Impact of Intermediate Layer on Quality of Experience of HTTP Adaptive Streaming

Michael Seufert*, Tobias Hoßfeld†, Christian Sieber‡
*University of Würzburg, Institute of Computer Science, Würzburg, Germany
†Now with: University of Duisburg-Essen, Chair of Modeling of Adaptive Systems, Essen, Germany
‡Now with: Technische Universität München, Institute for Communication Networks, Munich, Germany
Email: seufert@informatik.uni-wuerzburg.de | tobias.hossfeld@uni-due.de | c.sieber@tum.de

***Abstract*—HTTP Adaptive Streaming (HAS) adapts the video quality to the current network condition by switching between different quality layers. As HAS was shown to perform better than classical video streaming, it is becoming increasingly popular. Recent research showed that quality switch amplitude and time on layer have an impact on the Quality of Experience (QoE) of HAS. However, those studies focused only on adaptation between two layers so far. This work extends these findings by taking adaptation between three layers into account. Thereby, especially the impact of an intermediate layer on user perceived quality is investigated. Crowdsourcing experiments were conducted in order to collect subjective ratings for adaptation between three layers. The results indicate that the quality of each layer and the time on each layer are important QoE parameters. This encourages the usage of temporal pooling approaches for QoE prediction and QoE-aware traffic management. Therefore, mean pooling of per-frame metrics will be applied and its performance will be validated with the subjective crowdsourcing results.**

I. Introduction

Video streaming services are among the most popular and volume-demanding services in the current Internet. Nowadays, many video services have adopted HTTP adaptive streaming (HAS) technology, which was standardized as MPEG-DASH [1], to align the video quality to the current bandwidth conditions and to avoid the shortcomings of classical streaming (i.e., stalling). To utilize HAS, a video has to be available on the server in multiple bit rates, i.e., quality levels, and split into small segments each containing a few seconds of playtime. The client requests an appropriate segment of the video based on measurements of the current bandwidth or the buffer filling level. As the client selects segments, which can likely be downloaded before their playout deadline, stalling is avoided to the greatest possible extent.

However, the adaptation of the video bit rate influences the video quality, which can be perceived by end users. Much research has been conducted to quantify the impact of adaptation on the Quality of Experience (QoE). Ultimately, QoE models are needed, which can predict the subjectively perceived quality of users of a HAS service. Influence parameters, which are typically investigated, are initial delay, stalling delays and frequencies, played back video quality, as well as time on a video quality layer, and quality switching frequency. However, many research works are based on adaptation between a high and a low quality layer only, and thus, results cannot be directly transferred to currently deployed implementations, which utilize also intermediate quality layers.

In this work, the impact of an intermediate layer on the perceived quality of HAS is investigated. Extending the works in [2], crowdsourcing studies are conducted to collect subjective ratings for adaptation conditions, which include an intermediate quality layer. The questions are answered whether the time on the intermediate layer and the position of the intermediate layer have an impact on QoE. Thereby, our findings contribute to future QoE models, which take into account multiple quality layers. Moreover, temporal pooling for QoE prediction (e.g., [3], [4]) is revisited. A simple mean pooling of different per-frame metrics is applied and the resulting performances are compared to the subjective ratings obtained in the crowdsourcing study. The comparison validates the temporal pooling approach, which, thus, could be utilized in future QoE models and should be considered for automated QoE evaluation frameworks and QoE-aware traffic management.

The remainder of this work is structured as follows. Sec. II discusses related works on HAS QoE and the impact of an intermediate layer. Sec. III outlines details on the test setup as well as the crowdsourcing study. Sec. IV presents the results of the impact of the intermediate layer on QoE and validates the temporal pooling approach. Sec. V concludes our work.

II. Related Work

As the popularity of HAS is ever growing, also many research has been conducted towards its Quality of Experience [5]. Advanced results are available from works, which considered adaptation between two quality layers. [6] showed that the adaptation amplitude has a significant impact on the subjectively perceived quality. These findings were confirmed by [7], which also showed that the length of the interval between low quality sections has no significant impact. [2] extended these results showing for two quality layers that the MOS of the constant high/low quality profiles bounds the MOS of the adaptation patterns, and that the overall time on each quality layer has a significant impact on QoE.

