Video streaming with SP and SI frames

Eric Setton and Bernd Girod

Information Systems Laboratory,
Stanford University, Stanford, CA 94305

ABSTRACT

SP and SI frames are new picture types introduced in the latest video coding standard H.264. They allow drift-free bitstream switching and can also be used for error-resilience and random access. We investigate the benefits of SI and SP frames for error resilience as compared to periodic I frame insertion. We discuss the rate-distortion performance of SI and SP frames based on empirical rate-distortion curved obtained with our implementation of an SP/SI frame encoder. Experiments carried out over a simulated bandwidth-limited network analyze the influence of loss rate and delay on the congestion-rate-distortion performance of streaming with SI and SP frames. Our results help identify scenarios for which SI and SP frames provide an attractive alternative to streaming with I frames.

Keywords: Video compression, video streaming, H.264, SI frames

1. INTRODUCTION

The design of the latest video coding standard, H.264,[1] reflects the increasing need for video streaming solutions which adapt to varying network conditions. In addition to achieving superior coding efficiency, H.264 uses network-friendly syntax and incorporates several new encoding features which can be taken advantage of when designing flexible and adaptive streaming systems. The new picture types SP and SI are one of these features.

Based on the seminal work by Färber et al.,[2] SP and SI frames were proposed in 2001 by Karczewicz and Kurceren, as a solution for error resilience, bitstream switching and random access.[3, 4] They are now part of the extended profile of H.264. The main advantage of this new picture type is that, it can be reconstructed exactly by using different sets of predictors or no predictor at all. This allows drift-free bitstream switching applications such as refreshing a prediction chain or switching betweeen different quality streams as depicted in Fig. 1.

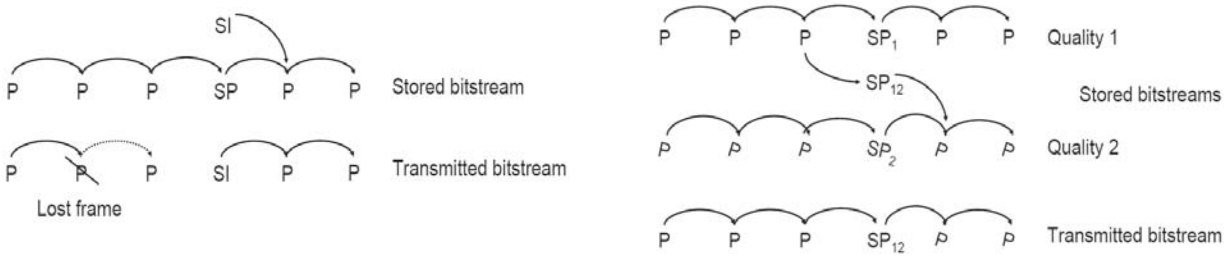

Figure 1. SI frames have the same instant refresh properties as I frames but only need to be transmitted if a transmission error occurs. Switching SP frames allow to switch streams using predictive frames only.

No work so far has analyzed how streaming solutions with SP and SI frames compare to more traditional systems. This is, in part, due to the fact that no reference implementation of an SP/SI encoder has been provided to the community. In this paper, we analyze the efficiency of a streaming scheme which uses SI frames for error resilience and compare it to a system which relies instead on I frames. We seek to identify under which circumstance streaming with SI and SP frames surpasses streaming with I frames and to quantify potential performance gaps. The results presented in this paper are based on experiments carried out using our publicly

Send correspondence to Eric Setton. E-mail: esetton@stanford.edu, phone: 1 650 723 3476 available implementation of an SP/SI frame codec, based on H.264.[5] Our empirical results complement the theoretical rate-distortion analysis presented in a related paper.[6]

In the next section, we describe main features of SP/SI frames encoders. In Sec. 3, we analyze how to tradeoff optimally the size of SP and SI frames for streaming applications. In Sec. 4, empirical rate distortion efficiency of SP and SI frames is compared to that of I, P and B frames and bounds on experimental performance gaps are shown for different sequences. Experimental results based on NS-2 simulations[7] are discussed in Sec. 5, where streaming with SI and SP frames is compared to streaming with I frames. We analyze the effect of bit rate, losses and delay on the performance of the two systems and conclude by identifying for which scenarios SI and SP frames provide an attractive alternative to I and P frames.

