Including the Effects of Video Content in the ITU-T G.1070 Video Quality Function

Jose Joskowicz
Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
Phone: +598 2 7110974
josej@fing.edu.uy

J. Carlos López Ardao
ETSE Telecomunicación
Campus Universitario, 36310
Vigo, Spain
Phone: +34 986 8212176
jardao@det.uvigo.es

Rafael Sotelo
Facultad de Ingeniería
Universidad de Montevideo
Montevideo, Uruguay
Phone: +598 2 7067630
rsotelo@um.edu.uy

ABSTRACT
In this paper we present an enhancement to the video quality estimation model described in ITU-T Recommendation G.1070 "Opinion model for video-telephony applications", including the effect of video content in the G.1070 video quality function. This enhancement provides a much better approximation of the model results with respect to the perceptual MOS values. SAD (Sum of Absolute Differences) is used as an estimation of the video spatial-temporal activity. The results are based on more than 1500 processed video clips, coded in MPEG-2 and H.264/AVC, in bit rate ranges from 50 kb/s to 12 Mb/s, in SD, VGA, CIF and QCIF display formats.

Categories and Subject Descriptors
H.4.3 [**Information Systems Applications**]: Communications Applications - *Computer conferencing, teleconferencing, and videoconferencing*; C.4 [**Performance of Systems**]: Design Studies, Modeling Techniques

General Terms
Algorithms, Performance, Design, Standardization

Keywords
Video perceptual quality, Video codecs, Video signal processing, VoIP Network design.

1. INTRODUCTION
Video-telephony applications are growing quickly in the market, using IP (Internet Protocol) as the underlying protocol. In these emerging applications it is critical to provide an appropriate QoE (Quality of Experience) for the end user, in accordance to the offered service. QoE can be defined as the overall performance of a system, especially from the user perspective. Many factors can affect the QoE in video-telephony, but the audio and video qualities are the most important aspects to consider.

ITU-T Recommendation G.1070 "Opinion model for video-telephony applications" [1] describes a computational model for videophone applications over IP networks that is useful as a QoE planning tool. This model takes into account the bit rate, frame rate and packet loss rate for video quality estimation, but does not take into account video content. Video perceived quality can have great variations for different video contents for the same video codec, bit rate and frame rate.

In previous works [2][3], we showed that the video content must be taken into account in order to provide an appropriate estimation of the perceived video quality. In this paper, we propose how to include the video content in the G.1070 model, taking into account the contributions of our previous works.

The new parameters and relation proposed are calculated for MPEG-2 [4] and H.264/AVC [5] in bit ranges from 50 kb/s to 12 Mb/s, and in display formats SD (Standard Definition, 720 × 576 pixels), VGA (Video Graphics Array, 640 x 480 pixels), CIF (Common Intermediate Format, 352 × 288 pixels) and QCIF (Quarter Common Intermediate Format, 176 × 144 pixels)

The MSE (Mean Square Error) and Pearson correlation for the original and the proposed model are calculated and compared.

The paper is organized as follows: Section 2 briefly describes the video quality function proposed in G.1070. Section 3 describes previous enhancements proposed for the G.1070 model. Section 4 shows the effects of video content in the perceived quality. Section 5 describes how to include the effect of video content in the G.1070 model, and a comparison is made between the original model and the proposed enhancements. Section 6 summarizes the main contributions.

2. VIDEO QUALITY FUNCTION IN ITU-T RECOMMENDATION G.1070
The ITU-T Recommendation G.1070 describes a computational model for point-to-point interactive videophone applications over IP networks that is useful as a QoE (Quality of Experience) and QoS (Quality of Service) planning tool for assessing the combined effects of variations in several video and speech parameters that affect the perceived quality. The model takes into account the speech and the video perceived quality [6], and combines both in an integration function for overall multimedia quality [7]. Speech quality estimation is based on the ITU-T Recommendation G.107 [8], known as the E-Model. Video quality estimation V_q is calculated as shown in Equation (1).

$$V_q = 1 + I_c e^{\frac{P_{plv}}{D_{Pplv}}} \qquad (1)$$

where I_c represents the basic video quality (as defined in Equation (2)), determined by the codec distortion and is a function of the bit rate and frame rate, P_{plv} is the packet loss rate and D_{Pplv} expresses the degree of video quality robustness due to packet loss, and can also depend on bit rate and frame rate.

