A Layered Model for Quality Estimation of HTTP Video from QoS Measurements

Toni Mäki, Martín Varela
Communication Systems
VTT Technical Research Centre of Finland
Oulu, Finland
Email: first.last@vtt.fi

Doreid Ammar
Institutt for Telematikk
NTNU
Trondheim, Norway
Email: doreid.ammar@item.ntnu.no

***Abstract*—HTTP video is quickly becoming a dominating type of traffic on the Internet, with popular services such as YouTube and Netflix being used by hundreds of millions of users daily, and showing ever-growing usage numbers. Understanding Quality of Experience (QoE) for these services is an important topic, and one that has been addressed in the literature. However, the available works focus on the impact of application-level events (e.g. stalls) on the perceived quality, but not on the underlying cause, i.e., network-level impairments, as the relation between Quality of Service (QoS) and QoE is significantly more complex than it was in the case of RTP/UDP based video, due to HTTP video being streamed over TCP. In this paper we present a first step in the direction of solving this QoS-to-QoE mapping for HTTP video, by providing a (parametric) layered model approach for network-side QoE monitoring.**

***Keywords*-QoS, QoE, HTTP Video, DASH**

I. Introduction

HTTP video is currently one of the most popular service types, and one of the largest contributors to overall traffic on the Internet. Some forecasts are predicting that by 2019, video will account for up to 80% of consumer traffic [1]. The businesses around online video are also massive, with YouTube alone having over 1Bn users[1], and several hundred million hours of video streamed per day, with yearly growth of ∼50%. Similarly, Netflix has over 60M subscribers worldwide (40M in the US alone), and accounts for a large portion of last-mile traffic in the US.

In order to keep all those users satisfied, the performance of the video streaming services needs to be good enough to achieve acceptable quality levels, lest users abandon a session, or worse still, abandon a service. The quality of HTTP video depends on many factors, including Content Delivery Network (CDN) and caching strategies, encoding parameters, player buffering and adaptation techniques, and of course, the performance of the network(s) over which it is transmitted. Unlike UDP/RTP video, where poor network performance (e.g., losses, or excessive delay and jitter) led to video artifacts due to dropped or late frames, in HTTP video, poor network performance translates into stall events (caused by playout buffer underruns), and in some cases lowered video quality as a result of adaptation, by using technologies such as Dynamic Adaptive Streaming over HTTP (DASH) [2], which can change the video's representation in order to overcome network impairments, e.g., using lower bit-rates when facing network congestion.

The fact that HTTP video is streamed over TCP, together with the different existing buffering and adaptation strategies, make external prediction of stall events (and their duration) hard to do. Yet, for some of the stakeholders in the delivery chain, notably network providers, it is important to be able to monitor service performance, as they are often the first ones to be blamed when a service does not work properly.

Ideally, network providers would be able to monitor network QoS for video traffic, and detect (or even better, predict) when QoS issues will lead to quality degradation for the users. This would enable them to take active steps in order to solve the problems in a timely fashion. To do this, one must first understand how network impairments affect the buffering performance of the video players, how the adaptation mechanisms in the players react to this, and how their reaction feeds back into the network performance. This is still an open problem.

In this paper we provide a first (and humble) step towards a solution for this problem, by showing that in the absence of adaptation, the perceived quality for HTTP video can be estimated with at least a one minute granularity and with minimal knowledge of player parameters (buffer size and segment lengths). We do this by layering two models, one for the buffer performance as a function of the network QoS, and another one for perceived quality as a function of buffer performance. This layered approach has several benefits, as it allows a clean separation of concerns and expertise in the modelling process, as well as the possibility of using several different models (or make the models parametric in order to accommodate different scenarios) in order to accommodate different services. QoE in the literature is widely understood to be of multidimensional nature, being influenced by many different factors (for example human and contextual factors in addition to technical ones). While this study covers only technical influence factors and perceptual dimensions of QoE, the presented layered model itself can be extended to cover other input factor spaces.

The rest of the paper is organized as follows. Section II provides an overview of related work and the layered modeling approach we took. In Section III, we describe the experiments carried out for this work, and their results. Section IV deals with the modelling and the performance

[1] https://www.youtube.com/yt/press/statistics.html

of the models obtained. We conclude the paper and discuss future work in Section V.