2. SP AND SI FRAMES ENCODING

A diagram of an SP frame encoder is shown on the left of Fig. 2. It is mainly composed of a traditional video encoder followed by an additional intra-frame encoder which operates on the reconstructed image signal s'[*]. It is this second quantization stage that allows identical reconstruction from different predictors and provides the switching and restart functionalities of SP frames.

The quantized coefficients output by the second intra-frame encoder, l_{rec}, are subsequently entropy-coded to produce SI frames. For switching SP frames, only the residual of a motion-compensated prediction of l_{rec} is entropy-coded, as depicted on the right of Fig. 2. As these steps are lossless, the coefficients l_{rec} may be obtained at the decoder whether an SP, SI or switching SP frame is transmitted. This ensures the reconstructed image is identical in all cases.

The design shown in the left of Fig. 2 differs slightly from the encoder originally proposed in,[3, 4] where motion-compensation is followed by an additional intra-frame encoder. It is comparable to a later design accepted by JVT[8, 9] where this additional step is only performed if it is beneficial for rate-distortion performance. The performance of the encoder shown in Fig. 2 is very close and much easier to analyze and understand, it is chosen for simplicity.

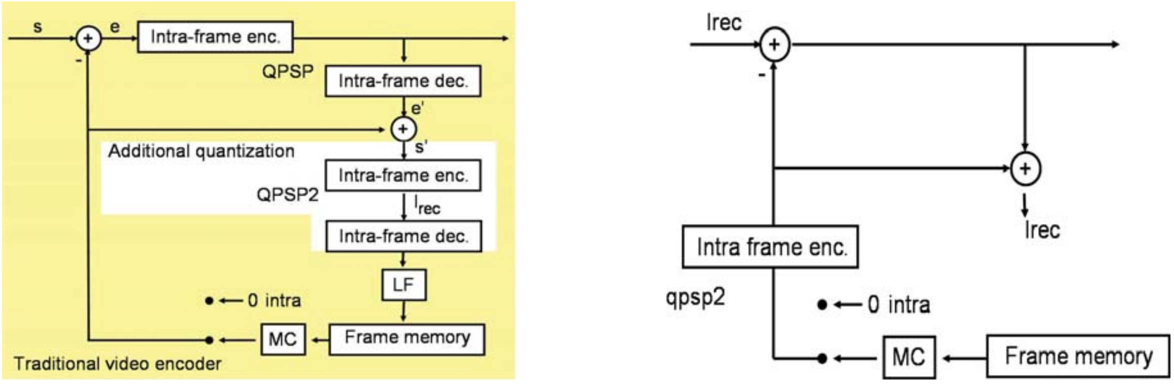

Figure 2. Left: SP frame encoder. Right: SI and switching SP frame encoder.

For a given quality, the size of SP frames and of SI frames may be traded off by varying the two quantization parameters shown in Fig. 2, $QPSP$ and $QPSP2$. When there is no quantization in the first encoder (i.e. $QPSP = 0$), SI frames become equivalent to I frames. When there is no quantization in the second encoder (i.e. $QPSP2 = 0$), SP frames are equivalent to P frames. The optimal tradeoff is usually an intermediate setting which depends on the application.

[*] We call intra-frame encoder the combination of a spatial transform followed by quantization. In the figures of the paper, we show next to this block a symbol (either QPSP or QPSP2) representing the value of the quantizer.

3. GENERATING SP AND SI FRAMES FOR STREAMING

In the rest of the paper, we focus on the use of SP and SI frames for error resilience. In this particular scenario, intra-coded pictures are sent to stop error-propagation, only if a packet has been lost. This results in a more flexible transmission of pre-stored encoded streams, and should lead to bit-rate savings compared to periodic I frame insertion which occurs regardless of the outcome of previous transmissions. For a given quality, the best performance is achieved when the quantizers step size $QPSP$ and $QPSP2$ are set to minimize the expected bit-rate, while keeping the quality of SP frames equal to that of the other frames of the stream. This obviously depends on the quality of the encoded stream and on how often SI frames are sent in lieu of SP frames. The optimal encoder settings were derived in,[6] based on theoretical rate-distortion analysis of the encoder.

