3rd International Workshop on Perceptual Quality of Systems (PQS 2010)
6-8 September 2010, Bautzen, Germany

Gain from Strain? Assessing the Impact of User Fatigue on the Quality of Subjective MOS Ratings

Raimund Schatz, Sebastian Egger, Kathrin Masuch

Telecommunications Research Center (FTW), Vienna, Austria
egger@ftw.at, schatz@ftw.at, kmasuch@ftw.at

Abstract

This paper addresses a central question in subjective test design: "What is the appropriate duration for QoE lab experiments that involve human participants?" Since user tests are time-consuming and costly, this is an important question as it exposes the fundamental challenge of maximizing test duration while avoiding problems arising from undue strain on test participants. We provide an extensive analysis of the influence that workload and fatigue have on the rating behavior of QoE test participants, based on empirical data. Our analysis is grounded on measurements and experiences made during two typical QoE lab studies which assessed the impact of different network conditions on perceived quality of data services (web browsing, file download). During both studies, we measured participants' workload and fatigue in two complementary ways, subjectively by means of a questionnaire and objectively by capturing eye blink rate (EBR). Our main finding is that even after 90 minutes of active testing, the mean and standard deviation of participants' MOS grades were not significantly affected by their temporal position within the test sessions (beginning vs. end). Thus, for comparable QoE lab user experiments, this duration represents a safe recommendation for balancing results quantity with results quality.

Index Terms: Subjective QoE Testing, Workload, Physiological Measurements

1. Introduction

During the last decade, subjective testing has become the primary method for assessing the Quality of Experience (QoE) of networked communication and media presentation systems. The main rationale behind subjective testing is that it requires a human being to perceive and experience something. In this respect, exposing people to the phenomena under investigation and subsequently eliciting their opinions about them represents an essential prerequisite for understanding and quantifying quality as it is actually experienced. On the other hand, the subjective nature of test participants' perception, opinions and quality ratings also represents a methodological challenge, threatening the objectivity required for scientifically valid study results. In this context, frequently cited causes for biased or distorted results are the phrasing of test assistants' briefings and interview questions, as well as the participants' internal mental and emotional states [1, 2]. Therefore, reliable and valid subjective quality tests require a careful experimental design that takes these factors into account in order to ensure that feedback from participants remains undistorted.

A central design parameter of QoE lab studies is the duration of the testing session, since subjective tests tend to be time-consuming and costly. Researchers thus strive for maximizing test duration and thus the number of test conditions per user, but obviously, there are practical limits to maximization such as personnel costs as well as decreasing participant motivation and energy. However, related literature on subjective testing in the domains of usability engineering [3, 2, 4] and speech/audio-visual quality [5, 6, 7, 8] falls short of providing concrete guidance in terms of upper limits for session duration.

In order to address this shortcoming, this paper analyzes the influence of workload and fatigue on the rating behavior of QoE test participants. It does so by triangulating data from three sources: user ratings, subjective reports on strain and objective physiological measurements of fatigue. Our results provide concrete methodological guidance, suggesting how much test duration can be increased without compromising the reliability of user ratings. The remainder of this paper is structured as follows: Section 2 provides an overview of related work, in particular concerning existing guidelines and recommendations for QoE study design. Section 3 describes the two QoE studies that form the basis of our work as well as the methods used for data capturing and analysis. Section 4 presents the results of our data analysis, followed by a more detailed discussion and interpretation in Section 5. Finally, Section 6 provides general conclusions as well as an outlook on future work.

2. Theory and Related Work

Participant strain is a natural consequence of any user test based on active involvement of participants. In general, it is an undisputed rule that exhaustive fatigue of test subjects should be avoided in order to gain reliable and unbiased results. Nevertheless, there is less agreement on when exactly users get too tired to rate reliably or on how to reliably measure participants' alertness state. To illustrate this problem, we discuss related work that has investigated the topics user test duration, fatigue and strain measurement as well as physiological measurements in the context of QoE studies.

