This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2011 proceedings

Quality of Experience measurement tool for SVC video coding

Kamal Deep Singh[1], Adlen Ksentini[2], Baptiste Marienval[1]

[1] INRIA - Rennes Bretagne Atlantique
Campus Universitaire de Beaulieu, 35042 Rennes, France
Email : (kamal.singh, baptiste.marienval)@irisa.fr

[2] IRISA – University of Rennes 1
Campus Universitaire de Beaulieu, 35042 Rennes, France
Email : adlen.ksentini@irisa.fr

***Abstract*– The scalable extension of H.264, known as Scalable Video Coding (SVC), is recently finalized and adapted by the Joint Video Team. Video scalability is achieved in the temporal, spatial, quality (SNR), or any combination of these domains. One example of using scalability is in saving bandwidth when the same media content is required to be sent simultaneously at different resolutions to support heterogeneous devices and networks. Meanwhile, Quality of Experience (QoE) is the key criteria for evaluating the video service such as SVC. Unlike Quality of Service (QoS) metrics (such as bandwidth, delay, jitters, etc.), QoE is more accurate to reflect the user experience as it considers human visual system and its complex behaviour towards distortions in the displayed video sequence. In order to evaluate QoE, objective assessment tools may not correlate well with the human perceived video quality and at same time, subjective quality assessment methods are costly and time consuming.**

In this paper, we design an automatic QoE measuring tool for SVC video coding mechanism. The proposed module is based on PSQA (Pseudo Subjective Quality Assessment tool), which is a hybrid (objective/subjective) assessment tool. PSQA uses RNN (Random Neural Network) to capture the non-linear relation between the video coding as well as the network parameters affecting the video quality, and QoE. The results clearly show that our module can accurately estimate QoE for SVC video streams.

Keywords: *QoE, SVC, PSQA, No-reference tool, Random Neural Network*

I. INTRODUCTION

One of the multimedia market trends is audiovisual service (TV or VoD) anywhere, at any time. To support such service, a Video Service Provider has to manage, store, and distribute content towards multiple kinds and scales of terminals, and over different and transient access technologies to reach the end user. To solve such issues, video scalability seems to be the most relevant solution. It encodes the video in multiple separated layers and this enables a large number of users with heterogeneous capability to view any desired video stream, at anytime, and from anywhere. One of the most well known scalable standards is the Scalable Video Coding (SVC) extension of H.264/MPEG-4 AVC video compression [1]. The scalability in SVC is achieved by taking advantage of layered approach already known from previous experiences with different video coding approaches. It has three fundamental types of scalabilities: spatial, temporal, and quality (Signal-to-Noise Ratio or SNR). A typical SVC stream includes one base layer and one or several enhancement layers. The removal of an enhancement layer still leads to reasonable quality of the decoded video at reduced temporal, spatial or/and SNR level. The base layer conforms to existing H.264/MPEG4-AVC profile, ensuring backward compatibility with existing receivers. Within SVC, Video Service Providers have the possibility to constitute a set of layer combinations to create the SVC video streams. This will allow them to target different spatial as well as temporal dimensions, in order to be aware of the user environment. To evaluate such combinations, and monitor the performance of the SVC encoding scheme, in term of user experience, there is a need to automatically estimate the Quality of Experience (QoE) of SVC video streams.

QoE is defined in [2] as "the overall acceptability of an application or service, as perceived subjectively by the end user" QoE is different from Quality of Service (QoS) network indicators in terms of bandwidth, loss rate, jitter, which are not sufficient to get a precise idea about the visual quality of a received video sequence. QoE instead focuses on the overall experience of the end user. It depends on the global system behaviour, going from the source of a given service up-to the final user, including the content itself and the network performance.