For our implementation of the SP codec,[5] these settings are given in Table 1 as a function of QP, the quantization parameter used for the P frames of the stream, when the probability x of transmitting an SI frame instead of an SP frame varies. Note that $QPSP$ and $QPSP2$ need to be integers; this limits the range of their values.

Table 1. Optimal settings for QPSP and QPSP2, for different probabilities of transmitting an SI frame.

x	≤ 0.2	≥ 0.2 and ≤ 0.5	≥ 0.5
QPSP	$QP - 1$	$QP - 2$	$QP - 3$
QPSP2	$QP - 10$	$QP - 5$	QP

4. PERFORMANCE ANALYSIS

In this section, we analyze the empirical rate-distortion efficiency of SI and SP frames and derive bounds on bit rate savings which could be achieved when streaming video with SI and SP frames.

Figure 3 illustrates the rate-distortion performance of SI and SP frames compared to I, P and B frames for 18 frames evenly spaced among the 300 first frames of the sequences *Foreman* and *Mother and Daughter*. These curves correspond to realistic rates for today's Internet video streaming of up to 600 kbps. SP frames and SI frames were encoded following the settings shown in the middle column of Tab. 1. For these settings, SP frames are typically larger than P frames by approximately 70% for both sequences. Similarly, SI frames are typically twice as big as I frames. These differences are less pronounced at higher rates as analyzed in.[6] The main difference between these two sequences is the relative size of I frames and P frames. For the sequence *Foreman*, I frames are approximately twice the size of P frames. For the sequence *Mother and Daughter*, I frames are approximately 3 times larger than P frames. This difference comes from larger motion in the sequence *Foreman* which increases the average size of P frames. As a consequence, the bit rates savings expected by sending intra-coded frames on an as-needed basis will be higher for the sequence *Mother and Daughter*.

The video encoding structure shown in Fig. 4 was chosen for the streaming experiments presented in the next section. The GOPs are 16 frames long with one SP frame (and its corresponding SI frame) per GOP and 3 B frames between P frames. This ensures good error resilience properties and allows to easily scale down the frame rate by 2 or even 4 if needed. The encoded video sequences used in this paper, as well as rate-distortion preambles characterizing the size and quality of the frames are made publicly available[10][†].

The curves on Fig. 5 show the rate-distortion characteristic of the video sequences compressed using the coding structure described in Fig. 4. As there are only one I, SP or SI frame in 16 frames, the difference between the curves is not as pronounced as in Fig. 3. For the first sequence, transmitting SP frames instead of I frames can lead to a performance gain of 1.5 dB at low bit rates and 1 dB at higher rates. For the second sequence, this gap is smaller and ranges from 1 dB to 0.8 dB. These gaps represent a bound on the performance improvement, achieved when streaming takes place with no losses. If SI frames are used instead of I frames, the rate distortion performance is reduced by approximately 1.5 dB at low rates and a little less than 1 dB at high rates.

[†] The preambles also indicate the distortion values obtained by concealing a frame with any other frame of the stream, allowing to simulate realistically video streaming without the overhead of encoding and decoding.