2.1. Guidelines for User Studies

Within the field of audio and video quality assessment, several ITU-T recommendations [6, 5, 9, 7, 8, 10] for subjective user test design exist. However, these recommendations differ significantly in the session length they suggest. The majority is found in the field of speech quality assessment, where the recommended overall test duration ranges from 20 minutes [6] up to 160 minutes [9]. Less guidance can be found in the video quality testing domain where video-sequence durations (i.e. condition durations) are recommended. However, the decision regarding overall test session length is left to the test designer himself. Another source of information on user test duration is the usability engineering literature [4, 3, 2]. Although a

10.21437/PQS.2010-9

number of factors to be considered in user testing are discussed, explicit guidance on optimizing test duration is missing.

2.2. Non-intrusive Measurement of Strain and Fatigue

In order to determine a subject's quality perception and rating competence in presence of fatigue it is necessary to assess the subject's overall alertness state. Relevant studies that address this challenge have been recently conducted in the domains of traffic telematics and highway safety. Their aim is to detect driver drowsiness in the context of car driving using minimal intrusive techniques. One of the most prominent approaches in this field is the observation of the drivers eye blink behavior.

Early stage work on psychophysiological states and their impact on eye blink behavior [11, 12] provided the basis for the utilization of eye blink analysis as an indicator of fatigue. Schleicher et al. [13] have utilized these results and conducted driving tests of considerable length (approx. 2.5 hours) in order to assess the relationship between subjectively reported and objectively measured alertness and eye blink statistics. The authors identified eye blink rate (EBR) as one major objective indicator that highly correlates with reported alertness ratings. Unfortunately, the Electrooculography (EOG) method used in their study is too intrusive for QoE testing. To bypass such intrusive sensory equipment one can alternatively analyze EBR by means of video recordings of the subjects' faces with minimal intrusiveness.

In addition to objective alertness indicators, participants' subjective reports and ratings can be used to complement and validate results gathered through objective fatigue measures. A widely used tool for such subjective workload and strain assessment is the NASA Task Load Index (TLX) [14]. Using a 20-point rating scale, the TLX addresses several dimensions that typically arise in test situations, including mental, physical and temporal demand.

2.3. Physiological Measurements in QoE assessment

Existing work that incorporates physiological measurements for QoE assessment and prediction can be classified into two approaches. The first approach attempts to directly correlate physiological measurements with QoE ratings [15, 16].The second approach on the other hand assesses the subjects mental and emotional states [17] and further relates these states to user experience of certain entertainment technologies (e.g. multiplayer games). Both approaches struggle to find a clear linkage between user QoE ratings and the measured physiological indicators, as these indicators are influenced by several parallel somatic and psychological processes and are also subject to considerable invidual variation. In contrast, our studies pursue a different goal: rather than trying to predict QoE by physiological measurements we use such measurements to gauge participant well-being in order to provide recommendations for test design. Therefore, we set out to correlate physiological measurements with indicators for the *quality and reliabiliy* of QoE ratings and not with the rating values themselves.

3. Experimental Setup and Methodology

This section describes the two QoE user studies that served as context for our research. Furthermore, it describes our method for subjective and objective assessment of test participant strain and fatigue.

3.1. QoE User Studies

Our research on the influence of user workload and fatigue on QoE rating behavior was conducted in the context of two lab studies, which assessed the impact of different network conditions on the perceived quality of VoIP and data services (web browsing). As shown in Figure 1, the overall test duration of both studies was approximately 120 minutes, with the active QoE testing session lasting 90 minutes, including a 10-minute break inbetween. The QoE testing session consisted of sequences of 2-3 minute conditions, with each condition representing a user task executed under certain network quality settings. After each condition, users were prompted for rating perceived network speed, overall quality, task difficulty, etc. using an electronic questionnaire. At the beginning of each test, a user briefing and training phase was performed, including a paper questionnaire on the participant's general background and internet usage habits. Each test was concluded by a final debriefing interview and a demographic questionnaire. In this respect, both studies followed the common format of lab based QoE studies.