The basic video quality I_c for each bit rate b (the video quality due to encoding degradation only, without packet loss) is expressed in Equation (2).

$$I_c = v_3\left(1-\frac{1}{1+\left(\frac{b}{v_4}\right)^{v_5}}\right) \qquad (2)$$

where b is the bit rate and v_3, v_4 and v_5 are three coefficients of the model.

Combining Equation (1) and (2), the video quality V_q without packet loss is expressed in Equation (3).

$$V_q = 1+v_3\left(1-\frac{1}{1+\left(\frac{b}{v_4}\right)^{v_5}}\right) \qquad (3)$$

According to the ITU-T Recommendation G.1070, coefficients v_3, v_4 and v_5 are dependent on codec type, video display format, key frame interval, and video display size, and must be calculated using subjective video quality tests. Provisional values are provided only for MPEG-4 in QVGA (Quarter VGA, 320 × 240 pixels) and QQVGA (Quarter QVGA, 160 × 120 pixels) video formats. In [9] a new set of parameters for the MPEG-2 codec are proposed.

In [10], the same model presented in equation (3) is proposed for IPTV services in HD (High Definition, 1440 x 1080 pixels), and coefficient values are provided for the H.264 codec.

The model coefficients are not dependant on video content.

3. ENHANCEMENT TO ITU-T RECOMMENDATION G.1070

In a previous work [3] we proposed some enhancements to the G.1070 model:

a) One of the model parameters is suppressed (v_3), and performance is not affected.

b) One new parameter is added (a), that takes into account the display format

c) We showed that the two remaining parameters (v_4, v_5) are highly correlated to subjective video movement content, and we defined three sets of these two parameters, one for "low movement content" applications, one for "medium movement content" and one for "high movement content" applications.

The Enhanced G.1070 model presented in the referred paper is expressed in Equation (4)

$$V_q = 1+4\left(1-\frac{1}{1+\left(\frac{ab}{v_4}\right)^{v_5}}\right) \qquad (4)$$

where V_q represents the video quality determined by the codec distortion, b is the bit rate (in Mb/s), a is a constant that depends on the display format and v_4 and v_5 are other model parameters with a strong dependence on video content. Video clips are classified in three categories, according to the subjective movement content, and the model parameters v_4 and v_5 are calculated for each category (High, Medium and Low movement content).

The best values for each coefficient were calculated in [3], and are presented here: The coefficient a depends on the display format, according to Table 1. The coefficients v_4 and v_5 depends on the "subjective movement content", and are presented in Table 2. In that paper, the movement content was derived only based on qualitative analysis, no quantitative analysis was proposed to the video classification into each category.

Table 1. Best values for *a*

Display Format	*a*
SD	1
VGA	1.4
CIF	3.2
QCIF	10.8

Table 2. v_4 and v_5 values for each movement content

Movement	v_4	v_5
Low Movement	0.366	1.32
Medium Movement	0.67	1.36
High Movement	1.088	1.56

4. EFECTS OF VIDEO CONTENT IN PERCEIVED QUALITY

For this paper, we used the NTIA (National Telecommunications and Information Administration) "General Full Reference Model", available in [11], and standardized in ITU-T Recommendation J.144 [12] for perceived MOS (Mean Opinion Score) estimation. For each video clips pair (original and degraded), the NTIA algorithm provides a VQM (Video Quality Metric), with values between 0 and 1 (0 when there are no perceived differences and 1 for maximum degradation) that can be directly associated with the DMOS (Differential Mean Opinion Score). The DMOS values returned from the NTIA model can be related to the MOS using Equation (5). The interpretation of the MOS values is presented in Table 3.

