4th International Workshop on Perceptual Quality of Systems (PQS 2013)
2-4 September 2013, Vienna, Austria

Asking Costs Little? The Impact of Tasks in Video QoE Studies on User Behavior and User Ratings

Andreas Sackl[1], Michael Seufert[2], Tobias Hoßfeld[2]

[1] Telecommunications Research Center Vienna (FTW), Vienna, Austria
Email: sackl@ftw.at
[2] University of Würzburg, Institute of Computer Science, Würzburg, Germany
Email: {seufert | hossfeld}@informatik.uni-wuerzburg.de

Abstract

Three crowdsourcing video QoE studies have been conducted to investigate the influence of tasks on user behavior and quality perception. Different dimensions of tasks have been applied to impaired video clips and their impact on task success, focus time, and QoE was measured. Our findings suggest that task presence or absence has no influence on the user. However, continuous tasks can improve the focus times of users, and hard tasks are suitable to filter unreliable participants.
Index Terms: Quality of Experience, Video Streaming, Task Execution

1. Introduction

As stated in [1], Quality of Experience should be considered under a holistic view but most of current research focuses on technical (network) factors and the influence of various application types on perceived quality, which neglects user related factors like expectations [2], user decisions and also context of use and tasks. In this paper we investigate the interplay of task fulfillment and HTTP video streaming (influenced by temporal impairments) and its impact on user behavior and user ratings.

Several related work was already conducted in this field. For web browsing QoE the difference between task [3] and non-task QoE [4] was studied. [5] investigated the impact of purpose on rating quality and user acceptance. Anchored tasks were used in voice quality studies and increases [6] as well as decreases [7] in listeners reliability were observed. In [8] the influence of simple and complex tasks on usability was investigated and a significant main effect of task complexity was found.

To analyze the impact of tasks in video QoE studies we designed three crowdsourcing experiments in which users had to watch short video clips which were impaired by stalling events, i.e. playback interruptions, which are typical artifacts of HTTP video streaming. We varied the given tasks in three different dimensions and examined the influence on task execution, focus time, and perceived quality. In particular, the following research questions were considered: *What is the impact of task fulfillment on HTTP video streaming, when the tasks are (a) present or absent, (b) related to the content or to the impairment, or (c) easy or difficult to fulfill?* As an additional outcome of this study, novel insights on task design and the identification of reliable users were gained which is in particular interesting for future crowdsourcing based user studies.

The remainder of this study is structured as follows: In Section 2 the study design is described in detail. Sections 3-5 will answer the presented research questions, and Section 6 will discuss the results and present an outlook on future work.

2. Design and Implementation of User Tests

The experiments were conducted via crowdsourcing, which is an emerging methodology for subjective user studies [9, 10]. The implementation was based on the framework used and described in [9]. To exclude the influence of network conditions, before the study started all videos were downloaded to the browser cache while the users filled out a personal data questionnaire. We used a video player which could be controlled via JavaScript commands to play the videos and to control the stalling. An animated icon with rotating circles was used as a visual indicator for stalling. Moreover, the framework collected browser-triggered events such as focus and blur events which are fired when the browser window gets or loses the focus. Thus the *focus time*, i.e. the time in which the video had the focus, could be computed.

For each experiment, our test participants consumed seven short video clips of 15 s length each. During the playback we introduced artificial stalling events to deteriorate the quality to a certain extent. To investigate the influence on user behavior, different viewing tasks were introduced to the users before and/or after the playback of the video clip, e.g., "What was the color of the car in the video clip?". After the playback, the user had to answer (1) the task question, (2) specify whether she had noticed stalling, and she had to (3) rate the impairment of the stalling on a standard DCR scale, ranging from 1

