(Not) the sum of its parts: relating individual video and browsing stimuli to web session QoE

Johannes Schleicher, Nikolas Wehner, Tobias Hoßfeld, Michael Seufert

Angaben zur Veröffentlichung / Publication details:

Schleicher, Johannes, Nikolas Wehner, Tobias Hoßfeld, and Michael Seufert. 2024. "(Not) the sum of its parts: relating individual video and browsing stimuli to web session QoE." In *2024 16th International Conference on Quality of Multimedia Experience (QoMEX), June 18-20, 2024, Karlshamn, Sweden*, edited by Markus Fiedler, Lea Skorin-Kapov, Hadi Amirpour, and Karel Fliegel, 104–10. Piscataway, NJ: IEEE.
https://doi.org/10.1109/QoMEX61742.2024.10598239.

2024 16th International Conference on Quality of Multimedia Experience (QoMEX)

(Not) The Sum of Its Parts: Relating Individual Video and Browsing Stimuli to Web Session QoE

Johannes Schleicher*, Nikolas Wehner†, Tobias Hoßfeld†, Michael Seufert*
*University of Augsburg, Institute of Computer Science, Augsburg, Germany
†University of Würzburg, Institute of Computer Science, Würzburg, Germany
{johannes.schleicher | michael.seufert}@uni-a.de, {nikolas.wehner | tobias.hossfeld}@uni-wuerzburg.de

***Abstract*—The integration of web and video applications as dominant Internet content has underscored the importance of Quality of Experience (QoE) for user satisfaction, retention, and digital service success. While current research has extensively studied QoE for individual stimuli, such as web page loading or video streaming, there exists a significant gap in understanding and quantifying QoE for mixed web browsing and video streaming sessions. This paper addresses the critical need to evaluate session QoE when web and video stimuli are combined within a single web session. Employing a crowdsourcing methodology, we investigate the impact of session length, content type, and individual stimuli QoE on the overall session QoE through a full factorial design with both unimpaired and impaired stimuli. Based on these results, we evaluate the applicability of various models to accurately estimate session QoE from information about individual stimuli, offering insights into optimizing the subjective experience in web sessions.**

***Index Terms*—Quality of Experience; Web Session QoE; Video Streaming QoE; Web Browsing QoE; Session QoE Model**

I. Introduction

The digital era is driven and self-reinforced by the increasing availability and usage of web and video applications, establishing them as the most dominant forms of content on the Internet. With videos seamlessly integrating into web pages, both web browsing and video streaming can be conveniently used from a web browser, even within the same web session. The ever-increasing popularity of both types of applications on the Internet has also revealed the importance of understanding and optimizing the Quality of Experience (QoE) for users, a crucial metric for network and service providers [1]. QoE directly influences user satisfaction, retention, and the overall success of digital services, making it a pivotal aspect in the competitive web and video service industry.

Despite the obvious importance of QoE, current research predominantly focuses on the assessment of QoE for individual stimuli, such as a single web page load or a single video stream. For web browsing, impact factors like the loading time and loading behavior of web pages are well understood [2], and similar results exist for video streaming, where factors such as initial delay, stalling, and quality adaptation play a crucial role [3]. The understanding of these individual elements is complemented by well-established QoE metrics, such as page load time (PLT) for web browsing [4] or the number of stalling events for video streaming [5]. Moreover, several QoE models exist - even standardized models like G.1030 [6] for web browsing or P.1203 [7] and P.1204 [8] for video streaming - which allow to map QoE metrics to a Mean Opinion Score (MOS), and thus, guide providers in enhancing the subjective experience of their end users.

979-8-3503-6158-2/24/$31.00 ©2024 IEEE

However, a critical gap exists in our understanding of session QoE, particularly when different types of stimuli are mixed within a session, as is typically the case in normal web usage. Not only was it shown that QoE models of web and video stimuli do not well align in mixed sessions [9], but it also still needs to be researched how the QoE scores of individual stimuli can be combined and aggregated to obtain a single QoE score for the entire session.

In this work, we conduct a QoE study to investigate the QoE in mixed web browsing and video streaming sessions using a crowdsourcing approach. In particular, we apply a full factorial design with unimpaired and impaired web and video stimuli using two different session lengths. This allows us to evaluate the impact of the session length as well as position, content type, and QoE of the individual stimuli on the overall session QoE. Finally, we investigate the applicability of several models to accurately estimate the resulting session QoE from the QoE scores of the individual stimuli, and discuss our results.

Therefore, this work is structured as follows. Section II outlines related works on web and video QoE as well as on the QoE of sessions. Section III describes the implemented crowdsourcing study, the data filtering, and the resulting dataset. Section IV presents the results on session QoE and evaluates the performance of selected session QoE models. Finally, Section V summarizes the findings and concludes.

