An Analytical Model of the End-to-End Performance for Linear Video Delivery Under Bandwidth Constraints

Anthony Trioux, M Gharbi, François-Xavier Coudoux, Patrick Corlay

An Analytical Model of the End-to-End Performance for Linear Video Delivery Under Bandwidth Constraints

Anthony TRIOUX, Mohamed GHARBI, François-Xavier COUDOUX, Patrick CORLAY
UMR 8520 - IEMN, DOAE, Univ. Polytechnique Hauts-de-France, CNRS, Univ. Lille, YNCREA, Centrale Lille, F-59313 Valenciennes, France

Abstract: *Recently, Linear Video Coding and Transmission (LVCT) schemes have been proposed as an alternative to traditional video transmission schemes, the latter experiencing cliff-effect [3] in wireless error-prone environments. In this paper, we propose an analytical model to estimate/predict the end-to-end performance of SoftCast [2] video transmission, pioneer of the LVCT schemes. This model, based on the PSNR metric, accounts for channel conditions (CSNR) as well as data compression due to bandwidth constraints. We show that regardless of the available channel bandwidth, the end-to-end video quality can be accurately modeled and predicted according to the characteristics of the video and the channel conditions.*

Keywords: Wireless Video Transmission, SoftCast, Distortion Model, Bandwidth constraints, Linear Video Delivery.

1 Introduction

The basic scheme of SoftCast [2] is introduced in Fig. 1. SoftCast first takes a Group of Pictures (GoP) and uses a 3D full-frame DCT as a decorrelation transform. These DCT frames are then divided into N small rectangular blocks of transformed coefficients called *chunks*. In the SoftCast scheme, the data compression can be done after the decorrelation transform. Specifically, when the available channel bandwidth for the transmission is less than the signal bandwidth, SoftCast discards chunks with less energy. This is generally the case especially for the transmission of High Definition (HD) content as mentioned in [5, 6]. At the receiver side, these discarded chunks are replaced by null values [2].

For ease of reading, the compression level denoted hereafter CR is usually used in SoftCast and in LVCT schemes. It is defined as follows [4]:

$$\mathrm{CR} = \frac{K}{N} \tag{1}$$

where K represents the number of transmitted chunks per GoP and N the total number of chunks within a GoP. This ratio is between 0 (no information transmitted, i.e., $K = 0$) and 1 (no compression, i.e., $K = N$).

The third block at the transmitter level called Power Allocation or Scaling is used to provide error resilience. SoftCast scales the magnitude of the DCT coefficients to offer a better protection against transmission noise. Since the total transmission power available P is limited and fixed, it is distributed to all the chunks in a way that minimizes the Mean Square reconstruction Error (MSE) between transmitted and decoded chunks. This is a typical Lagrangian problem which leads to the following solution [1, 2] given by:

$$g_i = \lambda_i^{-1/4} \cdot \sqrt{\frac{P}{\sum_i \sqrt{\lambda_i}}}, \tag{2}$$

where $g_i, i = 1, 2, \ldots, N$ is the scaling factor for the i^{th} chunk, and λ_i the energy of the ith transmitted coefficient (after 3D-DCT) [2].

The Hadamard transform is then applied to the scaled chunks to provide packet loss resilience. This process transforms the chunks into *slices*. Each slice is a linear combination of all scaled-chunks.

Finally, these packets are transmitted in a pseudo-analog manner using Raw-OFDM [2], i.e., classical coding (e.g., Forward Error Correction code) and modulation stages are skipped.

In parallel, the SoftCast transmitter sends an amount of data referred as metadata. These data consist of the mean and the variance of each transmitted chunk as well as a bitmap, which indicates the positions of the discarded chunks into the GoP. Metadata are strongly protected and transmitted in a robust way (e.g., BPSK [2]) to ensure correct delivery and decoding process.

At the receiver side, if an estimation of the channel noise is available, a Linear Least Square Error (LLSE) decoder can be used to get the best estimation of the received values. Otherwise, a Zero-Forcing (ZF) decoder is used. Using the metadata, the decoded values are then reassembled to form DCT-frames, which are then passed through an inverse 3D-DCT process.

In a recent paper [8], Xiong et al. modeled the end-to-end performance of SoftCast for any channel Signal-to-Noise Ratio (CSNR, expressed in decibels).