$$MOS = 5 - 4DMOS \qquad (5)$$

Table 3. MOS to perceived quality relation

Quality	Bad	Poor	Fair	Good	Excellent
MOS	1	2	3	4	5

The error obtained using the standardized NTIA model with respect to subjective tests can be estimated in +/- 0.1 in the 0-1 scale. This means that the order of magnitude of the standardized algorithm error is 0.1 in a DMOS scale from 0 to 1 [13]. MOS errors, using this model, can be estimated in +/-0.4 in the 1-5 scale (4 times the DMOS error).

Figure 1 shows the relation between MOS and bit rate, for the sixteen clips detailed in table 4, coded in MPEG-2 (using the coding parameters detailed in Table 5), in SD display format. Similar curves are obtained for other codecs (i.e. H.264/AVC) and display formats (i.e. VGA, CIF, QCIF). MOS values were derived from DMOS, using Equation (5). DMOS values were calculated using the NTIA Model. The video clips used (available in the VQEG web page [14]) have durations of 8 – 10 seconds and spans over a wide range of contents, including sports, landscapes, "head and shoulders" and so.

Table 4. Source video clips used

Source	Name	Source	Name
src 2	Barcelona	src 14	New York 2
src 3	Harp	src 16	Betes pas betes
src 4	Moving graphic	src 17	Le point
src 5	Canoa Valsesia	src 18	Autums leaves
src 7	Fries	src 19	Football
src 9	Rugby	src 20	Sailboat
src 10	Mobile & Calendar	src 21	Susie
src 13	Baloon-pops	src 22	Tempete

Table 5. MPEG-2 and H.264 coding parameters used

MPEG-2	H.264
Profile/Level: MP@ML Max GOP size: 15 GOP Structure: Automatic Picture Structure: Always Frame Intra DC Precision: 9 Bit rate type: Constant Bit Rate Interlacing: Non-Interlaced Frame Rate: 25 fps	Profile/Level: High/3.2 Max GOP size: 33 Number of B Pict between I and P: 2 Entropy Coding: CABAC Motion Estimated Subpixel mode: Quarter Pixel Bit rate type: Constant Bit Rate Interlacing: Non-Interlaced Frame Rate: 25 fps

As can be seen in Figure 1, all the clips have better perceived quality for higher bit rates, as expected. In MPEG-2, in SD, for bit rates higher than 6 Mb/s all the clips have an almost "perfect" perceived quality (MOS higher than 4.5). At 3 Mb/s all the clips are in the range between "Good" and "Excellent". However for less than 3 Mb/s the perceived quality strongly depends upon the clip content. For example at 2 Mb/s, MOS varies between 3.6 and 4.8, and at 0.9 Mb/s MOS varies between 1.9 (between "Bad" and "Poor") and 4.2 (between "Good" and "Excellent"). Similar results are obtained for different display formats and codecs.

