Chapter 10
Temporal Development of Quality of Experience

Benjamin Weiss, Dennis Guse, Sebastian Möller, Alexander Raake, Adam Borowiak and Ulrich Reiter

Abstract Most research on Quality of Experience treats QoE as a static event. As a result, QoE is measured for stimuli of delimited length, and the QoE which is associated with the stimulus is considered to be stable along its duration. However, this rarely happens in reality where usage episodes extend over several seconds and minutes (e.g. a phone call), hours (e.g. a video film), or regularly over periods of weeks or months (when considering QoE of a subscribed service). In this chapter, we will discuss the cognitive processes involved when QoE is integrated over usage episodes, and describe corresponding assessment methods. We will also review models for estimating episodic and multi-episodic QoE from momentary QoE judgments or their predictions.

B. Weiss (✉) · D. Guse · S. Möller
Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Berlin, Germany
e-mail: bweiss@telekom.de

D. Guse
e-mail: dennis.guse@telekom.de

S. Möller
e-mail: sebastian.moeller@telekom.de

A. Raake
Assessment of IP-based Applications, Telekom Innovation Laboratories, TU Berlin, Berlin, Germany
e-mail: alexander.raake@telekom.de

A. Borowiak · U. Reiter
Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
e-mail: adam.borowiak@iet.ntnu.no

U. Reiter
e-mail: ulrich.reiter@ntnu.no

S. Möller and A. Raake (eds.), *Quality of Experience*,
T-Labs Series in Telecommunication Services, DOI: 10.1007/978-3-319-02681-7_10,

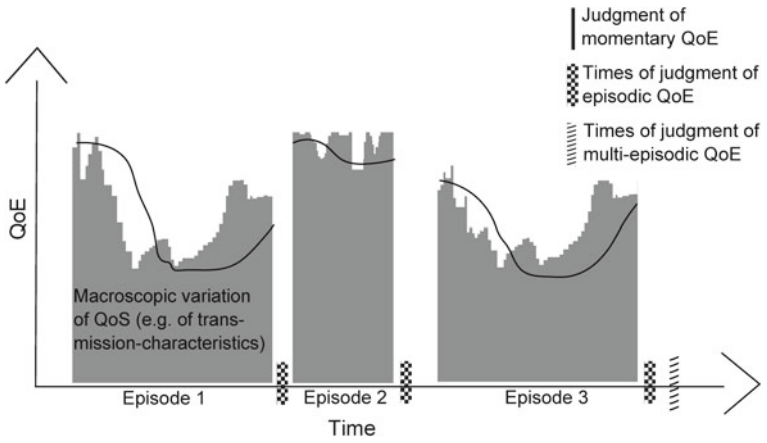

Fig. 10.1 Schematic illustration of QoE concepts

10.1 Introduction

With current communication services and networks, one major issue is the temporal variability of transmission characteristics as they are common in today's mobile and best-effort networks. From a Quality of Experience (QoE) perspective, a system with time-varying characteristics will have its impact on the user in terms of various aspects. Respective time spans have been defined for user experience [44], ranging from:

1. *momentary*/instantaneous experience of the system characteristics as a result of the current media quality level or of events like sudden changes in transmission characteristics,
2. over the retrospective appraisal of various events during an *episode* of usage like a call or video clip,
3. until the cumulative experience when evaluating a whole service after *multiple episodes*.

Of course, retrospective ratings can be asked for at any time during one episode. In this chapter, however, we consider only complete episodes. Confer Fig. 10.1 for an illustration of time spans and judgment events of momentary QoE, (remembered) episodic and multi-episodic QoE, and the macroscopic (see below) QoS variation.

These three time-spans have a close connection to Kahneman [31], who distinguishes between the *momentary*-based approach and the *memory*-based approach. This separation is made for the global field of human experience with concepts like joy or pain, and can be directly transferred to QoE (see also Chap. 2).

The momentary-based approach corresponds to momentary QoE and reflects direct instantaneous measures of experience. Variation over time of such measures corresponds to *macroscopic* changes of system characteristics, as these are actually

perceived as being varying in quality, as opposed to *microscopic* system behavior (cf. [42]). Accumulating (e.g. averaging) ratings of momentary experience represents a useful abstraction, not actual experience, called *total utility* in [31]. Momentary QoE itself is an important measure for addressing the instantaneous user reaction to changes in Quality of Service (e.g. types of transmission degradations). Assessment methods of momentary QoE include those for truly instantaneous assessment, but also "sampled momentary" QoE, i.e. assessment for short-term samples which exhibit no macroscopic, but microscopic variation. These methods are presented in Sect. 10.3.

