Macro-temporal Development of QoE: Impact of Varying Performance on QoE over Multiple Interactions

Dennis Guse [1], Sebastian Möller[1]

[1] *Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin,*
Email: {firstname.lastname}@telekom.de

Abstract

For IP-based multimedia service providers perceived Quality of Experience (QoE) is an important factor as it influences user satisfaction, which in turn affects future usage behavior. We propose to complement the current state-of-the-art approach to QoE from short-term, i.e. one interaction up to some minutes, to cover multiple interactions over longer periods to determine long-term effects on QoE. In this paper, we present a study employing an audio-speech communication system and an entertainment system. The study was conducted with 20 subjects over 14 days providing daily task-based interaction using two service performance profiles. We found that low performance events immediately influenced the per-interaction QoE-ratings and the following QoE-ratings. We further derived a simple model to estimate the integrated QoE over multiple interactions per service.

Introduction

The subjective perception of audio and audio-visual systems and the resulting Quality of Experiences (QoE) depends on personal preferences, prior experiences, expectations, context of use, and the task of the current user [7]. Research on QoE has a long history and a major impact on the development of technology like compression algorithms, end-user devices and transmission systems. Subjective experiments are used to compare different options or to find the key dimensions determining QoE. Subjective perceptional experiments must be reproducible, e.g. different test subjects should have a similar experience, and are therefore conducted under controlled laboratory settings. Such experiments focus on short time intervals up to several minutes that are meaningful for evaluating and comparing technologies, but are limited with regard to temporal effects over multiple full-length interactions.

In this paper, we report the results of a field study exploring the impact of repeated interactions of 5-15 minutes on QoE using speech communication and audio-visual entertainment over a period of 14 days. This paper is structured as follows. First, an overview on temporal effects in QoE and related fields is given. Second, we describe the study design and the used technical setup. Following, the study results are reported and future research discussed.

Related Work

Very little is known about temporal effects of time-varying system performance on QoE over longer or even multiple interactions. Short-term effects like the recency effect [9], i.e. more recent events have a higher impact than previous events, or the peak-end-rule [6], i.e. negative outliers are considered more important, have been adapted from cognitive science.

Gross et. al. [3] and also Weiss et. al. [13] investigated the impact of varying transmission performance during telephone calls up to two minutes in a laboratory setting.

Staelens et. al. [12] studied QoE of full-length movies and thus extending the typical timespan for current QoE research. The study was conducted in the home-environment of the test participants. It was found that not all errors were perceived and that the number of perceived impairments does not explain the QoE rating. Stalling et. al. showed that completely new models must be developed to estimate QoE for full-length movies.

In the 1960s Duncanson [2] conducted a study about the integration of QoE over several telephone calls and found that the “average telephone call is better than the average telephone call”, i.e. the QoE rating of a just completed call with average performance is higher than the QoE rating of for the “usual” telephone call with the system.

Möller et. al. [8] extended this by letting pairs of test subjects use an audio-visual communication tool over 14 days in their home-environment. The service performance was modified on a day-to-day basis. Two major findings were reported. First, after a low performing interaction the per-interaction QoE ratings require up to two interactions to recover to the level before the degraded interaction. Second, the QoE ratings have the tendency to increase slowly over the study period.

A completely different angle on temporal effects is shown in the business literature, e.g. Parasuramen et. al. [10], with regard to usage of services, like a bank or online shopping, mainly focusing on satisfaction and behavioral intentions like future-use and willingness-to-pay. Quality as something quantifiable, i.e. in a sense of performance, is considered as an important factor. However, little work has been done on the question how integration of quality events happens over multiple interactions.

Study

The goal of this paper is to investigate temporal effects on QoE over several interactions. We limited our work to macro-temporal variations of performance, i.e. variation only between interactions, and not micro-temporal variation, i.e. within-interaction.

Study Design

We chose two different types of services, which are used frequently and broadly in daily life: speech communication and audio-visual entertainment. The study was designed for a length of 14 days. This enables test subjects to familiarize with the systems and covers a typical period for service adaptation.