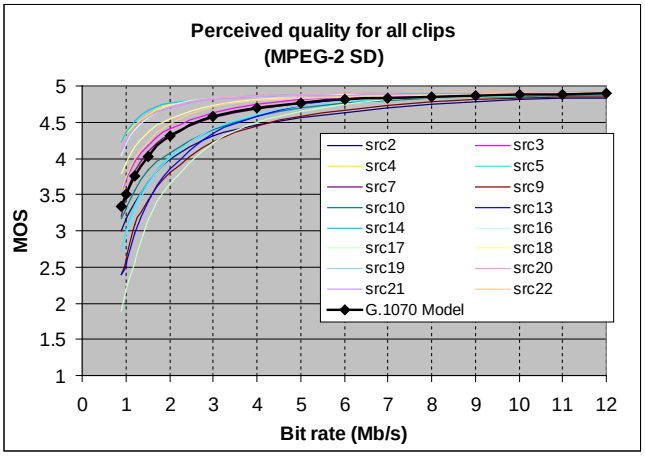

Figure 1. Perceived Quality as a function of the Bit Rate and video content

ITU-T Recommendation G.1070 does not take into account the video content. The best that the model can estimate is the *average* video quality for different contents, at each bit rate, as shown in Figure 1 in a bold black line. However, when we analyze the graph in Figure 1, we can see that video quality strongly depends on video content, especially for low bit rates. In our previous works [2][3] we proposed to classify the video clips into three categories, according to the subjective movement content, and different set of values for the model were calculated for each category. In the following sections, we will describe how to derive the Enhanced G.1070 model coefficients from objective video quality spatial-temporal estimations.

5. INCLUDING VIDEO CONTENT IN THE ENHANCED G.1070 MODEL

Each curve in Figure 1 can be modeled with Equation (4), and the best values for v_4 and v_5 can be obtained for each clip. Table 6 shows the values of v_4 and v_5 that best fits Equation (4) to each curve in Figure 1, as well as the MSE (Mean Square Error) and the subjective movement content ("Mov" column, classified into "Low", "Medium" and "High"). The MSE is related to the "distance" between the estimated and the actual values. Lower values of MSE means lower "distances", and therefore better estimations.

Subjective movement content is related to the video spatial-temporal activity. Different estimations for the video spatial-temporal activity were evaluated, and a strong correlation between the v_4 and v_5 parameters with the average SAD (Sum of Absolute Differences) of the original clip has been found. SAD is a simple video metric used for block comparison and for moving vectors calculations, and can be efficiently calculated [15][16]. Each frame is divided into small blocks (i.e. 8x8 pixels) and for every block in one frame the most similar (minimum SAD) block in next frame is found. This minimum sum of absolutes differences is assigned as the SAD for each block in each frame (up to the n-1 frame). Then all the SAD values are averaged for each frame and for all the frames in the clip, and divided by the block area, for

normalization. This value (average SAD/pixel) provides an overall estimation about the spatial-temporal activity of the entire video clip.

The relations between the v_4 and v_5 parameters for MPEG-2 encoding with the average SAD per pixel are presented in Table 6. The table also shows the MSE obtained with these values for each clip using Equation (4), for SD, VGA, CIF and QCIF display format, and bit rates from 50 kb/s to 12 Mb/s. The table is ordered by v_4 and the subjective movement content is also presented. A similar order is obtained when using H.264/AVC. The previous categories, based on subjective movement content, can be mapped to the SAD values, according to Table 7.

Table 6. v_4 and v_5 values that best fits to the actual NTIA curves and Avg SAD/pixel

Src	Name	v_4	v_5	MSE	Avg SAD/ pixel	Mov
4	Moving graphic	0.252	1.2	0.0288	0.684	Low
21	Susie	0.29	1.2	0.0476	1.251	Low
20	Sailboat	0.29	1.24	0.0334	1.303	Low
14	New York 2	0.252	1.2	0.0441	1.386	Low
18	Autums leaves	0.442	1.28	0.0395	1.599	Low
16	Betes pas betes	0.328	1.28	0.0537	1.804	Low
22	Tempete	0.594	1.48	0.0365	3.315	Med
3	Harp	0.594	1.4	0.0336	3.457	Med
10	Mobile & Calendar	0.784	1.32	0.0172	3.600	Med
7	Fries	0.708	1.44	0.0603	3.632	Med
2	Barcelona	0.86	1.24	0.0228	4.243	High
19	Football	1.05	1.68	0.0302	4.520	High
5	Canoa Valsesia	1.012	1.6	0.0465	5.148	High
13	Baloon-pops	1.24	1.84	0.0356	5.656	High
9	Rugby	1.24	1.6	0.0647	6.164	High
17	Le point	1.506	2.04	0.0686	8.256	High

Table 7. Movement content vs SAD

Movement	Avg SAD/pixel
Low Movement	$s < 2$
Medium Movement	$2 \leq s < 4$
High Movement	$4 \leq s$

Figure 2 shows the relation between v_4 and average SAD per pixel, for MPEG-2 and H.264/AVC codecs. Similarly, Figure 3 shows the relation between v_5 and average SAD per pixel, for MPEG-2 and H.264/AVC codecs. In these figures, the subjective movement content is graphically showed with different colors, confirming that low values for SAD/pixel are related to low subjective movement content and high values are related to high movement content.

