Toward Optimality in Scalable Predictive Coding

Kenneth Rose, *Member, IEEE,* and Shankar L. Regunathan, *Student Member, IEEE*

***Abstract*—A method is proposed for efficient scalability in predictive coding, which overcomes known fundamental shortcomings of the prediction loop at enhancement layers. The compression efficiency of an enhancement-layer is substantially improved by casting the design of its prediction module within an estimation-theoretic framework, and thereby exploiting all information available at that layer for the prediction of the signal, and encoding of the prediction error. While the most immediately important application is in video compression, the method is derived in a general setting and is applicable to any scalable predictive coder. Thus, the estimation-theoretic approach is first developed for basic DPCM compression and demonstrates the power of the technique in a simple setting that only involves straightforward prediction, scalar quantization, and entropy coding. Results for the scalable compression of first-order Gauss–Markov and Laplace–Markov signals illustrate the performance. A specific estimation algorithm is then developed for standard scalable DCT-based video coding. Simulation results show consistent and substantial performance gains due to optimal estimation at the enhancement-layers.**

I. Introduction

IT has become a common requirement of coding and transmission systems to provide a scalable bitstream. Many applications, including multiparty video conferencing and multicast over the Internet, require the compressed information to be simultaneously transmitted to multiple receivers over different communication links. The evolving global communication network is, in fact, a patchwork of transmission media, which is highly nonuniform in its communication capabilities, and is characterized by vast variations in the channel bandwidth available to different links and to the same link at different moments. Moreover, the feasible bit rate of each receiver is constrained by its computational power and memory capacity.

A scalable bitstream is one that allows decoding at a variety of bit rates (and corresponding levels of quality), where the lower rate information streams are embedded within the higher rate bitstreams in a manner that minimizes redundancy. We are chiefly concerned here with what is commonly referred to as "SNR scalability," but the work is extendible to include scalability via various forms of down-sampling.

In the most common approach to scalability [10], enhancement layers simply compress and transmit the reconstruction error of the lower (base) layers. In other words, the best reconstruction available so far is used as an *estimate* for the original signal, and the estimation error is compressed for the next enhancement layer. This estimate ensures that the compressed residual (prediction error) of the lower layers is fully utilized. In the case of *predictive coding*, this approach to scalability is suboptimal as there is potentially useful information available from prior reconstructed samples at the enhancement layer, which could be used to improve the enhancement-layer estimate of the current sample. A scalable coder that neglects the additional information available for enhancement-layer estimation can incur a significant penalty in compression performance.

In the specific case of (nonscalable) video coding, the standard compression technique predicts the current frame from the motion-compensated previous frame prior to transformation and quantization (see Fig. 1). Scalable video coding, therefore, suffers from the above mentioned suboptimality [9]. This problem has also led to proposals of nonpredictive scalable video coding such as the three-dimensional coding approach [18]. However, predictive coders are generally preferred in most practical applications because of their minimal requirements in terms of delay and memory and, further, because they allow straightforward incorporation and exploitation of motion compensation. An alternate approach to scalability in predictive video coding is to use the previous enhancement layer reconstruction for prediction at both the base and enhancement layers [5], [9]. Since the base-layer decoder does not have access to enhancement-layer reconstruction, this results in a *drift* between encoder and decoder reconstruction at the base-layer. This method provides efficient compression for the enhancement-layer, but the accumulating drift may lead to degradation in base-layer performance. While drift may not be a significant problem in some applications [2], we follow the trend of recent standards such as H.263+ and MPEG-4, and prefer to focus exclusively on drift-free coders that aim at true scalability, i.e., those that achieve efficient compression at the enhancement-layer without compromising the base-layer performance.

In this work, we develop an *estimation-theoretic* (ET) approach to enhancement-layer prediction in scalable coders. This prediction, or rather estimation, at the enhancement layer is shown to be optimal in the sense that it minimizes the mean squared prediction error given all the information available at the enhancement layer. In experiments, this optimality translates into substantial gains in compression efficiency at the enhancement-layer. The method is first derived and explained in the simpler and fundamental setting of two-layer differential pulse code modulation (DPCM). It is then adopted to and demonstrated in the context of predictive DCT-based video coding with multiple layers of scalability. The coders we develop in this work are tailored toward the broadcast scenario, i.e., in the context of two-layer scalable coding we assume that the base-layer bitstream is received error free for a subset of decoders, while both base and enhancement layer bitstreams are received error free

Manuscript received October 6, 1999; revised March 20, 2001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Luis Torres.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: rose@ece.ucsb.edu).

Publisher Item Identifier S 1057-7149(01)05432-X.

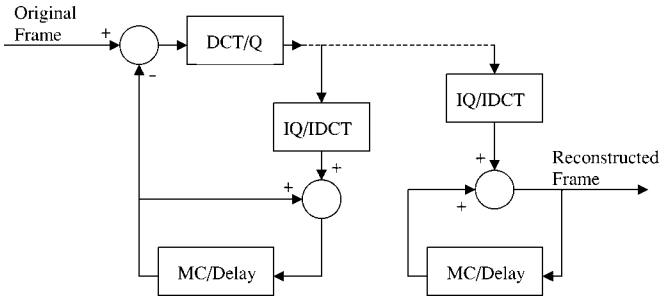

Fig. 1. Sketch of generic predictive coding scheme. Transform (DCT/IDCT) and MC modules are specific to video coding.

at the other decoders. The application of estimation-theoretic prediction to scalable video coding over packet loss channels is pursued in [21].

The paper is organized as follows. In Section II we state, discuss, and motivate the problem. Section III provides the derivation of our approach within an estimation-theoretic framework for the basic setting of a scalable DPCM coder. It also includes simulation results and high resolution analysis to substantiate the performance gains. Section IV adopts the optimal estimation approach to the problem of DCT-based scalable video compression. Simulation results demonstrate the performance advantage of our approach over standard scalable video coders.

II. Problem and Motivation

Let us consider a two-layer scalable coder. The prediction at the base-layer is that of a standard (nonscalable) coder, and is simply based on prior reconstructed base-layer samples. (In the case of video coding, it consists of motion-compensating the previous base-layer reconstructed frame). The main difficulty arises at the prediction module for the enhancement-layer where there are two candidate predictors. On the one hand, it is advantageous to predict the current sample (frame) from the *previous* reconstructed *enhancement*-layer sample (frame) since the enhancement layer offers better quality of reconstruction than the base layer. On the other hand, one may employ the base-layer prediction and complement it with the current compressed base-layer residual (prediction error), i.e., an estimate based on the *current base*-layer reconstruction. The two main existing approaches to enhancement-layer prediction amount to the exclusive use of either one of the above sources of information:

P1: Discard the additional information available from prior samples of the enhancement layer. Use the current base-layer reconstruction as the estimate. In other words, the enhancement layer directly compresses the base-layer reconstruction error (e.g., [19]). A coding system using P1 for enhancement-layer prediction is shown in Fig. 2.

P2: Discard the information contained in the compressed base-layer residual. Predict the current sample (frame) from prior enhancement-layer reconstructed samples (motion-compensated frames) as in [7]. Note that in this case the two layers are, in fact, separately encoded (simulcast) except for savings on shared side-information such as motion vectors. Fig. 3 shows a complete coding system using P2 for enhancement-layer prediction.

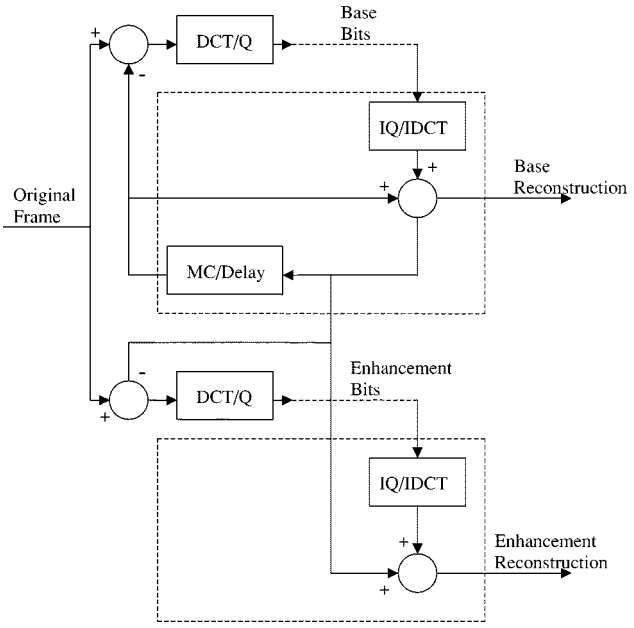

Fig. 2. Sketch of two-layer scalable (en)coder with P1 prediction at the enhancement-layer. Encoder of each layer contains the corresponding decoder (indicated by dotted lines). Transform (DCT/IDCT) and motion compensation (MC) modules are specific to video coding.