II. Background and Related Work

The Quality of Experience (QoE) of video streaming mostly depends on initial delay, stalling, and quality adaptation [3]. Stalling or rebuffering, i.e., playback interruptions due to buffer depletion, is considered the worst QoE degradation [5], [10], [11], and should be avoided. Furthermore, video streams should be played out with high visual quality [12]. In contrast, initial delay has a smaller impact on the QoE [4].

A large number of QoE models was proposed in literature, e.g., [3], [13], including the standardized P.1203 [7] and P.1204 [8] models. These models typically consider a set of different QoE factors, such as the total stalling length, the

number and duration of stalling events, the visual quality, number and amplitude of quality switches, and initial delay. In [14], the authors showed that many of those QoE models perform significantly different as they attach different weights to these QoE influence factors. Also artificial intelligence (AI) and machine learning (ML) are already widely used for video QoE modelling [15]. In this context, however, it was proposed to prefer explainable AI (XAI) over black-box ML models for QoE modelling [16].

With respect to QoE factors of web browsing, many studies found that response times (i.e., waiting times) are the most important QoE factor, as initially highlighted in [17]. Consequently, early models for web QoE mainly focused on page load time (PLT) [4]. To provide a more fine-granular temporal assessment of waiting times, a multitude of time instant metrics, e.g., Above the Fold (ATF) [18], and time integral metrics, e.g., Google's SpeedIndex (SI) [19], have been proposed, studied, and used for QoE models since then [2]. Most recently, Google proposed Core Web Vitals [20], which emphasize the overall user experience with web pages, however, were found to not well correlate to web QoE [21].

Traditional web QoE models are usually based on the IQX and WQL hypotheses. While the IQX hypothesis assumes an exponential relationship between waiting time and web QoE [22], the WQL hypothesis assumes a logarithmic relationship on a linear ACR scale [23]. Also, the standardized model G.1030 [6] relies on a logarithmic relationship. Recently, more complex methods to model web QoE were presented, including also machine learning-based approaches [24]. [2], [25] provided a comprehensive overview over waiting time based studies and models.

Despite this plethora of works on the QoE of individual web or video stimuli, there is much less work considering the QoE of a longer session with multiple, diverse stimuli. Considering the aggregation of multiple subsequent QoE ratings into an overall score, [26] investigated the memory effect for web QoE. They identified transient effects when the service quality decreased over a sequence of stimuli, as well as an influence of the preceding quality levels on the rating of the current stimuli. They proposed to use an Iterative Exponential Regression Model (IERMo) to model the evolution of the QoE over the course of a multi-stimuli web session. [9] used the IERMo model on mixed sessions with both video and web stimuli, and found that standard video and web QoE models are incompatible and might not be able to appropriately map the network conditions to MOS for mixed sessions. Also related, [27] studied the QoE of videos composed of 1-3 different scenes and assumed that the QoE updates in a non-linear and asymmetric fashion after every content, i.e., scene. Similarly, [28] proposed a cumulative video QoE model, however, without explicitly studying the impact of video content changes during model updates. To the best of our knowledge, the only study on a combination of web and video stimuli was conducted in [29]. They found that low web page load times did not affect the QoE of a subsequent video, concluding that users expected short delays when browsing to a video. Apart from these works, also general principles and cognitive biases from psychology should apply to session QoE, such as the anchoring effect [30] and the peak-end rule [31].

III. Methodology

For our session QoE study we developed a customized crowdsourcing framework using jsPsych [32], a JavaScript library designed for behavioral experiments in web browsers. This framework empowers us to seamlessly conduct experiments across a diverse participant pool, facilitating comprehensive data collection for crowdsourcing studies. The framework boasts a versatile array of plugins that can be effortlessly tailored to meet specific experiment requirements. Moreover, if needed, we have the flexibility to develop additional custom plugins. In our study, we fine-tuned the framework to adhere to crowdsourcing best practices [33] while retaining complete control over the presentation of web pages and video playback. Following a top-down approach, we simulate various PLTs and stalling events, ensuring the experiment's autonomy from individual participants network conditions by manipulating DOM elements and adjusting video playback timings.

In this study, we aim to investigate how individual stimuli influence the overall QoE during a session. Participants actively engage in sessions that encompass both web browsing and video streaming. Mixed sessions mirror typical online experiences, given that videos are nowadays even commonly embedded in standard webpages. Our examination focuses on the PLT for individual web pages, using a custom creation of a news and shopping page, under two scenarios: a fast PLT of 1 second (`W0`) and a slow PLT of 10 seconds (`WS`), representing diverse loading speeds. The video component, which consists of a 30 second excerpt from Big Buck Bunny, undergoes also two conditions: one with no stalling event (`V0`) and another with 3 uniformly distributed 4 seconds stalling events (`VS`). We propose the following naming convention for our study: '`V0`' represents a video without stalling, '`VS`' for a video with stalling, '`W0`' for a web with fast Page Load Time (PLT) of 1 second, and '`WS`' denotes a web with slow PLT of 10 seconds. The notations '`V/W`' indicate content type (video/web), and '`0/S`' describe the degradation type, where '`0`' signifies undisturbed (`V`: no stalling, `W`: fast PLT), and '`S`' indicates stalling (3 stalling events) or slow PLT (10 seconds). To encompass a wider range, we also investigate the influence of the stimuli on the session MOS for two distinct session lengths: three stimuli (base) and nine stimuli (long). For simplicity, the long session is created by repeating the short session three times. Therefore, by creating mixed sessions through the combination of two web stimuli with one video stimulus or vice versa, the experimental design encompasses a total of 48 unique sessions. This constitutes a factorial design, systematically exploring the impact of the examined conditions on the overall online session experience.

