Long-term Quality Evaluation in OTT Video

Shahid Mahmood Satti, Roland Bitto, Michael Keyhl, Matthias Obermann, Christian Schmidmer
OPTICOM GmbH, Erlangen, Germany
info@opticom.de

***Abstract*— Quality variations due to network bandwidth fluctuations is a common phenomenon in today's HTTP based over-the-top (OTT) video streaming services where video sessions often last several minutes. In addition to quality variations, OTT video playouts encounter buffer under-runs resulting in client-side video interruptions (a.k.a. stallings) which highly impact viewer's quality of experience (QoE). OTT service providers would like to know the overall quality score of a given video stream to optimize their service. Network provider's target is to maximize the overall video quality score for their network. In any case an accurate measurement of video stream quality is a key requirement for different stake holders in the OTT business. This paper investigates the design of a parametric model for estimating long-term (60 sec or longer) perceptual quality from short-term (10 sec) video quality estimates in H.264 based adaptive video streaming. Based on an advanced subjective test framework, the proposed model linearly combines the perceptual impact of three types of distortions (initial-loading, stalling and coding/quality-switching) observed in OTT video in an overall quality score on a 5-point opinion scale. The proposed model yields high accuracy of quality prediction when evaluated for up-to 3 minutes long video sequences.**

I. INTRODUCTION

The recent advent of over-the-top (OTT) video services, e.g., Netflix, YouTube, Amazon Video, etc., has triggered a paradigm shift in how video content will be delivered over the Internet in future. The flexibility and convenience of OTT video is making such services more popular among both providers and consumers in comparison to the IPTV alternative. Although OTT videos are distributed over dedicated content delivery network (CDN) infrastructures, OTT content is delivered over the more congestion prone open Internet (unmanaged network). This is in contrast to the IPTV services which run on dedicated managed networks. Thus, compared to the IPTV case, controlling the perceived video quality or quality of experience (QoE) is a major challenge for OTT service providers.

OTT services generally rely on streaming protocols such as HTTP adaptive bitrate streaming, where a video is coded in a number of quality layers, each of which is split into fixed length chunks (typically 2 to 10 seconds long) and stored on one or more HTTP servers or cached within the CDN for a quick access. Depending on the local client characteristics (e.g., screen resolution, processing power, supported video formats, etc.), the client determines a subset from the available set of quality layers which it can play. Then, depending on the network characteristics (e.g., maximum and average available bandwidth, network congestion), the client downloads appropriate chunks and smoothly switches between different quality layers at the chunk boundaries in order to provide an uninterrupted user-experience under varying network bandwidth conditions. Commonly known distortions in OTT are *initial-loading*: delay between a user pressing the *play* button to the actual start of the video, *stallings*: video interruptions due to client side buffer under-runs and *coding/quality-switching*: artefacts due to compression and due to downloading chunks from different quality layers.

To date, research and standardization efforts for the subjective and objective evaluations of video quality are generally limited to testing short video clips of 10-20 sec [4], [5]. However, the typical video session lengths observed in practice for OTT video are often well over 10-20 sec, proving these standardization efforts are limited in scope for the OTT case. Long-term quality measures can be constructed based on short-term scores [1]-[3] if one knows how the human visual system integrates instantaneous quality variations into a long-term quality score? The work presented here takes a parametric approach to address this question. For any objective model subjective mean opinion score (MOS) serves as the ground truth. This means that any kind of short-to-long term quality integration function should be derived from the subjective scores first before any kind of mapping to objective scores. Taking this approach, we have designed subjective databases of long (60 sec or more) videos (from here onwards referred to as long-term (LT) database) simulating the commonly encountered distortions in adaptive streaming services. These videos are then split into 10 sec non-overlapping pieces to create short length database – from here onwards referred to as short-term (ST) database. Note that it is very important to have LT database cover the complete range of distortions in OTT video streaming in order to draw any meaningful conclusions. To achieve this, the YouTube streaming service is taken as the benchmark to identify quality level bitrates, typical adaptation and stalling patterns, etc. Based on the obtained subjective scores for LT/ST databases, one short-to-long quality prediction function is determined for each of the three types of distortions. Later, these individual functions are linearly combined to create an overall quality model. For validation, new subjective databases of short and long video sequences are created. When short-term subjective MOS are used as input, the proposed model achieves a Pearson correlation coefficient (PCC) value of 0.95 and root mean squared error (RMSE) of 0.30 to predict the quality of 3 minute video. If short-term objective scores are used, the proposed model yields PCC of 0.94 and RMSE of 0.35. For computing short-term objective scores OPTICOM's full-reference PEVQ [8]-[11] was used.

