RECURSIVE END-TO-END DISTORTION ESTIMATION FOR ERROR-RESILIENT ADAPTIVE PREDICTIVE COMPRESSION SYSTEMS

Sina Zamani, Tejaswi Nanjundaswamy, Kenneth Rose

Department of Electrical and Computer Engineering, University of California Santa Barbara, CA 93106
E-mail: {sinazmn, tejaswi, rose}@ece.ucsb.edu

ABSTRACT

Linear prediction is widely used in speech, audio and video coding systems. Predictive coders often operate over unreliable channels or networks prone to packet loss, wherein errors propagate through the prediction loop and may catastrophically degrade the reconstructed signal at the decoder. To mitigate this problem, end-to-end distortion (EED) estimation, accounting for error propagation and concealment at the decoder, has been developed for video coding, and enables optimal rate-distortion (RD) decisions at the encoder. However, this approach was limited to the video coder's simple setting of a single tap constant coefficient temporal predictor. This paper considerably generalizes the framework to account for: i) high order prediction filters, and ii) filter adaptation to local signal statistics. Specifically, we propose to simultaneously track the decoder statistics of the reconstructed signal and the prediction parameters, which enable effective estimation of the overall EED. We first demonstrate the accuracy of the EED estimate in comparison to extensive simulation of transmission through a lossy network. Finally, experimental results demonstrate how this EED estimate can be leveraged, by an encoder with short and long term linear prediction, to improve RD decisions and achieve major performance gains.

***Index Terms*—** Error resilience, rate-distortion optimization, end-to-end distortion, adaptive prediction

1. INTRODUCTION

Exploiting correlation is a critical component of all compression and communication systems. One central approach to do so involves prediction, wherein typically linear short term and/or long term prediction filters are employed. Modern packet-switched networks, such as the Internet, provide limited or no end-to-end Quality of Service (QoS) [1] guarantees, where packets may be discarded due to buffer overflow at intermediate nodes of the network, or considered lost due to long queuing delays. Thus, robustness to packet loss is a crucial requirement, especially in the case of predictive coding, where the prediction loop propagates errors and causes substantial, and sometimes catastrophic, deterioration of the received signal.

The problem of packet loss is mitigated by adding redundancy in the bitstream to recover from errors, e.g., by resetting prediction [2] at appropriate intervals to stop propagation of error, or employing error correcting codes [3] to protect critical information. In such scenarios, the overall performance of coders depends on optimizing the trade-off between compression and redundancy for error resilience. A formal illustration of the problem setup is shown in Fig. 1. Encoder E compresses source $\mathcal{X}$ to $\mathcal{Z}$, while accounting for channel or network unreliability. The Decoder D receives $\hat{\mathcal{Z}}$ and decodes the reconstructed source $\hat{\mathcal{X}}$. The overall problem is formally posed as optimizing encoder parameters and decisions to minimize the end-to-end distortion (EED), which accounts for quantization, packet loss, error propagation and concealment at the decoder, given the prescribed bit rate. Clearly, effective EED estimation at the encoder is critical to solving this problem.

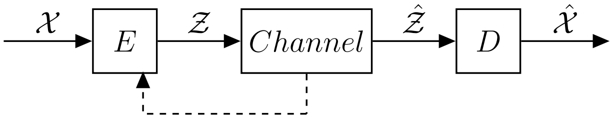

Fig. 1. A general compression and communication system

In [4] the recursive optimal per-pixel estimate (ROPE) of EED was proposed for video coders, wherein EED is estimated at the encoder via tracking the first and second moments of the reconstructed signal at the decoder, which are recursively updated. ROPE was demonstrably optimal for the video coding setting it addressed, and its superior accuracy yielded significant performance gains over earlier heuristic methods. Nevertheless, ROPE was derived for a rather simple setting, which limits its applicability to more general settings. Specifically, ROPE was derived for a predictor with single tap for every pixel in a video frame (pointing to a motion-compensated position in the previous frame), but many coders employ a combination of short term and long term prediction filters, which lead to complex dependencies across consecutive samples. Moreover, ROPE assumes a fixed temporal prediction coefficient, which is obviously not affected by packet loss, while many compression techniques use time varying prediction parameters adapted to the local statistics. When a packet generated by an adaptive predictive coder is lost, information necessary to determine the prediction parameters is lost as well. Thus recursively estimating the EED while accounting for this uncertainty entails considerable challenges. Some techniques were previously proposed in [5, 6] to extend ROPE for handling cross correlation terms that arise due to basic interpolation filters, by employing certain approximations relevant to the context of video coding, but they do not account for adaptive prediction parameters. Note that a somewhat related problem setup exists in networked control systems (NCS) [7, 8], wherein observations or innovations from sensors are transmitted over unreliable networks and the receiver performs state estimation to make controller decisions. These systems only account for channel unreliability for state estimation at the receiver, which is equivalent to packet loss concealment in networked compression systems. Instead we propose tackling the problem of

This work was supported in part by a gift from Mozilla.

