Analysis of Video Streaming with SP and SI Frames in UMTS Mobile Networks

Luca Superiori and Markus Rupp*
Institute of Communications and RF Engineering,
Vienna University of Technology
Gusshausstrasse 25/389
A-1040 Vienna, Austria
{lsuper,mrupp}@nt.tuwien.ac.at

Wolfgang Karner
mobilkom austria
Obere Donaustrasse 21
A-1020 Vienna, Austria
w.karner@mobilkom.at

ABSTRACT

In this paper we discuss possible benefits of transmitting SI frames as an error resilience tool in UMTS video streaming. SP and SI frames can be used to stop temporal error propagation, as an alternative to the regular insertion of I frames. Since their application relays on feedback information from the mobile equipment, different delay scenarios have been investigated. The performed transmission simulations show enhancement of the rate-distortion behaviour when using SP and SI frames with reduced feedback times. The error traces were measured at transport block level in a live UMTS mobile network.

Keywords

Video Streaming, UMTS, H.264/AVC, SP and SI Frames

1. INTRODUCTION

H.264/AVC [1] is the state-of-the-art video codec jointly standardized by ISO/MPEG (International Standard Organization Moving Picture Expert Group) and ITU/VCEG (International Telecommunication Unit Video Coding Expert Group) in 2003. The standard has been organized in different *profiles*, each designed for specific classes of applications (such as unicast, broadcast or storage) and defines a set of enabled features.

The 3GPP (3rd Generation Partnership Project) is the international association that defines the specifications of the third generation mobile phone system. In [2] the 3GPP defines the H.264/AVC in its *baseline* profile as optional video codec that has to be supported by standard compliant mobile devices. The baseline profile is the most basic one and its set of functionalities are included by all the other profiles. It has been selected because of the target device capabilities (mobile phones with limited computational power) and because of the transmission characteristics (unicast with strong delay constraints).

The baseline profile allows two kind of frames. The Intra (I) predicted frames exploit the spatial correlation of a single picture. The frame is segmented into *macroblocks* and each macroblock is then encoded using the neighbouring ones as prediction's reference. The Inter (P) predicted frames are encoded taking advantage of the similarities between consecutive pictures. The size of an encoded frame depends strongly on the prediction type used. The spatial prediction is much less effective than the temporal one, therefore it requires more bits for correcting the prediction. For this reason, a video stream consists of a group of inter encoded frames interleaved, with a given frequency, by intra encoded frames. The distance, in frames, between two consecutive I frames is defined as Group Of Picture (GOP).

Because of the previously mentioned delay constraints, incorrectly received packets cannot usually be retransmitted and are discarded. The decoder applies *error concealment mechanisms* to recover the missing information and reduce the impact of the damaged frames. Even though, for the specific single frame, error concealment methods might hide the missing data, the picture is used as a source of reference by the following Inter encoded pictures. This effect is shown in Fig. 1. One of the packets belonging to F_1 is incorrectly

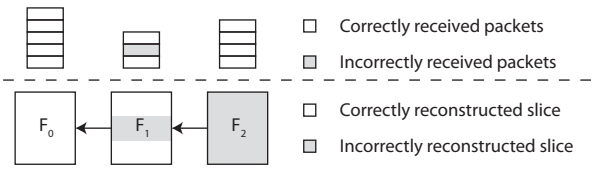

Figure 1: Temporal Error Propagation

received, causing a slice of the reconstructed picture to be concealed. Although all the packets belonging to F_2 are correctly received, the frame is incorrectly reconstructed as a concealed frame is used as a source of temporal prediction. We will refer to this effect as *temporal error propagation*.

The temporal error propagation ceases once an I frame is correctly received. The extent of the temporal error propagation strongly depends, therefore, on the size of the GOP. Small GOP sizes limit the propagation of the error in time

*The authors thank mobilkom austria AG for technical and financial support of this work. The views expressed in this paper are those of the authors and do not necessarily reflect the views within mobilkom austria AG.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MoMM 2009 December 14–16, 2009, Kuala Lumpur, Malaysia.
Copyright 2009 ACM 978-1-60558-659-5/09/0012 ...$10.00.

but result in increasing stream size as shown in Fig. 2.

