Video Quality Evaluation in IP Networks

Miguel Chin

Instituto Superior Técnico
Technical University of Lisbon
Lisbon, Portugal

***Abstract* – In this paper, no-reference objective quality metrics for encoded videos transmitted over a lossy channel are proposed and evaluated. The metrics consider both the effects of H.264/AVC video encoding and packet losses over IP networks. All the algorithms predict the video perceived quality based on elements extracted from the compressed bitstream and on the information about the packet losses, taken from the headers of the arriving packets. Five quality assessment metrics for H.264/AVC videos impaired by packet losses are presented: 1) a simple model that accounts for the packet loss ratio; 2) a model that considers the frame type where the correspondent losses occur; 3) a model that considers the frame type and the movement in the video under analysis; 4) a model that considers the frame type, the movement and the dependencies between frames; 5) a model that considers the frame type and statistical metrics taken from the packet loss pattern. The fifth model provided the best results with the resulting quality prediction being well correlated with subjective assessment data.**

Index Terms – No-reference video quality, H.264/AVC, Video over IP, Packet losses.

1. INTRODUCTION

Video transmission over Internet Protocol (IP) networks is a growing market. As competition between video service providers increases, the need to ensure that a service meets the end user's expectations becomes more relevant; since the success of a service provider is strongly dependent on the entire end user experience, there is as clearly a need for Quality of Experience (QoE) evaluation methods. The most reliable source of QoE data is naturally the end user. However, gathering QoE data from users requires expensive and time consuming subjective quality assessment tests. An alternative to subjective quality assessment is to automatically score the users perceived quality using objective metrics.

Objective picture quality metrics can be categorized into full reference (FR), reduced reference (RR) and no-reference (NR). FR metrics require both the original and distorted video to compute the video quality, while NR metrics only require the distorted video. RR metrics are somewhere in between the other two, since they require the distorted video and only some information about the original video.

This paper describes several bitstream based NR metrics for quality assessment of H.264/AVC video transmitted over IP networks. We start analyzing the effects of lossy H.264/AVC video encoding, resulting in a NR quality metric that accounts for the compression impact on video quality. Afterwards, we analyze the effects of video transmission (namely packet losses) and improve the previous NR model, by considering both encoding and transmission losses.

This paper is organized as follows: after the introduction, section 2 provides an overview of video quality assessment; the main reasons for video quality degradation in IP are described and methods, used by the H.264/AVC encoder and decoder to resist and conceal occurring errors, are outlined. Additionally, we overview already proposed NR metrics for transmitted video. In section 3, new NR quality metrics that address only the compression distortion are described, evaluated and compared. In section 4, new NR metrics for transmitted video over IP networks, considering both compression and transmission distortions, are described, evaluated and compared. Finally, in section 5, the main conclusions are given and some proposals of future work are put forward.

2. VIDEO QUALITY OVERVIEW

A. *Overview of the H.264/AVC*

H.264/AVC is a video compression standard developed by the ITU-T Video Coding Experts Group together with the ISO/IEC Moving Picture Experts Group (MPEG) in a partnership known as the Joint Video Team (JVT), formed in 2001 [WSBL03]. The objective of the JVT was to develop an advanced video coding specification capable of coding rectangular video with higher compression efficiency, when compared to existing standards such as H.263, MPEG-2 video and MPEG-4 Visual. Another objective was to have a good flexibility in terms of efficiency-complexity trade-offs, allowing the standard to be applied on a wide variety of applications such as broadcast, storage and multimedia streaming services over various networks.

To address the need for flexibility and customizability, the H.264/AVC design covers a Video Coding Layer (VCL) and a Network Abstraction Layer (NAL). The VCL was designed in a way to represent the video content efficiently; the NAL formats the VCL representation of the video so it can be compatible with various transport protocols or storage media. At the VLC, efficient video compression is achieved by exploiting the spatial and temporal redundancies of the video. Much like previous standards, H.264/AVC is based on a block-based coding approach. This means each frame of the video is represented by block shaped units called *macroblocks*, being each *macroblock* represented by 16×16 luminance pixels and by two 8×8 chrominance samples. The standard defines three types of frames:

- ***Intra frames*** (I-Frames), exploit spatial redundancy and are coded using only information within the frame. These frames are also used for random access since they do not require information from previous frames. I-Frames provide the less compression among the three frame types.
- ***Predicted Frames*** (P-Frames), not only exploit spatial redundancy but also temporal redundancy. This is done by using information from previous I or P frames.
- ***Bidirectionally Predicted Frames*** (B-Frames), also exploit spatial and temporal redundancies but they may use information from past, as well as from future I or P frames. B-Frames provide the highest compression among the three frame types.

The three frame types were also defined on previous standards, but H.264/AVC improved on them by adding some new features, such as multiple reference frame motion compensation and the ability to use a B-Frame as a reference frame.

As previously mentioned, the NAL adapts the compressed data from the VCL, so it can be compatible with various transport protocols. For video transmission over IP networks some protocols were defined for the three layers of the Open Systems Interconnection (OSI) model: Real-time Transport Protocol (RTP) on the application layer, User Datagram Protocol (UDP) on the transport layer and IP on the network layer [Weng03].

B. The origins of video losses

When a H.264/AVC encoder exploits the spatial redundancy it uses a type of coding based on the Discrete Cosine Transform (DCT). The coefficients resulting from the transform are quantized in order to remove irrelevancy. However, this quantization can reduce the quality of the video since some information is discarded resulting in compression losses, that can manifest as visible picture artefacts.

Concerning transmission losses in IP networks, packet losses occur mainly due to three factors [KGPL06]:

- Occasional bit errors caused by low noise margin or equipment failure.
- Buffer overflow or packet delay caused by congestion in the network.
- Rerouting to get around breakdowns or bottlenecks in the network.

Since the decoder on the receiving side needs that packets arrive in time to be displayed, packets too much delayed are discarded.

C. Error resilience and concealment techniques

The H.264/AVC standard provides error resilience schemes [KXMP06] in an attempt to minimize the consequences of transmission losses. These are mainly contained in the VCL and some of them have been used in previous standards. Some of the error resilience techniques are: the semantics and syntax used, Intra-frame refreshing, *slice* structuring and flexible macroblock ordering.

When the decoder detects an error, it may use an error concealment technique to try to make the error unnoticed. Some basic techniques are inter-frame prediction and intra-frame prediction. Inter-frame prediction uses information from previous frames, while intra-frame prediction uses information from the same frame in order to predict the content of lost MBs.