978-1-4673-7802-4/16/$31.00 ©2016 IEEE

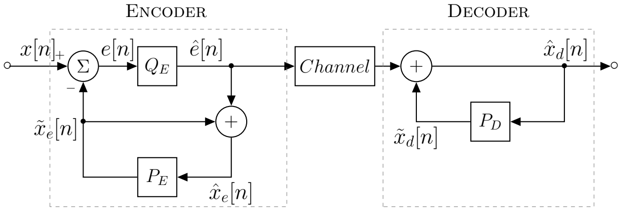

Fig. 2. A predictive compression system

accounting for the network reliability at the encoder.

In this paper we substantially generalize the ROPE framework to estimate EED at the encoder for a compression system which employs a higher order predictor with adaptive prediction parameters. We specifically derive a recursive procedure to estimate EED by separately tracking statistics of both prediction parameters and the reconstructed signal at the decoder, which are then effectively combined to estimate the overall EED. The accuracy and efficacy of the estimation is shown via simulation results which substantiate that incorporating such information in making RD optimal decisions of prediction resets at the encoder can achieve significant performance gains.

2. END-TO-END DISTORTION

Fig. 2 illustrates a predictive compression system, which is a specific case of the general system described in Sec. 1. Input signal samples, $x[n],\ 0 \leq n < N$, are processed by the encoder to generate and transmit a bitstream through a channel, which the decoder receives and decodes to generate the reconstructed samples. The encoder also computes the reconstructed sequence, $\hat{x}_e[n]$, and uses prior reconstructed samples to generate the prediction $\tilde{x}_e[n]$. The prediction error, $e[n] = x[n] - \tilde{x}_e[n]$ is quantized to obtain $\hat{e}[n]$, which is conveyed to the decoder. When the decoder receives $\hat{e}[n]$, it adds it to its predicted sample, $\tilde{x}_d[n]$, to generate its reconstructed samples, $\hat{x}_d[n]$. If $\hat{e}[n]$ is not received, the decoder performs loss concealment by generating an approximation or a "guess" for the reconstructed samples. Clearly, the decoder reconstructed samples, $\hat{x}_d[n]$, may differ from those at the encoder, $\hat{x}_e[n]$, and consequently the predicted samples at the decoder, $\tilde{x}_d[n]$, and the encoder, $\tilde{x}_e[n]$, may also differ. Hence the added subscripts to indicate encoder versus decoder quantities. As far as the encoder is concerned, the decoder reconstruction $\hat{x}_d[n]$ is a random variable, as the exact loss pattern cannot be known while encoding, but a statistical model for channel loss is presumably available to the encoder. The encoder's only feasible strategy is to estimate the expected end-to-end distortion (EED). For the squared error distortion metric, EED for the given source sequence is,

$$\begin{aligned} D &= \sum_{n=0}^{N-1} E\{(x[n] - \hat{x}_d[n])^2\} \\ &= \sum_{n=0}^{N-1} x^2[n] - 2x[n]E\{\hat{x}_d[n]\} + E\{\hat{x}_d^2[n]\}. \end{aligned} \tag{1}$$

Note that the encoder knows the source sequence so the only source of randomness is due to loss in the channel. Clearly, to estimate this distortion, first and second moments of the decoder reconstructions should be accurately estimated at the encoder. In [4], a recursive algorithm was proposed to optimally estimate decoder reconstructions' first and second moments, which were used to estimate EED for the setting of a single tap constant coefficient temporal predictor. This estimated EED was then used to optimally switch between inter-frame and intra-frame prediction to control the error propagation through frames.

