A UNIFIED FRAMEWORK FOR SPECTRAL DOMAIN PREDICTION AND END-TO-END DISTORTION ESTIMATION IN SCALABLE VIDEO CODING

Jingning Han, Vinay Melkote, and Kenneth Rose*

Department of Electrical and Computer Engineering, University of California Santa Barbara, CA 93106
{jingning,melkote,rose}@ece.ucsb.edu

ABSTRACT

A novel scalable coding approach is proposed for video transmission over lossy networks, which builds on two estimation-theoretic (ET) paradigms previously developed by our group: (1) an ET approach to enhancement layer prediction in scalable video coding (ET-SVC) that optimally combines all available information from both the current base layer and prior enhancement layer frames, and (2) the spectral coefficient-wise optimal recursive estimate (SCORE) of end-to-end distortion. SCORE provides the encoder with an estimate of distortion per decoder-reconstructed transform coefficient, accounting for the effects of quantization, concealment, packet loss and error propagation via the prediction loop. The current work significantly extends the scope of SCORE to encompass the setting of ET-SVC, whose prediction involves non-linear operations. This advance enables optimization of ET-SVC systems for transmission over lossy networks, thereby combining optimal prediction with optimal mode decisions at the enhancement layer. Experiments first demonstrate the estimation accuracy of SCORE in the settings of the ET-SVC coder. They then show considerable gains when SCORE is incorporated into ET-SVC to optimize encoding decisions under a wide range of packet loss and bit rates.

***Index Terms*—** Scalable video coding, error resilience, end-to-end distortion estimate, optimal prediction

1. INTRODUCTION

In scalable video coding (SVC) the base layer consists of information about the video sequence that can be decoded independently to obtain a reconstruction of coarse quality. Enhancement layers' information allows a decoder to successively refine the reconstruction. Enhancement layer packets may be dropped as necessary to adjust the transmission rate while still retaining a baseline decoding quality, and thus SVC obviates the need of generating and/or storing redundant versions of compressed video to accommodate at various bit-rates. Further, rate adjustment decisions can be made on the fly at intermediate network nodes. Thus SVC is suitable for applications that cater to receivers with diverse reception bandwidths or deployed over networks with diverse communication capabilities [1]. Throughout this paper, for exposition simplicity, we consider a two-layer quality-scalable bit-stream, although the proposed concepts are extensible to more layers and other types of scalability.

Base layer encoding is essentially the same as single layer video coding, and macroblocks are typically encoded after motion compensated temporal prediction. Prediction at the enhancement layer, however, has access to more information than motion compensated enhancement layer reconstruction of the prior frame, as it may exploit current frame information from the base layer. Standard approaches perform the enhancement layer prediction in the pixel domain and are inherently suboptimal: they cannot fully exploit information from both base and enhancement layers (more on this in Sec. 2.1). As an alternative, an optimal enhancement layer prediction approach was proposed by our group in [2], where the enhancement layer motion compensated reference is optimally combined with base layer quantization information, in a suitably derived estimation-theoretic (ET) framework, directly in the transform domain. This approach, which we henceforth refer to as ET-SVC, provides significant coding gains compared to current and prior standard pixel-domain enhancement layer prediction methods.

Practical deployment of video codecs often requires careful consideration of the impact of subsequent transmission over lossy packet networks. Errors due to packet losses propagate via the prediction loop, and can significantly affect the reconstruction quality. A major strategy to achieve error resilience is to judiciously select the prediction mode at the encoder (e.g., intra- or inter- in case of the base layer, or inter-frame or inter-layer in case of the enhancement layer) so that the end-to-end distortion (EED) versus rate tradeoff is optimized. EED measures the distortion in the decoder reconstruction, and includes the effects of quantization, packet loss and concealment at the decoder. Estimating EED at the encoder is central to optimize its decisions. The recursive optimal per-pixel estimate (ROPE) [3] is an optimal EED estimation method that recursively calculates the first and second moments of reconstructed *pixels* via update equations which explicitly account for motion-compensated/inter-layer prediction, packet loss rate, concealment, etc. In [4, 5] ROPE was employed to optimize encoding decisions in standard SVC coders, and achieved significant performance gains.