Other works focused on more layers in order to investigate whether smooth or abrupt adaptation was perceived better by end users. [8] found by using paired comparison that a stepwise decrease of image quality is rated slightly better than one single decrease. Also [9] confirmed that switching to a lower quality is generally considered annoying. However, abrupt up-switching could possibly increase QoE as users might appreciate the visual improvement. [10] investigated adaptation patterns, which contained both a quality decrease and an

978-3-901882-77-7

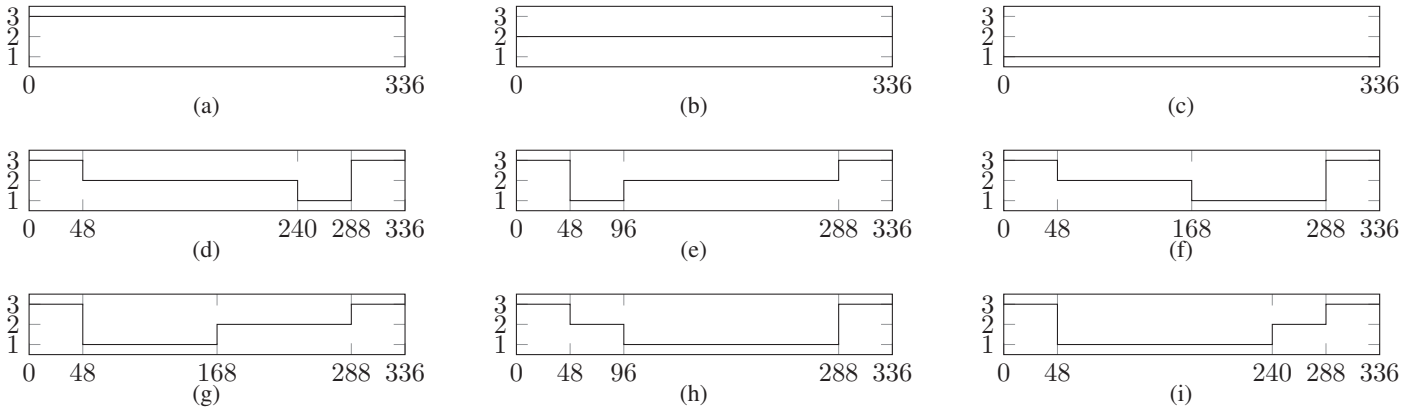


Fig. 1: Quality switching patterns with the quality layer on the y-axis and the video time in terms of frames on the x-axis.

increase. They found that gradual switching resulted in a worse QoE than abrupt switching. [11] found that neither smooth nor abrupt had a significantly better performance, although results in [12] indicate that slow gradual down switching was preferred. Also [13] did not find a significant difference for quality increases. Only for down switching smooth switching performed slightly better in terms of QoE than abrupt switching. [14] described that switching to an intermediate layer before switching to a higher layer was preferred over multiple large amplitude switches. They also showed that among streams that use an intermediate layer, those with a higher intermediate layer quality are rated significantly better in most videos. [15] showed that a constant intermediate layer stream is rated better than a stream with a drop from high to low, although the latter had a higher average bit rate.

Several QoE models for HAS were already proposed in literature. [16] used random neural network based on stalling statistics and average quantization parameter to estimate the QoE. The psychometric model in [17] additionally took frequency and amplitude of quality switches, objective content quality parameters, and buffer thresholds into account, but was tested only on content with two layers. [18] predicted the MOS of six layer adaptation patterns based on linear models of objective per-chunk metrics. [4] achieved a high prediction performance on the same six layer patterns by temporal pooling of per-frame metrics.

To sum up, the impact of an intermediate layer on the QoE of HAS was not explicitly investigated by subjective studies yet [19]. Existing works either were based on only two quality layers or considered intermediate layers only for comparing smooth to abrupt adaptation. However, when taking a closer look at these results and proposed QoE models, some suggest that intermediate layers have to be considered. Focusing explicitly on the impact of an intermediate layer, we close this gap with the results from our subjective crowdsourcing study.

III. Experimental Setup and Subjective Studies

Crowdsourcing Framework and Study Design: The subjective study was carried out by crowdsourcing, i.e., by distributing the experiment to a large and anonymous crowd of users through the Internet. The experiment was designed in cooperation with the crowdsourcing platform microworkers.com, which provides a large user base and international reach. The study was advertised on the platform and we offered a monetary compensation for the participation. To conduct the study, we utilize the web-based QoE framework QualityCrowd2 [20]. To ensure the quality of the results for distributed remote experiments, we follow the guidelines in [21], which summarize best-practices in the area of crowdsourcing-based QoE studies. For example, we asked simple content questions to check the concentration and reliability of the worker.