II. Background

A. Related Work

With the explosion of video streaming services like YouTube, Vimeo and later Netflix, HBO Go, etc., the importance of QoE for HTTP-based streaming services increased and a wealth of literature has been written on the topic. HTTP video presents different challenges for quality assessment than RTP/UDP video did, as the impairments to which the user is subjected are very different. Whereas in RTP/UDP[2], network impairments such as losses and jitter result in visible and audible artifacts (and hence clear quality degradation), HTTP streams manifest these QoS issues as stalls in the playback and initial waiting times. Hence, the quality experienced by the users is no longer dominated by artifacts[3], but rather by the pauses in playback — their frequency and duration — and to a lesser degree by the initial waiting time before playback starts, as unlike in the case of real-time streams, HTTP video players buffer several seconds' worth of content before starting playback.

With respect to waiting times, it has been shown that even relatively large initial waiting times do not have a large impact on quality [3], [4], and that the main source of annoyance for the users are playback stall events. Moreover, some results [5], [6] show that the distribution of the stalling times, in particular the number of stall events, has a significant impact on quality. This precludes simple metrics such as total stalling times from being sufficient for providing accurate quality estimates.

With the introduction of DASH [2] and other adaptive streaming technologies, new QoE influence factors related to the adaptation (e.g. the types of adaptation, and adaptation strategies used by the player) became relevant to the quality of HTTP streams.

An excellent overview of QoE issues for adaptive HTTP streaming is given by Seufert *et al.* in [7].

For the most part, the efforts found in the literature have been focused on understanding the relations between stalls, start up delays, adaptation-related factors, and perceived quality. The network aspects, however, are often ignored or subsumed into the buffer performance. An exception to this is the very interesting work of Mok *et al.* [6], which proposes something quite similar to what we propose in this paper, namely, the integration of network performance and perceived quality models. Our work differs from theirs in several important aspects. Firstly, we consider DASH streams instead of Flash-based ones (despite not considering adaptation in this work, we do consider other DASH parameters such as segment length). Secondly, their subjective assessment presents certain flaws (low number of subjects, use of a single source clip) which we have avoided. We have also considered the distribution of stalls and their relative durations, which was not a factor in their study. Our results show that the stall durations are significant (in line with e.g. [5]), whereas their results claim that they are not. On the network performance side, our approach is also different from theirs. They consider packet loss and round-trip time (RTT) as their QoS factors, whereas we considered packet loss and mean loss burst size, but not RTT. The reason for this is that bursty loss events are more common than excessively long RTTs in wireless networks, which are becoming a dominant use case for HTTP video.

[2]And other "real-time" streaming protocols and approaches.

[3]Though depending on the aggressiveness of the encoding, artifacts may be evident.

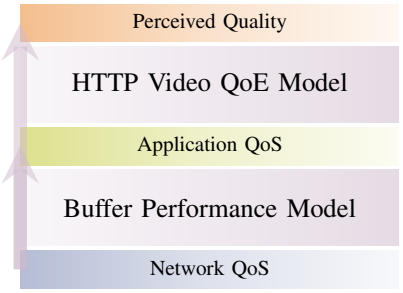


Figure 1: Layered models for mapping network QoS to perceived quality

B. Layered Models for QoE Monitoring

When dealing with HTTP video and transmission impairments, the mapping from QoS metrics to perceived quality is complicated by two facts, namely that a playout buffer sits in between, and that the streaming is done over TCP. TCP assures in-order delivery of all frames, and the video player, as opposed to RTP/UDP players, doesn't skip over late frames, but waits for them to be delivered. This leads to a situation where there are no visual quality impairments, but stalls and rebuffering events. QoE models for HTTP video, as mentioned in Section II-A, typically take these stall events (and other parameters such as start-up delay, and adaptation events) as a basis for making predictions of QoE. Integrating the QoS-to-buffer mapping into such a model is complex, and does not follow a "separation of concerns" strategy.

In [8], the authors propose a layered approach to building QoE models, whereby models for different parts of the overall system (including performance models for the technical parts of it, but also context and user models) can be stacked on top of each other, with the higher-level models feeding off the outputs of the lower-level models. This approach has also been adopted for a QoE monitoring architecture by ETSI [9] standardisation body.

Following that approach, our modelling strategy comprises two models, one mapping network-level QoS to application-level QoS (buffer behavior), and the second one mapping that application-level QoS to user perception, as depicted in Figure 1.

III. Experiment Design

The results presented in this paper stem from two separate experiments, one for understanding the impact of playout stall events on the subjective perception of the video streams, and the other one for understanding the

impact of network QoS on the playout buffer behavior (notably, stall events). In this section we describe both experiments and their respective results.