D. No-reference objective quality metrics

The effects of compression on video quality have been intensively studied and accurate metrics have been developed (e.g., [BrQu10][BrQu08] and the references included). On the other hand, the effects of transmission on video quality still need to be more investigated. Typical metrics used by service providers are the bit error rate (BER) or the packet loss ratio (PLR), which can be used to roughly predict the video quality.

At the present time, there are no standardized procedures for no-reference video quality assessment, although an intensive work on that subject is being performed by ITU-T and ITU-R through study/working groups SG9, SG12, and WP6C. The closest standard that is related with NR image quality assessment is ITU-T Recommendation G.1070 [ITUT07], which presents a quality model for video telephony applications.

Concerning scientific publications, three NR methods were proposed in [EVS04] to estimate mean squared error due to packet losses, directly from the video bitstream. Winkler and Mohandas proposed in [WiMo08] a no-reference metric – the V–factor –oriented to packetized transmission of MPEG-2 and H.264/AVC video. More recently, in [YWXW10], a quality measure for networked video is introduced using information extracted from the compressed bit stream without resorting to complete video decoding.

3. OBJECTIVE QUALITY ASSESSMENT OF ENCODED VIDEO

A. Objective quality models for encoded video

The main goal of all video quality prediction algorithms is to be able to predict the opinion a human observer would give, when evaluating a video's quality. Therefore, subjective video quality evaluation is essential to benchmark the objective video quality metrics. These subjective evaluations use human participants and specific evaluation methods. After a statistical analysis of the subjective scores, the Mean Opinion Score (MOS) of the human participants is obtained for every video sequence. The subjective data can then be used to calibrate or to validate the quality prediction algorithms. The subjective data, used to validate the objective metrics described in this section, were obtained through subjective video quality assessment tests performed at *Instituto Superior Técnico* [PQR09] with the purpose of studying the subjective quality of H.264/AVC and MPEG-2 encoded video.

The MOS depends on how much the video was compressed – video encoded with a high bit rate usually has a better quality (high MOS value) than a video encoded at a lower bit rate. The same behaviour is observed for the mean square error (MSE) of the encoded video. In fact, although the MSE is a rough measure of the perceived quality, its

correlation with the MOS tends to be high for the same sequence, when encoded at different bit rates, and using the same encoder. This conclusion may be confirmed by Figure 1, which shows the MOS versus the MSE values for the different video sequences, and encoding conditions, used in the subjective tests (described in [PQR09]). However, the previous conclusion does not hold when considering different video sequences – in this case, an increase of MSE may not correspond to a decrease in MOS values. For instance, looking once more at Figure 1, we may figure out that the increasing of MSE from 80 to 160 may be accompanied either by an increase in MOS from 0.46 ("Foreman") to 0.63 ("Tempete"), either by a decrease in MOS to 0.18 ("Football").

The MSE is defined as the mean squared difference between the original sequence and the coded sequence. When applied to the video luminance component, it is expressed as:

$$\text{MSE} = \frac{1}{M \times N \times T} \times \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{M} [Y_o(i,j,t) - Y_c(i,j,t)]^2 \quad (1)$$

where Y_o represents the luminance of the original t-th frame at pixel (i,j), Y_c the luminance of the compressed t-th frame at pixel (i, j), T the total number of frames and M,N the number of pixels per line and the number of lines of each frame, respectively.

With the MSE we can obtain the Peak Signal-to-Noise Ratio (PSNR), which is commonly used as an objective measure of video quality. The PSNR is the ratio between the maximum possible value of luminance (for pixels represented in 8 bit per sample this value is 255) and the MSE, and it is usually expressed in logarithmic units as:

$$\text{PSNR} = 10\, log_{10}\left(\frac{255^2}{MSE}\right), \text{ [dB]} \quad (2)$$

All the metrics developed in this section, explore the MOS versus MSE relation and use the error estimation module proposed in [BrQu10] for a no-reference MSE estimate of the encoded video sequences.

B. The MOS prediction models

Based on the relationship between MOS and MSE, four different MOS prediction models are described in this sub-section.

The first MOS prediction model was proposed by Bhat in [BRK09] as a FR model and it considers that the relationship between MOS and MSE can be seen as a straight line with slope $-k_s$ and a y-intercept of 1; mathematically, it can be expressed as:

$$\text{MOSp} = 1 - k_s(\text{MSE}) \quad (3)$$

where MOSp is the predicted MOS. By using linear regression of the MOS values, obtained from the subjective test, versus the corresponding MSE values, it is possible to obtain the straight line parameter (k_s value) for each video sequence. Figure 2 shows the subjective data and the straight line

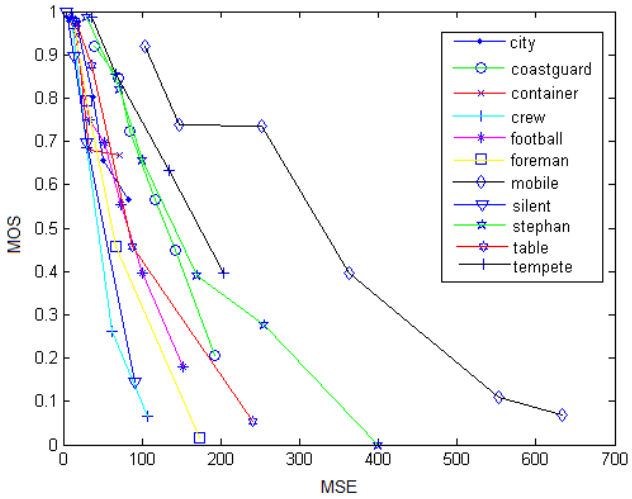

Figure 1 - MOS versus MSE

resulting from the linear regression for the "Crew", "Foreman", "Mobile" and "Stephan" video sequences. straight line parameter (k_s value) for each video sequence. Figure 2 shows the subjective data and the straight line resulting from the linear regression for the "Crew", "Foreman", "Mobile" and "Stephan" video sequences.

