MODELLING MULTI-EPISODIC QUALITY PERCEPTION FOR DIFFERENT TELECOMMUNICATION SERVICES: FIRST INSIGHTS

Dennis Guse, Benjamin Weiss, Sebastian Möller

Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
{Dennis.Guse, Benjamin.Weiss, Sebastian.Moeller}@telekom.de

ABSTRACT

Most studies on the perceptual quality evaluation of multimedia services are limited to short media samples or episodes, although typical usage of services occurs in a more regular way over longer periods of exposure (weeks, months or years). So far, it is unclear how the experience of individual episodic use is integrated to form an experience of one service as a whole, spanning over multiple usage episodes.
In this paper, we present two studies on the quality experienced during multi-episodic usage of different media and communication services, and first approaches to model multi-episodic quality judgments. An analysis of the data shows that the particular instance of episodic judgment strongly affects the integration function, which makes it difficult to derive models of general validity. Ways to overcome the observed limitations are discussed.

Index Terms—Quality of Experience, episodic use, Multi-episodic QoE, Temporal Integration

1. MOTIVATION AND INTRODUCTION

The evaluation of Quality of Experience (QoE) of multimedia services is mostly limited to the consideration of individual usage episodes or parts-of-episodes. For example, the quality of telephony services is frequently evaluated on the basis of speech files of 4…8 s, and in rare cases on the basis of entire conversations of 2…4 min length. In a similar vein, Audio on Demand (AoD) or Video on Demand (VoD) services are evaluated on the basis of short audio or audiovisual clips of several seconds, and more rarely on the basis of longer episodes (several minutes up to a typical length of a movie). One exception is a study done by Staelens et al. [1] focusing on full-length movie consumption in the user's home environment.

In contrast to these evaluation paradigms media and communication services are used on a regular basis spanning far beyond the considered time period of laboratory quality evaluations. Therefore it cannot be deduced how the user's expectations and usage behavior adapts over longer timespans, which might introduce changes in the internal quality references. Kujala et al. [2] studied the adaption of mobile phone usage over a period of one year and were able to show that users adapt their interaction from explorative and playful to being task-driven. However, very limited knowledge exists on how the memory of experiences made during longer usage periods affects the quality rating at a given time.

For shorter periods of time related to a single usage episode (e.g. one telephone call or one movie), several cognitive effects have been observed, see e.g. [3, 4, 5] for an overview. Typically, averaging of the quality judgments of all segments of an episodic stimulus only explains part of the variance of the episodic judgments. Temporal effects that are known in the literature are the *primacy effect*, *recency effect*, *peak-rule* [6] and *peak-end-rule* [7]. The first two effects focus on the higher importance of the beginning and, respectively, the end of an episode (or a part of an episode) with regard to the episodic judgment whereas the peak-rule highlights the importance of the lowest quality. The peak-end rule combines peak-rule and recency effect. These cognitive effects have been observed for different types of multimedia services, but could not necessarily be confirmed to appear for different types of services and (if at all) sometimes with different effect sizes.

The mentioned effects have been used for predicting episodic judgments using judgments of individual stimuli occurring during an episode. For example, call-quality models have been developed predicting the judgment at the end of a conversation and approaches been made to use "simulated conversations" composed of pre-defined parts to avoid an impact of user behavior [6]. Taking the average of all individual stimulus judgments as a basis (resulting in a correlation using Pearson's R of around 0.90), the prediction performance could be improved to around 0.98 by including a recency effect and a model of individual strongly negative ratings. Similar results were obtained for video telephony [3], increasing the correlation from 0.93 for the plain average to approx. 0.97 for models including recency and strongly negative samples effects.

Although some of the mentioned effects have also been observed for longer periods of time (e.g. the peak-end-rule for integrating emotions and well-being [7]), there have been very few investigations about their impact on the experience during continued use of multimedia services over longer usage periods.

Duncanson [8] observed that the quality rating of a single just finished telephone call is more positive than the quality rating of subjectively averaged telephone calls of the

similar performance made in the past. One reasonable explanation is that a user remembers past episodes of poor quality more readily than episodes of good quality, and attaches higher significance to poor episodes in the multi-episode judgment process.