The remainder of this paper is organized as follows: The design of LT/ST databases is described in Section II.

978-1-5090-5966-9/16/$31.00 ©2016 IEEE

Parametric model derivations are carried out in Section III. Model validation is presented in Section IV and Sections V draws the conclusions of this work.

II. Design of LT/ST Databases

For the LT test 13 one minute HD quality sources are used. Source videos are coded in a number of quality levels using H.264 (x264/ffmpeg implementation is used), the description of quality levels is given in Table 1. In the LT database we used 50 processed video sequences (PVSs) distributed in five classes. Details of each class are given in Table 2.

H.264 Codec Parameters	Values
Frame-rate	23.97 – 30 frames per second (FPS)
Preset	Medium
Scene-cut	No
Profile	High10
Coding type	Constant bitrate 2-pass coding
GOP size, type, segment length	1 sec, IBBP, 5 sec
Resolution/Bitrate	A) 1920x1080: 15 Mbps B) 1920x1080: 3 – 5 Mbps C) 1280x720: 1 – 2 Mbps D) 854x480: 600 – 900 Kbps E) 640x360: 200 – 400 Kbps F) 426x240: 100 – 200 Kbps Note: the above bitrate ranges B-F are typically observed for YouTube.

Table 1: Coding parameters for the LT database.

Class	Details
Reference (R)	5 PVSs: 5 source videos (SRCs) at quality A (see Table 1)
Initial-loading (I)	4 PVSs: 2, 6, 10, 15 sec initial-loading (IL) followed by quality A. 2 SRCs from class R are used, 1st for simulating 2 and 10 sec IL, 2nd for simulating 6 and 15 sec IL PVSs.
Stalling (S)	8 PVSs: Quality A throughout, for each case a unique SRC (different from the ones in class R) is used. PVS1: 2 stallings of 2 sec each, PVS2: 4 stallings of 1 sec each, PVS3: 1 stalling of 8 sec, PVS4: 2 stallings of 4 sec each, PVS5: 3 stallings 2, 3, 3 sec, PVS6: 8 stallings of 1 sec each, PVS7: 1 stalling of 12 sec, PVS8: 3 stallings of 4 sec each. Note: To avoid any primacy/recency bias - see Section IV, no stalling is simulated in the first or last 10 sec of the PVS.
Quality Change (Q)	24 PVSs: 24 adaptivity patterns varying between quality levels B-F are selected. Three subsets of 8 patterns each are mapped to three SRCs of class R to create 24 PVSs.
Mix (M)	9 PVSs: Out of the 24 class Q adaptivity patterns 9 are selected, combined with IL and stalling lengths values different from ones listed in class I and S, 9 PVSs were created. SRCs for class M are same as used for class Q.

Table 2: Test conditions for the LT database.

The ST database is created by splitting the PVSs from the LT database into 10 sec pieces of actual video, i.e., lengths for some pieces may be longer than 10 sec if the long PVS contains initial-loading or stalling(s).

For both LT and ST databases, A five-point categorical quality scale with quality ratings: 5=Excellent, 4=Good, 3=Fair, 2=Poor and 1=Bad, was used in Absolute Category Rating (ACR) test method of ITU-T P.910 [4]. For viewing, a calibrated Panasonic 42 inch 1080p TV display was used. A viewing distance of 3H, where H denotes the vertical size dimension of the display, is maintained for each subject during the subjective test. PVSs in both LT and ST databases were randomized for each subject. A total of 26 subjects, after pre-screening for visual acuity and color blindness, participated in each test. Subjective test started with a training phase in order to make subjects familiar with the rating procedure and the kind of qualities they should expect in the actual test. This is done in a careful manner in order to avoid biasing subject's judgement about a certain type of distortion. The videos used for the training phase were not included in the actual test. Subjects were asked to provide a video-only quality score after watching each PVS. In case PCC value for the scores given by a subject with respect to the average of the scores given by all the subjects is below 0.7, the scores of this subject were not considered in the final analysis. This procedure for discarding subjective scores is a common practice in standardization bodies, e.g., ITU-T SG12. For LT and ST databases scores from 24 valid subjects out of 25 and 26 total subjects, respectively, are taken to compute the subjective MOS for each PVS.