However, observing Figure 1, it becomes clear that the MOS versus MSE evolution has not the same behaviour for the highest values of MSE. The straight line parameter from the previous model doesn't seem to be constant and appears to decrease as the MSE increases. In other words, the quality seems to decrease faster on lower MSE values when compared to higher MSE values. Therefore, another possible model is to consider the relation between MSE and MOS as an exponential function, which can be expressed by:

$$\text{MOSp} = \exp\left(-\frac{\text{MSE}}{k_s}\right) \quad (4)$$

where MOSp is the predicted MOS and k_s is the exponential parameter. By using regression with the subjective data and the real MSE values, the exponential parameters were obtained for each sequence. Figure 3 shows the subjective data and the resulting regression curves for the "Crew", "Foreman", "Mobile" and "Stephan" sequences.

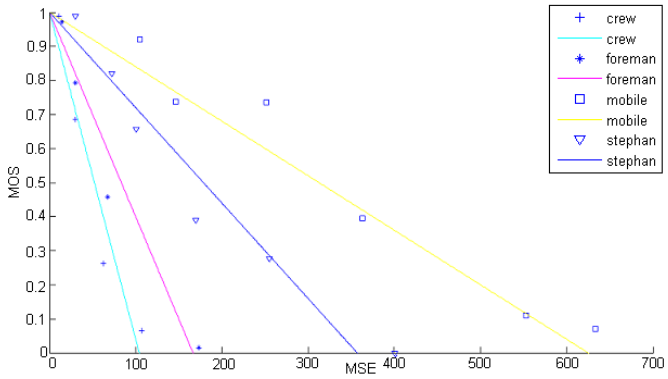

Figure 2 - Regression curves for the linear model for Crew, Foreman, Mobile and Stephan

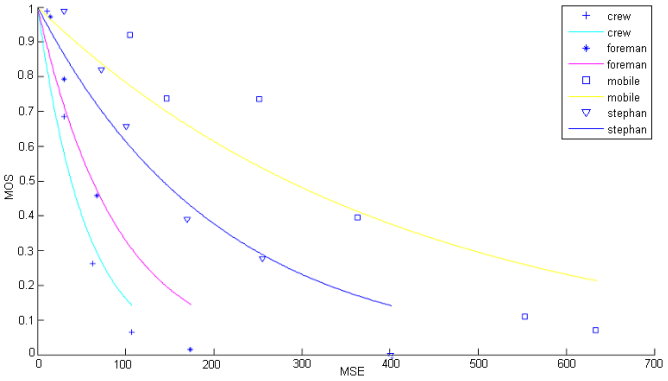

Figure 3 - Regression curves for the exponential model for Crew, Foreman, Mobile and Stephan

By taking a closer look at Figure 1, we can see that the MOS versus MSE relation is not a simple exponential function, as considered in the previous model. In fact, it seems to resemble a sigmoid function since the quality has a slower decrease on lower and higher MSE values when compared to mid MSE values. This third model (Sigmoid1 model) uses a sigmoid function which can be expressed as:

$$\mathrm{MOSp} = \frac{1+e^{-k_1 k_2}}{1+e^{k_1(\mathrm{MSE}-k_2)}} \qquad (5)$$

where MOSp is the predicted MOS and k_1 and k_2 are the sigmoid parameters. By using, once again, regression with the subjective data and the real MSE values, it is possible to obtain the sigmoid parameters for each sequence. Figure 4 shows the subjective data and the regression curves obtained for the "Crew", "Foreman", "Mobile" and "Stephan" sequences.

The fourth and final model was proposed in [WP02] as a FR model and it also follows a sigmoid function (Sigmoid2 model)

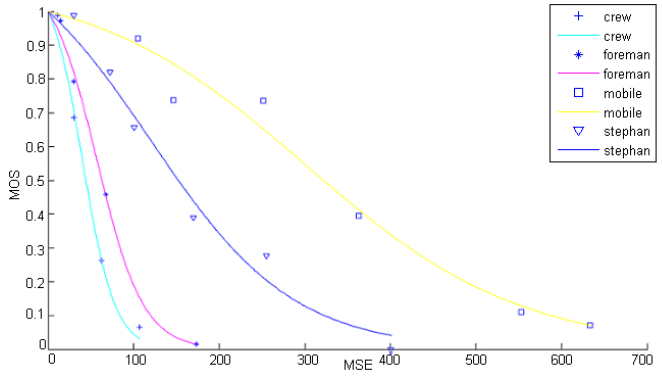

Figure 4 - Regression curves for the sigmoid model (MSE) for Crew, Foreman, Mobile and Stephan

However, it uses the PSNR measurement to estimate the MOS. Mathematically, this model is expressed as:

$$\mathrm{MOSp} = 1 - \frac{1}{1+e^{k_1(\mathrm{PSNR}-k_2)}} \qquad (6)$$

where MOSp is the predicted MOS and k_1 and k_2 are the sigmoid parameters. Applying, once again, regression with the subjective data and the real PSNR values, the sigmoid parameters were obtained for each sequence. Figure 5 shows the subjective data versus MSE (for a better comparison with the previous models, the plot is versus MSE and not PSNR values) and the regression curves obtained for the "Crew", "Foreman", "Mobile" and "Stephan" sequences.

We have seen that each model predicts the MOS by using the MSE and one or two parameters; these parameters were obtained by regression using the subjective data (MOS values). However, in a practical transmission scenario the subjective data is unavailable, so those parameters have to be estimated from the received data (video bitstream and/or decoded video). In [BRK09], where the straight line model was proposed, the authors showed that the required parameter, k_s, is related to the video content activity. This approach was also followed in this thesis, in order to estimate the model parameters required by the new models.

An estimation of the video spatial activity can be obtained with information taken from the bitstream, namely the DCT coefficients. We propose to compute the video activity by first estimating the spatial activity of each I-frame through:

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(DCTcoef_i - \mu)} \qquad (7)$$

where σ_j is the estimated spatial activity of the j-th I-frame, $DCTcoef_i$ is the i-th DCT coefficient of the frame, μ is the average value of the DCT coefficients in the frame and n is the number of DCT coefficients in the frame.

The estimated activity of the video results from the average of the spatial activity of all I-frames:

$$Estimated\ activity = \frac{1}{N}\sum_{j=1}^{N}\sigma_j \qquad (8)$$

where N is the number of I frames and σ_j is the estimated spatial activity of the j-th I frame.

In order to validate the estimated activities, the Pearson correlation between them and the activity values computed in the pixel domain, using the original video was computed, and a value of 0.97 was obtained. It was also verified that the effect of compression as a marginal impact on this value.