A more recent study on multi-episodic quality of video telephony [9] showed that users did not simply average over individual episodic experiences, but instead have the tendency to slowly increase their quality rating over the usage period, as long as the functionality of the service was maintained. In addition, it was observed that users remembered bad quality episodes happening within a day or two, but do not necessarily take them into account in their multi-episodic judgment. It was concluded that the surprisingly optimistic multi-episodic ratings, which contradict the recency and peak-end effect, might have been caused by cognitive effects different from those relevant in an episodic time frame.

A comparable study [10] using VoD and speech telephony showed that episodic QoE judgment for severe degradations like high packet-loss on speech quality is higher than expected, but reflected properly in the multi-episodic QoE judgments.

In this paper, we will take up the results of the video telephony study [9] and complement them with results from a new empirical study addressing AoD and VoD. Both studies were conducted in the field in which participants made regular use of the addressed service(s) during an observation period of 12…15 days, and to provide ratings on the quality of the service(s) *experienced so far* after several usage episodes. In both studies, only the effect of media or communication quality was addressed; other factors which might have an impact on multi-episodic quality judgments such as the usage situation, motivation, environmental conditions, costs and account were not considered; see [11] for a discussion of these factors.

The studies will be described in Section 2, and the obtained subjective results will be analyzed in Section 3. In Section 4, we will predict multi-episodic ratings on the basis of ratings of the individual episodes. Some general conclusions will be summarized in Section 5, leading us to suggestions for future work, which will be outlined in Section 6.

2. EMPIRICAL STUDIES

In order to keep other factors potentially carrying an influence on multi-episodic quality as far as possible constant, but still allowing for an observation period of roughly 2 weeks, we opted for a semi-controlled set-up in the participant's home environment. Participants had to fulfill certain tasks (such as performing a video telephony call over the Skype service, or watching a soap opera and listening to an audio book) once or twice a day, and had to rate the quality of each episodic user directly following the interaction (episodic judgments). In addition, participants were regularly asked to judge the quality of the episodic usages up to the current moment (multi-episodic judgments). In both studies the quality judgments were given on a continuous 7pt scale [12] shown in Figure 1.

Figure 1: Continuous 7pt scale (labels in German) used in Study 1 and 2. Labels from left-to-right: extremely bad, bad, poor, fair, good, excellent and ideal [12].

2.1. Study 1: Video telephony

In Study 1 (confer [9]) calls were set up through a purposely-modified Skype client, which was installed on the computers of the test participants at their home. In the client, the joint audio-and-video bandwidth was artificially restricted up to a certain maximal bandwidth (62.5, 18.75 or 4 kbps) for particular days of the study. The bandwidth limitation affected both audio and video signals; however, due to the larger proportion of bandwidth necessary to transmit video, the impact on video was already noticeable for the medium bandwidth, whereas audio quality was mostly affected only for the low bandwidth. Participants were selected to have fast and reliable DSL connections and flat rates, so that neither additional bandwidth restrictions nor effects of costs were expected.

Within the study period of 12 days, each pair of participants had to carry out two conversations daily: one between 6h and 15h and one between 15h and 24h. Within these periods, the bandwidth restrictions according to one of five pre-defined profiles applied as shown in Table 1.

Profile	Bandwidth (High/Low)	Low episodes
1	62.5 kbps / 4 kbps	5 - 6
2	62.5 kbps / 4 kbps	5 - 8, 17- 20
3	62.5 kbps / 4 kbps	5 - 6
4	18.75 kbps / 4 kbps	5 - 6, 17 - 18
5	62.5 kbps / -	-

Table 1: Performance profiles of Study 1.

For each call, test participants were given a scenario which should ensure that conversations had approximately the same length and structure, and which engaged them in a short role play of everyday situations (booking a train ticket, appointment with the doctor, etc.) [13]. Calls lasted for 9 min on average ranging from 3 to 12 min. After each call, participants rated the overall quality, the audio quality and the video quality of that particular call. After the 4th, 14th and 24th call (i.e. after 2, 7 and 12 days), the multi-episodic judgment for the same dimensions needed to be rated. The following analyses will be limited to the overall quality.

56 participants, aged 14-64 years (27 male, 29 female with mean 28 years) participated in this study and were assigned pairwise to one of the five profiles (16…10 participants per profile). Further details on the study can be found in [9].