146 workers participated in the crowdsourcing campaign. Each worker had to access our study via the URL provided on the crowdsourcing platform and watch nine sequences in random order. To prevent any abuse (e.g., fast skipping through the sequences) or problems related to insufficient bandwidth (e.g., stallings), the test sequences are first downloaded to the browser cache and the worker has to watch and rate the current sequence before being able to go to the next one. After each sequence, the user is asked *Did you notice any changes in quality during playback? If yes, did you feel annoyed by them?* and is requested to set a continuous slider, which was marked with the 5-point DCR (degradation category rating) options *Imperceptible (did not notice any)*, *Perceptible but not annoying (did notice, but did not care)*, *Slightly annoying*, *Annoying*, and *Very annoying*.

To detect unreliable participants, we computed inter- and intra-rater reliability scores as the Spearman rank order correlation coefficient (SROCC) of a user's ratings and the mean ratings of all users (inter-rater) and the SROCC of a users' ratings and the time-averaged quality level of the patterns (intra-rater). We filtered out 69 users, which had at least one of the reliability scores (SROCCs) below 0.5, and 4 users, which could not answer the content question correctly. After the filtering, 73 users (50%) remained and their ratings were used for the evaluation.

Adaptation Patterns and Video Content: The crowdsourcing experiment was designed to evaluate nine different adaptation patterns, which are presented in Figure 1. Three constant quality profiles a-c were tested as reference profiles. Moreover, three pairs of symmetric patterns were created. All these patterns start and end with 2 s of playtime of the high quality layer. Patterns d/e play out the intermediate layer for most of the time (8 s). Patterns f/g stay the same time on the intermediate layer as on the low layer (5 s), and patterns h/i have the shortest time on the intermediate layer (2 s). Those pairs are symmetric in that the low quality part is either directly after the initial high quality or before the ending high quality. Thus, they can be used to assess the impact of the position of the intermediate layer.

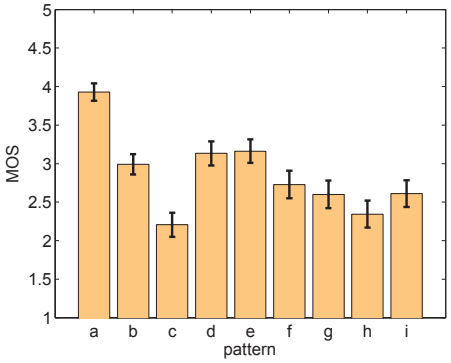


Fig. 2: MOS and 95 % confidence intervals of overall quality. Each test condition was rated by 73 different users.

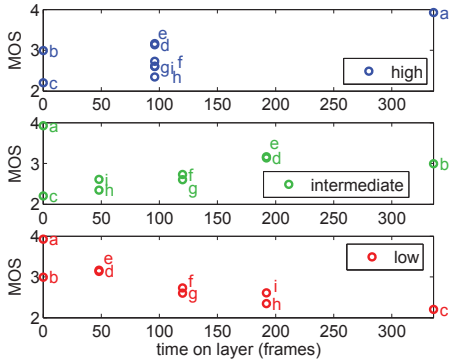


Fig. 3: MOS values of the tested adaptation profiles with respect to the time on each layer.

We chose a popular movie scene[1], which was available on YouTube, as content for the experiment. The selected sequence from that scene is 14 seconds long (336 frames) and shows a monologue of one character in front of another in a living room. The original 720p (YouTube *itag* 136) sequence was decoded to individual uncompressed pictures (YUV). Afterwards, the single pictures were encoded according to the specified patterns and the three quality levels 1 (low), 2 (intermediate), and 3 (high) using the codec H.264/AVC (libx264). The target resolutions for all three quality levels is 640x480, i.e., the size of the player window in the crowdsourcing study. We choose 28 as QP parameters for quality level 3, 36 for level 2, and 44 for level 1. The final test sequences include audio and the sizes of the sequences range from 444 KB for the constant high quality level to 266 KB for the constant low quality level. We computed the SSIM and PSNR values with quality level 3 as reference. For SSIM, we obtained scores of 1.0, ~0.96, and ~0.90 for level 3, 2, and 1, respectively. For PSNR, we calculated scores of 100, ~36.6, and ~31.4, respectively.

IV. Numerical Results

The impact of the intermediate layer is investigated based on the filtered subjective ratings from the crowdsourcing study. We analyze the Mean Opinion Scores (MOS) of the subjective ratings and derive the influence of the position of the intermediate layer and the time on the intermediate layer. Finally, we test the applicability of temporal pooling approaches for predicting the QoE results of the crowdsourcing study.