The memory-based approach is concerned with retrospective appraisal, i.e. remembered experience. This subsumes episodic and multi-episodic QoE, which can be differentiated regarding the time of assessment and the lengths of the experience: Episodic QoE is concerned solely with one episode and typically assessed directly after this one, whereas multi-episodic QoE might be assessed with or without temporal alignment to an episode, and the scope of the experienced quality is usually the whole service up to the event of assessment. This remembered QoE (*remembered utility* in terms of Kahneman [31]) reflects the users' integration processes and the establishment and development of attitudes towards a system or service (cf. Sects. 10.4 and 10.5).

Dependent on the research question, either momentary or (multi-)episodic QoE is in focus. But in order to study their relationship, instantaneous and retrospective ratings have to be analyzed together.

The focus in this chapter lies on the QoE as a result of time-varying transmission characteristics. We address this by presenting empirical results in a structured way, outlining assessment methods, and even presenting prediction models for auditory, visual and audio-visual quality. Waiting times are not a topic presented in this chapter, as they are covered in Sect. 22.3. Further information on cognitive aspects of QoE are presented in Chap. 2.

According to the three time spans presented, this chapter starts in Sect. 10.2 with cognitive processes related to the temporal development of QoE. Then, it provides an overview of methods to assess momentary, episodic and multi-episodic QoE (Sects. 10.3, 10.4, 10.5) and presents major principles of how retrospective appraisal is related to momentary QoE (mainly in Sect. 10.4). We conclude with an outlook on major issues in current research and applications (Sect. 10.6).

10.2 Cognitive Processes of Temporal QoE

One of the major topics in QoE research is the relationship between several momentary events and retrospective appraisal of the resulting QoE. Often, a weighted average of momentary ratings is correlated strongly with a single rating obtained directly after a session. Results for such weighted averages reveal a higher weight of the last ratings compared to the others. This observed effect is often called recency effect, implying a relation to cognitive processes of recall. The assumption

seems to be, that human processes in recalling information from the working memory will ground such a retrospective appraisal, based on cognitive models [1].

Within the time range of the working memory of about tenths of seconds, typically 5–9 unrelated items (like object names or numbers) can be recalled right after the presentation or after a short break without distraction. The exact time span used for recall cannot be defined, as the working memory is not just an information storage, but a multi-component system used for complex cognitive tasks. For example, names or digits can be rehearsed within the phonological loop as long as there is no distraction. The actual performance and time span depend on the individual, content of information, motivation and attention to the task, or modality of distractor tasks (see also Chap. 2).

The main effect observed in most studies addressing free recall from the working memory is the likelihood to recall information better or worse depending on the position of presentation: The first and the most recent items are recalled with a higher probability in a free recall task. The first, so-called *primacy effect* lessens somewhat with more items (e.g., 10–20 and thus also with longer time to recall), whereas the *recency effect* can be reduced or eliminated by distraction or delay (e.g., 15–30 s) between presentation and recall of items. Recency can thus be viewed as the retaining of information, where the recall has not been deteriorated by subsequent information.

The likelihood to recall information better or worse depending on the temporal position within an episode can also be taken up by instrumental quality prediction models. The rationale is to include such a cognitive process to model the rating at the end of an episode on the basis of ratings for relatively short stimuli or calls with varying quality.

Interestingly, such positional effects can also be observed for longer time spans like several minutes to hours (e.g., remembering content of a talk), or even for a whole season or year, determined by the number of events, not the time elapsed. A different cognitive approach is not directly linked to actually recalling or retrieving information from memory, but judging individual, even long-term experience in retrospect based on memories, the memory-based approach presented in Sect. 10.1. The *peak-end rule* models the heuristic of appraisal of one's experiences in terms of valence and intensity only by two remembered moments [30]. That is, retrospectively judging long-term experiences (tens of minutes, but also time spans of, e.g., years) are dominated by the most extreme and the most recent experiences. It could be shown, that other information are not lost, but just not included into the retrospection. Here, a primacy effect is not apparent.

This heuristic and the underlying peak-end rule seem to be closer to the actual task of rating (multi-) episodic QoE than recall from the working memory. In fact, the task of an episode-final QoE judgment is quite different to an instantaneous judgment of momentary QoE, and thus does not precisely require to recall the experienced quality, compare, judge and describe it. Instead, remembered QoE is bettercharacterized by

eliciting the *current* attitude towards the service based on those quality events in scope, may this be one or more episodes, and on which the peak-end rule is based on.

The relevance of intensity is already known for other judgment tasks based on heuristics, e.g. the topic of combining traits for interpersonal judgments. Here, a stronger impact of negative traits than positive ones is found [22], that can be modeled with a weighted average [32]. Accordingly, models for time-varying quality often take into account valence (i.e. special treatment of degradations) and variability itself (cf. Sect. 10.4).

