V_s\mathbf{X}_s(t)), \end{aligned} \tag{26}$$

and the back-propagated $\nabla_{Z_e}Z_e(t)$ is equal to an $N_e \times N_e$ diagonal identity matrix.

3.3 Gain-adaptive Nonlinear Predictive Coding

Nonlinear predictive neural networks can be trained to fit the signal with any scale of amplitudes at their outputs since no nonlinear activation function is imposed on their output-units, nonetheless, a gain normalisation step may still be desirable to smooth the outputs and reduce the output variances so as to improve the adaptive ability of coding. Unlike linear predictive speech coding, the gain-term is placed at the output of the nonlinear predictive filter as shown in Figure 6 instead of at the output of the excitation codebook (both places are equivalent to the linear case). Such an architecture leads to $g(t)$ being independent of the $Z_p(t)$ and its optimal solution can be found as

$$g(t) = \frac{[S(t)]^T[Z_e(t) + Z_l(t) + Z_s(t)]}{\| Z_e(t) + Z_l(t) + Z_s(t) \|^2}. \tag{27}$$

This results in the coding error

$$D = \| S(t) - g(t)[Z_e(t) + Z_l(t) + Z_s(t)] \|^2 \tag{28}$$

$$= \| S(t) \|^2 - \frac{\{[S(t)]^T[Z_e(t) + Z_l(t) + Z_s(t)]\}^2}{\| Z_e(t) + Z_l(t) + Z_s(t) \|^2}. \tag{29}$$

Therefore, the gain-adaptive nonlinear predictive speech coder can be trained by minimising the above D or maximising the second term of Eqn(29).

3.4 Coding Performance

Figure 7 compares the coding performances of the trained code-excited nonlinear predictive speech coder (TCENLP) to that of the code-excited linear predictive speech coder (CELP), using the same training and test speech data sets as described in Section 2.4.

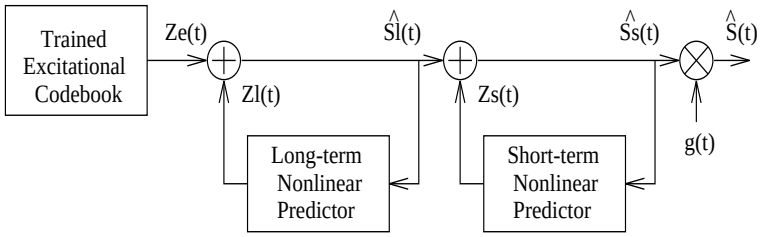

Figure 6: Trained code-excited nonlinear predictive (TCENLP) speech coder.

Parameters	*Bits/Sample*
Code of short-term predictive VQ	$\frac{6}{N_{sp}}$
Code of long-term predictive VQ	$\frac{6}{64}$
Code of excitation codebook	$\frac{B_e}{N_e}$
Pitch	$\frac{B_p}{64}$
Gain	$\frac{5}{N_e}$

Table 2: Bit allocation in speech coders. N_{sp}, the frame length of short-term prediction, is equal to 128 samples for nonlinear predictive coders and 256 for linear ones. B_e is the size of the excitation codebook and equals 7, 8 or 9 bits. B_p is the code-length of pitch-information, which is 8 bits if the pitch-period is not equal to zero and one bit for zero pitch-period. N_e is the length of excitation vectors which is set to 32 or 64.

Table 2 lists the bit allocation in the speech coders. In all the simulations, the transmitted rate varies from 0.34 bits/sample to 0.68 bits/sample, i.e. from 2720 *bps* to 5440 *bps*.

The excitation codebook was initialised by Gaussian samples and trained over the training set. To observe the sensitivity of the excitation codebook, its size was varied from 7 to 9 bits, and the dimension of codevectors was set to 32 or 64. Because the pitch-period was limited between 20 to 148 samples (from 2.5 to 18.5 *ms*), it could be coded with 7 bits without loss of information. The gain-value was quantised with a Max-Lloyd quantiser (Lloyd, 1957; Max, 1960). Its design was based on the same training data set. We found that the coding performance can be improved by sequentially re-optimising the excitation codebook and the gain codebook (Wu and Fallside, 1992). This was accomplished by recursive adjustment of the entries of each codebook. One codebook was fixed when another was updated.