Besides evaluating the quality of SVC streams, QoE could be used in optimizing the SVC transmission (i.e., the number of layers to send to an end user) through networks. In fact, adapting the SVC streams to user environment may be beneficial in situations when several users with diverse characteristics have to be served or when a gateway has to convey the streams to a downstream network offering lower bandwidth or suffering from congestion or increased packet losses; often, wireless access networks are subject to such conditions. It is possible, with a MANE (Media Aware Network Element) (in context of IP transmission) [3] to adjust the number of layers sent to an end user. The MANE could be associated with a gateway or a router, and it has the ability to parse the video stream and differentiate between SVC layers. Based on the QoE feedbacks, the MANE can discard portions of the original bitstream whereby the base layer is always compliant to AVC. For instance, the MANE can take an action to drop some SVC layers when the overall QoE is going low.

In this paper, we propose an automatic QoE estimation module for the SVC coding scheme that supports SNR scalability. The QoE module is based on PSQA (Pseudo Subjective Quality Assessment) [4], which is an hybrid (subjective and objective) evaluation technique using a specific learning tool (Random Neural Network) to capture the non-linear relationship between networks as well as video encoding parameters, and QoE. To the best of our knowledge this is the first work addressing automatic (and real-time) QoE measuring for SVC video coding.

The remainder of this paper is structured as follows. The next section gives a summary of the QoE concepts and related works. Section III, presents our proposed QoE tool for SVC. In section IV we present and discuss the results obtained by the proposed tool. Finally, we conclude the paper in section V.

978-1-61284-233-2/11/$26.00 ©2011 IEEE

II. QoE CONCEPTS AND BACKGROUND

A. QoE concepts

There are several factors that can influence QoE for video applications. Characteristics such as frame rate, of a video stream, impact the fluidity of the video and a lower frame rate means "choppiness" that can degrade the perceived quality. In addition, spatial video resolution is another significant factor and depending on the limitations of the end device, the users may prefer the highest available resolution. Quantization is another important factor, related to the (lossy) compression of the video stream. Thus, during compression, some amount of information is thrown away and this will introduce certain distortion in the video that in turn will have an impact on QoE.

In addition to the preceding remarks, the type of video content itself may have significant importance. For example, a documentary or news video might have lesser frame rate requirements, but better quantization requirements. On the other hand, a fast moving sports video will have higher frame rate requirements to ensure good QoE. Moreover, the network used to provide the service can significantly impact the video quality. For example, packet losses can strongly degrade the video perceived quality. Delays and jitter in the network can introduce, first, a long initial delay before video can start to play, and then, play-out disruptions and eventual data losses because of the late video packets that miss the play-out deadline. In addition, other parameters like network bandwidth impose limitations on the video characteristics because some quality of the video will have to be sacrificed, either by lowering the frame rate or by using more compression, to accommodate the video with the available bandwidth.

B. Related Works

Objective quality assessment tools, such as PSNR (Picture Signal to Noise Ratio), may not correlate well with the human perceived video quality and on the other hand the subjective quality assessment methods [5] are costly and time consuming. Thus, automatic QoE monitoring is highly desired. There has been considerable work in the literature to propose QoE models. Some signal based ones, such as in [6],[7] take full original video as a reference, or at least as a partial reference, to estimate the final QoE.

Some no-reference models have been proposed. An "opinion model" is proposed in ITU G.1070 [8]. The work in [9] showed that it is possible to enhance this model and make it more precise by replacing, for example, packet loss rate with packet loss event rate. A further extension is proposed in [10].

The full reference model is only applicable at the encoder where the original video sequence is available. As for the reduced reference model and no-reference model, especially the latter, they are extremely suitable for the wireless and an IP-based video service where the original reference sequences is absent. Nevertheless, synchronization between the original and impaired data is still necessary for the reduced reference metric. Thus, we choose the no-reference model for this paper.

C. Pseudo Subjective Quality Assessment (PSQA)

PSQA is a quality assessment tool that is a hybrid between subjective and objective evaluation techniques. The idea is to do subjective tests for several distorted videos and use the results of this evaluation to teach a RNN (Random Neural Network) the relation between the parameters that cause the distortion and the perceived quality. The PSQA methodology is shown in Figure 1 and is explained in the following text.

