Modeling Call Quality for Time-Varying Transmission Characteristics Using Simulated Conversational Structures

Benjamin Weiss, Sebastian Möller, Alexander Raake
Quality & Usability Labs, Deutsche Telekom Laboratories, TU Berlin, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany

Jens Berger, Raphael Ullmann
SwissQual AG, Allmendweg 8, CH-4528 Zuchwil, Switzerland

Summary
This study investigates the perception of speech quality over telephone channels with time-varying transmission characteristics for simulated conversational structures. The aim is to establish a relationship between subjective quality associated with short speech samples (5–6 seconds) and quality associated with overall conversations (1–2 minutes). Two two-part experiments were conducted. In the first part of each experiment, dialog-final ratings within the temporal structure of a telephone conversation were assessed. Varying transmission characteristics were realized with ten different degradation profiles of preprocessed speech samples obtained mainly from real mobile channels to ensure authentic types of degradation. The second part was carried out to obtain separate short-term ratings of the speech samples used in the first part. Experiments 1 and 2 tested different conversation durations (1 and 2 minutes). The results demonstrate that dialog-final ratings vary with respect to the degradation profile, revealing a recency effect and a strong impact of individual bad samples. Two related models which implement these findings are presented. With these models, dialog-final quality ratings can be estimated significantly better than by plain averaging of short sample ratings (about 10% absolute improvement). They also perform better than two algorithms taken from literature. Both models can be applied to the instrumental method described in ITU-T Rec. P.862 [1], resulting in about 13% absolute improvement. They were evaluated with the results of two different experiments, which were performed independently but on the basis of our test procedure. In these experiments similar profiles but a different type of quality degradation, different sample durations, and different speech material were used. The models proved to be valid and reliable for the time span investigated (1–2 minutes) and for the profiles used. One of them is now being recommended by the ETSI STQ mobile group.

PACS no.

1. Introduction

In contrast to the traditional landline infrastructure, a typical problem of recently-established telecommunication networks (mobile telephone, voice over IP) is the time-varying quality of the transmission channel. To assess and optimize the quality of experience, it is essential to estimate perceived quality despite this variation. According to the recommended procedure of ITU-T Rec. P.800 [2], subjective quality of short samples of about 5–12 s is assessed on a five-category scale with the ACR (Absolute Category Rating) method. These ratings are being averaged over different subjects to form a Mean Opinion Score, MOS [3]. For this time span, instrumental methods can also be used to estimate MOS values in a quite valid and reliable way (cf. [1]). Such models are applied to monitor network quality under operating conditions. The model according to ITU-T Rec. P.862 already takes the temporal position of disturbances into account, as well as the specific nature of silence intervals [4]. However, these temporal adjustments to human perception are limited to a specific sample duration, and telephone conversations mostly exceed the maximum duration of 30 s recommended by ITU-T Rec. P.862.3 (cf. [5]). Averaging several short estimates does not lead to satisfying results in modeling subjective ratings, as the estimates tend to be too optimistic [6].

The experiments presented here were carried out in order to further investigate the relationship between short-term listening quality ratings and dialog-final ratings of time-varying speech quality in simulated telephone conversations. Speech quality considered in this study emerges from a situation that is conversational and exhibits longer durations, so it cannot be estimated by directly averaging auditory or instrumental measures of short samples' quality. Our aim was to build an explanatory model that can be used to predict quality of telephone conversations out of such short samples, however disregarding the talking-quality- and interaction-structure-related effects caused by talker echo and delay, see Section 2 for details. The model developed in this study has now been recommended by the ETSI STQ mobile group [7].

1.1. Quality of longer speech samples

When assessing the quality of longer samples, a recency effect has been observed: Later occurring stimuli presented in a serial order are recalled sooner and more accurately, regardless of any distraction [8]. This leads to overemphasizing events with temporal proximity to the time of judgment. One typical explanation of this finding is that recent events are dominantly stored in the short-term memory, which is limited in capacity regarding the

Received 1 January 2XXX,
accepted 1 January 2XXX.

number of events and elapsed time (typically not more than 20 s [9]). However, ratings based on retrospective evaluation are not limited to such short episodes. Studies of Kahneman show that final judgments concerning emotional arousal of events with varying time-spans (about 10–40 min) are mostly based on the peak (maximum or minimum regardless of their time of occurrence) and the last instantaneous rating (cf. [10] for an overview of the "peak-end rule").