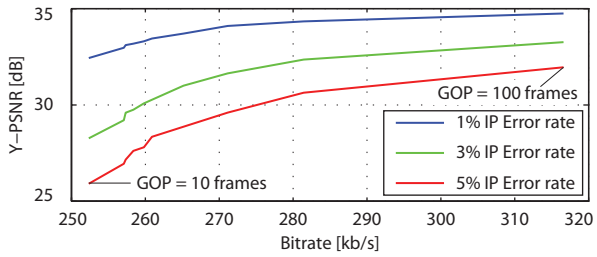

Figure 2: Rate-distortion behavior depending on the GOP size

In this article we consider the application of the SI frames in mobile environments to reduce the effect of the temporal error propagation. The SI frames are special intra encoded frames, that can be sent in place of SP frames, special encoded P frames, in case the terminal reports that an error has occurred. Although the SP and SI frames are not yet supported by the profile suggested in [2], we performed an analysis aimed at measuring possible benefits deriving from their application.

The paper is organized as follows. In Section 2 the SI and SP frames are briefly explained. Section 3 describes the considered channel realisation. The simulation setup is discussed in Section 4. Section 5 presents the performed analysis in terms of rate distortion. Section 6 offers conclusive remarks.

2. SP AND SI FRAMES

The temporal error propagation affects the inter encoded frames following the incorrectly received frame. It is stopped by the next Intra encoded frame, since it does not contain reference to the previous pictures. Moreover, this is true only if the Intra frame empties the reference pictures buffer or if the size of the buffer itself has been set equal to one.

The position of the I frames is set during the encoding and cannot be modified. The GOP size might be constant or, for some specific applications, variable. In case of sudden scene changes, in fact, the encoder might find beneficial to encode the whole frame as Intra, since the previous frames do not represent an efficient source of prediction.

Also considering the possibility of a feedback mechanism, in case temporal error propagation has been detected an I frame cannot be sent in place of a P frame and substitute it. The macroblock reconstruction is extremely sensitive to the correctness of the prediction source. After the Intra, or Inter, prediction, each 4×4 subblock of the original macroblock is subtracted from the correspondent one of the best found prediction. The *residual* block is then transformed by means of an horizontal and vertical Direct Cosine Transformation (DCT) and quantized, obtaining the coefficients matrix. The coefficients matrix is entropic encoded and used at the decoder side for improving the prediction using the inverse quantization and inverse DCT transformations and then summing up the result to the prediction. In order not to introduce drifts between the encoder and the decoder, the two predicted block have to be the same. For this reason, at the encoder side the *reconstructed* blocks are used for prediction in place of the, available, *original* blocks.

If an I frame is scheduled for transmission in place of the original P frame without being considered at the encoder side, the following frames are using a wrong source of prediction. This cause the incorrect reconstruction of the pictures, as the decoder and the encoder are using different sources of prediction.

To overcome this limitation, Karczewicz and Kurceren [3] proposed two new types of frames, SP and SI frames. The SP and SI frames are used for encoding the same picture and they reflect the prediction mechanism of the P and I frame, respectively. The two versions of the frame are differently encoded but, when decoded, offer the very same reconstructed picture. This allows, at the decoder side, to substitute transparently an SP frame with an SI frame.

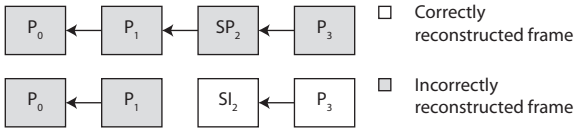

Figure 3: Application example of SP and SI frames

In Fig. 3 is shown how the approach based on the SP and SI frames work. The SP frames act as synchronization points for the SI frames and are interleaved regularly between P frames. Assume now a sequence containing I, P and SP frames. Consider the frame P_0 to be incorrectly received and reconstructed by concealment. Even though the packets belonging to P_1 are correctly received, the frame is incorrectly reconstructed, since its source of prediction is corrupted. As the frame SP_2 is encoded exploiting the temporal prediction, it is erroneously reconstructed as the following P frames. As the SI_2 frame is encoded exploiting the spatial prediction, its quality does not depend on the previous reconstructed frames. The reconstructed SI_2 frame offers the same prediction source to P_3, as the SP_2 would have offered in the error-free case, therefore stopping the temporal error propagation.

