USING OVERLAPPING SUBJECTIVE DATASETS TO ASSESS THE PERFORMANCE OF OBJECTIVE QUALITY METRICS ON SCALABLE VIDEO CODING AND ERROR CONCEALMENT

Y. Pitrey

University of Vienna, AUSTRIA
yohann.pitrey@univie.ac.at

R. Pépion, P. Le Callet, M. Barkowsky

LUNAM Université de Nantes,
IRCCyN UMR CNRS 6597, FRANCE

ABSTRACT

In this paper, four subjective video datasets are presented. The considered application is Scalable Video Coding used as an error-concealment mechanism. The presented datasets explore the relations between encoding parameters and perceived quality, under different network-impairment patterns and involve error-concealment on the decoder's side, to simulate a complete distribution channel. The datasets share a part of common configurations which enables, in the first part of the paper, to compare the outcomes from several Single Stimulus experiments and draw interesting correspondances between different types of distortion. In the second part of the paper, we analyse the performance of three common objective quality metrics on each step of the distribution channel, to identify the possible directions to be followed in order to improve their accuracy in predicting the perceived quality.

Index Terms— Subjective video databases, Video quality metrics, Inter-experiment alignment, Scalable Video Coding, Error-concealment, Packet-loss, Source characteristics.

1. INTRODUCTION

When dealing with video transmission over lossy communication networks, several factors are involved in the perception of quality of the end user. First, video coding is known to decrease the perceived level of quality by adding quantization artifacts. Then, packet-loss or corruption can compromise the decoding process and result in additional visual distortions. These distortions can in turn be corrected or made less visible by error-concealment techniques, which try to mask the missing parts in the decoded stream.

A possible way to enable effective error concealment is to use Scalable Video Coding (SVC) to provide redundant versions of the same video content. In this paper, we consider a scenario where a video stream is encoded using two scalable layers: one base layer and one enhancement layer. Under nominal transmission conditions (i.e. when no packet-loss occurs), the enhancement layer is decoded and displayed on the end-user's terminal. When packet-loss makes decoding the enhancement layer impossible, data from the base layer is used to conceal the visual artifacts produced by the missing data. As the base layer is usually encoded with relatively low bitrate, it is easier to protect it using conventional techniques and to make sure it is transmitted correctly. This assumption might not hold in extreme transmission conditions. One of the major advantages of SVC regarding coding efficiency is that the higher layer can use data from the base layer for prediction to reduce the overall needed bitrate. This advantage can turn into a major drawback in case of high error-rate, as losing data from the lower layer makes the whole stream impossible to decode. An alternative scenario can be to use SVC without this inter-layer prediction, at a price in terms of bitrate increase. The use of SVC still makes sense in such a configuration, because of the stream structure. The layers are indeed interlaced, together with synchronization information, that facilitates the switching operations between the different levels of the stream.

Such extreme transmission conditions are beyond the scope of this paper, and we assume in the following that the base layer is correctly decoded at all times. As a result, a lower-quality version of the video is displayed in case of loss, which represents an interesting alternative to conventional frame freezing and skipping, commonly used with single layer coding. Naturally an SVC solution comes at a price in terms of bitrate and encoding complexity, but this drawback can be compensated by a better acceptance from the users.

In this paper, we present four subjective video quality assessment experiments conducted in the context of scalable video transmission under various packet-loss configurations with error-concealment. The paper is constructed around two main contributions. First, we show how the four experiments can be combined through the use of common sets of configurations, to circumvent the classical corpus-effect of Single Stimulus experiments. The combination of the datasets can be used to analyse the relative impact of several types of distortions which can be met in the considered application context. As a second contribution, we analyse the performance of three common objective quality metrics on specific aspects of our experiments and show how these datasets can be exploited to build more accurate video quality models.