The recency effect is already well-documented in the field of psycho-acoustics (e.g. for ratings of overall loudness, cf. [11]). It is also found for subjective quality of speech transmission channels. In a basic experiment, Jekosch [12] identified methodological issues regarding time-varying quality of speech transmission. Results of her experiment revealed significant dependency of subjects' ratings on the pattern of degradation, including the temporal position of degradations. Therefore, averaging short stimuli (6 s) did not account well for ratings of longer stimuli.

One method to assess instantaneous instead of overall quality is the use of movable sliders for continuous quality rating. This method was presented by Hansen and Kollmeier [13] with speech recordings which had been degraded with noise modulation [14]. This method is now recommended to assess instantaneous ratings in the ITU-T Rec. P.880 in addition to final MOS values for longer speech samples with varying transmission quality [15]. Gros and Chateau [16] used it to rate 190 s long stimuli with varying quality (due to IP packet loss of 0%–30%). The reaction time to changes in quality was different for strong degradations (about 10 s) compared to strong improvements (about 30 s), and it lasted much longer than in [13] (about 1 s). The authors see these long reaction times originating from the nature of the stimuli: In contrast to added noise (in [13]) the degradation of quality in [16] was characterized by isolated short-term bad events. Thus, subjects had to integrate their impression over some time to fulfill the experimental task. Gros' and Chateau's results also revealed a recency effect up to 2 min duration, which exceeds the common time window typically associated with it.

Raake [17] compared different methods to estimate time-varying quality for use in the E-model [18], which is an algorithm to predict conversational speech quality on the basis of transmission parameters typically used in transmission planning. One of them is the model proposed by Clark [19], which implements the results from [16] and will be evaluated with the data presented in this paper.

A different approach to capture the influence of time-varying quality on the experience after an entire conversation is to use ratings for short-sample ratings instead of instantaneous ones. Rosenbluth [20] proposed to use a weighted average of 8 s samples, with higher weights for samples with a position closer to the time of judgment, and for samples with stronger degradations. His model precisely estimated the overall listening-only quality of 60 s samples. This model will also be evaluated with our data.

1.2. Influence of context factors

Whereas the just mentioned results are found in listening-only situations, speech quality experienced in telephone networks is conversational in nature, so this context factor has to be considered. Gros et al. [21] examined the conversational context as well the difference between laboratory and field experiments. Results show that there were neither great differences between conversational and listening set-ups, nor between laboratory and field experiments. In conversation, ratings of different degradation profiles displayed a slightly smaller range while the individual ratings of each profile showed a higher variability compared to a listening test. These differences were however not significant. The same amount of degradation – but at a later position within the stimulus – led to a lower rating for these 2 min samples. This recency effect was significant only for the listening condition. Failing to reach statistical significance for the recency effect in conversation could be caused by the particular stimuli used in this study, or it could be a result of the distraction due to the task of speaking. Summarizing the results of [21], laboratory listening experiments seem to have produced results that can be applied to field experiments, even to environments with background noise. Differences between listening tasks and conversation tasks in the field context are not easy to interpret as they depended on the degradation profile used.

Differences between interaction modes (speaking, listening, conversation) were also studied for telephone lines with echo and delay [22]. Conversational and listening-only quality ratings depend in a complex manner on the degree and combination of echo and delay. When these degradations are combined, quality ratings during conversations are lower than during listening-only. In a recent publication, Guéguin et al. [23] define conversational quality as integration of listening and talking quality, as well as quality influenced by the interaction. The latter was shown to be characterized by delay. With this approach, high correlations are obtained for both, subjective and objective estimates of conversational quality, that are superior to the E-model.

1.3. Structure of the paper

In Section 2, a test procedure is presented that was used to simulate the conversational structure of telephone dialogs in the laboratory. Simulating conversational structures was preferred over a standard listening-only test, because a conversational situation is more representative for the purpose of assessing quality of telephone calls. It was preferred over a real conversation test in order to control where the degradations occur in the dialog structure (see also Section 2).

