From Single-Call to Multi-Call Quality: A Study on Long-term Quality Integration in Audio-Visual Speech Communication

Sebastian Möller [1], *Chihuy Bang* [1], *Teele Tamme* [2], *Markus Vaalgamaa* [3], *Benjamin Weiss* [1]

[1] Quality and Usability Lab, TU Berlin, Germany
[2] Skype Labs, Skype, Tallinn, Estonia
[3] Skype Labs, Skype, Helsinki, Finland

sebastian.moeller@telekom.de, chihuy.bang@gmail.com, teele.tamme@skype.net, markus.vaalgamaa@skype.net, bweiss@telekom.de

Abstract

Speech quality is commonly assumed to be the most important factor for the quality of a speech communication service and solution. However, little is known about how the quality experienced during individual calls forms the quality perception of an entire service or solution. Taking the example of an audio-visual IP-based communication solution, a long-term study is presented in which we analyze this relationship in a controlled setting. Results show temporal integration effects in the users' response to time-varying quality levels and prove that simple averaging of call quality scores does not provide sufficiently accurate estimations of service quality.

Index Terms: speech quality, service quality, audio-visual quality, VoIP, peer-to-peer software

1. Introduction and Motivation

With the advent of mobile VoIP telecommunication services and peer-to-peer solutions, time-varying transmission characteristics proved to be a major challenge. In fact, frame erasures in mobile networks as well as packet loss in IP-based transmission commonly result in a time-varying quality perceived by the user. The effect can be observed for Voice-over-IP (VoIP) and peer-to-peer software solutions as well as for video-telephony and conferencing applications.

Up to now, time-varying speech transmission has mainly been analyzed on a short-term and medium-term level. With "short term", we refer to time constants which are related to the human short-term memory during which verbal information is rehearsed in the phonological loop (typically 2 s in the phonological storage from which information is passed to the articulatory rehearsal component), see [1]. This duration is in the order of magnitude of the length of speech samples which are commonly used for speech and audio-visual quality assessment (4…8 s). Temporal quality integration within such short samples is usually taken care of by averaging and emphasizing negative events, see e.g. [2] and [3].

When it comes to the quality of individual communication events (usually calls), several analyses addressing the integration of time-varying characteristics have been carried out. Using a continuous instantaneous assessment method with a slider, Hansen and Kollmeier [4] as well as Gros and Chateau [5] found time constants in the reaction to temporally changing degradations. In Gros and Chateau's study with 190 s long stimuli, these time constants were asymmetrical for strong degradations (typically 10 s) vs. strong improvements (30 s). When judging the quality at the end of a speech stimulus, the time span between the occurrence of a degradation and the judgment plays a role; this is commonly referred to as the "recency" effect, although it is not clear whether the "recency" refers to a constant time window before the judgment time (sometimes called "end effect") or a relative proportion of the duration of the speech stimulus itself. Using 1…2 min simulated conversations which consist of 8 s-long speech stimuli arranged in a meaningful order, Weiss et al. [6] found evidence for the former. Their finding is also in line with the so-called "peak-end-rule" [7] which postulates that retrospective final judgments involving events with emotional arousal happening within a 10…40 min time span are mostly based on the peak (minimum or maximum judgment) and the final instantaneous rating.

When judging the quality of a communication service or solution as a whole, the mentioned time constants are obviously too short. For example, a speech communication service will be used frequently over days, weeks or potentially months, and the user integrates this past experience when making up her mind regarding the quality of the service. Only one investigation is known to the authors which integrates over such a long period of time: Duncanson [8] carried out a study in 1969 with the AT&T telephone network available at that time where he asked users to rate the quality of specific calls, and compared these ratings to ratings regarding the "usual" quality of the same service. The results showed that the "usual" quality rating was lower than the average of the individual call ratings; this was explained by a psychological bias when users were directly asked for a rating of an individual call.

In addition to transmission quality, there are of course other factors influencing the user's perception of a service, including the communication partner, the communication situation, the motivation for calling, environmental conditions like background noise, costs involved in the call, account conditions, etc.; see [9] for a discussion. In case that only the impact of transmission quality is of interest, these factors have to be held constant in order not to spoil the results.