A. Subjective Assessment of HTTP Video

1) Evaluation methodology and test session structure: We chose the subjective assessment methodology following the results of Garcia et al.[10]. In line with other studies on HTTP video quality, we used a Single-Stimulus (SS) with Absolute Category Rating (ACR), in order to obtain results comparable with the available literature, and in accordance to the ITU-T P.913 recommendation [11].

For better granularity in the voting, we extended the standard five-point ACR scale into a nine-point scale by adding an intermediate (unlabelled) voting option between the original categories as illustrated in Table I. The voting prompt presented to participants was *Please rate your viewing experience*.

In order to provide a suitable variety to the number and duration of stall events, we used one minute-long samples.

Table I: The rating scaled used in the subjective assessment to answer the prompt: *Please rate your viewing experience*

Value	1	1.5	2	2.5	3	3.5	4	5.4	5
Label	Bad		Poor		Fair		Good		Excellent

For each sample, the following procedure was carried out: after the user clicked the *Start*-button, playback of the one-minute long video sample started immediately. As soon as the sample playback finished, user was presented a rating scale (radio buttons on grey screen on background). The voting time was not restricted. Once the user rated the sample and clicked the *Continue*-button, a new sample started after 3 seconds of displaying grey screen (with text *Next video sample starting...*) to let user prepare for the upcoming sample.

The overall test session had the following structure. When a participant arrived, she was greeted and offered some refreshments in order to help her detach from the daily work and thoughts. Then she was asked to read the instructions (possible questions were answered by the operator) after which she did a training session to familiarize with the test and voting tool (the training content was not present in actual test session). Then the actual test of 46 samples was performed by the participant, with enforced 2 minute pause after the 23rd sample. When the participant had finished all the tests, she was asked to fill a post-session survey and was rewarded with two movie tickets.

2) Material and conditions: The video sample material was chosen to represent typical video streaming content. The original videos are listed in Table II. Additionally, clips from the movie *Grace of Monaco* were used during training. The content was copied from the purchased Blue Ray discs and edited into clips of one minute length [4]. The video streams in final clips were encoded in high bitrate of approximately 15 Mbit/s, with frame rate of 25 fps and resolution of 1920x1080 pixels using H.264 video codec (High Profile, level 4.1). Audio tracks were encoded in bitrate of 192 kbit/s using AC-3 codec. The clips were chosen to be diverse in terms of scene cuts, movement and panning. The audio was considered in editing so that no clip starts or ends with interrupted utterance or sound (in some cases a controlled fade out was added).

The controlled variables (also called *independent variables*) are listed in Table III. The number of stalling events N_s refers to the number of events during the video sample, when the playback halts (because there is not enough data in player's buffer). The maximum of three stalls was chosen with presumption that it would be already intolerable for the most users (which was not interestingly the case). The total stalling time $T_{s,total}$ was the second controlled variable. It is the cumulative sum of lengths of stalls during the playback. The final controlled variable was the pattern in which the stalls of different lengths were presented to participants, P_s. Table III shows the exact patterns. There are patterns that have only equal length stalls, patterns that have increasing trend in stall length and patterns that have decreasing trend in stall length during the playback. As some combinations are not possible (e.g. pattern U:U[5] and single stalling) and as stalling lengths were integer values (for implementation reasons), there were total of 21 different conditions.

The stall patterns generation was based upon YouTube traces graciously provided by Columbia University, collected as part of the YouSlow [12]. From the trace data analysis, it emerged that a 2-state Markov chain is sufficient to generate stall patterns statistically similar to those observed in actual YouTube usage.

The low number of tested conditions allowed extra measures for decreasing statistical noise, which was found especially important as the actual buffering event instantiation had a random element due to use of Markov chain. To this end, participants evaluated each condition twice (but with different content and with different instantiation of stall pattern). We also added two repeated cases (with the same content) to allow consistency checking later on. Due to the way realistic stall patterns were generated (see below), 2 extra cases of 12 second/3 stalls conditions were added after closer examination of the generated stall patterns. The content for test conditions was randomly drawn from the pool of samples. The resulting total number of test conditions was *46*. The tested conditions were presented to each participant in randomized order in order to compensate any learning effect.

The test design allows analyzing also implicitly introduced independent variables (that were not controlled). The times, lengths and number of the stall events are known and can be used to calculate variables such as *average stall length* and *inter-stall length*. However, their

[4]The actual duration varied slightly, so as to preserve the semantic content of the scenes cut, and not to create interruptions in the dialog

[5]In order to refer to the approximate relative duration of stalls, refer to U as the "unit", or smallest stall observed, and G to mean "greater than the unit". A pattern of U:U means two equally long stallings, whereas a G:U:G pattern implies a long-short-long relation between three stalls.

