Robust Predictive Vector Quantizer Design

Hosam Khalil and Kenneth Rose
Signal Compression Laboratory
Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106
{hosam,rose}@ece.ucsb.edu *

Abstract

The design of predictive quantizers generally suffers from difficulties due to the prediction loop, which have an impact on the convergence and the stability of the design procedure. We recently proposed an asymptotically closed-loop approach to quantizer design for predictive coding applications, which benefits from the stability of open-loop design while asymptotically optimizing the actual closed-loop system. In this paper, we present an enhancement to the approach where joint optimization of both predictor and quantizer is performed within the asymptotically closed-loop framework. The proposed design method is tested on synthetic sources (first-order Gauss and Laplacian-Markov sequences), and on natural sources, in particular, line spectral frequency parameters of speech signals.

1 Introduction

Vector quantizers have been incorporated into several international speech coding standards, and have also found application in image and video compression. Such sources usually exhibit considerable correlation. In order to achieve better efficiency, quantizers with memory have been designed to exploit this correlation. Examples for such VQ structures include finite-state VQ, classified VQ, and predictive VQ (PVQ), which is the quantizer structure of interest in this paper. While the complexity of predictive VQ may not be higher than conventional memoryless VQ, the performance gains are often considerable.

In this work we introduce a technique for PVQ design based on asymptotic closed-loop optimization. We then present simulation results on different input sources.

*This work was supported in part by the National Science Foundation under grant no. IIS-9978001, EIA-9986057, the University of California MICRO Program, Conexant Systems, Inc., Lernout & Hauspie Speech Products, Lucent Technologies, Inc., Medio Stream, Inc., and Qualcomm, Inc.

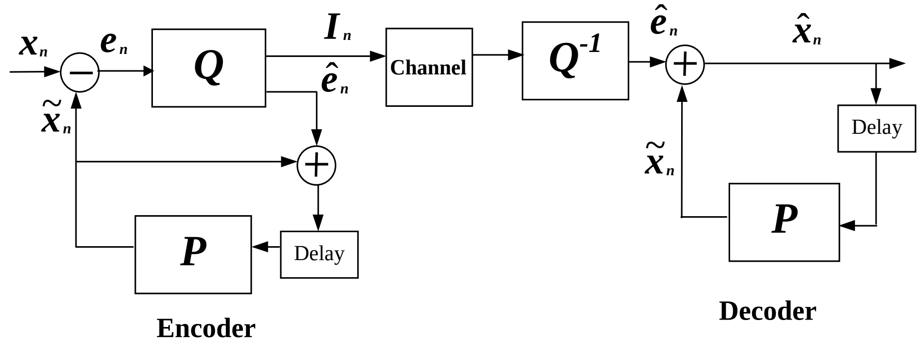

Figure 1: A basic predictive vector quantizer system.

2 Traditional PVQ Design

2.1 The Problem:

A typical PVQ system is shown in Fig. 1. Let $X : \{x_n\}_{n=0}^{N}$ be a vector-valued source over the k-dimensional Euclidean space R^k. A first-order linear vector predictor is used here to simplify the notation. In this case, the prediction process may be denoted by the simple matrix multiplication $\tilde{x}_{n+1} = A\hat{x}_n$, where A is of dimensionality $k \times k$. For an input sequence of vectors, the prediction error e_n, and reconstruction sequence $\hat{x}_n$ are defined recursively by $e_n = x_n - \tilde{x}_n$ and $\hat{x}_n = \tilde{x}_n + Q(e_n)$, respectively. The quantization operator is denoted by $Q(.)$.

Although the PVQ structure is well-known, its design is problematic, and often fails to produce an optimal (and even a good) predictor and quantizer. The feedback loop creates a complex relationship between predictor and quantizer. To design the quantizer, a representative training set of prediction errors is needed. But obtaining the prediction errors requires the system to run in closed-loop to produce such samples, which means that it depends on the quantizer to be designed. Further, there are open questions concerning the design of the predictor. Clearly, the choice of the best predictor is not independent of the quantizer in use. Two commonly-used approaches were introduced by Cuperman and Gersho [1], and a joint predictor and quantizer design method was introduced by Chang and Gray [2].