Expanding the transmission quality with factors of user perception leads to the term "Service Quality" which involves a comparison of user's expectations with the performance. In this paper, we present a study which systematically analyses how the long-term rating of a service quality is impacted by variations on transmission quality. For this purpose, we asked users to carry out regular calls in a relatively controlled setting over a period of 12 days. This period is a typical first-time trial period and is expected to be important for the final adoption of a new service or solution. The communication solution we address is an audio-visual IP-based video calling product operated from a computer terminal (typical Skype software) which was artificially degraded to provoke time-varying quality profiles. For each of the calls, quality ratings were solicited from the test participants, and in addition, a rating on the entire service was solicited after 2, 7 and 12 days. The test set-up is explained in Section 2. The results are analyzed in Section 3, first with respect to the time-varying judgment for

each profile and then with respect to the temporal integration towards a service quality judgment. Some conclusions and proposals for future work are given in Section 4.

2. Experiment

As we want to address the impact of transmission quality on the service quality of the audio-visual communication, we need to keep the system and the usage situation constant. However, carrying out such a long-term study in the laboratory would be hardly practical, and it would impact the ecological validity of the experiment. As a compromise, we opted for an experiment which is mainly carried out at the user's home environment, in a realistic situation, however with controlled and pre-checked system and network conditions.

2.1. Experimental design

The calls were set up through a video call client which was installed on the computers of the test participants at their home. This client was a modified Skype software in which the joint audio-and-video bandwidth was artificially restricted up to a certain maximal bandwidth (62500, 18750 or 4000 bytes per second) for particular days. 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.

Test participants were selected to have fast DSL connections (> 2Mbit/s) and flatrates from their Internet Service Provider so that no additional bandwidth restrictions of effects of costs were expected. Clients were installed by the experimenter on PCs with Windows XP or newer versions, at least 1 GHz CPU clock, and at least 256 Mb RAM; each PC was equipped with a high-quality USB-connected Skype headset and a HD-resolution and high frame rate capable web camera to ensure equal audio/video capturing and presentation conditions at all sites. Transmission conditions of the individual calls were analyzed in retrospect to verify that the target bandwidth was approximately reached – two participants were removed as the actual bitrates profile were not what was designed, but for the majority of the users no major deviations from the target call settings were detected.

The test took place within a 12-days period for each pair of participants. Within this period, each pair had to carry out at least two conversations daily, one in the first half (between 6 and 15h) and one in the second half of the day (between 15 and 24h). Within these periods, the bandwidth restrictions according to a pre-defined profile (see Section 2.2) applied. Test participants were free to carry out additional private calls, which were limited by the same bandwidth restrictions applying for the particular period, however these calls were excluded from the analysis.

For each intended call within the pair, test participants were given a scenario which should ensure that conversations had approximately the same length and structure. These scenarios engaged the participants in a short role play of everyday situations (booking a train ticket, appointment with the doctor, ordering a pizza, etc.) in which around half of the information was available to one conversation partner only, so that balanced conversations are necessary to resolve the scenario. The scenarios have been developed in [9] and are now recommended in ITU-T Rec. P.805 [10]. Calls lasted for 9 min on an average, and ranged between 3 and 12 min per profile; we assume that a stable quality judgment could be established during this period in each call.

2.2. Quality profiles

For each pair of test participants, transmission conditions were symmetric and controlled according to a profile which defines the maximum transmission bandwidth available for each call. 5 profiles have been defined for this study, see Fig. 2. Most of the profiles show a constantly high or medium bandwidth for most of the calls, with dedicated drops in bandwidth for series of 2 or 4 calls. These drops were introduced to have clear periods of bad quality during the duration of the experiment to which the participants were expected to react with low quality ratings. Profile 5 contained no such drops and served as a control condition.

2.3. Rating procedure

Before the experiment, test participants filled in a briefing questionnaire with some demographic data (age, experience with Skype software, etc.). Scenarios were then provided via PDF forms which had to be filled in during the call. After each call, another PDF form solicited ratings regarding the overall quality, the audio quality and the video quality of that particular call. For this purpose, continuous rating scales as outlined in Fig. 1 were used; compared to standard 5-point category scales [11], such a scale is expected to avoid saturation effects at the scale extremities and frequently shows a more Gaussian distribution of the judgments.