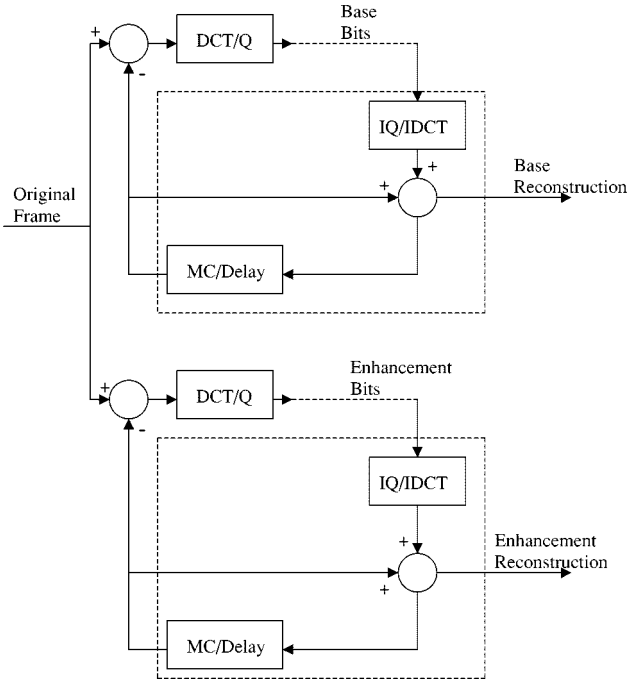

Fig. 3. Sketch of two-layer scalable (en)coder with P2 prediction at the enhancement-layer. Encoder of each layer contains the corresponding decoder (indicated by dotted lines). Transform (DCT/IDCT) and motion compensation (MC) modules are specific to video coding.

More sophisticated proposals are based on switching between these sources of information in order to adaptively select the better of the two. These include switching per macro-block as proposed in the H.263+ [22] and MPEG-4 [24] standards and switching per coefficient in the context of pyramid and subband techniques [3], [17].

The main observation is that all the above methods are restricted to exploit only one of the available information sources

(and hence discard the other) at any time instance. An important exception to this rule can be found in MPEG-2's spatial scalability [13], H.263+ [22], MPEG-4 [24], where the enhancement predictor switches per macroblock between P1, P2, and a weighted linear combination of the two. However, linear combination remains an ad-hoc method of combining the two information sources and requires transmission of the weights as side-information.

The above provides direct motivation for the work described in this paper. We propose an estimation-theoretic (ET) approach which ensures that all sources of information available to the enhancement-layer are optimally exploited.

III. Scalable DPCM Coder Design

Let us reformulate the problem as one of *estimation* and *coding* of the current sample at the enhancement layer given all available information. It is convenient to define the two sources of information as: (i) enhancement-layer reconstruction of prior samples, and (ii) values of all parameters and variables associated with the base-layer compression of the current sample (including the reconstruction value, the compressed residual, and the quantization parameters). Note that we assume that all relevant information from past base-layer reconstruction samples is subsumed by the enhancement-layer reconstruction of those samples. Finally, we assume the existence of a statistical model for the signal, which may be used for prediction. We will show that even naive models are sufficient to achieve significant gains in practical video coding systems.

The prediction error at both the base and enhancement-layers is assumed to be scalar quantized. The quantizer index is encoded by a lossless entropy code and transmitted over the channel. The distortion criterion is the commonly used mean-squared error. We first focus on the optimal estimation (prediction) of the sample at the enhancement-layer, and then discuss the optimal entropy coding of its prediction error.

A. Estimation-Theoretic Predictor Derivation

Let x_n, $\hat{x}_n^b$ and $\hat{x}_n^e$ be the current sample, its base and enhancement-layer reconstruction values, respectively.

1) Base-Layer: The optimal *base*-layer predictor of the current sample is obtained by expectation over the conditional density $p\left(x_n|\hat{x}_{n-1}^b,\hat{x}_{n-2}^b,\ldots\right)$

$$\tilde{x}_n^b = E\left[x_n|\hat{x}_{n-1}^b,\hat{x}_{n-2}^b,\ldots\right]. \tag{1}$$

The base encoder quantizes the residual

$$r_n^b = x_n - \tilde{x}_n^b$$

and transmits index i^b. Let (a,b) be the quantization interval associated with index i^b, i.e., $r_n^b \in (a,b)$. Clearly, the statement $x_n \in \left(\tilde{x}_n^b + a, \tilde{x}_n^b + b\right)$ captures *all the information* provided to the decoder on x_n by the received residual index. Therefore, the optimal base-layer reconstruction is given by

$$\hat{x}_n^b = E\left[x_n|x_n \in \left(\tilde{x}_n^b + a, \tilde{x}_n^b + b\right), \hat{x}_{n-1}^b, \hat{x}_{n-2}^b, \ldots\right]. \tag{2}$$

This estimate is computed by calculating the centroid of the interval $\left(\tilde{x}_n^b + a, \tilde{x}_n^b + b\right)$ with respect to the density $p\left(x_n|\hat{x}_{n-1}^b,\hat{x}_{n-2}^b,\ldots\right)$

$$\hat{x}_n^b = \frac{\int_{\tilde{x}_n^b+a}^{\tilde{x}_n^b+b} x_n p\left(x_n|\hat{x}_{n-1}^b,\hat{x}_{n-2}^b,\ldots\right) dx_n}{\int_{\tilde{x}_n^b+a}^{\tilde{x}_n^b+b} p\left(x_n|\hat{x}_{n-1}^b,\hat{x}_{n-2}^b,\ldots\right) dx_n}. \tag{3}$$

Note that (2) and (3) are well approximated by standard predictive coding. We have recast the derivation within an estimation-theoretic framework to prepare the approach for the case of the enhancement-layer, where common practice differs considerably from the optimal approach.

2) Enhancement-Layer: In addition to the information provided by the base-layer, the enhancement-layer decoder has access to prior enhancement-layer reconstructed samples: $\hat{x}_{n-1}^e, \hat{x}_{n-2}^e, \ldots$ Recall, further, that the compressed base-layer residual provides *precisely* the information: $x_n \in \left(\tilde{x}_n^b + a, \tilde{x}_n^b + b\right)$. Thus, taking into account all the available information, the optimal enhancement-layer predictor is

$$\tilde{x}_n^e = E\left[x_n|x_n \in \left(\tilde{x}_n^b + a, \tilde{x}_n^b + b\right), \hat{x}_{n-1}^e, \hat{x}_{n-2}^e, \ldots\right]. \tag{4}$$

It is reasonable to assume that $\hat{x}_{n-i}^b, \forall\, i > 0$ provide little or no information in addition to that contained in $\hat{x}_{n-i}^e$, and we therefore neglected to condition on prior base-layer reconstructed samples.

Hence, the ET predictor is computed by calculating the centroid of the interval $\left(\tilde{x}_n^b + a, \tilde{x}_n^b + b\right)$ with respect to the density $p\left(x_n|\hat{x}_{n-1}^e,\hat{x}_{n-2}^e,\ldots\right)$

$$\tilde{x}_n^e = \frac{\int_{\tilde{x}_n^b+a}^{\tilde{x}_n^b+b} x_n p\left(x_n|\hat{x}_{n-1}^e,\hat{x}_{n-2}^e,\ldots\right) dx_n}{\int_{\tilde{x}_n^b+a}^{\tilde{x}_n^b+b} p\left(x_n|\hat{x}_{n-1}^e,\hat{x}_{n-2}^e,\ldots\right) dx_n}. \tag{5}$$

This estimate is conditioned on prior enhancement-layer information but, at the same time, it is restricted to the quantization interval determined by the base layer. *Thus, the enhancement-layer ET predictor seeks the best estimate based on prior enhancement-layer reconstruction, which is consistent with the quantization interval specified by the current base-layer*. Note that the estimate takes advantage of all sources of information available to the enhancement-layer. Note, further, that the best estimate is a nonlinear combination of the available information in contrast to the simple weighted average of P1 and P2 as in [13].

The enhancement-layer encoder quantizes the residual

$$r_n^e = x_n - \tilde{x}_n^e$$

and transmits index i^e. Let (c,d) be the quantization interval associated with index i^e. Hence, $r_n^e \in (c,d)$ and $x_n \in (\tilde{x}_n^e + c, \tilde{x}_n^e + d)$. It is convenient to define

$$e = \max\left[\tilde{x}_n^b + a, \tilde{x}_n^e + c\right], \quad f = \min\left[\tilde{x}_n^b + b, \tilde{x}_n^e + d\right]. \tag{6}$$

The information provided by the two quantization intervals is compactly expressed by the statement

$$x_n \in (e,f). \tag{7}$$

The enhancement-layer reconstruction of the sample is given by

$$\hat{x}_n^e = E\left[x_n | x_n \in (e,f), \hat{x}_{n-1}^e, \hat{x}_{n-2}^e, \ldots\right] \tag{8}$$

or

$$\hat{x}_n^e = \frac{\int_{(e,f)} x_n p\left(x_n | \hat{x}_{n-1}^e, \hat{x}_{n-2}^e, \ldots\right) dx_n}{\int_{(e,f)} p\left(x_n | \hat{x}_{n-1}^e, \hat{x}_{n-2}^e, \ldots\right) dx_n}. \tag{9}$$

The above ET predictor derivation is extended in a straightforward manner to the multilayer coding scenario as follows. For prediction at the kth enhancement layer, we use the corresponding layer's reconstruction of previous samples while the quantization interval over which we evaluate the expectation is determined by the quantization intervals of all the layers below it. The information provided by each lower layer specifies an interval in which x_n lies. Thus the overall information provided by all the lower layers is that x_n lies in the intersection of all these intervals. Let us denote this interval by I_k. Thus

$$\hat{x}_n(k) = \frac{\int_{I_k} x_n p(x_n | \hat{x}_{n-1}(k), \hat{x}_{n-2}(k), \ldots) dx_n}{\int_{I_k} p(x_n | \hat{x}_{n-1}(k), \hat{x}_{n-2}(k), \ldots) dx_n}. \tag{10}$$

B. A Special Case: The First-Order Markov Process

To illustrate the workings of the procedure let us consider the important special case where the source is a first-order Markov process

$$x_n = \rho x_{n-1} + z_n \tag{11}$$

where ρ is the correlation coefficient, and z_n is zero-mean, white, wide-sense stationary, and independent of $x_{n-i}, \forall\, i > 0$.

