SPEECH CODING WITH NONLINEAR LOCAL PREDICTION MODEL

Ni Ma and Gang Wei

Department of Electronics and Communication Engineering,
South China University of Technology, Guang Zhou, 510641 P.R.China

ABSTRACT

A new signal process based on a nonlinear local prediction model(NLLP) is presented and applied to speech coding. With the same implemention, the speech coding based on the NLLP gives improved performance compared to reference versions of the standard ITU-T G.728 and linear local scheme. The computational efforts for the NLLP analysis does not increase over the conventional linear prediction(LP), and the NLLP supplies better prediction performance over the LP and linear local prediction.

1. INTRODUCTION

It has recently been proved that the state space based local prediction model is a better singal predictor[2][7]. In speech coding, the linear local modeling, which is developed from the useful linear prediction coding(LPC) technique with all pole autogressive(AR) model, gives improved performance over comparative linear model[6]. The effective strategy for the nonlinear speech modeling of this case involves fitting an AR model to the signal locally in a state space, that is, the model parameters vary as a function of the state. This nonlinear model can be viewed as a problem of interpolating from the noisy samples, therefore the accurate model is acquired by some linear interpolating functions.

However, from the approximation viewpoint, the nonlinear interpolating functions are capable of obtaining more efficacious outcomes for the nonlinear speech signal. Furthermore, some nonlinear functions, e.g., radial biasis function, provide regularized solutions, and then they can make the number of modeling parameters fairly low and guarantee the stability of the corresponding synthesis scheme[3]. For the computational efforts, the method supplied by [6] is a useful way to reduce the complexity of the linear local model and can also be used in the nonlinear local model. The other superiority of the nonlinear function is that the total compute amount is able to be reduced by cutting down the number of the modeling parameters.

In this paper, the backwardly adaptive technique is used in speech coding with a nonlinear local model and additional computational efforts of the pattern matching is decreased by a little number of the model parameters, as is distinguished from [3], where the nonlinear function was used as a global model and the predictor adaptation had been performed in a forward way.

2. PREDICTION OF NLAR PROCESS IN STATE SPACE

Let Θ and Ψ be the maps in state space $\Re^n$, a broad class of system, including AR model and other generalizations of the AR model, can be represented in a common state space form[2]:

$$\mathbf{x}_{k+1} = \Theta(\mathbf{x}_k, \mathbf{u}_k, k) \quad (1)$$

$$\mathbf{y}_k = \Psi(\mathbf{x}_k, \mathbf{u}_k, k) \quad (2)$$

where the $n \times 1$ vector $\mathbf{x}_k$ is the state, the $p \times 1$ vector $\mathbf{u}_k$ is the input, and the $m \times 1$ vector $\mathbf{y}_k$ is the output. Generalizing the model to include nonlinear system, while retaining the companion state variable structure, leads to systems described by a nth order nonlinear difference equation of the form:

$$y_{k+1} = F(y_k, y_{k-1}, \cdots, y_{k-n+1}) + u_k, \quad (3)$$

where $F(\cdot)$ maps $\Re^n$ to $\Re$, and u_k is stational white noise. We refer to the process (3) as a nonlinear autogressive process(NLAR).

It is clear from (1–3), that the state vector $\mathbf{x}_k$ can be reconstructed from the observations of the scalar output y_k,

$$\mathbf{x}_k = (y_{k-n+1}, \cdots, y_{k-1}, y_k)^T. \quad (4)$$

Thus the minimum mean square error(MMSE) estimate of y_{k+1} given its entire signal history is:

$$\hat{y}_{k+1} = F(\mathbf{x}_k). \quad (5)$$

Although $F(\mathbf{x})$ is a part of the system model, and therefore unavailable, the state dynamic of the system

can be observed through

$$y_{k+1} = F(\mathbf{x}_k) + u_k. \tag{6}$$

Thus given y_k and recovering $\mathbf{x}_k$ from (4), the signal history represents a set of noisy samples of $F(\mathbf{x})$, nonuniformly distributed in state space. Consequently, the estimation problem for y_{k+1} can be regarded as a solution of interpolating $F(\mathbf{x})$ from white noisy samples.