In the following, the used MSE values can be the ones computed using the original and encoded videos ("true" MSE) or using the no-reference MSE estimate from [BrQu10] ("estimated" MSE). As for the video activity, it can be obtained directly from the uncompressed videos ("true" activity) as described in [BRK09] or by using equations (7) and (8) ("estimated" activity).

Figure 6 shows how the model parameters, obtained using the true MSE values, relate with the video activity; this figure suggests that some model parameters have a linear relation with the video activity, while others have an exponential relation.

The parameters (k_s) for the exponential and the sigmoid1 models were considered to have an exponential relation with the video activity, which can be expressed as:

$$model\ parameter = \beta_1 . \exp(Activity \times \beta_2) \qquad (9)$$

The parameters (k_s) for the linear and sigmoid2 models were considered to have a linear relation with the video activity, which can be expressed as:

$$model\ parameter = \beta_1 \times Activity + \beta_2 \qquad (10)$$

Parameters β in (9) and (10) are obtained by regression using the subjective MOS, the MSE values, and the video activity, by substituting (9) or (10) in (3), (4), (5) and (6).

C. MOS prediction

Using the described NR models the predicted MOS, for each video sequence, can be obtained. To do so, it is required:

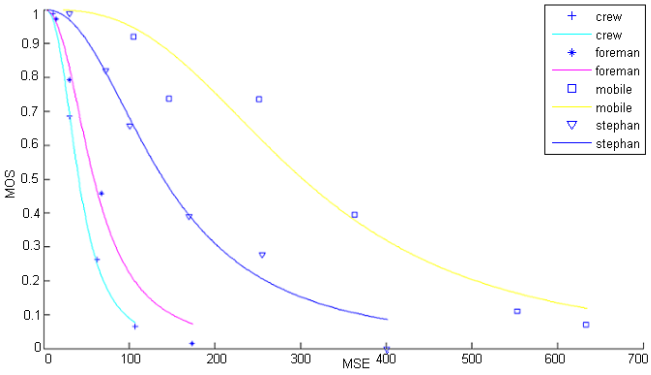

Figure 5 - Regression curves for the sigmoid model (PSNR) for Crew, Foreman, Mobile and Stephan

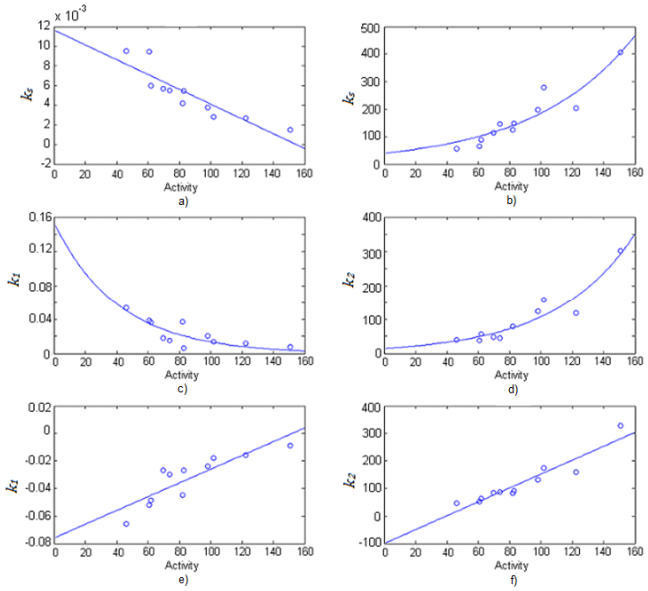

Figure 6 - Model parameters versus video activity, and resulting regression curves, for: a) linear model, b) exponential model, c) and d) sigmoid1 model, e) and f) sigmoid2 model, using the true MSE and video activity values.

- The functions of the prediction models.
- A training video set to train the models by calculating the parameters β.
- An estimated MSE value of the video sequence whose MOS we want to predict [BrQu10].
- An estimated (eq. (8)) of the activity for the video sequence whose MOS we want to predict.

To validate objective quality metrics, it is often used a training set and a validation set. The training set calibrates the metrics which, in our case, correspond to the finding of the β values; the validation set is used to evaluate the metrics. Since we have a small set of video sequences, the *leave-one-out cross-validation* was used. This method is done by turns, in each turn the different encoded versions of the same video sequence are used as the validation set while all the other video sequences are used as the training set. In each turn, the MOSp of the validation set are obtained. This is repeated until all the video sequences have been used as the validation set and the corresponding MOSp obtained.

After performing the *leave-one-out cross-validation* method on the eleven video sequences, their MOSp was calculated for each prediction model. Figure 7 shows the MOS versus MOSp for the different models using the estimated MSE and video activities.

D. Results and model comparison

To compare the four models, the VQEG performance metrics described in [VQEG03], namely Pearson and Spearman correlation coefficients and root mean squared (RMS) error, were used. Table 1 shows the three performance

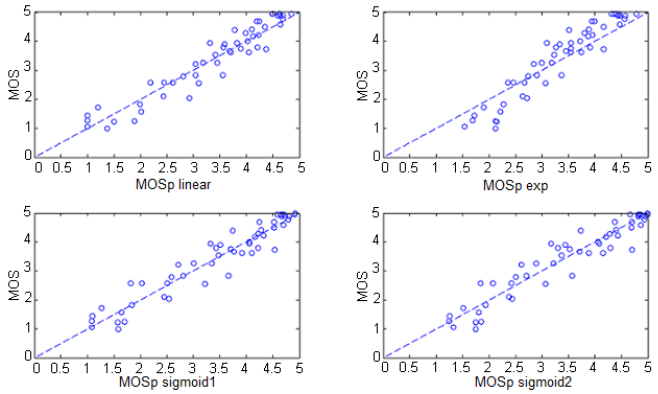

Figure 7 - MOSp versus MOS for the four NR prediction models using estimated MSE and estimated activity

metrics obtained with all eleven videos using the estimated MSE and video activities.

In Table 1 we have highlighted in green the model with the best results for each performance metric. The linear model is the one with best results in the RMS and Pearson coefficient and, together with the Sigmoid1 model, has also the highest Spearman coefficient. All the models produce very similar results, making them all acceptable. However, the linear model can be considered as best of the four, not due to the correlation coefficients obtained but also because of its lower complexity when compared to the other models.