10.3 Assessing Momentary QoE

Most of the existing mono- and multi-modal quality assessment techniques do not take into account fluctuations of the quality which happen to appear when a stimulus of extended duration is viewed (i.e. more than 10 s). The remembered quality rating provided after an episode (e.g. 10 min) should not be considered as an accurate measure of momentary QoE, which is continuously evolving. This is due to the fact that humans are more likely to make the overall quality judgment based on the most recent experiences which are assumed to be of a greater importance or significance (cf. Sect. 10.2). In order to overcome the mentioned inaccuracy two approaches can be applied: the long content can be divided (windowed) into a number of shorter episodes (e.g. 10 s each) and then evaluated separately (for an overview of parametric models for estimating QoE of such short samples, see Sects. 12.3, 14.3 and 19.2), or the quality can be judged on the fly, throughout the entire stimulus duration. In the first approach, standardized methods applicable for short sequences evaluation can be used. Examples of such methods, typically applied to 3–16 s long episodes, are: Double Stimulus Continuous Quality Scale (DSCQS) [26], Absolute Category Rating (ACR) [28] and Paired Comparison (PC) [28] (for more see [27, 29]).

However, the mentioned assessment techniques are lengthy in nature and hence impractical in real life applications, where e.g. content of 30 min duration needs to be evaluated. For this purpose, an appropriate, no-reference method (i.e. without a stimulus to compare) allowing for instantaneous quality evaluation should be employed. An example of such a method is the Single Stimulus Continuous Quality Evaluation (SSCQE) developed by the RACE MOSAIC project [19] and later incorporated into the ITU-R recommendation BT.500-7 [26], or the corresponding continuous assessment method for speech quality described in ITU-T Rec. P.880 [43]. The SSCQE allows participants to judge the perceived quality dynamically using a slider mechanism with associated interval scale (commonly from 0 to 100 with the range divided into five equal slots corresponding to the ordinal five-point quality scale). Although the method is capable of catching the quality variations instantaneously and over extended periods of time, it is not free from drawbacks and ambiguities. It has been reported that the continuous operation of the slider might divert the user's attention from the process of quality assessment [7] and that the differences in participant's reaction time to quality changes can reduce the accuracy

of the method [41]. Recently, there has been increased interest in the development of alternative methodologies capable of tracking the quality changes in a continuous manner by using different types of rating devices, i.e. a glove [8], a steering wheel [37], etc. Nevertheless, except for the type of rating instrument, those methods do not bring any major methodological changes compared to the SSCQE. This is due to the fact that all of them use the same type of rating scale (or in some cases even simplified versions with reduced resolution) associated with each of the device. Moreover, the improvement in the rating instruments' performance over the slider mechanism has not been proven and an effect of stimulus duration on users' fatigue related to usage of these devices has not been verified.

Modification of another standardized method (ACR) has been suggested for a quasi-momentary assessment of a longer episode by providing judgments after time-intervals of fixed window size [18]. The quality judgments are made during the appearance of segments with no degradations in contrast to the gray segments used for this purpose in the ACR method. This way, according to the authors statement, the continuity of the sequence is preserved making the viewing conditions more realistic.

A different approach towards continuous quality evaluation has been proposed by Borowiak et al. [5]. Instead of providing a numerical representation of the perceptual experience, the user is allowed to actively adjust the quality to the most appreciated level in case degradation occurs. The improvement in the quality is achieved by means of an adjustment device (e.g. rotary knob) and based on perception of quality changes solely (no tactile feedback from the device). The scale assigned to the assessment instrument in fact is a direct representation of the quality levels used in the test, and a translation of the perceived quality into a numerical score or position of the rating device is not required. There are no physical limits in the rotation mechanism as witnessed in the previously mentioned methods, and the maximum quality can be overpassed if not recognized by the user, causing gradual decrease in quality. This is a reversible process, so the user can return to the reference quality by rotating the device in the opposite direction again.

With this new technique, a measure of momentary QoE is achieved in relation to the desired QoE. Eliciting the user's behavioral reaction to experienced quality by means of the quality adjustment approach allows for gathering the data with less cognitive resources required compared to typical assessment methods [5]. In consequence, subjects' attention is on the presented stimuli rather than on the usually challenging task of quality evaluation. Although new findings are possible with the method, it should not be treated as a replacement of the existing ITU-R rating scales based methodologies, but rather as an additional source of information with respect to the user's cognitive experiences.