10.21437/PQS.2013-9

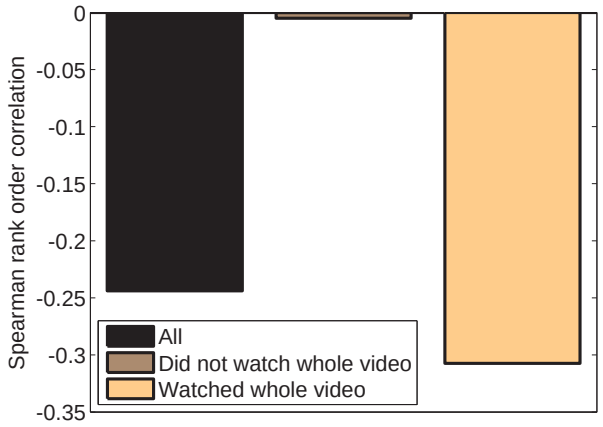

FIGURE 1: Spearman correlation between MOS ratings and number of stalling events.

(very annoying) to 5 (imperceptible).

We advertised three crowdsourcing campaigns and admitted 100 participating users each. To investigate the impact of tasks on the quality ratings, unreliable users which inadequately conducted the test had to be excluded from the result analysis as described in [9]. Using only a single criterion, for each video we filtered out users who did not watch the video completely. Therefore, a user is considered unreliable if her focus time was more than 1 s shorter than the video time, or if she could not answer the task question correctly. In Figure 1 the Spearman rank order correlation coefficient of users' ratings and number of stalling events is depicted for all users, users who did not watch the whole video, and users who watched the whole video. It can be seen that there is no correlation between the ratings of users who did not watch the whole video and the number of stalling events. Figure 1 shows that this filtering is an acceptable approach, as we remove users whose quality ratings can be considered as random noise. However, the results can only be considered quantitative results about the impact of tasks. Qualitative results about the video quality cannot be obtained with this filtering method. This issue is covered in more detail in Section 6.

Each of our experiments had a different focus regarding the presented question: in experiment 1 the influence of *task presence* is investigated, i.e. whether users are influenced by presenting task questions before and/or after the video is shown. In experiment 2, we assigned an *impairment related task* to the participants by asking them to count the occurred stalling events. Finally in experiment 3, the impact of *task difficulty* is examined, i.e. whether users are influenced by easy and hard tasks. The results for each experiment are shown in the next three sections.

Video ID	Stalling Impairment	Task Presence
Video 1	1x 2sec	after video
Video 2	1x 2sec	before and after video
Video 3	1x 2sec	after video
Video 4	1x 2sec	before and after video
Video 5	no	no
Video 6	1x 2sec	no
Video 7	1x 2sec	no

TABLE 1: Test conditions of experiment 1

3. Task Presence

In the first crowdsourcing experiment, each impaired 15 seconds video was compromised by a single stalling event with a duration of 2 seconds. Table 1 gives an overview about the used videos, the presence of stalling events, and the occurrence of content questions. After each video the users were asked about the quality of the video. Additionally we asked easy, content related questions, e.g., "Which region is shown in the video?" with the answer options "Mountains, Ocean, Desert" for a movie about mountains. In videos 1 and 3, these questions were shown *only after* the video playback, and the users had to answer it. However, in videos 2 and 4 users were informed about the task *before* the video started, e.g., "Please watch the video and answer the following question: Which region is shown in the video?". After the video playback the users had to answer the respective questions similar to videos 1 and 3. Video 5 had no temporal impairment and was used as reference, Videos 6 and 7 were impaired but not connected with any content questions.

In Figure 2, a bar plot is shown which indicates the ratio of users who wrongly answered the content questions for a presented video. Different colors represent the different question types and experiment numbers. The black bars show the ratio of wrongly answered questions in experiment 1. Only a small amount of users (below 5%) was not able to answer the easy content questions when the question was presented *only after* the video was shown. If the easy question was displayed already *before and after* the video was played, everyone was able to answer it. Thus, we assume that our content questions were too easy to influence task execution.