Table II: Contents used for the subjective assessment

Movie	Description	Visual stimulus	Audio stimulus
Metsän Tarina	Document about life in Finnish forests	Fauna, flora and nature details	Narrator voice, nature's sounds
Need For Speed 4	Action movie	Fast objects, people, scene cuts	Dialogues, sounds of sports cars
Stalingrad	War movie	People, buildings, weapons	Dialog, narrative, sounds of war
Toy story 3	Animation	Characters of Toy story, computer generated graphics	Dialogues, sound effects
Transcendence	Sci-fi drama	People, technology of future, sceneries	Dialogues
Transformers 3	Sci-fi action	Aliens, robots, people, dinosaurs, fast objects	Dialogues, sounds of large scale machinery

Table III: Controlled Variables

Controlled Variable	Description	Levels
N_s	Number of stalling events in sample	0, 1, 2, 3
$T_{s,total}$	Total stalling time in sample	0, 1, 2, 3, 6, 12
P_s	Stalling pattern	U, U:U, U:G, G:U, U:U:U, U:U:G, G:U:U (U=Unit, G=Greater)

Table IV: Test environment parameters

Parameter	Value
Lighting level (background)	2.0 lx
Screen illuminance (peak)	129 cd/m^2
Noise level	Quiet, background noise of 38 dBA due to air conditioning
Peak audio volume	79 dBA
Viewing distance	3.2H
Environment	Controlled
Monitor	Sony 55X8505B TV
Size of monitor	55", 16:9 aspect ratio
Audio system	TV's built in loudspeakers

analysis is beyond the scope of this paper. Initial delay was left out from the list of controlled variables. As discussed in Section II-A its impact on QoE is not as large as that of stalling and rebuffering events.

3) Environment and demographics: The room where the tests were done was a living room-like space built into a closed laboratory space. The instructions of ITU-T P.913[11] were followed where applicable. The lights and audio volume were adjusted according ITU-T P.911 [13]. The environment parameters are summarized in Table IV.

The samples were presented to participants by means of custom software driving the VLC media player. The stall events were generated by pausing the playback with timers. During a stall event a buffering indicator was rendered on top of the paused picture.

Table V summarizes the key demographics data for the assessment. This data was not used in the analysis related to the results of presented work.

4) Voting statistics: We did a Shapiro-Wilk normality test for the collected voting data from the user campaign and found the votes to be normally distributed with alpha level 0.05 (p = 0.832). The outlier detection described in ITU-R BT.500 [14] yielded no outliers among the test subjects. Next, we studied the Standard deviation of Opinion Scores (SOS) test proposed in [15]. Figure 2 shows the standard deviation as a function of Mean Opinion Score

Table V: Demographics of the subjective assessment campaign.

	Category	Count
Gender	Female	8
	Male	14
Age group	20 - 29	1
	30 - 39	17
	40 - 50	4
Familiarity with HTTP video streaming technologies	Not familiar	2
	Slightly familiar	2
	Moderately familiar	7
	Very familiar	5
	Extremely familiar	6
Video service consumption	Occasionally	4
	Several times a week	10
	Daily	8
Assumed criticality at home	Less critical	2
	Similar	15
	More critical	5

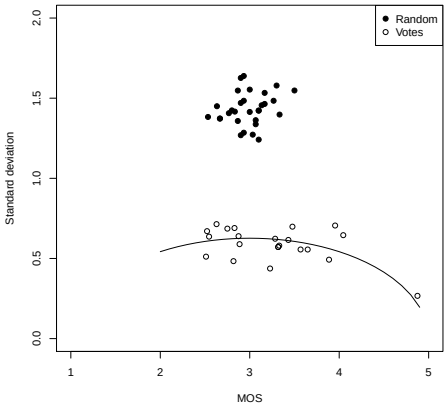


Figure 2: Standard deviation of Opinion Scores

(MOS) and the fit to the theoretical behavior of standard deviation. The figure shows that the users were not diverse in voting behaviour which implies good agreement and understanding of the task. Also the so called *a-value* (0.098) is just below, but rather close to the values in comparable studies (about video streaming) presented in [15].

