Fully Vector-Quantized Neural Network-Based Code-Excited Nonlinear Predictive Speech Coding

Lizhong Wu, Mahesan Niranjan, and Frank Fallside

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 exploit the 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 quantizers**
b) **the analysis of the output variance of the nonlinear predictive filter with respect to the input disturbance**
c) **the design of nonlinear predictive speech coders.**

The above studies have resulted in a fully vector-quantized, 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 in speech coding and the improvement in coding performance.

I. Introduction

In the past, most speech coding techniques have been 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 [1], [2].

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 [3]. This linear representation of the speech signal has been an important choice when no alternative is available. Linear simplification will doubtless lead to inaccurate representation, inefficient processing, and poor coding performance [4].

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 [5], [6].

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

1) *The development of a neural network-based nonlinear predictive model*:

   In Section II-A, 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 nonlinear predictors reported recently, e.g. [7], 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 [8].

2) *The design of a nonlinear predictive vector quantizer*:

   When applied to speech coding, the nonlinear predictive parameters should be quantized. Usually, a low transmitted rate cannot be achieved by directly quantizing the weight parameters of neural networks [9]. Instead, we train a finite set of nonlinear predictors to form a nonlinear predictive vector quantizer (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 II-B 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. A long-term nonlinear predictive model and the integration of short- and long-term predictors are studied in Section II-C of this paper.

4) *The tolerance of the nonlinear predictive filter to an excitation disturbance*:

   A fully vector-quantized predictive speech coder consists of two main parts: a predictive VQ and an excitation VQ. For each frame of speech, an excitation code vector 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 III-A of this paper.

Manuscript received March 24, 1992; February 18, 1994. This work was supported by the British Science Engineering Research Council. The associate editor coordinating the review of this paper and approving it for publication was Prof. Huseyin Abut.

L. Wu is with the Department of Computer Science and Engineering, Oregon Graduate Institute, Portland, OR 97291-1000 USA.

M. Niranjan is with the Engineering Department, Cambridge University, Cambridge, England.

F. Fallside, deceased, was with the Engineering Department, Cambridge University, Cambridge, England.

IEEE Log Number 9403978.

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. In another approach, instead of using a stochastic codebook, we train an excitation codebook directly with the analysis-by-synthesis method. We will concentrate on the second approach in this paper.

5) *The design of a fully vector-quantized, code-excited, nonlinear predictive speech coder*:

Unlike linear predictive coding, a closed-form solution 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 initialized by Gaussian samples and trained directly to minimize the error between the synthesized speech and the original one. This is described in Section III-B of this paper.

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

II. Neural Network-Based Nonlinear Prediction

Time series prediction can be defined in the following way: Given p previous observations of the signal $s(t)$, $X = (s(t-1), \ldots, s(t-p))^T$, find a function $g(\cdot)$ that minimizes the predictive residual

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

where $P(X, s)$ is the density function of the joint probability of X and s. The theoretical solution of (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 multilayer neural network, if the number of input units is p and there is only one output unit, the network can be trained as a pth-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 the posterior mean estimator $g(X)$. This is a problem of nonparametric estimation with neural networks. Several studies, e.g. [10], 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 modeling data without any specific prior assumption about the form of nonlinearity. Their advantages have been reported by a number of researchers, e.g. [7]. In this section, we study a general predictive model with a recurrent neural network and apply it to nonlinear predictive vector quantization and long-term (pitch) nonlinear speech prediction.

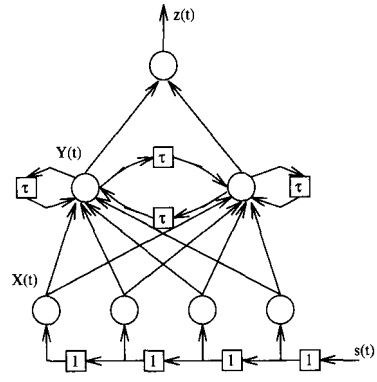


Fig. 1. Neural network-based nonlinear predictor. The small squares stand for time delay, and the numbers inside the squares represent their time delay units.