In Figure 3 the ratio of focus time and video time is investigated. If the user consumes the complete video, the ratio is 1. A lower value indicates, that the user aborted the video playback or changed the Internet browser tab during playback which resulted in a shorter focus time. In the figure, two CDFs of the ratio are plotted for videos with and without a precedent task. It can be seen, that the presence of a task has only marginal influence on the ratio of focus time and video time as both curves are very close. We assumed that assigning a task to the users, should force them to keep focused on the video, thereby

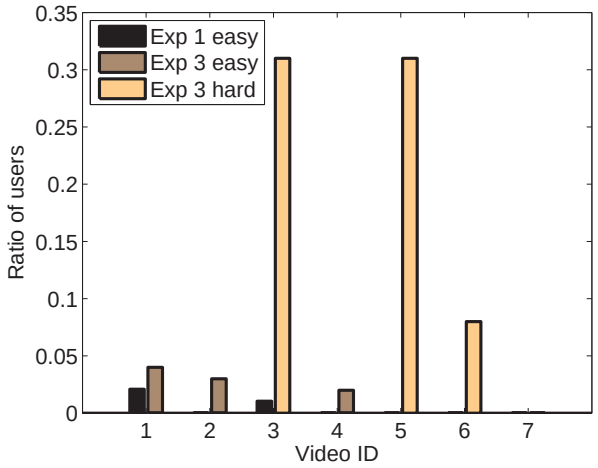

FIGURE 2: Ratio of users who wrongly answered the content question for different experiments and difficulty

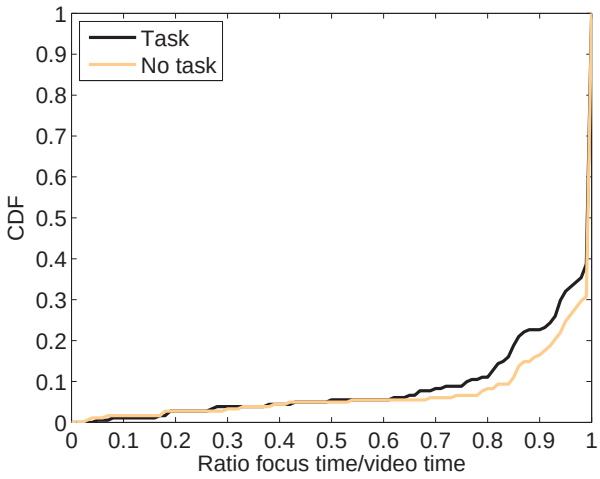

FIGURE 3: Cumulative distribution function of focus ratio, i.e. ratio of focus time and video time, for experiment 1

increasing the focus time. However, it seems that the presence of a task did not significantly influence the ratio and therefore tasks - at least our tasks which could be fulfilled easily within the first seconds of video playback - can not be used to increase focus times.

Additionally, we investigated the influence of task presence on quality perception. We hypothesized that the perception of temporal impairments is influenced if the test user has to focus on the content to answer a task question. The results of our first experiment can be seen in Figure 4 which shows the MOS for all videos which contained a single stalling impairment depending on task presence. All different task options, i.e. no task, task before and after video, and task only after video, resulted in similar quality ratings. Thus, we cannot observe any influence of easy task presence on the rating of stalling impairment.

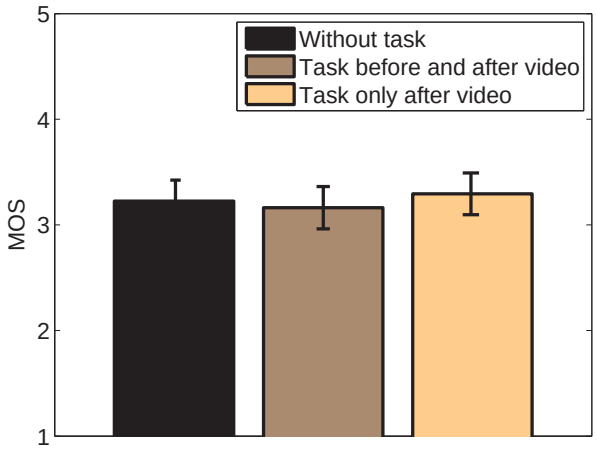

FIGURE 4: Mean opinion scores of videos with a single 2s stalling event for different presence of an easy task