In general, there has been relatively little attention devoted to the topic of continuous quality evaluation, resulting in a comparatively small number of related publications. However, some interesting research findings have been claimed with respect to the momentary quality assessment. In [33] it has been concluded that subjects react almost immediately when a change from good to poor quality occurs while in the reverse situation the adaptation process is much slower. This asymmetry in tracking

the quality has been confirmed in [9] where changes in momentary speech quality were evaluated. Other studies [4, 6], in which longer duration content (30 min) was employed, prove the time dimension not being the main factor influencing quality perception. The authors discovered that quality expectations over extended periods of time are rather constant and that the same holds for the absolute quality level at which the change is usually discovered. Moreover, it has been found that sensitivity to quality variations highly depends on the awareness of the process of quality changes, and is higher when the subjects are in charge of the quality adjustment, than when the process is controlled externally. These findings hold, no matter which way the degradations appear in the presented material; whether stepwise in time or immediately, in one move [3].

10.4 Assessing Episodic QoE

In contrast to momentary QoE, quality attributed to an episode of usage is commonly judged directly after the end of the episode, may it be after watching a video clip or movie, or after finishing a telephone or video call. Such ratings of remembered QoE collected at the end-point of an episode may be different from the remembered QoE judged upon at a later point in time, where effects such as distraction or contextual transformation of QoE come into play. Still, such an episode-final quality rating may not be easily related to momentary QoE experienced during the episode: For the case of time-varying transmission characteristics, the averaged momentary QoE, i.e. the total utility (cf. Sect. 10.1), is in many cases not an appropriate estimate of the quality of the whole episode, as given by an episode-final judgment. Usually, such an average is too optimistic (cf., e.g., [13, 46] for speech quality, and [20] for video quality).

The recency effect between momentary and episode-final ratings was confirmed for speech quality [14] and video quality [20]. There is evidence that the impact of recency is smaller when momentary QoE is assessed continuously, i.e. with sliders [21], so it may be advisable to assess momentary and episodic QoE separately.

A method which aims at achieving both short-term and episode-final ratings is described in [12] for speech quality. The method consists of two separate tests which are carried out on a specific type of stimulus material. The material consists of several stretches of speech which have a duration of 4–8 s (thus the typical duration of short speech stimuli), and which are related to each other by their content. 5 to 6 of such stimuli form a storyline of a *simulated conversation*, i.e. a virtual exchange of information between two parties in which the stimuli represent the contributions of one party only. The stretches of speech are first presented in a standard test set-up according to [27], this way obtaining (position independent) short-term QoE judgments for each stretch. Secondly, the stretches are presented in their logical order, with pauses of approx. 8 s between the stimuli. During each pause, the test participant is asked to orally answer a content-related question, usually in a multiple-choice fashion, to incite her concentrating on the content and engaging in a conversation-like

situation. At the end of the last stretch, the test participant is asked for an episode-final rating of the overall quality of the episode. This so-called simulated conversation test provides short-term and related episode-final QoE ratings, and thus allows to relate one to each other. The procedure has also been adapted to video calls (then putting the focus also on the visual modality [23]), and is currently considered for a future ITU-T Recommendation [25].

Although this method [12] is able to *simulate* conversations, it is methodologically difficult to assess momentary and episodic QoE in real interactive situations, as the duration of each turn cannot be controlled, and thus neither strength nor position of degradations can be systematically varied in order to obtain reliable averaged results. As a consequence, many results and models stem from data obtained in passive situations, and a valid transfer from judgments obtained this way to interactive quality cannot be guaranteed. For the example of speech quality, a recency effect found for passive listening was not replicated for the interactive session [16], and the method described above to simulate conversational structures obviously neglects the effects of echo and delay [12], which might be integrated differently from other sources of degradations like noise.

Another effect which was found to influence episodic QoE is the impact of extreme qualities [14, 20]. Consequently, models to describe episodic QoE with an integration of momentary ratings include weightings for each momentary rating, with stronger weighting for episode-final times of occurrence, and for stronger changes (or only stronger degradations, as mentioned in Sect. 10.2):

- In [10], from any individual rating of speech QoE, may it be continuous or a short-sample rating, a weighted mean is calculated, with a higher weight towards the end of the episode, and for extreme degradations.
- In [9] the asymmetric temporal delay to adjust to changes in momentary speech QoE observed by [14] (cf. Sect. 10.3) is integrated into a model of episodic QoE, that uses these estimated momentary ratings, integrates them by averaging, but also models a recency effect by taking into account the last significant degradation.
- In [12], there is also a two-step approach chosen for the speech domain. First, a weighted mean is calculated, taking into account a recency effect—in contrast to the two above with an absolute time window—and the impact of the strongest minimum is additionally subtracted as a difference to the episodic average, for estimating episode-final QoE.
- An alternative of the last model is presented in [46], using also the difference of momentary speech QoE to the average instead of absolute weighting values to obtain the weighted average.
- For picture QoE, [20] also propose a model incorporating a fixed recency effect and the strength of impairment to calculate a weighted mean.
- A simple regression model for picture QoE using continuous rating includes the contributions of the level of the extreme degradation only. Although the latest ratings (recency effect for the last 5 s) is also significant, its inclusion does not improve the model, and duration of the extreme degradation and residual mean quality do no contribute significantly at all [21].