We published our study on Microworkers[1] to take advantage of a large-scale crowdsourcing platform. Initially, participants

[1]https://www.microworkers.com/

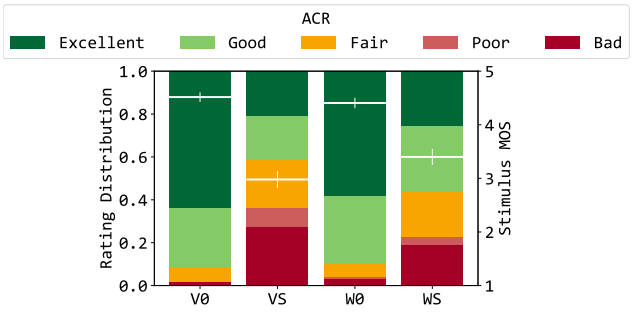


Fig. 1: Distribution of different stimuli ratings

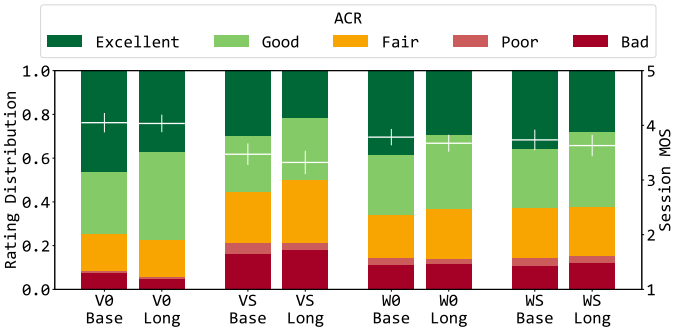


Fig. 2: Base vs Long: Distribution of different session ratings

were familiarized with the study and test content through clear instructions and illustrative stimuli. The study consisted of two parts. In the first part, the participants observed a complete session consisting of 3 or 9 stimuli and gave a comprehensive evaluation of the entire session. Subsequently, in the second part, the individual stimuli within the session were presented one after the other so that the participants could rate each stimulus independently. Ratings are based on a 5-point Absolute Category Ranking (ACR) scale (bad, poor, fair, good, excellent). Additionally, participants were required to memorize a letter included in a video or interact with the presented webpage as a simple validation check to ensure they had actually engaged with the content. Also, if a user exits the study tab, their participation is flagged as invalid. This precaution is taken because it cannot be guaranteed that the participant engaged thoughtfully in the study under such circumstances. A total of 2552 users participated in the crowdsourcing study. After excluding participants that did not pass the validation check, switched the study tab, and those who participated multiple times, we retained 613 valid runs. Among these, 329 runs are dedicated to the base case. In average, each of the 48 unique sessions got rated 6.8 times, while the median remains at 6.5. The number of valid ratings ranged from 1 to 15 per run. For the long case, we retained 284 runs. These sessions included a minimum of 2 and a maximum of 12 ratings per session. while the median and average number of participants per unique session is 5.0 and 5.9 respectively. In addition to collecting stimuli and session ratings, our study sought valuable insights into participants backgrounds by gathering personal information. Participants willingly shared details such as gender (male, female, other), origin, education level, and Internet usage patterns. Examining the gender distribution, we found that 60% of participants identified as male, 39% as female, and 1% chose 'other'.

IV. Evaluation

A. Influence of Web and Video QoE on Session QoE

Before assessing the session QoE, Figure 1 investigates the QoE of the individual stimuli, that comprise the sessions. The x-axis represents the investigated stimuli, while the y-axis shows the distribution of ratings and the corresponding Mean Opinion Score (MOS) for each stimulus. The rating distribution is shown as a percentage of participants who rated the stimulus, with the session MOS and its associated 95% confidence interval given by a white line within each bar. Notably, users consistently assign lower ratings to disrupted stimuli compared to their undisturbed counterparts. This distinct disparity underscores the robustness of the employed methodology. On a broader scale, we observe that video stallings lead to a reduced MOS, similar to the effect observed with an extended PLT during web browsing. This trend is further supported by the MOS values. Specifically, the MOS for video stalling (VS) lies at 2.98 (95% CI: 2.83, 3.13), contrasting with 4.52 (95% CI: 4.44, 4.60) for video without stalling (V0). A similar trend is observed in the web stimuli, where the MOS is 3.4 (95% CI: 3.26, 3.54) for slow PLT (WS) and 4.41 (95% CI: 4.32, 4.49) for fast PLT (W0). Notably, V0 and W0 applications exhibit no significant difference, as demonstrated by the overlap of the respective confidence intervals. Conversely, for the degree of degradation in this study, users see disruptions in video stimuli more bothersome than slow PLT in web browsing, as shown by a clear MOS difference of 0.44 and no overlapping confidence interval.