They showed that the total distortion that affects the reconstructed video quality without data compression can be obtained from:

$$\mathrm{D}_{[\mathrm{ZF/FB}]} = \sum_{i=1}^{N} D_i = \frac{\sigma_n^2}{P} \left(\sum_{i=1}^{N} \sqrt{\lambda_i} \right)^2, \tag{3}$$

where σ_n^2 is the noise power.

Based on the following definition of the CSNR and PSNR expressed in decibels:

$$\mathrm{CSNR} = 10\log_{10}(\bar{P}/\sigma_n^2), \quad \bar{P} = P/N, \tag{4}$$

$$\mathrm{PSNR} = 10\log_{10}(255^2/\bar{D}), \quad \bar{D} = \mathrm{D}_{[\mathrm{ZF/FB}]}/N. \tag{5}$$

They showed that the expected reconstructed video quality can be finally obtained from:

$$\mathrm{PSNR}_{[\mathrm{ZF/FB}]} = c + \mathrm{CSNR} - 20\log_{10}\big(H\big), \tag{6}$$

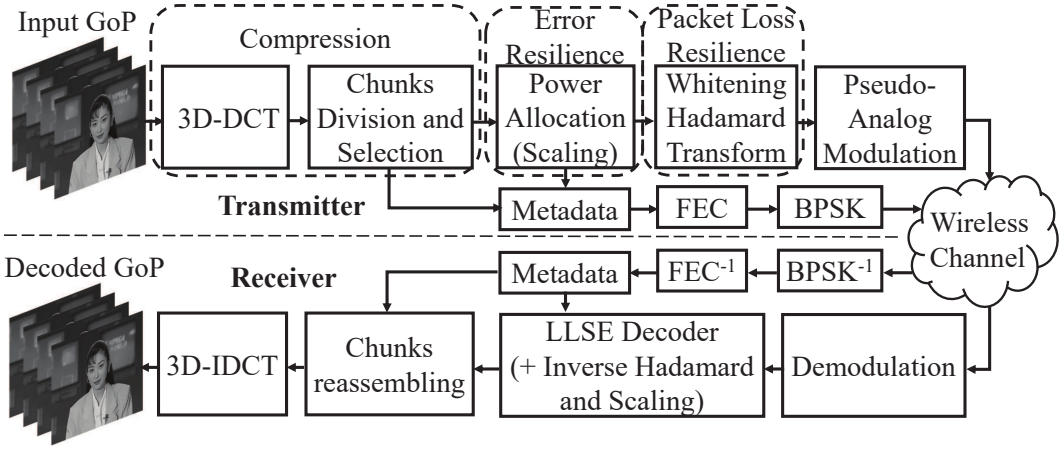

Figure 1: Block diagram of the SoftCast scheme.

where $c = 20\log_{10}(255)$ and

$$H = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\lambda_i}, \tag{7}$$

refers to the *data activity* of the video content [8]. The higher the data activity H, the lower the reconstructed PSNR, showing the importance of taking into account the characteristics of the transmitted video content in a SoftCast context. Note the linear characteristic of the $\mathrm{PSNR}_{[\mathrm{ZF/FB}]}$ that depends on the channel transmission conditions.

We note that Xiong's model relies on two assumptions: the first is that the available bandwidth of the application allows the transmission of the N elements of $\mathbf{x}$ (*i.e.*, CR=1, no compression applied). However, in practice, this is generally not the case since bandwith ressources are limited. This is especially true when considering the transmission of high resolution (HD, 4K, etc.) video formats. The second hypothesis assumes that a ZF estimator used at the receiver side. This is not valid when considering the original SoftCast scheme proposed by [2] which uses an LLSE decoder.

2 Proposed work

The first objective of this work is to consider the more realistic and general case *i.e.* only the $K \leq N$ largest energy chunks are transmitted due to bandwidth constraints.

The challenge consists in proposing a more realistic theoretical model than Xiong's model [8], i.e., addressing the weaknesses of the initial model (inaccurate prediction for bandwidth constrained applications).