4. Objective Video Quality Assessment in IP Networks

A. Objective quality models for transmission with packet losses

The subjective data used in this section was obtained through subjective tests performed in *Politecnico di Milano* (PoliMi) – Italy and *Ecole Polytechnique Fédérale de Lausanne* (EPFL) - Switzerland [SNTD09] resulting in two databases. The subjective tests addressed the effect of packet losses on a video's perceived quality when encoded with H.264/AVC. For this purpose, six video sequences in CIF format and with a frame rate of 30 fps, were considered, namely "Foreman", "Hall", "Mobile", "Mother", "News" and "Paris".

The MOS and the MSE values of each video obtained after decoding, allow the analysis of the MOS versus MSE behavior resulting from packet losses. By looking at Figure 8, it is possible to conclude that, as in the previous section, where only losses due to compression were considered, the MOS does not correlate well with the MSE if we consider all video sequences.

If we look at each sequence individually, we can see that the MOS tends to decrease as the MSE increases. However, a closer look shows that, unlike what happens after compression, the plot MOS(MSE) does not have a monotonous variation, since there are situations where the MOS clearly increases when the MSE increases.

Table 1 - Correlation coefficients using estimated MSE and estimated activity for all videos

Correlation Coefficient	*Model*	*All*
RMS	*Linear*	*0.355*
	Exponential	*0.478*
	Sigmoid1	*0.359*
	Sigmoid2	*0.377*
Pearson	*Linear*	*0.959*
	Exponential	*0.954*
	Sigmoid1	*0.957*
	Sigmoid2	*0.953*
Spearman	*Linear*	*0.947*
	Exponential	*0.945*
	Sigmoid1	*0.947*
	Sigmoid2	*0.943*
Outlier	*Linear*	*0.038*
	Exponential	*0.154*
	Sigmoid1	*0.038*
	Sigmoid2	*0.038*

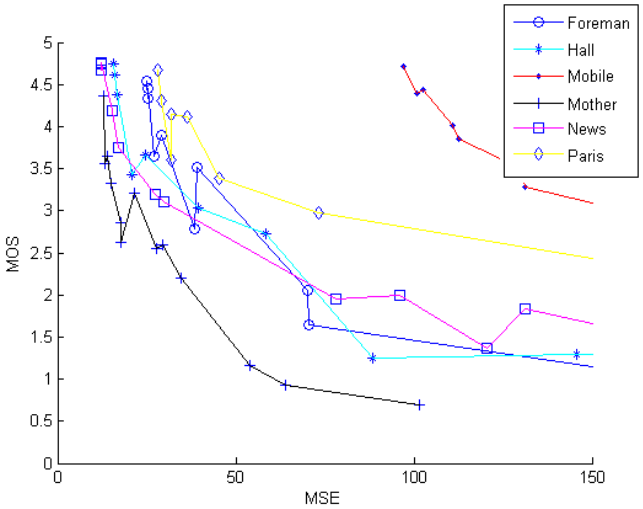

Figure 8 - MOS versus MSE for the PoliMi database and MSE values in the range 0 - 150

With these observations, a video quality prediction model based on the MSE seems to be potentially unreliable for quality prediction of videos affected by packet losses. Thus a different approach is necessary.

Since the new element introduced were the packet losses, characterized by the PLR, the relation between PLR and MOS was analyzed. We started by computing the actual PLR of each video sequence by analysing the bitstream and checking the syntax of the packet header on each transmitted packet. Figure 9 presents the resulting MOS values *vs* PLR for the PoliMi database.

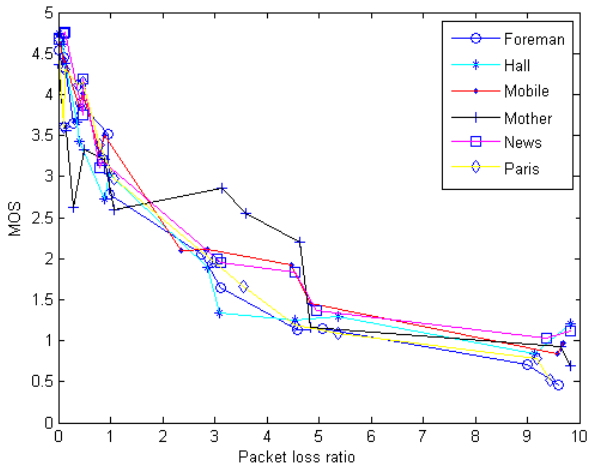

Figure 9 - MOS versus PLR for the PoliMi database

The MOS versus PLR suggest that MOS values are better correlated with PLR than with MSE. To further confirm this, the Spearman correlation metric between MOS and PLR or MSE were calculated and are presented in Table 2.The values of the Spearman metric confirm that the MOS has a better correlation with the PLR. Taking this into consideration, all the following prediction models are based on the MOS/PLR relationship.

B. Simple PLR Model

This model is based on the video quality prediction model proposed in ITU-T Rec. G.1070 [ITUT07] and it relates the MOS with the PLR. Figure 9 suggests that the MOS/PLR relation can be described by an exponential function, thus the model is mathematically given by:

$$\text{MOS} = MOS_{pl0} \times \exp\left(-\frac{PLR}{\theta}\right) \qquad (11)$$

where MOS_{pl0} is the MOS of the video without any transmission losses, PLR is the packet loss ratio and θ is a parameter. The θ parameter was considered a constant that can be obtained by regression. Since only six video sequences are available a cross-validation training method was used to train and test the model namely, the *leave-one-out cross-validation* was utilized. The model was also evaluated as a NR model by using an estimation of the MOS_{PL0}. Figure 10 plots the obtained MOSp versus the subjective MOS (PoliMi) values, using the NR model; Table 3 shows the correlation metrics and RMS values obtained. From these results we may conclude that the Simple PLR FR model scores acceptable values for the Pearson, Spearman and RMS metrics.