	SRC	HRC	OBS	HRC overview	Common PVS
T1	9	15	29	– 1 non-coded reference – 1 AVC-based error concealment – 2 SVC upscaled base layer, bitrate = 120,200 kbps – 2 SVC-based error concealments : patch, switch – combined with 2 SVC BL bitrates : 120,200 kbps – combined with 2 SVC BL FPS : 15,30 Hz	T2 : 36 T3 : 36 T4 : 0
T2	11	30	27	– 1 non-coded reference – 16 SVC constant QP scenarios : – base layer: QP0 = 26, 32, 38, 44 – combined with enhancement layer: QP1 = 26, 32, 38, 44 – 4 upscaled base layer scenarios: QP = 26, 32, 38, 44 – 4 AVC scenarios: QP = 26, 32, 38, 44	T1 : 36 T3 : 126 T4 : 50
T3	11	36	28	– 1 non-coded reference – 3 AVC scenarios : QP = 32, 38, 44 – 23 scenarios with a selection of the following parameters : – length of impairments = 2, 4, 8, 16, 32, 64, 128, 224 – number of impairments : 1, 2, 3, 4 – intervals between impairments : between 8 and 128 – base layer QP = 38, 44	T1 : 36 T2 : 126 T4 : 16
T4	60	5	27	– 1 non-coded reference – same HRCs as T2 – only 4 HRCs per content – 4 groups of HRCs were formed from the results of T2 – MOS < 2; MOS $\in [2;3]$; MOS $\in [3;4]$; MOS > 4 – 1 HRC from each group was randomly selected for each content	T1 : 0 T2 : 50 T3 : 16

Table 1. Overview of the four subjective experiments presented in this paper.

2. DESCRIPTION OF THE SUBJECTIVE DATASETS

Four subjective experiments were conducted in the context of Scalable Video Coding and error concealment. Table 1 summarizes the four experiments in terms of source contents and tested configurations. Here we quickly review the important aspects differentiating the four experiments. More details are available in [1, 2, 3].

The first experiment (T1) proposes an overview of the capabilities of SVC as an error concealment technique [1]. Nine original video sequences are encoded using two spatially scalable layers, with a base layer in QVGA format (320×240 pixels) and an enhancement layer in VGA format (640×480 pixels). A wide variety of contents is covered by the video clips with several genres such as documentary, sports, outdoors and news reports. We simulate a loss in the enhancement layer for one second and use data from the base layer to conceal the loss. Two SVC-based error-concealment techniques are simulated using the base layer upscaled to VGA. The "patch" technique replaces only the damaged areas in the frame, whereas the "switch" technique replaces the whole frame as soon as loss is detected. For comparison, an equivalent H.264/MPEG-4 AVC stream is added to the experiment and distorted under the same conditions as the SVC streams. A buffer-repetition AVC-based error-concealment technique is used to compare the behaviour of single-layer coding to multi-layer coding. Finally, several combinations of bitrate and number of frames per second are used for the base layer, in order to determine the best tradeoff between the two layers under a global bitrate constraint for the whole stream.

The second experiment (T2) studies with more details the influence of bitrate and quality repartition among the two scalable layers [2] on 11 video sources (9 of them are common to T1). The same QVGA-VGA configuration was used, under constant QP scenarios for the two layers. A total of 16 SVC scenarios were studied and compared to 4 VGA AVC and 4 upscaled QVGA AVC scenarios. The third experiment (T3) explores the influence of impairment temporal distribution on the perceived quality [3]. We use two constant QP scenarios extracted from T2 combined with the "switch" error concealment technique from T1, and simulate various combinations of length, number and interval between impairments. Finally, the fourth experiment (T4) focuses on the influence of the source content on perceived quality under SVC coding distortions only [2]. A total of 60 source contents are encoded using the same SVC and upscaled AVC QVGA scenarios as in T2. Ten video sequences are common between T2 and T4. The additional 50 clips used in T4 extend the variability of contents by including action scenes, sequences with scenecuts, high and low contrast, human faces and figures, animated clips, complex motion structures and small objects.

All four experiments were conducted under viewing conditions following the ITU BT.500 recommendation regarding lighting, display and room setup. The Absolute Category

Rating with Hidden Reference (ACR-HR) protocol was used with a 5-level scale, following the ITU P.910 recommendation. The original video material was converted from high quality HD sequences to VGA using a reference downscaling algorithm [4], making sure that the change in aspect ratio did not affect the aspect of the pictured scenes. The videos were then encoded using the JSVM Reference Software for scalable streams [4], and the JM Reference Software for AVC streams [5]. The notation T1 to T4 denotes the chronological order in which the experiments were conducted.

The subjective datasets are available for download at the address reported in [6]. The observer votes are provided along with the MOS values per Processed Video Sequence (PVS), per Hypothetical Reference Circuit (HRC) and per source content (SRC). The video data is also available through a dedicated FTP server.