We follow the task-driven approach by Möller et. al. [8] and Staelens et. al. [12] to achieve comparable usage patterns between participants. For speech communication, *short conversation tests* (SCT) [4] were used. Those two-person role-plays are solvable in about 3-7 minutes. For audio-visual entertainment, the task is to watch a movie of circa 15 minutes length and answer two content-related questions. Choosing longer content allows immersing into the story and is realistic for lean-back entertainment. To counter the imbalance in usage length, one entertainment task and two communication tasks were given per day. The communication tasks should be solved between 6am and 1pm, and 3pm and 12pm, respectively. The entertainment task should be solved between 3pm and 12pm.

We used two performance levels: *high performing* (HP) and *low performing* (LP), which should result in a high and low QoE, respectively. We used two performance profiles as shown in Figure 1. *Profile 1* (P1) is HQ only and *Profile 2* (P2) starts with 2 days HP, following 3x2 days of LP with a gap of 2 days HP and finishes with 2 days HP. All communication and entertainment interactions on one day were either HP or LP. The technical parameters are shown in Table 1.

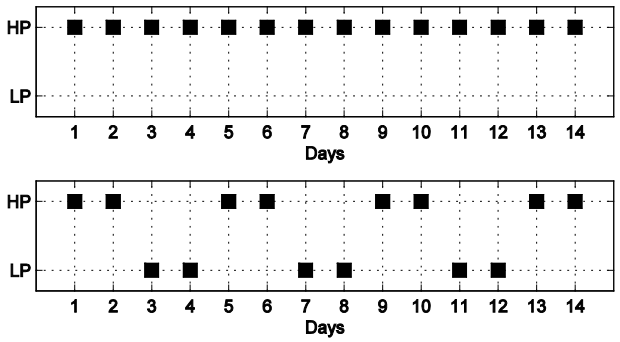

Figure 1: Performance Profiles 1 and 2.

As content for the entertainment task, we used *Friends Season 3 episodes 1-8*. Each episode was split into two parts, so that the storyline remained intact. The length of the split parts was 12-17 minutes. The movies were presented during the study in the chronological order.

Questionnaires

In the study, two questionnaires were used: one per usage presented directly after finishing one task and one to measure the integrated QoE over several interactions. To measure the QoE we used the scale shown in Figure 2 [5] that was also used by Möller et. al. [8]. For communication the overall QoE and for entertainment overall, audio and video QoE should be rated. The second questionnaire was presented after day 4, 7, 10 and 14 to measure the integrated QoE and determine the users satisfaction with each system. Satisfaction was measured using the *Net Promoter Score* (NPS) [11].

Figure 2: Rating scale for QoE feedback.
Labels from left-to-right: very bad, bad, poor, fair, good, excellent and ideal.

Table 1: High and low Performance Parameters for Communication and Entertainment interaction.

	Type	HP	LP
Speech Communication	Codec	G.722 (16kHz; 64 Kbit/s)	
	Random Packet-loss	0%	5%
Audio-visual Entertainment	Video Codec	H.264 – Profile Main	
	Video encoding bandwidth	2 Mbit/s	125 Kbit/s
	Resolution	720x576	
	Audio Codec	AAC (stereo): 48 kHz	
	Audio encoding bandwidth	165 Kbit/s	

Technical Setup

The communication system was implemented using SIP and RTP (UDP-based) using Asterisk 10[1] acting as a RTP-proxy running on FreeBSD 9.0[2] reachable via a public IPv4-Address. The open-source SIP client Jitsi 1.0[3] was used. Random packet-loss was inserted using Dummynet [1]. For the entertainment system a video player running in a Web browser was implemented using Microsoft Silverlight. The player behaves like an ordinary audio-visual streaming client, but videos are stored on the device. Seek and pause is not available. For the study, we provided each test subject one prepared USB-Stick with the video player incl. the videos and one Logitech PC120 headset. The test subjects used their own computer or notebook running Windows 7.

Study Procedure

The study consisted of three parts. First, a pair of test subjects was invited to the initial session. We recruited the test subjects in pairs of two, who have known each other before to avoid effects due to changing interaction partners for the communication systems. The initial session consisted of a short interview about demographic data and their experience with multimedia as well as internet technology. Then, the client software was installed on their notebooks, and it was verified that the devices were performant enough. To demonstrate the software and tasks the pair solved one SCT and one entertainment task under supervision of the experimenter. The 14-day study period started on the following day and was conducted by the tests subjects on their own in their home environment. At the end, a final interview was conducted. Each test subject received 70 € as compensation.