Next, we looked at the significance of the effects of the independent variables, N_s, $T_{s,total}$, P_s had on dependent variable QoE_p (denoting the perceived quality). A type I ANOVA was calculated in order to capture any interactions, and no interactions were detected. We then

Table VI: Main Effects (Type II ANOVA)

Independent Variable	F-value	p-value
N_s	28.148	**<0.001**
$T_{s,total}$	110.151	**<0.001**
P_s	1.703	0.198

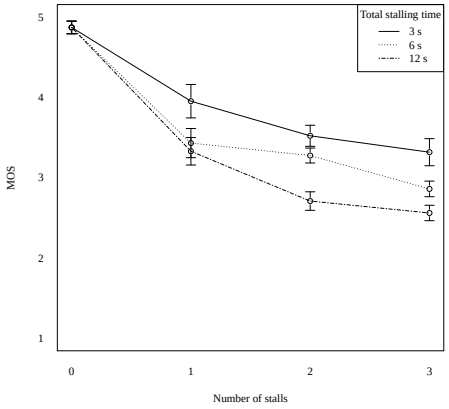


Figure 3: MOS as a function of number of stalls for different total stalling lengths

did a three-way type II ANOVA to calculate the main effects shown in Table VI (significant ones bolded). N_s and $T_{s,total}$ were found significant and P_s insignificant on alpha level 0.05.

Figure 3 illustrates how MOS behaves as a function of number of stalls, for each of the total stall times considered. The observed MOS values are comparable (albeit slightly higher) to previous results in the literature (e.g. [3], [16]). Both N_s and $T_{s,total}$ have an effect on perceived quality (in most cases no overlapping confidence intervals). The first stalling event has higher negative impact than subsequent ones. The calculated MOS conforms also to the Weber-Fechner Law (WFL) [17] regarding at least N_s (similar can be observed for $T_{s,total}$ when $N_s = 3$). According to WFL the relationship between the magnitude of a stimulus and perceived quality (or other perceived intensity of the stimulus) is of logarithmic nature.

Finally, we conducted an analysis of the possible fatigue and learning effects on the assessment. The fatigue test showed no evidence of fatigue on the results. We also studied the effect of the previous test condition on the current vote. Figure 4 illustrates the effect of the previous test condition on the mean deviation of MOS for controlled variable N_s. The results show that with no stalls in previous condition, the voting behaviour differs from the cases with one or more stalls experienced in previous condition with the highest deviation (around 0.5) found for $N_s = 1$. Furthermore, the results from Type I ANOVA show that the number of stalls (N_s) effect is in fact significant on alpha level 0.05 (mean deviations are different, p = 0.02). This effect is not observable in the MOS due to the random order in which we presented the test conditions to the

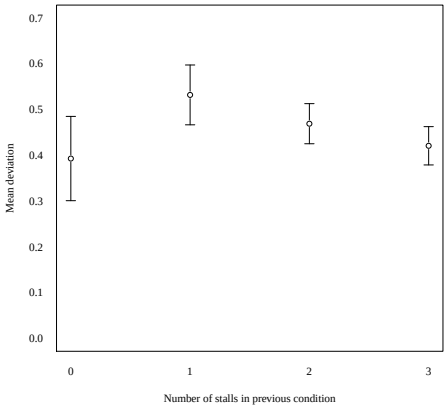


Figure 4: The effect of previous condition on the mean deviation of MOS for controlled variable N_s

subjects.

B. Measuring the Impact of QoS on Buffer Behavior

1) Simulation environment: We built a simulation environment for capturing the relation between buffering events in DASH streaming application and the network conditions considered. In these tests, some simplifications were done in order to make the variable space manageable. In line with the quality assessment campaign described in Section III-A, we left adaptations out of the scope in order to first understand how the player buffer behaves in single bitrate configuration.

The simulation environment is illustrated in Figure 5. It consists of three computers joined by a 1 Gbps Ethernet network. The network was isolated in order to prevent unwanted network fluctuations and disturbances. Host 3 was running an instance of Nginx HTTP server that served the DASH formatted media files to client executed on Host 1. Host 1 was connected to Host 3 via Host 2 that acted as a L2 network bridge and an emulated network. Network emulation was done with regular Netem and was controlled remotely by the Testing scripts. The streaming client in Host 1 is an in-house developed DASH simulation tool that can calculate the buffer utilization during a streaming session without the actual playback. The tests were configured and executed by a set of scripts executed in Host 1.

An execution of single tested condition had following steps: 1) Load the test condition containing player and network parameters, 2) configure network emulation, 3) configure DASH streaming application, 4) start DASH playback session and 5) after the 1-minute playback session, collect the results (statistics about buffering events).