A. A Neural Model of Nonlinear Prediction

A general structure of neural network-based nonlinear predictor is shown in Fig. 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 can be described by the equations

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

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 (3), if $W = 0$, the network is of feedforward type. This form of predictor has been widely studied, e.g. [7]. 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

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

$$\nabla_V z(t) = U \nabla_V Y(t). \tag{5}$$

By defining matrix $A = W\nabla_\Phi Y(t-\tau)$ and 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 $1 \le n \le N_i$.

B. Nonlinear Predictive Vector Quantization

Like the linear predictive case [11], 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, where r (in bits) is the size of the quantizer. The performance of the predictive quantizers 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 quantizer is therefore expected to cope with the nonstationarity of the predicted signal and to improve the predictive performance over that of an individual predictor. During quantization, each frame of speech is successively applied to all the predictors in the quantizer. The predictor with the least predictive error, averaged over the whole frame, is then selected to quantize the current frame. The process can be described by the following 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 quantizer 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 the updating equation becomes

$$\Delta\Phi_k = \eta \sum_{\mathbf{S(t)} \in k} \sum_{t}^{N} \|s(t) - F(\Phi_k, X(t))\| \nabla_{\Phi_k} z_k(t) \tag{9}$$

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 and the zero-input responses. The memory effect of the previous selected predictor cannot be directly removed by subtracting the current frame speech with the zero-input response. The memory of the previous frame should be taken into account during the predictor selection for the current frame. The initial value $Y(t)$ for the current frame is therefore set to the last value $Y(t)$ of the previous frame.

C. 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 [12].

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 Fig. 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:[1]

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

where T is the pitch period, which is found from the short-time autocorrelation function of the short-term residual [3]:

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

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 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. Therefore, in (10)

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

for the cascade form, and

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

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

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

The short- and long-term predictors are trained in sequence. First, the short-term predictor is trained to minimize the short-term residual $r_s(t)$. With the short-time autocorrelation function of $r_s(t)$, the pitch period is then estimated [3]. Finally, the long-term predictor is trained to reduce the combined short- and long-term residual $r(t)$.

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

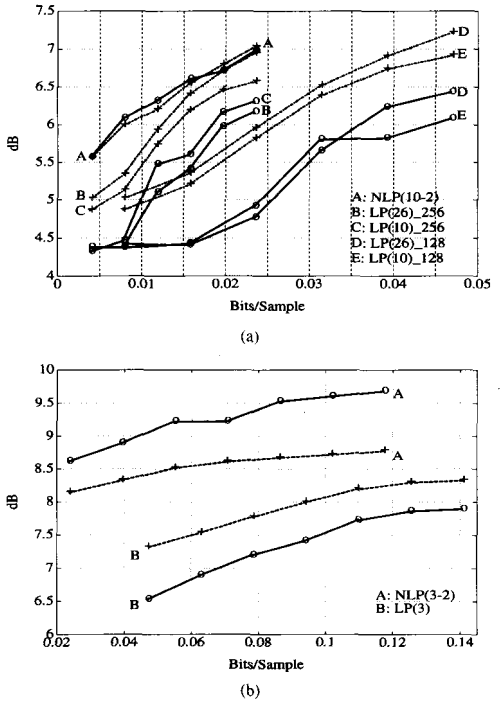


Fig. 2. Comparison of the predictive gain-rate functions of linear and nonlinear predictive VQ's: (a) VQ's with short-term prediction only; (b) VQ's with combined short- and long-term prediction in the cascade form. The training performance is connected by dashed curves, and the test performance is connected by solid ones. In (a), "-256" and "-128" represent the frame advance. The transmitted rate for coding the pitch information has not been included in this figure.

D. Nonlinear Predictive Quantization 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 quantization 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 pth-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.

The nonlinear predictive quantizer is evaluated with continuous speech data. The speech samples were from the TIMIT database [8] and were prefiltered to 8 kHz. The training set consists of ten different spoken sentences by ten different speakers (five females). The test set consists of four different spoken sentences by four speakers (two females) not included in the training set.

