Interactive Error Control for Mobile Video Telephony

Waqar Zia, Tauseef Afzal
Dept. of EE and IT, Technische Universität München, Arcisstraße 21, 80333, Munich, Germany
Email: waqar.zia@tum.de, tauseefafzal@hotmail.com

Wen Xu
Infineon Technologies AG Am Campeon 1-12 85579 Neubiberg, Germany
Email: wen.xu@ieee.org

Guenther Liebl, Thomas Stockhammer
Nomor Research GmbH Brecherspitzstrasse 8 81541 Munich, Germany
Email: {liebl,stockhammer}@nomor.de

Abstract—Error robust video communication for hand-held devices is a delicate task because of limited computational resources and hostile channel conditions in mobile environments. The loss of coded video data on the channel can result in spatio-temporal error propagation in the video. In addition, stringent end-to-end delays for conversational applications make this challenge even more difficult. In this work, we investigate several techniques which exploit feedback from the receiver to enhance the performance of conversational video in realistic mobile communication environments. Specifically, we show how a low-complexity interactive error tracking technique can be combined with multiple reference picture selection (RPS) based on the existing syntax of H.264/AVC. This technique outperforms other interactive error protection strategies by a margin of more than 2 dB for moderate channel loss rates, with minimal impact on end-to-end delays.

I. INTRODUCTION

Recent advances in mobile communication technology have enabled digital video communication via mobile devices. While being considered for a long-time, only the introduction of other video services, the availability of appropriate hardware and increased data rates will allow conversational video with sufficient quality. Currently, the enhancement of video telephony services is a subject of ongoing activities within 3rd generation partnership project (3GPP). While mobile transmission in general is challenging due to severe fading and burst errors on the radio channel, the task is even more difficult when it comes to video communication: the temporal and spatial dependencies in the sequence typically result in spatio-temporal error propagation in the case of a single loss. However, due to the already large bandwidth requirements of video applications, stronger channel coding is a difficult option to employ, since it increases the bandwidth requirements even further. As a consequence, regardless of the level of protection, residual bit errors still happen in mobile environments, leading to loss of coded video data. If the video application continues to make “loss unaware” encoding decisions, spatio-temporal error propagation will result in severely degraded video quality. Hence, some sort of error robustness must be added to the application.

H.264/AVC is the state-of-the-art codec of choice for robust video communication. It provides a range of tools that are tailored for error-resilient communication. These tools have been the focus of many previous investigations: for example, slice-structured coding and flexible macroblock ordering (FMO) [1] provide a basic degree of robustness in an open-loop video communication system. For a more robust system, long-term memory (LTM) motion compensated prediction (MCP) [2], along with average statistical information about the channel has been employed for selecting optimal mode decisions in [3], [4], [5]. Further techniques that use feedback from the receiver on the current packet loss have been investigated in [6], [7], [8], [9] in the context of ARQ-type solutions in combination with accelerated retroactive decoding (ARD). In [10] the feedback from the receiver is used to predict the lost regions in the reference frames in order to modify the encoding decisions. Finally, [11] states that the best performing techniques with least complexity are those that employ instantaneous feedback information about the current loss situation. However, despite their theoretical gains, most of these techniques cannot be directly applied to the next generation of packet-switched conversational (PSC) applications in 3GPP, mainly due to stringent end-to-end delay requirements and complexity constraints on mobile phones. Hence, in this work we investigate possible ways to integrate simple interactive error control (IEC) strategies into a realistic mobile communication scenario. Several objective and subjective assessments will be provided that allow to develop a better understanding of the system and to identify the best alternative for error-resilient video transmission. In addition, we will present a novel IEC approach which combines feedback-based error tracking with multiple reference picture selection (RPS) and compare it to existing alternatives.