Based on the interploating viewpoint, the kernel-based strategies involving splines or radial basis functions can be used to create a global approximation of $F(\mathbf{x})$[2]. One benefit of such a scheme is that the model for $F(\mathbf{x})$ can be precomputed, making signal prediction a simple function evalution. However the performance of this scheme depends critically on the choice of the kernel, since rather strong assumptions are imposed on $F(\mathbf{x})$ between the observations, and the global approximation requires great amount computation and intensive convergence. Hence a philosophical approach which makes fewer assumptions about the behavior of the function between the samples and has little computation is using the local models. In a manner reminiscent of vector quantization, we can view the signal as a codebook of pairs(state-vector, signal-value) of the form$(\mathbf{x}_k, y_{k+1})$. Because each codebook entry must satisfy (6), the prediction strategy is to use the present state of the system $\mathbf{x}_k$ to "look up" $F(\mathbf{x}_k)$ in the codebook. Thus The method for predicting y_{k+1}, given $y_i, 0 \leq i \leq k$ is[2][6]: (1) Form a codebook of pairs $(\mathbf{x}_i, y_{i+1})$ from the signal history, (2) Select pairs $(\mathbf{x}_i, y_{i+1})$ from the codebook for $\mathbf{x}_i$ neighbouring on $\mathbf{x}_k$, (3) Fit a local model $y_{i+1} \approx \hat{F}(\mathbf{x}_k)$ to the selected pairs, (4) Apply the local model to obtain $\hat{y}_{i+1} = \hat{F}(\mathbf{x}_k)$.

3. SELECTION OF LOCAL MODEL

$F(\mathbf{x})$ were approximated as the linear functions near $\mathbf{x}_k$ by[2][6][7], as resulted in a linear local prediction model(LLP), which can be shown to be a generalization of the AR process and had good approximations. Unfortunately the linear estimation solutions for $F(\mathbf{x})$ sometimes exhibit unstable behaviors due to the problem of the singularities. Thus the singular value decomposition or other techniques has to be employed, which increases extensive computation.

Instead, if a nonlinear function is selected as the basis function, the local model becomes the university of the NLAR model. Radial basis function(RBF) has been reported to be universial approximation capability and a regularization form. Specially, the nonlinear local prediction model(NLLP) retains inherent advantages of the RBF due to its nonlinear nature.

The NLAR model based on RBF can be expressed by

$$y_{i+1} = \lambda_0 + \sum_{i=1}^{m} \lambda_i g_i(\| \mathbf{x}_i - c_i \|) + u_i, \tag{7}$$

where $\{g_i\}$ are the RBF, $\| \cdot \|$ is a norm(e.g., L_2-norm) in $\Re^n$, $\{c_i\}$ are the RBF centers, $\{\lambda_i, i = 0, 1, \cdots, m\}$ are the weights of the linear combination and m is the number of the RBF. It can be easily verified that Gaussian RBF $g_i(x) = exp(-\frac{x^2}{\sigma^2})$, σ^2 being the variance associated to each RBF, makes the synthesis system with this NLAR stable. Hence the evaluation y_{i+1} is:

$$\hat{y}_{i+1} = \lambda_0 + \sum_{i=1}^{m} \lambda_i g_i(\| \mathbf{x}_i - c_i \|), \tag{8}$$

For the RBF, the estimation accuracy is crucially governed by the number and position of the centers. In order to obtain the trade-off between the computational efforts and prediction precision, the orthogonal least squares(OLS) learning algorithm[5], which is a simple and efficient means for fitting RBF networks, is used to retrieve a small number of data points as the centers. For a special purpose, the OLS algorithm has the property that each selected center maximizes the increment to the explained variance or energy of the desired output and suffers little numerical ill-conditioning problems[4].

4. SPEECH CODING WITH NLLP

Like the LLP, the NLLP application to speech coding has been impeded by two major obstacles: i) the quantization of the predictor's parameters, and ii) the prohibitive computational efforts. Here the first problem is solved by applying it to a predictor backwardly adaptive speech coding algorithm, i.e., ITU-T G.728 LD-CELP[1]. As for the second point, the use of a local model and relevant little number of centers reduce the computational complexity.

The standard coder does not use long-term predictor and the short-term predictor order is increased to 50 to compensate for the loss in speech quality. Since a local predictor is optimized over neighborhood vectors that are close to the "target" vector $\mathbf{x}_k$ in the state space, which also includes those vectors which are approximately an integral number of pitch period away, it has the ability to model long-term or pitch period correlations as well. Therefore the local model coding scheme need not long-term predictor either.