Fig. 2 compares the predictive gain-rate functions of the linear and nonlinear predictive VQ's. The prediction gain is measured by the total speech energy over the total predictive error (in decibels). The sizes of the VQ's vary from 1–6 b. 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 $\frac{r_s}{256}$ b/sample for the nonlinear predictive information.

The linear predictive VQ was designed using the approach developed in [13]. Each frame consists of 256 or 384 Hamming-windowed samples and is overlapped by 128 samples. Therefore, the transmitted rate is $\frac{r_s}{128}$ b/sample for the frame size of 256 and $\frac{r_s}{256}$ b/sample for the frame size of 384.

Fig. 2 also shows the predictive gain-rate functions of the combined short- and long-term VQ's. The sizes of long-term predictive VQ's are also from 1–6 b. The long-term predictive VQ is cascaded with a 6-b, 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 [3]. 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 [12] and [14]. The predictor parameters were vector quantized using a one-step identification/compression technique. A codebook of long-term predictor parameters is exhaustively searched to identify which predictor minimizes 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 $\frac{r_l}{64}$ b/sample, where r_l is the size of long-term predictive VQ in bits.

From inspection of Fig. 2, we conclude that all nonlinear predictive quantizers outperform linear ones either with the same predictive order or with equal numbers of weights. The test performance of linear predictive VQ's is 0.5–1 dB worse than the training performance. In contrast, with the same training and test data sets, Fig. 2 shows that the test performance of nonlinear predictive VQ is close to, or even better than, the training performance.[2]

III. Code-Excited Nonlinear Predictive Speech Coding

In a code-excited linear predictive speech coding (CELP) system [15], speech is synthesized by passing an excitation vector through a cascade of short- and long-term linear predictive filters. The excitation codebook usually performs the code vector search in an analysis-by-synthesis manner [2]. The excitation codebook may either be stochastic, i.e., pseudo-randomly populated, or predesigned over some training data.

In a code-excited nonlinear predictive speech coding system, the predictive filters are nonlinear. In this section, we investigate the problems caused by such a replacement and

[2]Both the training and test data sets are randomly chosen from TIMIT database. Compared with the training data set, the test data set contains more voiced sounds, which usually achieves better predictive gain than unvoiced speech does. Therefore, the test performance is better than the training performance.

study how to achieve better coding performance based on the capability of nonlinear predictive vector quantization.

A. 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 subsection, we will analyze the tolerance of nonlinear predictive filter to the 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 pth-order linear predictive coefficient vector. We easily find

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

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

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

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, $\gamma = \max \mid f'(O^*(t)) \mid$, $O^*(t) \in (O(t), O_0(t))$, 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 $\rho = \rho_1 + \rho_2$.

Equations (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

$$0 \le \gamma\|W\| < 1$$

and therefore

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

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.

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 minimize the sensitive parameters at the same time as minimizing the predictive error during the design of nonlinear VQ's
2) instead of using the stochastic codebook to use the trained excitation codevectors in the code-excited nonlinear predictive coder.

We will focus on the second approach in this paper.

B. Trained Code-Excited Nonlinear Predictive Coding

For a nonlinear predictive coder, the closed-form solution 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] Our predictive speech coder can be expressed as shown in the following:

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

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 quantizer only. For the case of a combined short- and long-term predictive quantizer, 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

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

then

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

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

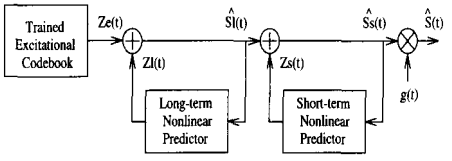


Fig. 3. Trained code-excited nonlinear predictive (TCENLP) speech coder.

C. 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 normalization step may still be desirable to smooth the outputs and reduce the output variarces 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 Fig. 3 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

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

This results in the coding error

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

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

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

D. Coding Performance

Fig. 4 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 II-D.