After we found out that degradation of stimuli results in deteriorated MOS values, the subsequent analysis delves into a more detailed examination of the influence of individual stimuli on the overall MOS. Figure 2 illustrates the session MOS when a specific stimulus is included in the base session or the long session. Consistent trends are observed for each session length. Degrading the video content is clearly associated with lower overall satisfaction, evident in the session MOS for both session lengths: 4.04 for V0 (Long: 4.03) and 3.47 for VS (Long: 3.32). In contrast, for the web browsing stimuli, such a direct correlation is not recognizable. The session MOS for base sessions is 3.78 for W0 (Long: 3.67) and 3.73 for WS (Long: 3.63). Consequently, it can be inferred that a high-quality fluid video positively influences user satisfaction, while video disruptions have a negative effect. However, no such discernible influence is noticed for the web browsing stimuli based on the presented plots. Upon examining the MOS per session for a single stimulus across various session lengths, the current obtained results indicate, that base sessions have a slightly higher MOS compared to long sessions.

In the subsequent step, the influence of the session length on the overall session MOS is examined in more detail. Figure 3 presents the average ratings across all 48 unique

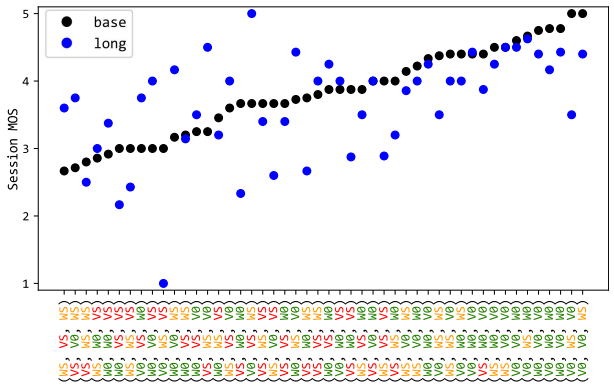


Fig. 3: Base vs Long: Average rating per stimulus

viewed sessions for both session lengths. Black dots represent base sessions, while blue dots correspond to long sessions. The y-axis indicates the MOS, and the x-axis shows the sessions sorted based on the MOS score for the base sessions. Additionally, the single stimuli of the corresponding session are color-coded, allowing investigations into whether the MOS score increases when undisturbed stimuli are included in the session. The `W0` and `V0` stimuli are colored green, `WS` stimuli are colored orange, and `VS` stimuli are colored red. Upon investigating the influence of session length on the MOS score, it is evident that in 30 of the 48 considered sessions, base sessions result in higher MOS scores. In 16 sessions, the long session receive higher ratings, with the remaining 2 sessions being equally rated. Consequently, it seems like participants experience slightly higher satisfaction when viewing base sessions compared to long sessions, which underlines the previous indications based on Figure 2. Furthermore, an investigation into the effects of both undisturbed and disrupted stimuli on the overall session MOS can be conducted. A thorough analysis of the coloring of the stimuli reveals that sessions with a higher prevalence of undisturbed stimuli generally result in higher MOS values. Conversely, degraded stimuli are primarily concentrated on the left side of the figure. In the top 17 sessions, only a single session includes a `VS` stimulus, while in each of the lowest 10 rated sessions, at least one `VS` is present. This indicates a noticeable trend, suggesting that the `VS` stimulus significantly influences the overall user experience. For the slow web stimulus `WS` it also seems like most of the sessions with at least one `WS` stimulus are contained in the bottom half. However, a closer look reveals that almost a third (9 of 30) of `WS` containing sessions appear in the top third. The graph also implies that bad stimuli at the end reduce the MOS, as these sessions are mainly in the lower half (17 of 24).

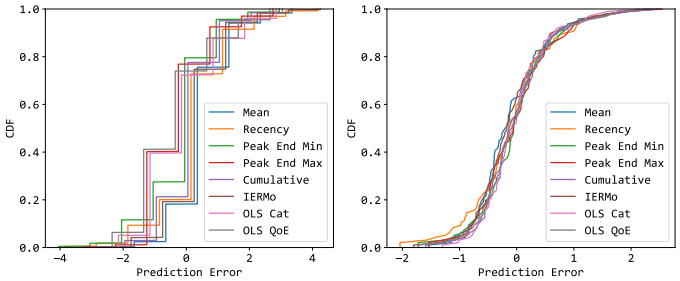
B. Statistical Investigation

To delve deeper into our investigation of session QoE and validate our findings, we employ statistical tests on the dataset. As multiple tests are performed, the single p-values are corrected using the Holm-Sidak correction where required. To enhance the robustness of our results, we once oversampled the individual participant responses, ensuring an equal number of ratings for all sessions. Subsequently, the oversampled data points are utilized in the statistical tests. Additionally, to maintain the integrity of the comparisons, redundant data points shared across multiple sets undergoing statistical tests are systematically eliminated, whenever feasible, to maintain the integrity of the analyses. All conducted tests with their corresponding results are given in Table I. If the difference between the groups is significant, it is highlighted in color, and the significance test result is mentioned after the corresponding p-value if it is significant.