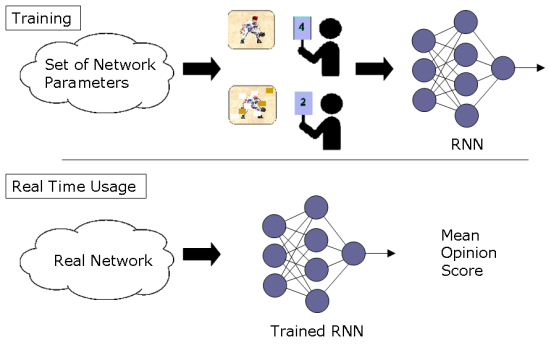

Figure 1: PSQA methodology

The procedure consists in first identifying the parameters that have an impact on QoE in the given context. These parameters may vary from one context to another and some examples of these parameters are: type of codec, packet loss rate, delay, jitter, etc. A video database is then generated by simulating the identified parameters after choosing a set of representative values for each of them, together with an interval for the parameter, according to the conditions under which we expect the system to work. Then, a uniformly sampled subset *S* of this video database is subjectively evaluated by a panel of humans. After statistical processing of the answers (designed for detecting and eliminating bad observers whose answers are not statistically coherent with the majority), each video sequence in *S* receives a QoE value (often, this is a Mean Opinion Score, or MOS). It results in *S* configurations of the parameters and a corresponding QoE score or MOS. Then some of the configurations are used for training the RNN and remaining ones are used for validation that, in turn, are not shown to RNN during training.

In order to validate the trained RNN, a comparison is done between the value given by the trained RNN at the point corresponding to each configuration in the validation set and to its actual MOS value; if they are close enough (having low mean square error), the training is validated. If the validation fails, a review of the chosen architecture and its configurations is needed.

III. QoE AFFECTING PARAMETERS FOR SVC

To Build PSQA for SVC, we need to identify clearly the parameters having an impact on the perceived quality of the video streams. These parameters could belong to the network QoS metrics (ex. bandwidth, packet loss rate, jitter, etc.) or/and video encoding parameters (ex. IDR (Instantaneous Decoder Refresh) frequency, Quantization Parameters, number of layers, etc.). After that, the relevant parameters will have to be simulated, which will result in distorted video sequences. These distorted video sequences will be used to train the PSQA tool with the help of a panel of human observers. The trained PSQA will then be used in real time to estimate the subjective video quality.

The next two subsections will describe the parameters that clearly affect the quality of the SVC video streams. It is important to note that these parameters are independent on the scalability type. Even though we considered quality scalability in our implementation, these parameters are still relevant for other types of scalability solutions (temporal, spatial, or a combination of those). Moreover, it should be noted that either

other parameters, such as frame rate, are assumed constant or are converted to other parameters, such as delay and jitter leading to a delayed packet are converted into loss.

A. IDR Frequency

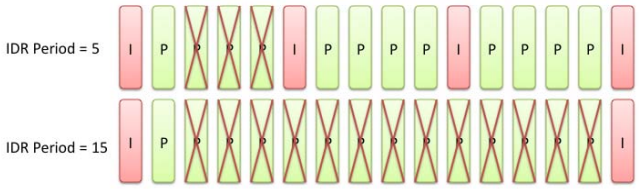

Figure 2: Relation between IDR period and losses

The frequency of IDR pictures (f_{IDR}) is an essential factor for the final quality. An increase in f_{IDR} means an increase in the number of IDR frames that in turn decreases the number of P-frames. Increased number of I frames are beneficial because an error will propagate only until the next I-frame arrives. Nevertheless, there are 2 problems with the increase of f_{IDR}: first of all, an encoded I-frame is larger in size as compared to a P-frame. Thus, the final size of the video will be increased. Furthermore, an increase in the number of I-frame involves more risk of error propagation. In fact, an error on an I-frame will be propagated on all the following P-frames (Figure 2).