However, ROPE's applicability is inherently limited to account for operations that are recursive in the pixel domain. While this is sufficient for standard (suboptimal) SVC coders, the ET-SVC approach achieves optimality by performing its prediction directly in the transform domain. Thus an error-resilient variant of ET-SVC for transmission over lossy networks would greatly benefit from a ROPE-like EED estimate that accounts for transfrom domain operations. The *spectral coefficient-wise optimal recursive estimate* (SCORE) [6] recently proposed by us for single layer video coding is exactly the tool for this purpose. SCORE recursively computes up to second moments of decoder reconstructed *transform coefficients*, in rough analogy to what ROPE does per-pixel in the spatial domain. This work extends the scope of SCORE to encompass ET-SVC prediction. In particular, the non-linear recursive transform domain operation inherent to ET-SVC prediction is incorporated into the SCORE update equations via a quadratic approximation. Experiments first demonstrate the estimation accuracy of such extended

*Vinay Melkote is now with Dolby Laboratories Inc., 100 Potrero Avenue, San Francisco, CA 94103

SCORE in conjunction with ET-SVC. Subsequently, coding mode decisions in the ET-SVC scheme are optimized while exploiting the accurate EED estimates provided by SCORE. The proposed overall ET-SVC-SCORE coder substantially outperforms standard (pixel domain) SVC optimized by ROPE, as well as "regular" ET-SVC that incorporates no rate-EED optimization, across a broad range of packet loss and bit rates.

2. RELATED BACKGROUND

2.1. Standard Scalable Video Coding Methods

The H.264/SVC coder compresses the base layer as a single bitstream, and employs a single-loop design to code the enhancement layer, where the decoder need not buffer its base layer reconstruction to produce the enhancement layer signal. Particularly, the enhancement layer coder starts with motion compensation from previously reconstructed frames in the same layer to generate a prediction residual block. It then adaptively decides whether to further subtract the base layer reconstructed prediction error from this residual block before transformation and quantization (see [1, 7] for details). In earlier standards such as H.263++ the enhancement layer prediction switches between prior enhancement layer motion compensated reference and current base layer reconstruction, or a linear combination thereof, in what is referred to as a multi-loop design. It has been recognized that multi-loop design performs better than single-loop at the expense of more complexity [7]. ROPE has been successfully incorporated in existing SVC schemes to optimize encoding decisions for better end-to-end coding efficiency [4, 5].

2.2. Estimation-Theoretic Enhancement Layer Prediction

In [2] an ET approach for optimal prediction at the enhancement layer was proposed which we briefly describe here. Let x_n denote the value of a particular transform coefficient in a block of the current frame. For any unitary transform, one may equivalently calculate the residual in the spatial or the transform domain. Let $\hat{x}^b_{n-1}$ denote the reconstructed transform coefficient of the same frequency as x_n, but of the base layer motion compensated reference. Thus the operation of the standard base layer encoder is equivalent to quantization of $x_n - \hat{x}^b_{n-1}$ to produce the index i^b_n. Let $[a_n, b_n)$ be the quantization interval associated with index i^b_n. Thus, $x_n \in \mathcal{I}^b_n = [\hat{x}^b_{n-1} + a_n, \hat{x}^b_{n-1} + b_n)$, i.e., all the information on x_n provided by the base layer is captured by specifying the interval in which it must reside.

When encoding the enhancement layer of x_n, the encoder may access enhancement layer information from previous frames. Specifically, it has access to transform coefficient $\hat{x}^e_{n-1}$ of the motion compensated reference block. In [2], an approach is proposed to combine the prior enhancement layer information $\hat{x}^e_{n-1}$, with the base layer interval $\mathcal{I}^b_n$ to obtain the optimal enhancement layer prediction for the coefficient x_n. Note that although the enhancement layer information $\hat{x}^e_{n-1}$, which is a reconstruction value, can be equivalently obtained in the spatial pixel domain, the quantization interval $\mathcal{I}^b_n$ does not simply map to the spatial domain. Thus, ad hoc spatial domain linear combinations of base layer residual or reconstruction, with prior enhancement layer reconstruction, as employed by current and prior standard SVCs cannot achieve optimal enhancement layer prediction.