III. Parametric Model Derivation

Based on the subjective data, we firstly study the relationship of ST and LT MOS for each of the different types of distortions. Let ∇_1 and ∇_2 denote the 1st and 2nd order mappings, respectively, i.e.,

$$\nabla_1(x) = d_1 x + d_0, \quad \nabla_2(x) = e_2 x^2 + e_1 x + e_0$$

where, constants d_n, e_n, $n = 1,2,3$, belong to the set of real numbers. MOS_{RL}, MOS_{DL}, MOS_{RS} and MOS_{DS} denote the LT reference, LT degraded, ST reference and ST degraded MOS, respectively. K denotes the number of short pieces contained in a long duration PVS, MOS_{DS}^{k} denotes the degraded MOS of the k^{th} short piece. RMSE for a database or a subset of database is computed after ∇_1 mapping (slope and offset corrections) of objective scores to subjective MOS.

A. Initial-loading (IL)

Figure 1 plots the drop in MOS as a function of IL time for 4 PVSs in class I. It can be assumed that this relationship is linear. The absolute value of the MOS drop is smaller for longer videos. This is understandable as the subjects only score a PVS after watching it completely and they are more forgiving for IL time after having watched a longer video of high quality.

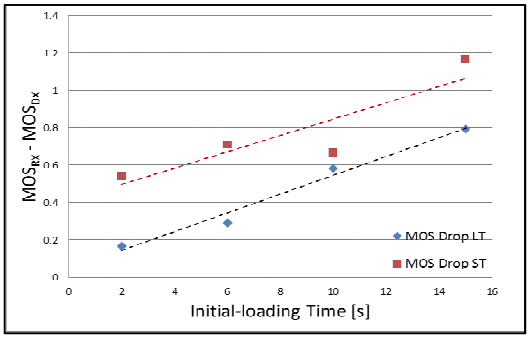


Figure 1: Drop in MOS in a high quality playout due to IL, X ϵ {S, L}.

The parametric relationship between MOS drop and IL time (T_{IB}) is observed to be as follows:

$$(MOS_{RL} - MOS_{DL}) = \nabla_1(T_{IL}), (MOS_{RS} - MOS_{DS}) = \nabla_1(T_{IL})$$

978-1-5090-5966-9/16/$31.00 ©2016 IEEE

$$MOS^{*}_{DL} = \nabla_1\left(\alpha T_{IL} + \frac{1}{K}\sum_{k=1}^{K} MOS^{k}_{DS}\right) \quad (1)$$

MOS^{*}_{DL} in the above equation and from here onwards denotes the predicted value of MOS_{DL}. We experimentally observed that $\alpha = -0.05$ yields best mapping in this case.

B. Stalling/Buffering

To our observation MOS_{DL} inverse-linearly relates with the product of number of stallings (N_B) and total stalling time (T_B in sec), see Figure 2 which plots MOS_{DL} for 8 class B PVSs with respect to this product. To derive the relationship between MOS_{DL} and MOS_{DS} this knowledge is used.

To our observation, MOS_{DL} highly correlates with the average MOS_{DS} minus some factor of the product $N_B T_B$, i.e.,

$$MOS^{*}_{DL} = \nabla_1\left(\beta N_B T_B + \frac{1}{K}\sum_{k=1}^{K} MOS^{k}_{DS}\right) \quad (2)$$

Figure 3 plots Equation 3 with respect to MOS_{DL}. $\beta = -0.0308$ yields the best mapping in this case.

C. Coding/Quality-Switching

For coding/quality-switching cases, MOS_{DL} is correlated with the average of MOS_{DS} – see Figure 4. This is in line with findings of previous research studies [6].