- A prediction model for QoE of streamed video over mobile networks is referred to in Sect. 19.3.3 [40]. It uses parametric estimates of momentary QoE, which are adapted to cover context effects (e.g., [14]) and integrated to estimate overall, i.e. episodic QoE.

Applying the simulated conversation test method of [12], an evaluation of three of these models [10, 12, 46] could confirm their relevance also for scenarios with changing audio bandwidth and packet loss [36]. Applying these models unmodified even on audio-visual simulated conversation QoE, all three resulted in satisfying estimates of episodic ratings, with [10] and [46] performing better than [12] (cf. also Sect. 27.4). A comparable evaluation for audio-visual QoE including all models mentioned above showed similar results [2]. The models show slightly lower performance when estimates of momentary speech and/or video quality are used instead of subjective momentary quality ratings, depending on the type of model used for the momentary quality estimation (see e.g. [36] for a comparison, and Sects. 12.3,14.3 and 19.2 for an overview of models for estimating momentary QoE).

Incorporating also delay and echo as interaction degradations, [15] propose a simple model to estimate episodic QoE for speech telephony based on instrumental estimates of momentary QoE. In addition to echo and delay, noise and packet loss are taken into account as listening degradations. The method to elicit authentic conversational situations uses short conversation tests as described in [24, 38] to obtain episodic QoE for the interactive scenario. Although not dealing with time-varying QoE, this approach provides a method and subsequently a model to integrate listening degradations with echo and delay as a basis to study temporal aspects for realistic interactive scenarios.

As a summary, the averaged momentary ratings typically account for most of the variance explained. Based on [2, 36, 46], Pearsons' r are about 0.84–0.90 for the plain average. Adapting momentary ratings to context effects (e.g. [14] improves this a little bit (r increase of about 0.05). Accounting for recency and extreme qualities results in values typically about or even over 0.95. Of course, with instrumental estimates of momentary QoE, correlations are typically lower (r over 0.9). Still, for data which is covered already very well by the plain average, additional modeling does not improve the correlations that much further.

It seems that the recency effect itself is not as strong as the impact of the strength of a degradation [2, 20, 46], although such a conclusion is dependent on the media stimuli used, and therefore difficult to draw. Yet, both systematic effects seem to resemble the mechanism described by [30] for remembered utility, and it would be interesting to compare the models explicitly defined for episodic QoE with the peak-end rule incorporating only the most extreme experience in addition to that for the last portion.

Apart from the question of the cognitive processes involved in integrating momentary to episodic QoE, there is of course the issue of valid and reliable assessment methods. There is already much knowledge available for this area, resulting in a number of standards (e.g. [25]). Still, most methods define and recommend laboratory settings for quality assessment, although these do not represent typical (and

thus ecologically valid) usage situations. For example, assessing video quality for an entire movie in the living room of actual test participants [45] revealed a systematic difference compared to the laboratory setting. The authors conclude, that their more authentic method and test environment provides more valid results, e.g. a lesser impact of picture degradations and a stronger impact of degradations interrupting the flow of the movie.

10.5 Assessing Multi-Episodic QoE

Quality of Experience should be considered over multiple episodes as continued usage influences the user's expectations and future behavior towards a system.

Interacting with a system for the very first time, the experience made by a user is determined by his prior expectations and experience, which are used to form his internal reference. The comparison between the experience and the internal reference leads to a quality judgment of the experience. The internal reference will then be updated according to the user's individual experiences, and will influence the perception of future interactions with the system. The update process of the internal reference happens during and after each interaction with the system (see Sect. 2.3).

The change of the user's internal reference will influence not only the QoE of future episodic interactions, but also the user's acceptance and thus behavior towards the system. This includes likeliness to use and attitude towards the system, but also task selection and task solving strategies (see Sect. 2.3).

Adaptation effects have also been found in research on User Experience. It could be shown that usage behavior of a user with a system changes over longer time periods [34]: in the beginning, new features are explored and the interaction is playful, whereas with time interactions become more task-driven and practical.

Methods to assess short-term QoE are neither designed, nor suited to study multi-episodic QoE. In fact, sequence effects are frequently balanced out by randomizing the order of stimuli over multiple participants. This is due to the targeted performance comparison of different systems like codecs or media network configurations. Furthermore, some experimenters try to replace an unknown, participant-specific internal reference, which is used in the quality judgment process, by priming the participants in the beginning of the experiment using a fixed set of anchor stimuli of pre-defined levels of quality. This is especially important if participants are habituated to "better" systems than the ones under study.