The *base*-layer predictor becomes

$$\tilde{x}_n^b = E\left[x_n | \hat{x}_{n-1}^b\right] \approx \rho \hat{x}_{n-1}^b. \tag{12}$$

The above "commonly used" approximation is based on the assumption that quantization errors are zero-mean and nearly independent, and that the "closed-loop" prediction error density (prediction based on reconstructed samples) is approximated by the "open-loop" prediction error density (based on unquantized samples). These issues have been extensively discussed in the predictive coding literature (see [4], [6], and [11] for such treatment). We will use the above simplifying approximation since it allows the derivation of explicit analytic expressions for the various expectations, while noting that it is sufficient to demonstrate substantial performance gains in the experiments.

The base-layer reconstruction is

$$\begin{aligned} \hat{x}_n^b =& E\left[x_n | x_n \in \left(\tilde{x}_n^b + a, \tilde{x}_n^b + b\right), \hat{x}_{n-1}^b\right] \\ \approx& \rho \hat{x}_{n-1}^b + E\left[z_n | z_n \in (a,b)\right]. \end{aligned} \tag{13}$$

The optimal enhancement-layer predictor becomes

$$\tilde{x}_n^e = E\left[x_n | x_n \in \left(\tilde{x}_n^b + a, \tilde{x}_n^b + b\right), \hat{x}_{n-1}^e\right]. \tag{14}$$

which may be closely approximated as

$$\begin{aligned} \tilde{x}_n^e \approx& \rho \hat{x}_{n-1}^e \\ &+ E\left[z_n | z_n \in \left(\tilde{x}_n^b + a - \rho \hat{x}_{n-1}^e, \tilde{x}_n^b + b - \rho \hat{x}_{n-1}^e\right)\right]. \end{aligned} \tag{15}$$

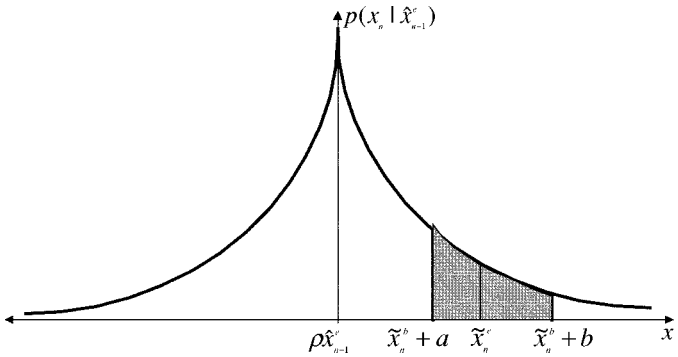

Fig. 4. Computation of ET predictor. The estimate is computed as centroid of the interval specified by the base-layer, $(a + \tilde{x}_n^b, b + \tilde{x}_n^b)$, with respect to the enhancement-layer prediction pdf centered at $\rho \hat{x}_{n-1}^e$.

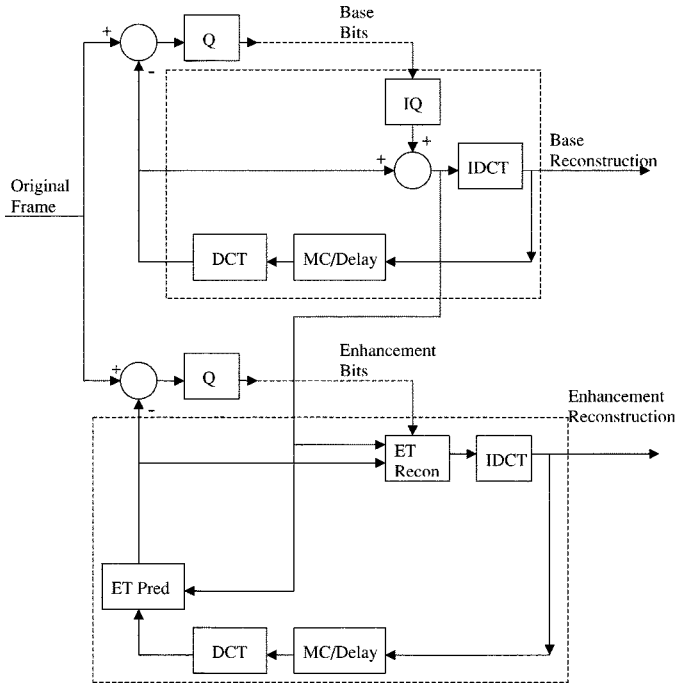

Fig. 5. Sketch of two-layer scalable (en)coder with ET prediction at the enhacement-layer. Encoder of each layer contains the corresponding decoder (indicated by dotted lines). Transform (DCT/IDCT) and motion compensation (MC) modules are specific to video coding.

The formulation in (15) allows direct calculation of the ET predictor from the density $p(z_n)$. Fig. 4 illustrates that the ET predictor can be obtained by computing the centroid of the quantization interval obtained from base-layer with respect to the density $p(z_n)$ whose mean is derived from the previous enhancement-layer reconstruction.

Finally, the enhancement-layer reconstruction is given by

$$\hat{x}_n^e = E\left[x_n | x_n \in (e,f), \hat{x}_{n-1}^e\right] \tag{16}$$

where e and f are given in (6). This is conveniently approximated by

$$\hat{x}_n^e \approx \rho \hat{x}_{n-1}^e + E\left[z_n \in \left(e - \rho \hat{x}_{n-1}^e, f - \rho \hat{x}_{n-1}^e\right)\right]. \tag{17}$$

Fig. 5 shows a complete two-layer coding scheme that uses ET prediction at the enhancement-layer.

We conclude the subsection by showing that the ET predictor degenerates to the conventional prediction schemes, P1 and P2, under certain limiting conditions.

- Total Rate ≈ Base-Layer Rate: If the total rate is approximately the same as base-layer rate, the quality of the base-layer is comparable to that of the enhancement layer and thus $\hat{x}^e_{n-1}$ in (14) may be replaced by $\hat{x}^b_{n-1}$. Hence

$$\tilde{x}^e_n \approx E\left[x_n | x_n \in \left(\tilde{x}^b_n + a, \tilde{x}^b_n + b\right), \hat{x}^b_{n-1}\right] = \hat{x}^b_n \qquad (18)$$

  and the ET predictor is approximated by P1 in this case.
- Low Correlation: If $\rho \approx 0$ then time-prediction provides little gain. It can be readily seen from (13) and (14) that $\tilde{x}^e_n \approx \hat{x}^b_n$. Thus, in this case too, P1 is nearly optimal.
- Base-Layer Rate $\ll$ Enhancement-Layer Rate: The base quantizer is very coarse in comparison to the enhancement-layer quantizer. Thus the quantization interval specified by (a, b) is very large and captures almost all the probability of z_n. We have from (15)

$$\tilde{x}^e_n \approx \rho\hat{x}^e_{n-1} + E[z_n | z_n \in (-\infty, \infty)] = \rho\hat{x}^e_{n-1} \qquad (19)$$

  where the right hand side follows from the fact that z_n is zero-mean. Thus P2 approximates the ET predictor.

In summary, P1 and P2 provide close to optimal performance for either extreme target rates or for extremely low correlation. At most rates of practical interest and for most sources, however, neither P1 nor P2 approximate the ET predictor well enough, and this is the main shortcoming of conventional scalable coders.

C. Conditional Entropy Encoding at the Enhancement-Layer

Let us next consider the encoding of the prediction error at the enhancement-layer. Recall that the optimal predictor (15) is computed by calculating the centroid of interval $I = \left(\tilde{x}^b_n + a - \rho\hat{x}^e_{n-1}, \tilde{x}^b_n + b - \rho\hat{x}^e_{n-1}\right)$ with respect to the conditional density $p(z_n)$. Equivalently, it may be viewed as simple expectation with respect to the density obtained by truncation of $p(z_n)$ to the above interval, followed by normalization

$$p(z_n | z_n \in I) = \begin{cases} \frac{p(z_n)}{\int_I p(z_n)dz_n}, & \forall\, z_n \in I \\ 0, & \text{otherwise.} \end{cases} \qquad (20)$$

It follows that the density of the estimation error, $r^e_n = x_n - \tilde{x}^e_n$, is directly obtained as the zero-mean, shifted version of the density in (20). Thus, the prediction error statistics may vary considerably depending on the position of the base quantizer interval (as shown in Fig. 4). The rate for encoding the residual at the enhancement-layer can be substantially reduced by exploiting this fact via *conditional entropy coding*.