2.2 The Open-Loop and Closed-Loop Approaches

In the open-loop method (OL) described in [1], both the predictor and quantizer are designed based on the original *unquantized* source vectors. The auto-regressive predictor is obtained from the autocorrelation of the input source (t denotes transposition):

$$A_{OL} = R_1 R_0^{-1} = (\sum_{n=1}^{N} x_n x_{n-1}^t)(\sum_{n=1}^{N} x_{n-1} x_{n-1}^t)^{-1}. \tag{1}$$

A training set of prediction error vectors, $T = \{e_n\}_{n=1}^{N}$, is generated from the original source vectors: $e_n = x_n - \tilde{x}_n = x_n - A_{OL} x_{n-1}$. The design of Q_{OL} is a straightforward design based on the training set T.

In the closed-loop approach (CL) of [1], a closed-loop (real) system is used to generate the prediction errors for the *quantizer design* in an iterative fashion. The predictor, however, is designed using the open-loop method, i.e., $A_{CL} = A_{OL}$. Given a quantizer at iteration $i-1$, which we denote by $Q^{(i-1)}$, a training set of prediction errors is generated for iteration i, $T^{(i)} = \{e_n^{(i)}\}_{n=1}^N$ where, $e_n^{(i)} = x_n - A_{CL}\hat{x}_{n-1}^{(i)}$, and $\hat{x}_n^{(i)} = A_{CL}\hat{x}_{n-1}^{(i)} + Q^{(i-1)}(e_n^{(i)})$.

Comparing OL with CL, the main advantage of the OL approach is that the training set, T, is fixed. Therefore, we can design the PVQ by applying a standard optimization technique such as GLA [3]. Since the training set remains unchanged, the design algorithm is expected to converge to a locally optimal solution. However, the decoder does not have access to the original source vector for prediction. Therefore, during the actual operation of the compression system, prediction is performed using reconstructed source vectors. Thus, the training set of prediction errors is statistically different from the prediction errors to be quantized in practice. The statistical mismatch, which is exacerbated by feedback through the prediction loop, results in poor performance. For the CL method, since the training residuals were generated by the same closed-loop coder that will be used in the actual mode of operation, the input residual error statistics are expected to be similar to those used to train the quantizer. However, convergence of the algorithm is not guaranteed, as the training set changes from iteration to iteration in an unpredictable fashion, and the predictor is not re-optimized for a new quantizer. In fact, the CL system sometimes exhibits catastrophically unstable behavior.

2.3 The Steepest Descent Approach

The previous two methods assumed that the bit rate is high, and thus the open-loop design of the predictor is sufficient. Two gradient algorithms for designing predictive vector quantizers were developed by Chang and Gray [2]; the steepest descent algorithm and the stochastic gradient algorithm. Both of these methods, which are based on known adaptive filtering techniques, attempt to jointly optimize quantizer and predictor. We choose to present an overview of the steepest descent (SD) algorithm since it is less complex than the stochastic gradient method, while producing similar results ([2]). This method proposes to improve over the CL method by including optimization of the predictor. Since changing the predictor affects the training residual and thus the quantizer, and vice versa, a joint optimization is needed. Such joint optimization is achieved through an iterative procedure. The step sizes used in [2][4] are assumed for the following algorithm:

The predictor in this case is variable from iteration to iteration, $A_{SD} = A^{(i)}$. Given a quantizer at iteration $i-1$, $Q^{(i-1)}$, a training set of prediction errors $T^{(i)} = \{e_n^{(i)}\}_{n=1}^N$ is generated for iteration i where,

$$e_n^{(i)} = x_n - A^{(i)}\hat{x}_{n-1}^{(i)}, \tag{2}$$

and

$$\hat{x}_n^{(i)} = A^{(i)}\hat{x}_{n-1}^{(i)} + Q^{(i-1)}(e_n^{(i)}). \tag{3}$$

The solution for the optimal predictor is obtained by taking the gradient of the average distortion:

$$\nabla_{A^{(i)}} D(x, \hat{x}) = \nabla_{A^{(i)}} \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \|x_n - \hat{x}_n\|^2 = 0. \tag{4}$$

To solve (4), it is necessary to make the assumption that changing $A^{(i)}$ will have no effect on $\hat{x}_n^{(i)}$. This postulation is only very approximate. Defining $P^{(i)}$ and $R^{(i)}$ as the cross-correlation and correlation matrices, respectively, at iteration i, the optimal predictor is defined to be $A^{(i)} = P^{(i)}{R^{(i)}}^{-1}$, where

$$P^{(i)} = \frac{1}{N}\sum_{n=1}^{N}(x_n - Q^{(i-1)}(e_n^{(i)}))(\hat{x}_{n-1}^{(i)})^t \tag{5}$$

and

$$R^{(i)} = \frac{1}{N}\sum_{n=1}^{N}\hat{x}_{n-1}^{(i)}(\hat{x}_{n-1}^{(i)})^t. \tag{6}$$