The mobile video communication system we consider in this work is introduced in Section II along with its main constraints. The suitability of different error mitigation techniques for this system is also discussed. Next, we give a detailed description of the investigated IEC techniques in Section III. The performance evaluation criteria and simulation parameters are presented in Section IV, followed by a discussion of the various performance results. Some concluding remarks can be found in Section V.
II. CONVERSATIONAL PACKET SWITCHED VIDEO SERVICES

A considerable performance enhancement is expected from the next generation of mobile video communications [12]. Therefore, 3GPP mobile video telephony services will be enabled by conversational packet switched multimedia services [13], which are all based on the IP multimedia subsystem (IMS). Figure 1 shows a block diagram of a possible end-to-end application scenario:

![Block diagram of a possible end-to-end scenario for conversational video applications.](image)

**Fig. 1.** A possible end-to-end scenario for conversational video applications.

Terminal A is a mobile device that is connected via a bidirectional packet-data access link, such as provided by high speed packet access (HSPA), to the base station. Compressed and packetized video data, as well as other control information from the terminal, is transferred to the base station, which conveys it over the core network to the remote Terminal B. The latter could be a terrestrial video phone, a computer system, or a hand-held device itself. A similar information transfer occurs in the reverse direction from Terminal B to Terminal A.

The protocol stack for this packet-switched conversational video service is depicted in Figure 2: The conversational application sits atop an H.264/AVC video codec and/or other multimedia, like audio, text, etc. The network abstraction layer of the video codec handles compressed video data that is contained within RTP packets [14], which are communicated to and from the lower layers. The latter take care of forwarding packets to other network elements, as well as of detecting corrupted packets and discarding them at the receiving side.

![Protocol stack for PSC applications.](image)

**Fig. 2.** Protocol stack for PSC applications.

Control information can be generated either by the application or the codecs and is transported to the remote terminal via the audio-visual profile with feedback (AVPF) extension of RTP [15]. This profile is well-suited for transporting feedback information from the video decoder to the encoder side. In addition to the bit rate limitation and high loss encountered on the wireless link in Figure 1, the application is subject to stringent end-to-end delay requirements in order to provide a meaningful conversational service. For example, in [16], this value is prescribed to be 100 ms for video to achieve lip-synchronization. At the same time, at least one of the terminals in the above mentioned scenario shall be a hand-held device, which is supposed to perform real-time H.264/AVC encoding and decoding simultaneously, along with other processes like speech coding and decoding. It is thus needless to say that this leaves little room for employing error-resilience techniques with considerable complexity on a battery operated device with a low power processor.

At the same time the stringent delay requirement also provides opportunities: it is known that feedback-based error resilience techniques perform better for smaller delays [11]. In addition, a bidirectional communication link with the possibility of allowing control traffic makes low-complexity feedback-based techniques an ideal choice for adding error robustness.

Hence, several interactive error control (IEC) strategies have already been proposed in the literature. However, employing statistical information on the channel condition for optimal mode selection is reported in [11] to have significant complexity, as well as worst performance for longer delays compared to techniques that are based on instantaneous feedback information. Moreover, ARD as proposed in [8] cannot be assumed to work in a complexity-constrained environment. Thus, any packet arriving after its nominal decoding time will generally be of no use in mobile conversational environments. Finally, error tracking with pixel-based accuracy using motion vector information, as proposed in [10], also seems far too complex to be employed directly. For this reason, we will introduce some practical error control techniques in the following section that can be integrated in such a complexity- and delay-constrained environment.

III. IEC STRATEGIES

The different interactive error control techniques that have been investigated in this work are described below. All of them are based on feedback from the receiver using RTCP messages [14]. The encoder then uses this feedback to update the encoding process accordingly. However, the frequency and the level of detail of the feedback information may vary for each strategy from simple estimated packet loss rates to indication of the exact packet that was lost.

A. Average RIR Tuning

Random intra refresh provides a simple means to reduce error propagation in error-prone video transmission. The only parameter is the random intra macroblock refresh (RIR) rate defined as the ratio of intra-coded macroblocks (MBs) to the total number of MBs in a picture. For **Average RIR Tuning**, feedback messages are used to report estimated packet error rates (PER) from the receiver to the transmitter. The encoder then adjusts its RIR rate according to the reported PER. In this work we have used the following piece-wise linear relationship...
between PER and RIR:
\[
RIR = \begin{cases} 
1.33 \cdot PER & \text{if } PER \leq 0.15, \\
0.2 & \text{otherwise.} 
\end{cases} 
\tag{1}
\]
Note that Eq. 1 represents a good heuristic approximation of the exact relationship, for which a closed-form solution usually not exists.