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

The excitation codebook was initialized 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 b, 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 b without loss of information. The gain value was quantized with a Max–Lloyd quantizer [16], [17]. Its design was based on the same training data set. We found that the coding performance can be improved by sequentially reoptimizing the excitation codebook and the gain codebook [18]. This was accomplished by recursive adjustment of the entries of each codebook. One codebook was fixed when the other was updated.

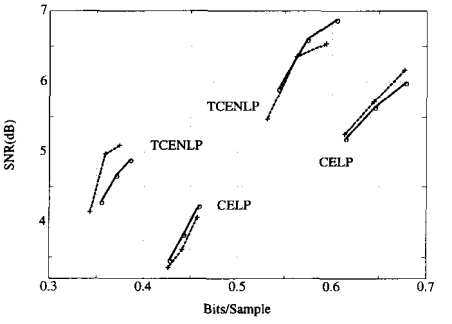


Fig. 4. Comparison of the coding performance of the TCENLP and CELP. The solid curves plot the test performance, and the dashed curves plot the training performance. There are two groups of curves for each coder: one corresponds to $N_v = 64$ and another to $N_v = 32$. The three points on each curve refer to B_t =7, 8, and 9 b. The bit allocation of other coding parametersis listed in Table I.

TABLE I
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_t IS THE SIZE OF THE EXCITATION CODEBOOK AND EQUALS 7, 8, OR 9 b. B_p IS THE CODE LENGTH OF PITCH-INFORMATION, WHICH IS 8 b IF THE PITCH PERIOD IS NOT EQUAL TO ZERO AND 1 b FOR ZERO PITCH PERIOD. N_v IS THE LENGTH OF EXCITATION VECTORS, WHICH IS SET TO 32 OR 64.

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	$\frac{B_t}{N_v}$
Pitch	$\frac{B_p}{64}$
Gain	$\frac{5}{N_v}$

The CELP speech coder design was based on [2], [14], and [19]. The short- and long-term linear predictive quantizers were described in Section II-D. The excitation codebook was formed by Gaussian samples and fully searched during coding.

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 II. 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, where each was 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 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. One subject said that the two kinds of reconstructed speech were not significantly different.

TABLE II
SMALL RATING OF OPINION SCORES OF SPEECH QUALITY AND LISTENING EFFORT [20]

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

IV. CONCLUDING REMARKS AND FURTHER WORK

A. 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 quantization 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 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. 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 points 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 quantization. Compared with the single mapping of linear prediction, the nonlinear predictor network needs two mappings: first from the inputs to the hidden units and then from the hidden units to the output. Because we use the network that 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-quantized TCENLP speech coder with neural networks.

B. Further Work

In conventional CELP speech coders, e.g., DoD CELP [21], 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 [1]. Another is the use of postfiltering to enhance the quality of reconstructed speech [22].

In the predictive quantizer and the excitation codebook of the TCENLP coder, their elements are fully searched. Structured quantizers and codebooks might greatly reduce the computational requirement [11].

A larger data set is required for more formal performance evaluation of the TCENLP coder.

APPENDIX
OUTPUT VARIANCE OF A RECURRENT NEURAL NETWORK TO AN INPUT PERTURBATION

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

$$Y(t) = f(WY(t-\tau) + VX(t))$$
$$Z(t) = UY(t) \tag{30}$$

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 $O(t) = WY(t-\tau) + VX(t)$ and using the mean value theorem and Schwarz's inequality [23], we find

$$D_z \le \gamma^2 \|U\|^2 \|O(t) - O_0(t)\|^2 \tag{31}$$

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

The time delay τ is usually small, and therefore, the neural network satisfies the following dynamic function:

$$\tau \frac{dO(t)}{dt} = Wf(O(t)) - O(t) + VX(t). \tag{32}$$

If we let $D_o(t) = \|O(t) - O_0(t)\|^2$, then

$$\frac{dD_o(t)}{dt} = 2(O(t) - O_0(t))^T \left(\frac{dO(t)}{dt} - \frac{dO_0(t)}{dt} \right). \tag{33}$$