3. PROPOSED END-TO-END DISTORTION ESTIMATION

In this section we describe EED estimation for a compression system employing a higher order predictor with adaptive prediction parameters by simultaneously tracking statistics for relevant decoder quantities, namely, prediction parameters and the reconstructed signal. First we explain the general estimation algorithm, then we describe extension for a compression system using a cascade of short term and long term predictor, and finally we describe how EED is employed to optimize RD decisions at the encoder.

3.1. General EED Estimation Framework

The common approach for adapting to local statistics is to divide the input signal into frames and employ different prediction parameters for each frame. Let $x^f[n]$ and $\hat{x}_e^f[n]$ denote the original and encoder reconstruction value of sample n in frame f, respectively. The predicted samples of frame f using a higher order predictor are given by,

$$\tilde{x}_e^f[n] = \sum_{i=1}^{P} \gamma_e^f[i]\hat{x}_e^f[n-i], \tag{2}$$

where P is the prediction order and $\gamma_e^f[i]$ is the ith prediction coefficient used for samples in frame f. Given the predicted samples, the quantized prediction error, $\hat{e}^f[n]$, is generated as in Sec. 2, and is transmitted along with the prediction parameters, $\boldsymbol{\gamma}_e^f = [\gamma_e^f[1],\ \ldots,\ \gamma_e^f[P]]$, in a single packet over the channel. Due to lossy nature of the channel the packet may either be received by the decoder, or lost. For simplicity of presentation (and without loss of generality) let us model the channel loss with a Bernoulli model, where each packet is lost independently of other packets, with probability p, called packet lost rate (PLR). Upon receiving the packet, the decoder adds the quantized error, $\hat{e}^f[n]$, to its predicted sample, $\tilde{x}_d^f[n]$, and generates its reconstructed sample, $\hat{x}_d^f[n]$. The predicted samples at the decoder are given by,

$$\tilde{x}_d^f[n] = \sum_{i=1}^{P} \gamma_d^f[i]\hat{x}_d^f[n-i], \tag{3}$$

where $\gamma_d^f[i]$ is the ith prediction coefficient employed in frame f at the decoder. If a packet is lost, concealment is done by assuming the quantized prediction error was zero and copying the prediction parameters from the previous reconstructed frame, $\boldsymbol{\gamma}_d^f = [\gamma_d^{f-1}[1],\ \ldots,\ \gamma_d^{f-1}[P]]$. Recall that because of packet loss, the predicted samples, reconstructions, and prediction parameters employed ($\tilde{x}_d^f[n]$, $\hat{x}_d^f[n]$, and $\boldsymbol{\gamma}_d^f$, respectively) at the decoder differ from corresponding quantities at the encoder, and must be viewed as random variables by the encoder.

Let f_R denote the event that the packet containing information of frame f is received and let f_L denote the event that it is lost. Then the first moment of the reconstructed sample at the decoder can be expressed as,

$$E\{\hat{x}_d^f[n]\} = (1-p)E\{\hat{x}_d^f[n]|f_R\} + pE\{\hat{x}_d^f[n]|f_L\}, \tag{4}$$

where

$$\begin{aligned} E\{\hat{x}_d^f[n]|f_R\} &= E\{(\hat{e}^f[n]+\sum_{i=1}^{P}\gamma_e^f[i]\hat{x}_d^f[n-i])|f_R\} \\ &= \hat{e}^f[n]+\sum_{i=1}^{P}\gamma_e^f[i]E\{\hat{x}_d^f[n-i]|f_R\} \end{aligned} \tag{5}$$

$$E\{\hat{x}_d^f[n]|f_L\} = E\{\sum_{i=1}^{P}\gamma_d^{f-1}[i]\hat{x}_d^f[n-i])|f_L\}. \tag{6}$$

Note that in (6), both $\gamma_d^{f-1}[i]$ and $\hat{x}_d^f[n-i]$ are random variables. If we further assume them to be uncorrelated, we can approximate the first moment for event f_L as,

$$E\{\hat{x}_d^f[n]|f_L\} \approx \sum_{i=1}^{P}E\{\gamma_d^{f-1}[i]\}E\{\hat{x}_d^f[n-i]|f_L\}. \tag{7}$$

Note that while one may object from a source coding perspective that a source and its prediction parameters would normally be correlated, but it is important to keep in mind that the only uncertainty of the encoder (and hence the only source of randomness) about the reconstructed samples and the prediction parameters is due to unreliability of the channel. The validity of this assumption is verified in the experimental results in Sec. 4. Based on the concealment strategy adopted, the first moment for the prediction parameter vector employed at the decoder is also estimated recursively as

$$E\{\boldsymbol{\gamma}_d^f\} = (1-p)\boldsymbol{\gamma}_e^f + pE\{\boldsymbol{\gamma}_d^{f-1}\}. \tag{8}$$