An additional observation is that for two PVSs with same average bitrate, one having quality-switching and the other having a constant quality throughout, subjects tend to score latter slightly higher than the former; this means that switching in quality by definition is not preferred from perceptual point of view if the underlying network bandwidth is constant. This is precisely the reason why we required a non-linear function ∇_2 to map average of MOS_{DS} to MOS_{DL} – see Figure 4, i.e.,

$$MOS^{*}_{DL} = \nabla_2\left(\frac{1}{K}\sum_{k=1}^{K} MOS^{k}_{DS}\right) \quad (3)$$

D. Mix Distortions

Now that we have studied the perceptual behavior of individual distortions and derived an ST to LT parametric mapping function for each, we can evaluate a more realistic setting, i.e., when all three types of distortions are present at once in a video. Assuming that the perceptual characterization of individual distortions derived in previous sections will stay stable in the combined case as well, we take a linear combination of all three quality functions, i.e.,

$$MOS^{*}_{DL} = \nabla_1\left(\alpha T_{IL} + \beta N_B T_B + \gamma\frac{1}{K}\sum_{k=1}^{K} MOS^{k}_{DS}\right) \quad (4)$$

and see how this function works for class M PVSs. Figure 5 plots the function given in the brackets in Equation 4 with respect to MOS_{DL} for the nine PVSs in class M. The mapping function seems to predict MOS_{DL} from MOS^{k}_{DS} quite accurately. There may be a certain bias due to how the database was produced and ∇_1 mapping should be allowed to correct the obtained objective scores w-r-t the subjective data. Note that the initially found best fitting constants $\alpha = -0.05, \beta = -0.0308, \gamma = 1$ still do a reasonably good job in the mixed distortion case. This seems to indicate that the derived constants are fairly general, although evaluation on more subjective databases may be needed to check the universal applicability of the derived constants. This suggests that the

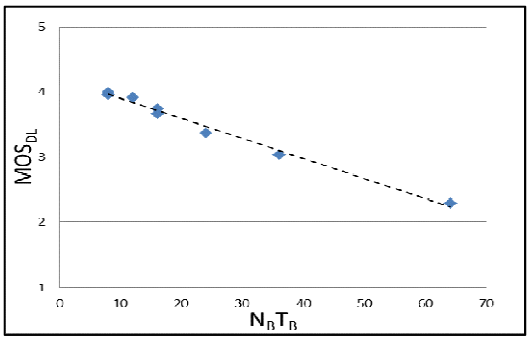


Figure 2: Linear relationship of MOS_{DL} with product $N_B T_B$.

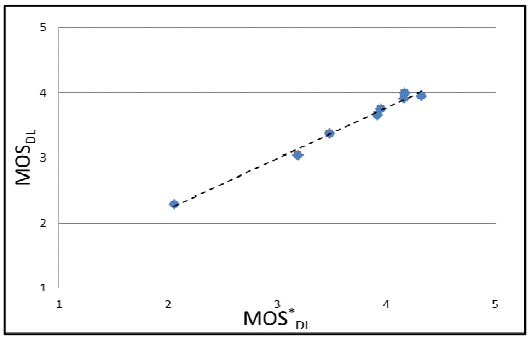


Figure 3: Parametric relationship between MOS_{DL} and MOS_{DS} for stalling distortions, $\beta = -0.0308$.

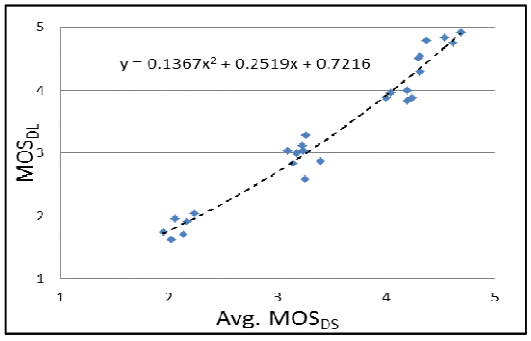


Figure 4: Parametric relationship between MOS_{DL} and MOS_{DS} for coding/quality-switching distortions.