Using (32), we find

$$\begin{aligned} \frac{dD_o(t)}{dt} &= \frac{2}{\tau} \Big\{ \|O(t) - O_0(t)\|^T W[f(O(t)) - f(O_0(t))] \\ &\quad - \|O(t) - O_0(t)\|^2 \\ &\quad + \|O(t) - O_0(t)\|^T V(X(t) - X_0(t)) \Big\} \\ &\le \frac{2}{\tau} \left[\xi D_o(t) + \zeta \sqrt{D_o(t)} \right] \end{aligned} \tag{34}$$

or

$$D_o(t) \le \left\{ \left[\sqrt{D_o(0)} + \frac{\zeta}{\xi} \right] \exp\left(\frac{\xi}{\tau} t \right) - \frac{\zeta}{\xi} \right\}^2 \tag{35}$$

where

$$\xi = \gamma \|W\| - 1, \tag{36}$$
$$\zeta = \sqrt{D_x} \|V\|. \tag{37}$$

If $O_0(t)$ is the unique equilibrium for the input X_0, the network satisfies [24]

$$\gamma \|W\| < 1 \tag{38}$$

and therefore, $\xi < 0$ and

$$\sup_{t \to \infty} D_o(t) = \left(\frac{\zeta}{\xi} \right)^2. \tag{39}$$

By letting

$$\rho = \frac{\gamma\|U\|\|V\|}{\gamma\|W\| - 1} \tag{40}$$

(31) becomes

$$D_z \leq \rho^2 D_x. \tag{41}$$

Therefore, after the network has settled, its output variance D_z to an input perturbation D_x can be approximated by (41).

ACKNOWLEDGMENT

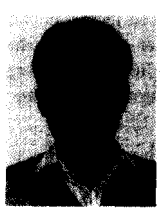
The authors wish to dedicate this paper to the memory of their associate, Prof. F. Fallside. They also thank their colleagues in the Speech Laboratory of Cambridge University Engineering Department for participating in the subjective test of speech quality. C. Giguere and C. Seymour helped conduct this experiment. T. Robinson, C. Giguere, and Y. Yan have proofread the manuscript. We would also like to thank the reviewers for their comments.