Video ID	Stalling Impairment	Stalling Question
Video 1	2x 2sec	no
Video 2	3x 2sec	no
Video 3	2x 2sec	yes
Video 4	3x 2sec	yes
Video 5	no	no
Video 6	1x 2sec	no
Video 7	1x 2sec	yes

TABLE 2: Test conditions of experiment 2

4. Impairment Related Task

In experiment 2, we investigated if an impairment related task, i.e. counting the number of stalling events which occurred during video playback ("How many stalling events have you seen?"), had an influence on the task execution and ratings. In Table 2, the different stalling patterns are shown: In contrast to experiment 1, in the second experiment we varied the amount of stalling whereas the duration of the interruptions remained constant at 2 seconds. So, a stalling occurred one, two or three times during a 15 second video. Video 5 was used as a reference, there were no stalling events included. Before videos 3, 4, and 7 started, a hint was presented to the users ("Please count the stalling events in the following video"). For videos 1, 2 and 6, no task were given. Thus, these 3 videos are directly comparable with the videos 3, 4, and 7 respectively, which were accompanied by the stalling-counting task.

After the count task, the users had to rate on a scale from 0-4, how many stalling events they noticed. In Figure 5, the distribution of difference between users answers and actual stalling events is shown. If the difference between the counted stalling events and the real amount of stalling events is 0, the count task was done successfully. It can be seen that more than 77% of the impairment related questions have been answered correctly,

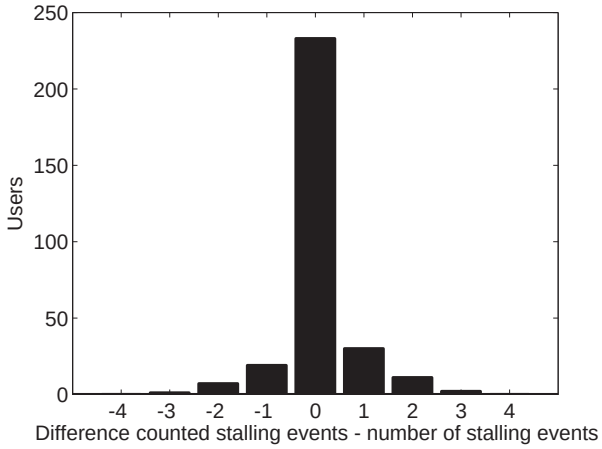

FIGURE 5: Results of count task execution

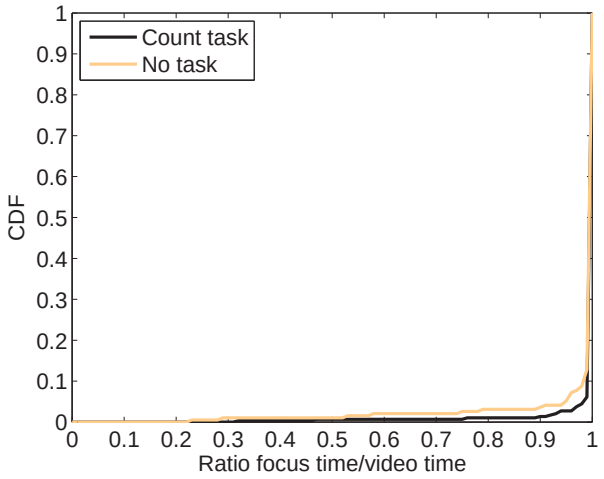

FIGURE 6: Cumulative distribution function of focus ratio for experiment 2

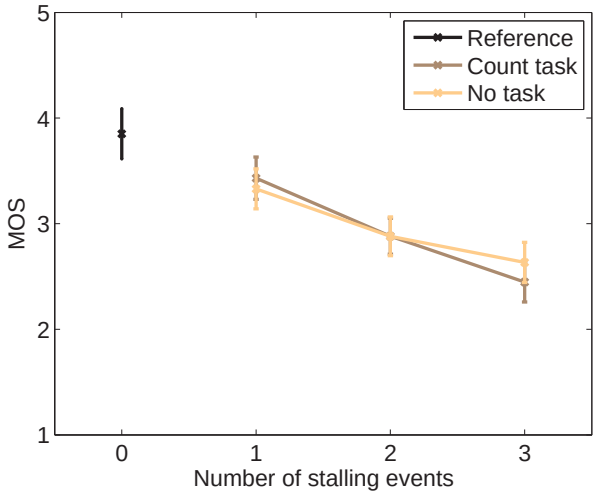

FIGURE 7: Mean opinion scores for different number of stalling events and different task presence

therefore we assume that this task was not perceived as hard by the users.