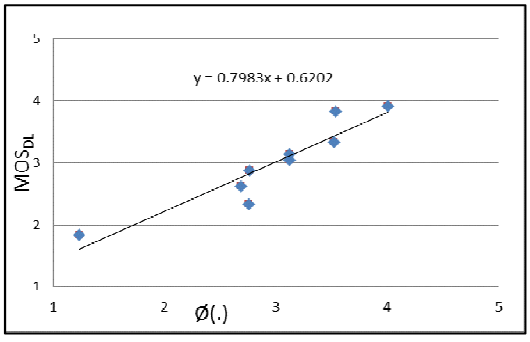


Figure 5: Parametric relationship between MOS_{DL} and MOS_{DS} for the mix distortion case, $\alpha = -0.05, \beta = -0.0308, \gamma = 1$. PCC=0.93, RMSE=0.23.

derived parametric model is stable and can be applied to unknown conditions. For the training databases we tried to avoid the recency aspect, more about recency will be discussed in the next section.

IV. Model Validation

In this section we will validate the previously derived parametric model for new validation databases. We created three validations databases: LTV1 (60 one minute long videos with mix distortions), LTV2 (22 three minutes long videos with mix distortions) and STV2 (10 sec pieces of LTV2). For validation databases, coding parameters (except the bitrates) are the same as for training LT/ST databases and are given in Table 1. Unlike for the training databases, video sources are not repeated in validation databases. Bitrates are different from training databases but are chosen from the bitrate ranges specified in Table 1. Based on YouTube as the benchmark service, we determined commonly observed ranges of initial-loading, stalling times and quality-switching patterns, which were employed to create the three validation databases. Test setup for validation databases is exactly the same as for LT/ST. Subjective MOS scores were computed based on the average of 24 valid subjects.

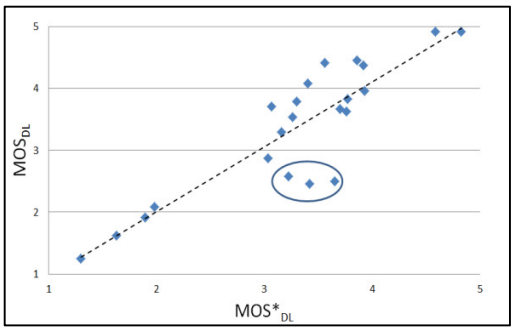
A. Subjective Score based Validation

For subjective scores based validation, LTV2 and STV2 are used because we know the subjective MOS scores for both long PVSs and short pieces of these long PVSs. Figure 6(a) plot the output of the model against the MOS_{DL} of LTV2 using the model constants derived in the previous section. There are certain cases (shown inside the circle) which the model fails to predict accurately. Looking closely at these cases reveal that these are PVSs which have long parts of poor or bad quality towards the end of the video. Although earlier parts of these PVSs may have good or excellent quality, subjects tend to rate these PVSs slightly lower for being ending at a *bad note*. This phenomenon in the field of psychology is known as primacy/recency effect [7] which in a video viewing context would mean that subjects would remember the beginning (primacy) and end (recency) parts of the video and hence will give more importance to the quality of these parts in the overall quality score. We did not observe any contribution of this effect in our one minute LT training database; this was mainly due to the fact that there we deliberately tried to avoid this effect in the tested conditions. Although the same was the intention for LTV2, we were not completely successful and a few PVSs seem to exhibit the recency effect. However, we did not observe any primacy effect in our analysis. We conclude that for longer duration videos, only the ending (and not the beginning) parts contribute higher than the average to the overall quality score. To take into account the observed recency effect we used a generalized recency function,

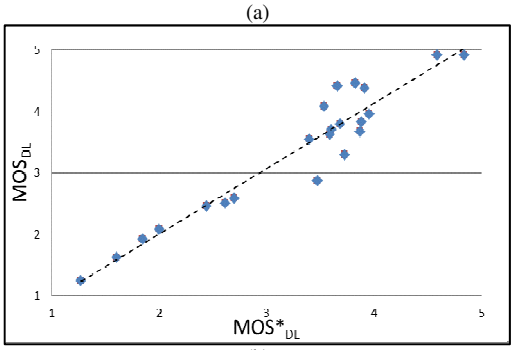
$$R = \{\dots, 1, 1, 1, 2, 3, 4\}, \quad r_k = \frac{R_k}{\sum_{k=1}^{K} R_k} \qquad (5)$$