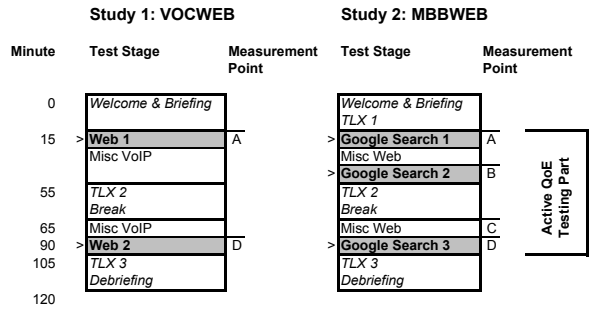

Figure 1: Overview of the main stages of the QoE studies. The test stages used for rating comparison are marked in grey.

Content-wise, the two studies had different foci, resulting in different task (and thus strain) profiles: Study 1 (VOCWEB) was a mixed VoIP (Voice over IP) and Web QoE study, while Study 2 (MBBWEB) was a pure Web study. At the beginning (minute 15) as well as at the end (minute 105) of the active testing phase, participants executed a sequence of 4-5 web surfing conditions during which they browsed through an online photo album at different network speeds (64-2048 kBit/s) using a laptop. The comparison of the statistics of users' gradings between both web sequences was made possible by the following test design: in both sequences, average network speeds were the same and we factored in two shared control conditions. This way, any differences in rating results between the two sequences can be validly put into a causal relationship with their temporal placement within the test (beginning vs. end). As the VoIP part of the active QoE testing phase inbetween both web sequences is regarded, participants executed a series of voice quality assessment tasks featuring conversational activities causing only little visual strain.

In contrast to Study 1, Study 2 (MBBWEB) was a pure Web QoE study: during the whole testing phase participants executed web browsing conditions causing primarily visual and cognitive strain. Compared to the web conditions in Study 1, users were exposed to a greater variety of sites (photo album, search, news, travel) and tasks (search, browse, read and understand). In order to again assess the impact of temporal placement (and thus user fatigue) on rating results, we placed three sequences of four conditions (all of them based on the

same google search task) at the beginning (minute 15), middle (minute 50, before the break) and end (minute 100) of the test session. The three sequences shared at least one control condition, with sequence 1 and 3 even consisting of completely identical conditions (however using randomized order).

In both studies, objective fatigue measurement was conducted by analyzing 10 minute periods of the video material, starting at measurement points A (beginning) and D (end) of the test. In order to additionally assess the influence of the break, we decided to add measurement points B (before) and C (after) for Study 2[1](see Figure 1). 68 people participated in total. As we did not record all of them on video and some video recordings were of poor technical quality (lighting issues), we came down to 14 analyzable users for MBBWEB and 9 analyzable users for VOCWEB. From these 10 were female and 13 male test users, with an age distribution between 19 and 45 years (mean 27.9 years, median 25 years).

3.2. Workload and Fatigue Measurement

During both studies, we measured participants' workload and fatigue in two complementary ways, subjectively by means of a questionnaire [14] and objectively by capturing eye blink rate (EBR).

3.2.1. Subjective Strain Assessment

For the subjective assessment of strain, we adapted the NASA TLX (Task Load Index) questionnaire [14] regarding the questions asked to the user. Mental demand and effort were the dimensions we chose for our tests, as our task design suggested changes on these dimensions while other dimensions, such as temporal demand were steady throughout a test session.

3.2.2. Objective Strain Assessment

As shown in Section 2, eye blink rate (EBR) has been demonstrated to be a reliable indicator of workload and fatigue that can be objectively measured in non-intrusive ways. In comparison to other physiological stress and fatigue measurement approaches (such as galvanic skin resistance), this method has the advantage of being completely unobtrusive (when done via video analysis), which is an essential feature for investigating natural, unbiased behavior [18]. For capturing the users' faces, we placed the camera at a distance of about two meters in front of the user. The majority of the users did not notice the camera although we told them in advance that the user test will be video captured. Hence, we can conclude that we adhered with the minimal intrusiveness criteria demanded for QoE testing as mentioned above.