Using the new $T^{(i)}$ and the new $A^{(i)}$, a new $Q^{(i)}$ is optimized. For better stability, after calculating a new predictor, a new training set is generated using the new predictor (and latest quantizer) before the new quantizer is designed. Similarly, a new reconstructed sequence is generated using the new quantizer (and latest predictor) before a new predictor is calculated. Even so, there is a stability problem, as there is a complex interaction between predictor and quantizer. Efforts to reduce the updating step sizes may improve the stability problem but do not significantly improve the final results. In fact, in several applications for PVQ, the improvement obtained by SD over the simpler CL method was not significant [2],[5].

3 Asymptotic Closed-Loop PVQ Design

Motivated by the shortcomings of the existing methods, we recently proposed the Asymptotic Closed-Loop [6][7] approach (ACL), which offers improved design stability. In [6][7], our objective was quantizer design only and did not include optimization of the predictor, as the predictor in video coding applications usually takes the fixed form of motion compensation. In this paper, we enhance the ACL design by proposing a joint optimization of predictor and quantizer, making the algorithm applicable to a wider array of applications. The ACL design procedure inherits the design stability of open-loop techniques while ultimately optimizing the system for closed-loop operation. The ACL procedure for PVQ design is illustrated in Fig. 2, and is explained below in more detail.

The main objective of the design procedure is to avoid accumulation of errors due to mismatched quantization through the prediction loop. We therefore base our prediction on the reconstructed vectors of the *previous iteration*. By basing prediction on an "older" version of reconstructed vectors, the prediction residuals are in effect calculated in open-loop and we can thus circumvent the destabilizing effects of the feedback of the closed-loop system. As a direct consequence of this effectively open-loop approach, a monotonic optimization procedure is possible for jointly optimizing both quantizer and predictor (for the given set of fixed vectors on which prediction is based). Once such quantizer and predictor are optimized, an improved set of reconstructed vectors is produced (also in open-loop). The new set of reconstructed vectors is fixed for the next design iteration, and the steps are repeated until convergence.

To start the algorithm, a set of reconstructed vectors at iteration $i = 0$, $\{\hat{x}_n^{(0)}\}_{n=0}^{N}$, is needed. Denote the prediction residuals and absolute value for the distortion achievable by the current predictor and quantizer as follows:

$$e_n^{(i,s)} = x_n - A^{(i,s)}\hat{x}_{n-1}^{(i-1)} \tag{7}$$

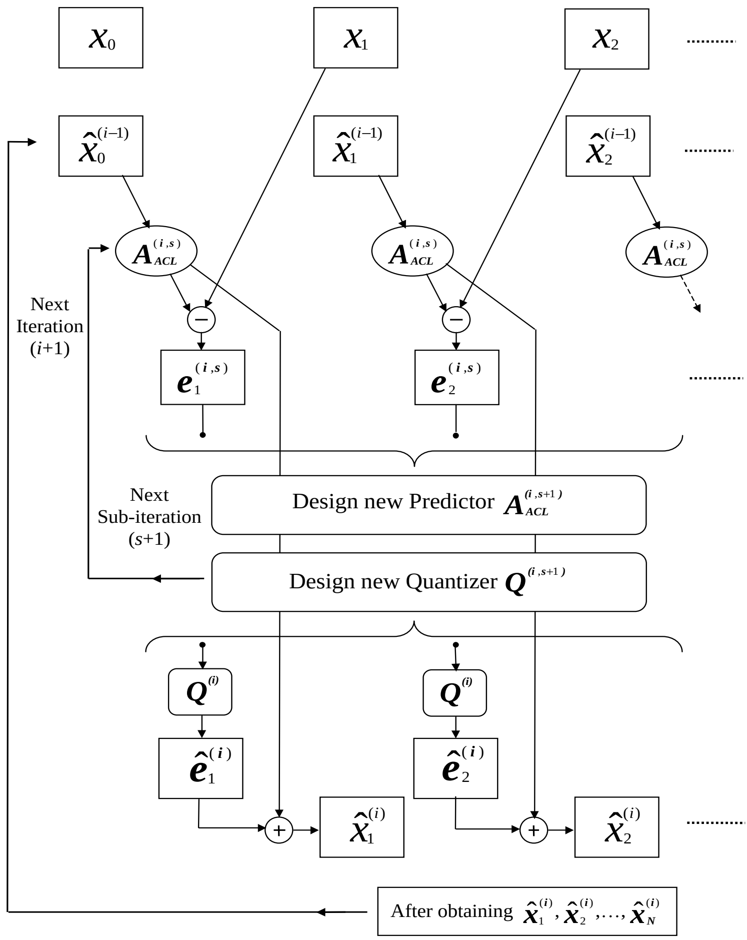

Figure 2: Proposed ACL procedure.

and

$$D^{(i,s)} = \sum_{n=1}^{N} \left\| e_n^{(i,s)} - Q^{(i,s)}(e_n^{(i,s)}) \right\|^2. \tag{8}$$

The iteration order indicated by i is supplemented here by a sub-iteration number indicated by s to indicate that for an iteration i, further optimization by iterating over s is possible. In (8), distortion is only function of two variables: The prediction parameters and quantizer parameters. But note that we have no closed-loop dependencies in the above equation, as the prediction is based on the fixed reconstructed set from the previous iteration. An alternate minimization algorithm is proposed to minimize (8) as follows: For the current quantizer, the best predictor to minimize the distortion is obtained by descending along the gradient $\nabla_{A^{(i,s)}} D^{(i,s)}$. Using the new predictor, a new residual set $e^{(i,s)}$ is obtained, and the distortion is minimized by optimizing a new quantizer $Q^{(i,s)}$ using GLA. These two distortion-lowering steps can be repeated by iterating over s to convergence.

Specifically, for the design of a predictor at iteration i and sub-iteration s, the optimal predictor

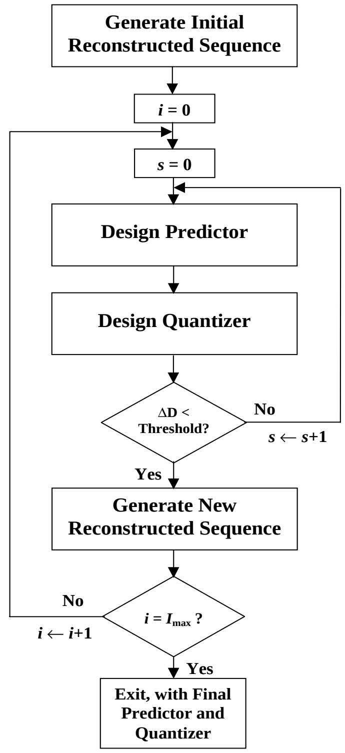

Figure 3: Flow-diagram for the ACL training procedure.

is $A^{(i,s)} = P^{(i,s)} {R^{(i)}}^{-1}$, where

$$P^{(i,s)} = \sum_{n=1}^{N} (x_n - Q^{(i,s)}(e^{(i,s)}))(\hat{x}_{n-1}^{(i)})^t, \tag{9}$$

$$R^{(i)} = \sum_{n=1}^{N} \hat{x}_{n-1}^{(i)} (\hat{x}_{n-1}^{(i)})^t. \tag{10}$$

Also, for the new optimal quantizer, the new training set is generated by $T^{(i,s)} = \{e_n^{(i,s)}\}_{n=1}^N$ where $e_n^{(i,s)} = x_n - A^{(i,s)} \hat{x}_{n-1}^{(i-1)}$.

When the predictor and quantizer iterations reach convergence, the final quantizer becomes $Q^{(i)} = Q^{(i,s_{final})}$ and predictor $A^{(i)} = A^{(i,s_{final})}$.

Increment i, reset s, and calculate the new set of reconstructed vectors by

$$\hat{x}_n^{(i+1)} = A^{(i)} \hat{x}_{n-1}^{(i)} + Q^{(i)}(e_n^{(i,s_{final})}). \tag{11}$$

Fix the reconstructed set, assign $Q^{(i+1,0)} = Q^{(i)}$, $A^{(i+1,0)} = A^{(i)}$, and repeat algorithm till the reconstructed set converges. At convergence, exit routine with $A_{ACL} = A^{(i_{final})}$, and $Q = Q^{(i_{final})}$. See Fig. 3 for a flow-diagram of the above training procedure.

