LD-CELP Speech Coding with Nonlinear Prediction

Arun Kumar and Allen Gersho, *Fellow, IEEE*

***Abstract*— A technique for nonlinear prediction of speech via *local linear prediction* (LLP) is presented and applied to LD-CELP at 16 kbps. With 18th-order backward adaptive LLP for voiced frames, the hybrid LD-CELP coder gives higher segmental signal-to-noise ratio (SNR) compared to a reference version of the ITU-T G.728 LD-CELP algorithm, which has a 50th-order backward adaptive linear predictor. The computational complexity for LLP analysis is significantly less than that of a conventional one-step recursive LLP, and the LLP method gives better prediction gain and a remarkably "whiter" residual compared to backward adaptive linear predictor. With an appropriate state space neighborhood for *local linear* analysis, the short-delay predictor is also able to effectively model long-term correlations without requiring pitch estimation.**

I. Introduction

IT HAS BEEN shown that a one-step recursive local linear predictor (LLP) of speech, which is effectively a nonlinear predictor, gives improved performance over comparative linear prediction (LP) in terms of prediction gain and "whiter" residuals [1]–[3]. However, its application to CELP speech coding has been impeded by two major obstacles: i) the prohibitive computational complexity of local predictors and ii) the excessive bit rate required for the transmission of predictor parameters. Here we solve the first problem with a modified version of LLP and the second problem by applying it to a backwardly adaptive CELP coding algorithm. The modified LLP can be represented as a linear filter structure that allows us to use standard complexity reduction measures of CELP. We will outline the structure of a hybrid coder at 16 kbps, which uses LLP for voiced frames, and assess its performance relative to an appropriately modified version of the ITU-T G.728 LD-CELP coder.

II. Local Linear Prediction in State Space

Unlike LP analysis, which optimizes parameters over a contiguous time frame of data, the method of *local prediction* optimizes the predictor over *local* volumes in state space [1]. Consider an observed scalar time series $x_i, i = 0, 1, \cdots$. For one-step LLP, the problem is to predict $\hat{x}_n$ based on an *analysis frame* $x_i, i = n - N_f, \cdots, n-1$, where N_f is the analysis frame length. The first step is to reconstruct a vector time series $\mathbf{x}_i, i = n - \hat{N}_f, \cdots, n-1$; $\hat{N}_f = N_f - d + 1$ in d-dimensional state space. We choose the state variables as a vector of contiguous observables

$$\mathbf{x}_i = [x_{i-d+1}, x_{i-d+2}, \cdots, x_{i-1}, x_i]^T. \tag{1}$$

A one-step local predictor can be expressed as

$$\hat{\mathbf{x}}_n = \mathbf{g}(\mathbf{x}_{n-1}) \tag{2}$$

where $\mathbf{g}$: $\mathcal{R}^d \rightarrow \mathcal{R}^d$ is a state based local predictor. Specifically, for the present work

$$\mathbf{g}(\mathbf{x}_{n-1}) = [x_{n-d+1}, x_{n-d+2}, \cdots, x_{n-1}, \hat{x}_n] \tag{3}$$

and for the case of local *linear* prediction

$$\hat{x}_n = f(\mathbf{x}_{n-1}) = [a_1, \cdots, a_d]\mathbf{x}_{n-1}. \tag{4}$$

To estimate the coefficients $a_1, \cdots, a_d$, we select N_l ($N_l > d$) *nearest* neighbors of $\mathbf{x}_{n-1}$ from $\mathbf{x}_i$, $i = n-\hat{N}_f, \cdots, n-2$, and form N_l pairs $\{\mathbf{x}_{k_j}, x_{k_j+1}\}\, j = 1, \cdots, N_l$. Next we perform a weighted minimization of E with respect to $a_1, \cdots, a_d$ where

$$E = \frac{\sum_{j=1}^{N_l} w_{k_j}^2 [x_{k_j+1} - f(\mathbf{x}_{k_j})]^2}{\sum_{j=1}^{N_l} w_{k_j}^2} \tag{5}$$

and w_{k_j} is an appropriate weighting factor. For example

$$w_{k_j} = [||\mathbf{x}_{n-1} - \mathbf{x}_{k_j}||]^{-1}, \qquad j = 1, \cdots, N_l \tag{6}$$

where $\|\cdot\|$ is the L_2-norm in d-dimensional state space. A local linear predictor can be shown to be a generalization of the usual linear predictor [1]. The results of a detailed study of the prediction performance of one-step LLP as a function of model parameters N_f, N_l, and d and comparison with short-term LP and short-term plus long-term LP are given in [3].