Since $N-K$ chunks are discarded due to bandwidth constraints, the total distortion $\mathrm{D}_{[\mathrm{ZF/CB}]}$ now consists of two parts:

- The distortion D_i that affects each of the K transmitted coefficients x_i, given by: $D_i = E\left[(\hat{x}_i - x_i)^2\right]$. For ease of reading, let us denote the total distortion due to the transmitted coefficients $D_s = \sum_{i=1}^{K} D_i$.
- The distortion D_j due to each of the $N-K$ discarded coefficient x_j, given by: $D_j = E\left[(0 - x_j)^2\right]$. Likewise, we denote the total distortion due to the discarded coefficients $D_d = \sum_{j=K+1}^{N} D_j$.

Therefore, the overall distortion (3) becomes:

$$\begin{aligned}\mathrm{D}_{[\mathrm{ZF/CB}]} &= D_s + D_d, \\ &= \frac{\sigma_n^2}{P}\left(\sum_{i=1}^{K}\sqrt{\lambda_i}\right)^2 + \sum_{j=K+1}^{N}\lambda_j.\end{aligned} \tag{8}$$

We note that the average transmission power in (4) becomes $\bar{P} = P/K$ as the total transmission power is here distributed over the K transmitted coefficients and in (5) $\bar{D} = \mathrm{D}_{[\mathrm{ZF/CB}]}/N$.

By inserting (8) into (5), we get:

$$\begin{aligned}\mathrm{PSNR}_{[\mathrm{ZF/CB}]} &= 10\log_{10}\left(\frac{255^2 \cdot N}{D_s + D_d}\right), \\ &= c - 10\log_{10}\left(1 + \frac{D_d}{D_s}\right) + 10\log_{10}\left(\frac{\bar{P}}{\sigma_n^2}\right) \\ &\quad - 10\log_{10}\left(\frac{1}{NK}\left(\sum_{i=1}^{K}\sqrt{\lambda_i}\right)^2\right).\end{aligned} \tag{9}$$

By analogy with (6), we identify the new data activity of the transmitted coefficients as:

$$H_t = \frac{1}{\sqrt{NK}}\sum_{i=1}^{K}\sqrt{\lambda_i}. \tag{10}$$

For ease of reading we also define E_d, the overall energy of all dropped coefficients:

$$E_d = \frac{1}{N}\sum_{j=K+1}^{N}\lambda_j. \tag{11}$$

According to these new definitions, the end-to-end video quality considering bandwidth constraints for the ZF estimator is finally given by:

$$\begin{aligned}\mathrm{PSNR}_{[\mathrm{ZF/CB}]} &= c + \mathrm{CSNR} - 20\log_{10}\left(H_t\right) \\ &\quad - 10\log_{10}\left(1 + \frac{\mathrm{CSNR}_{lin} \cdot E_d}{{H_t}^2}\right).\end{aligned} \tag{12}$$

where $\mathrm{CSNR}_{lin} = \frac{\bar{P}}{\sigma_n^2}$.

The above equation includes a new term in comparison to (6) that reflects the effect of the data compression applied. The PSNR now depends on three parameters: first, the CSNR which depends on the transmission conditions, and then the two other terms E_d and H_t that are directly related to the video data characteristics. For a given bandwidth, the higher E_d, the greater degradation. However, as E_d is multiplied by the CSNR_{lin}, the degradation becomes less noticeable at low CSNR environments.

When $K = N$, *i.e.*, CR=1, (12) and (6) are identical. In other words, the video quality scales linearly with the CSNR as stated in [8].

3 Results

To evaluate the effectiveness of the proposed model, we perform full end-to-end simulations. We create a Mixed sequence by slicing the first 128 frames of ten HD 720p sequences (*Ducks, Four People, In to tree, Johnny, Kristen and Sara, Old town, Parkjoy, Parkrun, Shields* and *Stockholm*). We use GoPs of 16 frames and divide each frame into 64 chunks as it represents the original and mostly used configuration [2]. We verified similar results for other GoP-sizes and chunk-sizes. We then consider four different compression ratio $\text{CR} = 1, 0.75, 0.5, 0.25$.

As observed in Fig. 2, the proposed model (colored lines) perfectly matches the simulation results (colored dots), regardless of the considered CR or CSNR values. When CR=1 (red color), we logically obtain the same linear characteristic as in [8].