In a first step, we tried to automatically detect eye blinks by using the process engine from [19]. Unfortunately, the results obtained were not perfect due to several reasons: first the unconstrained movement of the subjects poses a huge difficulty to automatic blink detection. Furthermore, some subjects were too short and not optimally seated, thus parts of their faces were occluded by the monitor. However, the required setup using an unobtrusive camera position did not always allow for capturing every subject's complete face.

For those reasons, we performed manual coding of the same video sequences. The coding person went through the video sequences and recorded the number of visible blinks per minute . Furthermore, any other incidents or unusual behaviors that affect EBR measurement (e.g. user ceasing eye blinking at all, face becoming covered by the laptop screen due a shift in body position) were recorded. For each measurement point (MP), 10 minutes of video were analyzed and then averaged, resulting in a mean EBR per user and per measurement point. This has to be considered as raw data, as blink rates are person specific and vary from individual to individual. In addition, seeing aids such as contact lenses and eye glasses exert considerable influence on blinking patterns. Therefore, intrapersonal evaluation by normalization is a necessity. We thus calculated the differences in each person's EBR for measurement points B, C, D from measurement point A which serves as the reference EBR. Then, we related each measurement point difference score to the mean EBR for the respective user in order to get the relative EBR change for the measurement points B, C and D.

4. Results

This section discusses the measurement results we obtained from letting test subjects grade equivalent QoE conditions at different stages of both tests while measuring their strain and fatigue levels.

4.1. User Ratings

As described in Section 3.1, we exposed subjects to equivalent sequences of test conditions at the beginning (measurement point A) and the end (measurement point D) of both studies. Figure 2 shows the comparison of network speed MOS ratings[2] for each condition setting. As visible in both diagrams, the user gradings given at the end of the studies differ from those made at the beginning only to a small extent. Although the mean values of each condition pair do deviate, these differences lie within the range of statistical error and thus do not indicate the presence of any statistically significant influence on rating behavior, including fatigue [3]. The only exception was the 8 seconds page load time condition in Study 2 (see Figure 2 bottom) where the ratings made at the beginning and at the end differed with statistical significance. In addition, equal magnitudes of the error bars imply that also the *variance* of gradings, another reliable indicator of a change in rating behavior, do not differ between the beginning and the end of both studies.

These results suggest that for a QoE test with an active testing phase of up to 90 minutes duration, the quality of user ratings does not change in any statistically significant way. Thus, only marginal influences resulting from the temporal placement of conditions within the test could be detected considering user gradings only. This result raises the question, whether external influences such as workload and strain were present at all. In order to answer this question, the following two subsections discuss the results from our strain measurements performed during both studies.

4.2. Subjective Strain Assessment

Figure 3 shows participant ratings of perceived mental workload and overall strain (mental and physical) as inquired at the

[1]This was not possible for Study 1 (VOCWEB), since EBR measurement requires a visual task of at least 10 minutes duration. However, suitable visual tasks were not present in the middle part (VoIP) of this study.

[2]At the end of each condition, subjects graded several QoE dimensions. Nonetheless, in this paper we only use network speed MOS ratings for comparisons, as they had the strongest discriminatory power.

[3]We performed Wilcoxon signed rank tests where no significant differences were obtained except for the 8 second page load time condition in Study 2.