TABLE I: Statistical investigation of session length and positional stimuli influence on session MOS

Influence	Test	Result - Base	Result - Long
Base vs Long	MWU	$p = 0.0356$ Base>Long	
Position `V0`	KW	$p = 0.0709$	$p = 0.5060$
Start vs Middle	MWU	$p = 0.0846$	$p = 0.5614$
Start vs End	MWU	$p = 0.3398$	$p = 0.7110$
Middle vs End	MWU	$p = 0.2498$	$p = 0.7334$
Position `VS`	KW	$p = 0.1036$	$p = 0.0001$
Start vs Middle	MWU	$p = 0.6454$	$p < 10^{-3}$ M>S
Start vs End	MWU	$p = 0.3470$	$p = 0.8205$
Middle vs End	MWU	$p = 0.0766$	$p < 10^{-3}$ M>E
Position `W0`	KW	$p = 0.0036$	$p = 0.3605$
Start vs Middle	MWU	$p = 0.0031$ M>S	$p = 0.3278$
Start vs End	MWU	$p = 0.0895$	$p = 0.1736$
Middle vs End	MWU	$p = 0.1536$	$p = 0.6228$
Position `WS`	KW	$p = 0.3505$	$p < 10^{-5}$
Start vs Middle	MWU	$p = 0.3872$	$p = 0.0589$
Start vs End	MWU	$p = 0.5444$	$p = 0.0060$ E>S
Middle vs End	MWU	$p = 0.7718$	$p < 10^{-5}$ E>M

To evaluate the influence of session length on the session MOS, the Mann-Whitney-U (MWU) test is employed. The resulting p-value of $p = 0.0356$ indicates a significant difference, signifying that base sessions receive higher ratings when compared to their longer counterparts. In addition to the impact of session length on the session MOS, our analysis reveals that the video performance significantly influences the session MOS. Further investigation is conducted with the Kruskal-Wallis (KW) test to explore whether the placement of individual stimuli within a session contributes to variations in the session MOS. For `V0`, no significant influence is observed. The same applies to `VS` for base session, while for long sessions the KW test identifies a significant difference with a p-value of $p = 0.0001$. Subsequent pairwise MWU tests demonstrate that the session MOS is lower when the `VS` stimulus is placed at the start ($p < 10^{-3}$) or end ($p < 10^{-3}$) of the session compared to the middle. In comparison to the video stimuli, a significant difference is evident for the `W0` stimulus. For the base session length, the KW test reveals a significant difference, with a p-value lower than $p = 0.0036$, emphasizing the importance of positional placement. The pairwise MWU test results indicate that the session MOS is lower when the `W0` stimulus is presented at the start position ($p = 0.0031$) compared to the middle position. Regarding the long sessions, no significant difference could be discovered. Turning to the `WS` stimulus, no positional influence is evident for base sessions, whereas for long sessions, the p-value is



(a) Individual session rating (b) Session MOS

Fig. 4: MOS estimation models

less than $p < 10^{-5}$, warranting a pairwise positional analysis. The MWU test demonstrates that the WS stimulus significantly increases the session MOS when placed at the end. The associated p-values are $p = 0.0060$ for the start and less than $p < 10^{-5}$ for the middle position when compared to the end position, respectively. The positional influence of the web stimuli on session MOS is a bit surprising given the trend from Figure 3, where it appeared that the position and PLT of the website had no such influence on session MOS as identified here. The statistical tests confirm the significant difference for the different session lengths. However, as the positional influences are different for both session lengths, no further general conclusion can be drawn here.

C. Adaptation of Existing QoE Models

To estimate the overall session QoE, eight different model types are examined. The mean model calculates the mean of the individual stimuli scores. The recency model, as its name implies, uses the score of the last presented stimulus as the session MOS. The peak-end model considers the peak score, either the minimum or maximum value, and the last score. Both possible minimum and maximum peaks are examined in our analysis. We further select a cumulative model based on [28], which calculates the session score based on the minimum, average, and last score. The Iterative Regression Model (IERMo), as described in [26] can be used to estimate the MOS of a session. Originally designed for sessions involving web browsing exclusively, this model requires adaptation, as proposed in [9], to use it for our studied sessions. The modified IERMo model processes a sequence of scores from a given session to estimate the session score. It uses exponential regression to simulate a decreasing weight over successive stimuli with similar scores. However, if a stimulus with a different score occurs, the current rating significantly influences the session score.