Note that in (11), the quantizer $Q^{(i)}$ is used to encode *exactly* the same prediction error vectors used for its design. Neglecting the possible local-optimality of the quantizer design algorithm, this is the best quantizer for these vectors. We are thus assured that the resulting reconstruction is improved, and this results in better prediction. Under the reasonable and common assumption that smaller prediction errors lead to smaller quantization errors (and vice versa), we obtain monotonic improvement throughout the process.

Note that the entire design is in open-loop mode since we compute prediction errors for the entire sequence before quantization. As the distortion is generally decreasing, we expect the process to converge. At convergence, further iterations do not modify the quantizer and predictor, i.e., $Q^{(i+1)} = Q^{(i)}$, and $A^{(i+1)} = A^{(i)}$, respectively. This immediately ensures that the reconstruction sequence is non-changing, i.e., $\hat{x}_n^{(i+1)} = \hat{x}_n^{(i)}$, and that the next-vector prediction sequence is non-changing:

$$A_{ACL}\hat{x}_{n-1}^{(i)} = A_{ACL}\hat{x}_{n-1}^{(i-1)}. \tag{12}$$

This, in turn, implies that the prediction would be unchanged if it were based on the reconstruction of the current iteration, instead of on the reconstruction from the previous iteration. In other words, the procedure is asymptotically equivalent to closed-loop design, but the algorithm is running at all times in open loop. The procedure is thus "open-loop" in nature, yet it converges to optimization of the closed-loop performance.

4 Simulation Results

The advantages of the original ACL algorithm as presented in [7] were especially apparent in the application of video coding, where stability was an important issue at low coding rates. When the original ACL algorithm was tested on synthetic sources in [7], the major benefit achieved was robustness at very low bit rates. However, when the available bit rate is high, the accumulation of error is greatly reduced and thus, the gains over traditional design techniques tend to diminish. With the new approach presented in this paper, where the predictor and quantizer are jointly optimized, the new ACL outperforms all other methods consistently over a variety of sources and operating points.

4.1 Experiments on Synthetic Sources

In the first set of experiments, the proposed ACL approach is compared to the OL, CL, and SD methods on several synthetic sources. Specifically, sources with the following characteristics were synthesized. Source GM-A: a 6-dimensional first-order Gauss-Markov source with intra-vector and inter-vector correlation coefficients of 0.9. Source GM-B: a 6-dimensional Gauss-Markov source with intra-vector and inter-vector correlation of 0.8 and 0.95, respectively. Source GM-C: a 6-dimensional first-order Gauss-Markov source with inter-vector correlation of 0.8 and (in an effort to produce a more realistic source) intra-vector correlation varying along the components, in the range of 0.5 to 0.95. Source LM: a 10-dimensional Laplace-Markov source with intra-vector and inter-vector correlation of 0.95. A first-order predictor is used in the PVQ design.

We present results for a sample of bit rates. Table 1 summarizes the gains in dB achieved by ACL at various bit rates. Figure 4 shows the average distortion over the training set and test set of source GM-C, at a coding rate of 8 bits/vector, and its evolution with the iterations. The distortion shown is for normal closed-loop operation of the PVQ system when using the predictor and quantizer available at the corresponding iteration of the design. It is can be seen that both