3. ALIGNING EXPERIMENTS ON A COMMON QUALITY SCALE

Despite its popularity, the ACR test methodology has a major drawback. It is not possible to directly compare the outcomes of different tests to each other. Indeed, it has been identified that human subjects tend to rate a stimulus relatively to the range spanned in the whole test [7]. This phenomenon, known as the corpus-effect, is caused by the fact that the observers have to calibrate their judgement of quality using anchor conditions, which are not explicitly displayed when rating a stimulus. Thus, the observers calibrate their judgement using data from the test itself, making their ratings relative to the current experiment.

It is possible to compare the outcomes from several ACR experiments, by aligning them on a common quality scale. The experiments have to share a set of common configurations, in order to determine how to project the votes on the common scale. Several contributions have considered this problem using different approaches. In [8], Pinson and Wolf present the approach used in the VQEG Multimedia test plan to align the outcomes on a single scale. The average score obtained over all experiments for the common set is first processed. This set of "Grand Means" is then used as a common scale onto which the single outcomes are mapped using a linear function.

A similar approach is presented by Garcia and Raake [7], where a reference test is selected as the reference quality scale. The outcomes of the remaining tests are mapped on this reference scale, using a set of common stimuli. The reference test is chosen as the one covering the widest range of qualities and types of distortions in order to provide a robust calibration.

In the four experiments presented in this paper, we used a slightly different approach to align the ratings on a common quality scale. Our main goal was to provide an approach suitable for the regular activities in a subjective quality assessment laboratory. In the two mentioned contributions, the set of experiments was planed at once and the design of the common set was included in the initial effort. For our experiments, we did not have such an insight on the upcoming test campaigns. Each experiment was designed as an extension of the previous ones and choices were made in their design after analysing the results of the previous experiments. The experimental material was constructed in successive rounds and the content of T4 was not known when T1 was designed. Therefore, it was not possible for us to design an optimal common set of configurations for the four datasets.

As a result, the common set was designed progressively, including significant parts of the previous experiments in each new experiment. Table 1 gives an overview of the number of common PVS between each pair of experiments. As one can see, T2 shares the highest number of PVS with the three other experiments. Therefore we use this experiment as a reference and map the votes of the three other experiments onto the scale of T2.

The design of the common set is a critical step. In [8], Pinson and Wolf draw a list of constraints to address in order to build a reliable common set. It is specified that the common set should span the entire range of contents and qualities included in all the experiments, and that the common PVS should be evenly distributed along the quality scale. Figure 1 shows the distribution of the common PVS in our four experiments. We observe that the full MOS scale is covered and that the PVS are evenly distributed along the scale. An exception has to be made of the common PVS between T3 and T4, for which the middle of the scale is not properly covered. This is one of the motivations for choosing T2 as a reference experiment, as we can use it as an intermediate between T3 and T4. The number of configurations to include in the common set is recommended to be comprised between 10% and 20% of the total number of PVS in a subjective test. Including less configurations might lead to a less robust fitting, whereas including more configurations introduces a bias in the data. In our experiments, T2 shares between 7% and 17% of common PVS with the three other tests. Considering the number of HRC and SRC we needed to test in the main parts of the experiment, we had to choose to decrease the number of common PVS in order to keep the duration of a test session under the acceptable limit for observers. However, we assume that the highly linear tendency of the relation between common PVS and the full coverage of the quality scale provides a reliable fitting between the experiments.

In order to avoid adding more bias in the data, the merging process must be carried out carefully. Before merging several experiments on a single quality scale, one should remove the duplicates formed by the common configurations. A common way to perform this step is to keep either the version of the PVS that was contained in the reference experiment, or the closest PVS to the grand mean in case the experiments are aligned on a mean quality scale. In our case, we kept the votes

for the PVS from T2, which is our reference experiment.

After merging experiments, one gains access to new comparisons, for instance between different types of distortion or different source contents. Such as presented in [2], an example of new comparison is the possibility to use 4 HRC from T4 to predict the behaviour of a source in terms of quality on the 20 SVC HRC from T2. Also, by comparing the PVS from T1 and T2, one can observe that an upscaled QVGA AVC stream encoded with a QP of 26 is equivalent to several two-layers configurations impaired by different loss patterns. However, the upscaled version only needs an average bitrate of 0.84 Mb/s to be encoded, which is about half the bitrate needed to transmit one of the equivalent two-layer streams. This result illustrates the tradeoff between AVC and SVC when no network impairments appear. Additionally, links between the loss of quality due to network impairments and coding distortions can be drawn by comparing the outcomes of T2 to the outcomes of T3. For instance, we observed that two impairments of 32 frames separated by an interval of 64 frames in a SVC stream encoded with a QP0 of 44 and QP1 of 32 are equivalent in terms of quality to an AVC stream encoded with a QP of 38 with no impairments. These three observations illustrate how comparisons on the relative and combined influence of coding distortions and packet loss can be evaluated by merging multiple tests on a common quality scale. The merged data is available for download with more details on the alignment procedure at the address in [6].