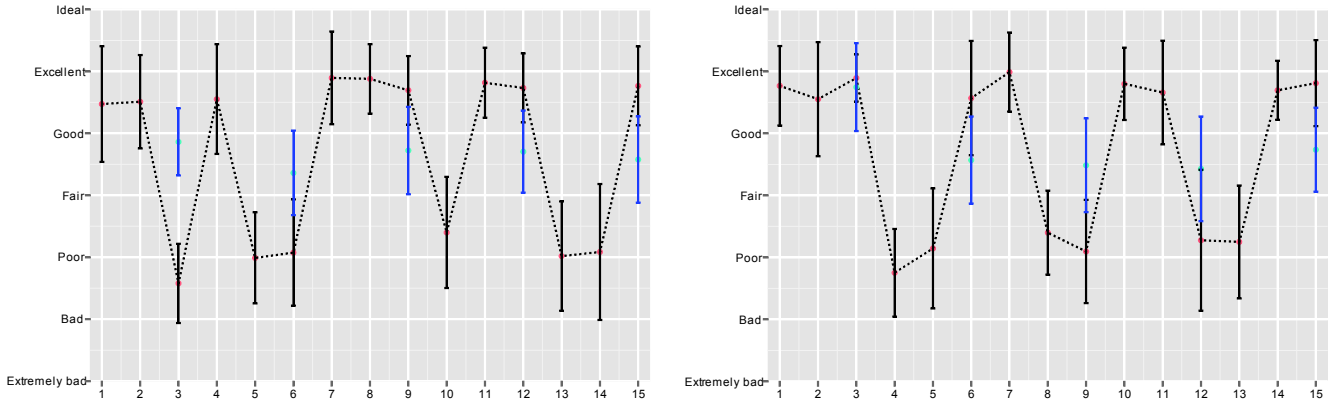

Figure 3: Mean episodic (red) and multi-episodic quality ratings (blue) with standard deviation for Study 2 combining AoD and VoD ratings. Left profile 1 and right profile 2. Y-axis denotes the quality rating on the 7pt scale and X-axis the episode.

2.2. Study 2: Audio and Video on Demand

In Study 2 AoD and VoD were used in a mobile setup. Participants were provided with a Samsung Galaxy S II LTE (GT-i9210), which has 4.3" screen (AMOLED+) with a resolution of 480x800, and a pair Sure SRH240 headphones.

The two services provided two different performances namely *high performance* (HP) and *low performance* (LP). In the HP case audio was encoded with MP4A at 320 kbps and video with h264 at 5 Mbps at device resolution. In the LP case the audio was encoded using GSM full-rate and a video bandwidth of 250 kbps.

The study period consisted of 15 consecutive days using the AoD service between 7 h and 13 h and the VoD service between 17 h and 23 h. For each interaction consecutive content was defined with a length of 13 to 17 min. After each episodic service use the participant rated the overall quality. After every 3rd episodic uses in addition the multi-episodic quality for each service needed to be judged.

The two service performance profiles were applied asymmetrically, e.g. participants having profile 1 for AoD had profile 2 for VoD and vice versa. In each profile 6 out of 15 episodic usage where provided in LP.

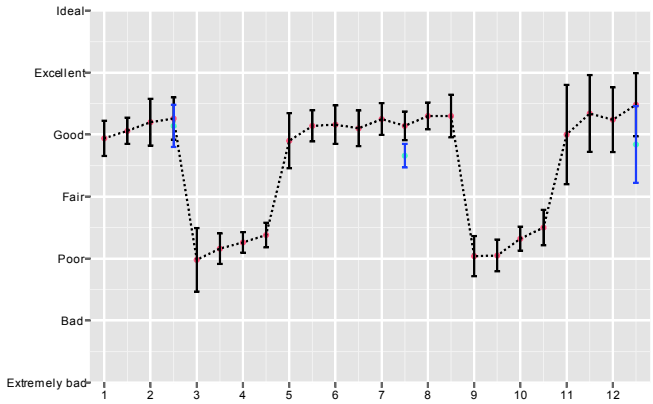

Figure 2: Mean episodic (red) and multi-episodic quality ratings (blue) with standard deviation for Study 1, Profile 2. Y-axis denotes the quality rating on the 7pt scale and X-axis the episode.