Traditionally, blocks of pixels along the same motion trajectory in consecutive video frames are modeled as an autoregressive (AR) process. Motion compensation is employed to align these pixel blocks, and pixel domain subtraction (prediction) removes temporal redundancies. In [2], the equivalent viewpoint (assuming unitary transform), that corresponding blocks of DCT coefficients form an AR process, is adopted. Thus x_n (at any given frequency) and the corresponding motion-compensated reference transform coefficient x_{n-1} conform to the first order AR model: $x_n = \rho x_{n-1} + z_n$, where z_n are independent and identically distributed (i.i.d) innovations of the process with probability density function (pdf) $p_Z(z)$. To mimic what is implicitly assumed by pixel domain motion-compensated prediction, we will arbitrarily assume here the maximum correlation coefficient $\rho = 1$ at all frequencies. The above transform domain AR process perspective provides the advantage that the motion compensation $\hat{x}^e_{n-1}$, and the quantization interval $\mathcal{I}^b_n$, can now be combined to produce the optimal estimate.

Assuming that $\hat{x}^e_{n-1} \approx x_{n-1}$, we obtain the conditional pdf $p(x_n|\hat{x}^e_{n-1}) \approx p_Z(x_n - \hat{x}^e_{n-1})$. In the absence of additional base layer information, the best prediction of x_n would just be $\hat{x}^e_{n-1}$, the default enhancement layer estimate. But the base layer indicates that $x_n \in \mathcal{I}^b_n$, which refines the conditional pdf of x_n to

$$p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) \approx \begin{cases} \frac{p_Z(x_n - \hat{x}^e_{n-1})}{\int_{\mathcal{I}^b_n} p_Z(x_n - \hat{x}^e_{n-1}) dx_n} & x_n \in \mathcal{I}^b_n \\ 0 & else \end{cases} \quad (1)$$

Note that the above is equivalent to centering the Laplacian pdf at $\hat{x}^e_{n-1}$, restricting it to (intersecting it with) the interval $\mathcal{I}^b_n$ (a highly non-linear operation), and then renormalizing to obtain a pdf. The optimal predictor $\hat{\hat{x}}^e_n$ at the enhancement layer is given by [2]

$$\hat{\hat{x}}^e_n = E[x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n] \quad (2)$$

the centroid of the above pdf in the interval $\mathcal{I}^b_n$. The residual $x_n - \hat{\hat{x}}^e_n$ is then quantized and encoded in the enhancement layer. We refer to the scalable video coder which incorporates this ET prediction as ET-SVC.

In simulations, but without loss of theoretical generality, we will assume that the innovation pdf is Laplacian, i.e., $p_Z(z_n) = \frac{1}{2}\lambda e^{-\lambda|z_n|}$, where the parameter λ is itself frequency dependent. Laplacian innovations offer an easily derived closed form of the above expectation [2].

3. SPECTRAL DOMAIN DISTORTION ESTIMATION IN SCALABLE VIDEO CODING

An important class of SVC applications involves transmission over lossy networks, where error-resilience is a crucial requirement. Errors due to enhancement layer packet losses generally propagate through the enhancement layer prediction loop. Thus, a natural approach to achieve error-resilience is to provide an option to occasionally cut off temporal prediction at the enhancement layer, via a prediction mode that is solely based on current base layer information. This section proposes a new variant of ET-SVC coder where the enhancement layer prediction is switched between two modes - the ET prediction of Sec. 2.2, and base-layer only prediction (also called inter-layer prediction) where $\hat{\hat{x}}^e_n = \hat{x}^b_n$. Encoding mode decisions are made to optimize the rate-distortion tradeoff, and critically depend on accurate estimation of EED. For this purpose we extend the SCORE approach [6] to the scalable setting. We assume a guaranteed base layer transmission, and focus exclusively on the optimization of the enhancement layer prediction mode. Thus, if an enhancement layer packet is dropped, the decoder uses the corresponding base layer reconstruction as concealment.