A critical aspect of inter-experiment comparison is then how to assess the statistical significance of the difference between configurations. Intervals of confidence are usually employed to perform this verification for single experiments, and often confirmed using student t-tests. For merged experiments, these statistical tools might not be perfectly suited. Stimuli from different experiments have not necessarily been evaluated by the same number of observers. The paired student t-test allows to compare two mean values obtained from populations with different sizes, which makes it an acceptable statistical tool for comparing MOS obtained from different experiments. However, the observations made from this type of data should always be considered as indications and not indisputable conclusions.

4. OBJECTIVE METRICS PERFORMANCE

In this section we present possible exploitations of the four subjective datasets, focused on metric design and improvement. To this end, we analyse the accuracy of three common quality metrics under SVC coding distortions, transmission errors and error-concealment artifacts. The three metrics involved are the PSNR, the VQM and the TetraVQM.

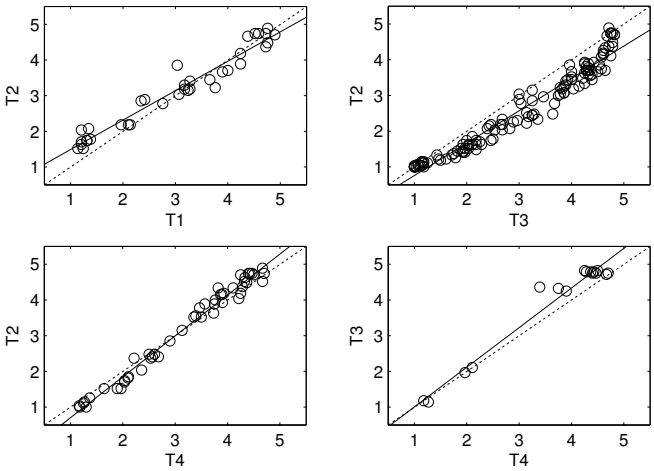

Fig. 1. Relation between the common stimuli scores between the four experiments. Each dot corresponds to a PVS MOS. The solid lines represent the linear approximations, while the dotted lines represent the identity function (perfect alignment).

4.1. Tested quality metrics and motivation

The PSNR is calculated on the luminance component of the videos. In order to get a single value for a whole video sequence, we calculate the arithmetic average of all individual frame PSNR values. This simplistic quality metric only looks at pixel differences between a reference and a distorted stimuli. It does not take into account the properties of the Human Visual System and its performance is known to be relatively lower than perceptual metrics. Nevertheless it remains a popular method and the scores expressed in decibels (dB) are usually well understood by a large community.

The VQM is used in its full-reference version with the general model [9]. This metric is known as one of the most effective for predicting the human rating behaviour for most digital video delivery systems. It takes into account spatial and temporal artifacts and source content information for the calculation of the quality scores. Nevertheless, the authors acknowledge that it was not designed to handle artifacts due to packet-loss or error-concealment.

The TetraVQM is based on the VQM and adds considerations about the possible pausing and skipping introduced by transmission over lossy networks [10]. It is therefore able to predict a particular type of error concealment, which consists in displaying the last non-impaired frame until the distortions due to packet-loss disappear. A spatial distortion module is also included as in the classic VQM, which should allow the metric to detect spatial artifacts due to switching to the base layer.

We would like to emphasize that our goal in this section is not to criticize the performance of these metrics on our datasets. We are particularly aware that the two perceptual metrics were not designed with the scenarios we consider in

mind. On the opposite, we aim at illustrating the potential for improvement of these metrics using the four presented datasets.

4.2. SVC coding artifacts

The T2 dataset can be used to evaluate the performance of objective metrics in predicting the quality perceived by the observers under SVC coding artifacts. Our dataset covers a wide range of quality levels, with QP values comprised between 26 and 44 for both layers. Eleven video sequences are involved, ensuring a fair variability of contents.