The CELP speech coder design was based on (Trancoso and Atal, 1986; Davidson and Gersho,

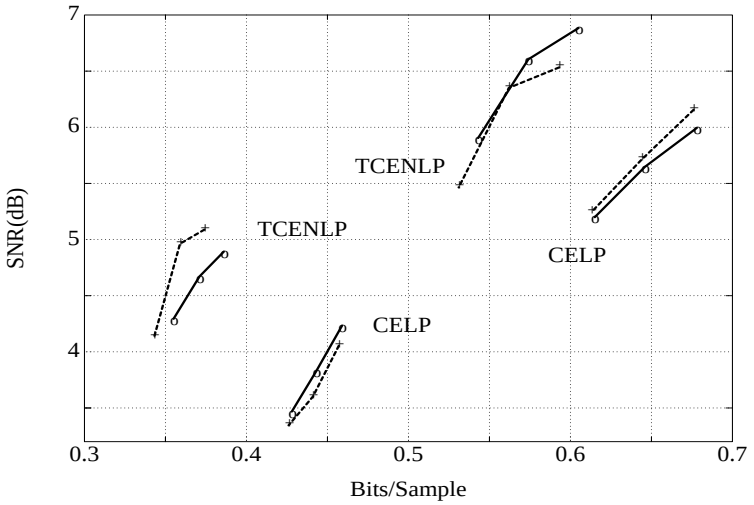

Figure 7: Comparison of the coding performance of the trained code-excited nonlinear predictive speech coder (TCENLP) and the code-excited linear predictive speech coder (CELP). The solid curves plot the test-performance and the dashed curves plot the training-performance. There are two groups of curves for each coder, one corresponds to $N_e = 64$ and another to $N_e = 32$. The three points on each curve refer to $B_e = 7$, 8 and 9 bits. The bit allocation of other coding parameters is listed in Table 2.

1987; Kroon and Deprettere, 1988). The short-term and long-term linear predictive quantisers were described in Section 2.4. The excitation codebook was formed by Gaussian samples and fully searched during coding.

Figure 8 shows the waveform and spectrogram of a segment of reconstructed speech in the test set. We see that the basic information, such as the formant positions, amplitudes and transitions, has been reproduced in the reconstructed spectrogram.

Subjective evaluation was carried out to compare speech quality of the nonlinear and linear predictive coders. A Mean Opinion Score with five-point categories was used to evaluate the speech quality. The rating of speech quality, as described by its corresponding listening effort, is shown in Table 3. Ten subjects participated in the experiment and asked to judge three sections of speech. These three sections were respectively the 8 kHz-downsampled original speech from the TIMIT database, the reconstructed speech from a 4840 *bps* TCENLP speech coder and the reconstructed speech from a 4920 *bps* CELP speech coder. The speech was from the test data set. It consisted of four different sentences, each spoken by a different speaker. On average, the nonlinear predictive coder scored 3.00 points and the linear one 2.65 points. As a reference, the original speech scored

Rating	Quality	Listening Effort
5	Excellent	Complete relaxation possible, no effort required
4	Good	Attention necessary, no appreciable effort required
3	Fair	Moderate effort required
2	Poor	Considerable effort required
1	Bad	No meaning understood with any feasible effort

Table 3: Rating of opinion scores of speech quality and listening effort (Natvig, 1988).

4.00 points. Among the ten subjects, eight judged that the nonlinear predictive coder produced higher-quality reconstructed speech. Only one subject gave an inverse judgement. Another subject said that both kinds of reconstructed speech were not significantly different.

Figure 8: From the top downwards: the spectrogram and the original speech, the spectrogram and the residual with a 940 *bps* nonlinear predictive VQ, and the spectrogram and the reconstructed speech with a 4840 *bps* TCENLP speech coder. The frequency range is $0 - 5\ kHz$. The waveform consists of 10240 samples (1.28 seconds).

4 Concluding Remarks and Further Work

4.1 Concluding Remarks

For a non-Gaussian signal, like speech, nonlinear prediction will always achieve better predictive gain than linear prediction. This paper has studied the implementation of nonlinear predictive vector quantisation with a set of recurrent neural networks. It was demonstrated that the nonlinear predictive VQ with combined short-term (formant) and long-term (pitch) prediction outperforms the linear counterpart by $2-2.5$ dB in the predictive gain when the transmitted rate is between 360 bps and 940 bps.

The excitation codebook in nonlinear prediction speech coding cannot be formed from stochastic samples due to the poor tolerance of the output variance of nonlinear predictive filters to input disturbances. Instead, a trained excitation codebook should be used. Trained code-excited nonlinear predictive (TCENLP) speech coders achieve better coding performance than conventional CELP speech coders. For a coder with about 4800 bps, the advantage in performances is about 1.5 dB in SNR and 0.35 point in a five-category Mean Opinion Scale, on the test data set.

The complexity of the excitation codebook search procedure in the TCENLP coder is the same as that of the CELP. More computational effort is required in neural computation of nonlinear predictive vector quantisation. Compared to the single mapping of linear prediction, the nonlinear predictor network needs twice mappings, first from the inputs to the hidden-units and then from the hidden-units to the output. Because we use the network which contains only two hidden-units, increase in computation is not very high.

In summary, the studies in this paper have shown the applicability of nonlinear prediction to speech coding, the improvement of the coding performance over linear prediction, and the implementation of a fully vector-quantised TCENLP speech coder with neural networks.

