Embracing Uncertainty: A Probabilistic View of HTTP Video Quality

Martín Varela[†]
EXFO
Oulu, Finland
Email: martin.varela@exfo.com

Toni Mäki
Aalto University
Espoo, Finland
Email toni.j.maki@aalto.fi

***Abstract*—When dealing with Quality of Experience (QoE) and in particular perceptual quality assessment and modeling, averaging is a common occurrence. For instance, the most commonly used measure of QoE is the aptly-called Mean Opinion Score (MOS), which is intended to represent an idealized average subject's rating of the quality. Another form of averaging occurs when choosing and preparing the samples used for the assessment, which are supposed to be representative of an average viewing situation. This leads to nice, smooth scalar representations of quality, but at the same time, it leads to a loss of information. In this paper we present a first step towards working with all the information available in an explicit way, rather than averaging it away. We do so in the context of constructing layered quality models for HTTP video streaming (using Dynamic Adaptive HTTP Streaming — DASH, excluding its adaptation feature at this stage), mapping network-level QoS measurements to probability distributions of different MOS values for a given set of conditions.**

***Index Terms*—QoS, QoE, HTTP Video, DASH**

I. Introduction and Related Work

The Mean Opinion Score (MOS), and its variations (DMOS, etc.) have been the *de-facto*, go-to measure for quality assessment and modeling for a long time. It provides a neat, simple way to deal with quality estimates, by giving an easy to understand scalar value to quality. More recently, however, the limitations of dealing with averages in this context have become clear (1; 2). While in the field of quality estimation the use of averages continues to be dominant, in some other domains the use of distributions as a prediction outcome has been considered. In (3), Zhao et al. propose a method for predicting probability distribution of image emotions in Valence-Arousal space. Tian et al. present a method for predicting travel time distributions in the field of transportation research(4).

In this work we look at the behavior of DASH video streams, in the absence of adaptation, and in particular, how the presence of losses in the network can lead to different playout behaviors, and therefore different quality ratings. Our previous work (5) proposed a layered model for DASH streams which maps network QoS values to perceptual quality by first estimating the number and duration of the stall events that will happen during playback, and then producing a MOS estimate from those. Similar stall-to-quality models can be found in the literature, and a good overview of them can be found in (6).

The proposed layered model works well, but it has two inherent limitations. Firstly, the output is a MOS estimate, thereby not allowing a host of potentially useful applications for service providers (e.g., for determining the ratios of satisfied to unsatisfied users under a given condition). Secondly, the models mapping the QoS metrics to playout behavior, that is, to the expected number of stalls and their expected duration, deal with averages, whereas in practice there can be non-trivial (with respect to the observed quality) variations in those two values.

This paper presents our currently on-going work towards developing new quality models able to estimate the probability distributions of quality assessments from the observed network and application conditions. At this stage, we are able to predict the distribution of MOS values, by first obtaining the distributions for number of stalls and their duration, and then applying the quality model developed in (5). The results obtained so far are by no means perfect, but indeed promising, and they provide a stepping-stone towards a fully probabilistic view of quality, which would be of significant interest for content and network providers, as it would allow e.g., to predict how many users might complain about poor quality, or eventually churn. The separation of concerns provided by the layered approach taken (7) also allows the re-use of the models obtained in different contexts.

The rest of the paper is organized as follows. In Section II, we describe how we model the playback behavior from network QoS and basic information about the stream and the player. Section II-A describes the experiments carried out in this work. Sections II-B and III describe the model created and its performance. We conclude the paper and discuss future work in Section IV.

II. From Network QoS to Playback Behaviour

The performance of video streaming over HTTP depends on a variety of factors, of which some make modeling said performance difficult. In terms of quality, the main aspects to consider are related to the number of stalls during the playback, and their duration (the other obvious factor, start-up time, has been shown (8) to have only a minor impact on the perceived quality of the playback, and is therefore

[†] M. Varela's work was done during his tenure at VTT Technical Research Centre of Finland.

TABLE I: Streamed content and average bit rates (60s clips)

Content	Average bit rate (Mbps)
Need for Speed	8.05
Stalingrad	8.17
Toy Story 3	8.06
Transformers — Age of Extinction	9.00

TABLE II: Network parameter values

Parameter	Values
Bandwidth (Mbps)	6, 8, 9, 10, 15
Loss Rate – LR (%)	0, 1, 5, 7.5, 10, 12.5, 15
Mean Loss Burst Size – MLBS (packets)	1, 1.25, 1.5

not considered here). These, in turn depend on the player's buffer size, buffering strategy, the segment size used, and of course, the video bit rate and the network performance. Other aspects, such as DASH quality adaptation strategies also have an impact on the way network QoS affects the video playout. For simplicity's sake, we did not consider adaptation at this stage, as it was not clear when starting out whether modeling even the simpler, non-adaptive case was feasible.