We also use two linear models to predict the session score. These models are fitted using ordinary least squares (OLS) regression, once with the conditions of the stimuli at each position in the session as categorical parameters (OLS Cat) and once with the MOS output of QoE models specific to each stimulus as a parameter (OLS QoE), respectively. Both models also consider session length as a categorical parameter. To derive the MOS using QoE models, we apply P.1203 for video stimuli and the WQL PLT model for web sessions, as detailed in [7] and [23], respectively.

The eight introduced model types can now be applied for two tasks: 1. prediction of the individual session rating of a user based on his or her ratings of the stimuli in that session 2. prediction of the session MOS, i.e., the average session rating over all users, using the QoE scores of the stimuli in that session. Note that the OLS models will output the same session scores for both tasks as they only take session characteristics as input. This means they do not explicitly consider stimuli ratings for the prediction although the ratings were implicitly considered during model fitting.

First, we will investigate the performance of the models for the prediction of the individual session rating based on the ratings of the stimuli within the session. Figure 4a shows the cumulative distribution functions (CDF) that visualize the distributions of the prediction error for the described models on the x-axis. The prediction error is calculated by the prediction rating minus the actual session one, so that a positive prediction error corresponds to an overestimation and a negative one to an underestimation. In our study, each participant rated the QoE of the session on the ACR scale, requiring our models to predict categorical session ratings. However, as the model outputs can be continuous values, rounding is applied where necessary. To prevent graphs from overlapping, we decided to apply a small horizontal shift to each CDF, so that all prediction error distributions are clearly visible. Considering the results obtained from Figure 4b, it is noticeable that both OLS models and the peak-end-max model are the least likely to correctly estimate the session score, while the increase is maximal for the cumulative, IERMo, and mean model. To gain a better overview of all models, the Mean Absolute Error (MAE) is calculated and compared between the single models. The MAE is lowest for the mean model (0.57) followed by IERMo (0.60), cumulative (0.61), peak-end-min (0.71), recency (0.71), peak-end-max (0.74), while for both OLS models it is above 0.9.

We now consider the task of predicting the session MOS, i.e., average rating of the session, from the QoE scores of the individual stimuli. Figure 4b presents the corresponding results for the different models for all of the 96 different session types. The results show that all models perform similarly, except that the CDF for the recency and mean model are slightly shifted to the left. Consequently, MAE of the individual models is examined. Here, the OLS Cat model has the lowest MAE of 0.43 followed by the OLS QoE (0.48), cumulative (0.48), peak-end-min (0.49), IERMo (0.50), mean (0.52), peak-end-max (0.52), and recency (0.56) model. Considering the prediction errors of the models in both tasks, the mean model seems to be the best choice, as it achieved the lowest MAE in predicting the individual session rating and also provides a low MAE for the second task of predicting the session MOS. If, however, the focus is on session MOS estimation only, the OLS Cat model performed best. Still, as no model performed particularly better than the others for session MOS prediction, it suggests that more studies are required to further investigate

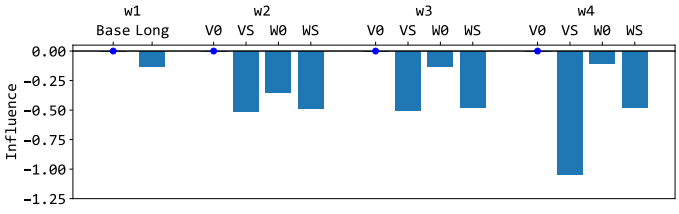


Fig. 5: Fitted parameters of OLS Cat model (intercept: 4.81)

the influence of the individual stimuli of a session as well as the characteristics of the session as a whole. As the OLS Cat model performed best for the session MOS prediction, this model will be subject to further analysis, as it can also be used to carry out further investigations, i.e., it allows to explore the influence of parameters on the session MOS using analysis of variance (ANOVA). The p-values associated for all stimuli positions are low, and thus significant, with the p-value for the start stimulus being the largest at $p < 10^{-8}$. To quantify the effect sizes, we can calculate η^2 based on the proportion of variance in the MOS estimation from the single parameters. The corresponding η^2 values are 0.002 for session length, 0.025 for the category of the start, 0.029 for the category of the middle and 0.103 for the category of the end stimulus of the session. According to the η^2 effect size classification in [34], the start and middle position have a small effect, while the the end position has a medium effect on the session MOS. This result is surprising in that it suggests that the end stimulus has a significant effect on the session MOS and therefore the peak-end model should predict the session MOS fairly accurately, although its estimate is worse than for the OLS model.