An estimation of v_4 and v_5 for MPEG-2 and H.264/AVC, as a function of the average SAD/pixel can be performed as presented in Equation (6).

$$v_4 = c_1 s^{c_2} + c_3$$
$$v_5 = c_4 s^{c_5} + c_6 \qquad (6)$$

where s is the original video average SAD per pixel. The best values for c_1.. c_6 were calculated (using least squares method) and are presented in Table 8.

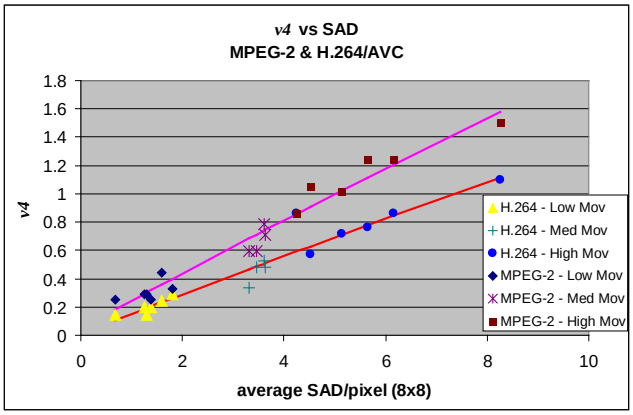

Figure 2. Relation between v_4 with respect to SAD

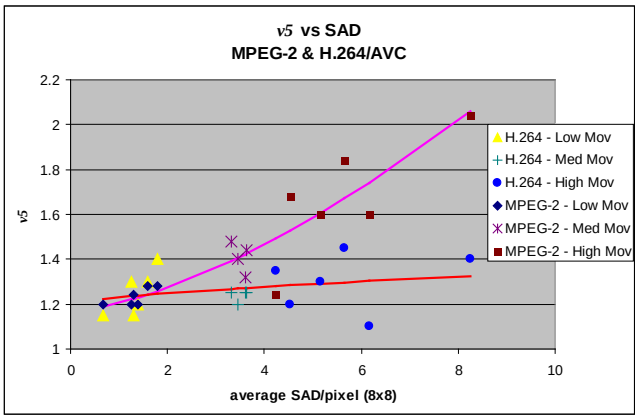

Figure 3. Relation between v_5 with respect to SAD

Table 8. MPEG-2 and H.264 coding parameters used

Codec	c_1	c_2	c_3	c_4	c_5	c_6
MPEG-2	0.208	0.95	0.036	0.036	1.52	1.17
H.264/AVC	0.150	0.95	0	0.030	0.68	1.20

Using Equation (6), the video quality estimation presented in

Equation (4) only depends on the encoded bit rate and the spatial-temporal activity of the original video clip, measured as the average SAD/pixel. The dispersion between the MOS values derived using Equations (4) and (6) and the perceived MOS values (using the NTIA VQM), for the sixteen video clips used, coded in MPEG-2 and H.264/AVC (using the coding parameters detailed in Table 5), in SD, VGA, CIF and QCIF display format, with bit rates from 25 kb/s to 12 Mb/s are plotted in Figure 4. In this figure, each point represents a video clip coded in a specific combination of codec, bit rate and display format. It is worth noting that subjective rating scales have ranges of 1 unit, in the 1-5 MOS scale. On the other hand, the NTIA algorithm standardized by ITU has errors in the order of +/- 0.4 regarding to MOS measures of subjective quality. In Figure 4, the dotted lines represent the estimated +/- 0.4 error margin of the NTIA model. Only 27 from the 2064 (about 1%) points are outside the dotted lines, meaning that the predicted MOS values have the same degree of precision than the VQM standard that has been used. The Pearson correlation between the values derived using Equations (4) and (6) and the perceived MOS values (using the NTIA VQM) is 0.985. The Pearson correlation metric evaluates the precision of the prediction. It varies from 0 to 1, where 1 indicates a direct relationship and 0 indicates no relationship at all. In this case, 0.985 indicates a very high correlation between the values derived using Equations (4) and (6) and the perceived MOS values.