B. NALU loss rate for each layer

The NALU (Network Abstraction Layer Unit) is the transport unit of video packet. A NALU can only transport information of one layer. The loss of a NALU affects only a single layer. However, it is important to mention that loosing a NALU belonging to the base layer has more impact on the video quality, than the loss of NALU belonging to other enhancing layers, as is explained in section IV. A loss of base layer NALU impacts the other layers as all the other layers in SVC use the base layer as reference and any error in this layer propagates to other layers, which reduce seriously the video quality.

Figure 3: NALU format

Usually, a NALU packet consists of one header (as AVC header), and a specific header extension (Figure 3). This extension has particular fields D, Q and T, which are used to identify the spatial, quality and temporal layers, respectively.

We denote the NALU loss rate of a layer n as L_n with only layer 0, or base layer, having a different name as L_{BL}.

C. Measuring SVC QoE parameters

In order to estimate QoE, the tool needs to be able to measure the identified parameters automatically and in real-time. Usually, f_{IDR} is a static value; it can be obtained when encoding the video stream. Another way of obtaining f_{IDR} is by tracking the appearance of the IDR frames in the video stream. An IDR frame in turn can be identified by parsing the NALU headers that contain the information about whether its payload corresponds to an IDR frame or not. As said before, higher value of f_{IDR} means higher resilience to the error propagation. Nevertheless, the size of the video will increase.

On the other hand, the NALU loss rate for each SVC layer is obtained by relying on the RTP layer. Combined with RTP simple packetisation mechanism (Single NAL Unit) [3], we propose to use a multi-session RTP connection for each layer. In other words, for each layer, an independent RTP session is established, where one NALU is conveyed in one RTP packet. Thus, we can obtain the NALU loss rate for each SVC layer at the RTP level, either at the decoder side, or by enabling the RTCP protocol for a remote monitoring.

IV. RESULTS

In order to train the RNN function for QoE Monitoring, 5 different video sequences were considered (Table 1). We encoded the different videos by using the JSVM encoder [11]. For the decoder side, we used the openSVC soft [12].

The resolution is 4CIF (704 x 576), frame rate is 30, and the values of QP are 34, 28, 24, respectively for layer 0 (Base Layer), 1 and 2 as shown in Table 2. Here, the QP parameter represents the SNR scalability.

Video	NAL number
CITY	300
CREW	300
HARBOUR	300
ICE	240
SOCCER	300

Table 1: Videos used for Training

Fixed Parameter	Value		
Resolution	704 x 576 (4CIF)		
Frame Rate	30		
Layer/QP	0/34	1/28	2/24

Table 2: Video parameters

As NALU losses have a serious impact on the final quality of the video, pertinent values of loss rate have to be defined. To obtain these values, the loss rate is simulated by varying values from 0 to 10% (quality is already very bad at 10% loss). For IDR frequency, 3 values are defined: 75, 150 and 300.

Parameter	Set of values
NALU loss rate for Base Layer (%)	0, 0.3, 0.5, 1, 3, 5, 10
NALU loss rate for Layer 1 (%)	0, 0.3, 0.5, 1, 3, 5, 10
NALU loss rate for Layer 2 (%)	0, 0.3, 0.5, 1, 3, 5, 10
IDR Frequency	75, 150, 300

Table 3: QoE affecting parameters values

A large number of videos were created using different combinations of the above parameters and their values given in Table 3. The videos were reduced to 500 using uniform sampling. In the next step, using manual evaluation, around 100 videos were selected for the subjective test session such that 25 videos corresponded to a MOS (Mean Opinion Score) score between 1 and 2, 25 videos between 2 and 3 and so on. A MOS scale shown in Table 4 was used for all quality evaluations.