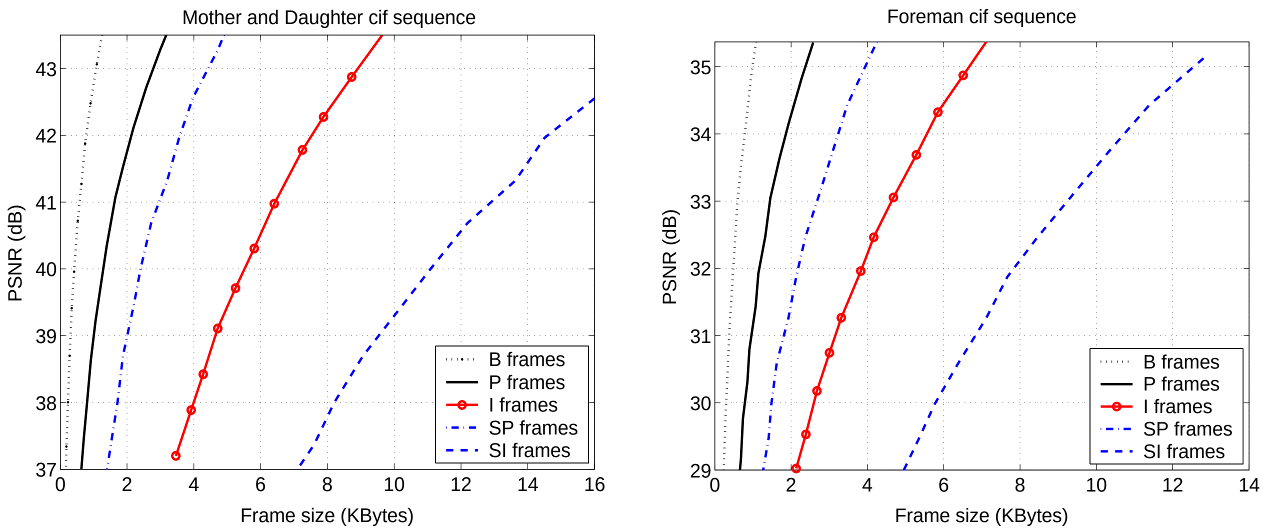

Figure 3. Rate-distortion performance for different frame types.

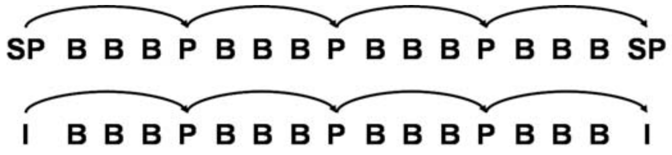
SP B B B P B B B P B B B P B B B SP
I B B B P B B B P B B B P B B B I

Figure 4. GOP structures used for streaming with SP and SI frames and for periodic I frame insertion.

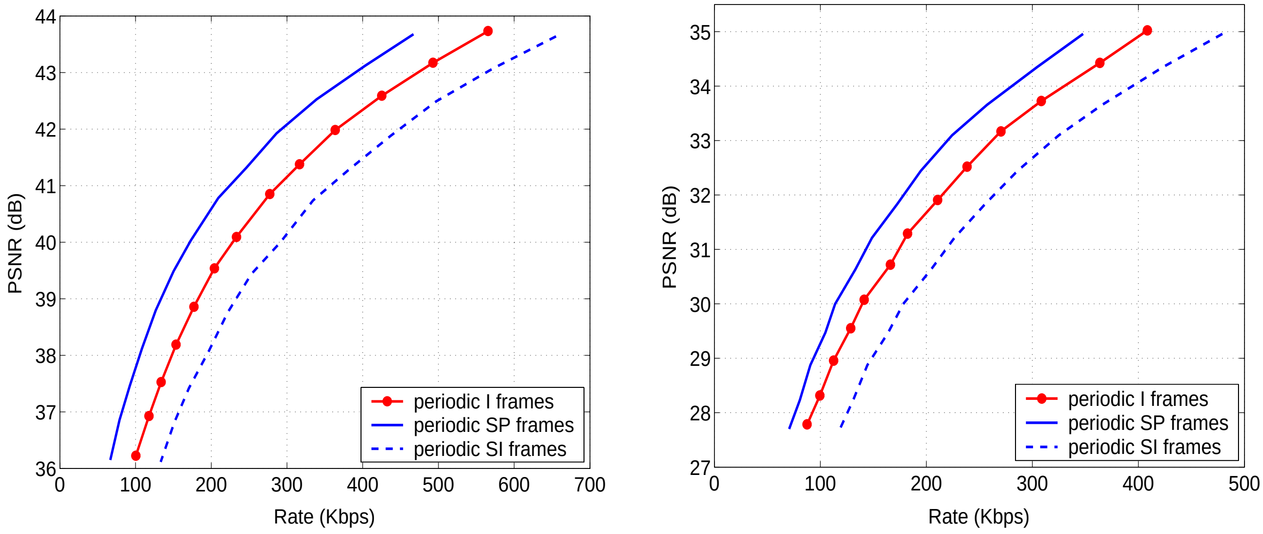

Figure 5. Rate-distortion performance with periodic I frame, SP frame or SI frame insertion, for *Mother & Daughter* (left) and *Foreman* (right).