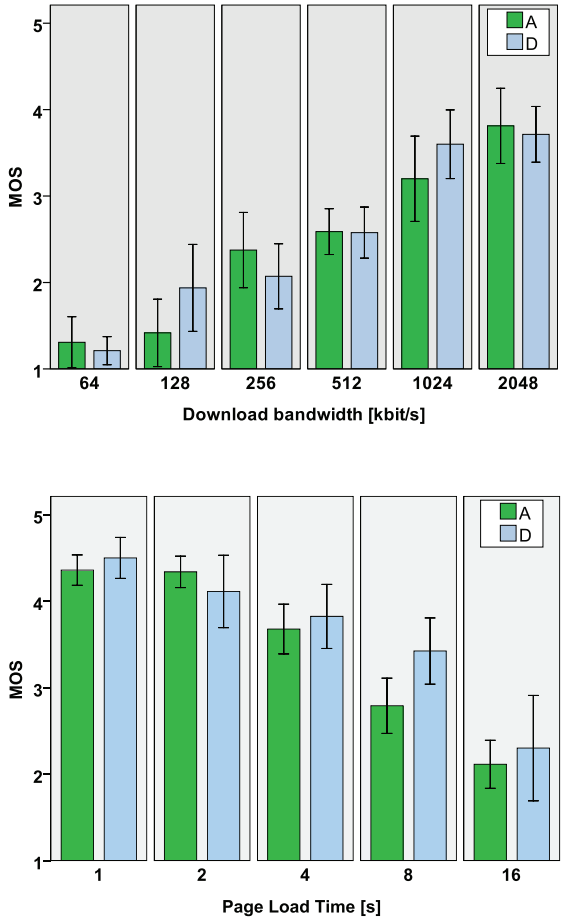

Figure 2: Comparison of mean MOS ratings between measurement points A and D for photo web browsing conditions at different network speeds in Study 1 (top) and for google search tasks at different page load times in Study 2 (bottom).

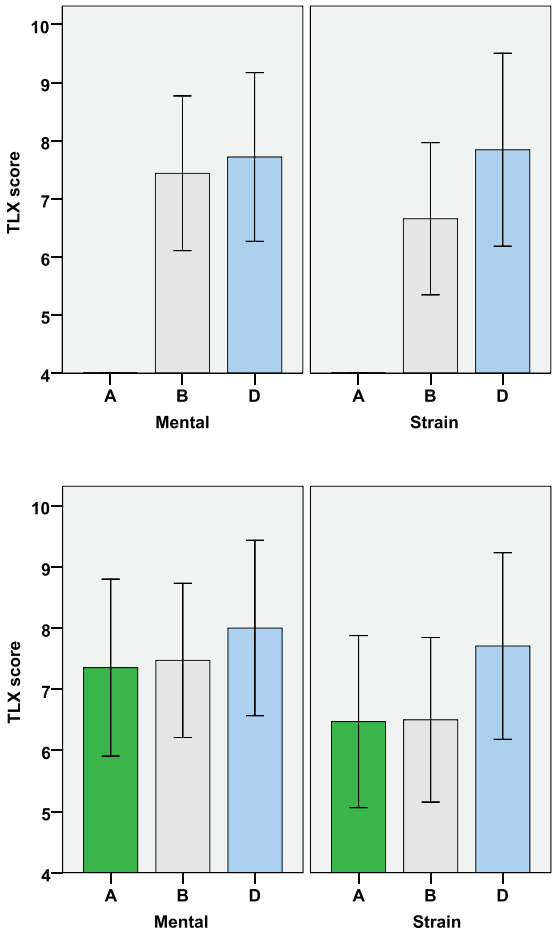

Figure 3: Mean user ratings of perceived mental load and overall strain as performed at measurement points A (Study 2 only), B and D.

beginning, middle and end of the active testing phase (i.e. measurement points A, B and D)[4] using a TLX-based questionnaire.

The TLX user ratings demonstrate that perceived workload and strain do visibly increase towards the end of both studies (albeit only the 'overall strain' ratings differ with statistical significance). Interestingly, mean ratings (and their variances) at points A (Study 2 only) and B are almost equal, thus the subjects did not feel any sort of fatigue increase at the end of the first half of the active testing period. However, since self-reporting of mental and emotional states is known to have its limitations [13], we expected the objective measurement of fatigue symptoms to yield further insights.