Summarizing, in case of trusted transmission channels, SP and SI frames can be substituted transparently to the remaining P frames. In case of error prone channels, one SI frame can be scheduled in place of the equivalent SP frame for stopping the temporal error propagation. In order to ensure an identical reconstructed image, the encoding and decoding of these special frames considerably differs from the ones of the I and P frames. The two most significant differences are: (i) The difference between the original block and the predicted one is calculated in the Direct Cosine Transform (DCT) domain and (ii) The encoding consists of a two stages quantization-dequantization step, using two different Quantization Parameters (QP), QPSP1 and QPSP2. Further details on the implementation of the SP and SI frames in H.264/AVC can be found in [3, 4].

3. CHANNEL MODEL

In literature, channels with a given bit error rate or a given packet loss rate are often considered for simulations. For UMTS wireless channels, they represent a non realistic worst-case scenario. Uncorrelated error models do not reflect the bursty characteristics of the real channels, as shown in [5].

In order to build a more realistic simulation scenario, an error model based on the correctness of the Transport Blocks

(TB) is used. The protocol architecture for video streaming over UMTS networks is drawn in Fig. 4. The entities pro-

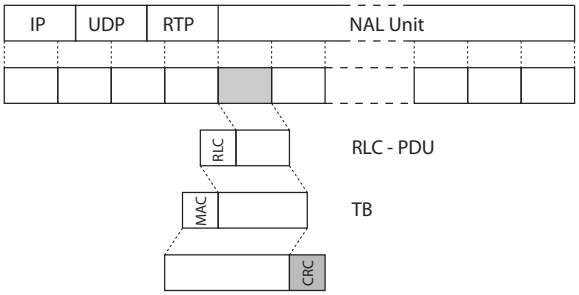

Figure 4: Protocol Stack

duced by the video encoder, the Network Abstraction Layer Unit (NALU), are further encapsulated into the Real Time Protocol (RTP) [6]. To each RTP packet an Universal Datagram Protocol (UDP) [7] as well as an Internet Protocol (IP) [8] header are further attached. This results in a 40 bytes protocol overhead. The IP packets are then segmented into Radio Link Control (RLC) Protocol Data Units (PDUs). For the bearer used for video streaming, it is suggested an RLC payload of 40 bytes. These Transport Blocks (TB) are then mapped onto the transport channel by the Medium Access Control (MAC). To evaluate the correctness of the MAC packets at the receiver, a Cyclic Redundancy Check (CRC) is attached.

For the measurements, a UDP data stream with bit rates of up to 360 kb/s in downlink was sent from a PC over the UMTS network to a notebook using a UMTS terminal as a modem via a USB connection. In Fig. 5 is given a schematic illustration of the setup for the measurements in the live networks. In the notebook, the traces were recorded

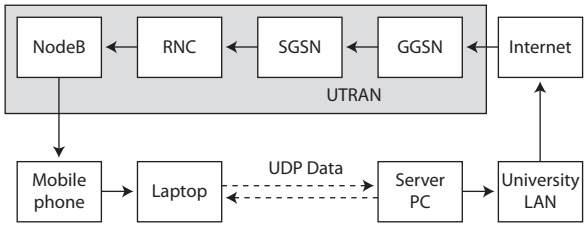

Figure 5: Measurement setup

using the TEMS investigation software by Ericsson. The CRC information of the received TBs were parsed from the traces and used for the analysis of the UMTS link error characteristics.

4. SIMULATION SETUP

The scope of the investigation is to measure the performance of the application of SI and SP frames in H.264/AVC with respect to the classical I-P-P scheme. We simulate the transmission of streams encoded with both strategies over the channel realizations described in Section 3.

4.1 Generation of the encoded sequences

In order to generate the encoded sequences, the standard reference software JM [9] ver. 11.0 has been used. Although at the time of the submission the ver. 15.0 of the software was available, it was not used as the implementation of SP and SI frames in the newer versions is known to be not properly working.

A football video sequence in DVD format consisting of around 1000 frames has been encoded considering the typical configuration of unicast video streaming over third generation mobile networks. The most stringent constraint is represented by the bandwidth that is limited to 180 kb/s. To match this restriction, the resolution of the video has been reduced to QVGA (Quarter Video Graphic Array), 320×240 pixels (typical for most mobile phones). The sequence is also decimated in time, reducing the frame rate (fr) from 30 to 15 frames per second (f/s). The compression level is tuned by means of the quantization parameter, having impact on both the resulting quality and datarate.