A. Experiment Design

The experiment data was collected on a testbed comprising three computers running Linux (Ubuntu 14.04 LTS), connected by an isolated 1Gbps Ethernet network. One of the hosts runs the Nginx HTTP server, providing the content in DASH format to another host . In between these hosts, sits a third host providing a layer-2 bridge and emulation for different network conditions, via NetEm.

The measurement experiment was conducted by streaming different contents (four source sequences, encoded at between 8 and 9Mbps, cf. Table I) over the emulated network, to an in-house developed tool (*dashsimu*) that does DASH streaming without rendering the video.

For the network emulation, we introduced losses, using NetEm's built-in simplified Gilbert model, and limited the bandwidth of the link. The segment size and the buffer size were fixed at 2s. The network parameters were varied as shown in Table II. As can be expected by the choice of representations and rate limits, the resulting playouts ranged from very smooth to extremely impaired.

A total of 10471 streaming sessions (covering each parameter and content configuration multiple times, each playing back 60s of video) were carried out sequentially during the experiment. The process was scripted, and for each condition, the script set up the network emulation, configured the streaming application, ran it, and afterwards collected the results.

A second, independent experiment was run for validation of the model created. The validation experiment covered 400 sessions with conditions set up with the parameter values described in Table III

B. Modeling

The goal of this work was to provide estimates for the distributions of the number of stalls and their duration, with a one-minute playback window. To this end, we collected the playout statistics for all the streaming sessions from the *dashsimu* tool. With those, we created histograms, which we then used as training data for our models. For the number of stalls, we binned the data for each integer between 0 and 9, with an extra bin for all values > 9. In practice, given the results obtained in previous campaigns, values higher than 3 or 4 stalls in a minute are assessed as unacceptable by users, and hence are not particularly interesting.

For the total stall duration, we binned the data in 5-second intervals, with a total of 19 bins considered (as in the number of stalls case, this provides a far larger range than is needed in practice, as stalls that long lead to session abandon in most cases), the last one covering all larger values.

We used simple 3-layer feed-forward neural nets (implemented with R's *neuralnet* package) for training. We also tried more complex NN architectures, but they did not offer any meaningful improvement over a 3-layer one. The training was carried out in a ten-fold cross-validation fashion, doing random splits (90% training, 10% verification) of the experimental data. All the performance results presented in the next section correspond to conditions observed during the separate validation experiment, of which the data was not used during training.

TABLE III: Network parameter values for the validation experiment

Parameter	Values
Bandwidth (Mbps)	8, 12
LR (%)	0, 3, 8, 12
MLBS (packets)	1, 1.3

TABLE IV: Kullback-Leibler divergence summary for the number of stalls distribution estimations (only non-infinite values)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.01181	0.06342	0.09561	0.19770	0.32130	0.51870

III. Results

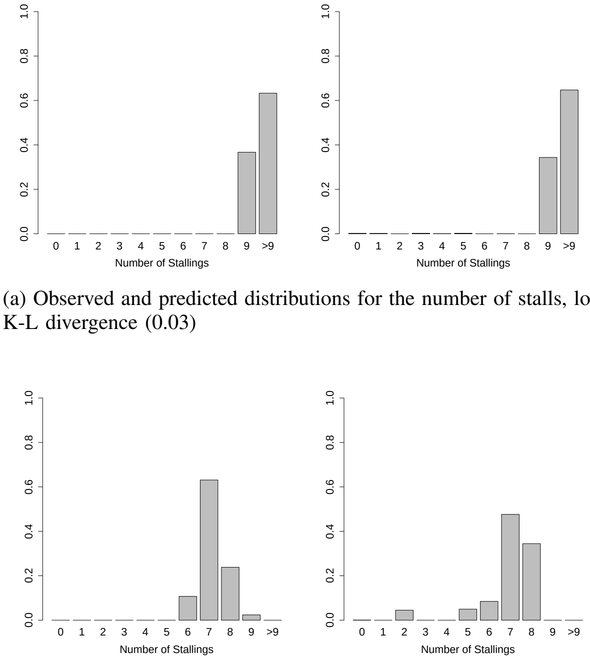
The resulting models for the number of stalls and their duration were tested against the data from the validation experiment described above. Visually, the resulting histograms appear to be, in most cases, a good approximation of those observed in practice, but in order to quantify their closeness, we computed Kullback-Leibler divergence on them. For the number of stalls, we found that out of the sixteen validation conditions, two of them had an infinite value for K-L divergence (i.e., the predicted distribution did not resemble the original), whereas the remaining fourteen had low K-L values (cf. Table IV). The two failed cases corresponded to the same combination of LR and bandwidth values (3% and 8Mbps, respectively).