A. Results of the Crowdsourcing Study

Based on the ratings of 73 reliable users, Fig. 2 shows the overall quality of the videos in terms of MOS and 95% confidence intervals ranging from excellent (5) to bad (1) quality. The constant quality profiles are rated as expected. The high quality profile achieves the highest MOS of 3.93, the low quality profile results in the lowest QoE (MOS: 2.21). The intermediate quality profile is rated fair (MOS: 2.99). The clear, almost linear separation of the ratings also indicates that the selected three quality layers are appropriate as they can be visibly distinguished by the participants. When taking a look at the symmetrically designed patterns, it can be seen that those conditions belonging together achieve similar MOS values (d/e, f/g, h/i). It can be concluded that the position of the intermediate layer has no significant impact on the perceived quality. Moreover, the MOS of these three pairs is different suggesting that the time on the intermediate layer has an impact on the resulting QoE.

In Fig. 3, this effect can be observed more clearly. The figure shows the MOS of each pattern depending on its time on the high layer (top), the intermediate layer (middle), and the low layer (bottom), respectively. The constant patterns a-c only provide a reference as they have all frames on the same quality layer. It follows from the top subplot that the time on the high layer is not the only influence factor because the six patterns d-i result in different MOS values. L at the time on the intermediate and low layer, it can again be seen that symmetric conditions achieve similar QoE ratings. Moreover, a positive effect of the intermediate layer can be observed as the MOS increases when the time on the intermediate layer increases. Analogously, a negative effect of the low layer is visible.

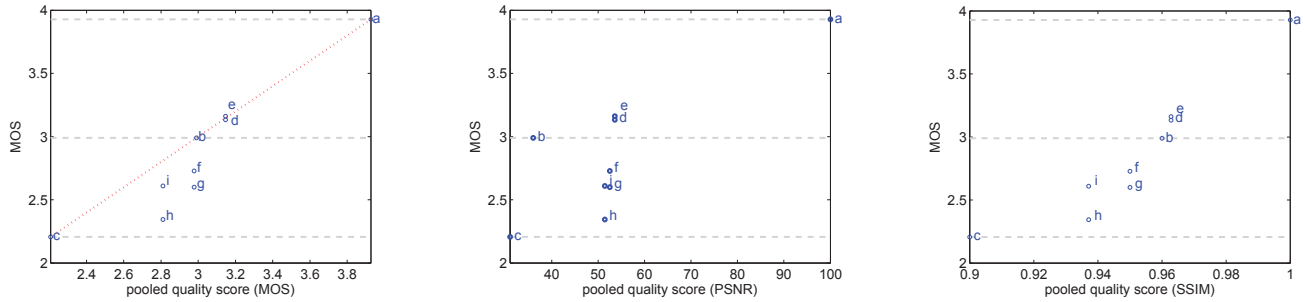
These results support the findings of [2]. We find that the MOS of the constant high and low quality profiles bounds the MOS of the adaptation patterns. Moreover, the position of each layer has no impact on the quality ratings. Finally, it could be observed that the time on each layer has an impact on QoE, and thus, is an important parameter for QoE assessment.

B. Validation of Temporal Pooling Approach

As the results of the crowdsourcing study suggest that the time on each layer has to be considered for QoE, time-weighted QoE models come into consideration. Being part of this class of models, [4] investigated temporal pooling of per-frame metrics, i.e., the aggregation of per-frame quality indicators, weighted over time, into a single quality score. They found that the simple mean of per-frame metrics already delivered a respectable prediction performance. In the following, we apply mean pooling to the crowdsourcing results in order to validate this QoE model.

First, the MOS of the constant profiles is considered as per-frame metric. Figure 4a shows a scatter plot having the pooled MOS (i.e., time average of the constant profile MOS values) on the x-axis and the MOS of the crowdsourcing study on the y-axis. It is evident from the definition of the mean pooling that the constant profiles a-c lie on the bisecting line. The scatter plot indicates that the mean pooling achieves a good MOS prediction reaching a high Spearman rank order correlation coefficient (SROCC) of 0.9537. Although also the

[1] https://www.youtube.com/watch?v=IFSAbxfLBYU, 1:05 to 1:19



(a) Mean pooling of MOS of constant profiles. (b) Mean pooling of PSNR values. (c) Mean pooling of SSIM values.

Fig. 4: Scatter plots of temporal pooling (x-axis) and MOS of crowdsourcing study (y-axis).