For the I-P-P scheme, 17 different GOP sizes varying from 2 to 50 frames have been considered. The GOP size strongly affects datarate, being the encoded I frames much bigger than the P frames. Because of the different encoding mechanism, even if the same QP for both the I frames and the P frames has been used, the quality of the Intra predicted frames was, in average, higher than the one of the Inter encoded frames. That makes the error free sequence quality a function of the GOP size, introducing an unnecessary complexity in the investigation. In order to make the quality of the I-P-P sequence not dependent on the GOP size a QP of 35 for the P frames and QP 37 for the I frames has been used.

For the I-P-S sequences, a single I frames was encoded at the beginning of the sequence. Two sets of sequences have been generated for different S-frames Distances (SD), one containing I-P-SP frames and one containing I-P-SI frames.

As for the I-P-P scheme, the same 17 different distances from 2 to 50 were considered. The quality and the size of the SP and SI frames strongly depend on the selected QPSP1 and QPSP2. At this step, without knowing the probability of needing an SI frame, the both QPSPs have been set to 33. The QPSP of the SP/SI frames has been set slightly smaller than the one of the P frames in order to guarantee the same quality to the reconstructed frames.

4.2 Transmission of the sequences

As suggested in [10] the packet size of the I and P frames was set to 750 bytes, offering a good compromise between header overhead and impact of a lost packet. The bigger the packet, in fact, the less efficient the error concealment.

The SP and SI frames were not sliced into packets since the current implementation of the JM software does not support a fixed size of SP and SI frames in Byte. In order to guarantee the same reconstructed picture, the encoding of the SI frame should be aware of the slice segmentation of SP frames in order to tune the allowed prediction modes, and vice versa.

The transmission was simulated inserting the errors (measured as described in Section 3) at transport block level. Each RTP packet of the stream has been discarded in case one of its transport blocks was marked as flawed. For the I-P-P scheme, a damaged packet was removed from the video file and the standard concealment strategies of JM were used to recover the missing information.

For the I-P-S scheme, a switching mechanism was implemented as shown in Fig. 6. The H.264/AVC encoder pro-

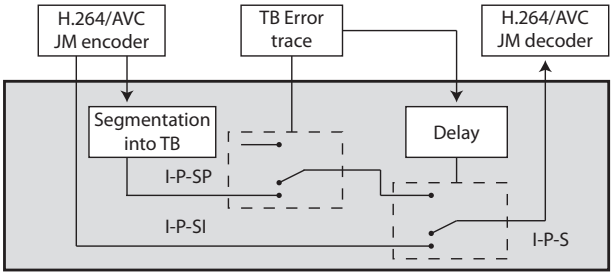

Figure 6: Implementation of the switching

duces the sequences containing the I-P-SP and the I-P-SI frames. As the transmission of the I-P-SP sequence is simulated, it is segmented into Transport Block, which are then compared with the measured error traces. In case one transport block of the IP packet is damaged, the whole IP packet is discarded. Additionally, after a given delay, the script substitutes the SP frames with the corresponding SI frame, in case an error has been signalized.

In a realistic transmission scenario, the delay time has to be considered, depending both on how the feedback information is transmitted and where the SI frames are stored. Two possible implementations are shown in Fig. 7.

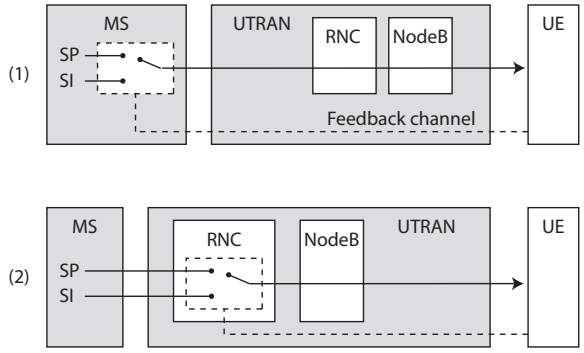

Figure 7: Implementation schemes in real networks

The easiest implementation (1) considers the SI frames stored in the Media Server (MS) where the whole sequence is stored. The SI frames are sent in place of SP frames if the mobile terminal feedbacks errors in the link. A smarter solution (2) considers the SI frames to be sent from the media server to the Radio Network Controller (RNC) each time the appropriate SP frame is scheduled for transmission. Assuming RNC improved capabilities, this network element decides which frame is delivered further to the Node-B without the need of forwarding the feedback information to the media server. Although redundant information is sent through the core network, in this second implementation the feedback information concerns the correctness of the single transport block, and not the correctness of the whole packet, further reducing the required delay. In order to cover a variety of technical implementations, different delay times varying from 0.1 to 2 s have been considered. As soon as the SI frame request has been received, the SI frame is sent in place of the next scheduled SP frame. In order to obtain sufficient statistics, 100 different realizations of each channel have been considered, randomly selecting the starting point of the error trace.