Substituting (5) and (7) in (4) we obtain a recursive estimate for the first moment of reconstructed samples at the decoder. Similarly, for the second moment,

$$E\{(\hat{x}_d^f[n])^2\} = (1-p)E\{(\hat{x}_d^f[n])^2|f_R\}+pE\{(\hat{x}_d^f[n])^2|f_L\}, \tag{9}$$

where

$$\begin{aligned} E\{(\hat{x}_d^f[n])^2|f_R\} &= E\{(\hat{x}_d^f[n](\hat{e}^f[n]+\sum_{i=1}^{P}\gamma_e^f[i]\hat{x}_d^f[n-i]))|f_R\} \\ &= \hat{e}^f[n]E\{\hat{x}_d^f[n]|f_R\} + \\ &\quad \sum_{i=1}^{P}\gamma_e^f[i]E\{\hat{x}_d^f[n]\hat{x}_d^f[n-i]|f_R\} \end{aligned} \tag{10}$$

$$\begin{aligned} E\{(\hat{x}_d^f[n])^2|f_L\} &= E\{(\hat{x}_d^f[n](\sum_{i=1}^{P}\gamma_d^{f-1}[i]\hat{x}_d^f[n-i]))|f_L\} \\ &\approx \sum_{i=1}^{P}E\{\gamma_d^{f-1}[i]\}E\{\hat{x}_d^f[n]\hat{x}_d^f[n-i]|f_L\}. \end{aligned} \tag{11}$$

The correlation terms in (10) and (11) can be calculated from the past correlation terms as

$$\begin{aligned} E\{\hat{x}_d^f[n]\hat{x}_d^f[n-i]|f_R\} &= E\{(\hat{x}_d^f[n-i](\hat{e}^f[n]+ \\ &\quad \sum_{j=1}^{P}\gamma_e^f[j]\hat{x}_d^f[n-j]))|f_R\} \\ &= \hat{e}^f[n]E\{\hat{x}_d^f[n-i]|f_R\}+ \\ &\quad \sum_{j=1}^{P}\gamma_e^f[j]E\{\hat{x}_d^f[n-i]\hat{x}_d^f[n-j]|f_R\} \end{aligned} \tag{12}$$

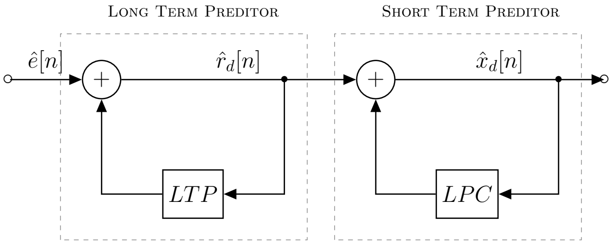

Fig. 3. Decoder section of a speech coder with cascade of predictors

$$\begin{aligned} E\{\hat{x}_d^f[n]\hat{x}_d^f[n-i]|f_L\} &= E\{(\hat{x}_d^f[n-i](\sum_{j=1}^{P}\gamma_d^{f-1}[j]\hat{x}_d^f[n-j]))|f_L\} \\ &\approx \sum_{j=1}^{P}E\{\gamma_d^{f-1}[j]\}E\{\hat{x}_d^f[n-i]\hat{x}_d^f[n-j]|f_L\}. \end{aligned} \tag{13}$$

Overall, equations (10) to (13) are employed to recursively estimate the second moment of reconstructed samples at the decoder. Given the first and second moments, EED is estimated using (1).