MOS	Quality	Level of Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 4: Mean Opinion Score (MOS)

It is necessary to reduce the number of videos because 100 videos, of 10 seconds duration each combined with the time required to give quality scores, already correspond to around 30 minutes. These 100 videos were evaluated by 5 humans using DSIS (Double Stimulus Impairment Scale) methodology

[5]. The obtained scores were then pre-processed, as in [4], to check for inconsistent scorers. Then the data was used to train the RNN. Moreover, after a first training, the significant 5% outliers were removed because we were interested only in a 95%-level accuracy. Three RNN layers were used. The first layer has 4 neurons corresponding to the 4 inputs, the hidden layer has n_h neurons and final layer has 1 neuron for the single needed output (quality). The number n_h was chosen after several training iterations. The data set was divided into 2 sets with 80% points as training data and 20% as validation data. Only training data was used for training and validation set was not shown to RNN while training. The untrained RNN will obviously give high mean squared error (MSE) for both sets, whereas an over trained RNN, also undesirable, will give high MSE on the validation set. Based on these requirements, after several iterations, the chosen value of n_h was $n_h = 5$.

From (i:j)	To (i:j)	W^+	W^-
0:0	1:0	0.096701	5.872150
0:0	1:1	0.024633	1.741510
0:0	1:2	0.002058	4.198560
0:0	1:3	0.000367	3.388390
0:0	1:4	0.000365	3.329620
0:1	1:0	0.367032	0.531753
0:1	1:1	2.036930	0.539114
0:1	1:2	0.002047	0.669375
0:1	1:3	0.000649	0.373859
0:1	1:4	0.000648	0.363866
0:2	1:0	0.135111	1.648410
0:2	1:1	0.632614	0.304190
0:2	1:2	0.053647	1.486570
0:2	1:3	0.000026	0.854922
0:2	1:4	0.000026	0.888809
0:3	1:0	0.325495	1.526620
0:3	1:1	0.747832	0.452583
0:3	1:2	0.083876	0.831486
0:3	1:3	0.002404	1.385160
0:3	1:4	0.002398	1.345570
1:0	2:0	0.027058	1.091710
1:1	2:0	2.353470	2.157500
1:2	2:0	0.050469	0.218778
1:3	2:0	0.000974	0.000119
1:4	2:0	0.000967	0.000120

Table 5: Weights for the RNN function

In Table 5 we provide the details of the trained RNN and details of training are provided later in this section. Some very compressed details are given here to understand it (please see [4][13], for more details on RNN.): the output of the network is a fraction where the numerator is the sum in h of the "state" ∂_h of hidden neuron h, weighted by $W^+(h, o)$, the positive weight from h to the single output neuron o; the denominator is the sum of the rate of neuron o and the sum of the ∂_h weighted by the $W^-(h, o)$. State ∂_h is in turn equal to a similar fraction using layers 0 and 1.

In Table 5, there are three layers indexed 0, 1 and 2 with 4, 5 and 1 neurons respectively. The weights connecting the neurons in different layers are denoted as W^+ and W^-. The 4 inputs to the RNN are f_{IDR}, L_{BL}, L_1 and L_2 normalised as follows: $f_{IDR}/300$, $L_{BL}/10$ $L_1/10$ and $L_2/10$; the values are set to 1.0 if more than 1.0. The service rates at the input layer are 1.0, for hidden layer are computed from weights [4] and for the single output neuron, it is 0.01. In the table, (i, j) means the j^{th} neuron in layer i. The RNN is trained such that it gives *(1 – normalized MOS)* as output. Let us denote the RNN output as q_{inv} then the predicted MOS = $5(1 - q_{inv})$ on a MOS scale from 1 to 5.

The final results obtained are shown in Figures 4, 5 and 6. Figure 4 shows that the video quality is very sensitive to losses in the base layer and the value of MOS quickly decreases to 1.0 with an increase in NALU loss rate of base layer. Even a loss rate of around 1%, of base layer NALUs, degrades the value of MOS to be lower than 3 out of 5 that, in turn, corresponds to the impairment in the video quality as slightly annoying (Table 4). This fast degradation with NALU loss rate of base layer is because all the other layers in SVC use the base layer as reference and any error in this layer propagates to other layers. In addition, it can be seen that MOS decreases slightly with increasing value of f_{IDR}, IDR frame frequency. This is because a high value of f_{IDR} means that a particular error has higher chances of propagating to a large number of frames until an IDR frame arrives and refreshes the decoder. For example a value of f_{IDR} = 300 means that an error occurring in 1st frame can propagate up to maximum 299 frames that in turn would mean around 10 seconds of video that has 30 frames per second.