5. SIMULATION RESULTS

To illustrate realistically the benefits of streaming with SP and SI frames we consider a low latency video streaming scenario, suitable for live streaming or for video-on-demand, where a sender transmits video frames sequentially to a receiver which sends acknowledgements (ACKs) back. We strive for end-to-end delays of no more than a few hundred milliseconds. When a packet arrives at the receiver after its playout deadline, it is discarded by the decoder as if it were lost. To avoid interruptions, the errors due to packet loss or to excessive delays are concealed by freezing the previous frame until the next decodable frame and the playout continues at the cost of higher distortion. The sender retransmits lost packets when ACKs are received out of sequence, and when there is still enough time to retransmit a packet before its playout deadline. For the case when SP frames are used, if a P frame or an SP frame is lost and cannot be retransmitted, an SI frame is sent at the beginning of the next GOP as depicted in Fig. 1.

We consider the route between sender and receiver as a succession of high bandwidth links ended by a bottleneck last hop which can support up 800 kbps. Packet losses are simulated on this last hop in some of the following experiments. Packets containing an entire video frame are generated by our video encoder and are fragmented, if required by the transport layer. When a loss occurs, the entire frame is discarded, even though, in most cases, only one packet is lost. In such environments, it is important to consider realistic packetization as different frame types have vastly varying sizes, as illustrated in Fig. 3. At low rates, for example, B and P frames may fit into one MTU size packet, whereas SP frames may necessitate 2 packets, I frames 3, and SI frames 6. Consequently, different frame types may experience different loss rates. The impact on the resulting PSNR may be significantly different from that induced by independent losses identically distributed among all the frames.

The sequences are encoded at 30 frames per second with the GOP structure shown in Fig. 4. The first 288 frames of the sequences are encoded, in order to have an even number of GOPs and the encoded sequence is looped 50 times when collecting results. Video quality is measured by taking the average of the peak-signal-to-noise-ratio (PSNR) over all the decoded frames. Performance is also evaluated in terms of the total transmitted rate and of the average end-to-end delay between the server and the client. This quantity reflects the congestion created by the stream on the network. The fact that this metric, unlike rate, depends on the capacity of the network path makes it well-suited to performance evaluation in a bandwidth-limited environment. It reflects the delay another stream would experience if it was sharing the link with the video stream. End-to-end delay is measured by taking the average end-to-end delay of header packets transmitted every 20 ms from the server to the client, as in.[11]

We first analyze the influence of packet losses. We consider a fixed 50 ms propagation delay and a 500 ms latency tolerance. In Fig. 6, both the congestion-distortion performance and the rate-distortion performance are shown for two sequences. In the absence of losses, the gains in terms of rate and distortion are close to those predicted in the previous section. The rate-distortion performance gap is a little smaller due in part to the fact that I frames are inserted every 10 seconds, each time the sequence is looped. For the sequence *Mother & Daughter* the performance gap is approximately 0.6 dB, and 0.4 dB for the sequence *Foreman* for different bit rates. The congestion-distortion performance gap is larger, it varies from 2 dB for low levels of congestion to 1 dB for higher levels of congestion for the sequence *Mother & Daughter*. The gap is not as large for the sequence *Foreman*. This illustrates the queueing delay spikes caused by I frames, which are not captured by the average rate of the sequence. When a 5% loss rate is introduced on the bottleneck link, the performance drops for all the curves. This drop is more significant at high bit rates as the packet loss rate translates into a higher frame loss rate. For higher rates than those shown, the average decoded video quality decreases. Surprisingly, the rate-distortion performance gap increases when losses are introduced. This is due to the effect of I frame retransmission on rate-distortion efficiency which is larger than that of occasional SI frame insertion. The congestion-distortion performance gap remains almost the same. These experiments show that streaming with SI and SP frames is beneficial in this experimental setup regardless of packet loss rate.