3.2. EED Estimation for Cascaded Predictors

In many real-world predictive coders, higher order predictors are implemented as a combination of multiple predictors. For example, speech coders [9] employ a cascade of a short term prediction filter (known as the linear predictive coding (LPC) filter) and a long term prediction (LTP) filter. Fig. 3 illustrates the decoder section of an example speech coder. The decoder processes the received quantized prediction error, $\hat{e}^f[n]$, through the LTP synthesis filter to reconstruct the excitation signal, $\hat{r}_d^f[n]$, as,

$$\hat{r}_d^f[n] = \sum_{i=0}^{P_1-1}\beta_d^f[i]\hat{r}_d^f[n-T_d^f-i]+\hat{e}^f[n], \tag{14}$$

where $\beta_d^f[i]$ is the ith LTP filter coefficient, T_d^f is the lag parameter, and P_1 is the number of LTP filter taps. The LPC synthesis filter uses the reconstructed excitation signal to generate the reconstructed samples as

$$\hat{x}_d^f[n] = \sum_{j=1}^{P_2}\alpha_d^f[j]\hat{x}_d^f[n-j]+\hat{r}_d^f[n], \tag{15}$$

where $\alpha_d^f[i]$ is the jth LPC prediction coefficient and P_2 is the LPC filter order. We can easily combine (14) and (15) to form a single prediction filter, $\hat{x}_d^f[n]=\hat{e}^f[n]+\tilde{x}_d^f[n]$, where

$$\begin{aligned} \tilde{x}_d^f[n] &= \sum_{j=1}^{P_2}\alpha_d^f[j]\hat{x}_d^f[n-j]+\sum_{i=0}^{P_1-1}\beta_d^f[i](\hat{x}_d^f[n-T_d^f-i]- \\ &\quad \sum_{j=1}^{P_2}\alpha_d^f[j]\hat{x}_d^f[n-T_d^f-i-j]). \end{aligned} \tag{16}$$

Clearly, (16) is similar to (3), wherein P and $\gamma_d^f[i]$ of (3) can be written in terms of P_1, P_2, $\beta_d^f[i]$, $\alpha_d^f[j]$, and T_d^f of (16). Thus, as would be expected, the estimation framework proposed in Sec. 3.1 is applicable to coders with cascaded predictors.

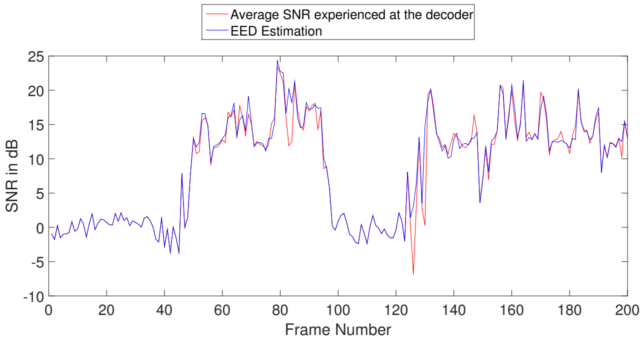

Fig. 4. Comparison of average SNR experienced at the decoder (red) and estimated SNR (blue) for a the speech file *English Male* with a PLR of 5 %

3.3. Employing Estimated EED for Encoder Decisions

A common approach to combat error propagation through the prediction loop is to introduce prediction resets [10] at the encoder to halt dependency on past frames. While these resets stop the error propagation due to packet losses, they come at the cost of increased bit rate, so optimizing the number and location of resets plays a crucial role in achieving the right balance between compression efficiency and robustness to packet losses. Conventional methods use random resets at a rate equal to the PLR to stop error propagation. Instead, we leverage the proposed EED estimate as computed by the encoder to directly minimize the EED for the prescribed bit rate all within the encoder RD optimization framework. This results in optimal selection of location and number of resets. Specifically to encode frame f, we choose the mode (reset or no reset) to minimize the rate-distortion cost function,

$$J^f = D^f + \lambda R^f, \tag{17}$$

where R^f is the bit rate, D^f is the estimated EED, and Lagrange multiplier λ controls the RD operating point.

4. EXPERIMENTAL RESULTS

To validate the accuracy and efficacy of our proposed method we employed it in a coder with cascade of predictors similar to Sec. 3.2. The 6 speech files available in the EBU SQAM database [11] were used to conduct the experiments. We set $P_1 = 5$ and $P_2 = 12$, while operating with frames of 20ms sampled at 16 kHz. Pitch lags from 2ms (32 samples) to 20ms (320 samples) were allowed. We estimated the LPC and LTP parameters in an open-loop for each frame and used them to generate the open-loop prediction error. A fixed rate 4-bits scalar quantizer was then designed for the entire prediction error sequence. Finally, we employed the prediction parameters and the designed quantizer in a closed-loop system to generate the quantized prediction error that is sent to the decoder along with all the parameters every frame at a fixed rate.