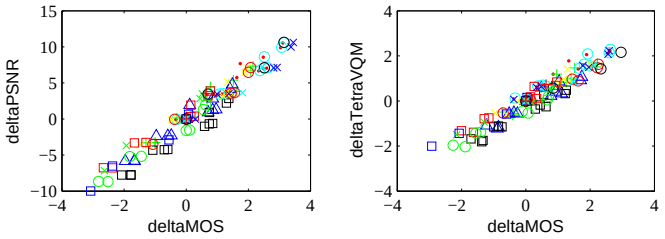

Fig. 2. Difference in MOS values between pairs of HRCs, compared to the difference in metric score on the same pairs of HRCs. Similar behaviour have been observed for the VQM.

Our goal in this section is to show that in some cases, two significantly different PVS in terms of MOS can obtain very similar metric scores. Therefore we look at the differences between pairs of HRCs, first in terms of MOS, then in terms of metric score, such as illustrated in Figure 2. Comparing pairs of HRCs allows us to have more data points to support our analysis (i.e. $16 \times (16-1)/2$ pure SVC HRC pairs). Furthermore it allows us to analyse results in the original scale of the metrics (e.g. in dB for the PSNR), whereas a direct comparison of PVS MOS values would require a fitting, and the results would not be linked to a realistic scale. One can observe that the mismatch between MOS and metric score does not evolve with the magnitude of the difference between HRCs. Therefore, we report only the maximum mismatch between difference in MOS and difference in metric score, under the constraint that the measured difference is below 5% of the quality range covered by the metric. Table 2 reports the maximum values for MOS difference observed for each content, as well as the average over the 11 contents. For an average of 0.052 dB in PSNR, an average of 1.101 on the MOS scale can be observed in the worst case over the 11 contents. This represents more than 20% of the MOS scale, which illustrates the inaccuracies of the metric. Similar results can be observed for the VQM and the TetraVQM, although surprisingly a slight difference in favour of the PSNR can be noticed.

In the case of coding distortions only, the main parameters influencing the perceived quality are the QP values. Under constant QP coding, the distortion level is relatively constant throughout the sequence and its impact on the judgement of observers can be extracted easily. As a result, one could imagine that a bitstream-based quality metric could perform well in this context. However, as soon as packet-loss artifacts are to expect, the performance of such a model would decrease rapidly, as we will illustrate in the next section.

Table 2. Maximum difference between pairs of HRC in terms of MOS and metric score.

Metric	Video Content	δ MOS	δ Metric
PSNR	ShadowBoxing	1.296	0.062
	Stream	1.333	0.086
	Skatefar	0.962	0.076
	Family	0.592	0.004
AVERAGE		1.101	0.052
VQM	ShadowBoxing	1.296	0.005
	Stream	1.777	0.024
	Skatefar	1.555	0.010
	Family	1.777	0.029
AVERAGE		1.461	0.021
TetraVQM	ShadowBoxing	1.555	0.195
	Stream	1.296	0.020
	Skatefar	1.592	0.185
	Family	1.111	0.016
AVERAGE		1.262	0.134

4.3. Error-concealment techniques

Error concealment is a post processing added by the decoder when detecting irregularities in the bitstream. As a result, parametric quality measures based on bitstream analysis might not be able to anticipate it, as no information about the error-concealment technique is usually included in the bitstream. Perceptual metrics are more likely to identify the artifacts due to error concealment, as they look at the PVS directly. In our previous work [1], we identified significant differences between the local SVC concealment (patch) and global SVC concealment (switch). Here we want to question the ability of the metrics to identify these differences. To this end, Figure 3 displays the differences between the two SVC-based error concealment techniques, both in terms of MOS and metric score, for the three metrics. The experiment includes 4 configurations for the base layer with 15 and 30 frames per second and a bitrate equal to 120 kbps and 200 kbps. One can easily observe that the metrics are not able to reproduce the difference in MOS between the local and global concealemnt techniques. Additionally, it seems that the behaviour of the metrics cannot be easily predicted, as no particular pattern appears in the figure.

We are aware that the involved metrics were not built to cope with SVC-based error concealment. Unlike frame pausing or skipping, the patch and switch methods introduce only blurring in the concealed frames. In case of a difference in the number of frames per second between the two layers, temporal discontinuities are also introduced, which are perceived as

quite disturbing by the observers, as shown in our previous work. The metrics do not seem to capture the impact of these discontinuities, possibly because the amount of blurring is not severe enough.