Each rating followed immediately after the last sample of a dialog (dialog-final). Using this procedure, two experiments have been carried out. We develop models which predict the experimental data in Section 3.1. These models use typical short samples (5–6 s) to cover short-term ratings as input, and estimate dialog-final quality judgments of the simulated dialogs as output. Their performance is compared to the one of two existing models in Section 3.2. In addition to auditorily-obtained MOS values, the models are also tested with MOS values estimated by ITU-T Rec. P.862, Section 3.3. Two further experiments have been carried out to perform an independent verification of our models (see Section 4). They were conducted with a different type of quality degradation, different sample durations, and with different speech material. The conclusions and some consequences for future work are summarized in Section 5.

2. The relationship between instantaneous and dialog-final ratings

Two experiments were conducted. Both were divided into two parts. In the first part of the experiments, the perceived quality of simulated telephone conversations was assessed. Subjects had to listen to one or two short sentences of a normal conversation, and verbally answer questions regarding the content of the sample they just heard (see the Chapter 2.1 for more information).

One of such simulated telephone conversations is called "dialog". It consists of 5 samples and the pauses used for answering the questions. After the last sample, participants had to rate the quality of this entire dialog on an ACR scale. The answering part was introduced to come close to a real conversation with its turn-takings, and to distract subjects from concentrating on the quality until rating it. Subjects were instructed to try to put themselves into the position of an interlocutor. In Experiment one the 5 short samples had a duration of 5–6 s, resulting in one simulated dialog of 1 min duration, including the participants interaction. In Experiment two, each simulated dialog lasted 2 min, as two consecutive samples were presented in one longer sample (12 s) and the time reserved for participants answering the questions was doubled as well.

In the second part of both experiments, only the short-term samples (5–6 s) from both experiments were rated separately, so this part was identical for the 1 min and 2 min test.

The chosen method does not assess full conversational quality, but listening-only quality which reflects the effects of transmission-related artifacts on the quality perceived in a conversational situation. The advantages of the listening-only situation are that

- the conversation structure can be controlled, so that fixed degradation profiles can be used, which are identical for each test listener
- the attention of the listeners – although being directed to the contents of the sentences by means of the parallel task – is still sufficient for the quality-judgment task; in a real conversation, instead, the participant's attention is likely to be fully taken up by the conversation task, so that reliable judgments of quality are unlikely.

Of course, these advantages come with the inconvenience that talking-related degradations (like talker echo) and interaction-related degradations (like pure delay affecting the conversational structure) cannot be covered. The results of [22] show, that those factors influence perceived quality in such a complex manner, that they have to be investigated separately before including them in a controllable way in a study like the one presented here. Instead, the focus lies on the conversational structure and authentic (one way) transmission qualities of the samples.

Still, we think that the method is reasonable for assessing and modeling the impact of the temporal occurrence of degradations on a call-final judgment. This is why simulated conversations are preferred over real ones for this particular task.

2.1. Material

High quality recordings (48 kHz, 16 bit quantization) of four German speakers (two male, two female) were used in the experiments. Each simulated one dialog partner of a telephone conversation regarding a unique topic (zoo visit, car rent, kitchen purchase order, making an appointment with a dentist). Each recording consisted of 10 shorter samples of 5–6 s length each, representing one or two sentences. As regarding content, these 10 samples have a strict order. One example is:

> *'Hallo Marion, gestern war ich mit den Kindern meines Bruders im Zoo. Das war vielleicht lustig.'*
> (Hi Marion, yesterday I have been to the zoo with the children of my brother. This was really fun.)

The correspondent 12 s sample for Experiment two does also include the subsequent short sample:

> *'Die Schimpansen haben ihnen am besten gefallen. Die Kleinen haben sich total gefreut und waren überglücklich.'*
> (They liked the chimpanzees best. The little ones have been thrilled and were very happy.)

The recordings were down-sampled to 8 kHz, filtered by an IRS (Intermediate Reference System, see [24]) characteristics, and processed over real mobile channels. This processing was repeated 8 times in order to obtain 9 different versions of one sample with varying degrees of degradation.