A major problem that prevents the use of one-step LLP in analysis-by-synthesis coding is the prohibitive computational complexity. We overcome this problem by reducing the update rate of the predictor to $N_s > 1$ samples. Thus, a frame of data, x_i, for $i = n - \hat{N}_f, \cdots, n-1$ is analyzed to derive an LLP that predicts $\hat{x}_i$, $i = n, \cdots, n + N_s - 1$. In this case, LLP analysis based on a neighborhood of $\mathbf{x}_{n+[N_s/2]}$ instead of $\mathbf{x}_{n-1}$ will generally provide better prediction. (Here, $[x]$ denotes the greatest integer less than or equal to x). This is because a local neighborhood of $\mathbf{x}_{n-1}$ will be optimal only for predicting x_n. To use a fixed predictor for $i = n, \cdots, n + N_s - 1$, a local neighborhood centered around the middle sample $\mathbf{x}_{n+[N_s/2]}$ can give improved prediction gain. However, $\mathbf{x}_{n+[N_s/2]}$ is not available for LLP analysis

Manuscript received February 22, 1996. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. V. Viswanathan.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA.

Publisher Item Identifier S 1070-9908(97)02516-9.

because the predictor is to be designed as backward adaptive. An intuitive strategy that markedly improves prediction performance is to find the neighborhood of $\mathbf{x}_{n-1}$ as before and shift the corresponding pairs by $[N_s/2]+1$ samples. Thus, after determining the N_l nearest neighbors of $\mathbf{x}_{n-1}$ in state space and their corresponding indices $j = 1, \cdots, N_l$, "lookahead" pairs $\{\mathbf{x}_{k_j+[N_s/2]+1}, x_{k_j+[N_s/2]+2}\}$ are obtained for LLP analysis.

We compared the prediction gain of the modified LLP with backward adaptive LP for voiced speech. The voiced/unvoiced decision is made from previously buffered data, $x_i, i = n-160, \cdots, n-1$, using a combination of tests based on the peak of the autocorrelation function and signal energy. For LP, we used a hybrid window and frame size $N_f = 105$ as in LD-CELP [5], [6].

For the update rate $N_s = 5$, the segmental SNR (segSNR) over voiced segments increases from 12.52 to 13.48 dB as the LP order is increased from 10 to 50. In comparison, an 18th-order LLP (i.e., $d = 18$) with N_f =120, $N_l = 60$ gives segSNR of 13.72 dB. A one-step 18th-order LLP ($N_s = 1$) performs better than a one-step backward adaptive LP by 0.65 dB in voiced segments. All quantitative results reported in this letter are based on tests with 46.5 s of speech comprising of eight male and eight female sentences sampled at 8 kHz with 16 b/sample accuracy.

Since a local predictor is optimized over neighborhood vectors that are close to the "target" vector $\mathbf{x}_{n-1}$ in state space, which also includes those vectors which are approximately an integral number of pitch periods away, it has the ability to model long-term or pitch period correlations as well. An 18th-order LLP adapted every five samples (i.e., $N_s = 5$) is significantly more capable in removing long-term correlations compared to a 50th-order backward adaptive LP (model parameters as given above, for both cases). This can be seen from Fig. 1, which compares plots of the relative number of segments (of length 160) of prediction residuals of an 18th-order LLP and various backward adaptive LP that have peak normalized autocorrelation value (for lags between 20–140) greater than different threshold values. The short-term nonlinear predictor proposed in [4] is also capable of removing harmonics of the pitch frequency in the residual spectrum.