Similar to the results from experiment 1, the presence of a task has no huge influence on the ratio of focus time and video time, see Figure 6. However, in contrast to experiment 1, the focus times are generally higher. In the first campaign it was possible to answer the content related questions after a partial consumption, e.g. if the user had to specify the geographical region of a video, the answer was evident after a few seconds. However, in this experiment, users had to watch the entire video to count all stalling events. Therefore, to increase the focus time, a continuous task can be beneficial in order to keep the users focused.

In Figure 7, the influence of stalling events with respect to the presence of the counting tasks is shown. Unsurprisingly, a higher amount of stalling events lead to lower quality ratings, but the degradation of the ratings is not independent from impairment related questions. According to [9], the relation between stalling events and MOS should be logarithmic, but in our case the results describe a more linear coherence. We noticed that the Pearson linear correlation coefficient between user rating and number of stalling events increased from -0.26 (no task) to -0.38 (count task). Thus, a quality related task does have an influence on the resulting quality ratings and therefore should be avoided if qualitative results shall be obtained.

5. Task Difficulty

In the third crowdsourcing campaign, the difference between easy and hard content related questions was examined. All questions were presented as a task before the video was played and the question had to be answered afterwards. In contrast to the easy tasks/questions of experiment 1, e.g. "Which sport was presented in the video?" for a video snippet of a basketball game, more difficult questions were chosen for some videos, e.g. "Which team is playing in white sport suits?". These questions required the users to focus on a specific part of the video, e.g. a close-up of a player of the white team where the team name could be read easily. Thus, even hard questions could be easily answered by a user was attentive towards the task and the video. According to Table 3, video 7 was used as a reference, i.e. there was no stalling applied and no content related question was asked. Similar to experiment 2, different amount of stalling events have been applied, so the users were confronted with 1, 2 or 3 stalling events each with a duration of 2 seconds in a 15 seconds video.

In Figure 2, the ratio of wrong answering users is shown for each video ID. The hard questions in video 3 and 5 led to a high amount of wrong answers. More than

Video ID	Stalling Impairment	Content Question
Video 1	3x 2sec	easy
Video 2	1x 2sec	easy
Video 3	1x 2sec	hard
Video 4	2x 2sec	easy
Video 5	2x 2sec	hard
Video 6	3x 2sec	hard
Video 7	no	no

TABLE 3: Test conditions of experiment 3

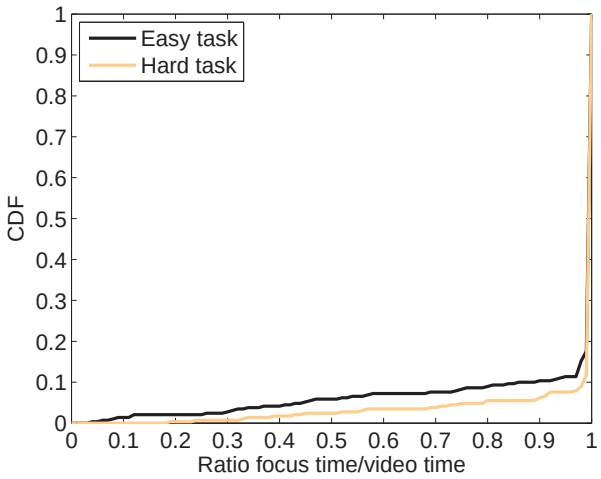

FIGURE 8: Cumulative distribution function of focus ratio for experiment 3

30% of the users were not able to answer them correctly. The question in video 6 could not be answered correctly by roughly 10% of the participants. As expected from the results of experiment 1, the easy content related questions were no problem for the users, as most of them were able to answer them correctly. According to Figure 8, the presence of an easy or hard task had no influence on the ratio of focus time and video time. There was only a small shift towards longer video consumption if a hard task had to be fulfilled. This little difference might be explained by the fact, that the answers to the hard questions were not so obvious (and thus could not be found so fast) in contrast to easy questions.