If we make the further approximation that $\left(\tilde{x}^b_n + a - \rho\hat{x}^e_{n-1}, \tilde{x}^b_n + b - \rho\hat{x}^e_{n-1}\right) \approx (a, b)$, then we may condition the entropy directly on the base-quantizer index. In our simulations, we used the simplified setting of two entropy coders for the enhancement-layer. One was designed for the case of "zero" base quantizer index (selected quantization interval contains the origin). The other entropy coder was designed for the complementary case of "nonzero" index. Our simulation results demonstrate that significant gains in compression performance can be achieved by conditional entropy coding, especially, for the Laplace–Markov process.

In principle, conditional entropy coding of the residual may also be used with the conventional prediction method P1. However, the enhancement-layer residual in this case is simply the base-layer reconstruction error, and its statistics show lesser variation with base quantizer interval. Therefore, conditional entropy coding in conjunction with P1 prediction does not provide significant compression gains, (as will be verified by simulations), and, this may explain why it is not implemented in standard coding algorithms.

D. Simulation Results

To demonstrate the performance of the proposed approach we consider the scalable coding of first-order Gauss–Markov and Laplace–Markov sources. In the simulations, we used a uniform threshold quantizer with a central dead zone. Such quantizers are often used in image and video compression [16]. The rate is calculated as the first order entropy of the quantizer indices.

Results compare the performance of scalable coders with the following prediction methods at the enhancement-layer:

1) prediction using current base-layer reconstruction (P1) but using only single entropy coder;
2) prediction P1 where two conditional entropy coders are used;
3) prediction from previous enhancement-layer reconstruction (P2);
4) proposed estimation-theoretic (ET) prediction but using only a single entropy coder;
5) ET prediction where the residual is encoded with two conditional entropy coders.

The base-layer is identical in all coders, and the performance is shown for various enhancement-layer rates. Also provided for reference is the performance of a nonscalable coder at the same total rate.

1) Gauss–Markov Process: The zero-mean unit-variance Gauss–Markov process can be defined according to (11) which we repeat here

$$x_n = \rho x_{n-1} + z_n \qquad (21)$$

where x_n, and z_n are stationary zero-mean Gaussian random processes with variances 1 and $1 - \rho^2$, respectively.

High-resolution analysis for scalable coding of Gauss–Markov sequences is given in the Appendix. It provides insight into the performance difference between standard prediction methods P1 and P2, the potential for gains over them, and the circumstances under which such gains may be realized.

Fig. 6 depicts the simulation results for the compression of Gauss–Markov sequences. The signal-to-noise ratio (SNR) versus enhancement-layer rate is shown for all the competing approaches. The base-layer rate is identical in all the coders. For reference, the performance of the nonscalable coder is shown. The proposed ET prediction provides significant gains over prediction methods P1 and P2. Note that the gains saturate with increasing bit rate. These results are in agreement with the high resolution analysis of the Appendix. Note, further, that

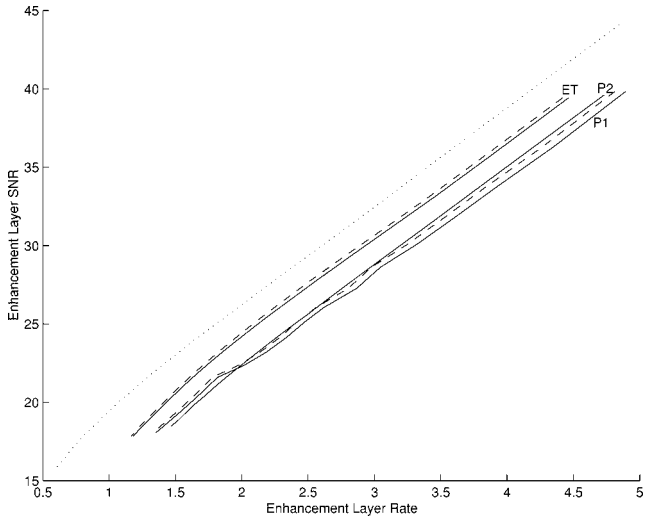

Fig. 6. Performance of two-layer scalable DPCM coding for Gauss–Markov source with $\rho = 0.99$. SNR of enhancement-layer versus enhancement-layer rate (bits/sample) is shown for different prediction methods. For P1 and ET prediction, solid lines and dashed lines show performance with single entropy coder and two entropy coders respectively. Base-layer rate was 0.59 bits/sample. Performance of nonscalable coder with the same total rate is indicated by dotted line.

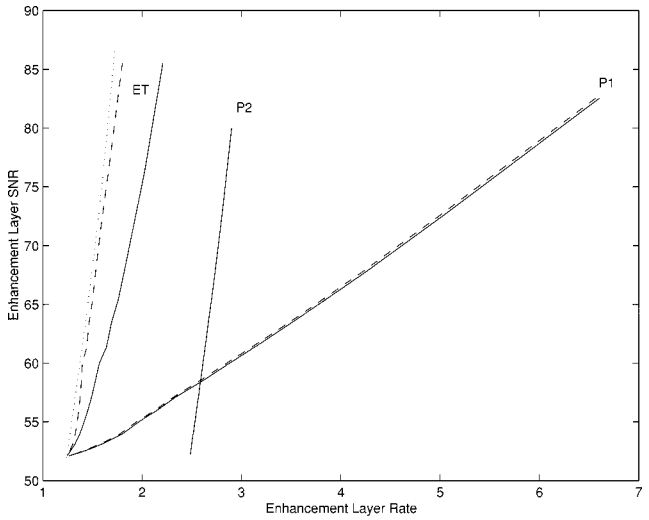

Fig. 7. Performance of two-layer scalable DPCM coding for Laplace–Markov source with $\rho = 0.95$. SNR of enhancement-layer versus enhancement-layer rate (bits/sample) is shown for different prediction methods. For P1 and ET prediction, solid lines and dashed lines show performance with single entropy coder and two entropy coders respectively. Base-layer rate was 1.14 bits/sample. Performance of nonscalable coder with the same total rate is indicated by dotted line.

the gains due to conditional entropy coding are modest for the Gauss–Markov process.

2) Laplace–Markov Process: The zero-mean unit-variance Laplace–Markov process (see e.g., [4]) is defined as the first order Markov process of (11) where the marginal density of x_n is Laplacian

$$p_x(x_n) = \frac{1}{\sqrt{2}} \mathrm{e}^{-|x_n|\sqrt{2}} \tag{22}$$

and, therefore, z_n has the distribution

$$p(z_n) = \rho^2 \delta(z_n) + (1-\rho^2) \frac{1}{\sqrt{2}} \mathrm{e}^{-|z_n|\sqrt{2}}. \tag{23}$$

Consideration of this process is motivated by the observation that speech, image and video signals possess marginal densities that are closely approximated by Laplacian densities [12], [14], [16].

We provide high-resolution analysis for scalable coding of Laplace–Markov sequences in the Appendix. Fig. 7 summarizes the simulation results for Laplace–Markov sequences. The SNR versus enhancement-layer rate is given for all the competing approaches. The base-layer rate is identical for all coders. For additional reference, the performance of the nonscalable coder is shown. As expected, P1 outperforms P2 at small ratios of enhancement to base rate, and underperforms P2 at the other extreme. It is seen that ET prediction provides substantial gains over prediction methods P1 and P2. In particular, the gain over P1 does not saturate and is asymptotically unbounded, as expected from the high-resolution analysis of the Appendix . Our intuitive explanation of the increasing gains hinges on the property of the Laplace–Markov sequence, which allows surprisingly good prediction. In particular, the innovation process density of (23) is a mixture of a Laplacian and a delta function impulse. The presence of the delta function implies that, with probability well above zero, the prediction is perfect in the absence of quantization error feedback. For a given base-layer rate, the amount of quantization noise that is fed back via the prediction loop in P1 is independent of the enhancement-layer rate. On the other hand, the quantization noise that is fed back in ET and P2 prediction decreases with increasing enhancement-layer rate. Thus, the rate-distortion curve of P1 prediction differs considerably in slope from ET and P2 prediction, and hence the large gains at high rates. We finally note that conditional entropy coding in conjunction with ET prediction provides significant additional gains and performs almost as well as nonscalable coding.

IV. Scalable Video Coder Design

A. Derivation

This section adopts the proposed ET approach for the problem of scalable video compression. We restrict our attention to predictive DCT-based coding because of its dominance in all current standards, but it should be emphasized that the approach is general and applicable to virtually any form of predictive coding.

In standard predictive video coding, the frame is divided into macroblocks. These blocks are coded either with interframe prediction (“intermode”), or without such prediction (“intramode”). Intramode coding is used infrequently and, due to the absence of time-prediction, does not represent a significant challenge for scalability. We, therefore, focus our attention on the more important and interesting case of scalable coding of intermode macroblocks.

For each intermode macroblock, a motion vector is transmitted. Note that skipping a macroblock implies that its motion vector is zero. At the base-layer, the previous reconstructed block is used as predictor. DCT is applied to the prediction error, and the resulting transform coefficients are scalar quantized, en-

tropy coded and transmitted. At the receiver, the residual is decoded and added to the prediction to form reconstruction of the current frame.