Table 2 - Spearman metric for MOS/MSE and MOS/PLR

Correlation metric	MOS versus MSE	MOS versus PLR
Spearman	-0.7511	-0.9518

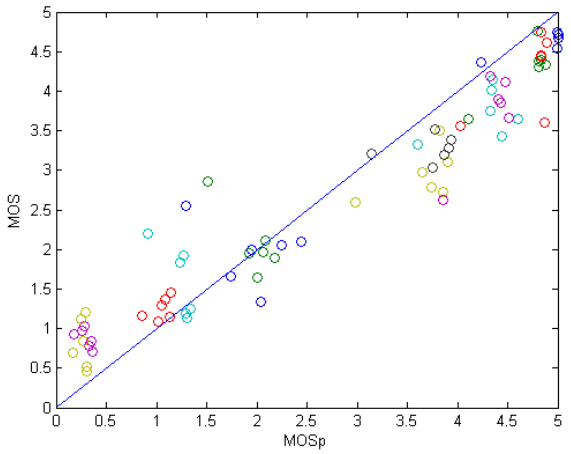

Figure 10 - MOS versus MOSp for the PoliMi database using the NR Simple Model

Despite these positive results, there are a few particular cases where the Simple PLR Model didn't have a good performance. An example is the sequence "Mother" where the model scored a Pearson value of 0.88 for PoliMi and 0.87 for EPFL; also, for low PLR values (*i.e.*, PLR < 1%), there are some cases in which an increase in PLR is accompanied by an increase in MOS. In the next sub-sections we will address possible modifications to this simple model.

C. Frame Type Model

Since the H.264 uses frame dependency, it is expected that the degradation caused by a packet loss on a video sequence will depend on the type of frame where the loss occurs. For the used encoding, a packet loss in an I-frame is expected to be more relevant than a loss in a B-frame since an I-frame has frames which depend on it, while a B-frame does not. Moreover, the decoder uses, as error concealment, intra-frame prediction for I-frames and inter-frame prediction for P and B frames. This difference in the concealment technique reinforces the idea that losses should be discriminated by the frame type where they occur.

The model described and analyzed in this section, separates the packet losses according to the type of frame where they occur, giving them different weights. It tries to improve the Simple PLR Model by using a modified PLR

$$\text{MOSp} = MOS_{pl0} \times exp(-fPL) \qquad (12)$$

where,

$$fPL = \left(\frac{\omega_I.\sum I\ Block\ loss + \omega_P.\sum P\ Block\ loss + \omega_B.\sum B\ Block\ loss}{\sum total\ blocks}\right) \qquad (13)$$

being, *fPL* the modified PLR, MOS_{pl0} the MOS of the video sequence without any transmission losses, ω_j the weight of the j-type frames, $\sum j\ Block\ loss$ the total of lost 4×4 blocks belonging to a *j*-type frame and $\sum total\ blocks$ the total number of 4×4 blocks in the video.

Figure 11 represents the resulting *fPL* values versus the MOS for the PoliMi database. Besides the exponential relation between *fPL* and MOS, Figure 11 shows a more monotonically relation between the two when compared with the MOS and PLR relation. In order to validate the model, the *leave-one-out cross-validation* method was used. Figure 12 shows the MOS (PoliMi) versus MOSp for the Frame Type Model while Table 3 shows the resulting correlation metrics.

When compared with the Simple PLR Model, the Frame Type Model scored slightly worse results in all three correlation coefficients. A reason for this may be the fact that this model only considers the frame type where the losses occur and ignores the subjective impact cause by error propagation (due to frame dependency).

D. *Frame Type and Movement Model*

As previously mentioned, video decoders use error concealment techniques to try to prevent video degradation caused by packet losses. Some techniques work better than others however, all of them can more efficiently conceal a loss when the video sequence doesn't have much movement. The model described and analyzed in this sub-section, adds this information to the Frame Type Model. This is done by only considering a lost 4×4 block in a P or B frame as an actual loss, if the norm of its motion vector (*MVabs*) is higher than a threshold value. Losses occurring in an I-frame are always considered as actual losses.

The *norm of a motion vector* is computed by:

$$MVabs = \sqrt{MVx^2 + MVy^2} \tag{14}$$

where *MVx* and *MVy* are, respectively, the *x*-axis component and the *y*-axis component of the motion vector.

Mathematically, the model is given by:

$$\text{MOSp} = MOS_{pl0} \times \exp\left(fPL_{mv}\right) \tag{15}$$

where,

$$fPL_{mv} = \frac{\omega_I . \sum I\ Block\ loss + \omega_P . \sum P\ Block\ loss\ \omega_B . \sum B\ Block\ loss}{\sum total\ blocks} \tag{16}$$

being fPL_{mv} the modified PLR, MOS_{pl0} the MOS of the video sequence without any transmission losses, ω_j the weight of the *j*-type frames, $\sum j\ Block\ loss$ the total of actual lost 4×4 blocks belonging to a *j*-type frame and $\sum\ total\ blocks$ the total number of 4×4 blocks in the video.

To choose the value of the threshold, the model was tested with various values and a threshold of 10 was the one producing the best results. It should be noted that the MVabs were computed using the MVs associated to the lost blocks.

Once again, the *leave-one-out cross-validation* method was used to validate the model. Figure 13 shows the MOS (PoliMi) versus the obtained MOSp, while Table 3 shows the resulting correlation metrics. The results show that the Frame Type and Movement Model has scored acceptable values for the correlation coefficients. However, when compared with the Simple PLR Model, this model scored worse results, particularly for the RMS metric.