For generating defined degradation profiles – i.e. a predefined sequence of specific levels of degradation, we used ITU-T Rec. P.862 as an indicator for the degree of degradation. Appropriately degraded versions of the original samples were selected to build up realizations of ten different degradation profiles, see Figure 1. They contained constant good, fair and poor qualities (profiles 1–3); continuously rising and falling qualities (4, 5); constantly good qualities with single degradations at the third or last sample position (9, 10); and burst degradations at the first, third or last position (6–8). Bursts are samples without constantly bad quality, but sudden interruptions

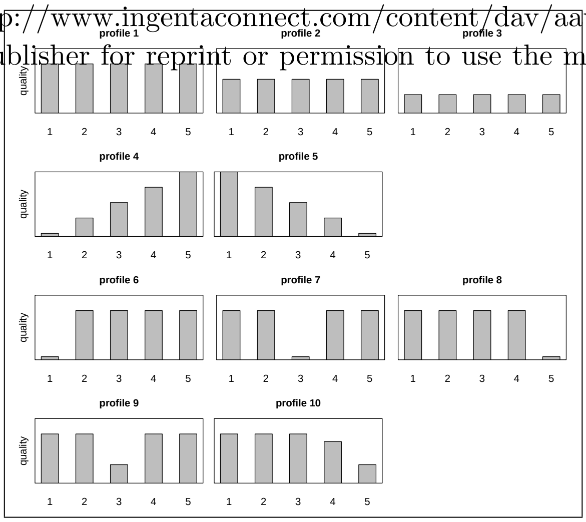

Figure 1. Schematical presentation of different profiles.

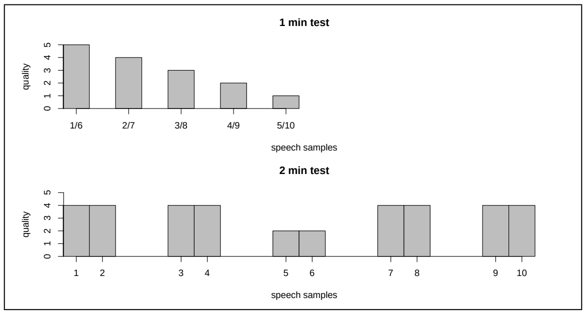

Figure 2. Structure of the samples in experiments 1 & 2. Pauses indicate subjects' interaction.

due to noise or silence. Not all profiles could be generated using real-life networks, as it was difficult to produce highly-impaired samples. When necessary, highly-impaired samples were produced with an AMR codec (Adaptive Multi-Rate) with frame losses, or with a GSM-HR codec (Global System for Mobile communication, Half Rate). Altogether, 16% of the used samples in the 1 min test and 19% in the 2 min test were produced in this way.

Each "dialog" consists of five samples and four breaks for subject interaction, see Figure 2.

The samples and pauses were equally long, with a fixed break duration. Thus, for the 2 min test, two succeeding short samples from the material were used as one sample (11–12 s), and the breaks were 12 s long, resulting in 105–108 s before rating. For the 1 min test, each sample formed a sample with 6.5 s long breaks, resulting in 53–56 s before rating. As the textual content of the material requires a strict order, there are two possible combinations of samples for the 1 min test (sample no. 1–5 and 6–10) for every of the four speakers with their 10 short samples, but only one possible combination for the 2 min test (sample no. 1+2, 3+4, 5+6, 7+8. 9+10, cf. Figure 2).

This resulted in 40 dialogs for the 1 min test (10 profiles · 4 speakers) with each profile realized once for each speaker. For the 2 min test each profile was realized once for each gender (5 for each speaker), resulting in 20 dialogs. In this way the repetitions of samples for the participants was limited to 5 times in both experiments.

2.2. Subjects

24 naïve subjects participated in every experiment, 9 females and 15 males, aged between 17 and 48. All of them were paid and recruited outside of the laboratory. All reported normal hearing, which was controlled with an audiometer (DIN EN 600645-2) prior to the experiments.

2.3. Procedure

Each experiment consisted of two parts. After the audiometer test, the subjects were individually seated in a silent room (cf. [2]) in front of a computer screen. The samples were presented via a standard handset (Post FeTAp 752). The subjects were told to concentrate on the content of the samples. After the presentation of each sample, a short question popped up on the screen with three possible answers (one correct) and the option "I could not understand / cannot remember":

Question: *Wann war die Frau im Zoo?* (When has the woman been visiting the zoo?)

- *heute?* (today?)
- *gestern?* (yesterday?)
- *noch nie?* (never?)
- *Ich habe nicht verstanden / kann mich nicht erinnern.* (I did not understand / cannot remember)

The subjects had to provide a verbal answer to this question within the pause interval until the next sample. After the last sample, no question appeared. Instead, the subjects had to rate the overall speech quality on a 5-point ACR scale displayed on the screen. This rating was done with a computer mouse to stop the simulated conversation explicitly with this change in modality. Due to the short time for the participants to answer the question in Experiment one, the questions were carefully worded and the provided answers were very short. Due to the training, no participant had problems answering them.