Let $x_n^{k,m}$ denote the original value of transform coefficient m in block k of frame n, $\hat{x}_{n,b}^{k,m}$ the base layer reconstruction, and $\hat{x}_{n,e}^{k,m}$ and $\tilde{x}_{n,e}^{k,m}$ the encoder and decoder enhancement layer reconstructions of this coefficient, respectively. Similarly, let $\hat{r}_{n,e}^{k,m}$ denote the quantized transform domain enhancement layer prediction residual, whose value is encoded and transmitted to the decoder. The enhancement layer motion compensated reference for this block is potentially "off-grid" in the prior frame. Let $u_n^{k,m}$ denote the original value of coefficient m in this (possibly off-grid) reference block. The encoder and decoder reconstructions of this coefficient are consistently denoted $\hat{u}_{n,e}^{k,m}$ and $\tilde{u}_{n,e}^{k,m}$. As far as the encoder is concerned $\tilde{x}_{n,e}^{k,m}$ and $\tilde{u}_{n,e}^{k,m}$ are random variables, due to stochastic loss in the channel. Thus the encoder *estimates* the expected enhancement layer distortion at this transform coefficient as

$$\begin{aligned}\delta_n^{k,m} &= E\{(x_n^{k,m} - \tilde{x}_{n,e}^{k,m})^2\} \\ &= (x_n^{k,m})^2 - 2x_n^{k,m}E\{\tilde{x}_{n,e}^{k,m}\} + E\{(\tilde{x}_{n,e}^{k,m})^2\}, \end{aligned} \quad (3)$$

where expectation is over packet loss events. The computation of $\delta_n^{k,m}$ only requires the first and second moments of the decoder reconstruction $\tilde{x}_{n,e}^{k,m}$ at the enhancement layer. SCORE recursively evaluates these moments for every transform coefficient in the frame, where update equations depend on the prediction mode.

Inter-Layer Prediction Mode: The packet containing transform coefficient prediction residual $\hat{r}_{n,e}^{k,m}$ is received correctly with probability $1-p$, producing $\tilde{x}_{n,e}^{k,m} = \hat{x}_{n,e}^{k,m}$. It is lost with probability p, where the decoder uses base layer reconstruction to conceal producing $\tilde{x}_{n,e}^{k,m} = \hat{x}_{n,b}^{k,m}$. Hence,

$$\begin{aligned}E\{\tilde{x}_{n,e}^{k,m}\}(IL) &= (1-p)(\hat{x}_{n,e}^{k,m}) + p\hat{x}_{n,b}^{k,m}\ , \\ E\{(\tilde{x}_{n,e}^{k,m})^2\}(IL) &= (1-p)(\hat{x}_{n,e}^{k,m})^2 + p(\hat{x}_{n,b}^{k,m})^2. \end{aligned} \quad (4)$$

ET Prediction Mode: The packet containing residual $\hat{r}_{n,e}^{k,m}$ and motion vector is received with probability $1-p$, and the decoder first refers to previously reconstructed frame (potentially distorted by prior packet losses) for the motion compensated transform coefficient, i.e., $\tilde{u}_{n,e}^{k,m}$. This along with base layer quantization interval, $\mathcal{I}_n^b$, is plugged into (2) to generate the ET prediction at the decoder. Note that this prediction is potentially different from that at the encoder due to the uncertainty in $\tilde{u}_{n,e}^{k,m}$. Since base layer information $\mathcal{I}_n^b$ is assumed to be available undistorted at the decoder, we henceforth represent the ET prediction (2) by the truncated notation $f_{\mathcal{I}_n^b}(\tilde{u}_{n,e}^{k,m})$, indicating its dependence on the random variable $\tilde{u}_{n,e}^{k,m}$. The residual $\hat{r}_{n,e}^{k,m}$ is added to $f_{\mathcal{I}_n^b}(\tilde{u}_{n,e}^{k,m})$ to produce the enhancement layer reconstruction. The packet is lost with probability p, in which case the decoder conceals with the base layer reconstruction:

$$\begin{aligned}E\{\tilde{x}_{n,e}^{k,m}\}(ET) &= (1-p)(\hat{r}_{n,e}^{k,m} + E\{f_{\mathcal{I}_n^b}(\tilde{u}_{n,e}^{k,m})\}) + p\hat{x}_{n,b}^{k,m} \\ E\{(\tilde{x}_{n,e}^{k,m})^2\}(ET) &= (1-p)((\hat{r}_{n,e}^{k,m})^2 + 2\hat{r}_{n,e}^{k,m}E\{f_{\mathcal{I}_n^b}(\tilde{u}_{n,e}^{k,m})\} \\ &\quad + E\{(f_{\mathcal{I}_n^b}(\tilde{u}_{n,e}^{k,m}))^2\}) + p(\hat{x}_{n,b}^{k,m})^2\ . \end{aligned} \quad (5)$$

The above update equations involve the first and second moments of $f_{\mathcal{I}_n^b}(\tilde{u}_{n,e}^{k,m})$, whose *exact* evaluation via recursive update equations is highly complex due to its inherent non-linearity via (2). Therefore, we approximate $f_{\mathcal{I}}(\cdot)$ by its Taylor series expansion around $E\{\tilde{u}_{n,e}^{k,m}\}$, retaining only up to the quadratic term:

$$\begin{aligned}f_{\mathcal{I}}(u) &\approx f_{\mathcal{I}}(E\{\tilde{u}_{n,e}^{k,m}\}) + (u - E\{\tilde{u}_{n,e}^{k,m}\})f_{\mathcal{I}}^{(1)}(u)|_{u=E\{\tilde{u}_{n,e}^{k,m}\}} \\ &\quad + \frac{(u - E\{\tilde{u}_{n,e}^{k,m}\})^2}{2}f_{\mathcal{I}}^{(2)}(u)|_{u=E\{\tilde{u}_{n,e}^{k,m}\}}\ . \end{aligned} \quad (6)$$

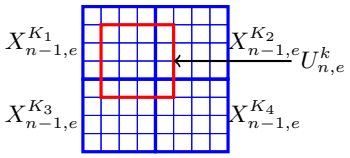

Fig. 1. An off-grid block (red) overlaps 4 on-grid blocks (blue).

Here, $f_{\mathcal{I}}^{(1)}(\cdot)$ and $f_{\mathcal{I}}^{(2)}(\cdot)$ denote, respectively, the first and second order derivatives of $f_{\mathcal{I}}(\cdot)$. As mentioned in Sec. 2.2, $f_{\mathcal{I}}(u)$ can be written in closed form involving u, the base layer interval $\mathcal{I}$, and Laplacian parameter λ, and thus $f_{\mathcal{I}}^{(1)}(\cdot)$ and $f_{\mathcal{I}}^{(2)}(\cdot)$ can be evaluated. Taking expectation of either side of (6) and plugging $u = \tilde{u}_{n,e}^{k,m}$ and $\mathcal{I} = \mathcal{I}_n^b$ yields the first moment of $f_{\mathcal{I}_n^b}(\tilde{u}_{n,e}^{k,m})$,

$$\begin{aligned}E\{f_{\mathcal{I}_n^b}(\tilde{u}_{n,e}^{k,m})\} &\approx f_{\mathcal{I}_n^b}(E\{\tilde{u}_{n,e}^{k,m}\}) \\ &+ \frac{1}{2}(E\{(\tilde{u}_{n,e}^{k,m})^2\} - (E\{\tilde{u}_{n,e}^{k,m}\})^2)f_{\mathcal{I}_n^b}^{(2)}(u)|_{u=E\{\tilde{u}_{n,e}^{k,m}\}}. \end{aligned} \quad (7)$$

The second moment can be obtained similarly. The above implies that the required moments of the prediction can be obtained if the moments $E\{\tilde{u}_{n,e}^{k,m}\}$ and $E\{(\tilde{u}_{n,e}^{k,m})^2\}$ of the potentially off-grid motion compensation reference are available. These latter moments can themselves be derived, as per the original SCORE derivation of [6], from the available moments of on-grid blocks in frame $n-1$. We recall this procedure below.