We propose to apply our estimation-theoretic (ET) paradigm for the prediction of the enhancement-layer block. We chose to implement the estimate in the transform domain, i.e., predict the transform coefficients of the current block rather than the pixels themselves as is commonly done. The motion estimation/compensation is performed in the pixel domain as usual, and the corresponding block in the enhancement-layer reconstruction of previous frame is identified. This block is transformed by DCT. The transform coefficients are combined with quantization information available, in the transform domain, from the base-layer reconstruction of the current block. This combination is performed within the ET framework to form the estimate of DCT coefficients of the current block. The prediction error between the original DCT coefficients and the estimated DCT coefficients is obtained and quantized. At the decoder, the quantized prediction error is added to the estimated DCT coefficients. An inverse DCT is applied to obtain the enhancement-layer reconstruction of the current block.

The DCT domain is more convenient for the ET predictor design because the DCT coefficients of the residual are almost uncorrelated. Further, the base-layer quantization interval of each DCT coefficient is readily available. Thus, the predictor can be independently implemented for each DCT coefficient with virtually no loss of optimality. However, one additional DCT computation is required for each block to calculate the transform coefficients in the previous enhancement-layer reconstruction. Note that implementation of the prediction in the DCT domain will produce no change in the performance of standard prediction methods, P1 and P2 or H.263+. We re-emphasize that the motion estimation/compensation for base and enhancement layers are implemented in the pixel domain as in conventional video coders. Fig. 5 provides a sketch of the proposed coder.

We assume that the evolution of a DCT coefficient in time (i.e., from frame to frame) can be modeled by the first-order Markov process

$$x_n = \rho x_{n-1} + z_n \tag{24}$$

where x_n is a DCT coefficient in the current frame and x_{n-1} is the corresponding (after motion compensation) DCT coefficient in the previous frame. The transform coefficients are zero-mean except for the DC coefficient. We assume that z_n is stationary, and independent of x_{n-1}. Note that our choice of notation is made so as to relate directly to the derivation in the previous section for the DPCM case. We now proceed in a similar fashion. While our model for interframe evolution of video is simple, it is sufficiently accurate to allow ET prediction to achieve significant gains.

The optimal base layer predictor is given by

$$\tilde{x}_n^b = E\left[x_n|\hat{x}_{n-1}^b\right] \approx \rho\hat{x}_{n-1}^b. \tag{25}$$

The base encoder quantizes the residual, $r_n^b = x_n - \tilde{x}_n^b$, and transmits index i^b. Let (a,b) be the quantization interval associated with index i^b. The optimal base-layer reconstruction is given by

$$\hat{x}_n^b = E\left[x_n|\hat{x}_{n-1}^b, x_n \in \left(\tilde{x}_n^b + a, \tilde{x}_n^b + b\right)\right]. \tag{26}$$

We note that the optimal prediction and reconstruction for the base-layer is nearly the same as that employed by standard video coding schemes. The main advantage of the estimation-theoretic approach is at the enhancement-layer.

The enhancement-layer decoder has access to $\hat{x}_{n-1}^e$, the corresponding enhancement-layer reconstructed DCT coefficient of the previous frame. The optimal enhancement-layer predictor is

$$\tilde{x}_n^e = E\left[x_n|\hat{x}_{n-1}^e, x_n \in \left(\tilde{x}_n^b + a, \tilde{x}_n^b + b\right)\right] \tag{27}$$

or

$$\tilde{x}_n^e = \rho\hat{x}_{n-1}^e + E\left[z_n|z_n \in \times\left(\tilde{x}_n^b + a - \hat{x}_{n-1}^e, \tilde{x}_n^b + b - \hat{x}_{n-1}^e\right)\right]. \tag{28}$$

Note how the ET predictor combines information from prior enhancement-layer reconstruction, and from the base-layer quantization interval. The enhancement-layer encoder quantizes the residual, $r_n^e = x_n - \tilde{x}_n^e$, and transmits index i^e. Let (c,d) be the quantization interval associated with index i^e, and let $e = \max\left[\tilde{x}_n^b + a, \tilde{x}_n^e + c\right]$ and $f = \min\left[\tilde{x}_n^b + b, \tilde{x}_n^e + d\right]$. The enhancement-layer reconstruction of the DCT coefficient is

$$\hat{x}_n^e = E\left[x_n|\hat{x}_{n-1}^e, x_n \in (e,f)\right]. \tag{29}$$

To evaluate such expectations we employ an appropriate probabilistic model for z_n, the innovation error process. It is well known that the marginal density function of the DCT coefficient may be approximated by a Laplacian distribution [16]. Hence, modeling x_n by a Laplace–Markov process, we obtain the density of z_n

$$p_{z_n}(z) = \rho^2\delta(z) + \frac{1}{2}(1-\rho^2)\alpha e^{-|z|\alpha}. \tag{30}$$

The parameters ρ and α may be estimated from a training set. We found that $\rho \approx 1$ for "low and intermediate frequency" DCT coefficients. The ET prediction consists of computing the centroid of the quantization interval (specified by the base layer) with respect to the density of (30) for each DCT coefficient. A closed form solution to the centroid computation is given in terms of the interval limits

$$E(z_n|z_n \in (s,t)) = \begin{cases} \frac{se^{-s\alpha} - te^{-t\alpha}}{e^{-s\alpha} - e^{-t\alpha}} + \frac{1}{\alpha}, & \text{if } s > 0 \\ \frac{te^{t\alpha} - se^{s\alpha}}{e^{t\alpha} - e^{s\alpha}} - \frac{1}{\alpha}, & \text{if } t < 0 \\ \frac{e^{s\alpha}(1-\alpha s) - e^{-t\alpha}(1+\alpha t)}{\alpha\left(2 - e^{s\alpha} - e^{-t\alpha} + \frac{2\rho^2}{1-\rho^2}\right)}, & \text{otherwise.} \end{cases} \tag{31}$$

Despite its imposing form, this expression is computationally benign. Therefore, the ET predictor can be implemented with a modest increase in complexity.

TABLE I
PERFORMANCE OF *TWO-LAYER* SCALABLE CODERS, WHICH DIFFER IN THEIR ENHANCEMENT-LAYER PREDICTION MODULE, AND NON-SCALABLE CODER. ENCODED SEQUENCE: *CARPHONE* AT QCIF RESOLUTION. THE ENTRIES PROVIDE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES VERSUS TOTAL RATE OF BASE AND ENHANCEMENT LAYERS (Kbps). TOTAL NUMBER OF FRAMES WAS 267 AT FRAME SKIP OF 3. FOR ALL THE METHODS, THE BASE-LAYER RATE WAS FIXED AT 32 Kbps, AND THE CORRESPONDING PSNR WAS 31.52 dB

Rate	P1	P2	H.263+	ET	Non-scalable
64	32.80	31.99	33.26	33.70	34.46
80	33.43	33.41	34.27	34.79	35.43
96	34.03	34.50	35.13	35.65	36.28
128	35.08	36.17	36.62	37.13	37.68
160	35.98	38.54	38.85	39.20	39.57

TABLE II
PERFORMANCE OF *TWO-LAYER* SCALABLE CODERS, WHICH DIFFER IN THEIR ENHANCEMENT-LAYER PREDICTION MODULE, AND NON-SCALABLE CODER. ENCODED SEQUENCE: *SALESMAN* AT QCIF RESOLUTION. THE ENTRIES PROVIDE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES VERSUS TOTAL RATE OF BASE AND ENHANCEMENT LAYERS (Kbps). TOTAL NUMBER OF FRAMES WAS 449 AT FRAME SKIP OF 3. FOR ALL THE METHODS, THE BASE-LAYER RATE WAS FIXED AT 32 Kbps, AND THE CORRESPONDING PSNR WAS 34.02 dB

Rate	P1	P2	H.263+	ET	Non-scalable
64	34.66	34.14	35.96	36.71	37.65
80	34.97	36.36	37.23	38.18	39.03
96	35.26	37.71	38.57	39.63	40.16
128	35.81	40.16	40.51	41.72	42.07
160	36.40	42.03	42.35	42.77	43.09
192	36.99	43.03	43.27	43.65	43.76

The quantization interval of any DCT coefficient can be determined from the quantized prediction error. It should be noted that the recovered quantization interval may not be accurate if thresholding is performed on the DCT coefficients at the base-layer. However, thresholding is usually less beneficial, and hence less likely to be used in scalable video coders, than single-layer coders. It is also important in the ET implementation to account for the fact that the quantization interval around origin (dead band) is larger than the other quantization intervals.

B. Simulation Results

We developed a test bed for scalable video coding by using the publicly available H.263 coder [23]. The H.263 algorithm was used for motion estimation, and for compression of the prediction error of the base and enhancement layers. The advanced motion compensation and arithmetic encoding options were turned off.

The following prediction modules for the enhancement-layer were implemented for the comparisons

1) P1 (proposed in [19]);
2) P2 (proposed in [7]);
3) an H263+ based coder;
4) proposed estimation-theoretic (ET) predictor.