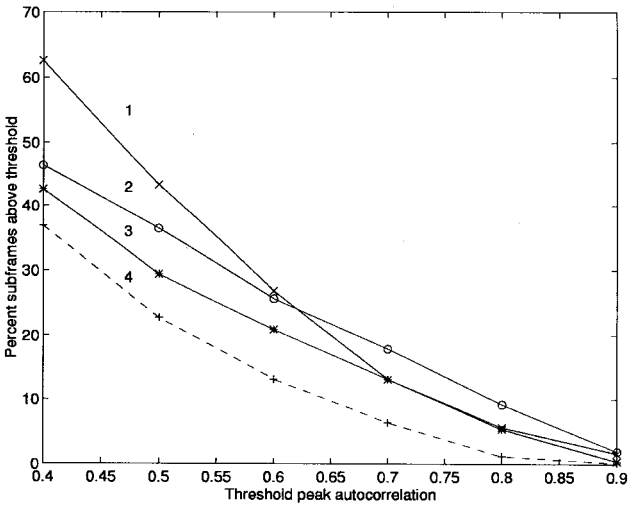

Fig. 1. Comparison of pitch period correlations in various prediction residual sequences through plots of relative number of segments (of length 160) having peak normalized autocorrelation (for lags between 20–140) above different threshold values. The corresponding residual sequences are: 1) 50th-order LP backward adapted every frame of 20 samples; 2) 18th-order LP backward adapted every subframe of five samples; 3) 50th-order LP backward adapted every subframe, and 4) 18th-order LLP backward adapted every subframe.

III. Low Delay Hybrid Codec at 16 kbps

We have designed and studied a hybrid codec at 16 kbps. The basic structure of the codec is the same as that of LD-CELP [5], [6], and we adhere to its nomenclature here. The main distinguishing features of the hybrid codec are as follows.

1) A voiced/unvoiced decision is made every subframe (length $N_s = 5$) based on immediately previous *decoded* speech. In case of voiced decision, a LLP ($N_f = 120$, $N_l = 60$, $d = 18$) is used for the subframe, otherwise a backward adaptive LP, as in LD-CELP but of 30th order is used.
2) The covariance method is used for LLP analysis. A "white noise correction" factor of 257/256 is applied to the diagonal terms of the covariance matrix to improve filter stability.
3) Since the statistics of the LLP residual is different from that of the LP residual, a trained excitation codebook designed using closed-loop analysis, is used for voiced subframes. For other subframes, the LD-CELP excitation codebook is used.

The algorithmic buffering delay is five samples, which is the same as in LD-CELP. This gives the codec its low delay property.

IV. Performance Comparison and Conclusion

To objectively compare the performance of the hybrid codec with LD-CELP using only standard LP, a slightly modified version of the G.728 LD-CELP was needed. Specifically, the updated LP coefficients are made available immediately for the next subframe instead of the two subframe delay. Also, the LP is adapted every subframe instead of a frame and the order is varied from 10–50. These two modifications improve the segSNR performance compared to standard LD-CELP and make comparison tests with the hybrid codec more meaningful. As the LP order of the modified LD-CELP is increased from 10 to 50, the segSNR over voiced subframes increases from 18.53 dB to 19.39 dB. In comparison, the hybrid codec that uses 18th-order LLP in the voiced subframes gives 19.79 dB, which is an improvement of 0.40 dB over a 50th-order LP-based modified LD-CELP. Informal listening tests show that the quality of decoded speech of the hybrid codec is comparable to that of LD-CELP.

The prediction residual results of Section II and comparative listening tests lead to the important and striking observation that short-term nonlinear predictors are capable of significantly modeling long-term linear or pitch period correlation. This property is worthy of a detailed investigation, because it has the potential of eliminating a hard decision about the pitch period. The results suggest that alternative versions of state-

based local prediction suited for lower rate speech coding may have a significant impact in future speech coding algorithms.

REFERENCES

[1] A. C. Singer, G. W. Wornell, and A. V. Oppenheim, "Codebook prediction: A nonlinear signal modeling paradigm," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1992, vol. 5, pp. 325–328.

[2] B. Townshend, "Nonlinear prediction of speech," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1991, pp. 425–428.

[3] A. Kumar, "Nonlinear dynamical analysis and predictive coding of speech," Ph.D. dissertation, Indian Institute of Technology, Kanpur, India, 1994.

[4] S. Wang, E. Paksoy, and A. Gersho, "Performance of nonlinear prediction of speech," in *Proc. Int. Conf. Spoken Language*, Kobe, Japan, 1990, pp. 29–32.

[5] J.-H. Chen *et al.*, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard," *IEEE J. Select. Areas Commun.*, vol. 10, no. 5, pp. 830–849.

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