REFERENCES

[1] B. S. Atal, "High-quality speech at low bit rates: Multi-pulse and stochastically excited linear predictive coders," in *Proc. Int. Conf. Acoust. Speech Signal Processing*, 1986, pp. 1681–1684.
[2] P. Kroon and E. F. Deprettere, "A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 Kbits/s," *IEEE J. Selected Areas Commun.*, vol. 6, no. 2, pp. 353–363, Feb. 1988.
[3] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice–Hall, 1978.
[4] S. Wang, E. Paksoy, and A. Gersho, "Performance of nonlinear prediction of speech," in *Proc. Int. Conf. Spoken Language Processing* (Kobe, Japan), Nov. 1990, pp. 2.1.1–2.1.4.
[5] N. Tishby, "A dynamical systems approach to speech processing," in *Proc. Int. Conf. Acoust., Speech Signal Processing*, 1990, pp. 365–368.
[6] B. Townshend, "Nonlinear prediction of speech," in *Proc. Int. Conference on Acoust., Speech Signal Processing*, 1991, pp. 425–428.
[7] A. Lapedes and R. Farber, "Nonlinear signal processing using neural networks: Prediction and system modeling," Tech. Rep. LA-UR-87-2662, Los Alamos Nat. Lab., Los Alamos, NM, June 1987.
[8] S. Seneff and V. W. Zue, "Transcription and alignment of the TIMIT database," in *Getting Started With the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database* (J. S. Garofolo Ed.). Gaithersburgh, MD: Nat. Inst. of Standards Technol. (NIST), 1988.
[9] Y. Xie and M. A. Jabri, "Analysis of the effects of quantization in multilayer neural networks using a statistical model," *IEEE Trans. Neural Networks*, vol. 3, no. 2, pp. 334–338, Mar. 1992.
[10] H. White, " Learning in artificial neural networks: A statistical perspective," *Neural Comput.*, vol. 1, pp. 425–464, 1989.
[11] L. Z. Wu and F. Fallside, "On the design of connectionist vector quantizers," *Comput. Speech Language*, vol. 5, no. 3, pp. 207–230, July 1991.
[12] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 37, no. 4, pp. 467–478, Apr. 1989.
[13] A. Buzo, A. H. Gray, Jr, R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech Signal Processing*, vol. ASSP-28, no. 5, pp. 562–574, Oct. 1980.
[14] G. Davidson and A. Gersho, "Real-time vector excitation coding of speech at 4800 bps," in *Proc. Int. Conf. Acoust., Speech Signal Processing*, 1987, pp. 2189–2192.
[15] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. Int. Conf. Acoustics, Speech Signal Processing*, 1985, pp. 937–940.
[16] S. P. Lloyd, "Least squares quantization in PCM," Bell Lab. Tech. Note, 1957; published in *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[17] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar 1960.
[18] L. Z. Wu and F. Fallside, "Source coding and vector quantization with codebook-excited neural networks," *Comput. Speech Language*, vol. 6, no. 3, pp. 243–276, July 1992.
[19] I. M. Trancoso and B. S. Atal, "Efficient procedures for finding the optimum innovation in stochastic coders," in *Proc. Int. Conf. Acoust., Speech Signal Processing*, 1986, pp. 2375–2378.
[20] J. E. Natvig, "Evaluation of six medium bit-rate coders for the Pan-European digital mobile radio system," *IEEE J. Selected Areas Commun.*, vol. 6, no. 2, pp. 324–331, Feb. 1988.
[21] J. P. Campbell, T. E. Tremain, and V. C. Welch, "The DoD CELP 4.8 kbps standard (Proposed Federal Standard 1016)," in *Advances in Speech Coding* (B. S. Atal, V. Cuperman, and A. Gersho, Eds.). Boston: Kluwer, 1991.
[22] J. H. Chen and A. Gersho, "Real-time vector APC speech coding at 4800 bps with adaptive postfiltering," in *Proc. Int. Conf. Acoustics, Speech Signal Processing*, 1987, pp. 2185–2188.
[23] G. A. Korn and T. M. Korn, *Mathematical Handbook for Scientists and Engineers*. New York: McGraw-Hill, 1968.
[24] A. F. Atiya, "Learning on a general network," in *Proc. Neural Inform. Processing Syst.* (D. Z. Anderson Ed.). New York: Amer. Inst. of Phys., 1987, pp. 22–30.

Lizhong Wu was born in China on September 30, 1963. He received the B.Sc. and M.Sc. degrees from South China University of Technology in 1983 and 1986 and the Ph.D. degree from Cambridge University Engineering Department, Cambridge, England, in 1992.

He was with Cambridge University Engineering Department as a research associate and has been with the Department of Computer Science and Engineering at Oregon Graduate Institute since November 1993. His research interests include signal processing, information and communications theory, time series analysis, and prediction.



Mahesan Niranjan was born in Sri Lanka on November 28, 1959. He received the B.Sc. degree from the University of Peradeniya, Sri Lanka, in 1982, the M.E.E. degree from the Netherlands Universities Foundation, Eindhoven, in 1985, both in electronic engineering, and the Ph.D. degree from the University of Cambridge in 1990.

He is currently lecturer in information engineering at the University of Cambridge. His research interests include speech processing, time series analysis, and neural computing.

Frank Fallside was professor and head of the Information Engineering Division at Cambridge University. He received the B.Sc. degree from Edinburgh University and the Ph.D. degree from the University of Wales. He died in March 1993.

The Information Engineering Division at Cambridge University covers the computer-related aspects of engineering including computer science, communications, and control. He had built up a research group specializing in neural networks since 1984. The group had worked with most of the common neural net architectures on theoretical and training aspects and with applications in speech, vision, robotics, control, radar, and medical data.

Dr. Fallside was a member of Institute of Electrical Engineers, joint editor of the Journal *Computer Speech and Language*, and subeditor of the IEEE TRANSACTIONS ON NEURAL NETWORKS.