2) Tested conditions: The work was done iteratively in order to identify the most important controlled variables and their ranges. After each of the iterative runs (in total 5 of them), the intermediate results were analyzed and range(s) of controlled variables were adjusted. For example, the original test set included three different one minute video samples. Preliminary analysis showed that

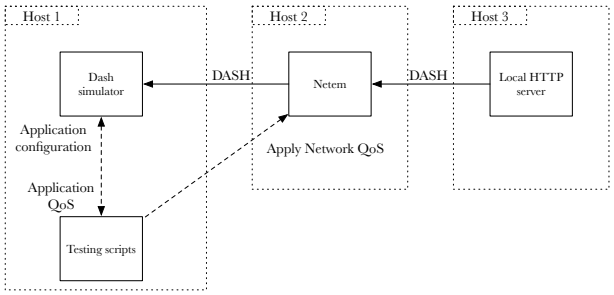


Figure 5: Network QoS to DASH events simulation environment

Variable	Description	Levels
Controlled variable		
$L_{segment}$	DASH segment length	2, 6, 10 s
T_{buffer}	Buffer length	2, 4, 8, 16 s
PL	Packet loss rate	0, 0.6, 1.2, 4.8, 9.6, 14.4, 19.2 %
B	Available network bandwidth	5, 6, 7, 8, 8.5, 9, 9.5, 10, 15 Mbit/s
Dependent variables		
N_s	Number of stalling events during a condition	
$T_{s,total}$	Total stalling time during a condition	

Table VII: Controlled QoS and dependent application variables

there was a minor content effect, but other factors were much stronger. Furthermore, we realized that the most important factor was the relation between nominal video bitrate and available bandwidth (as discussed in [5]). As a result, only one H.264 video clip, with average bitrate of 9 Mbit/s was included, allowing the inclusion of new data points to input factor B, available bandwidth.

Table VII lists the controlled (independent) variables and outcome (dependent) variables of the QoS to DASH events simulation (originally mean loss burst size, $MLBS$, was one of the controlled variables, but due to observed inaccuracy of netem it was dropped out of the analysis). During each test all the 2268 combinations (including MLBS) were iterated, resulting in run times of about 38h per test.

3) Results: Figure 6 illustrates the nature and the magnitude of the main effects and the strongest detected interactions. The individual main effects are extracted by aggregating the full result data set without filtering (therefore for example zero packet loss shows some stalling events). As can be observed, both N_s and $T_{s,total}$ are affected by all the independent variables, while $T_{s,total}$ is least affected by $L_{segment}$. Table VIII shows the interactions observed by the Type I ANOVA. Because of the numerous interactions in the result data (especially regarding dependent variable N_s), we do not present the significance of individual main effects. Also, the detailed analysis of the interactions is left out of this paper. Instead we use a machine learning approach to model the relation between network QoS and DASH events, for capturing the effects of interactions into a mapping function.

Independent Variable	N_s		$T_{s,total}$	
	F-value	p-value	F-value	p-value
$L_{segment}$ - T_{buffer}	**77.2**	**<0.001**	0.25	0.620
$L_{segment}$ - PL	**81.8**	**<0.001**	2.519	0.113
$L_{segment}$ - B	*3.812*	*0.051*	0.026	0.872
T_{buffer} - PL	**90.58**	**<0.001**	**22.50**	**<0.001**
T_{buffer} - B	**5.825**	**0.016**	0.009	0.923
PL - B	**11.35**	**<0.001**	0.303	0.582

Table VIII: Interactions between independent variables per dependent variables

IV. Modelling

A. Perceived Quality Model

The modelling of the dependent variable QoE_p as a function of the number of stallings N_s and the total stalling time $T_{s,total}$ was done with linear regression. It was possible to choose such a simple approach as there were no interactions, the main effects were clear and number of independent variables was small. The fit was verified with cross validation (5-fold with 80 %/ 20 % split between training and test sets). Equation (1) shows an instance of linear regression model from one of the training/test set splits.

$$QoE_p = 4.56 - 0.36N_s - 0.09T_{s,total} \mid N_s \in 0\ldots3 \; and \; T_{s,total} \in 0\ldots12 \quad (1)$$

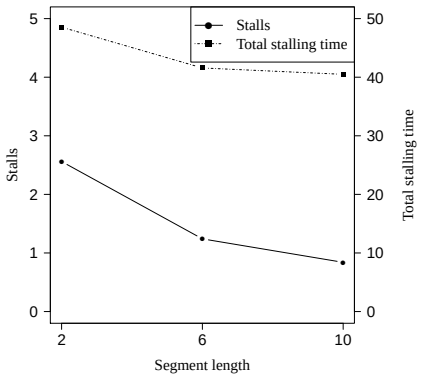
B. Buffer Performance Model

The high complexity and many interactions described in Section III-B3 precluded use of simple modelling approaches such as linear regression. Instead, two neural networks were trained and tested using the data from measurements described in Section III-B; one for predicting N_s and another one for predicting $T_{s,total}$ from independent variables $L_{segment}$, T_{buffer}, PL and B. Each neural network (created using the *neuralnet* package in R) had 5 input neurons, 10 hidden neurons and 1 output neuron. 10-fold cross-validation was performed (with 90 % / 10 % split to training and test sets). The performance and accuracy of the model is discussed in next section together with QoE model.