B. Instantaneous RIR Tuning

For this technique, the loss of a packet (without referencing its sequence number) is indicated to the transmitter. Upon receiving such feedback, the RIR rate is increased instantaneously to a peak value and is then reduced with each (correctly received) frame according to a specific percentage. The instantaneous RIR rate \( \rho \) for a given frame index \( s \) is calculated as
\[
\rho = \alpha \cdot \beta^{s-s'}, 
\tag{2}
\]
where \( s' \) is the index of the frame considered for encoding when the most recent loss event was received by the encoder. The first parameter \( \alpha \) is the maximum allowed additional RIR rate per frame that may be used for error mitigation. The second parameter \( \beta < 1 \) determines how quickly the transient state of generating additional intra MBs disappears. Both \( \alpha \) and \( \beta \) are also selected heuristically to yield maximum performance enhancement while minimizing the impact of the additional encoding buffering delays due to the low compression efficiency of intra-coded MBs.

C. Intra-Frame Update

Instead of increasing the rate of intra-coded MBs upon receiving an indication of a lost packet from the receiver, a complete I-frame (picture that is encoded entirely in INTRA mode) is sent here to stop error propagation. However, I-frames typically require several times more bits than P-frames (pictures encoded with reference to previous frames) for encoding. Thus, compared to broadcasting scenarios, where I-frames are mandatory such that receivers can randomly access the video stream, their use over low-bit-rate point-to-point connections, as considered here, should be restricted as much as possible. Nevertheless, we have added this technique as a benchmark for comparison.

D. IEC with Error Tracking and Multiple Reference Frames

We will now present a novel IEC technique that uses error tracking in combination with multiple reference picture selection (RPS). The main idea is depicted in Figure 3: in this example, a packet transmitted by the encoder at time \( t - 3T \) is lost. The receiver detects this and reports the sequence number of the lost packet immediately to the transmitter. Note that reporting may be done differentially, i.e. multiple collocated losses can be contained within a single report. Experiments show that for slice-structured coding with small slice size, multiple collocated losses are most common and this technique saves feedback bandwidth. Furthermore, the encoder keeps a record of the most recent packets it has transmitted and the corresponding reference areas used for predictive encoding of the content in each packet. Now assume that the loss report reaches the transmitter just before it encodes the frame at time \( t \). Hence, the sequence number of the lost packet is translated into the corresponding lost reference region of frame \( t - 3T \), and error tracking can be applied. Since pixel-accurate error tracking, as described in [10], is computationally too complex for hand-held devices, we propose the following suboptimal strategy: the lost area in a reference frame shall be assumed to grow at most with a rate equal to the motion vector search range for each temporally predicted frame. For our simple example in Figure 3, the arrows indicate the assumed maximum propagation of errors in each of the subsequently predicted images at \( t - 2T \) and \( t - T \). The shaded region in each image is therefore considered to be lost by the encoder, and will not be used in the mode selection and motion estimation process when encoding the frame at time \( t \). For example, if a total of 4 reference frames is allowed, the whole frame at \( t - 4T \), as well as the unshaded parts of the following three frames may be used in the inter-prediction process. Hence, the spatio-temporal error propagation is terminated at time \( t \).

The impact of assuming a larger region of the image to be lost on the encoding efficiency is not too significant, if a small round trip time for the feedback is possible. The simulation results in Section IV will show the suitability of our low-complexity error tracking algorithm for the given scenario.

IV. PERFORMANCE RESULTS

The structure of the simulation environment we have used for generating the results has been documented in [17]. While we have done investigations for different combinations of video sequence, frame rate, and bearer type, only results for one specific combination are shown here due to the limited space available. The respective parameters are as follows:
The video sequence (seq 1) is a concatenation of three test sequences approved by the video ad-hoc group within 3GPP: Stunt, Walk, and Friends.