4.3. Objective Strain Measurement

As described in Section 3.2, we performed eye blink rate (EBR) analysis during 10 minute periods starting at measurement points A/D and B/C (Study 2 only) in order to objectively measure participant strain. Figure 4 shows the mean relative change of EBR over the course of the active testing phase[5]. Our measurement results show a considerable increase of EBR (approximately +20% in VOCWEB and +30% in MBBWEB) from point A to D, indicating a significant increase of fatigue levels towards the end of both studies. However, the magnitude of the error bars at point D also indicates that the amount of relative EBR change varies significantly across individuals. This phenomenon can be explained by the fact that individuals vary in terms of fatigue patterns as well as translation of their fatigue into EBR change. Nonetheless, we measured a notable increase in objective fatigue, in line with the participants' subjective TLX ratings.

In order to obtain a more fine-grained picture of strain development over time, we assessed the influence of the 10-minute break by additionally measuring EBR at points B and C in Study 2. The results reveal a strong influence of the break: while at the end of the first half of the testing phase (point B), EBR reaches levels similar to point D, the presence of a break makes the mean EBR change drop down to approximately 10% (point C, see Figure 4). The consequences of these results are that firstly, after 40 minutes of active testing participants should have a break. Secondly, a break duration of min. 10 minutes is

[4] In order to obtain a further reference point for comparison, a third TLX questionnaire was inserted in Study 2 at point A.

[5] Since absolute EBR varies considerably between subjects, only relative EBR change can be used as reliable strain indicator.

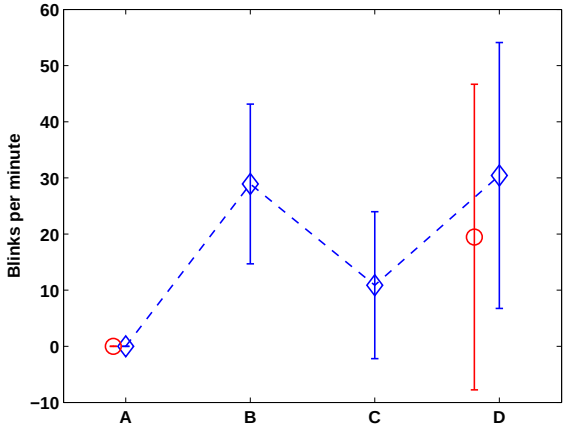

Figure 4: Mean percentual change of EBR relative to measurement point A for Study 1 (red) and 2 (dashed, blue).

needed to induce a significant recovery effect.

5. Discussion

The results presented in the previous section show that in both studies, the quality of MOS QoE ratings was affected only to a small extent by their temporal placement within the test (beginning vs. end). This was the case despite the fact that most subjects showed or reported signs of fatigue and exhaustion towards the end of their session. When comparing measurements from both studies, in Study 1 (VOCWEB) changes in EBR and rating quality were less present than in Study 2 (MBBWEB), where for one condition even a statistically significant difference of mean ratings was found. The most plausible explanation for this difference is that Study 2 consisted of visual tasks (web surfing) only. In contrast, Study 1 mixed auditive, conversational and visual activities, thus causing less eye strain and task monotony. Our results thus on the one hand support the hypothesis that different types of studies cause different strain profiles with varied impact on ratings. On the other hand, our EBR measurements detected similar strain levels for both studies, which is in line with the results from [12] that demonstrate that eye blink behavior changes are invariant to task profiles.

Our results also suggest that the test duration is not the only factor influencing user fatigue. Adequately chosen breaks are as important as the pure test durations and can be used to reduce user fatigue in long test sessions. Sufficient break insertion can thereby help to ensure high quality results for long enduring test sessions. By being able to conduct user tests with longer duration one can also overcome between subject comparison of results which is necessary if not all test conditions can be presented to each user due to time constraints. Such within subject comparison would further improve result quality.