In Fig. 4 we plot the actual SNR experienced at the decoder (averaged over 200 different loss patterns) and the estimated SNR obtained at the encoder by our proposed framework, for 200 frames of the speech file *English Male*, operating at 5% PLR. It is clearly evident that our estimate is fairly accurate in tracking the actual SNR experienced at the decoder.

Sequence	Average SNR in dB Random Resets	Average SNR in dB Proposed Approach
English Female	9.72	11.87
English Male	8.99	12.57
French Female	8.78	11.89
French Male	9.7	12.62
German Female	4.64	7.68
German Male	8.81	11.61
Average	8.44	11.37

Table 1. Comparison of average SNR experienced at the decoder for random reset versus the proposed reset strategy at PLR = 5%

Sequence	Average SNR in dB Random Resets	Average SNR in dB Proposed Approach
English Female	5.38	8.52
English Male	5.87	8.38
French Female	-1.21	6.72
French Male	5.31	8.48
German Female	-5.46	1.79
German Male	5.01	8.89
Average	2.48	7.13

Table 2. Comparison of average SNR experienced at the decoder for random reset versus the proposed reset strategy at PLR = 10%

We then compared our proposed strategy for deciding resets to that of using random resets at a rate equal to PLR. Since we employ fixed rate quantizers in our experimental setup, the cost used to decide resets, as explained in Sec. 3.3, simplifies to only the EED estimate. The evaluation is limited to 8 seconds of each speech file for time efficient evaluation. In Table 1 and Table 2 we compare SNR experienced at the decoder (averaged over 50 loss patterns) for the two competing prediction reset strategies at 5% and 10% PLR, respectively. For the random reset strategy, we additionally tried 10 different reset patterns, thus obtaining the final SNR as an average over 500 simulations. Clearly, the proposed approach consistently outperforms the random reset scheme under all testing scenarios, with gains of up to 7.8 dB, and an average gain of 2.9 dB and 4.6 dB for 5% and 10% PLR, respectively. A crude implementation of the proposed approach in MATLAB is 10X slower than the original codec on an Intel Core i5-4570R 2.7 GHz machine with 8 GB of RAM, which is largely due to operations involving large correlation matrices. However, note that correlation matrices are often sparse and structured, and we plan to exploit these properties in future work that will focus on complexity reduction for the proposed approach.

5. CONCLUSION

In this paper we proposed an effective technique to estimate EED in an adaptive predictive compression system. Specifically, we proposed to account for the effect of packet losses on distortion at the decoder by separately tracking statistics of the employed prediction parameters and the reconstructions at the decoder. We then demonstrated incorporating the estimate obtained by the proposed approach in an RD framework to decide the number and location of prediction resets to achieve the right balance between compression and addition of redundancy to combat packet losses. Significant performance improvements seen in experimental evaluation results demonstrate the utility of the proposed approach.

6. REFERENCES

[1] Q.-F. Zhu and L. Kerofsky, "Joint source coding, transport processing, and error concealment for H. 323-based packet video," in *Electronic Imaging'99*. International Society for Optics and Photonics, 1998, pp. 52–62.

[2] G. Côté and F. I. Kossentini, "Optimal intra coding of blocks for robust video communication over the internet," *Signal Processing: Image Communication*, vol. 15, no. 1, pp. 25–34, 1999.

[3] Y. Wang and Q.-F. Zhu, "Error control and concealment for video communication: A review," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 974–997, 1998.

[4] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 966–976, 2000.

[5] Z. Chen, P. V. Pahalawatta, A. M. Tourapis, and D. Wu, "Improved estimation of transmission distortion for error-resilient video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 636–647, 2012.

[6] H. Yang and K. Rose, "Advances in recursive per-pixel end-to-end distortion estimation for robust video coding in H. 264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 7, pp. 845–856, 2007.

[7] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, and S. Sastry, "Kalman filtering with intermittent observations," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1453–1464, 2004.

[8] R. T. Sukhavasi and B. Hassibi, "The Kalman-like particle filter: optimal estimation with quantized innovations/measurements," *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 131–136, 2013.

[9] A. S. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[10] B. S. Atal, V. Cuperman, and A. Gersho, *Advances in speech coding*, vol. 114, Springer Science & Business Media, 1991.

[11] G. Waters, "Sound quality assessment material-recordings for subjective tests: Users handbook for the EBU-SQAM compact disk," *European Broadcasting Union (EBU), Tech. Rep*, 1988.