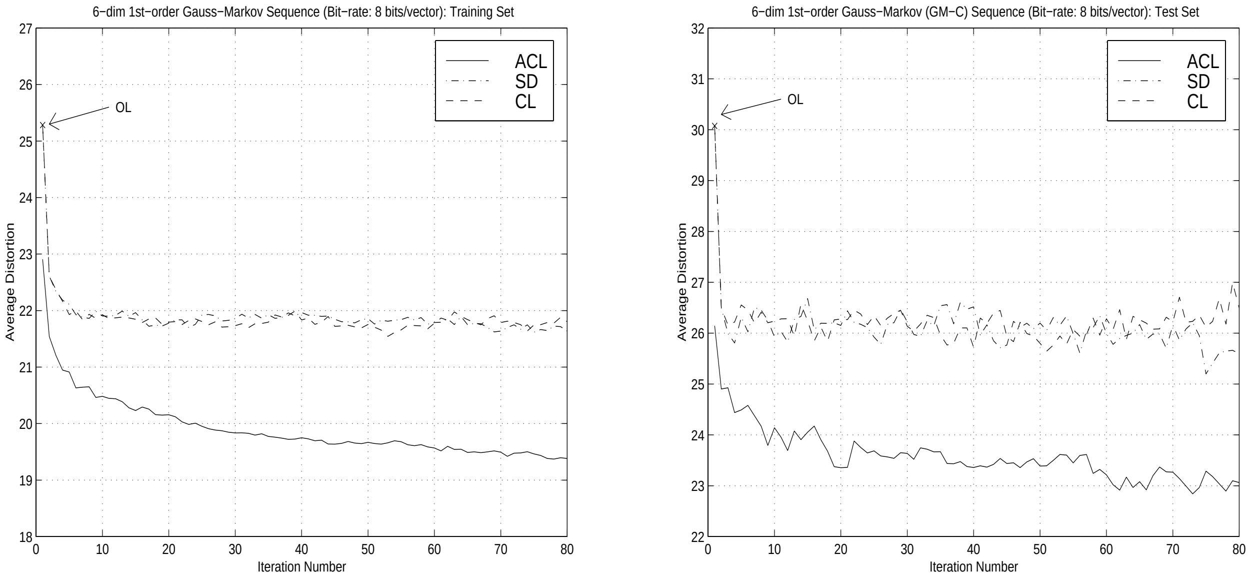

Figure 4: Synthetic source GM-C: Average distortion of OL, CL, and SD design and the proposed ACL approach to predictive quantizer design at a rate of 8 bits/vector, for a) Training set, and b) Test set.

CL and SD improve over OL. However, ACL achieves much more substantial gains. By carefully studying Table 1, it can be seen that the gains achievable by ACL over traditional methods for the simplest Gauss-Markov source GM-A at high rate are relatively small.

4.2 Experiments on Line Spectral Frequencies

Several speech coding standards are centered around the encoding of line spectral frequency (LSF) parameters. LSF parameters typically constitute 25% to 50% of the speech coding rate. In addition, LSF parameters have significant and direct impact on the intelligibility and quality of decoded speech. Thus, high-quality encoding of LSF parameters is an important objective, and careful design of PVQ is warranted. A frame of speech generally includes a 10-dimension LSF

Table 1: Gain in dB of ACL over traditional methods for various synthetic source processes.

		Training Set			Test Set		
Source	Bits	ACL/OL	ACL/CL	ACL/SD	ACL/OL	ACL/CL	ACL/SD
GM-A	5	1.38	0.65	0.46	1.35	0.51	0.47
GM-A	9	0.34	0.13	0.15	0.49	0.03	0.08
GM-B	8	0.55	0.47	0.47	0.58	0.46	0.46
GM-B	9	0.42	0.35	0.34	0.54	0.42	0.41
GM-C	5	2.05	0.83	0.83	2.10	1.02	1.09
GM-C	8	1.16	0.48	0.46	1.20	0.55	0.42
LM	6	1.23	1.23	1.23	0.22	0.22	0.22

Table 2: Gain in dB of ACL over traditional methods for LSF source.

		Training Set			Test Set		
Source	Bits	ACL/OL	ACL/CL	ACL/SD	ACL/OL	ACL/CL	ACL/SD
LSF	6	1.35	0.68	0.68	0.84	0.55	0.55
LSF	9	0.86	0.35	0.25	0.54	0.41	0.38
LSF	18	0.70	0.57	0.59	0.56	0.49	0.47
LSF	21	0.66	0.54	0.53	0.57	0.47	0.47

Table 3: Average spectral distortion (in dB) for LSF quantization using CL and ACL.

	Training Set			Test Set		
Bits	CL	SD	ACL	CL	SD	ACL
18	1.48	1.49	1.41	1.47	1.47	1.41
21	1.22	1.22	1.17	1.22	1.22	1.18