TABLE III
PERFORMANCE OF *TWO-LAYER* SCALABLE CODERS, WHICH DIFFER IN THEIR ENHANCEMENT-LAYER PREDICTION MODULE. ENCODED SEQUENCE: *CARPHONE*AT QCIF RESOLUTION. THE ENTRIES PROVIDE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES VERSUS TOTAL RATE OF BASE AND ENHANCEMENT LAYERS (Kbps). TOTAL NUMBER OF FRAMES WAS 898 AT FRAME SKIP OF 3. FOR ALL THE METHODS, THE BASE-LAYER RATE WAS FIXED AT 16 Kbps, AND THE CORRESPONDING PSNR WAS 29.30 dB

Rate	P1	P2	H.263+	ET	Non-scalable
32	30.16	29.85	30.64	31.01	31.52
40	30.62	30.78	31.52	31.86	32.37
48	31.07	31.66	32.25	32.61	33.16
64	31.84	33.07	33.51	33.78	34.46
128	34.21	36.54	36.81	36.98	37.68

TABLE IV
PERFORMANCE OF *TWO-LAYER* SCALABLE CODERS, WHICH DIFFER IN THEIR ENHANCEMENT-LAYER PREDICTION MODULE. ENCODED SEQUENCE: *MOTHER–DAUGHTER* AT QCIF RESOLUTION. THE ENTRIES PROVIDE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES VERSUS TOTAL RATE OF BASE AND ENHANCEMENT LAYERS (Kbps). TOTAL NUMBER OF FRAMES WAS 898 AT FRAME SKIP OF 3. FOR ALL THE METHODS, THE BASE-LAYER RATE WAS FIXED AT 64 Kbps, AND THE CORRESPONDING PSNR WAS 34.36 dB

Rate	P1	P2	H.263+	ET	Non-scalable
128	35.11	34.83	35.47	36.05	36.97
160	35.52	36.03	36.29	36.92	37.85
192	35.92	36.94	37.10	37.67	38.51
224	36.30	37.70	37.79	38.31	39.11
256	36.65	38.30	38.40	38.86	39.63
288	36.99	38.86	38.94	39.33	40.12
320	37.32	39.31	39.40	39.76	40.54
384	37.91	40.16	40.26	40.54	41.20

The H.263+ scalable coder can choose one of three prediction modes for each macroblock [22]

1) prediction from current base-layer block;
2) prediction from previous enhancement-layer reconstruction;
3) prediction from weighted sum of current base and previous base-layer blocks.

The best prediction mode for each macroblock is sent as side information. A similar prediction mode strategy is used in the MPEG-4 [24] scalable coder.

The z_n model parameters for each DCT coefficient were estimated from a training set extracted from the *Miss America* sequence. The frame-skip was three, and we present the average PSNR of the luminance component of reconstructed frames. (Significant PSNR gains were also obtained in the chrominance components.)

Tables I–V shows the results for *two layer* scalable compression on the sequence *Carphone*. The base-layer rate was fixed at 16 Kbps and coded in an identical manner by all the methods. PSNR results for the enhancement-layer are presented for various ratios of enhancement-layer to base-layer rates. It is easily seen that the proposed ET prediction outperforms all

TABLE V
PERFORMANCE OF *TWO-LAYER* SCALABLE CODERS, WHICH DIFFER IN THEIR ENHANCEMENT-LAYER PREDICTION MODULE. ENCODED SEQUENCE: *LTS* AT QCIF RESOLUTION. THE ENTRIES PROVIDE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES VERSUS TOTAL RATE OF BASE AND ENHANCEMENT LAYERS (Kbps). TOTAL NUMBER OF FRAMES WAS 487 AT FRAME SKIP OF 3. FOR ALL THE METHODS, THE BASE-LAYER RATE WAS FIXED AT 64 Kbps, AND THE CORRESPONDING PSNR WAS 27.18 dB

Rate	P1	P2	H.263+	ET	Non-scalable
128	28.06	27.90	28.66	29.04	29.60
160	28.54	28.98	29.58	29.88	30.63
192	29.01	29.83	30.37	30.60	31.47
224	29.45	30.54	31.06	31.24	32.21
256	29.84	31.19	31.66	31.81	32.86
288	30.23	31.77	32.22	32.34	33.45
320	30.59	32.28	32.73	32.82	34.02
384	31.27	33.21	33.62	33.69	34.96

TABLE VI
PERFORMANCE OF *MULTILAYER* SCALABLE CODERS WITH DIFFERENT PREDICTION METHODS, AND NON-SCALABLE (SINGLE-LAYER) CODER. ENCODED SEQUENCE: *CARPHONE* AT QCIF RESOLUTION. TOTAL NUMBER OF FRAMES WAS 267 AT FRAME SKIP OF 3. THE ENTRIES INDICATE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES OF A LAYER VERSUS THE TOTAL RATE (IN Kbps) OF THAT LAYER. NOTE THAT THE TOTAL RATE INCLUDES RATE OF ALL THE LAYERS UP TO THIS LAYER

Layer	Rate	P1	P2	H.263+	ET	Non-scalable
1	16	29.30	29.30	29.30	29.30	29.30
2	32	30.16	29.85	30.64	31.01	31.52
3	48	30.77	29.85	31.38	31.98	33.16
4	64	31.31	29.85	31.94	32.71	34.46
5	96	32.40	31.66	33.25	34.29	36.28
6	128	33.32	31.66	34.18	35.43	37.68
7	192	35.18	34.14	36.14	37.63	39.57
8	256	36.71	34.14	37.69	39.12	41.01

the competing approaches, and achieves substantial gains in reconstructed PSNR of the enhancement layer. As expected (see Section III-B), P1 outperforms P2 at small ratios of enhancement to base rate, and underperforms P2 at the other extreme. The H.263+ predictor outperforms P1 and P2 and gradually approaches the performance of the proposed ET predictor at high enhancement layer rates.

Tables VI–XIII show the performance for *multilayer* scalable coding on several video sequences of "video conference" type as well as "nonvideo conference" type. The base layer is identically encoded for all competing methods as is evident from the first row of the Tables. Note that the ET predictor substantially outperforms the other coders. It is important to emphasize how the prediction gains of ET build up with the number of layers and, in most cases, result in major performance improvements. In all cases, the ET predictor provided gains between 0.5 dB and 1.9 dB over H.263+ based prediction at the higher layers, and much larger gains over P1 and P2.

For rough evaluation of the complexity costs of ET prediction, we recorded the execution time in the experiment. The

TABLE VII
PERFORMANCE OF *MULTILAYER* SCALABLE CODERS WITH DIFFERENT PREDICTION METHODS, AND NON-SCALABLE (SINGLE-LAYER) CODER. ENCODED SEQUENCE *SALESMAN* AT QCIF RESOLUTION. TOTAL NUMBER OF FRAMES WAS 449 AT FRAME SKIP OF 3. THE ENTRIES INDICATE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES OF A LAYER VERSUS THE TOTAL RATE (IN Kbps) OF THAT LAYER. NOTE THAT THE TOTAL RATE INCLUDES RATE OF ALL THE LAYERS UP TO THIS LAYER

Layer	Rate	P1	P2	H.263+	ET	Non-scalable
1	16	31.28	31.28	31.28	31.28	31.28
2	32	31.70	31.47	32.33	32.87	34.02
3	48	32.02	31.47	33.09	34.16	36.22
4	64	32.30	31.47	33.97	35.01	37.65
5	96	32.86	34.08	35.63	37.59	40.16
6	128	33.43	34.08	37.32	39.30	42.07
7	192	34.57	37.60	40.10	42.08	43.09
8	256	35.72	37.60	42.18	43.62	44.99

TABLE VIII
PERFORMANCE OF *MULTILAYER* SCALABLE CODERS WITH DIFFERENT PREDICTION METHODS, AND NON-SCALABLE (SINGLE-LAYER) CODER. ENCODED SEQUENCE *CONTAINER* AT QCIF RESOLUTION. TOTAL NUMBER OF FRAMES WAS 300 AT FRAME SKIP OF 3. THE ENTRIES INDICATE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES OF A LAYER VERSUS THE TOTAL RATE (IN Kbps) OF THAT LAYER. NOTE THAT THE TOTAL RATE INCLUDES RATE OF ALL THE LAYERS UP TO THIS LAYER

Layer	Rate	P1	P2	H.263+	ET	Non-scalable
1	16	31.18	31.18	31.18	31.18	31.18
2	32	31.73	31.46	32.07	32.85	33.89
3	48	32.20	31.46	32.77	33.91	35.54
4	64	32.60	31.46	33.30	34.68	36.72
5	96	33.38	33.80	34.76	36.43	38.73
6	128	34.12	33.80	35.76	37.69	39.93
7	192	35.62	36.54	38.10	39.82	41.70
8	256	36.94	36.54	39.51	41.28	43.26

TABLE IX
PERFORMANCE OF *MULTILAYER* SCALABLE CODERS WITH DIFFERENT PREDICTION METHODS, AND NON-SCALABLE (SINGLE-LAYER) CODER. ENCODED SEQUENCE *HALL-OBJECTS* AT QCIF RESOLUTION. TOTAL NUMBER OF FRAMES WAS 330 AT FRAME SKIP OF 3. THE ENTRIES INDICATE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES OF A LAYER VERSUS THE TOTAL RATE (IN Kbps) OF THAT LAYER. NOTE THAT THE TOTAL RATE INCLUDES RATE OF ALL THE LAYERS UP TO THIS LAYER

Layer	Rate	Conventional Pred.		H.263+ based	ET	
Layer	Rate	P1	P2	H.263+	ET	Non-scalable
1	16	32.50	32.50	32.50	32.50	32.50
2	32	33.31	32.64	34.13	34.70	34.87
3	48	33.83	32.64	35.17	36.16	36.88
4	64	34.26	32.64	36.11	37.27	38.28
5	96	35.10	34.95	37.94	39.41	39.86
6	128	35.84	34.95	39.08	40.50	41.25
7	192	37.31	38.36	40.87	42.06	42.53
8	256	38.56	38.36	41.92	43.01	43.15