21 participants took part in this study, aged 20-35 with an average age of 25.9 (11 females, 10 males). 11 participants were assigned for AoD with profile 1 and VoD with profile 2 and 10 vice versa.

3. DATA ANALYSIS

In the following, we will briefly review the results of the episodic and multi-episodic quality judgments. As the results of Study 1 have already been analyzed with respect to the bandwidth profiles in [9], we will only show an exemplary profile here, and will concentrate on the results of the (new) Study 2. Figure 2 shows the episodic and multi-episodic overall quality judgments for Profile 2 of Study 1, which showed drops in bandwidth on days 3-4 and 9-10. The episodic quality ratings are strongly affected by the bandwidth drops during these days, but also within a time window of 1...2 days after the drop, indicating a small recovery effect (see [9]). The drop in multi-episodic quality ratings on day 7 (episode 15) and day 12 (episode 25) clearly reflect the impact of the degraded episodes before. Interestingly, the 6 non-degraded episodes before the multi-episodic rating after episode 15 suggests that a longer interval might be necessary to recover from the 4 degraded episodes. The same effect can be found at the multi-episodic rating after episode 25 on the 14th day.

The results of the two profiles used in Study 2 are shown in Figure 3. The service performance was experienced by the participants for each individual episode as designed. An increase over the study period with regard to episodic judgments could not be observed as in Study 1 (confer [9]). Due to the perceived differences in episodic service performance a higher impact on the multi-episodic judgments after the 6th episodic use was expected but not found in the experiment. This indicates an adaption of the participants towards the variation in service performance.

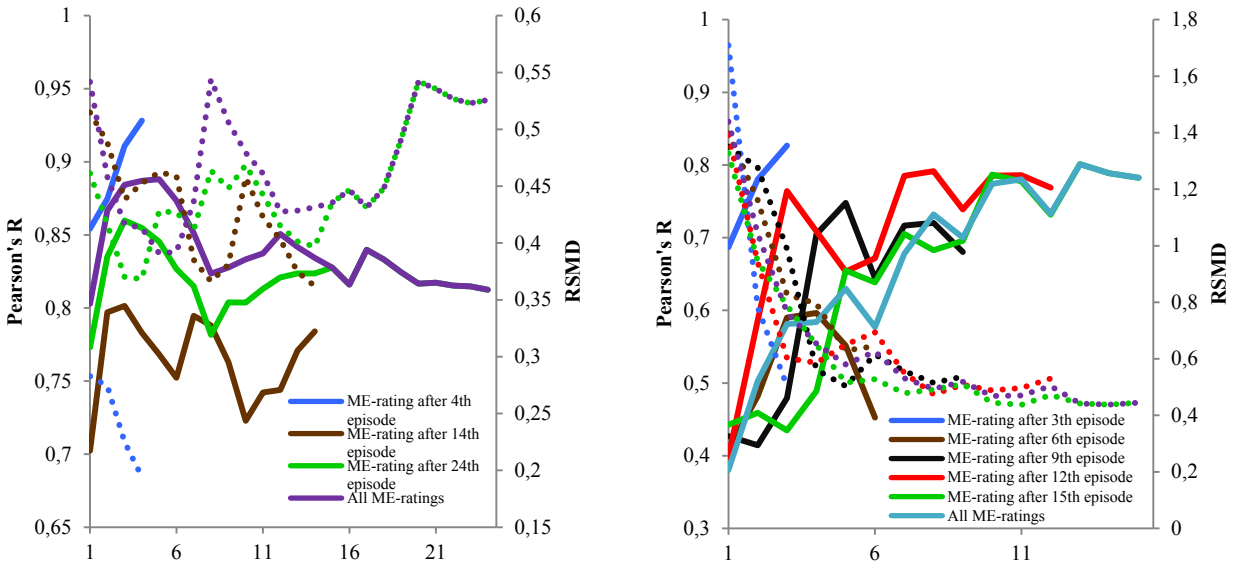

Figure 4: WA for multi-episodic quality judgment prediction depending on the window length. Study 1 (left) and 2 (right). Correlations (solid lines) and RMSD (dashed lines).