In this context it is important to remember that fatigue is not the only influence factor that becomes relevant when increasing test duration: habituation and learning effects as well as emotional states such as boredom affect the participants' vigilance, quality perception and their motivation to pay full attention to the test situation, particularly when tasks and content are of repetitive nature. Therefore, future studies should also assess motivation, boredom and general task performance in parallel to fatigue.

The reader might dislike the fact that the magnitude of the error bars of our strain measurement plots appears relatively high in comparison to actual mean values. We account that to the relative low number of subjects (9 for Study 1, 14 for Study 2) that we included in the analysis. Further studies should consider a larger sample of users to test our hypotheses and assumptions with higher confidence.

We also found a noteworthy discrepancy between subjective and objective fatigue scores. For measurement point B (before the break), subjective TLX ratings do not show fatigue whereas the objective EBR changes already show a significant increase of strain. This difference can be explained by the fact that user perception of emotional and psychophysiological states is not always correct [15]. Finally, we have to mention that we tested (negatively) for statistical differences in the ratings rather than for equality.

6. Conclusion

In this paper we reported on the link between QoE test participant fatigue and the reliability of their quality ratings. Our results show that for two QoE studies with an active test period of 90 minutes, the reliability of user gradings declined only slightly, with differences of mean values remaining below statistical significance thresholds. We therefore conclude that 90 minutes (break included) can be considered as a safe duration for active QoE testing sessions, despite the presence of significant physical and mental strain as detected by our measurements. In this respect, NASA TLX based self-reporting of strain served as a reliable but nonetheless crude instrument, which needed to be complemented by objective measurements. Video-based eye blink rate (EBR) counting proved itself as an unobtrusive, easy-to-implement way of fatigue detection, at least as is conserned the data gathering phase. However, this method also comes with a number of practical challenges, such as the need to perform robust blink detection as well as the presence of considerable variation among individuals' eye blink patterns. Consequently, we recommend the total number of test subjects to be larger than 20 in order to be able to obtain statistically significant EBR results while leaving enough room for excluding those (inevitable) cases where this method fails.

As far as future work is concerned, the next step is determining the citical point beyond which user ratings definitely become unreliable, i.e. differ with statistical significance. This point can be found by applying the method presented in this paper to QoE studies that either exceed the recommended duration of 90 minutes or omit the break in middle at all in order to make participants tire more rapidly. Secondly, we have only evaluated a mixed VoIP/Web and pure Web QoE study so far. We therefore plan to investigate the influence of fatigue in the context of other types of QoE studies, e.g. pure VoIP, IPTV video quality or audio-visual conferencing. Differences can be expected to arise in terms of overall fatigue patterns (and thus maximum test duration) as well as type of strain, since pure audio conferencing causes significantly less eye strain than continuos image interpretation and video quality judgement. Thirdly, fatigue detection itself can be significantly improved, not only by using an automated EBR counting algorithm (such as [20]) instead of manual coding, but also by augmenting the analysis with additional features. For example, eye-blink duration and delay of lid reopening are also very important variables for strain detection, as has been demonstrated by Schleicher et al. [13].

Furthermore, the robustness of fatigue assessment can be increased by using complementary modalities e.g. by analyzing and comparing participants' voice patterns over time. Meth-

ods as described in [21, 22] might be used to assess user fatigue from voice samples recorded via user tests. In a further step, triangulation of voice methods and the methods described in this paper could yield even better results in fatigue assessment. If they prove to be robust enough, these methods can serve as the foundation of an unobtrusive, real-time strain detector that timely suggests remedial actions (such as inserting a break) to the test assistants in case of participant fatigue. We therefore cordially invite those researchers who engaged in subjective QoE testing to jointly investigate this topic of importance and share their experiences with measuring and managing participant strain.

7. Acknowledgements

This research has been performed within the projects ACE and U-0 at the Telecommunications Research Center Vienna (FTW) and has been funded by the Austrian Government and the City of Vienna within the competence center program COMET. The authors would like to thank Stefan Scherer from the University of Ulm for his work on the automated eye blink detection algorithm and their colleagues from Areas U and N for their valuable input and discussions.