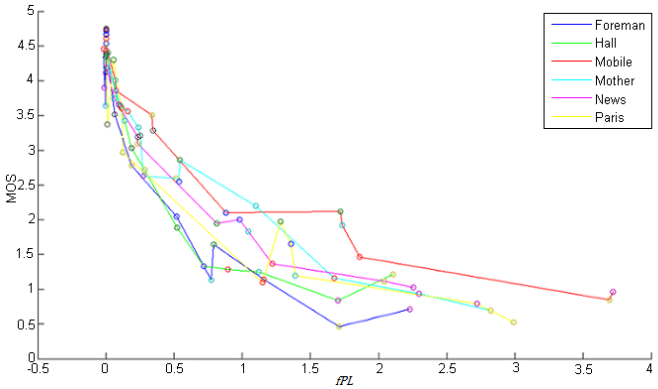

Figure 11 - MOS versus modified PLR for the PoliMi database

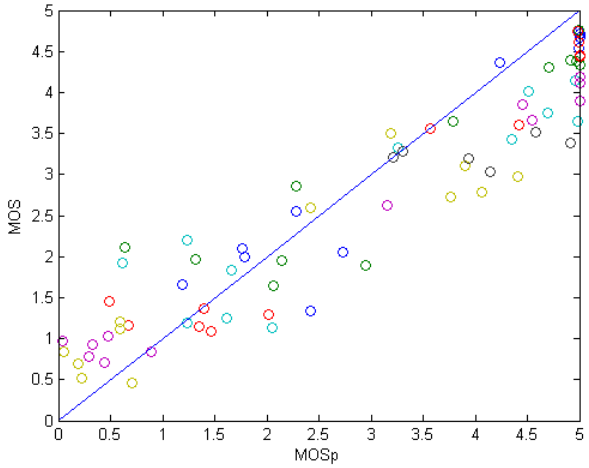

Figure 12 - MOS versus MOSp for the PoliMi database for the Frame Type Model

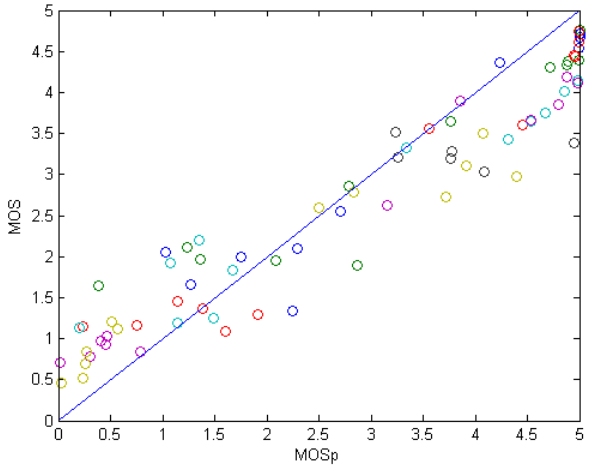

Figure 13 - MOS versus MOSp for the PoliMi database using the Frame Type and Movement Model

E. Frame Type, Dependencies and Movement Model

This model takes into account the frame type where the losses occur, the additional losses as a result of the dependency between I, P and B-frames and the movement in the area where the losses occurred. Once again, a 4×4 block is only considered as an actual loss if its *MVabs* is higher than a threshold (which assumes that the concealment technique used by the decoder is able to properly conceal a loss in a low movement area). This is also done to the additional losses resulting from error propagation.

This model is mathematically given by:

$$\text{MOSp} = MOS_{pl0} \times exp(-fPL2) \qquad (17)$$

where,

$$fPL2 = (\omega_I.\sum I\ Blk\ loss + \omega_P.\sum P\ Blk\ loss + \omega_B.\sum B\ Blk\ loss + \omega_{DI}.\sum Dep\ I\ Blk\ loss\ + \omega_{DP}.\sum Dep\ P\ Blk\ loss\ )/(\sum total\ blocks) \qquad (18)$$

being *fPL2* the modified PLR considering frame dependency, MOS_{pl0} the MOS of the video sequence without any transmission losses, ω_j the weight of the j-type frames, $\sum$ *j Block loss* the total of actual lost 4×4 blocks belonging to a *j*-type frame, $\sum$ *Dep j Blk loss* the total of 4×4 blocks received and with a *MVabs* higher than the threshold (but dependent on lost 4×4 blocks belonging to a *j*-type frame) and $\sum$ *total blocks* the total number of 4×4 blocks in the video.

To choose the value of the threshold, the model was tested with various values and a threshold of 25 was the one producing the better results. It should be once again noted that the *MVabs* were calculated with the MVs of the lost blocks.

Once again the *leave-one-out cross-validation* method was used to validate the model. Figure 14 shows the MOS (PoliMi) versus the obtained MOSp while Table 3 shows the resulting correlation metrics. The results show that the Frame Type, Dependencies and Movement Model has scored acceptable values for the correlation coefficients. However, when compared with the Simple PLR Model, this model is worse in all the correlation metrics and it is also more complex.

F. Statistical model

The modified PLR models weren't fully able to address the situation they were initially trying to solve. Characteristics such as the frame type where the losses occur are relevant, but there is another characteristic that affects a video's perceived quality, the packet loss pattern. By analyzing the syntax of the packet headers on each transmitted packet, this pattern can be obtained and various statistical metrics of the losses distribution can be computed. The ones that prove to be helpful in predicting a video's perceived quality are selected to be part of the statistical model.

By analyzing the packet loss pattern various statistical metrics were computed and the correlation between each metric and the MOS was determined. To select the appropriate model it is necessary to determine which variables should be used in the model. At the end, it is expected a model with

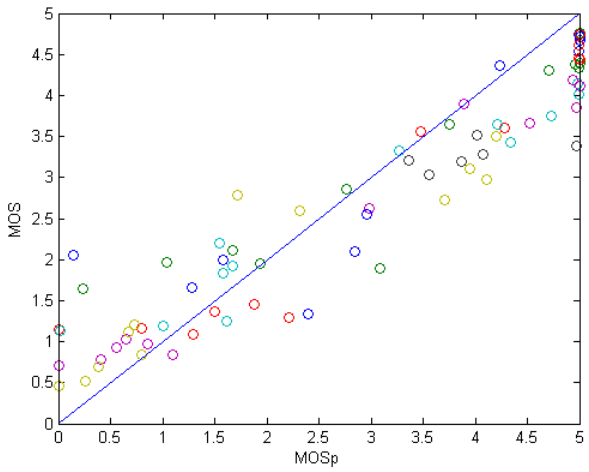

Figure 14 - MOS versus MOSp for the PoliMi database using the Frame Type, Dependencies and Movement Model

enough variables so that it can perform satisfactorily; however, too many variables may overcomplicate the model. The variable selection was based on a *stepwise regression* [MoRu03]. The *stepwise regression* was applied (using the PoliMi database) and the following statistical metrics were kept:

- Maximum number of lost packets on the same I-frame.
- Average number of lost packets on frames with more than one loss
- Maximum number of consecutive lost packets on the same P-frame
- Average number of consecutive lost packets per I-frames
- Average number of consecutive lost packets per P-frame, ignoring single losses
- Average distance between frames with losses (considering single losses)
- Modified PLR from the Frame type Model

The final statistical model is mathematically given by:

$$\text{MOSp} = MOS_{pl0} \times \exp(\sum_{i=1}^{n} \omega_i \times stat_i) \qquad (19)$$

being *n* the number of statistical metrics (7 in this case), ω_i the weight of the *i*-th statistical metric and $stat_i$ the value of the *i*-th statistical metric.

In order to validate the model the *leave-one-out cross-validation* method was once again used. In each turn, the weights are recalculated with the training set and the estimation of MOS values (MOSp) is obtained using the validation set. Figure 15 depicts the MOS (PoliMi) versus the MOSp values, while Table 3 shows the resulting correlation metrics.