The largest difference between two consecutive multi-episodic judgments is found between the 3rd and 6th use in profile 2 were the first two consecutive degraded episodes occurred. This shows that the first longer degradation has a high impact on the multi-episodic quality despite the HP episodic use on day 6 directly before the multi-episodic judgment. This effect is also visible in profile 1 between the 3rd and the 6th rating, but with a smaller effect size, which seems to be due to the degradation of the 3rd episode. Anyhow, the impact of the degradation in profile 1 on the 3rd rating is smaller than in profile 2 after the 6th episode, indicating that the first degraded episode in profile 1 is perceived and judged different than the two consecutive degraded episodes in profile 2.

Overall, for both profiles a faster adaption of the multi-episodic judgments with regard to LP episodes was expected showing that temporal integration is taking place slowly. An indication is given by the small recovery in profile 1 between the 6th and the 9th multi-episodic judgment due to three consecutive HP episodes. However, even this period only leads to an increase of 0.5 point on the 7pt scale.

In both studies the multi-episodic judgments are affected by the degraded episodes. In Study 1 however, the impact on multi-episodic judgments is smaller compared to Study 2 although the quality rating of degraded episodes for the shown profile are roughly comparable.

4. MODELLING APPROACHES

Prior work [4, 5] on temporal modeling for short-term episodic quality has shown that the average is useful as a baseline, but the prediction performance can be improved by incorporating the recency effect and the peak-rule.

Three modeling approaches for multi-episodic quality prediction using episodic quality judgments are evaluated: the windowed average (WA) and linear weighted moving average (LWMA) to allow for an impact of the recency effect and, finally, the full-average minus the weighted lowest episodic judgment (AP) to include a peak-effect. As performance metrics Pearson's R and RMSD are used.

The results for WA are shown in Figure 4 for Study 1 and 2. In case of Study 1 averaging over the complete history achieves a correlation from 0.75 to 0.93. The prediction performance decreases for later multi-episodic judgments (confer [9]). For Study 2 the correlation shows a higher variation ranging from 0.39 to 0.78. By increasing the win-

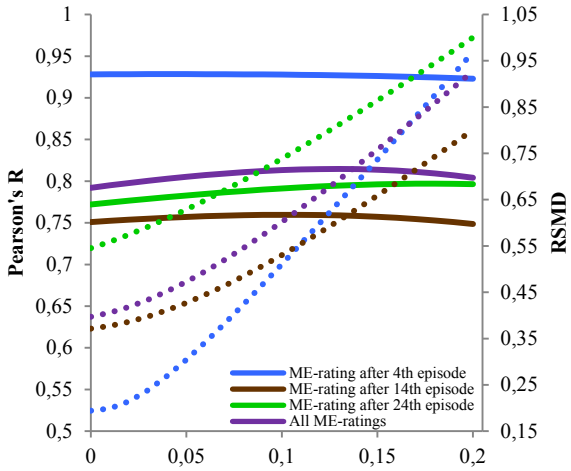

Figure 5: AP for multi-episodic quality judgments prediction depending on the weighting for the minimum for Study 1. Correlations (solid lines) and RMSD (dashed lines).

dow size from 1 to 7 days a substantial decrease in RMSD is achieved leading to a correlation of 0.65...0.78. Increasing the window size further leads to a slight increase in correlation and very slight decrease in RMSD. It is evident that the prediction performance depends on the temporal position of the multi-episodic judgment towards the LP episodes and therefore smaller windows sizes than 3 are not performing very well.

For WA in case of Study 1 a window length of 1.5 days, respectively 3 episodes, leads to $R>0.8$ for all three multi-episodic judgments. Allowing for a recency effect by using LWMA does only slightly improve the prediction performance – for one instance up to $R\approx0.85$ –, but does not reduce the difference between the temporal positions of the multi-episodic judgments. In case of Study 2 LWMA does not improve the prediction performance.

Investigating the impact of the peak-effect of the worst quality event by applying AP reduces the prediction performance if used at all for Study 1 and 2 in a similar manner. By introducing this factor the prediction performance degrades monotonously for RMSD as shown in Figure 5 for Study 1. For Study 1 the RMSD increases by up to 0.6 points for a minimum weighting factor of 0.2 compared to WA. Unexpected was the large increase for the prediction of the multi-episodic judgment on the 2nd day for Study 1 as degraded episodes were only scheduled later in all profiles. Similar results were obtained for Study 2 using AP.