In the following, we analyze the influence of the propagation delay. We consider a fixed 2% packet loss rate and a 500 ms latency tolerance. The propagation delay is varied between 20 ms and 200 ms. This delay occurs on the high bandwidth links and reflects the time needed for signal to propagate along links which can potentially be very long (e.g. transoceanic or transcontinental links). For short propagation delays the performance is only slightly worse than the performance in the absence of loss, shown in Fig. 6. The slight loss in performance is due

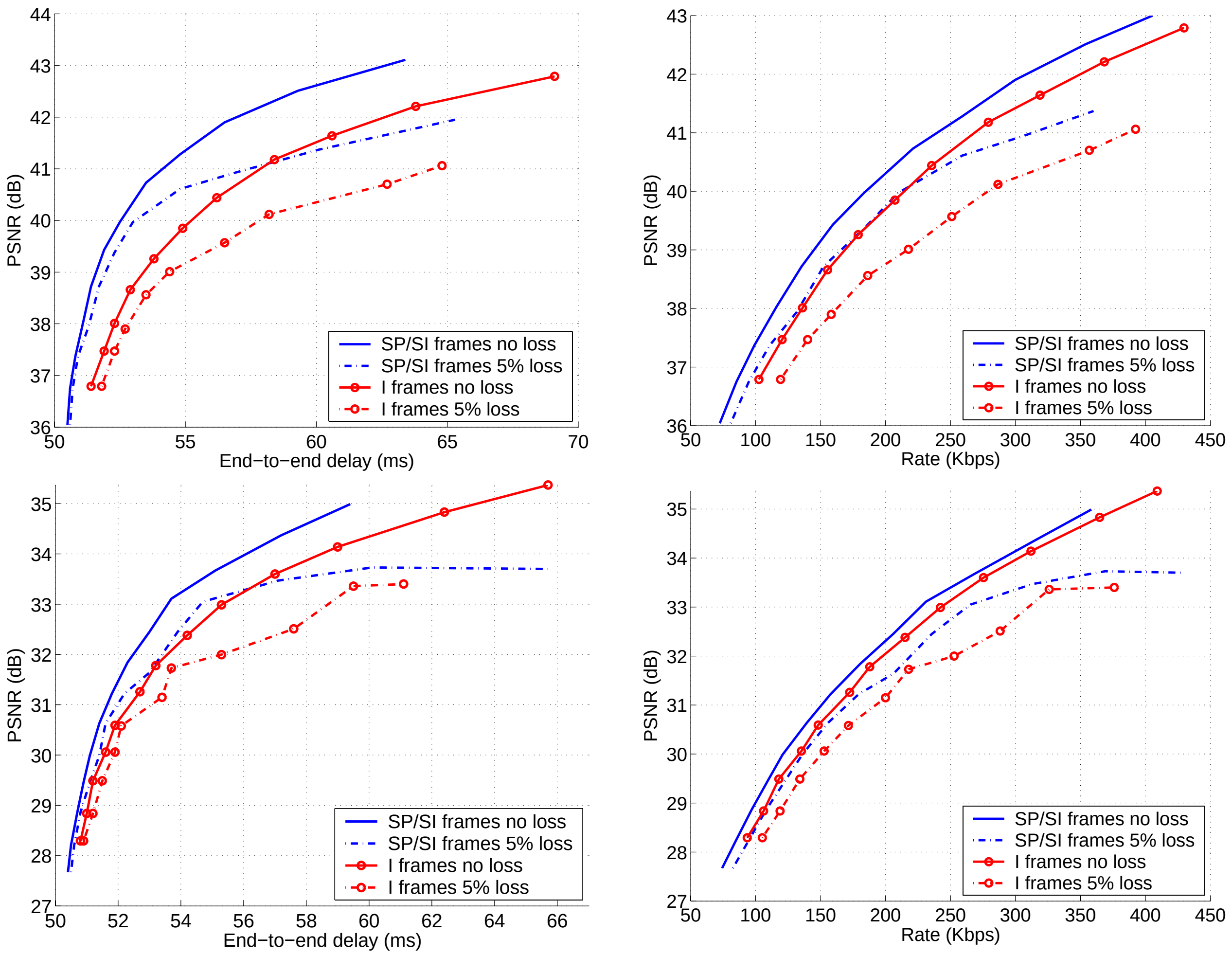

Figure 6. Congestion-Rate-Distortion performance for *Mother & Daughter* (top) and *Foreman* (bottom) for varying loss rates.