Furthermore, by examining the coefficients of the fitted OLS model, we can gain insights into the individual contributions of each parameter to the session MOS. Figure 5 illustrates the parameters of the OLS Cat model, with its formula given in Equation 1. The intercept corresponds to a baseline that represents the predicted value of the session MOS when all variables are set to zero. In our equation the baseline corresponds to a base session with only `V0` stimuli, although such a session is absent in our study. sl corresponds to the categorical session length indicator, which has the value 0 for base and 1 for long sessions. $cond(stimX)$ indicates which stimulus (`V0`, `VS`, `W0`, or `WS`) was at position $X \in 1, 2, 3$ inside the session, thus, adding the impact of position and stimulus type to the equation. Note that the OLS model fits only three weights for the four stimuli, as one stimulus (here: `V0`) is considered as baseline, which is identical to using a fixed weight of 0 in these cases.

$$\begin{aligned} MOS = intercept + w_1 * sl + w_2 * cond(stim1) \\ + w_3 * cond(stim2) + w_4 * cond(stim3) \end{aligned} \quad (1)$$

The fitted intercept has the value 4.81, which represents the MOS of the base session with only `V0` stimuli. All other fitted parameter values adjust this value based on their respective influences on the session MOS, as depicted in Figure 5. As all values are negative, they have a negative impact on the session MOS in the presence of the corresponding condition compared to the baseline conditions (i.e., base session length and all `V0` stimuli). This confirms the previous finding that long sessions have a negative influence on the session MOS (long: -0.13). Furthermore, the OLS model reveals that `VS` (-1.05) has a substantial negative impact on session MOS when presented as the end stimulus of a session, compared to the start (`VS`: -0.51) and middle (`VS`: -0.50) positions, where the influence is smaller and nearly identical. The influence of `WS` appears consistent across all positions, which confirms the previous observation for the base sessions, while it contradicts it for the long sessions. On the other hand, concerning `W0`, the OLS model indicates that the later the stimulus appears in the session, the less it negatively influences the session MOS (Start: -0.35, Middle: -0.14, End: -0.10). The influence of session length and the fact that `VS` and `WS` mainly worsen user satisfaction are in line with the previously obtained results. A new finding is that the stimulus `VS` as the final stimulus has an even stronger negative influence on the MOS than at the other positions.

V. Conclusion

Although various models exist to estimate the QoE of a specific web stimulus like browsing a web page or streaming a video, this is not the case for the QoE of web sessions consisting of several individual stimuli. However, it is important to close this gap and to obtain a holistic understanding how the individual stimuli, their types, and conditions impact the overall web session QoE to be able improve the overall QoE on the Internet. In this work, we took the first steps on this path by conducting a crowdsourcing study on the overall QoE of mixed web sessions containing both undisturbed and degraded web browsing and video streaming, which is a typical situation for web users. Specifically, we investigated whether the session length, the stimulus type and the position of the stimuli within the session have an influence on the session MOS.

Considering the session length, it was found that short sessions are overall better rated than longer sessions. Regarding the influence of the stimuli which compose the session, we found that the position and the QoE of the individual stimuli had an impact on the session QoE, in particular, degraded stimuli at the end lower the MOS the most. However, simple, generic relationships could not be derived and thus require follow-up studies in the future.

Since it is important to predict the user satisfaction in sessions, different models types were considered and fitted to the ratings obtained from our study. Here, simple averaging of the individual ratings worked best to predict the individual session ratings, while an OLS regression model (OLS Cat) could most accurately predict the session MOS. However, it became evident that no model was completely convincing, underscoring the necessity for further research, given the limited focus of this study. Still, this work provides initial results on the important topic of session QoE, which can be extended in future works, e.g., by considering a broader and more diverse range of session lengths and stimuli types.

Acknowledgement

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grant SE 3163/3-1, project number: 500105691. The authors alone are responsible for the content.

References

[1] K. Brunnström, S. A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hoßfeld, S. Jumisko-Pyykkö, C. Keimel, M.-C. Larabi *et al.*, "Qualinet White Paper on Definitions of Quality of Experience," 2013.

[2] H. Z. Jahromi, D. T. Delaney, and A. Hines, "Beyond First Impressions: Estimating Quality of Experience for Interactive Web Applications," *IEEE Access*, vol. 8, pp. 47 741–47 755, 2020.

[3] 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 COMST*, 2015.

[4] S. Egger, T. Hossfeld, R. Schatz, and M. Fiedler, "Waiting Times in Quality of Experience for Web Based Services," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, 2012, pp. 86–96.

[5] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of youtube qoe via crowdsourcing," in *2011 IEEE International Symposium on Multimedia*. IEEE, 2011, pp. 494–499.

[6] International Telecommunication Union, "ITU-T Recommendation G.1030: Estimating End-to-end Performance in IP Networks for Data Applications," 2009.

[7] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia *et al.*, "HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P. 1203: Open Databases and Software," in *ACM MMSys*, 2018.