TABLE X
PERFORMANCE OF *MULTILAYER* SCALABLE CODERS WITH DIFFERENT PREDICTION METHODS, AND NON-SCALABLE (SINGLE-LAYER) CODER. ENCODED SEQUENCE *COASTGUARD* AT QCIF RESOLUTION. TOTAL NUMBER OF FRAMES WAS 300 AT FRAME SKIP OF 3. THE ENTRIES INDICATE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES OF A LAYER VERSUS THE TOTAL RATE (IN Kbps) OF THAT LAYER. NOTE THAT THE TOTAL RATE INCLUDES RATE OF ALL THE LAYERS UP TO THIS LAYER

Layer	Rate	P1	P2	H.263+	ET	Non-scalable
1	16	25.94	25.94	25.94	25.94	25.94
2	32	26.61	26.25	27.00	27.14	27.63
3	48	27.12	26.25	27.58	27.81	28.97
4	64	27.57	26.25	28.03	28.33	29.98
5	96	28.49	27.69	29.12	29.46	31.49
6	128	29.28	27.69	29.89	30.29	32.74
7	192	30.91	29.47	31.59	31.96	34.69
8	256	32.22	29.47	32.86	33.18	36.23

TABLE XI
PERFORMANCE OF *MULTILAYER* SCALABLE CODERS, WITH DIFFERENT PREDICTION METHODS, AND NON-SCALABLE (SINGLE-LAYER) CODER. ENCODED SEQUENCE *GRANDMA* AT QCIF RESOLUTION. TOTAL NUMBER OF FRAMES WAS 869 AT FRAME SKIP OF 3. THE ENTRIES INDICATE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES OF A LAYER VERSUS THE TOTAL RATE (IN Kbps) OF THAT LAYER. NOTE THAT THE TOTAL RATE INCLUDES RATE OF ALL THE LAYERS UP TO THIS LAYER

Layer	Rate	P1	P2	H.263+	ET	Non-scalable
1	16	34.48	34.48	34.48	34.48	34.48
2	32	34.88	34.66	35.28	35.87	36.85
3	48	35.22	34.66	36.04	36.80	38.59
4	64	35.52	34.66	36.71	37.65	39.75
5	96	36.15	36.83	38.21	39.27	41.39
6	128	36.78	36.83	39.26	40.52	42.60
7	192	37.93	39.60	41.16	42.42	43.86
8	256	39.02	39.60	42.39	43.49	45.26

TABLE XII
PERFORMANCE OF *MULTILAYER* SCALABLE CODERS, WITH DIFFERENT PREDICTION METHODS, AND NON-SCALABLE (SINGLE-LAYER) CODER. ENCODED SEQUENCE *MOTHER–DAUGHTER* AT QCIF RESOLUTION. TOTAL NUMBER OF FRAMES WAS 869 AT FRAME SKIP OF 3. THE ENTRIES INDICATE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES OF A LAYER VERSUS THE TOTAL RATE (IN Kbps) OF THAT LAYER. NOTE THAT THE TOTAL RATE INCLUDES RATE OF ALL THE LAYERS UP TO THIS LAYER

Layer	Rate	P1	P2	H.263+	ET	Non-scalable
1	64	34.36	34.36	34.36	34.36	34.36
2	128	35.11	34.83	35.47	36.05	36.97
3	192	35.71	34.83	36.22	37.07	38.51
4	288	36.55	36.03	37.31	38.40	40.12
5	384	37.31	36.03	38.17	39.36	41.20

overall complexity was observed to increase by 10% relative to that of H.263+.

TABLE XIII
PERFORMANCE OF *MULTILAYER* SCALABLE CODERS, WITH DIFFERENT PREDICTION METHODS, AND NON-SCALABLE (SINGLE-LAYER) CODER. ENCODED SEQUENCE *LTS* AT QCIF RESOLUTION. TOTAL NUMBER OF FRAMES WAS 487 AT FRAME SKIP OF 3. THE ENTRIES INDICATE THE AVERAGE PSNR (IN dB) OF RECONSTRUCTED FRAMES OF A LAYER VERSUS THE TOTAL RATE (IN Kbps) OF THAT LAYER. NOTE THAT THE TOTAL RATE INCLUDES RATE OF ALL THE LAYERS UP TO THIS LAYER

Layer	Rate	P1	P2	H.263+	ET	Non-scalable
1	64	27.18	27.18	27.18	27.18	27.18
2	128	28.06	27.90	28.66	29.04	29.60
3	192	28.70	27.90	29.39	30.02	31.47
4	288	29.63	28.98	30.45	31.33	33.45
5	384	30.41	28.98	31.24	32.30	34.96

V. SUMMARY AND CONCLUSIONS

This paper presents a new approach to optimal scalability in predictive coding. The predictor is designed within an estimation-theoretic framework. The current sample prediction is optimal given both the past enhancement layer reconstruction and all base-layer parameters and variables including the reconstruction and quantization interval. To emphasize its generality, the approach was first derived for the simple case of scalable DPCM systems. Its potential was then demonstrated on the application of scalable video coding. Simulation results show that the proposed scalable coding technique offers substantial performance gains over conventional approaches over a wide range of bit rates. The gains increase with the number of layers.

Although ET prediction was applied to video coding here in conjunction with standard DCT-based coding systems, it is easily extendible to subband-based, and pixel-domain coders, as is evident from the basic DPCM derivation. Work in progress shows that ET prediction has applications in error concealment for scalable video coding [21] and in scalable coding of stereophonic (two-channel) audio [1]. Note that we have only considered conditional entropy encoding of the enhancement-layer residual for the DPCM case. It has produced substantial gain in the case of Laplacian–Markov sources, and achieved performance close to that of nonscalable coding. The extension to conditional entropy coding of DCT coefficients in scalable video coders is a topic that deserves further study.

APPENDIX
HIGH-RESOLUTION ANALYSIS

A. Gauss–Markov Sequences

We derive asymptotic (high-rate) results for scalable DPCM coding of a Gauss–Markov sequence. We first review the corresponding nonscalable results [4]. For the Gauss–Markov source, the prediction error possesses a normal density. If h_e and σ_e^2 are the differential entropy and variance of prediction error, we have

$$h_e = \frac{1}{2}\log 2\pi e\sigma_e^2. \tag{32}$$

If the prediction error is encoded by an optimal quantizer whose output entropy is R, the quantization distortion D (by the Gish-Pierce result [8]) is

$$D = \frac{1}{12} 2^{2(h_e - R)}. \tag{33}$$

We also have that

$$\sigma_e^2 = \sigma_z^2 + \rho^2 D_p \tag{34}$$

where σ_z^2 is the variance of the innovation process (see (11)), and D_p is the quantization distortion in the previous reconstructed sample (used as predictor).

For a nonscalable coder, $D_p = D$, and hence, (32), (33) and (34) may be combined to yield

$$D = \frac{\frac{\pi e}{6} 2^{-2R} \sigma_z^2}{1 - \frac{\pi e}{6} \rho^2 2^{-2R}}. \tag{35}$$

For a scalable coder, let R_b and $R_e - R_b$ be the entropy of the quantizer output at the base and enhancement layers respectively. The reconstruction distortion for the base-layer is given by

$$D^b = \frac{\frac{\pi e}{6} 2^{-2R_b} \sigma_z^2}{1 - \frac{\pi e}{6} \rho^2 2^{-2R_b}}. \tag{36}$$

For prediction method P1, the reconstruction error of the base-layer is encoded by a quantizer of entropy $R_e - R_b$ at the enhancement-layer. Thus, the distortion at the enhancement-layer is

$$D^e_{P1} = D_b 2^{-2(R_e - R_b)} = \frac{\frac{\pi e}{6} 2^{-2R_e} \sigma_z^2}{1 - \frac{\pi e}{6} \rho^2 2^{-2R_b}}. \tag{37}$$

For prediction method P2, the previous enhancement-layer reconstruction is used as the estimate and the prediction error encoded by a quantizer of entropy $R_e - R_b$. The corresponding distortion is given by

$$D^e_{P2} = \frac{\frac{\pi e}{6} 2^{-2(R_e - R_b)} \sigma_z^2}{1 - \frac{\pi e}{6} \rho^2 2^{-2(R_e - R_b)}}. \tag{38}$$

The potential gains over P1 and P2 are limited by the performance of the single-layer coder of rate R_e. It is therefore of interest to consider this bound on the performance of the ET predictor as an indicator of circumstances under which large gains may be recouped. The corresponding distortion is given by

$$D^e_{\min} = \frac{\frac{\pi e}{6} 2^{-2R_e} \sigma_z^2}{1 - \frac{\pi e}{6} \rho^2 2^{-2R_e}}. \tag{39}$$

It is easy to see that the potential performance gains over P1 increase with R_e, and decrease as $\rho \to 0$. Further, the gains decrease with increasing R_b. The potential performance gains over P2 increase with R_b and decrease with increasing R_e. Further, as $R_e \to \infty$, all the distortion curves (in log scale) become parallel with slope 2. These asymptotic results accurately predict the simulation results presented in Fig. 6.