For the stall duration, all K-L divergence values were bounded, and slightly higher than those of the number of stalls, but still show that the predicted distributions are close to the original ones. The results are summarized in Table V.

TABLE V: Kullback-Leibler divergence summary for the stall duration distribution estimations

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0274	0.1209	0.1889	0.3510	0.3248	1.4230

Going over the ten-fold cross-validation data, we found comparable results for both models. Roughly 13% of the cases showed an infinite K-L divergence value, while mostly still showing a significant overlap in the modes of the distributions, whereas the rest of the cases show that the predicted distributions are indeed similar to the observed ones. Figures 1 and 2 show the predicted and observed distributions for good (low K-L divergence) and bad (high K-L divergence) performing cases of the number of stalls and total stall duration distribution estimations, respectively. for a case with low K-L divergence value (0.02, for a condition with LR=15%, MLBS=1.5 packets and bandwidth of 9Mbps), and another with an infinite K-L divergence value (for a condition with LR=1%, MLBS=1.5 packets, and bandwidth of 9Mbps). We could not discern any obvious commonalities in the conditions where the K-L divergence goes to infinite, so the causes for this remain to be understood as of this writing.

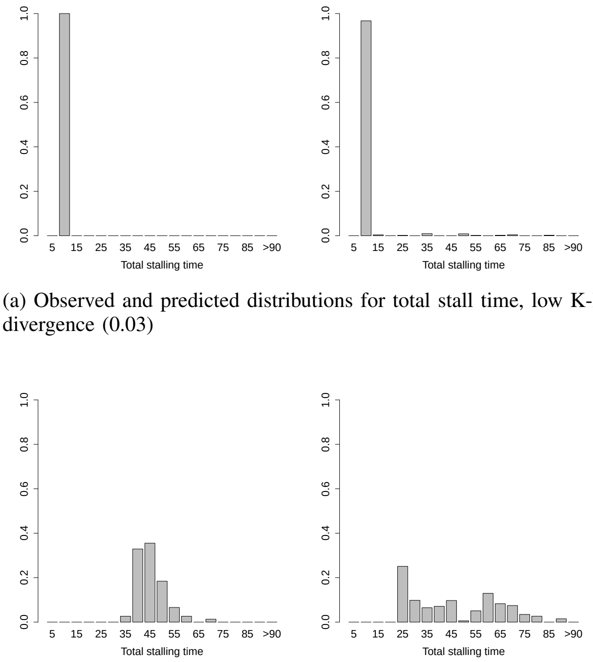


(a) Observed and predicted distributions for the number of stalls, low K-L divergence (0.03)

(b) Observed and predicted distributions for the number of stalls, infinite K-L divergence

Fig. 1: Performance of the distribution prediction for the number of stalls, (a) good fit (LR=15 %,MLBS=1.25, BW=9Mbps, K-L divergence = 0.03) and (b) bad fit (LR=7.5%, MLBS=1.5, BW=8Mbps, infinite K-L divergence) examples.

At any rate, the results obtained seem promising, in that we can successfully estimate the distributions for the number of stalls and their duration in at least $\sim 87\%$ of the cases. With these values, and the HTTP video quality model from (5) (or any other such model from the literature), it is possible to determine the distribution of MOS values from the estimated joint probability distribution for the number of stalls and their duration.



(a) Observed and predicted distributions for total stall time, low K-L divergence (0.03)

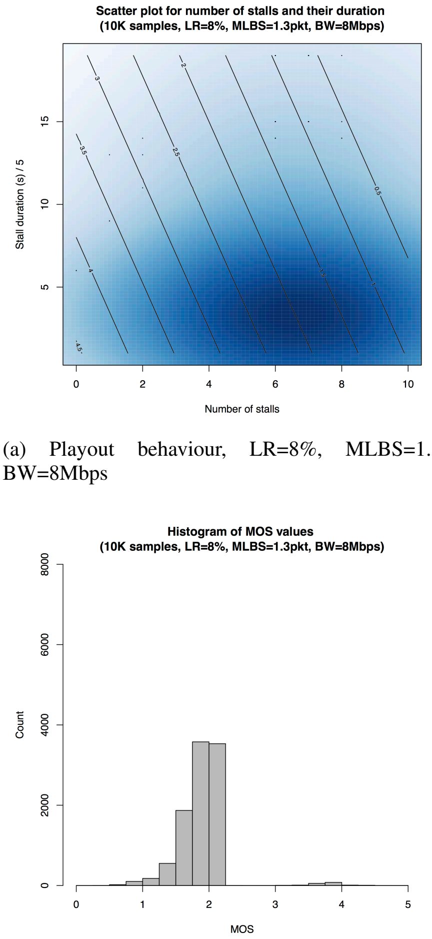
(b) Observed and predicted number distributions for total stall time, high K-L divergence (1.54)

Fig. 2: Performance of the distribution prediction for the total stalling time, (a) good fit (LR=0%, BW=6Mbps, K-L divergence = 0.03) and (b) bad fit (LR=12.5%, MLBS=1.5, BW=15Mbps, K-L divergence = 1.54) examples.