G. Model comparison

Figure 10 shows the MOS vs. MOSp plots for the Simple PLR Model. Here it can be seen that the Simple PLR Model has a good performance, which translates into high correlation coefficient values, as shown in Table 3. However, a few

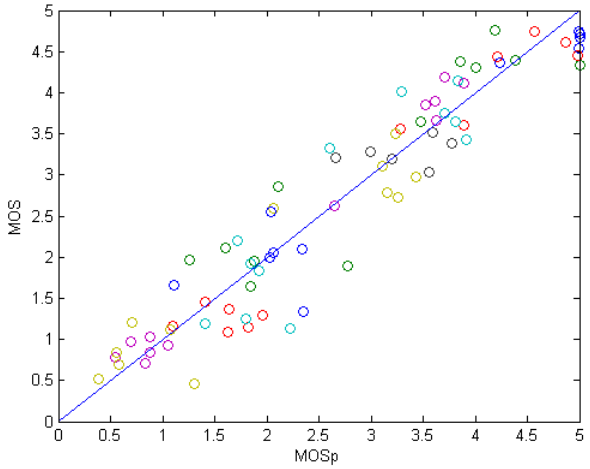

Figure 15 - MOS versus MOSp for the PoliMi database using the Statistical model

predictions were far from the true values and that resulted in the development of the Modified PLR models. All the Modified PLR models have acceptable performances, as shown in Table 3, but were unable to significantly improve the Simple PLR Model. In fact, their RMS values are higher than the Simple PLR Model's RMS values.

The Statistical model has a good performance (Pearson = 0.950 (PoliMi), 0.945 (EPFL); Spearman = 0.950 (PoliMi), 0.942 (EPFL)) when compared to the other models. For the Pearson and Spearman metrics, the model scored slightly lower than the Simple PLR Model. But, as a plus, the model was able to address the situations where the Simple PLR Model failed. This translates into better RMS values, since the Statistical model obtained a RMS of 0.426 (PoliMi) and 0.479 (EPFL) while the Simple PLR Model obtained a RMS of 0.581 (PoliMi) and 0.591 (EPFL).

5. CONCLUSIONS AND FUTURE WORK

In this paper, bitstream-based NR quality metrics for H-264/AVC encoded video, when transmitted over IP networks, were proposed and evaluated. The results achieved have shown that the Statistical model lead to the best performance. The model uses the information taken from bitstream and from the packet headers. Although the Statistical model has shown a good performance, there is still room for improvements. As previously mentioned, video decoders use error concealment techniques to prevent video degradation caused by packet losses. Some techniques work better than others. However, all of them can more efficiently conceal a loss when the video sequence doesn't have much temporal and/or spatial activity. Accordingly, it is expected that by better quantifying the video spatio-temporal activities, more accurate objective video quality metric could be developed at the expense of increased complexity. Additionally, the video sequences used on the subjective quality tests are quite limited and a new database that allows the study of the impact of the different network and coding parameters would be extremely useful.

Table 3 - Performance of each model

Model	Model performance			
	Database	Pearson	Spearman	RMS
Simple PLR Model	PoliMi	0.959	0.956	0.581
	EPFL	0.960	0.963	0.591
Frame Type Model	PoliMi	0.941	0.935	0.699
	EPFL	0.949	0.952	0.688
Frame Type and Movement Model Threshold 10	PoliMi	0.958	0.956	0.637
	EPFL	0.933	0.933	0.752
Frame Type, Dependencies Movement Model Threshold 25	PoliMi	0.947	0.945	0.678
	EPFL	0.952	0.955	0.703
Statistical Model	PoliMi	0.950	0.950	0.426
	EPFL	0.945	0.942	0.479

6. REFERENCES

[WSBL03] T. Wiegand, G. J. Sullivan, G. Bjontegaard, A. Luthra, "Overview of the H.264/AVC Video Coding Standard", *IEEE Trans. on CSVT*, vol. 99, no. 4, pp. 626-642, July 2003.

[Weng03] S. Wenger, "H.264/AVC over IP", *IEEE Trans. on CSVT*, vol. 13, no. 7, pp. 645-656, July 2003.

[KGPL06] Y. Kukhmay, K. Glasman, A. Peregudov, A. Logunov, "Video Over IP Networks: Subjective Assessment of Packet Loss", *IEEE Tenth International Symposium*, IEEE, 2006.

[BrQu10] T. Brandão and M. P. Queluz, "No-reference quality assessment of H.264 encoded video," *IEEE Trans. on CSVT*, vol. 20, pp. 1437, November 2010.

[BrQu08] T. Brandão and M. P. Queluz, "No-reference PSNR estimation algorithm for H.264 encoded video sequences," in proc. of EUSIPCO, Lausanne, Switzerland, August 2008.

[ITUT07] ITU-T, "Opinion model for video-telephony applications", ITU-Recommendation G.1070, April 2007.

[EVS04] A. R. Reibman, V. A. Vaishampayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," *IEEE Trans. on Multimedia*, vol. 6, no. 2, pp. 327–334, Apr. 2004.

[WiMo08] S. Winkler and P. Mohandas, "The evolution of video quality measurement from PSNR to hybrid metrics," *IEEE Trans. on Broadcasting*, vol. 54, no. 3, pp. 660–668, September 2008.

[YWXW10] F. Yang, S. Wan, Q. Xie and H.R. Wu, "No-Reference Quality Assessment for Networked Video via Primary Analysis of Bit Stream", *IEEE Trans. on CSVT*, Vol. 20, no. 11, November 2010.

[PQR09] M. P. Queluz, T. Brandão and L. Roque, "Subjective Video Quality Assessement", http://amalia.img.lx.it.pt/~tgsb/H264_test/, 2008-2009.

[BRK09] A. Bhat, I. Richardson, S. Kannangara, "A novel perceptual quality metric for video compression", *in proc. of Picture Coding Symposium* (PCS) 2009, California Illinois, USA, May 2009.

[WP02] S. Wolf and M. Pinson, "Video Quality Measurement Techniques", NTIA Report, June 2002.

[VQEG03] VQEG. "Final report from the video quality expertsgroup on the validation of objective models of video quality assessment, phase II". Technical report, www.vqeg.org, August 2003.

[SNTD09] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro and Y. Ebrahimi, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel" in *proc. of QoMEX*, S. Diego, USA, 2009

[MoRu03] D. C. Montgomery, G. C. Runger, *"Applied Statistics and Probability for Engineers"*, Third Edition, 2003.