[8] A. Raake, S. Borer, S. M. Satti, J. Gustafsson, R. R. R. Rao, S. Medagli, P. List, S. Göring, D. Lindero, W. Robitza *et al.*, "Multi-model Standard for Bitstream-, Pixel-based and Hybrid Video Quality Assessment of UHD/4K: ITU-T P. 1204," *IEEE Access*, 2020.

[9] M. Seufert, N. Wehner, V. Wieser, P. Casas, and G. Capdehourat, "Mind the (qoe) gap: On the incompatibility of web and video qoe models in the wild," in *2020 16th International Conference on Network and Service Management (CNSM)*. IEEE, 2020, pp. 1–5.

[10] D. Ghadiyaram, J. Pan, and A. C. Bovik, "A Time-varying Subjective Quality Model for Mobile Streaming Videos with Stalling Events," in *Proceedings of SPIE Applications of Digital Image Processing XXXVIII*, San Diego, CA, USA, 2015.

[11] K. Zeng, H. Yeganeh, and Z. Wang, "Quality-of-experience of Streaming Video: Interactions between Presentation Quality and Playback Stalling," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, 2016.

[12] M. Seufert, T. Hoßfeld, and C. Sieber, "Impact of Intermediate Layer on Quality of Experience of HTTP Adaptive Streaming," in *Proceedings of the 11th International Conference on Network and Service Management (CNSM)*, 2015.

[13] N. Barman and M. G. Martini, "QoE Modeling for HTTP Adaptive Video Streaming–A Survey and Open Challenges," *IEEE Access*, 2019.

[14] A. Seufert, F. Wamser, D. Yarish, H. Macdonald, and T. Hoßfeld, "QoE Models in the Wild: Comparing Video QoE Models Using a Crowdsourced Data Set," in *IEEE QoMEX*, 2021.

[15] G. Kougioumtzidis, V. Poulkov, Z. D. Zaharis, and P. I. Lazaridis, "A Survey on Multimedia Services QoE Assessment and Machine Learning-Based Prediction," *IEEE Access*, 2022.

[16] N. Wehner, A. Seufert, T. Hoßfeld, and M. Seufert, "Explainable data-driven qoe modelling with xai," in *2023 15th international conference on quality of multimedia experience (QoMEX)*. IEEE, 2023, pp. 7–12.

[17] J. Nielsen, *Usability Engineering*, 1993.

[18] J. Brutlag, Z. Abrams, and P. Meenan, "Above the fold time: Measuring web page performance visually," in *Velocity: Web Performance and Operations Conference*, 2011.

[19] Catchpoint WebPageTest.org, "Speed index," 2012, Accessed: Jan 19, 2024. [Online]. Available: https://docs.webpagetest.org/metrics/speedindex/

[20] Google web.dev, "Web Vitals," 2020, Accessed: Jan 19, 2024. [Online]. Available: https://web.dev/articles/vitals

[21] N. Wehner, M. Amir, M. Seufert, R. Schatz, and T. Hoßfeld, "A vital improvement? relating google's core web vitals to actual web qoe," in *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2022, pp. 1–6.

[22] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, 2010.

[23] E. Ibarrola, I. Taboada, and R. Ortega, "Web qoe evaluation in multi-agent networks: Validation of itu-t g. 1030," in *ICAS*, 2009.

[24] Q. Gao, P. Dey, and P. Ahammad, "Perceived performance of top retail webpages in the wild: Insights from large-scale crowdsourcing of above-the-fold qoe," in *Proceedings of the Workshop on QoE-based Analysis and Management of Data Communication Networks (Internet-QoE)*, 2017, pp. 13–18.

[25] S. Baraković and L. Skorin-Kapov, "Survey of research on quality of experience modelling for web browsing," *Quality and User Experience*, vol. 2, pp. 1–31, 2017.

[26] T. Hoßfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler, "The memory effect and its implications on web qoe modeling," in *2011 23rd international teletraffic congress (ITC)*. IEEE, 2011, pp. 103–110.

[27] A. Rehman and Z. Wang, "Perceptual experience of time-varying video quality," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2013, pp. 218–223.

[28] H. T. Tran, N. P. Ngoc, T. Hoßfeld, M. Seufert, and T. C. Thang, "Cumulative quality modeling for http adaptive streaming," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1, pp. 1–24, 2021.

[29] M. Seufert, O. Zach, M. Slanina, and P. Tran-Gia, "Unperturbed Video Streaming QoE under Web Page Related Context Factors," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017.

[30] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty." *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[31] D. Kahneman, B. L. Fredrickson, C. A. Schreiber, and D. A. Redelmeier, "When more pain is preferred to less: Adding a better end," *Psychological science*, vol. 4, no. 6, pp. 401–405, 1993.

[32] J. de Leeuw, "jsPsych," 2012, Accessed: Jan 19, 2024. [Online]. Available: https://www.jspsych.org/7.3/

[33] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best practices and recommendations for crowdsourced qoe," *Qualinet White Paper*, 2014.

[34] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.