Pearson linear correlation coefficient (PLCC) is high (0.9365), it can be seen from the distances of d-i to the bisector that the mean pooling is not yet highly accurate. Instead, this simple temporal pooling overestimates the actual MOS ratings of the profiles. Thus, a better pooling function has to be utilized or a mapping of the pooled quality scores to MOS values (i.e., a QoE model function) has to be applied in order to achieve a more accurate QoE prediction.

Finally, two objective quality metrics are used as per-frame metric for the mean pooling. Fig. 4b shows a scatter plot of MOS versus time average PSNR values for each adaptation profile. The plot visualizes that the pooled PSNR values (PLCC: 0.7974, SROCC: 0.7849) are not suitable for QoE prediction as the constant intermediate layer profile b is rated worse than all non-constant profiles. Additionally, the pooling of PSNR is not capable of clearly separating the non-constant profiles d-i. The analogous scatter plot of mean pooling of SSIM is depicted in Fig. 4c. Compared to Fig. 4b, it can be seen that SSIM clearly outperforms signal error based PSNR, as it was especially designed to incorporate human eye perception. The usage of SSIM results in a much better QoE prediction (PLCC: 0.9383, SROCC: 0.9537) comparable to the pooling of MOS values of constant quality profiles. Again, future work has to find appropriate QoE model functions to obtain an accurate MOS prediction. Nevertheless, a well-performing automated QoE evaluation based on objective per-frame metrics is conceivable. Another advantage of the usage of SSIM for temporal pooling compared to the MOS approach could be that it is an instantaneous per-frame metric, whereas MOS is a single retrospective score for an entire quality layer. Thereby, SSIM can possibly comprise quality fluctuations within a layer, which are neglected by the MOS approach.

All in all, temporal pooling proved suitable for QoE prediction. Confirming the findings of [4] for mean pooling, aggregating quality metrics over time resulted in a high correlation with subjective ratings. An accurate pooling or QoE model function remains an open issue for future research. However, as temporal pooling incorporates the main QoE influence parameters, amplitude (quality metric) and time on each layer (aggregation over time), it should be considered for future QoE models and could also facilitate automated QoE evaluation frameworks for QoE-aware traffic management.

V. Conclusions and Future Work

In this work, we investigated the impact of an intermediate quality layer on the subjectively perceived quality of adaptive video streaming with a crowdsourcing study. The results indicate that the intermediate layer interacts with all other layers, and thus, influences the subjectively perceived quality. Thereby, the position of the intermediate layer has no significant impact on the QoE. The time on the high layer is not the only QoE influence factor, but a positive effect of the intermediate layer and a negative effect of the low layer were visible. It follows that the quality of each layer and the time on each layer are important QoE parameters.

These findings support temporal pooling models, in which per-frame metrics (considering layer quality) are weighted over time (considering time on each layer) and aggregated into a single quality score. To validate this approach, we applied a simple mean pooling of different metrics (MOS of constant profiles, PSNR, SSIM) and compared the scores to the crowdsourcing MOS. We find that the quality scores for MOS of constant profiles and SSIM have high correlations with the collected subjective ratings. However, it was obvious that mean pooling is not perfectly accurate. Thus, a better pooling or additional QoE model function has to be found in future research to obtain more accurate QoE predictions based on temporal pooling. Nevertheless, the temporal pooling approach proved suitable for the application in QoE-aware traffic management solutions. An automated QoE evaluation framework could monitor the requested quality layers (i.e., the client's HTTP requests for chunks) or utilize in-network quality modules (e.g., SSIM computation as network function) to obtain objective per-frame/per-chunk metrics. Pooling these metrics results a decent QoE prediction, which could provide a basis for QoE-aware traffic management decisions.

To sum up, this work provides valuable insights into the QoE of currently deployed HAS technology, which mainly relies on multiple quality layers. The most important QoE parameters, i.e., layer quality and time on each layer, could be identified. This supports the applicability of temporal pooling for QoE prediction and a simple mean pooling of per-frame metrics could be compared to the obtained subjective ratings. These findings have to be extended and consolidated by further evaluations and studies in order to create a holistic QoE model for multiple layer HTTP adaptive streaming. Moreover, the integration of the proposed temporal pooling methodology for automated QoE evaluation into traffic management solutions has to be investigated and implemented.

Acknowledgements

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grants HO4770/1-2 and TR257/31-2 (OekoNet) and in the framework of the EU ICT Projects SmartenIT (FP7-2012-ICT-317846) and INPUT (H2020-2014-ICT-644672). The authors alone are responsible for the content.