Figure 9 shows the MOS over the number of stalling events for easy and hard tasks. It can be seen that the presence of stalling events has a negative influence on quality perception for both easy and hard tasks. We hypothesized, that the presence of a hard and strenuous task would influence quality perception in contrast to an easy task: if a test user has to focus on the content to fulfill the task, either each interruption is obstructive and distracting, or the test user is so focused that interruptions are not perceived. According to our findings, we have to discard this assumption.

However, as there are no significant differences between hard and easy content related questions regarding the quality scores but there is a difference in goodness of task execution (more wrong answers), hard task questions are a useful approach to recognize users who do not work thoroughly.

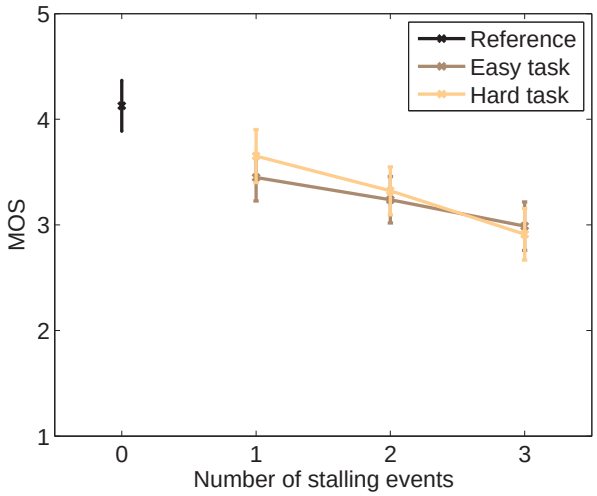

FIGURE 9: Mean opinion scores for different number of stalling events and different task difficulty

6. Discussion and Outlook

In this paper we examined the influence of different tasks on video quality perception impaired by stalling events. In general, task success (i.e. correctly answered task question) was only influenced by the difficulty of the task. The best focus time ratio could be achieved if the users had to fulfill a continuous stalling count task. Compared with more spotty content questions, this task naturally enforced users to consume the complete video not to overlook a stalling event.

The quality ratings were more or less independent from any task we presented to the user. The presence of a task did not positively or negatively influence the quality ratings. However, quality results from our experiments can only be considered qualitative. In general, in crowdsourcing studies filtering and the validity of the gained ratings are crucial to get also quantitative results. In Figure 10, the impact of different filtering approaches for all videos are shown. The results from experiment 2 (MOS over number of stalling events) are plotted both for normal and strict filtering. In the normal filtering approach, ratings are considered if the *current* video was consumed completely and the *current* task question was answered correctly. In the *strict filtering* approach, only ratings are considered if the concerning user *always* watched the complete videos and *always* answered the content or impairment related questions correctly. It can be seen, that the quality results of both filterings are qualitatively equal. For both, the quality deteriorates if the number