Assessing multi-episodic, and thus remembered, QoE is challenging for three reasons:

1. First, the order of episodic use and their perception is important due to the update process of the internal reference.
2. Second, the time-scales that must be taken into account are greater, and thus the experiment has to be longer, sometimes spanning over days, weeks or months. Using such long experimental periods, it is very difficult to control for other

(external) factors which might also influence the quality judgment. To validly study temporal effects on multi-episodic QoE also pauses between episodic usage periods must be considered, so that the update process of the internal reference can happen under realistic conditions.

3. Third, the usage behavior is dependent on the user himself, e.g. his personal preferences, socialization, context and attitude towards the system. This will influence the user's approach to use the system, including task selection and usage patterns, and ultimately his level of acceptance.

A practical issue occurs if a system is evaluated over a long period, where participants also use other comparable systems during the same period. The user's internal reference is influenced by all systems experienced during the usage period.

First work on multi-episodic QoE which is known to us has been conducted by Duncanson in 1969 for an oversea telephony system. He could show that the remembered QoE for multiple prior episodic uses is underestimated in comparison to the judgment to a just-finished episode with the same actual performance [11]. This suggests that low-quality episodes have a greater impact on remembered QoE over multiple episodes.

In [39] a method to study multi-episodic QoE with one system over multiple days is presented. A comparable system usage is achieved by providing task scenarios for each usage episode, so that not only each episodic use is similar but also the interaction lengths and timing are comparable. This reduces the impact of the participant's behavior in the quality evaluation process. Each participant has to perform several (in this case 24) usage episodes in fixed time intervals (here twice a day) within a certain usage period (here 12 days). Two types of questionnaires are used: one for assessing the QoE after each episodic use and one to assess the multi-episodic QoE after several days. In addition, an initial interview and a final interview are conducted.

Möller et al. [39] used this method in a field study over 12 days providing two tasks per day that should be fulfilled using a video telephony system. The system performance was controlled on a day-to-day basis. Overall, 5 system performance profiles were used with a total of 56 participants.

Two effects on QoE were found: first, a recovery effect after low performance episodes, showing that prior episodes influence the QoE rating of following episodes. The recovery interval was approx. 2 days long. Second, a general rise in QoE ratings over the usage period of 12 days was noticeable. In [39] also the integration of individual episodic QoE ratings into an overall QoE judgement for the system was studied: the multi-episodic QoE judgment could be estimated by the average of episodic QoE ratings of all prior episodes only for several days, whereas it was not a good predictor for the multi-episodic QoE judgment on day 12.

This method was used in [17] to study multi-episodic QoE in a multi-service scenario addressing audio-visual entertainment and telephony. In this study, the results of [39] have been confirmed. In addition, it was found that the impact of performance limitations depends on the type and use-case of the system.

Multi-episodic QoE is of special relevance for service providers, especially in telecommunication and entertainment, because those services are used very frequently and the switching costs for customers are low [35]. Thus, it is important to provide good QoE over longer usage periods in order to avoid customer churn.

10.6 Discussion and Conclusion

With time-varying Quality of Service, the primary issues for assessing QoE are different for the three time spans presented, *momentary, episodic* and *multi-episodic*.

For *momentary* QoE, the focus lies on the instantaneous assessment using, e.g., a slider, so that a relationship between instantaneous service performance and user-perceived QoE can be determined. As a drawback of continuous assessment, the validity for authentic service usage is not ensured by such a permanent and untypical secondary task, which directs attention from content to the quality judgment process. Here, reliability and validity of alternative methods have to be studied. An alternative would be to use, e.g., physiological or non-permanent methods, like assessing only extreme moments of QoE or even actively controlling quality like presented in Sect. 10.3.

Methods to assess *episodic* and *multi-episodic* QoE do not seem to be very different from each other. However, it is important to distinguish the "mere" integration process of an episodic experience when rating the (retrospective) quality from the attitude towards a service built within multiple episodes over a longer time span. This attitude towards the service is much more affected by individual needs and preferences and stronger linked to personal life. Therefore, authentic situations are much more important when assessing *multi-episodic* QoE compared to established and valid laboratory methods for *episodic* QoE. Multi-episodic QoE assessment thus requires field tests, although this results in less control over the test set-up, e.g. in creating specific quality profiles.

Relying on valid data might not only provide enough material, i.e. "profiles" of time-varying service performance, to build valid models, but will also provide insight into realistic distributions of time-varying performance to validate such models even better. Thus, the ongoing merging of service monitoring with QoE research is expected to solve several issues presented here.