4.2 Further Work

In conventional CELP speech coders, several techniques are important to practical implementations, but these have not been studied for the TCENLP coder in this paper. Among them is the use of a perceptual weighting filter to compute a more meaningful measure of distortion (Atal, 1986). Another is the use of postfiltering to enhance the quality of reconstructed speech (Chen and Gersho, 1987).

In the predictive quantiser and the excitation codebook of the TCENLP coder, their elements

are fully searched. Structured quantisers and codebooks might greatly reduce the computational requirement (Wu and Fallside, 1991).

We are currently working on more formal performance evaluation of the TCENLP coder, comparing it to the US government 4.8 *kbps* standard (Campbell et al., 1991).

Acknowledgements

We wish to thank our colleagues in the Speech Laboratory of Cambridge University Engineering Department for participating in the subjective test of speech quality. C. Giguere and C. Seymour helped conduct this experiment. T. Robinson, C. Giguere and T.Burrows have prove-read the original manuscript. We would also like to thank the reviewers for their comments. This research is supported by the British Science Engineering Research Council.

5 Appendix: Output Variance of a Trained, Time-delayed, Recurrent Neural Network to an Input Perturbation

For a multi-layer perceptron with time-delayed feedback connections as described by

$$\begin{aligned} Y(t) &= f\left(WY(t-\tau)+VX(t)\right) \\ Z(t) &= UY(t), \end{aligned} \tag{30}$$

this appendix derives the output variance $D_z = \| Z - Z_0 \|^2$ to an input perturbation $D_x = \| X - X_0 \|^2$ with respect to its weight parameters U, V and W.

By defining a state vector $O(t) = WY(t-r) + VX(t)$ and using the mean value theorem [4]

$$f(O(t)) - f(O_0(t)) = f'(O^*(t))(O(t) - O_0(t)), \tag{31}$$

and using Schwarz's inequality (Korn and Korn, 1968)

$$\| f(O(t)) - f(O_0(t)) \| \leq \| f'(O^*(t)) \| \, \| O(t) - O_0(t) \|, \tag{32}$$

we find

$$\begin{aligned} D_z &= \| Uf(O(t)) - Uf(O_0(t)) \|^2 \\ &\leq \gamma^2 \| U \|^2 \| O(t) - O_0(t) \|^2 \end{aligned} \tag{33}$$

where $\gamma = \max \mid f'(O^*(t)) \mid$ and $O^*(t) \in (O(t), O_0(t))$.

The time-delay τ is usually small [5] , therefore the neural network satisfies the following dynamic function [6]

$$\tau \frac{dO(t)}{dt} = Wf(O(t)) - O(t) + VX(t). \tag{34}$$

If we let

$$D_o(t) = \| O(t) - O_0(t) \|^2, \tag{35}$$

[4] Assume that $f(\cdot)$ is continuous and continously differentiable.

[5] With this assumption,

$$\begin{aligned} \frac{O(t+\tau) - O(t)}{\tau} &\approx \frac{dO(t)}{dt} \\ X(t-\tau) &\approx X(t). \end{aligned}$$

[6] The variable t of $X(t)$ and $Z(t)$ indicates that the input and the output are time varying. The t in this dynamic function represents the settling time of the evolution-process of the network from initial state to target state. During the evolution-period, the input can be assumed as fixed since the updating rate of the network activation is usually much greater than the updating rate of the input.

then

$$\frac{dD_o(t)}{dt} = 2(O(t) - O_0(t))^T \left(\frac{dO(t)}{dt} - \frac{dO_0(t)}{dt} \right). \tag{36}$$

From Eqn(34), we find

$$\frac{dO(t)}{dt} - \frac{dO_0(t)}{dt} = \frac{1}{\tau}\{W[f(O(t)) - f(O_0(t))] - (O(t) - O_0(t)) + V(X(t) - X_0(t))\}. \tag{37}$$

Using the mean value theorem and Schwarz's inequality once again, we find

$$\begin{aligned} \frac{dD_o(t)}{dt} &= \frac{2}{\tau}\{\| O(t) - O_0(t) \|^T W[f(O(t)) - f(O_0(t))] \\ &\quad -\| O(t) - O_0(t) \|^2 + \| O(t) - O_0(t) \|^T V(X(t) - X_0(t))\} \\ &\leq \frac{2}{\tau}\left[\xi D_o(t) + \zeta\sqrt{D_o(t)}\right], \end{aligned} \tag{38}$$

where

$$\xi = \gamma \| W \| - 1, \tag{39}$$

$$\zeta = \sqrt{D_x} \| V \|. \tag{40}$$

If we let $D_o(t) = Q^2(t)$ and $Q(t) > 0$, then Eqn(38) becomes

$$\frac{dQ(t)}{dt} \leq \frac{1}{\tau}[\xi Q(t) + \zeta], \tag{41}$$

which yields

$$Q(t) \leq \left[Q(0) + \frac{\zeta}{\xi}\right] \exp\left(\frac{\xi}{\tau}t\right) - \frac{\zeta}{\xi} \tag{42}$$

or

$$D_o(t) \leq \left\{ \left[\sqrt{D_o(0)} + \frac{\zeta}{\xi}\right] \exp\left(\frac{\xi}{\tau}t\right) - \frac{\zeta}{\xi} \right\}^2. \tag{43}$$