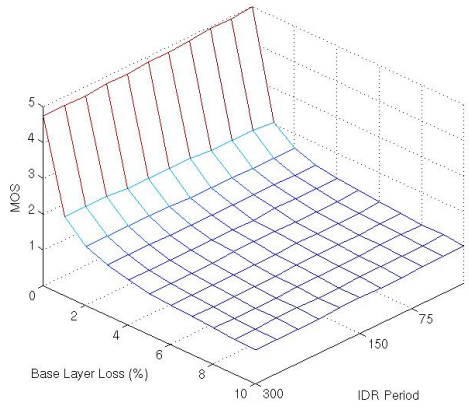

Figure 4: PSQA scores vs. L_{BL} and f_{IDR} with L_1 = 0% and L_2= 0%

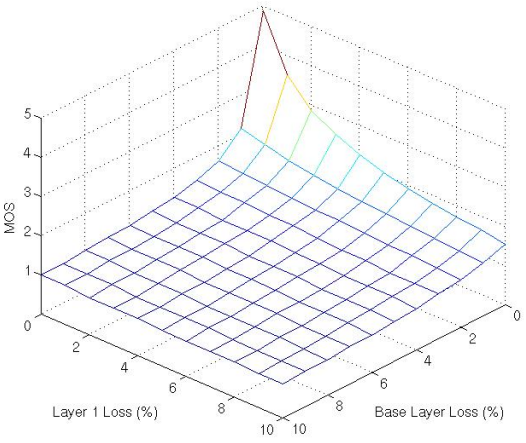

Figure 5: PSQA scores vs. L_{BL} and L_1 with f_{IDR}=300 and L_2= 0%.

Figure 5 shows the value of MOS predicted by PSQA with varying L_{BL} and L_1. It can be seen that the quality is more sensitive to an increase in NALU loss rate of base layer as compared to that of layer 1. Whereas, it can be seen in Figure 6

that QoE is, relatively, only slightly more sensitive to an increase in L_1 as compared to L_2. This is due to the fact that base layer in SVC is the most sensitive to the losses as compared to other layers. If there is no loss in base layer, and there is a loss in other layers, then decoder has a better chance of error concealment as compared to the case when the data from the base layer itself is lost.

Figure 7 shows the scatter plot with estimated MOS vs real MOS obtained from the subjective tests. The diagonally plotted line corresponds to the case when the estimated value of MOS is equal to the real value of MOS. Thus, points lying close to the line would indicate the precision of the estimation tool. As seen in Figure 7, the scatter plot shows a good accuracy of the estimation (also reflected by the overall mean square error of about 0.0071 on the MOS scale from 0 to 1.0).

Our QoE estimation tool is interesting because it can be used to provide real-time QoE feedback that in turn can be used for network dimensioning, SVC layer adaptation and QoE optimization. For example when the QoE value becomes low then the enhancement layers can be discarded at the sender, or at some other control point. In addition some bandwidth, thus freed, can be used to adaptively apply more error correction to the base layer, in order to protect it from network losses.

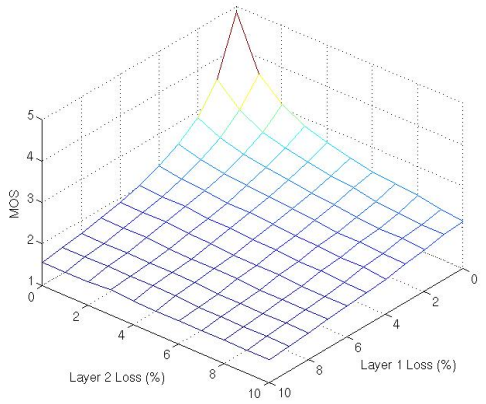

Figure 6: PSQA scores vs. L_1 and L_2 with f_{IDR} =300 and L_{BL} = 0%.