to the 2% loss rate which induces retransmissions and an increase in rate. For long propagation delays, there is no time for retransmission and the performance drop for all the schemes is 2 to 4 dB. This causes significant quality impairments for streaming with SI and SP frames as well as for streaming with I frames. The performance gap in this case is reversed. Indeed, as SI frames need to be inserted almost constantly, the congestion-rate-distortion performance is worse than for periodic I frame insertion. The performance gap ranges from 1.5 to 2 dB for different rates and congestion levels for *Mother & Daughter*. Likewise, for *Foreman*, it ranges from 1 dB to 2 dB. For high propagation delays, in the absence of retransmission, streaming with periodic I frames is more efficient and the performance gap is significant. Other experiments, run for an intermediate propagation delay of 100 ms, show comparable performance for both schemes.

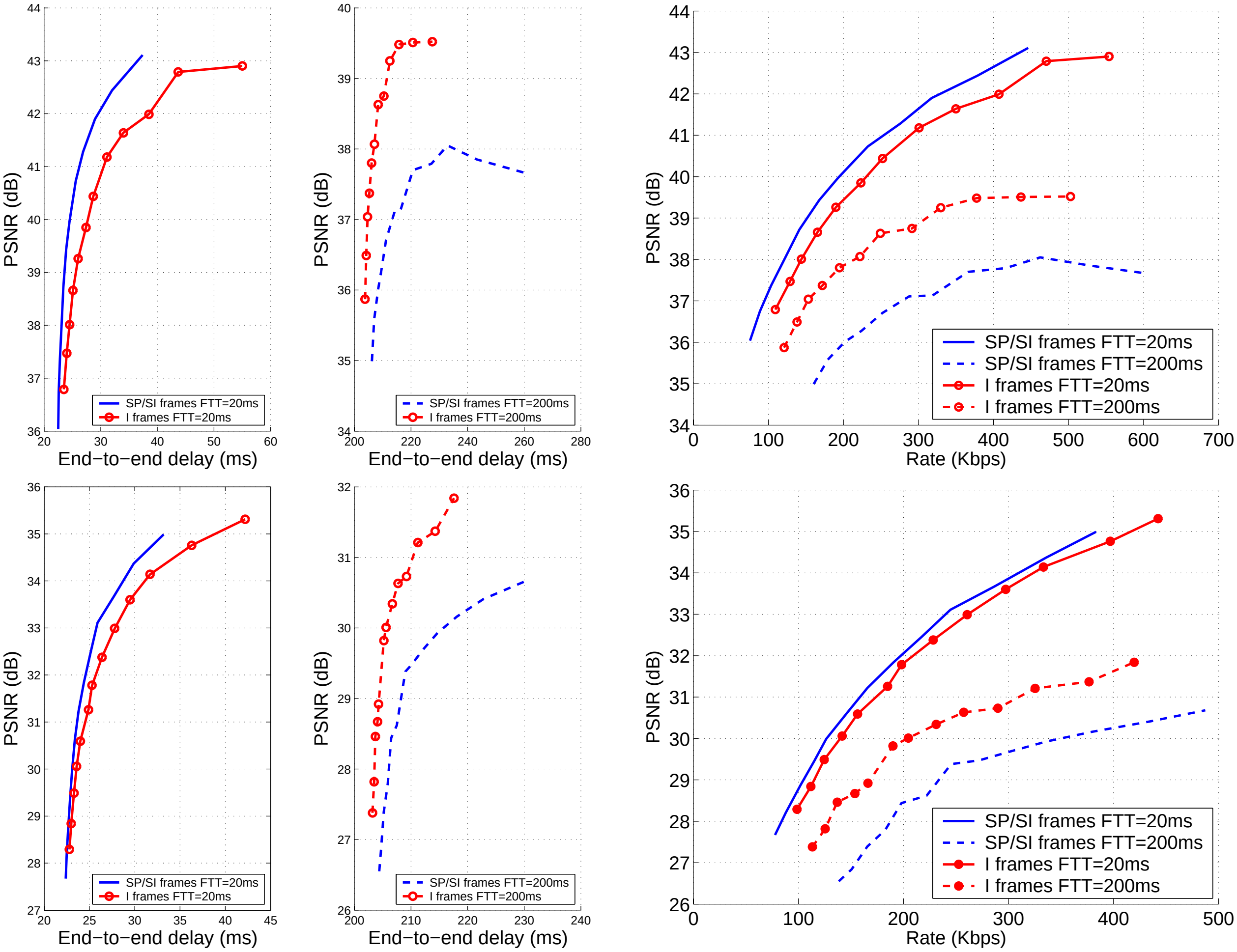

Figure 7. Congestion-Rate-Distortion performance for *Mother & Daughter* (top) and *Foreman* (bottom) for varying propagation delays

As a summary to this analysis, SP and SI frames provide an attractive alternative to streaming with I frames when feedback is available and propagation delay is small compared to the maximum tolerable latency. In these cases, the performance is superior both in terms of congestion-distortion and in terms of rate-distortion. The performance gap is larger for low motion sequences as this causes larger differences between I and P frames. It is also more pronounced at lower bit rates and can reach up to 1.5 dB. For the case when retransmissions are not possible, periodic I frame insertion remains the best alternative and the performance gap is over 1 dB.

6. CONCLUSIONS

In this paper we discuss the benefits of streaming with SP and SI frames. We first describe SP and SI frame encoding and explain how perfect reconstruction may be achieved by using our implementation of a simple SP frame encoder, based on H.264 and made publicly available. The performance analysis reveals that gains of