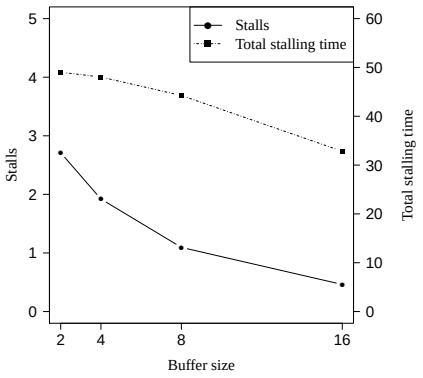
C. Accuracy of the Models

The performances of both perceived quality and DASH event prediction models were verified via series of cross-validations. For the subjective model, we used a five-fold cross-validation with 80 % / 20 % split between training sets and test sets. The average prediction performance indicates good fit with Pearson correlation of 0.94. The two DASH buffer event models (stalls prediction model and total stalling time prediction model) were verified with ten-fold cross-validation with a 90 % / 10 % split. Both models verify well with Pearson correlations of 0.988 and 0.997, respectively.

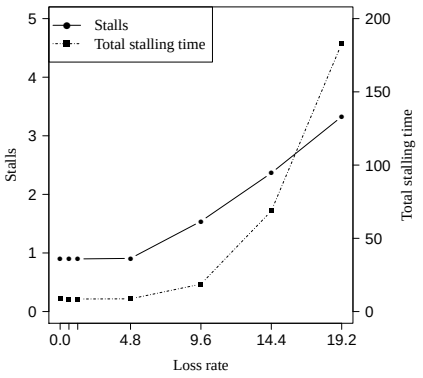
The results we present here rely on the assumption that the models will provide accurate results when chained. Given the remarkable accuracy of both DASH buffer event models, it seems reasonable that the outputs of these models can be fed into the perceived quality model without causing significant variability beyond the accuracy of the



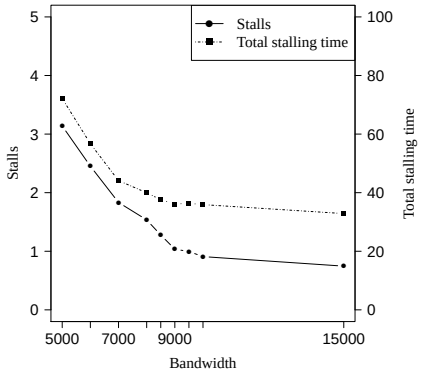
(a) Stalls/Total stalling time per Segment length



(b) Stalls/Total stalling time per buffer size



(c) Stalls/Total stalling time per loss rate



(d) Stalls/Total stalling time per bandwidth

Figure 6: Main effects

quality model itself. However, proper validation of the layered model via a new, specifically crafted subjective assessment campaign with samples created under specific network conditions is still needed before making stronger claims about the models' performance.

V. Conclusions

In this paper we have proposed a novel layered model mapping network QoS metrics to DASH streaming quality as perceived by the users, in the absence of adaptations. The model is composed of two simpler models, one for mapping the QoS metrics to the DASH buffer's performance (in terms of number of stalling events and overall stall duration), and a second one, mapping that buffer performance into perceived quality (which as for other media services, will likely be one of the most influential dimensions of QoE). A subjective assessment campaign was carried out to train the quality model, and a large-scale measurement campaign in a purpose-built test-bed was used to train the buffer performance model.

The subjective assessment results show that both the number of stalls and stalling time have a significant effect on perceived quality. From the perspective of the assessment itself, an effect of number of stalls in previous condition was also observed, which may deserve further analysis.

The QoS-to-Buffer measurement results show that DASH buffer events depend on various network QoS parameters in a non-trivial manner, as well as some application-dependent factors. The resulting models (QoS-to-Buffer and Buffer-to-Quality) were verified via cross-validation, and found to have very good correlation with observed results (≈ 0.99 and 0.94 average Pearson correlation over the cross-validation test sets, respectively).