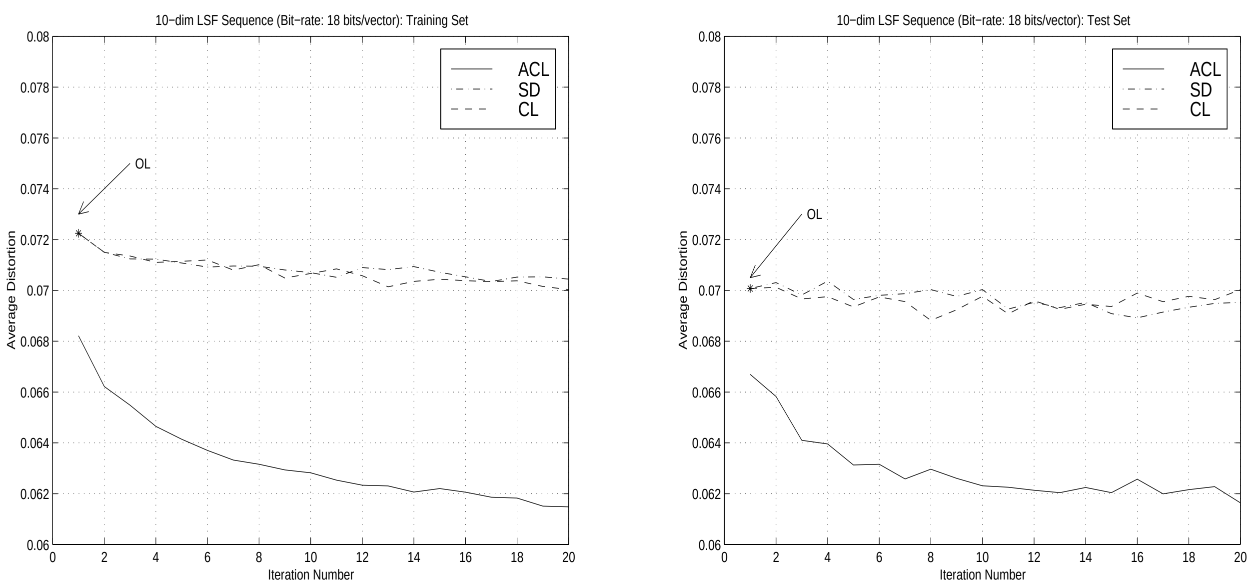

Figure 5: Line spectral frequency parameters: Average distortion of OL, CL, and SD design and the proposed ACL approach to predictive quantizer design at a rate of 18 bits/vector, for a) Training set, and b) Test set.

vector that usually exhibits significant correlation with previous frames. A common approach is to use a first or second-order interframe predictor to remove such redundancies. In this experiment, we use a first-order predictor and quantizers that operate at a variety of bit rates. In particular, two examples are shown in Table 2 for the bit rates of 6 and 9 bits/vector which are realizable using a single stage VQ. For higher bit rates, we show results for the two cases of 18 and 21 bits/vector, which are realizable using multi-stage VQ of 3 stages and bit assignments of 6+6+6, and 7+7+7, respectively.

Figure 5 shows the average distortion over the LSF training set and test set, at a coding rate of 18 bits/vector, and its evolution with the iterations. The spectral distortion in dB for the 18 and 21 bits/vector cases are specified in Table 3. The spectral distortion is believed to capture approximately the perceptual quality of the speech signal. However, the design here did not directly attempt to optimize this criterion. It should be noted that the gains achieved by ACL are at absolutely no cost to the encoder and decoder, and are due only to the design approach.

5 Conclusions

This paper describes a new approach to training predictive vector quantizers, which does not suffer from the statistical mismatch typical of OL training algorithms, nor from the instability experienced in the CL and SD approaches. Simulation results show the superiority of the proposed design algorithm over conventional approaches.

References

[1] V. Cuperman and A. Gersho, "Vector predictive coding of speech at 16 kbits/s," *IEEE Trans. Communications*, vol. COM-33, no. 7, pp. 685-696, July 1985.

[2] P. C. Chang and R. M. Gray, "Gradient algorithms for designing predictive vector quantizers," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 679-690, Aug. 1986.

[3] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. COM-28, pp. 84-95, Jan. 1980.

[4] K. Zeger, "Corrections to "Gradient algorithms for designing predictive vector quantizers," *IEEE Trans. Signal Processing*, vol. 39, no. 3, pp. 764-765, March 1991.

[5] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Boston, MA: Kluwer, 1992.

[6] H. Khalil and K. Rose, "Asymptotic Closed-Loop Design of Predictive Multi-Stage Vector Quantizers," *Proc. IEEE Int. Conf. on Image Processing*, ICIP'00, Vancouver, Canada, Sept. 2000.

[7] H. Khalil and K. Rose, "The Asymptotic Closed-Loop Approach to Predictive Vector Quantizer Design with Application in Video Coding," *IEEE Trans. Image Processing*, vol. 10, no. 1, pp. 15-23, Jan. 2001.