up to 1.5 dB can be obtained for video rates between 100 kbps and 600 kbps. Experimental results obtained on a simulated bandwidth limited network, for low latency streaming, with varying loss rates and propagation delays confirm these expectations. The experiments also illustrate that streaming with SP and SI frames reduces the congestion created by the stream on the network by up to 40%. The use of SP and SI frames is beneficial for scenarios where feedback is available from the receiver and the propagation delay is low enough to allow for ACK-based retransmissions.

REFERENCES

1. ITU-T and ISO/IEC JTC 1, *Advanced Video Coding for Generic Audiovisual services, ITU-T Recommendation H.264 - ISO/IEC 14496-10(AVC)*, 2003.
2. N. Färber and B. Girod, "Robust H.263 Compatible Video Transmission for Mobile Access to Video Servers," *Proc. ICIP-97, Santa Barbara, CA, USA* **2**, pp. 73–76, Oct. 1997.
3. M. Karczewicz and R. Kurceren, "A Proposal for SP-Frames," *Video Coding Experts Group Meeting, , Doc. VCEG-L-27, Eibsee, Germany* , Jan. 2001.
4. M. Karczewicz and R. Kurceren, "The SP- and SI-frames design for H.264/AVC," *IEEE Trans. on Circuits and Systems for Video Technology* **13**, pp. 637–644, July 2003.
5. "H.264 SP frame codec," *http://www.stanford.edu/˜ esetton/H264_2.htm* .
6. E. Setton, P. Ramanathan, and B. Girod, "Rate-distortion analysis of SP and SI frames," *submitted to ICIP 2005* .
7. "The Network Simulator - ns-2," *www.isi.edu/nsnam/ns/* .
8. X. Sun, S. Li, F. Wu, J. Shen, and W. Gao, "The improved SP frame coding technique for the JVT standard," *International Conference on Image Processing, Barcelona, Spain* **3**, pp. 297–300, Sept. 2003.
9. X. Sun, F. Wu, S. Li, and R. Kurceren, "The improved JVT-B097 SP coding scheme," *ISO/IEC JTC1/SC29/ WG11 and ITU-T SG16 Q.6, JVT-C114, Fairfax, Virginia, USA* , May 2002.
10. "Encoded sequences with SP/SI frames," *http://ivms.stanford.edu/˜ esetton/sequences.htm* .
11. E. Setton and B. Girod, "Congestion-Distortion Optimized Scheduling of Video," *Multimedia Signal Processing Workshop (MMSP), Siena, Italy* , pp. 99–102, Oct. 2004.