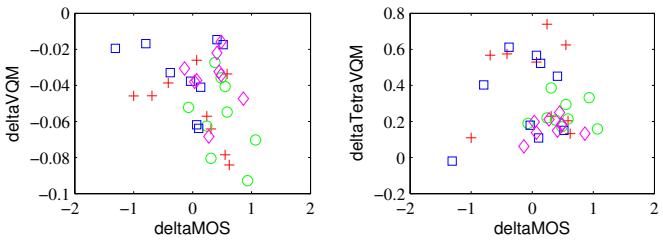

Fig. 3. Difference in terms of MOS and in terms of metric score between the "patch" and the "switch" error-concealment techniques. Each point corresponds to one PVS. "+": base layer encoded at 120 kbps, 15 fps; "□": 200 kbps, 15 fps; "◯": 120 kbps, 30 fps; "◇": 200 kbps, 30 fps. Similar results have been observed for the PSNR.

4.4. Impairment temporal distribution

It is well accepted that the distribution of artifacts due to network impairments has an influence on the perceived quality. The T3 dataset was designed to determine the parameters involved in a SVC based-scenario where the switch error-concealment technique is used to conceal impairments. In our previous work [3], we identified that the main parameters involved in the perception of quality in this context are the number of frames displayed from the base layer, the quality of the base layer itself and the number of impairment events.

This experiment was designed in a systematic way to facilitate the extraction of models for the influence of the impariment distribution on the perceived quality. The influence of each parameter can be analysed both independently and jointly with the other parameters. This dataset is therefore particularly suited for the design of an hybrid model, analysing both the bitstream to identify missing packets and encoding parameters, and the decoded video to determine the severity of the loss and get indications on the performance of the error-concealment technique.

4.5. Influence of source content

The T4 dataset uses the same SVC conditions as T2, on a set of 60 source contents covering a wide range of genres and complexity levels. In our previous work [2], we identified significant differences in terms of MOS between different source contents under the same coding conditions. A preliminary analysis of the behaviour of the PSNR on these datasets exhibited a common mismatch with the MOS which commonly reached one MOS category after a conventional third-order fitting. This early observation suggests that the source characteristics in this context have an impact on the performance of the quality metrics. Our dataset can therefore be used to train the metrics on a wide set of source contents and improve the way the characteristics of the source are taken into account in their models.

5. DISCUSSION AND CONCLUSION

In this paper we presented four subjective video datasets and their possible exploitation for the design and improvement of objective quality metrics. We showed that the influence of error-concealment artifacts was particularly difficult to predict using the considered metrics. We also identified possible directions to improve the metrics in their ability to handle Scalable Video Coding under network impairments.

6. REFERENCES

[1] Y. Pitrey, M. Barkowsky, P Le Callet, and R. Pepion, "Evaluation of MPEG4-SVC for QoE protection in the context of transmission errors," *SPIE Optical Eng.*, 2010.

[2] Y. Pitrey, M. Barkowsky, and R. Pepion, "Influence of the Source Content and Encoding Configuration on the Perceived Quality for Scalable Video Coding," in *SPIE Human Vision in Elect. Imaging*, 2012.

[3] Y. Pitrey, U. Engelke, M. Barkowsky, and R. Pepion, "Subj. Quality Of Svc-Coded Videos With Different Error-Patterns Concealed Using Spatial Scalability," *IEEE EUVIP*, 2011.

[4] Joint Video Team, "JSVM Reference Software, Version 9.18," http://ip.hhi.de/imagecom_G1/savce/downloads/.

[5] Joint Video Team, "Joint Video Model Reference Software," http://iphome.hhi.de/suehring/tml/.

[6] IRCCyN IVC Research Group, "Subjective Video Databases," http://www.irccyn.ec-nantes.fr/spip.php?article491.

[7] M.N. Garcia and A. Raake, "Normalization of subjective video test results using a reference test and anchor conditions for efficient model development," in *QoMEX*. 2010, IEEE.

[8] M. Pinson and S. Wolf, *Techniques for evaluating objective video quality models using overlapping subjective data sets*, US Dept. of Commerce, Nat. Telecom. and Information Admin., 2008.

[9] ITU-R, "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference," *ITU-R BT.1683*, 2004.

[10] M. Barkowsky, J. Bialkowski, and R. Bitto, "Temporal Trajectory Aware Video Quality Measure," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–13, 2008.