If $O_0(t)$ is the unique equilibrium for the input X_0, the network satisfies (Atiya, 1987)

$$\gamma \| W \| < 1, \tag{44}$$

so $\xi < 0$ and

$$\sup_{t\to\infty} D_o(t) = \left(\frac{\zeta}{\xi}\right)^2. \tag{45}$$

By letting

$$\rho = \frac{\gamma \| U \| \| V \|}{\gamma \| W \| - 1}, \tag{46}$$

Eqn(33) becomes

$$D_z \leq \rho^2 D_x. \tag{47}$$

Therefore, after the network has settled, its output variance D_z to an input perturbation D_x can be approximated by Eqn(47).

References

Atal, B. (1986). High-quality speech at low bit rates: multi-pulse and stochastically excited linear predictive coders. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 1681–1684.

Atiya, A. (1987). Learning on a general network. In Anderson, D., editor, *Proceedings of Neural Information Processing Systems*, pages 22–30. American Institute of Physics, Denver.

Buzo, A., Gray.Jr, A., Gray, R., and Markel, J. (1980). Speech coding based upon vector quantization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(5):562–574.

Campbell, J., Tremain, T., and Welch, V. (1991). The DoD CELP 4.8 kbps standard (proposed federal standard 1016). In Atal, B., Cuperman, V., and Gersho, A., editors, *Advances in Speech Coding*. Kluwer Academic Publishers.

Chen, J. and Gersho, A. (1987). Real-time vector APC speech coding at 4800 bps with adaptive postfiltering. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 2185–2188.

Davidson, G. and Gersho, A. (1987). Real-time vector excitation coding of speech at 4800 bps. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 2189–2192.

Korn, G. and Korn, T., editors (1968). *Mathematical Handbook for Scientists and Engineers*. McGraw-Hill Book Company.

Kroon, P. and Deprettere, E. F. (1988). A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbits/s. *IEEE Journal on Selected Areas in Communications*, 6(2):353–363.

Lapedes, A. and Farber, R. (1987). Nonlinear signal processing using neural networks: prediction and system modelling. Technical Report LA-UR-87-2662, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

Lloyd, S. (1957). Least squares quantization in PCM. Bell Laboratory Technical Note; Published on IEEE Transactions on Information Theory, IT-28, 2, March 1982, pp129-137.

Max, J. (1960). Quantizing for minimum distortion. *IRE Transactions on Information Theory*, IT-6:7–12.

Natvig, J. (1988). Evaluation of six medium bit-rate coders for the Pan-European digital mobile radio system. *IEEE Journal on Selected Areas in Communications*, 6(2):324–331.

Rabiner, L. and Schafer, R., editors (1978). *Digital processing of speech signals*. Englewood Cliffs, Prentice-Hall.

Ramachandran, R. P. and Kabal, P. (1989). Pitch prediction filters in speech coding. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-37(4):467–478.

Schroeder, M. and Atal, B. (1985). Code-excited linear prediction (CELP): high-quality speech at very low bit rates. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 937–940.

Seneff, S. and Zue, V. W. (1988). Transcription and alignment of the TIMIT database. In Garofolo, J. S., editor, *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database*. National Institute of Standards and Technology (NIST), Gaithersburgh, MD.

Tishby, N. (1990). A dynamical systems approach to speech processing. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 365–368.

Townshend, B. (1991). Nonlinear prediction of speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 425–428.

Trancoso, I. and Atal, B. (1986). Efficient procedures for finding the optimum innovation in stochastic coders. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 2375–2378.

Wang, S., Paksoy, E., and Gersho, A. (1990). Performance of nonlinear prediction of speech. In *Proceedings of International Conference on Spoken Language Processing, Kobe, Japan*, pages 2.1.1–2.1.4.

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1:425–464.

Wu, L. and Fallside, F. (1991). On the design of connectionist vector quantizers. *Computer Speech and Language*, 5(3):207–230.

Wu, L. and Fallside, F. (1992). Source coding and vector quantization with codebook-excited neural networks. *Computer Speech and Language*, 6(3):243–276.

Xie, Y. and Jabri, M. (1992). Analysis of the effects of quantisation in multilayer neural networks using a statistical model. *IEEE Transactions on Neural Networks*, 3(2):334–338.