Having this type of model at our disposal enables us to do network-side perceived quality (and hence, by proxy, QoE) estimations for over-the-top HTTP video streaming, which is a topic of particular interest to network operators.

The subjective assessment campaign, as well as the buffer performance measurements, were carried out over one-minute intervals, which means that the resulting models are accurate to within one-minute granularity. Future work (on-going as of this writing) includes shorter-term (say, five to ten seconds) prediction of stall events based on the occurrence of network impairments.

Having verified that this approach yields good results, we have two further lines of research in our road map. The first one is the inclusion of adaptation in the DASH player,

which is the most glaringly missing item in this work. The second one, is the development of more statistically useful models (as described, e.g., in [18]), providing for instance estimations of the distribution of MOS values, or values for key percentiles, rather than simply a MOS.

ACKNOWLEDGMENTS

Martín Varela's and Toni Mäki's work was partially funded by Tekes, the Finnish agency for research innovation, in the context of the CELTIC+ project NOTTS. Doreid Ammar's work was carried out during the tenure of an ERCIM 'Alain Bensoussan' Fellowship Programme.

REFERENCES

[1] Cisco, Inc., "Cisco Visual Networking Index: Forecast and Methodology, 2014-2019 White Paper," May 2015, Accessed: 2015-09-03. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf

[2] ISO/IEC, "Standard 23009-1:2014 — Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats," 2014. [Online]. Available: \url{http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=65274}

[3] T. Hoßfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial delay vs. interruptions: Between the devil and the deep blue sea," in *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, Jul. 2012, pp. 1–6.

[4] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching," *Broadcasting, IEEE Transactions on*, vol. 59, no. 1, pp. 47–61, March 2013.

[5] T. Hoßfeld, D. Strohmeier, A. Raake, and R. Schatz, "Pippi Longstocking Calculus for Temporal Stimuli Pattern on YouTube QoE: 1+1=3 and 1·4≠4·1," in *Proceedings of the 5th Workshop on Mobile Video*, ser. MoVid '13. New York, NY, USA: ACM, 2013, pp. 37–42. [Online]. Available: http://doi.acm.org/10.1145/2457413.2457422

[6] R. Mok, E. Chan, and R. Chang, "Measuring the quality of experience of HTTP video streaming," in *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*, May 2011, pp. 485–492.

[7] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hobfeld, and P. Tran-Gia, "A survey on quality of experience of http adaptive streaming," *Communications Surveys Tutorials, IEEE*, vol. 17, no. 1, pp. 469–492, Firstquarter 2015.

[8] M. Varela, L. Skorin-Kapov, F. Guyard, and M. Fiedler, "Meta-Modeling QoE - Towards a Generic Methodology for Building QoE Models," *PIK - Praxis der Informationverarbeitung und -kommunikation*, vol. 37, pp. 265–274, 2014.

[9] ETSI (Speech and Multimedia Transmission Quality — STQ), "TS 103 294: Quality of Experience A Monitoring Architecture," 12 2014.

[10] M.-N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnstrom, and A. Raake, "Quality of experience and HTTP adaptive streaming: A review of subjective studies," in *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, Sep. 2014, pp. 141–146.

[11] International Telecommunication Union, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," *ITU-T Recommendation P.913*, January 2014.

[12] H. Nam, K.-H. Kim, D. Calin, and H. Schulzrinne, "Youslow: A performance analysis tool for adaptive bitrate video streaming," in *Proceedings of the 2014 ACM Conference on SIGCOMM*, ser. SIGCOMM '14. New York, NY, USA: ACM, 2014, pp. 111–112. [Online]. Available: http://doi.acm.org/10.1145/2619239.2631433

[13] International Telecommunication Union, "Subjective audiovisual quality assessment methods for multimedia applications," *ITU-T Recommendation P.911*, December 1998.

[14] ——, "Methodology for the subjective assessment of the quality of television pictures," *ITU-R Recommendation BT.500-12*, Mar. 2009.

[15] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *QoMEX 2011*, Mechelen, Belgium, Sep. 2011.

[16] A. Sackl, S. Egger, and R. Schatz, "Where's the music? comparing the QoE impact of temporal impairments between music and video streaming," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 64–69.

[17] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo, "The Logarithmic Nature of QoE and the Role of the Weber-Fechner Law in QoE Assessment," in *2010 IEEE International Conference on Communications (ICC)*, May 2010, pp. 1 –5.

[18] T. Hoßfeld, P. Heegaard, and M. Varela, "QoE beyond the MOS: Added Value Using Quantiles and Distributions," in *QoMEX 2015*, Costa Navarino, Greece, 5 2015.