The experiment started with a training dialog where one additional profile was presented. This profile had a falling quality contour, so the subjects could get an impression of the entire quality range. After each dialog, the next one started immediately, but there was a pause of 5–10 min in the middle of the test session in order to avoid fatigue, dependent on the participant, not the experiment. The dialogs were presented in five different pseudo-randomized orders preventing directly succeeding samples of one speaker and series of profiles. Each dialog topic was presented 5 times with different profiles to one subject. For each of theses topics, 2 or 3 different questions were formulated. To make every screen picture unique, the incorrect answers were always different,

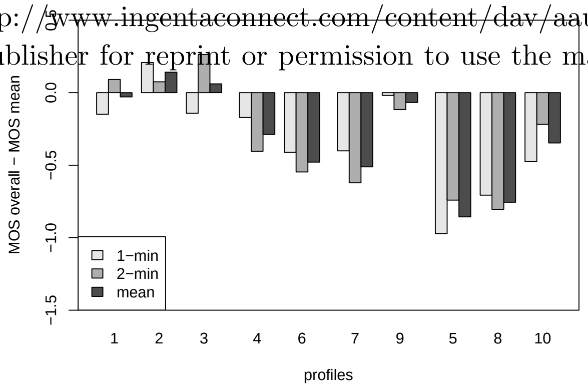

Figure 3. Differences between dialog ratings and averaged sample ratings for each degradation profile.

as well as the order of the three answers. As the 2 min test contains samples which are doubled in duration compared to the 1 min test, but also half of the number of dialogs, one performance lasted nearly equally long in both experiments (about 55–65 min, including audiometer test and briefing).

In the second part of the experiments, each sample used in both experiments has been presented in random order and rated on the same 5-point ACR overall quality scale [2]. This part also included one break of 5–10 min (35–45 min altogether). We obtained 45–48 ratings for each sample, as the first three subjects only rated the samples used in the 1 min test, and not the ones appearing solely in the 2 min test.

2.4. Results

Considering the MOS values of the first part (dialog-final ratings), results from both experiments (1 min and 2 min tests) are comparable regarding the profiles, even though both experiments differed in their dialog duration by a factor of two. They will be presented together as one set of data in the following section. Figure 3 shows the difference between dialog-final ratings and mean judgments of individual samples, averaged over each degradation profile.[1].

T-tests were used to compare the mean MOS values of the individual samples – as obtained by averaging all ratings from the second part of both experiments for each sample – to the final ratings for each dialog ($\alpha = 0.01$). Because of the real-life sample processing, not all of the dialogs turned out as good realizations of the profiles, as can be seen from the individual-sample MOS ratings. In particular, the realizations of profile 3 with constantly poor quality were not as constant as intended. For this profile, the use of even barely fitting processed samples was preferred over simulated ones to focus on real-life network degradations.

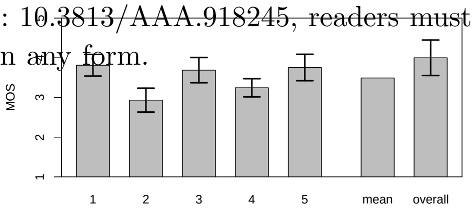

Figure 4. Individual sample ratings (1–5), mean, and overall dialog ratings for speaker W2 with conf. int. (profile 2).

2.4.1. Profiles with constant quality

11 out of 18 realizations of profiles 1–3[2] had approximately constant quality (by visual inspection of the individual-sample MOS values). The mean ratings of the individual samples do not differ from dialog ratings for these 11 dialogs. From the seven remaining realizations, only one was rated significantly better (two-sample t-tests for all single samples' MOS values (approx. 240) and the final ratings for each dialog (24): $t(28) = -3.03, p < 0.01$) than the mean (cf. Figure 4). The other six dialogs of profiles 1–3 were rated worse than the mean (one-sample t-tests regarding the mean of the individual samples' MOS and the final ratings for each dialog (24), $\alpha = 0.01$), probably due to the non-optimum realizations of the profiles.