8. References

[1] J. Bortz and N. Doering, *Forschungsmethoden und Evaluation: fuer Human- und Sozialwissenschaftler*, 4th ed. Heidelberg: Springer, 2006.

[2] J. Rubin, *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, T. Hudson, Ed. New York, NY, USA: John Wiley & Sons, Inc., May 2008, vol. 2nd edition.

[3] J. Nielsen, *Usability Engineering*. San Francisco, California: Morgan Kaufmann Publishers, October 1994.

[4] J. S. Dumas and J. C. Redish, *A Practical Guide to Usability Testing*. Exeter, UK, UK: Intellect Books, 1999.

[5] International Telecommunication Union, "A method for subjective performance assessment of the quality of speech voice output devices," *ITU-T Recommendation P.85*, June 1994.

[6] ——, "Methods for subjective determination of transmission quality," *ITU-T Recommendation P.800*, Aug. 1996.

[7] ——, "Subjective video quality assessment methods for multimedia applications," *ITU-T Recommendation P.910*, April 2008.

[8] ——, "Subjective audiovisual quality assessment methods for multimedia applications," *ITU-T Recommendation P.911*, December 1998.

[9] ——, "Subjective evaluation of conversational quality," *ITU-T Recommendation P.805*, July 2007.

[10] ——, "Interactive test methods for audiovisual communications," *ITU-T Recommendation P.920*, May 2000.

[11] V. Hargutt, "Das lidschlussverhalten als indikator fr aufmerksamkeits- und mdigkeitsprozesse bei arbeitshandlungen," Ph.D. dissertation, Julius Maximilians Universitaet Wuerzburg, 2003.

[12] P. E. Meinold, "Psychologie des lidschlags eine literatur- und methodenkritische studie," Ph.D. dissertation, Universitaet Koeln, 2005.

[13] R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired?" *Ergonomics*, vol. 51, no. 7, pp. 982–1010, 2008.

[14] S. G. Hart and L. E. Stavenland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds. Elsevier, 1988, ch. 7, pp. 139–183. [Online]. Available: http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000004342_1999205624.pdf

[15] G. M. Wilson and M. A. Sasse, "Do users always know what's good for them? utilising physiological responses to assess media quality," in *In: The Proceedings of HCI 2000: People and Computers XIVUsability or Else! (HCI 2000)*. Springer, 2000, pp. 327–339.

[16] ——, "Investigating the impact of audio degradations on users: Subjective vs. objective assessment methods," in *Proceedings of OZCHI 2000*, 2000, pp. 327–339.

[17] R. L. Mandryk, K. M. Inkpen, and T. W. Calvert, "Using psychophysiological techniques to measure user experience with entertainment technologies," *Behaviour & IT*, vol. 25, no. 2, pp. 141–158, 2006.

[18] W. N. Campbell, *On the use of non verbal speech sounds in human communication*, ser. Lecture Notes in Computer Science. Springer, 2007, vol. 4775, pp. 117–128.

[19] S. Scherer, V. Fritzsch, and F. Schwenker, "Multimodal real-time conversation analysis using a novel process engine," in *in Proceedings of International Conference on Affective Computing and Intelligent Interaction 2009 (ACII '09)*, 2009, pp. 253–255. [Online]. Available: http://acii2009.nl

[20] M. Divjak and H. Bischof, "Eye blink based fatigue detection for prevention of computer vision syndrome," in *IAPR Conference on Machine Vision Applications*, 2009.

[21] J. Krajewski, R. Wieland, and A. Batliner, "An acoustic framework for detecting fatigue in speech based human-computer-interaction," in *ICCHP '08: Proceedings of the 11th international conference on Computers Helping People with Special Needs*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 54–61.

[22] S. Scherer, H. Hofmann, M. Lampmann, M. Pfeil, S. Rhinow, and F. Schwenker, "Emotion recognition from speech: Stress experiment," in *Proceedings of Language Resources and Evaluation Conference 2008*, 2008.