The MSE between the values derived using Equations (4) and (6) and the perceived MOS values is 0.015, meaning a very low "distance" between the model and the standard VQM.

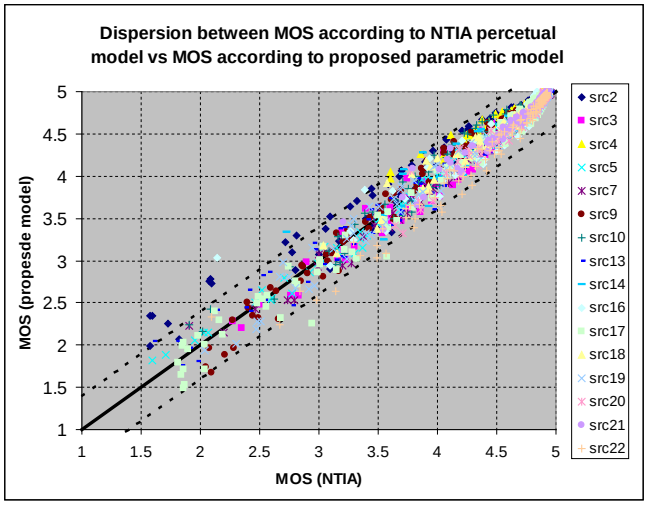

Figure 4. MOS dispersion in proposed model

Figure 5 represents the same data and shows the dispersion using the original G.1070 model, where only one value of v_4 and v_5 is used for all the clips of each codec. In this case, 411 from the 2064 (about 20%) points are outside the dotted lines. The Pearson correlation between the values derived from the original G.1070 model and the perceived MOS values (using the NTIA VQM) is 0.863. The MSE is 0.13. A comparison summary is presented in Table 9.

Table 9. Original and proposed model comparison

Model	*Pearson Correlation*	*MSE*
Proposed model, considering content (derived from average SAD/pixel)	0.985	0.015
Original G.1070	0.863	0.13

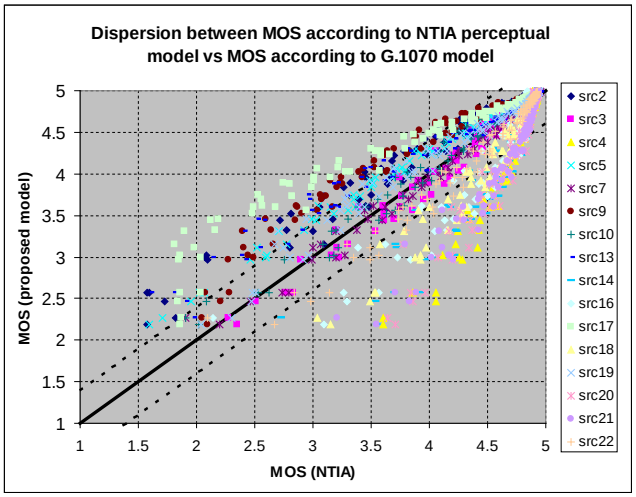

Figure 5. MOS dispersion in G.1070 original model

6. CONCLUSION

The video quality estimation proposed in ITU-T Recommendation G.1070 "Opinion model for video-telephony applications" does not take into account video content. We have shown that, specially in the low bit rate range, the video quality have high variations depending on video content for same bit rate and other codec parameters. In this work, an enhancement has been proposed to the model, including the characteristics of video content based on the average SAD/pixel of the original clip as a spatial-temporal estimation. It was found a strong relation between two model parameters (v_4 and v_5) and the average SAD/pixel of the original video, and a mathematical equation was proposed for modeling this relation.