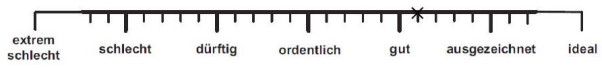

Figure 1: Rating scale. Label translations: Extremely bad; bad; poor; fair; good; excellent; ideal.

After the 4th, 14th and 24th call (i.e. after 2, 7 and 12 days), an extended questionnaire was provided. In addition to the mentioned ratings related to the individual call, this questionnaire also contained questions regarding the (overall, audio and video) quality of the service experienced so far, as well as any expected future service usage and recommendation to friends, in the format of standard Net Promotor Score questions [12]. In the following analyses, we will only analyse the quality ratings, both for the individual calls and for the solution rated up to a specific point in time.

2.4. Test participants

56 participants, aged 14-64 years (27 male, 29 female with mean 28 years) participated in this study. Participants were mostly recruited in the university area and were remunerated for their service via vouchers and via the headset and camera set they could keep after the experiment. Reflecting the university recruitment, most (35 persons) participants were experienced with Skype software. Participants were grouped in pairs of two who were familiar with each other. 16 participants were confronted with Profile 1 of Fig. 2, and 10 participants with each of the other profiles.

3. Data Analysis

Data from two subjects had to be excluded, as for them the delivered bandwidths did not follow the profile. All participants completed the study, but not all participants performed all calls (n = 6 calls missing). Judgments were analyzed with respect to extremities and outliers, but no participants were excluded due to unreasonable behavior. A two-sided t-test showed no statistically significant differences between the ratings of the experienced and the inexperienced