Results

The study was conducted in Berlin, Germany from August to October 2012. 10 pairs successfully finished the study, 5 per performance profile, with an average age of 22.9 years (min: 17 and max: 28). 11 male and 9 females participated. Of the 280 scheduled calls, 5 were not done and 18 per-interaction questionnaires were not filled. For the entertainment tasks, 13 per-interaction questionnaires and one integrated QoE questionnaire were not filled. In the following, only the ratings on QoE are reported.

[1] https://www.asterisk.org/
[2] http://www.freebsd.org/
[3] https://jitsi.org/

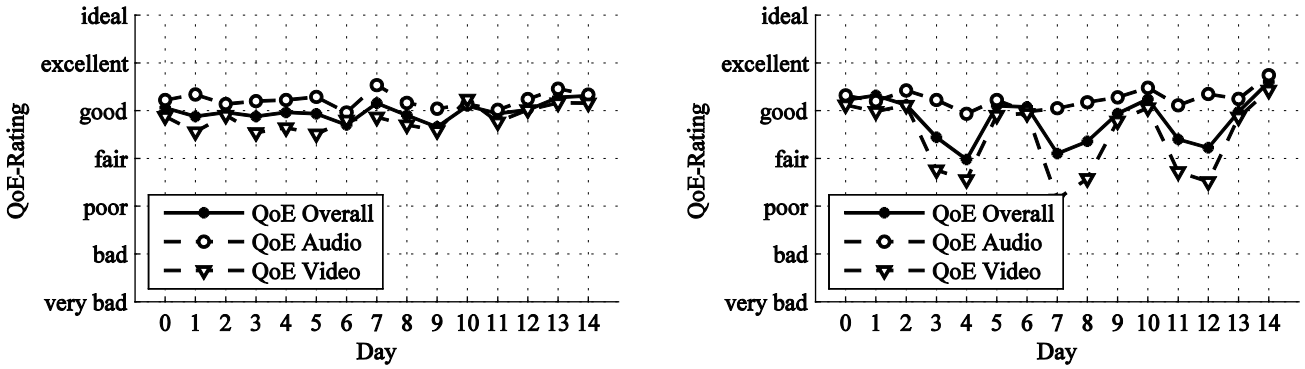

Figure 3: Average per-interaction ratings of overall, audio and video QoE for the audio-visual entertainment tasks (from left-to-right: P1 and P2).

Audio-visual Entertainment

The entertainment system worked as expected except one subject reported problems with choppy video playback. Thus, the performance profile could be delivered for P1 and P2. Figure 3 shows the average per-interaction ratings for overall as well as audio and video QoE. P2 shows that reducing the video encoding bandwidth produced, as expected, perceived artifacts. However, with regard to LP even lower video-QoE ratings were expected. It is noticeable that there is an impact of low video QoE on audio QoE. As expected, the overall QoE is between audio and video QoE and might be successfully estimable using the average of audio and video QoE.

In both profiles the effect reported by Möller et. al. [8] that per-interaction QoE ratings increase slightly over the study period is noticeable. The recovery effect of per-usage QoE rating after a low performance event could be observed in P2 after the 2nd and 3rd performance gap.

In Figure 4, the average overall per-interaction QoE ratings and the integrated overall QoE ratings are shown. P2 shows that the integrated QoE ratings react slower than the per-interaction ratings. Furthermore, it must be noted that P2 shows an increase to almost the same level as P1 in the integrated QoE ratings from day 7 on even though the degradations were applied and perceived.

We found that the integrated overall QoE QI_j rating on day j for $j \in \{4,7,10,14\}$ can be estimated using the *simple moving average* (SMA) of the per-usage overall QoE ratings (Q) of n days for $0 \leq n \leq j$ before. This model achieves a RMSE between 0.12 and 0.34 for P1 almost independent of n. Correlation is for all intervals greater than 0.76. With regard to P2, we found that this model achieves an RMSE between 0.26 and 0.74. Correlation is greater than 0.75 except of QI_{10} for $n = 0$ and $n = 1$, which lead to 0.43 and 0.44. For QI_{10} and QI_{14}, we found that for higher n the performance increases, e.g. RMSE tends to decreases slightly.