The proposed model enhancement has been evaluated, using sixteen different video sources, spanning over a wide range of contents, including sports, landscapes, animated pictures, "head and shoulders", and so on. These video clips were coded in MPEG-2 and H.264/AVC, in SD, VGA, CIF and QCIF display, and with bit rates from 50 kb/s to 12 Mb/s. In total 96 different formats were used, and more than 1500 processed video sequences were analyzed. The Pearson correlation of the original model is 0.863 for all these processed video clips and the MSE is 0.13, while the Pearson correlation of the proposed enhancement is 0.985 and the MSE is 0.015.

The result shows that the proposed enhancements perform much better than the original model and fits very well with respect to the perceptual video quality estimations derived from the standardized ITU-T VQM.

7. REFERENCES

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

[2] Jose Joskowicz, José-Carlos López-Ardao, Miguel A. González Ortega, and Cándido López García: A Mathematical Model for Evaluating the Perceptual Quality of Video, FMN 2009, LNCS 5630, pp. 164–175, 2009

[3] Jose Joskowicz, J. Carlos López Ardao: Enhancements to the Opinion Model for Video-Telephony Applications, LANC'09, September 24–25, 2009, Pelotas, Brazil , pp. 87-94

[4] ISO/IEC 13818-2:2000. Information technology – generic coding of moving pictures and associated audio information: Video.

[5] ITU-T H.264 Advanced Video Coding for Generic Audiovisual Services, March 2005

[6] Kazuhisa YAMAGISHI and Takanori HAYASHI: Opinion Model for Estimating Video Quality of Videophone Services, IEEE Global Telecommunications Conference, Nov. 27 2006

[7] Takanori Hayashi, Kazuhisa Yamagishi, Toshiko Tominaga, and Akira Takahashi: Multimedia Quality Integration Function for Videophone Services, IEEE Global Telecommunications Conference, 26-30 Nov. 2007

[8] ITU-T Recommendation G.107 The E-model, a computational model for use in transmission planning,March 2005

[9] Belmudez, B.; Möller, S: Extension of the G.1070 video quality function for the MPEG2 video codec, 2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)

[10] K Yamagishi and T Hayashi: Parametric Packet-Layer Model for Monitoring Video Quality of IPTV Services, IEEE International Conference on Communications 2008 (ICC 08), 19 May 2008

[11] Video Quality Metric (VQM) Software [Online]. Available at: www.its.bldrdoc.gov/n3/video/vqmsoftware.htm

[12] ITU-T Recommendation J.144 Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference, February 2004

[13] Margaret H Pinson and Stephen Wolf: A New Standardized Method for Objectively Measuring Video Quality, IEEE Transactions on Broadcasting, Volume 50, Issue 3, September 2004, pp. 312-322

[14] VQEG Phase I Test Sequences. [Online]. Available at: ftp://vqeg.its.bldrdoc.gov/SDTV/VQEG_PhaseI/TestSequences/Reference/

[15] J. Vanne, E. Aho, T.D.Hamalainen, K. Kuusilinna: A High-Performance Sum of Absolute Difference Implementation for Motion Estimation, IEEE Transactions on Circuits and Systems for Video Technology, Volume 16, Issue 7, July 2006 Page(s):876 – 883

[16] Joaquín Olivares, Javier Hormigo, Julio Villalba and Ignacio Benavides: Minimum Sum of Absolute Differences Implementation in a Single FPGA Device, Lecture Notes in Computer Science (LNCS), Volume 3203/2004, pp 986-990