Prediction models incorporating empirical results are at hand for short sample aggregations of momentary QoE and episodic QoE. However, there are still many relevant factors that are not considered to satisfy the demands of applications like monitoring or planning. The principal problem is the trade-off between incorporating factors like content, attention etc. in a reasonable, i.e. generic and scalable, way. And for multi-episodic QoE research related to modeling has just started. Here, especially context factors affecting attention (secondary tasks of users, interruptions, parallel usage of different devices) or the attitude towards the service (availability, mobility, and even more than for episodic usage: content) will also play an important role.

References

1. Baddeley A (2005) Human memory: theory and practice. Psychology Press, Hove (revised edn.)
2. Belmudez B, Lewcio B, Möller S (2012) Call quality prediction for audiovisual time-varying impairments using simulated conversational structures. Acta Acustica united Acustica 99:792–805
3. Borowiak A, Reiter U (2013) Long duration audiovisual content: impact of content type and impairment appearance on user quality expectations over time. In: Proceedings of 5th International Workshop on Quality of Multimedia Experience (QoMEX), Klagenfurt, pp 200–205
4. Borowiak A, Reiter U, Svensson UP (2012) Evaluation of audio quality requirements over extended periods of time using long duration audiovisual content. In: Advances in multimedia information processing. PCM, Lecture notes in computer science, vol 7674, pp 10–20
5. Borowiak A, Reiter U, Svensson UP (2012) Quality evaluation of long duration audiovisual content. In: Proceedings of the 9th annual IEEE consumer communications and networking conference. Special session on quality of experience (QoE) for multimedia communications, Las Vegas, pp 353–357
6. Borowiak A, Reiter U, Tomic O (2012) Measuring the quality of long duration AV content. Analysis of test subject/time interval dependencies. In: EuroITV—Adjunct Proceedings, Berlin, pp 266–269
7. Bouch A, Sasse MA (2000) The case for predictable media quality in networked multimedia applications. In: Procedings of ACM/SPIE multimedia computing and networking (MMCN), San Jose, pp 188–195
8. Buchinger S, Robitza W, Nezveda M, Sack M, Hummelbrunner P, Hlavacs H (2010) Slider or glove? Proposing an alternative quality rating methodology. In: Proceedings of the 5th international workshop on video processing and quality metrics for consumer electronics (VPQM), Scottsdale, Arizona
9. Clark A (2001) Modeling the effect of burst packet loss and recency on subjective voice quality. In: Proceedings of the internet telephony workshop (IPtel 2001), New York
10. Delayed Contribution ITU-T,D.064 (1998) Testing the quality of connections having time varying impairments. Source AT&T, USA (J. H. Rosenbluth) ITU-T Study Group 12 Meeting. International Telecommunication Union, Geneva
11. Duncanson JP (1969) The average telephone call is better than the average telephone call. Public Opin Q 33(1):112–116
12. ETSI TR 102 506: Speech Processing (2007) Transmission and quality aspects (STQ); Estimating speech quality per call. European Telecommunications Standards Institute, Sophia Antipolis
13. Gray P, Massara R, Hollier M (1997) An experimental investigation of the accumulation of perceived error in time-varying speech distortions. In: Proceedings of audio engineering society, 103rd convention, New York
14. Gros L, Chateau N (2001) Instantaneous and overall judgements for time-varying speech quality: assessments and relationships. Acta Acustica united Acustica 87:367–377
15. Guéguin M, Le Bouquin-Jeannès R, Gautier-Turbin V, Faucon G, Barriac V (2008) On the evaluation of the conversational speech quality in telecommunications. EURASIP J Adv Sig Proc 8:1–15
16. Guéguin M, Gautier-Turbin V, Gros L, Barriac V, Le Bouquin-Jeannès R, Faucon G (2005) Study of the relationship between subjective conversational quality, and talking, listening and interaction qualities: towards an objective model of the conversational quality. In: Proceedings of measurement of speech and audio quality in networks, Prague
17. Guse D, Möller S (2013) Macro-temporal development of QoE: impact of varying performance on QoE over multiple interactions. In: Proceedings of AIA-DAGA conference on Acoustics, Merano, Deutsche Gesellschaft für Akustik, Berlin