The spatial resolution of the sequence is QCIF, the temporal resolution is 15 fps, and the length is equal to 850 frames, i.e. about 1 minute length.

A single reference frame and no slices are used in the encoder, unless stated otherwise.

The radio bearer has bit rate 128 kbps, a TTI of 20 ms, an RLC-PDU size of 320 byte, and adjustable RLC-PDU loss rate.

Note that all sequences and software tools are available from 3GPP [18].

In the following, we will compare the performance of the first three simple IEC techniques described in Section III-A-III-C to the No Feedback case (with fixed RIR rate of 5%) using different quality evaluation metrics. Then we will show the gain achievable by our proposed technique in Section III-D over the best one among the simple strategies. In the end, we will also present some visual results for all of these techniques.

A. Achievable Reconstruction Quality at the Transmitter

We will start by investigating the effect of considering feedback in the encoding process on the maximum achievable reconstruction quality of the sequence. This will give us an idea of the price we have to pay in terms of source coding efficiency, if we want to add error resilience to our system. The respective quality metric will be the so-called Encoded PSNR of MSE, which is the Y-PSNR value corresponding to the mean squared error (MSE) averaged over the luminance component of each frame in the (fully) reconstructed sequence at the transmitter.

Figure 4 shows the Encoded PSNR of MSE vs. increasing RLC-PDU loss rate on the wireless link (which in turn leads to increased RTP packet loss) for the No Feedback case and the three simple IEC techniques. The values of the parameters for Instantaneous RIR Tuning technique were selected by iterative tuning as $\alpha = 0.6, \beta = 0.5$. As expected, for the No Feedback method the Encoded PSNR of MSE does not change with increasing loss rate (i.e. it is a flat line), since no feedback messages are considered in the encoding process. Hence, the maximum achievable reconstruction quality here is governed by the fixed RIR rate of 5%. For the other three feedback-based techniques, lost packets at the receiver trigger feedback messages, based on which the encoding process at the transmitter is adapted: if there is no loss of RLC-PDUs (leftmost points in the diagram), the maximum achievable reconstruction quality is thus higher than in the No Feedback case, since neither intra-MBs nor I-frames are required. However, as higher loss rates directly lead to either an increase in the intra-MB refresh rate or an increase in the number of sent I-frames, the Encoded PSNR of MSE starts to decrease slightly due to the growing inefficiency of the source coding process. Finally, it is interesting to note here that the Intra-Frame Update technique achieves the lowest decrease in source coding efficiency compared to both techniques employing intra-MB update.

B. Achievable Reconstruction Quality at the Receiver

Now we want to investigate the performance of the different techniques when actual losses are applied to the transmitted video sequence. The respective quality metric will be the PSNR of MSE, which is the Y-PSNR value corresponding to the mean squared error (MSE) averaged over the luminance component of each frame in the (partially) reconstructed sequence at the receiver. Figure 5 depicts the PSNR of MSE at the receiver vs. increasing RLC-PDU loss rate on the wireless link for the No Feedback case and the three simple IEC techniques. As can be seen, the performance of Instantaneous RIR Tuning and Intra-Frame Update is better than for the No Feedback case, which indicates that these two techniques allow to react appropriately to information on individual packet loss. Especially the Instantaneous RIR Tuning technique achieves a
gain of more than 3 dB in PSNR for loss rates above 0.2%, while the Intra-Frame Update technique becomes more and more inefficient for loss rates above 0.7%. Finally, it should be noted that the performance of the Average RIR Tuning technique is not very convincing, especially for loss rates below 1%. This can be explained by the fact that the feedback in this case only contains an average indication of the current loss situation, which does not seem to be sufficient for the error recovery process.

C. Instantaneous PSNR

In order to get an impression of how fast the different IEC techniques recover from lost video content, we will next examine the Instantaneous PSNR, which is the Y-PSNR value corresponding to the mean squared error (MSE) of the luminance component of each individual frame in the (partially) reconstructed sequence at the receiver.