5. RESULTS

As first outcome of the transmissions, the average temporal error propagation of the two schemes has been compared, as shown in Fig. 8. The duration of the propagation depends

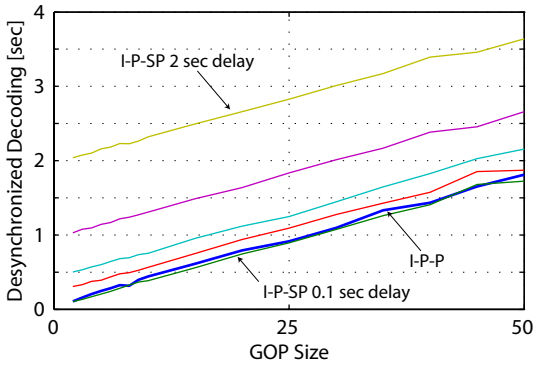

Figure 8: Average desynchronized decoding

strongly on the GOP size and on the feedback delay. The performance of the I-P-P scheme is almost identical to the I-P-S scheme with small delay times. Since, in average, the error occurs in the middle of the GOP, the propagation time increases linearly with the increasing GOP or, equivalently SD, being the delay a fixed offset as

$$\text{propagation\_time} = \text{delay} + [\text{GOP}, \text{SD}]/(2 \cdot fr). \quad (1)$$

The graph in Fig. 9 show the results of the simulations in

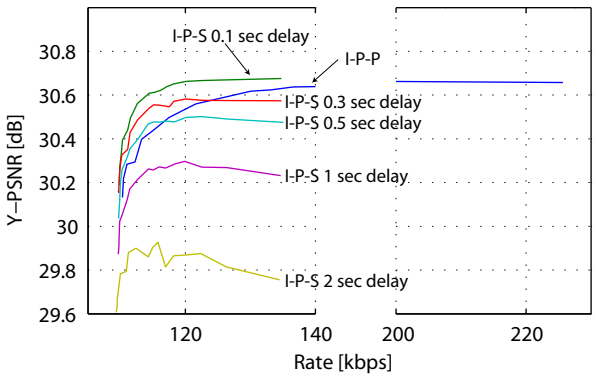

Figure 9: Rate-Distortion evaluation

terms of rate-distortion when comparing the I-P-P scheme with the I-P-S. This first graph considers both QPSPs equal to 33. The I-P-SP scheme outperforms the I-P-P scheme only in case the considered delay is around 0.1 s. The performance of the I-P-S scheme decreases considerably with increasing delay and, surprisingly, the quality decreases with decreasing SD. This is due to the fact that, for high delays, the slightly lower quality of the S frames was dominant over shorter desynchronization times.

This approach, however, does not take into consideration neither the channel characteristics nor the delay constraints. In order to optimize the rate-distortion behavior, the QPSP1

and QPSP2 values have to be adapted to the transmission characteristics.

In order to measure the rate-distortion behavior when varying the QPSP1 and QPSP2, the resulting frame quality and size has been measured for SP, SI and P frames. In Fig. 10, the average size of the SP and SI frames, respectively, has been measured. For the both QPSPs, the quantization

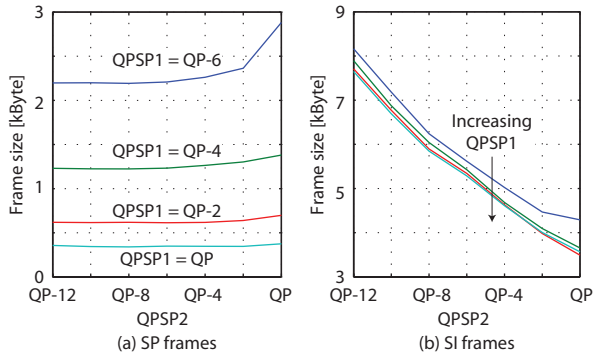

Figure 10: Size of the S frames as a function of the QPSPs

parameter has been expressed as a value relative to the P frames' QP, i.e. 35. The curves in Fig. 10(a) show the size of the SP frames, which is strongly dependent on the QPSP1 and starts depending on QPSP only for lower QPSP2. The behavior of the size of the SI frames, Fig. 10(b), shows a strong dependency on the QPSP2 and a moderate dependency on the QPSP1.