18. Gutierrez J, Perez P, Jaureguizar F, Cabrera J, Garcia N (2011) Subjective evaluation of transmission errors in IPTV and 3DTV. In: Proceedings of visual communications and image processing, Tainan
19. Hamberg R, de Ridder H (1995) Continuous assessment of perceptual image quality. J Opt Soc Am A 12:2573–2577
20. Hamberg R, de Ridder H (1999) Time-varying image quality: modeling the relation between instantaneous and overall quality. SMPTE Motion Image J 108:802–811
21. Hands D, Avons S (2001) Recency and duration neglect in subjective assessment of television picture quality. Appl Cognitive Psychol 15:639–657
22. Ito TA, Larsen JT, Smith NK, Cacioppo JT (1998) Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. J Pers Soc Psychol 75:887–900
23. ITU-T Contr. COM 12-340 (2012) Methodology for the assessment of audiovisual quality for simulated video calls. ITU-T Study Group 12 Meeting. International Telecommunication Union, Geneva
24. ITU-T Contr. COM 12-35 (1997) Development of scenarios fort a short conversation test. ITU-T Study Group 12 Meeting. International Telecommunication Union, Geneva
25. ITU-T Contr. COM 12-38 (2013) Proposal for a subjective method for simulated conversation tests addressing speech and audio-visual call quality (P.ACQ). ITU-T Study Group 12 Meeting. International Telecommunication Union, Geneva
26. ITU-R Recommendation BT.500-7 (1996) Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, Geneva
27. ITU-T Recommendation P.800 (1996) Methods for subjective determination of transmission quality. International Telecommunication Union, Geneva
28. ITU-T Recommendation P.910 (2008) Subjective video quality assessment methods for multimedia applications. International Telecommunication Union, Geneva
29. ITU-T Recommendation P.911 (1998) Subjective audiovisual quality assessment methods for multimedia applications. International Telecommunication Union, Geneva
30. Kahneman D (1999) Objective happiness. In: Kahneman D, Diener E, Schwarz N (eds) Well-being: the foundations of hedonic psychology. Russel Sage, New York, pp 3–25
31. Kahneman D (2000) Experienced utility and objective happiness: a moment-based approach. In: Kahneman D, Tversky A (eds) Choices, values and frames. Cambridge University Press, New York
32. Kenny DA (2004) PERSON: a general model of interpersonal perception. Pers Soc Psychol Rev 8:265–280
33. Kokotopoulos A (1997) Subjective assessment of a multimedia system for distance learning. In: Multimedia applications, services and techniques—ECMAST, Lecture notes in computer science, vol 1242, pp 395–408
34. Kujala S, Roto V, Väänänen-Vainio-Mattila K, Karapanos E, Sinnelä A (2011) UX curve: a method for evaluating long-term user experience. Interact Comput 23(5):473–483
35. Lee J, Lee J, Feick L (2001) The impact of switching costs on the customer satisfaction-loyalty link: mobile phone service in France. J Serv Mark 15:35–48
36. Lewcio B (2013) Management of speech and video telephony quality in heterogeneous wireless networks. Doctoral dissertation, Technische Universität zu Berlin, Berlin
37. Liu T, Cash G, Narvekar N, Bloom J (2012) Continuous mobile video subjective quality assessmentusing gaming steering wheel. In: Proceedings of the 6th international workshop on video processing and quality metrics for consumer electronics (VPQM), Scottsdale, Arizona
38. Möller S (2000) Assessment and prediction of speech quality in telecommunications. Kluwer Academic Publishers, Boston
39. Möller S, Bang C, Tamme T, Vaalgamaa M, Weiss B (2011) From single-call to multi-call quality: A study on long-term quality integration in audio-visual speech communication. In: Proceedings of interspeech, florence, International SpeechCommunication Association, pp 1485–1488
40. Next generation mobile networks alliance (2013). In: Wennesheimer M, Robinson D (eds) Service quality definition and measurement—a white paper, Frankfurt, Germany

41. Pinson M, Wolf S (2003) Comparing subjective video quality testing methodologies. Proc SPIE 5150:573–582
42. Raake A (2006) Short- and long-term packet loss behavior: towards speech quality prediction for arbitrary loss distributions. IEEE Trans Audio Speech Lang 14(6):1957–1968
43. Recommendation ITU-T,P.880 (2004) Continuous evaluation of time varying speech quality. International Telecommunication Union, Geneva
44. Roto V, Law E, Vermeeren A, Hoonhout J (Eds) (2011) User experience white paper: bringing clarity to the concept of user experience. Result from Dagstuhl seminar on demarcating user experience, 15–18 Sep 2010. www.allaboutux.org/uxwhitepaper
45. Staelens N, Moens S, Van den Broeck W, Mariën I, Vermeulen B, Lambert P, Van de Walle R, Demeester P (2010) Assessing quality of experience of IPTV and video on demand services in real-life environments. IEEE Trans Broadcast 56:458–466
46. Weiss B, Möller S, Raake A, Berger J, Ullmann R (2009) Modeling call quality for time-varying transmission characteristics using simulated conversational structures. Acta Acustica united Acustica 95(6):1140–1151