The parametric model derived in the last section is slightly modified, to taken into account the recency effect, as follows:

$$MOS^{*}_{DL} = \nabla_1 \left(\alpha T_{IL} + \beta N_B T_B + \gamma \frac{1}{K} \sum_{k=1}^{K} r_k MOS^{k}_{DS} \right) \quad (6)$$

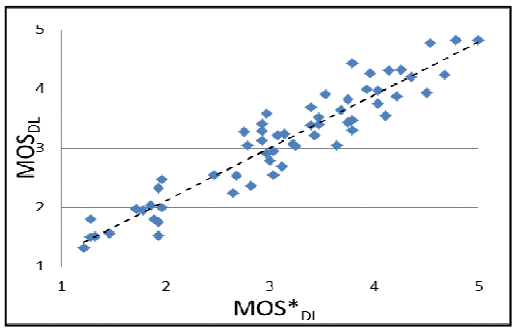


(a)

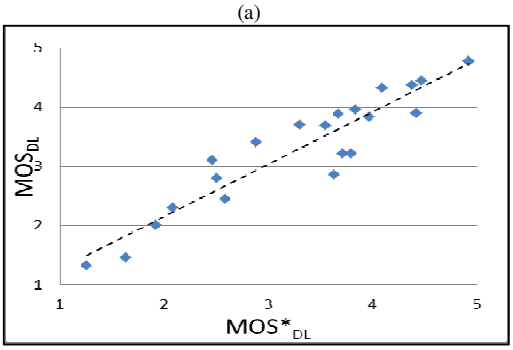


(b)

Figure 6: Subjective score based model validation for LTV2, $\alpha = -0.05, \beta = -0.0308, \gamma = 1$. PCC=0.84, RMSE=0.48. (a) without recency fix, (b): with recency fix, PCC=0.95. RMSE=0.30.



(a)



(b)

Figure 7: Objective score based model validation. (a) LTV1, PCC= 0.94, RMSE=0.31, (b) LTV2, PCC=0.94, RMSE=0.35.

Figure 6(b) plots the predicted score with respect to MOS_{DL} after accounting for the recency effect. It is evident that the overall prediction accuracy is improved significantly both in terms of PCC and RMSE. Additionally, the prediction efficiency for other cases for which recency did not play a role did not deteriorate.

978-1-5090-5966-9/16/$31.00 ©2016 IEEE

B. Objective Score based Validation

For objective score based validation, objective prediction of MOS_{DS} (i.e., MOS^{*}_{DS}) is used as the input to the parametric model. For MOS^{*}_{DS} computation we employed OPTICOM's Perceptual Evaluation of Video Quality (PEVQ) algorithm. PEVQ is a state-of-the-art method for quality estimation for SD/VGA/WVGA/HD resolutions and was standardized by ITU-T – see [8]-[11]. PEVQ produces a single quality score for each 10 sec piece of the video. No initial-loading or stalling instances are fed to PEVQ, meaning it only produces coding-only quality scores considering the quality-switching aspects alone. In order to take into account the instantaneous effect of initial-loading and stalling, objective scores of PEVQ for the first 10 sec piece and subsequent pieces which contain stalling events need to be corrected. We employ the following exponential decay function to correct the PEVQ objective scores for these distortions:

$$MOS^{k**}_{DS} = \max\left(1, MOS^{k*}_{DS} e^{(\rho T_{IB} + q N_S T_S)}\right) \quad (7)$$

where $k \in \{1,2,3 \ldots K\}$, ρ and q denote two database specific constants. This decay function was experimentally observed to give best results for the overall MOS prediction.

Figure 7 plots predicted MOS^{*}_{DL} with respect to MOS_{DL} for LTV1 and LTV2. In here, we take into account the recency correction of Equations 5 and 6. The prediction efficiency is quite good, though not as high as it was in the subjective score based prediction. This is understandable as an objective metric cannot predict MOS_{DS} with 100% accuracy. Also in subjective score based validation no correction of scores for IL/stalling contained pieces is required, as they were scored by subjects. Despite these two aspects contributing to the prediction error, the performance of the parametric model is only marginally worse compared to the subjective score based validation.