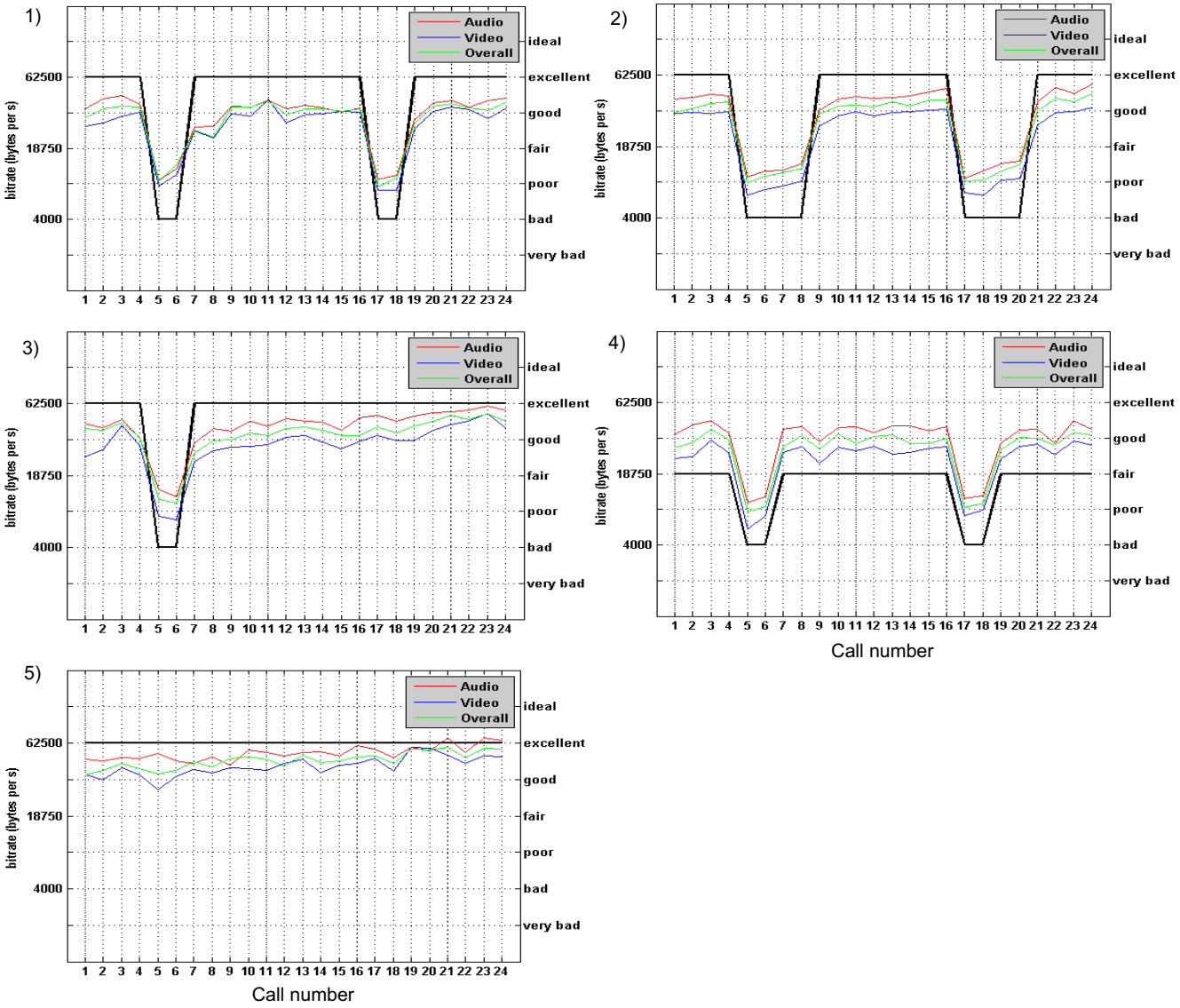

Figure 2: Maximum transmission bandwidth (solid grey line) and averaged quality ratings for the five quality profiles.

participants (p = 0.30), so that the data from both groups has been merged. No further effects of age or gender were detected. Individual judgments were averaged for the individual calls (overall MOS_O, audio MOS_A and video MOS_V) as well as for the entire solution after N calls ($SQ_{O,N}$, $SQ_{A,N}$, $SQ_{V,N}$).

3.1. Quality profiles

Fig. 2 shows that averaged MOS ratings over all participants for each bandwidth profile. In all Profiles 1-4, the drops in bandwidth are clearly associated with drops in subjective quality. As expected from the bandwidth restriction applied, video quality drop was slightly bigger than audio and overall quality, in most cases.

Overall, the ratings on all profiles show a slight tendency towards a rise from the beginning to the end (average increase 0.31 points for all profiles). Thus, ignoring the effects of bandwidth restrictions, there seems to be a slight increase in the quality judgments over the 12-days usage period. The effect is most prominent in Profile 5 and least in Profile 4 (which does not contain the highest bandwidth). It may be due to the participants getting satisfied with the offered quality level, as it offers full communication functionality. A similar effect is also noticeable in periods of low bandwidth, except for Profile 3. It can be expected that the increase will not be strictly linear, but curvilinear (e.g. in the shape of an exponential function), but further data is necessary to determine the exact shape of the rising function.

The decreases of the subjective ratings corresponding to the dips in the bandwidth are not limited to the duration of the dip, but also beyond this period quality ratings are negatively affected; this finding is similar to the one in [5], obviously with much longer time constants. Apparently, test participants remember the bad quality even in 1…2 subsequent calls, separated by 0.5…1 days from the bad experience. This effect is most prominent in Profiles 1 and 3 where short, deep dips in the bandwidth occur. Once again, this may be described by an exponential function with a negative exponent which is superposed to the otherwise step-shaped profile. In Profile 1, the time constant for this exponential decay seems to be longer for the first occurrence than for the second occurrence.

Comparing Profiles 1 and 4 which differ with respect to the maximum achievable bitrate, the judgment curves are very

similar. Apparently, test participants have the tendency to map the range of experienced quality to the scale range offered to them; this leads to different bandwidth levels resulting in the same rating curves. For all profiles, initial ratings were around the "good" quality label.

3.2. Service Quality

Averaging the MOS values over all calls of a particular profile does only provide a rough estimate of the service quality. The correlations between MOS and SQ ratings in Table 1 show that the average provides still a good estimate after only 4 calls, but that the correlation gets considerably lower after 14 and 24 calls. The effect is visible for all audio, video and overall quality ratings, to approximately the same degree.

Table 1: Correlations between averaged MOS and SQ ratings after *N* calls

N	Audio	Video	Overall
4	0.904	0.826	0.928
14	0.672	0.758	0.751
24	0.737	0.687	0.772

This relationship can be further analyzed by inspecting the scatter plots, see Fig. 3 as an example. It becomes obvious that averaging MOS values for individual calls mostly leads to too pessimistic estimations of overall service quality, in particular for high quality services and solutions. This finding is in clear contrast to the observations found in literature [2,3,5,6,8]; this difference may be due to changes in the users' quality experience within the last 40 years.