B. Laplace–Markov Sequences

Here, we derive the corresponding asymptotic results for the scalable DPCM coding of Laplace–Markov process. Consider a memory-less source whose pdf is given by (23) and let h_l denote the differential entropy of the continuous (Laplacian) component. If this source is encoded by a quantizer of output entropy R, the resulting distortion is given by [4]

$$D_{PCM} = \frac{1 - \rho^2}{12} 2^{\frac{2(h_l - (R - \mathcal{H}(\rho^2)))}{1 - \rho^2}} \tag{40}$$

where $\mathcal{H}(\cdot)$ is the binary entropy function

$$\mathcal{H}(\alpha) = -\alpha \log(\alpha) - (1-\alpha)\log(1-\alpha), \quad 0 \le \alpha \le 1. \tag{41}$$

For nonscalable DPCM coding of Laplace–Markov sequence, we have

$$D = D_{PCM} + \rho^4 D \tag{42}$$

or equivalently

$$D = \frac{1}{12(1+\rho^2)} 2^{\frac{2(h_l - (R - \mathcal{H}(\rho^2)))}{1 - \rho^2}}. \tag{43}$$

The differential entropy is given by

$$h_l = \frac{1}{2} \log\left(2e^2 \sigma_e^2\right) = \frac{1}{2} \log\left(2e^2 \left(\sigma_w^2 + \rho^2 D\right)\right). \tag{44}$$

From(43) and (44), it follows that

$$D = \frac{\frac{e^2}{6(1+\rho^2)} 2^{-2\frac{(R - \mathcal{H}(\rho^2))}{(1-\rho^2)}}}{1 - \frac{e^2 \rho^2}{6(1+\rho^2)} 2^{-2\frac{(R - \mathcal{H}(\rho^2))}{(1-\rho^2)}}}. \tag{45}$$

Consider a two-layer scalable coder with base and enhancement layer rates of R_b and $R_e - R_b$. For the base-layer, the reconstruction distortion is given by

$$D^b = \frac{\frac{e^2}{6(1+\rho^2)} 2^{-2\frac{(R_b - \mathcal{H}(\rho^2))}{(1-\rho^2)}}}{1 - \frac{e^2 \rho^2}{6(1+\rho^2)} 2^{-2\frac{(R_b - \mathcal{H}(\rho^2))}{(1-\rho^2)}}}. \tag{46}$$

The enhancement-layer distortion for P1 is given by

$$D^e_{P1} = \frac{\frac{e^2}{6(1+\rho^2)} 2^{-2\frac{(R - \mathcal{H}(\rho^2))}{(1-\rho^2)}} 2^{-2(R_e - R_b)}}{1 - \frac{e^2 \rho^2}{6(1+\rho^2)} 2^{-2\frac{(R - \mathcal{H}(\rho^2))}{(1-\rho^2)}}}. \tag{47}$$

For prediction method P2, the corresponding distortion is

$$D^e_{P2} = \frac{\frac{e^2}{6(1+\rho^2)} 2^{-2\frac{(R_e - R_b - \mathcal{H}(\rho^2))}{(1-\rho^2)}}}{1 - \frac{e^2 \rho^2}{6(1+\rho^2)} 2^{-2\frac{(R_e - R_b - \mathcal{H}(\rho^2))}{(1-\rho^2)}}}. \tag{48}$$

The gains that can be obtained by ET prediction are bounded by the performance of a single-layer coder operating at rate R_e. The corresponding distortion is given by

$$D^e_{\min} = \frac{\frac{e^2}{6(1+\rho^2)} 2^{-2\frac{(R_e - \mathcal{H}(\rho^2))}{(1-\rho^2)}}}{1 - \frac{e^2 \rho^2}{6(1+\rho^2)} 2^{-2\frac{(R_e - \mathcal{H}(\rho^2))}{(1-\rho^2)}}}. \tag{49}$$

Again, the potential performance gains over P1 increase with R_e, decrease as $\rho \to 0$, and decrease with increasing R_b. Similarly, the potential performance gains over P2 decrease with increasing R_e and increase with increasing R_b. Further, as $R_e \to \infty$, the distortion curves for P2 and the (nonscalable) bound (in

log scale) become parallel with slope $2/(1-\rho^2)$. However, for the Laplace–Markov process, it is important to note that the distortion curve of P1 decays much more *slowly*, i.e., with slope 2. Thus the gains over P1 never saturate. The simulation results presented in Fig. 7 verify these asymptotic results.

REFERENCES

[1] A. Aggarwal, S. L. Regunathan, and K. Rose, "Optimal prediction in scalable coding of stereophonic audio," in *Proc. 109th AES Conv.*, 2000.

[2] R. Aravind, M. Civanlar, and A. Reibman, "Packet loss resilience of MPEG-2 scalable video coding algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 426–435, Oct. 1996.

[3] J. F. Arnold, M. R. Fracter, and Y. Wang, "Efficient drift-free signal-to-noise ratio scalability," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 70–82, Feb. 2000.

[4] N. Farvardin and J. W. Modestino, "Rate-distortion performance of DPCM schemes for autoregressive sources," *IEEE Trans. Inform. Theory*, vol. 31, pp. 402–18, May 1985.

[5] M. Ghanbari and V. Seferidis, "Efficient H.261-based two-layer video codecs for ATM Networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, pp. 171–175, Apr. 1995.

[6] A. Gersho, "Stochastic stability of delta modulation," *Bell Syst. Tech. J.*, vol. 51, pp. 821–842, Apr. 1972.

[7] B. Girod, U. Horn, and B. Belzer, "Scalable video coding with multiscale motion compensation and unequal error protection," in *Multimedia Communications and Video Coding*, Y. Wang, S. Panwar, S.-P. Kim, and H. L. Bertoni, Eds. New York: Plenum, 1996, pp. 475–482.

[8] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 676–683, Sept. 1968.

[9] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to MPEG-2*. New York: Chapman & Hall, 1997.

[10] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewoods Cliffs, NJ: Prentice-Hall, 1984.

[11] J. C. Kieffer, "Stochastic stability for feedback quantization schemes," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 248–54, Mar. 1982.

[12] M. D. Paez and T. H. Glisson, "Minimum mean-squared-error quantization in speech PCM and DPCM systems," *IEEE Trans. Commun.*, vol. COM-20, pp. 225–230, Apr. 1972.

[13] A. Puri and A. Wong, "Spatial domain resolution scalable video coding," in *Proc. SPIE*, vol. 2094, 1993, pp. 718–729.

[14] R. C. Reininger and J. D. Gibson, "Distributions of the two-dimensional DCT coefficients for images," *IEEE Trans. Commun.*, vol. COM-31, pp. 835–839, June 1983.

[15] K. Rose and S. L. Regunathan, "Toward optimal scalability in predictive video coding," in *Proc. IEEE Int. Conf. Image Processing*, 1998.

[16] G. J. Sullivan, "Efficient scalar quantization of exponential and Laplacian random variables," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1365–1374, Sept. 1996.

[17] T. K. Tan, K. K. Pang, and K. N. Ngan, "A frequency scalable coding scheme employing pyramid and subband techniques," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 4, pp. 203–207, Apr. 1994.

[18] D. Taubman and A. Zakhor, "Multirate 3-D subband coding of images," *IEEE Trans. Image Processing*, pp. 572–588, Sept. 1994.

[19] D. Wilson and M. Ghanbari, "Transmission of SNR scalable two layer MPEG-2 coded video through ATM networks," in *Proc. 7th Int. Workshop Packet Video*, Mar. 1996, pp. 185–189.

[20] ——, "Optimization of two-layer SNR scalability for MPEG-2 video," in *1997 IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1997, pp. 2637–2640.

[21] R. Zhang, S. L. Regunathan, and K. Rose, "Optimal frequency-domain error concealment for the enhancement layer in scalable video coding," in *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, 2000.

[22] "Video coding for low bit rate communication draft ITU-T Recommendation H.263 Version 2 (H.263+)," ITU Telecommunications Standardization Sector, 1997.

[23] "TMN (H.263) coder version 2.0," Telenor R&D, Norway, 1996.

[24] "Coding of audio-visual objects: Video," MPEG-4 Video group, JTC1/SC29/WG11, 1999.

Kenneth Rose (S'85–M'91) received the B.Sc. (summa cum laude) and M.Sc. (magna cum laude) degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1983 and 1987, respectively, and the Ph.D. degree in electrical engineering from the California Institute of Technology, Pasadena, in 1991.

From July 1983 to July 1988, he was with Tadiran, Ltd., Israel, where he carried out research in the areas of image coding, image transmission through noisy channels, and general image processing. In January 1991, he joined the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently a Professor. His research interests are in information theory, source and channel coding, image coding and processing, speech and general pattern recognization, and nonconvex optimization in general.

Dr. Rose is currently Editor for Source/Channel Coding for the IEEE TRANSACTIONS ON COMMUNICATIONS. He is co-recipient of the William R. Bennett Prize Paper Award of the IEEE Communications Society (1990).

Shankar L. Regunathan (S'95) received the B.Tech degree in electrical and communication engineering from the Indian Institute of Technology, Madras, in 1994, and the M.S. and Ph.D. degree in electrical engineering from the University of California, Santa Barbara, in 1996 and 2001, respectively.

He is with the Digital Media Division, Microsoft Corporation.