As the quality of the SP and SI frames coincide, a single graph is shown in Fig. 11(a). Consistently with theoretical

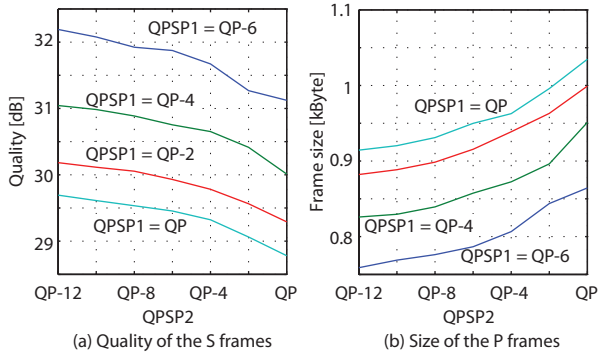

Figure 11: Qautlity of the S frames and size of the P frames as a function of the QPSPs

assumption, the quality is increasing both with decreasing QPSP1 and decreasing QPSP2. The size of the P frames is increasing with decreasing QPSP1 and QPSP2, as the quality of the pictures they are using as reference is decreasing as well. The quantization parameters have to be, therefore, optimized considering the probability of switching as small values of QPSP1 and QPSP2 would decrease the size of SP and P frames but increase the one of the SI frames. The optimization of the rate-distortion behavior has already been analytically solved by Setton and Girod in [10]. The optimal QPSP1 and QPSP2 values have been designed as a function of the QP of the P frames as well as the switching probability.

As the suitable QP of the P frames has been already defined, the probability of switching for the proposed channel measurements has been considered. The application of SI frames has been modeled as a state transition model with four states, as indicated in Fig. 12. The transition proba-

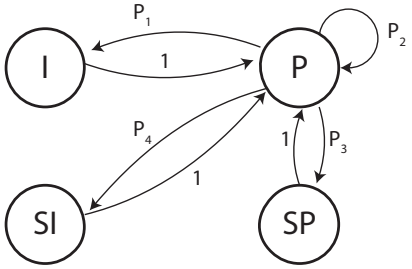

Figure 12: State transition model

bilities P are defined as follows:

- $P_1 = f(\text{GOP}) = 1/\text{GOP}$
- $P_3 = f(\text{SD,GOP,delay}) = 1/\text{SD}\cdot P_{SI}$
- $P_4 = f(\text{SD,GOP,delay}) = 1/\text{SD}\cdot(1 - P_{SI})$

They depend on the GOP size, on the S-Frame distance as well as on the probability of sending an SI frame in place of an SP frame (P_{SI}) as indicated in Fig. 13. For distances

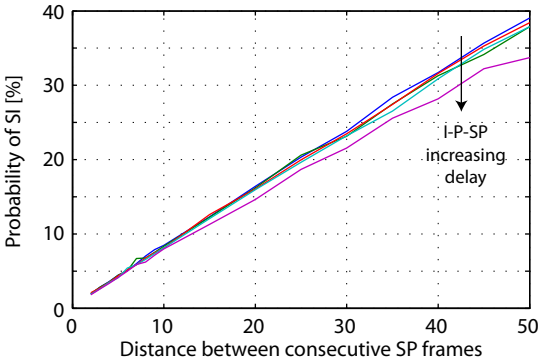

Figure 13: Probability of scheduling an SI

smaller then 15 frames between S frames, the probability of scheduling an SI frame remains below 10%. For increasing SD, the probability increases, since the probability that an error has occurred between two consecutive SP frames becomes higher. It reaches 40% for SD = 50. The probability is depending on the considered delay as well. With increasing delay, more SI frames are not sent because of feedack response issues, therefore the probability is slightly diminishing with increasing delay.