The designed nonlinear local prediction speech coding scheme is shown in Fig.1, which is little different from the LD-CELP except for the backwardly adaptive LP changing into the backwardly adaptive NLLP and

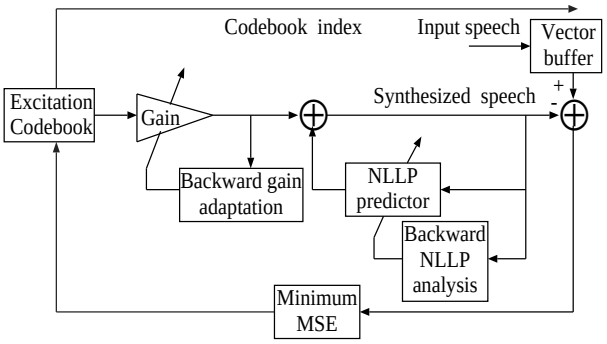

Figure 1: LD-CELP based on nonlinear local prediction model

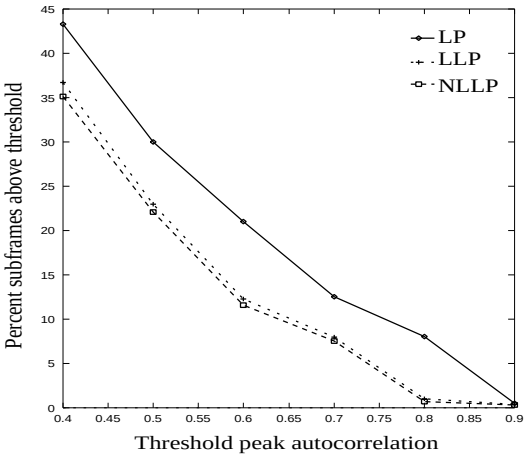

Figure 2: Comparisons of pitch period correlations

lacking perceptual weighting filters. The predictor's coefficients are adapted by performing NLLP analysis on the previously quantized speech, as is the same as the standard scheme. That is, the decoded speech(every subframe's length $N_s = 5$) constitutes analysis frame $y_i, i = k - L_f - 1, \cdots, k$, where L_f is the frame length. The state vectors $\mathbf{x}_i, i = k - L_k - 1, \cdots, k; L_k = L_f - n + 1$, are formed from y_i, and the $N_k (N_k > n)$ nearest neighbours of $\mathbf{x}_k$ from $\mathbf{x}_i, i = k - L_k - 1, \cdots, k - 1$, are selected to compose N_k pairs $(\mathbf{x}_j, y_{j+1}), j = 1, \cdots, N_k$, with which, the parameters in (8) can be achieved by the OLS algorithm. In coding process, the fitted local predictor is used to predict next subframe $\hat{y}_i, i = k+1, \cdots, k+N_s$, instead of only $\hat{y}_{k+1}$ in order to reduce the computational complexity while prediction gain decrease little due to small N_s.

For this NLLP analysis being comparable to the LP analysis, the number of the RBF centers m is chosen as 4 and the state vector's dimension n is 10, making the total number of the parameter 50. As proposed in [6], the analysis buffer parameters $L_f = 120$ and $N_k = 60$ are to reach acceptable computational efforts and coding accuracy. Since the statistics of the NLLP is different from that of the LP, a trained excitation codebook designed using closed-loop analysis[1] is substituted for that of the LD-CELP. As to the transmitted bit rate and algortihmic buffering delay are the same as those in the LD-CELP, which gives its low delay property and 16kbps channel rate.

5. PERFORMANCE COMPARISIONS AND CONCLUSION

5.1. Prediction Performance

As an effictive predictor, the NLLP should give improved performance, that is, it can provide better prediction gain and remarkably "whiter" residual. The one-step recursive prediction residuals and corresponding gains obtained in three cases(backward LP, LLP, NLLP)with the same number of the coefficients for one frame(30ms) speech sampled at 8kHz with 16b/sampe accuracy(all speech data used in this paper are gotton by this means) are shown in Fig.3 as an illustrative example, where the LP is with Hamming window, both the LLP and the NLLP are based on the identical analysis frame style explained in Section 4, and the LLP analysis adopts a weighted cost function way[6]. Obviously the NLLP gives the best result.

Fig.2 compares plots of the relative number of segments(of length 160) of prediction residuals of three backwardly prediction schemes that have peak normalized autocorrelation value(for lags between 20–140, and the analyzed speech is a segment of 48 seconds data comprising of ten males and ten females) greater than different threshold values, as is a example to show that the local short-term prediction is capable of modeling long-term correlation. This method is introduced by [6] to illustrate the LLP's capability modeling long-term dependency. The results shows the NLLP scheme has more accuracy.

5.2. Coding Performance

Because the perceptual weighting to the nonlinear prediction filter need studying further, a slightly modified version of the G.728 LD-CELP is done to make the comparisons more meaningful. For example, the perceptual weighting and post filtering in the LD-CELP are removed, decreasing the signal to noise ratios(SNRs) of the coding to a small extent.