Speech Communication

Figure 5 shows the QoE ratings for the communication system. P1 is not stable, i.e. has a high variance on QoE ratings from call to call, and in P2 LP cannot be clearly distinguished. In the data evaluation after the study, it became evident that the speech communication system did not perform as planned. This was because the Internet was used as transport medium and the used client, Jitsi, did not perform as expected. The later added an intense circuit noise and had some stability issues. Thus, P2 could not be delivered as planned. However, P1 and P2 are different with regard to per-interaction QoE. P1 is above 4 and has 8 several ratings over 5 whereas P2 has only 2 ratings above 5 and 6 around 4. Therefore, P2 has performed worse than P1 with regard to QoE. In addition, the integrated QoE-ratings of P2 are unexpected as they are below the per-usage QoE-ratings, which is different from P1 and the entertainment system. This leads to the assumption that the per-interaction QoE ratings are more positive, but that LP events and problems with the client are nevertheless expressed in the integrated QoE-ratings. This is somehow in line with the findings of Duncanson [2]. In this case, the average is not appropriate to model this effect. Because the reason for this is not clear, we must leave the modeling task for P2 unsolved.

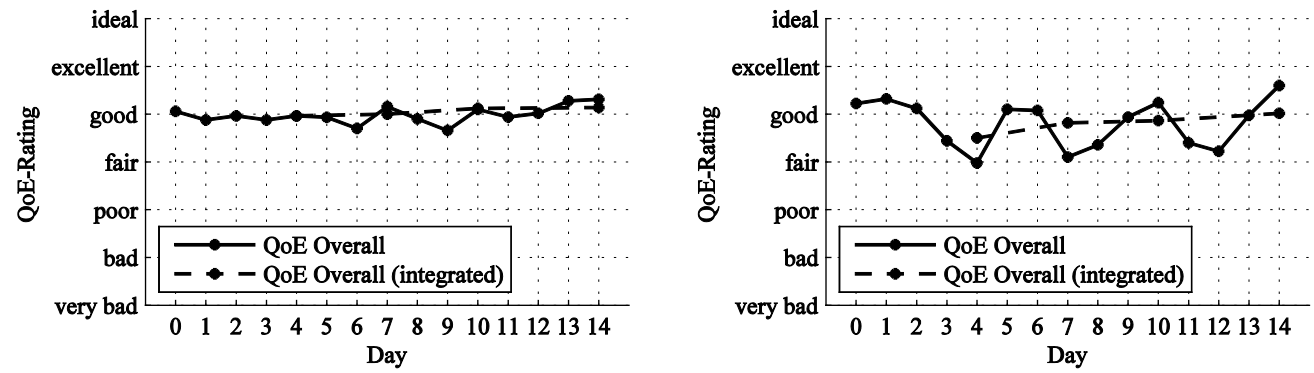

Figure 4: Average per-interaction and integrated QoE ratings for the audio-visual entertainment tasks (from left-to-right: P1 and P2).

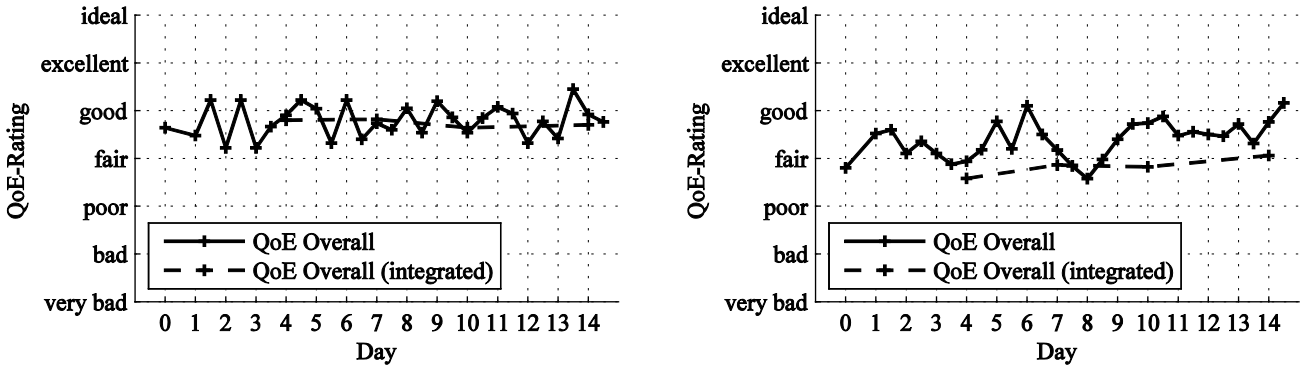

Figure 5: Average per-interaction and integrated QoE ratings for the audio-speech communication tasks (from left-to-right: P1 and P2).