We would like to emphasize that, although the parametric model is only tested with PEVQ, it can principally be coupled with other short-term objective quality metrics as well. How the accuracy of long-term prediction relates to the accuracy of objective metric remains to be seen though. In general, full-reference metrics are recommended over bitstream-based no-reference models as they are widely independent of the underlying coding architectures, GOP structures, profiles, etc.

In general, OTT video streaming sessions can be longer than the durations tested in this work. However, creating subjective databases of longer than 3 minutes videos would only test a few conditions in 60-75 minutes, which is considered to be a normal testing time without inducing viewing fatigue in subjects. It is foreseen that a more comprehensive handling of the recency effect may be required for longer duration videos. Model validation for over 3 minute videos is left for future consideration. Due to the linear architecture of the proposed model, we believe that it can also be adapted to the mobile case, i.e., evaluating OTT video quality on mobile terminals such as smartphones or tablets. In general, compared to the fixed (PC monitor or TV) case, for the mobile case subjects may combine the effect of individual distortions differently or may even have a different expectation of video quality. These two effects can easily be accounted for in the proposed model by using a different combination of α, β, γ or using a multiplicative monotonic expectation function.

V. Conclusions

A novel parametric model for short- to long-term quality is proposed to measure QoE for H.264 based adaptive streaming services, such as YouTube, VIMEO, etc. The proposed model takes into account commonly observed distortions in OTT video streaming. The perceptual impact of different types of distortions is linearly combined to yield an overall quality score. The proposed model is validated for unknown test conditions on both subjective as well as objective short-term scores. For both cases, the proposed framework yields high accuracy of prediction of long-term subjective MOS. We observed that recency plays an important role for videos longer than 1 minute; a generalized recency function was proposed to mitigate the insufficiency of our model for recency affected cases. Future extensions of this work for longer duration sequences and for video quality evaluation on smartphones/tablets are currently under consideration.

Acknowledgement

We acknowledge Margaret Pinson from Institute for Telecommunication Sciences NTIA USA, for providing the subjective rating software.

References

[1] J. Xue, D.-Q. Zhang, H. Yu, and C. W. Chen, "Assessing Quality of Experience for Adaptive HTTP Video Streaming," Proceedings of ICME, 2014.

[2] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, "To pool or not to pool: A comparison of temporal pooling methods for HTTP adaptive video streaming," Proceedings of QoMEX, 2013.

[3] J. De Vriendt, D. De Vleeschauwer, and D. Robinson, "Model for estimating QoE of Video delivered using HTTP adaptive streaming," Proceedings of IFIP, 2013.

[4] ITU-T Rec. 910, "Subjective Video Quality Assessment Methods for Multimedia Applications", International Telecommunication Union, Geneva, 02/2008.

[5] ITU-T Rec. 911, "Subjective Audiovisual Quality Assessment Methods for Multimedia Applications", International Telecommunication Union, Geneva, 12/1998.

[6] K. Seshadrinathan, A. Bovik, "Temporal Hysterisis Model of Time Varying Subjective Video Quality", Processings of Int. Conf. on Accoustics Signal Processing (ICASSP), vol. 2011, pp. 1153-1156, 2011.

[7] M. Bennet, "Serial Position Effect of Free Recall". Journal of Experimental Psychology, vol. 64, no. 2, pp. 482–488. 1962.

[8] M. Keyhl, M. Obermann, S. M. Satti, G. Rousell, "Perceptual Quality Evaluation of OTT Streaming Video TV Services", Proceedings of the 68th NAB Broadcast Engineering Conference, vol. 2014, pp. 252-260, 2014.

[9] ITU-T Rec. J.144, "Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference", International Telecommunication Union, Geneva, 03/2004.

[10] ITU-T Rec. J.247, "Objective Perceptual Multimedia Video Quality Measurement in the Presence of a Full Reference", International Telecommunication Union, Geneva, 08/2008.

[11] ITU-T Rec. J.343.5/6, "Hybrid Perceptual/Bitstream Models for Objective Video Quality Measurements", International Telecommunication Union, Geneva, 11/2014.

978-1-5090-5966-9/16/$31.00 ©2016 IEEE