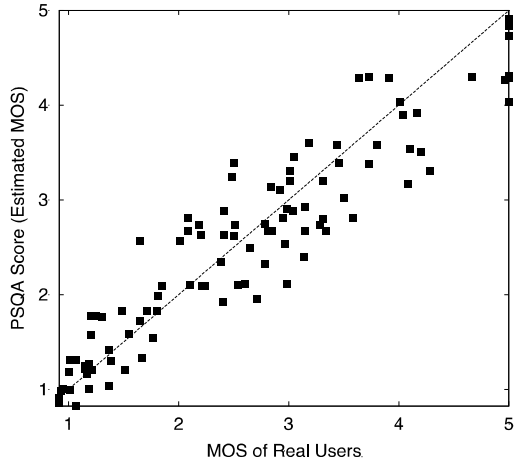

Figure 7: PSQA scores vs. scores given by real users

V. CONCLUSION

In this paper, we designed an automatic QoE measurement tool for SVC. After describing the parameters that affect the quality of SVC streams and their relationship with QoE, we presented how to capture this non-linear relation with PSQA. The results showed that our QoE estimation is very accurate and the obtained scores are very close to those given by real users. Despite the fact that SVC scalability used in our implementation is based on SNR, the parameters affecting the QoE are the same for other scalability methods. Indeed, from the obtained results, it is clear that the most affecting parameter is the loss of NALU belonging to the base layer.

To the best of our knowledge, this automatic QoE tool measurement is the only one addressing the SVC video encoding technique. This tool is interesting not only for measuring the quality of the transmitted flow, but also to optimize the network resources and for network dimensioning. Adaptation points like MANE can rely on QoE feedbacks to adapt the SVC layers transmitted to end-users, or to increase the network protection for the base layer NALUs (by increasing the FEC redundancy). Investigating issues related to MANE actions and QoE feedback will constitute excellent directions for future works. Furthermore, we intend to improve the accuracy of our QoE estimation tool by adding more parameters to it.

VI. REFERENCES

[1] T. Wiegand, G Sullivan, J. Reichel and M. Wien, joint Draft 9 of SVC Amendment, Joint Video Team, JVT-V201, Marrakech, Morocco, Jan 2007.

[2] ITU-T SG12, "Definition of Quality of Experience", COM12 – LS 62 – E, TD 109rev2 (PLEN/12), Geneva, Switzerland, 16-25 Jan 2007.

[3] S. Wenger, Y. Wang and T. Schierl, "RTP Payload Format for SVCVideo", IETF Internet Draft, draft-ietf-avt-rtp-svc-20.txt, Dec 2009.

[4] S. Mohamed and G. Rubino, "A study of Real-Time Packet Video Quality using Random Neural Networks", IEEE Transaction on Circuits and Systems for Video Technology, vol. 12, 12, pp 1071-1083, Dec 2002.

[5] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures", June 2002.

[6] C.J. van de B. Lambrecht and O. Verscheuren "Perceptual Quality measure using a spatio-temporal model of the human visual system", in Proc. SPIE, vol. 2668, pp. 45°-461, 1996.

[7] Yubing Wang, "Survey of objective video quality measurements", Technical Report WPICS-TR-06-02, EBU Technical Review, Feb. 2006.

[8] ITU-T Telecommunication G. 1070, "Opinion model for video-telephony applications". Apr 2007.

[9] K. Yamaghachi and T. Hayashi, "Parametric Packet-Layer Model for monitoring Video Quality of IPTV services", In Proc. Of ICC 2008.

[10] F. You, W. Zhang, and J. Xiao, "Packet Loss Pattern and Parametric Video Quality Model for IPTV," ICIS'09. In Proc. Of 2009 Eight IEEE/ACIS International Conference on Computer Information Science, 2009, Washington, DC, USA.

[11] JSVM, http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm

[12] OpenSVC, http://sourceforge.net/projects/opensvcdecoder/

[13] E. Gelenbe, "Random neural networks with negative and positive signals and product from solution", Neural Computer, vol. 1, pp. 502-511, 1989.