For P1 the RSME for SMA is between 0.23 and 0.66. QI_4 has an avg. RMSE of 0.36 independent of n. For QI_7 and $QI_{10}$the RMSE decreases from around 0.6 for $n = 0$ down to 0.27 for the maximum of n. For QI_{14} the RMSE decreases from 0.62 for $n = 0$ down to 0.47 for $n = 9$.

Future-use

In the final interview, we asked the test subjects, if they would continue to use the provided systems. The entertainment system would be used by 7 subjects of P1 and by 6 of P2 in the future whereas for the communication system 6 of P1 and only 2 of P2 would continue to use it. Therefore, the degradations of P2 influenced the future-use for the speech-communication system. This is congruent with the low ratings for the integrated QoE.

Conclusion

In this paper, we investigated macro-temporal effects on QoE using a speech communication system and an entertainment system. Our results show that our audio-visual entertainment tasks were stable against constant limitations of the video bandwidth, and that integrated QoE can be modeled with a moving average of prior per-interaction QoE ratings. Our findings with regard to speech communication are very different due to limitations of the system. Nevertheless, we found that per-interaction QoE is rated more positively for very low performing services compared to the integrated QoE. In future work it is necessary to determine key influence factors on macro-temporal QoE and study differences between entertainment and communication services.

Acknowledgements

This work made possible by the Telekom Innovation Laboratories, Deutsche Telekom AG. Thanks to Videoload and Warner Entertainment for providing the media content.

References

[1] Carbone M. and Rizzo L.: Dummynet Revisited, ACM SIGCOMM Computer Communication Review, 40(2) pg.12-20, March 2010.

[2] Duncanson, J. P.: The average telephone call is better than the average telephone call, The Public Opinion Quarterly, Vol. 33, No. 1, pp. 112–116, Spring 1969.

[3] Gros, L., Chateau: Instantaneous and overall judgments for time-varying speech quality: Assessment and Prediction, Acta Acustica united with Acustica, Vol. 87, No. 3, May/June 2001.

[4] ITU-T: Recommendation P.805 Subjective evaluation of conversational quality, April 2007.

[5] ITU-T: Recommendation P.851 Subjective quality evaluation of telephone services based on spoken dialogue systems, November 2003.

[6] Kahneman D.: Objective Happiness. In: Well-Being: The Foundations of Hedonic Psychology. D. Kahneman, E. Diener, N. Schwarz (eds.). Russel Sage, pp. 3-25, 1999.

[7] Le Callet P., Möller S., Perkis A. (eds.): 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), Lausanne, Switzerland, Version 1.1, June 3, 2012.

[8] Möller S. et. al.: From Single-Call to Multi-Call Quality: A Study on Long-term Quality Integration in Audio-Visual Speech Communication, Interspeech 2011.

[9] Murdock, B. R.: The Serial Position Effect of Free Recall, Journal of Experimental Psychology, Vol. 64, No. 5, pp. 482-288, 1962.

[10] Parasuramen A., Zeithaml V.A., Berry L. L.: SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality, Journal of Retailing, Vol. 64, No. 1, pp. 12-40, Spring 1988.

[11] Reichheld F. R.: The One Number You Need To Grow, Harvard Business Press, March 2003.

[12] Staelens N. et al.: Assessing the Influence of Packet Loss and Frame Freezes on the Perceptual Quality of Full Length Movies, 4th International workshop on Video Processing and Quality Metrics for Consumer Eletronics, January 2009.

[13] Weiss B. et. al.: Modeling Call Quality for Time-Varying Transmission Characteristics Using Simulated Conversational Structures, Acta Acustica 2009.