However, Xiong's model represented with the red line [8] is no longer valid when the available channel bandwidth decreases (cyan, green and blue dots) as the data compression is not considered. In practice, it is mandatory to consider such loss since it drastically degrades the received video quality and implies non-linear characteristics at high CSNR values. This is the well-known leveling-off effect [5] that appears and implies huge decibel losses (e.g. ΔPSNR up to 20dB for the considered case). Unlike Xiong's model, ours (colored lines) perfectly predicts and models this leveling-off effect regardless of the amount of discarded chunks.

4 Conclusion and perspectives

In this paper, we present a theoretical model that can be used in the context of linear video delivery under bandwidth constraints. In contrast to Xiong's model [8], the proposed model takes into account the losses due to data compression/bandwidth constraints. Regardless of the available channel bandwidth, results show that the model accurately represents the full end-to-end performance by predicting the leveling-off [5] effect that appears when some chunks are discarded. This model can help for parameters optimization in an LVCT transmission context subject to bandwidth limitations as in [6]. It can also be used to quickly evaluate schemes without requiring extensive end-to-end simulations. Further works concern the extension of the model to different versions of SoftCast that bring additional PSNR gain (e.g., by taking into account the LLSE estimator) or the extension of the model to other objective metrics such as SSIM [7].

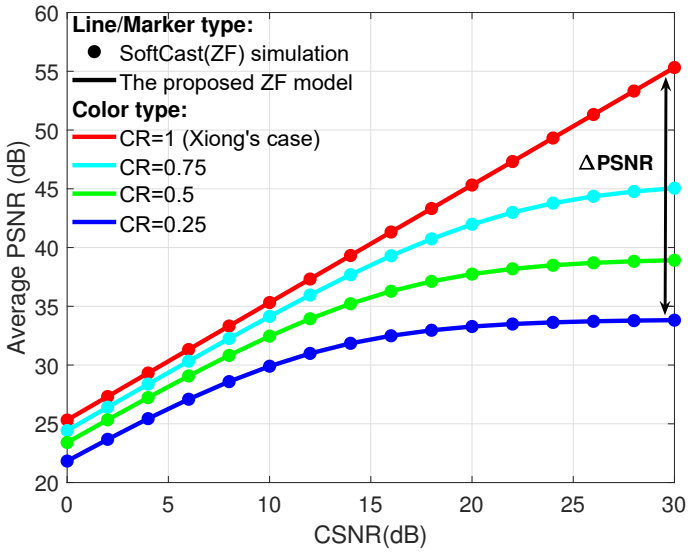

Figure 2: Average PSNR results for the proposed theoretical model (solid lines) and SoftCast simulations with ZF estimator: (dots markers) for the *Mixed HD720p* sequence. Configuration: GoP-size=16 frames, 64 chunks/frame. The colors red, cyan, green and blue represent CR=1, 0.75, 0.5 and 0.25, respectively.

References

[1] T. Fujihashi, T. Koike-Akino, T. Watanabe, and P. V. Orlik. High-Quality Soft Video Delivery With GMRF-Based Overhead Reduction. *IEEE Transactions on Multimedia*, 20(2):473–483, February 2018.

[2] Szymon Jakubczak and Dina Katabi. SoftCast: Clean-slate scalable wireless video. *MIT Technical report*, February 2011.

[3] S. Kokalj-Filipović and E. Soljanin. Suppressing the cliff effect in video reproduction quality. *Bell Labs Technical Journal*, 16(4):171–185, March 2012.

[4] Zexue Li, Hancheng Lu, and Yanglong Wu. Compressed uncoded screen content video transmission in bandwidth-constrained wireless networks. In *IEEE Int. Conf. Wireless Commun. & Signal Process. (WCSP)*, pages 1–5, November 2016.

[5] F. Liang, C. Luo, R. Xiong, W. Zeng, and F. Wu. Superimposed Modulation for Soft Video Delivery with Hidden Resources. *IEEE Trans. Circuits Systems Video Technol.*, 28(9):2345–2358, September 2018.

[6] Anthony Trioux, François-Xavier Coudoux, Patrick Corlay, and Mohamed Gharbi. Temporal information based GoP adaptation for linear video delivery schemes. *Signal Processing: Image Communication*, 82:115734, March 2020.

[7] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, April 2004.

[8] Ruiqin Xiong, Feng Wu, Jizheng Xu, Xiaopeng Fan, and al. Analysis of decorrelation transform gain for uncoded wireless image and video communication. *IEEE Trans. Image Process.*, 25(4):1820–1833, April 2016.