2.4.2. Profiles with continuous quality changes

Dialogs with increasing quality (profile 4) were rated equally to the mean (one-sample t-tests regarding the mean of the individual samples' MOS and the final ratings for each dialog (24), $\alpha = 0.01$), except for one dialog that was judged significantly worse. In contrast to this, dialogs with decreasing quality (profile 5) were mostly rated significantly worse than the mean (four out of six).

2.4.3. Profiles with single degradations

Profile 9 (with one medial strong degradation) was rated equally to the mean of the short samples in all cases. However, the other four profiles that comprise one strong degradation (poor quality or burst) are rated differently. To get an impression of systematical changes with the position and the degree (burst or poor quality) of the degradation, the number of significantly lower dialog ratings compared to the mean of individual-sample ratings is listed below:

- profile 6: 1 of 6
- profile 7: 3 of 6
- profile 8: 5 of 6
- profile 9: 0 of 6
- profile 10: 3 of 6

All other realizations show no significant difference in the dialog rating.

[1] Bad realizations of one of the profiles with constant quality (profiles 1–3) are not included into Figure 3 (see the following section for details on the 7 excluded values)

[2] Every degradation profile was realized 4 times for the 1 min test and 2 times for the 2 min test.

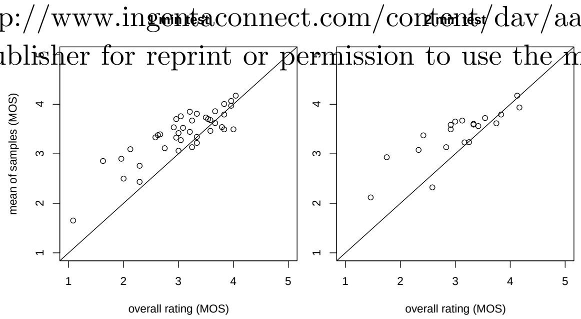

Figure 5. Comparison of dialog-final MOS values ("overall": mean over listener ratings for each dialog) and the averaged samples' MOS values ("mean": mean over the five MOS values obtained for each sample).

Those profiles with one strong degradation show more often significant results and lower ratings (ANOVA including the corresponding profiles, dialog nested in profile), if its position is closer to the time of judgment (profile 8 to 7 to 6, $F(2,414) = 9.7, p < 0.0001$; 10 to 9, $F(1,276) = 12.6, p < 0.001$) and if its quality is worse due to the burst characteristics (profile 7 to 9, $F(1,276) = 29.7, p < 0.0001$; 8 to 10, $F(1,276) = 29.2, p < 0.0001$). Additionally, visual examination suggests that the amount of degradation introduced by a burst or a sample of constantly poor quality has an impact. Because of the variation in the realizations of each profile, no statistical tests regarding differences between both experiments could be carried out to support the latter assumption.

Considering the number of significant results for each profile, continuous quality changes do not differ much from non-continuous ones with strong degradations at the same position (profiles 4 compared to 6, and profile 5 compared to 8 and 10), as the low number of significant differences shows:

- profile 4: 1 of 6
- profile 5: 4 of 6

3. Modeling dialog-final judgments with ratings of short samples

Using averaged MOS values of short samples to estimate dialog-final judgments is not satisfying for profiles with varying quality. Still, analyses show that the mean ratings are highly correlated ($r = 0.85$, $E_p = 0.26$; $r = 0.83$, $E_p = 0.28$, for the 1 min and 2 min test, respectively).[3] Figure 5 shows this comparison.

We will now develop two models which are able to cover the recency effect, and the effect of bad momentary speech quality. We then compare the prediction accuracy of these models with others known from literature. Finally, we try to base the predictions on estimations of instrumental models like the one defined in ITU-T Rec. P.862.

3.1. Data modeling

The observed effects on the dialog-final ratings are modeled in two separate steps:

3.1.1. Modeling the recency effect

MOS values are weighted differently according to their position in the dialog. The increasing weighting factor is not implemented relative to the entire dialog duration, but absolutely with regard to the individual samples' temporal distance to the end of the dialog. This is done because the model shall only describe the results of the analysis of both data sets combined:

We used a linear model for this purpose and determined the weighting coefficients for each sample of the simulated conversational structure. The results show that for the 1-min. test, only the last two samples are affected by the recency effect, and for the 2-min. test only the last sample. This can efficiently be modeled by correspondingly weighting a fixed-length window of 20 s from the end of the simulated conversation. As a side effect, this limits the recency effect found in the data to the time span we observed to show recency (up to about 20 s for one and two minute dialogs), which also corresponds to results from cognitive research [9]. If further experiments reveal other positional effects for longer sample durations, this model needs to be extended accordingly.