REFERENCES

[1] International Standards Organization/International Electrotechnical Commission (ISO/IEC), "23009-1:2012 Information Technology – Dynamic Adaptive Streaming over HTTP (DASH) – Part 1: Media Presentation Description and Segment Formats," 2012.

[2] T. Hoßfeld, M. Seufert, C. Sieber, and T. Zinner, "Assessing Effect Sizes of Influence Factors Towards a QoE Model for HTTP Adaptive Streaming," in *Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX 2014)*, Singapore, 2014.

[3] S. Rimac-Drlje, M. Vranjes, and D. Zagar, "Influence of Temporal Pooling Method on the Objective Video Quality Evaluation," in *Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB'09)*, Bilbao, Spain, 2009.

[4] 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," in *Proceedings of the 5th International Workshop on Quality of Multimedia Experience (QoMEX 2013)*, Klagenfurt, Austria, 2013.

[5] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[6] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen, "Flicker Effects in Adaptive Video Streaming to Handheld Devices," in *Proceedings of the 19th ACM International Conference on Multimedia (MM 2011)*, Scottsdale, AZ, USA, 2011.

[7] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, and P. Le Callet, "Subjective Quality of SVC-coded Videos with Different Error-patterns Concealed Using Spatial Scalability," in *Proceedings of the 3rd European Workshop on Visual Information Processing (EUVIP 2011)*, Paris, France, 2011.

[8] M. Zink, J. Schmitt, and R. Steinmetz, "Layer-encoded Video in Scalable Adaptive Streaming," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 75–84, 2005.

[9] M. Grafl and C. Timmerer, "Representation Switch Smoothing for Adaptive HTTP Streaming," in *Proceedings of the 4th International Workshop on Perceptual Quality of Systems (PQS 2013)*, Vienna, Austria, 2013.

[10] D. C. Robinson, Y. Jutras, and V. Craciun, "Subjective Video Quality Assessment of HTTP Adaptive Streaming Technologies," *Bell Labs Technical Journal*, vol. 16, no. 4, pp. 5–23, 2012.

[11] S. Tavakoli, J. Gutiérrez, and N. Garcia, "Subjective Quality Study of Adaptive Streaming of Monoscopic and Stereoscopic Video," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 4, pp. 684–692, 2014.

[12] S. Tavakoli, K. Brunnström, K. Wang, B. Andrén, M. Shahid, and N. Garcia, "Subjective Quality Assessment of an Adaptive Video Streaming Model," in *Proceedings of IS&T/SPIE Electronic Imaging Conference on Image Quality and System Performance XI*, San Francisco, CA, USA, 2014.

[13] S. Egger, B. Gardlo, M. Seufert, and R. Schatz, "The Impact of Adaptation Strategies on Perceived Quality of HTTP Adaptive Streaming," in *Proceedings of the 1st Workshop on Design, Quality and Deployment of Adaptive Video Streaming (VideoNext 2014)*, Sydney, Australia, 2014.

[14] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[15] B. J. Villa, K. De Moor, P. E. Heegaard, and A. Instefjord, "Investigating Quality of Experience in the Context of Adaptive Video Streaming: Findings from an Experimental User Study," in *Proceedings of the Norsk Informatikkonferanse (NIK 2013)*, Stavanger, Norway, 2013.

[16] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of Experience Estimation for Adaptive HTTP/TCP Video Streaming Using H. 264/AVC," in *Proceedings of the 2012 IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, NV, USA, 2012.

[17] C. Alberti, D. Renzi, C. Timmerer, C. Mueller, S. Lederer, S. Battista, and M. Mattavelli, "Automated QoE Evaluation of Dynamic Adaptive Streaming over HTTP," in *Proceedings of the 5th International Workshop on Quality of Multimedia Experience (QoMEX 2013)*, Klagenfurt, Austria, 2013.

[18] J. De Vriendt, D. De Vleeschauwer, and D. Robinson, "Model for Estimating QoE of Video Delivered Using HTTP Adaptive Streaming," in *Proceedings of the IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*, Ghent, Belgium, 2013.

[19] M.-N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnström, and A. Raake, "Quality of Experience and HTTP Adaptive Streaming: A Review of Subjective Studies," in *Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX 2014)*, Singapore, 2014.

[20] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd: A Framework for Crowd-based Quality Evaluation," in *Proceedings of the Picture Coding Symposium (PCS 2012)*, Krakow, Poland, 2012.

[21] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.