For better predicting service quality judgments, the slow increase in the quality ratings which was observed in Section 3.1 should be taken into account. In addition, the users might not remember bad quality events when judging the service quality, as long as they did not happen with the last day before the overall service quality judgment. In this case, the remembering effect apparently following an exponential quality increase after bad events has to be ruled out from the average judgment.

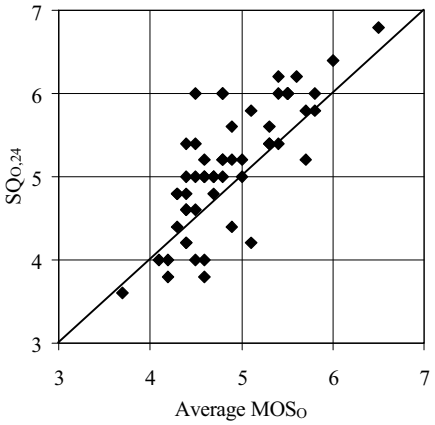

Figure 3: Relationship between $SQ_{O,24}$ and averaged MOS_O.

4. Conclusions and Future Work

We presented a first study addressing the impact of individual calls on the overall service quality of an audio-visual IP-based speech communication solution. Using time-varying quality profiles with exaggerated drops in quality for particular calls, we showed that users do not simply average over individual call experiences. Instead, they have the tendency to slowly increase their quality rating over the usage period, as long as the speech communication functionality of the service or solution could be maintained. In addition, we observed that users remembered bad quality events happening within a day or two, but do not necessarily take them into account in their service-final judgment. The effects were visible equally well for the audio, video and audio-visual quality judgments.

Our investigations were performed with a different time frame compared to most of the other experiments referenced in the introduction. The surprisingly optimistic final ratings which contradict the recency and end effects observed in other studies might be caused by cognitive effects different from those relevant in a medium time frame.

The results which could be obtained in the frame of this pilot study are apparently limited to the particular bandwidth profiles used and the observation period. In the future, we would like to collect more data with other (more natural) profiles, and with different observation periods. We would also like to separately address the impact of audio and video quality for overall service quality, as the importance of each modality has shown to depend on the interaction scenario, the degree of interactivity and alike. We will try to predict service quality judgments in a reasonably way, amending the average MOS ratings by considering the slope in the quality judgments, as well as by ruling out the remembering of negative events as long as they do not occur close in time to the service quality judgment. We would also like to transfer these findings to other speech-based services and solutions, such as spoken and multimodal dialogue system interactions.

5. References

[1] A.D. Baddeley, "Short-term Memory for Word Sequences as a Function of Acoustic, Semantic and Formal Similarity", Quarterly Journal of Experimental Psychology 18:362–365, 1966.

[2] A. Raake, "Speech Quality of VoIP – Assessment and Prediction", John Wiley & Sons, Chichester, West Sussex, 2006.

[3] P. Gray, R. Massara, M. Hollier, "An Experimental Investigation of the Accumulation of Perceived Error in Time-varying Speech Distortions", Proc. 103rd Convention of the Audio Engineering Society, New York NY, 1997.

[4] M. Hansen, and B. Kollmeier, "Continuous Assessment of Time-varying Speech Quality, J. Acoust. Soc. Am. 106:2888-2899, 1999.

[5] L. Gros, and N. Chateau, "Instantaneous and Overall Judgements for Time-varying Speech Quality: Assessments and Relationships", Acta Acustica united with Acustica 87:367-377, 2001.

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

[7] D. Kahneman, "Objective Happiness", in: Well-Being: The Foundations of Hedonic Psychology, D. Kahneman, E. Diener, N. Schwarz [Eds.], Russel Sage, New York NY, 3–25, 1999.

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

[9] S. Möller, "Assessment and Prediction of Speech Quality in Telecommunications", Kluwer Academic Publ., Boston MA, 2000.

[10] ITU-T Rec. P.805, "Subjective Evaluation of Conversational Quality", Int. Telecomm. Union, Geneva, 2007.

[11] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality", Int. Telecomm. Union, Geneva, 1996.

[12] F. Reichheld, "The One Number You Need to Grow", Harward Business Review, December 2003.