An off-grid block overlaps at most four on-grid blocks (Fig. 1). Let block U_n^k shown in the figure be the reference block for the current block k in frame n. This block, located in frame $n-1$, overlaps with on-grid blocks $X_{n-1,e}^{k_i}$ in the frame. The decoder enhancement layer reconstruction of block U_n^k is associated with coefficients $\tilde{u}_{n,e}^{k,m}$. Due to linearity of the transform, there exists a set of constants $a_{i,m}$ named *construction constants*, such that

$$\tilde{u}_{n,e}^{k,m} = \sum_{i=1}^{4}\sum_{m=0}^{15} a_{i,m}\tilde{x}_{n-1,e}^{k_i,m}\ .$$

The construction constants only depend on the relative position of U_n^k in this four block grid. The first moment of $\tilde{u}_{n,e}^{k,m}$ is given by

$$E\{\tilde{u}_{n,e}^{k,m}\} = \sum_{i=1}^{4}\sum_{m=0}^{15} a_{i,m}E\{\tilde{x}_{n-1,e}^{k_i,m}\}\ .$$

Computation of the second moment of $\tilde{u}_{n,e}^{k,m}$ involves cross-correlations of pairs of transform coefficients in on-grid blocks:

$$E\{(\tilde{u}_{n,e}^{k,m})^2\} = \sum_{i=1}^{4}\sum_{j=1}^{4}\sum_{m=0}^{15}\sum_{l=0}^{15} a_{i,m}a_{j,l}E\{\tilde{x}_{n-1,e}^{k_i,m}\tilde{x}_{n-1,e}^{k_j,l}\}\ .$$

The computationally intensive calculation of these cross-correlations is circumvented by the following 'uncorrelatedness' approximation which has been shown to hold well in the DCT domain [6]: $E\{\tilde{x}_{n,e}^{k_i,m}\tilde{x}_{n,e}^{k_j,l}\} \approx E\{\tilde{x}_{n,e}^{k_i,m}\}E\{\tilde{x}_{n,e}^{k_j,l}\}$ when $j \neq i$ or $l \neq m$. Thus the recursions (4) and (5) are complete.

4. ESTIMATION AND CODING PERFORMANCE

We first demonstrate the estimation accuracy of SCORE in ET-SVC, by embedding it in the encoder solely to track end-to-end distortion, i.e., SCORE's EED estimate is not used to optimize encoding decisions. The test sequence is $foreman$ at CIF resolution. Base layer packets are transmitted losslessly, while the enhancement layer packets are randomly dropped with probability 5%. Packet losses

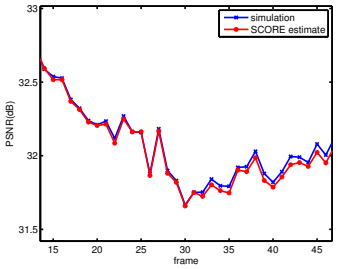

Fig. 2. Comparison of simulated and estimated PSNRs for the $foreman$ sequence at CIF resolution encoded by ET-SVC: the base layer bit-rate is $70kbps$, enhancement layer bit rate is $150kbps$, and enhancement layer packet loss rate is 5%.

are independent and identically distributed. The transmission is simulated over 30 different realizations of the lossy channel. The distortion of each frame is averaged over all realizations and converted to PSNR. The simulated and estimated PSNRs are shown in Fig. 2. It is evident that SCORE provides an accurate estimate of EED for ET-SVC.

Having established the accuracy of SCORE in the ET-SVC setting, we next present the end-to-end coding performance obtained when SCORE's EED estimates are employed to optimally select encoding parameters. Let $D^k_{n,e}(q,\mu)$ and $B^k_{n,e}(q,\mu)$ denote the EED and bit costs incurred in encoding macroblock k of frame n at the enhancement layer with quantization parameter (QP) q and prediction mode μ (inter-layer or ET). All macroblocks in the frame share the same QP, denoted by $q_{n,e}$. The optimization problem is formulated as the per-macroblock minimization:

$$\mu^k_{n,e}(\lambda,q) = \arg\ \min_{\mu}\{D^k_{n,e}(q,\mu) + \lambda B^k_{n,e}(q,\mu)\}, \qquad (8)$$

and the subsequent per-frame minimization:

$$q_{n,e}(\lambda) = \arg\ \min_{q} \sum_k D^k_{n,e}(q,\mu^k_{n,e}) + \lambda B^k_{n,e}(q,\mu^k_{n,e}), \qquad (9)$$

where λ is a Lagrange parameter whose value is fixed for the entire sequence in our simulation. Varying λ provides an operational rate-distortion curve.The ET-SVC encoder whose coding modes are optimized using SCORE is referred to as ET-SVC-SCORE. It is compared to a conventional (spatial domain) SVC optimized using EED provided by ROPE (SVC-ROPE, see [4, 5]), and to ET-SVC without additional error resilience. The rate-distortion performance of sequence $mobile$ at CIF resolution is shown in Fig. 3, where the base layer bit-stream is fixed and is identical for all three systems under comparison, and the enhancement layer is transported at packet loss rate 5%. To demonstrate the coding performance under various channel conditions, sequence bus at CIF resolution is encoded with fixed enhancement layer bit-rate and is evaluated with different packet loss rate. The performance shown in Fig.4 demonstrates that ET-SVC provides significant compression gains compared to the conventional SVC in the lossless channel scenario (i.e., $p = 0$), and in comparison to SVC-ROPE is fairly robust at low packet loss rates. The proposed ET-SVC-SCORE scheme inherits the compression efficiency of ET-SVC, while accounting for the channel condition and potential error propagation, and substantially outperforms the competing schemes across a wide range of packet loss rates.

5. CONCLUSION

A novel error-resilient SVC scheme is proposed that achieves two optimality goals. It subsumes optimal (non-linear) enhancement

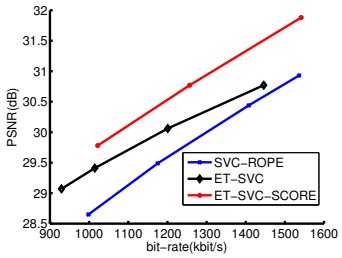

Fig. 3. End-to-end coding performance for sequence $mobile$ at CIF resolution: the base layer is encoded at bit-rate $500kbps$ and the enhancement layer has a packet loss rate of 5%.

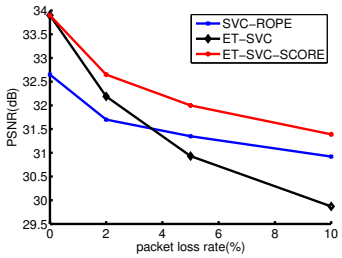

Fig. 4. Coding performance versus packet loss rate for sequence bus at CIF resolution: the base layer bit-rate is $430kbps$; the enhancement layer bit rate is $1020kbps$.

layer prediction that exploits all available information from both base and enhancement layer sources. It complements this with the necessary recursive estimate of end-to-end distortion that operates in the spectral domain, which accounts for compression, packet loss, error propagation, and concealment. Simulations provide evidence for the accuracy of the estimate and for substantial performance gains of the overall SVC system.

6. REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 17, pp. 1103–1120, Sep 2007.

[2] K. Rose and S. L. Regunathan, "Toward optimality in scalable predictive coding," *IEEE Trans. Img. Proc.*, vol. 10, no. 7, pp. 965–976, Jul 2001.

[3] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Jrnl. Sel. Areas Comm.*, vol. 18, pp. 966–976, June 2000.

[4] S. L. Regunathan, R. Zhang, and K. Rose, "Scalable video coding with robust mode selection," *Sig. Proc.: Image Comm.*, vol. 16, pp. 725–732, 2001.

[5] Y. Guo, Y. Chen, Y.-K. Wang, H. Li, M. M. Hannuksela, and M. Gabbouj, "Error resilient coding and error concealment in scalable video coding," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 19, pp. 781–795, 2009.

[6] J. Han, V. Melkote, and K. Rose, "A recursive optimal spectral estimate of end-to-end distortion in video communications," in *Proc. Packet Video*, Dec 2010.

[7] C. A. Segall and G. J. Sullivan, "Spatial scalability within the H.264/AVC scalable video coding extension," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 17, no. 9, pp. 1121–1135, September 2007.