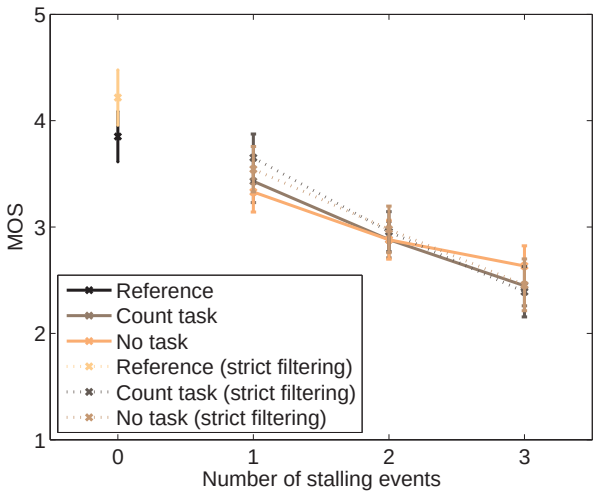

FIGURE 10: Results from experiment 2 for different filtering approaches

of stalling events increase, and the presence or absence of a count task had no influence on the ratings. However, only if strict filtering is applied, the resulting MOS scores are in line with previous findings which stated a big impact of stalling events (i.e. a more severe quality degradation) and a logarithmic coherence between temporal impairments and quality perception [9]. Thus, filtering mechanisms have to be designed carefully in order to obtain valuable results. From our experiments, we recommend to use *hard questions* which can improve filtering. As shown in the previous sections, the presence of a hard task does not influence MOS ratings, but the correct answering of the question can be used as a filtering criterion to filter out noise (see Figure 1) and thus increase the reliability and validity of the results.

We hypothesized, that tasks influence the attention of the users, and therefore the presence of temporal impairments should be perceived differently. But our findings revealed, that our test users were not affected by this. No influence of task presence could be recognized in terms of task execution and task success. We only observed that focus times could be improved by a continuous task, and that quality ratings could be slightly influenced by a quality related task. However, the obtained findings are by no means general. Future work will have to continue to investigate the impact of tasks on task execution and task results. Different dimensions of task design, reliability of workers, filtering mechanisms, and validity of results are still open issues in this field and have to be examined thoroughly.

7. Acknowledgements

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grants HO 4770/1-1 and TR257/31-1 (”OekoNet”) and the COST QUALINET Action IC1003. The authors alone are responsible for the content.

8. References

[1] P. Le Callet, S. Möller, and A. Perkis, “Qualinet White Paper on Definitions of Quality of Experience (2012),” European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Tech. Rep., 2012, Lausanne, Switzerland.

[2] A. Sackl and R. Schatz, “Evaluating the impact of expectations on end-user quality perception,” in *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*, Vienna, Austria, 2013.

[3] D. Strohmeier, S. Jumisko-Pyykko, and A. Raake, “Toward task-dependent evaluation of web-QoE: Free exploration vs. ”Who Ate What?”,” in *IEEE Globecom Workshops*, Anaheim, CA, USA, 2012.

[4] S. Egger, T. Hoßfeld, R. Schatz, and M. Fiedler, “Waiting Times in Quality of Experience for Web Based Services,” in *4th International Workshop on Quality of Multimedia Experience (QoMEX 2012)*, Yarra Valley, Australia, 2012.

[5] J.-L. Farh, A. A. Cannella, A. G. Bedeian *et al.*, “The impact of purpose on rating quality and user acceptance,” *Group & Organization Management*, vol. 16, no. 4, pp. 367–386, 1991.

[6] B. R. Gerratt, J. Kreiman, N. Antonanzas-Barroso, and G. S. Berke, “Comparing internal and external standards in voice quality judgments,” *Journal of Speech, Language and Hearing Research*, vol. 36, no. 1, p. 14, 1993.

[7] K. M. Chan and E. M. Yiu, “The effect of anchors and training on the reliability of perceptual voice evaluation,” *Journal of Speech, Language and Hearing Research*, vol. 45, no. 1, p. 111, 2002.

[8] I. F. Ince, Y. B. Salman, and M. E. Yildirim, “A user study: the effects of mobile phone prototypes and task complexities on usability,” in *2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, Seoul, Korea, 2009.

[9] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, “Quantification of YouTube QoE via Crowdsourcing,” in *IEEE International Workshop on Multimedia Quality of Experience - Modeling, Evaluation, and Directions (MQoE 2011)*, Dana Point, CA, USA, 2011.

[10] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, “CrowdTesting: A Novel Methodology for Subjective User Studies and QoE Evaluation,” University of Würzburg, Tech. Rep. 486, 2013.