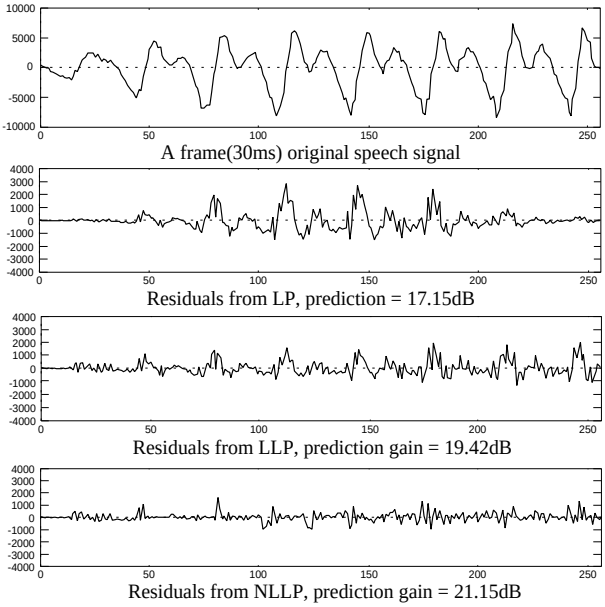

Figure 3: Comparisons of 3 cases' prediction using a frame speech

The results of reconstructed speech waveform and SNRs with the same frame speech for three schemes are presented in Fig.4, where the backwardly LLP coding scheme is based on [6]. The results clealy show that the reconstructed speech using the proposed approach provides the best approximation to the actual speech signal.

Using the continuous 48s speech to compare coding performance, the same conclusion can be obtained that the SNR of the backward NLLP is 11.23dB, which is an improvement of 0.4dB over the LLP and 0.7dB over the LP. Meanwhile, during the coding procedure, the ill-posed occured in the NLLP is three times, less than that in the LLP(eight times), which make the NLLP scheme have a better performance as well.

5.3. Conclusion

Speech signal has powerful nonlinearities and "local" properties, hence the NLLP based on the state space will be a more fine speech model. The practice of applying it to the speech coding shows that alternative versions of state based local prediction suited for lower rate speech coding may have a significant impact in future speech coding algorithm.

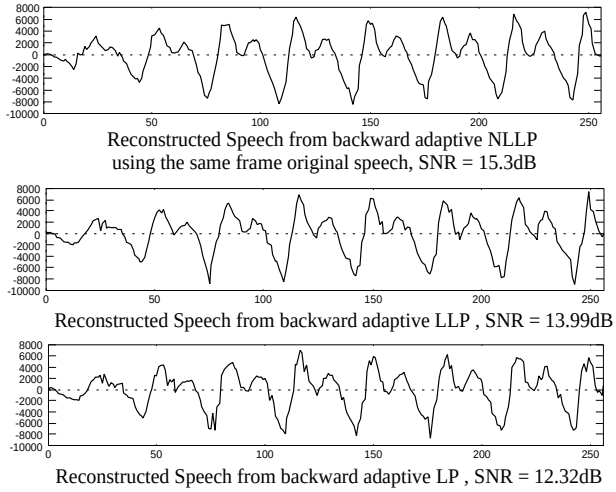

Figure 4: Comparisons of reconstruction performance with 3 coding schemes using the same frame speech

6. REFERENCES

[1] CCITT, "Coding of speech at 16kbit/s using low-delay code excited linear prediction recommendation G.728," Int. Telcommun. Union, Geneva, Switzerland, Spt. 1992.

[2] A. C. Singer, G. W. Wornell and A. V. Oppenheim, "Codebook prediction: a nonlinear signal modeling paradigm," in *Proc. ICASSP'92*, 1992, pp.V-325-328.

[3] D. M. Fernando and A. R. F. Vidal, "Nonlinear prediction for speech coding using radial basis functions," in *Proc. ICASSP'95*, 1995, pp.788-791.

[4] Y. Liguni, I. Kawamoto and N. Adachi, "A nonlinear adaptive estimation method based on local approximation," *IEEE Transactions on Signal Processing*, Vol.45, No.7, July 1997, pp.1831-1841.

[5] S. Chen, C. F. N. Cown and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function metworks," *IEEE Transactions on Neural Networks*, Vol.2, No.2, March 1991, pp.302-309.

[6] A. Kumar and A. Gersho, "LD-CELP speech coding with nonlinear prediction," *IEEE Signal Processing Letters*, Vol.4, No.4, April 1997, pp.89-91.

[7] B. Townshend, "Nonlinear prediction of speech," in *Proc. ICASSP'91*, 1991, pp.425-428.