Figure 6(a)-6(d) contain the Instantaneous PSNR of the first 200 frames in the received sequence for an average RLC-PDU loss rate of 1.5%.

![Fig. 6. Instantaneous PSNR](image)

It can be observed that all the feedback-based methods recover earlier than without feedback. While Average RIR Tuning and Instantaneous RIR Tuning exhibit a quite similar speed of recovery, the ultimately achievable PSNR (without loss) is larger for the second strategy. The best results in terms of fast error recovery, however, are found for Intra-Frame Update.

D. Buffer Occupancy

The cost of fast recovery are often large fluctuations in the output bit rate of the video encoder. In Figure 7, this tendency is visualized by showing how the codec buffer content varies for the different techniques: while Instantaneous RIR Tuning only causes medium spikes, Intra-Frame update leads to large deviations from the average frame size in the video sequence. If the offered data rate on the transmission links is relatively fixed, this typically leads to increased end-to-end delay and resulting late losses.

![Fig. 7. Buffer Occupancy Comparison](image)

E. Performance Gain of IEC with Error Tracking and Multiple Reference Frames

From the above results, it can be concluded that providing the encoder with feedback on the loss situation at the receiver is definitely beneficial. Next, we will show how our proposed strategy in Section III-D performs in comparison to the simple IEC strategies.

Slice-structured coding is used with 200 byte per slice for all compared strategies. In addition, a total of 5 reference frames is now allowed at the encoder. As a benchmark, we used the previously best technique, i.e. Instantaneous RIR Tuning, allowing the same number of reference frames and slice size for the sake of fairness. Iterative tuning for the given codec configuration yields the most suitable parameter values of $\alpha = 0.6$ and $\beta = 0.25$.

Figure 8 contains the respective results in terms of the PSNR of MSE at the receiver vs. increasing RLC-PDU loss rate. We want to note here first that the slight degradation of both strategies for the case of no RLC-PDU loss (the leftmost point) is due to the in general poorer compression efficiency of slice-structured coding. However, the benefit of introducing slices in the encoding process becomes obvious for nonzero loss rates: our proposed strategy achieves a gradual
descent in average reconstruction quality and outperforms the previous best solution by more than 2 dB in PSNR of MSE for loss rates above 0.8%. Hence, we are able to improve the quality of conversational video applications without incurring an additional end-to-end delay.

F. Visual Results

Finally, we want to show some visual results for the implemented techniques in Figure 9. Therefore, we have selected a frame in the received sequence which directly follows a loss event. In order to achieve a fair comparison, all investigated strategies now use the same parameters as in Section IV-E. We have arranged the plots in the order of ascending visual quality (except the first one, which represents the frame in the undistorted sequence). The visual impression closely matches the tendencies observed in the previously discussed quantitative results: clearly, all feedback-based techniques contain less block artefacts than the No FeedBack method. Moreover, IEC with Error Tracking also yields best subjective performance.

V. CONCLUSIONS

In this work we have investigated the performance of various interactive error control (IEC) techniques for a mobile video telephony environment when exploiting the available feedback channel. It could be shown that feedback-based strategies yield a considerable gain compared to the case with no feedback. Among the interactive error control techniques, solutions employing multiple reference frames perform better than their simpler counterparts which employ only a single reference frame. The cost, however, is increased complexity and buffer size. In addition, we found that the faster and more specific the feedback messages are, the larger is the achievable gain.

We want to mention at this point that during our investigations, we also experienced a large dependence of the achievable performance of the IEC strategies both on the image content and the bearer characteristics (we omitted the respective results here due to the limited space available). Therefore, an optimized solution should be able to dynamically adapt its parameters to both the source and the underlying transport channel.

A possible extension of this work is to drop the assumption of a perfect feedback channel from the receiver, i.e. also take into account practical timing aspects, control protocols, link errors, and delays of the feedback messages. Furthermore, since techniques that employ multiple reference frames are a new concept in a mobile video communication environment, further research on practical implementation issues, like available memory space on the terminals, is necessary.

REFERENCES