5. CONCLUSIONS

Approaches to model multi-episodic quality in a period of approximately 2 weeks cannot be directly deduced from episodic-models. Well-known short-term effects like the recency effect and the peak-rule did not improve the prediction accuracy. Further research is required to understand how quality evolves over meaningful time periods from the user's perspective that cannot be deduced from short stimuli in the range of seconds.

The success of modeling multi-episodic quality depends heavily on the available profiles and the position of the degraded episodes towards the multi-episodic judgments as shown by Study 2. Therefore, we believe that at least five profiles should be used in a modeling approach to cover a sufficient number of possible degradation patterns. Moreover, the profiles should resemble realistic and meaningful scenarios.

6. FUTURE WORK

The understanding of effects in multi-episodic judgments and modelling requires more studies applying different profiles. A standardization of a unique study procedure would enable gathering more data, so that the described modelling procedure as a basis can be followed.

The profiles should be designed such that the expected effects can be studied individually. This includes for example the frequency of episodes in a period, so it can be distinguished if either the number of degraded episodes or the temporal distance is more meaningful. Different types of services as well as usage situations must be taken into account using the same profile to investigate if there are service type related characteristics are important.

In addition, the multi-episodic quality should be extended to not only investigate one service only, but also takes services bundles into account and how the multi-episodic quality of service bundles arises.

7. REFERENCES

[1] Staelens, N., Stefaan, M., Van den Broeck, W., Marien, I., Vermeulen B., Lambert, P., Van de Walle, R., Demeester, P., Assessing Quality of Experience of IPTV and Video on Demand Services in Real-Life Environments, in: *IEEE Transactions on Broadcasting* 56, no. 4, 2010.

[2] Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., Sinnelä, A., UX Curve: A Method for Evaluating Long-Term User Experience, in: *Interacting with Computers* 23, no. 5, 2011.

[3] Weiss, B., Guse, D., Möller, S., Raake, A., Borowiak, A., Reiter, U., *Temporal Development of Quality of Experience*, in: Quality of Experience: Advanced Concepts, Applications and Methods (S. Möller and A. Raake, eds.), Springer, Heidelberg, 2014.

[4] Weiss, B., Möller, S., Raake, A., Berger, J., Ullmann, R., Modeling Conversational Quality for Time-varying Transmission Characteristics, *Acta Acustica united with Acustica* 95:1140-1151, 2009.

[5] Belmudez, B., Lewcio, B., Möller, S., Call Quality Prediction for Audiovisual Time-Varying Impairments Using Simulated Conversational Structures, *Acta Acustica united with Acustica* 99:792-805, 2013.

[6] Ito, T. A., Larsen, J.T., Smith, N.K., Cacioppo, J.T., Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations, *Journal of Personality and Social Psychology* 75:887–900, 1998.

[7] Kahneman, D., Objective Happiness, in: *Well-Being: The Foundations of Hedonic Psychology* (D. Kahneman, E. Diener and N. Schwarz, eds.), pp. 3–25. Russel Sage, New York, 1999.

[8] Duncanson, J. P., The Average Telephone Call Is Better than the Average Telephone Call, *The Public Opinion Quarterly* 33(1):112–116, 1969.

[9] Möller, S., Bang, C., Tamme, T., Vaalgamaa, M., Weiss, B., From Speech Quality to Service Quality: A Study on Long-term Quality Integration in Audio-Visual Speech Communication Services, in: *Proc. 12th Ann. Conf. of the Int. Speech Communication Assoc. (Interspeech 2011)*, 27-31 Aug., Firenze, 2011.

[10] Guse, D., Möller, S., Macro-temporal Development of QoE: Impact of Varying Performance on QoE over Multiple Interactions, AIA-DAGA 2013, Mera, 18-21. March 2013.

[11] *Qualinet White Paper on Definitions of Quality of Experience*. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003) (Le Callet, P., Möller, S. and Perkis, A., eds.), Lausanne, Version 1.2, Novi Sad, March 2013.

[12] ITU-T Recommendation P.851, Subjective quality evaluation of telephone services based on spoken dialogue systems, *International Telecommunication Union*, Geneva, 2003.

[13] ITU-T Recommendation P.805, Subjective evaluation of conversational quality, *International Telecommunication Union*, Geneva, 2007.