The graphs in Fig. 14 compares the rate distortion behaviour of the proposed schemes, when optimizing the QPSP2 as suggested in [10]. Major improvements are appreciated when using the refined I-P-S scheme. The quality, that previously saturated for increasing SD, is now increasing for shorter distance between consecutive S frames. I-P-S

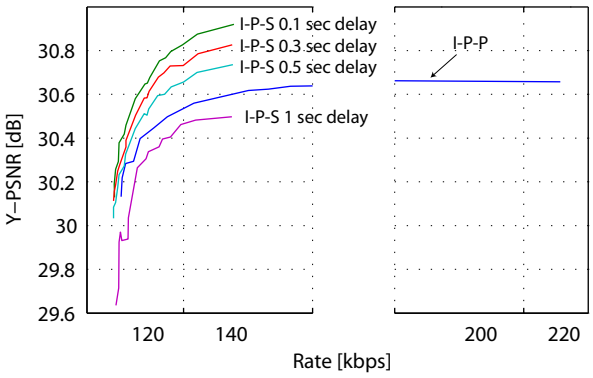

Figure 14: Rate-Distortion evaluation

schemes with feedback delays around 0.5 s still provide better performance than the classical I-P-P scheme.

The performed simulations bring us to the conclusion that the SP and SI frames might be beneficial under specific conditions. As shown in Fig. 14 their implementation is convenient as soon as the feeback delay remains smaller than 0.75 s. This time is sufficient to convey the feedack information to the RNC, but it remains questionable if the whole core network till the MS may be crossed in this interval. Although outside the scope of this work, complexity and power consumption issues have to be considered in order to select the optimal SD.

6. CONCLUSION

In this paper we analyzed the effectiveness of S frames in UMTS networks as a tool to reduce the temporal propagation, compared with the classical video streaming consisting of I and P frames. The investigation has been performed using measured error traces at transport block level. Different delay times, according to different the practical implementations, have been considered. For delay times smaller than 0.5 s, the application of SI frames is beneficial and increases the quality up to 0.3 dB for delay times equal to 0.1 s.

The benefits arising by the application of the SP and SI frames should also compared with an increasing level of complexity, when considering devices with limited capabilities. Such an investigation, however, cannot be performed using a development software such as JM. Another issue regards the usage of packet retransmission in place of SP and SI frames. Retransmitting the damaged packet is convenient when limiting the observation to the net data rate. Retransmission, however, has to cope with strict delay constraints set by the playout buffer. In case the round trip time is higher than the time span covered by the playout buffer, the retransmitted packet is received already outdated. SP and SI frames, despite the increase in complexity and average data rate, relax the delay constraints. The decoding resinchronization occurs by replacing packets *in the future*, whereas retransmission requires the replacement of a specific damaged packet. Hybrid broadcasting schemes can be therefore implemented by means of SP and SI frames. One single sequence, consisting of I, P and SP frames is broadcasted to all the users. In case one user receives damaged packets, the appropriate SI frame can be transmitted using a unicast parallel channel.

7. REFERENCES

[1] ITU-T. ITU-T Recommendation H.264 : Advanced video coding for generic audiovisual services, November 2007.

[2] 3GPP. Transparent end-to-end transparent streaming service; Protocols and codecs. TS 26.234, 3rd Generation Partnership Project (3GPP), 2008.

[3] M. Karczewicz and R. Kurceren. The sp- and si-frames design for h.264/avc. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13(7):637–644, July 2003.

[4] Eric Setton and Bernd Girod. Rate-distortion analysis and streaming of SP and SI frames. *IEEE Trans. Circuits Syst. Video Techn.*, 16(6):733–743, 2006.

[5] W. Karner, O. Nemethova, P. Svoboda, and M. Rupp. Link Error Analysis and Modeling for Video Streaming Cross-Layer Design in Mobile Communication Networks. *ETRI Journal*, 29(5):569–595, 2007.

[6] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A Transport Protocol for Real-Time Applications. Technical Report 3550, July 2003. Updated by RFC 5506.

[7] J. Postel. User Datagram Protocol. Technical Report 768, IETF, August 1980.

[8] J. Postel. Internet Protocol. Technical Report 791, IETF, September 1981. Updated by RFC 1349.

[9] H.264/AVC JM Reference Software, August 2008.

[10] E. Setton and B. Girod. Rate-Distortion Analysis and Streaming of SP and SI Frames. 16(6):733–743, June 2006.