The models allow operators and service providers to better understand how the streaming behaves as a function of network QoS. In terms of perceived quality, layering the distribution models with the quality model from (5) allows us to predict the distribution of MOS values, and to relate the expected playout behavior with the perceived quality, as shown in Figure 3. While there is a conceptual difference (and practical shortcoming) between predicting a distribution for MOS values and actual ratings, the results show that this is a viable way towards the latter. The layering of models used here and proposed in (5) is reliable, and given a QoE model able to predict rating distributions (or at least quantiles), the layered model would provide an accurate probabilistic view of QoE.

IV. Conclusions and Future Work

In this paper we have presented a simple probabilistic view of HTTP video quality in terms of network performance. We do so by estimating, given a set of network conditions, the expected distributions for the number of stalls occurring during playback, and their duration. We then use a layered approach to estimating the perceived quality, by using a simple model that maps the predicted playout behavior to quality estimates, in the form of MOS. This leads to predicting the distribution of



(a) Playout behaviour, LR=8%, MLBS=1.3, BW=8Mbps

(b) MOS distribution, LR=8%, MLBS=1.3, BW=8Mbps

Fig. 3: Estimation of MOS distributions. Density plot for number of stalls and their duration, overlaid with MOS contours, and MOS histogram (LR=8%, MLBS=1.3, BW=8Mbps)

MOS values for the network conditions considered. This type of approach allows content and network providers to better understand the levels of user satisfaction in their user base, by exploiting more of the information available, instead of simply averaging it away.

The ultimate goal of this approach is to give operators and service providers a holistic view of service quality, by providing them with an estimation of the expected user ratings. This allows them to adjust their operations to reach whatever satisfaction objectives are required by their business (e.g., 95% of users should have good or better ratings), within their budget capabilities. In practice, this requires suitable network QoS monitoring to be deployed, and depending on the actors involved, it may also require cooperation (e.g., between ISP and OTT providers).

The work, in its current form shows that this type of approach is feasible, but it has two important limitations. The first of these is the lack of adaptation in the DASH streams (and its impact on the network behaviour). This is the subject of on-going work. The second one is that for a small fraction ($\sim 13\%$) of cases the prediction is still poor. We believe this can be addressed with larger data sets.

The first one is that it predicts a distribution of possible MOS values, whereas the ultimate goal is to predict the distribution of user ratings. The solution for this requires having a sufficiently large number of user ratings, so as to be able to estimate their distribution with sufficient accuracy. This will require some large-scale (most likely crowdsourced) study, as laboratory-scale studies are not sufficient for this purpose. Secondly, there is a fraction (roughly 13%) of seemingly unrelated cases where the performance of the prediction is still poor. The vast majority of the cases, however, produce very accurate results, which is promising. We believe that this can be solved by applying more advanced ML techniques, in particular ones that allow to express the relationship between each of the outputs, which the simple neural net used herein does not allow for. Finally, both models considered here do not take adaptation into account. This is being addressed as of this writing.

References

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

[2] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, "QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS," *Quality and User Experience*, vol. 1, no. 1, p. 2, 2016.

[3] S. Zhao, H. Yao, and X. Jiang, "Predicting continuous probability distribution of image emotions in valence-arousal space," in *Proceedings of the 23rd ACM MM*. New York, NY, USA: ACM, 2015, pp. 879–882.

[4] D. Tian, Y. Yuan, H. Xia, F. Cai, Y. Wang, and J. Wang, "A route travel time distribution prediction method based on markov chain," in *Smart Cities Conference (ISC2), 2015 IEEE First International*, Oct 2015, pp. 1–5.

[5] T. Mäki, M. Varela, and D. Ammar, "A Layered Model for Quality Estimation of HTTP Video from QoS Measurements," in *SITIS / QUAMUS 2015*, Bangkok, Thailand, Nov. 2015.

[6] 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.

[7] 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.

[8] 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," *IEEE Trans. Broadcasting*, vol. 59, no. 1, pp. 47–61, March 2013.