$$\overline{MOS}_{\text{A\_mod1}} = \sum_{n=1}^{N} (a_n \cdot MOS_n) / \sum_{n=1}^{N} a_n \qquad (1)$$

n : sample index
a_n : weighting factor
N : number of samples

The weighting factor a_n depends on the temporal distance from the end of the call towards the center of a sample. It is assumed to be constant (0.5) for a distance greater than 19 s. From that point onwards, a_n increases linearly towards the end of a call. This implementation of the recency effect comes close to an "end-effect":

$$a_n = \begin{cases} \frac{1}{2} \cdot \frac{(19-t_n)}{19} + \frac{1}{2}; & t_n \leq 19 \\ \frac{1}{2} & \text{otherwise} \end{cases} \qquad (2)$$

t_n : temporal distance of the sample n from the end of the call, in s
a_n : weighting factor
n : sample index

3.1.2. Modeling single strong degradations

The difference between the mean of individual ratings and the minimum quality is subtracted from the result of the first modeling step:

$$\overline{MOS}_{\text{A\_mod2}} = \overline{MOS}_{\text{A\_mod1}} - 0.3(\overline{MOS} - min(MOS_n)) \qquad (3)$$

n : sample index

[3] Here, Pearson's r is used. The prediction error E_p is calculated as the Root Mean Squared Error of the regression: $E_p = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$.

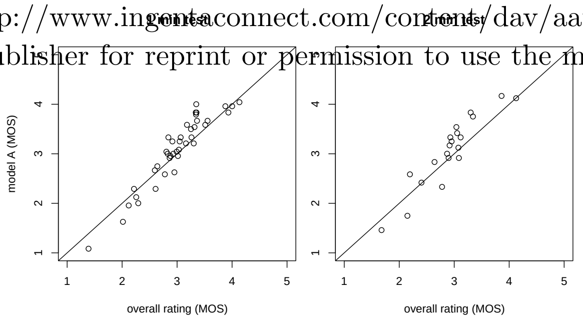

Figure 6. Predicted ("model A") and obtained dialog-final MOS values ("overall").

With this model, the linear correlation increases to $r = 0.94$ ($E_p = 0.19$, 1 min test) and $r = 0.93$ ($E_p = 0.21$, 2 min test). Modeling results compared to the actual dialog-final ratings are shown in Figure 6.

In an alternative version of this model, all deviation from the mean is taken into account. Before weighting short samples in the first step, the difference between an individual MOS value and the mean is subtracted from its individual value:

$$\overline{MOS}_{\text{B\_mod1}} = 2 \cdot \sum_{n=1}^{N} (a_n(MOS_n - .5\overline{MOS})) / \sum_{n=1}^{N} a_n \qquad (4)$$

n : sample index
a_n : weighting factor
N : number of samples

For this model (B), a_n stays constant (0.7) for samples with temporal distance towards the end of the call of 24 s and greater. Within this distance, a_n increases towards the end of a call:

$$a_n = \begin{cases} 0.3 \cos \frac{\pi t_n}{48} + 0.7; & t_n < 24 \\ 0.7 & \text{otherwise} \end{cases} \qquad (5)$$

t_n : temporal distance of the sample n from the end of the call, in s
a_n : weighting factor
n : sample index

Compared to model A, the recency effect is not as strong due to this modification. The second step is identical to that of model A:

$$\overline{MOS}_{\text{B\_mod2}} = \overline{MOS}_{\text{B\_mod1}} - 0.3(\overline{MOS} - min(MOS_n)) \qquad (6)$$

n : sample

With this alternative model (model B), correlations similar to model A are obtained ($r = 0.95$, $E_p = 0.19$, 1 min test and $r = 0.93$, $E_p = 0.21$, 2 min test). Both models are not very different from each other, and without any additional data, especially from additional degradation profiles, differences between them cannot be evaluated. A comparison of different model predictions and subjective ratings is shown in Figure 7, and the results of the correlation analyses are summarized in Table I.[4] We propose model B as an alternative to model A because it assigns a lower weighting to the recency effect. Still, the models perform equally well because of the high number of profiles with low quality in the final position (which is rated similarly by the two models).

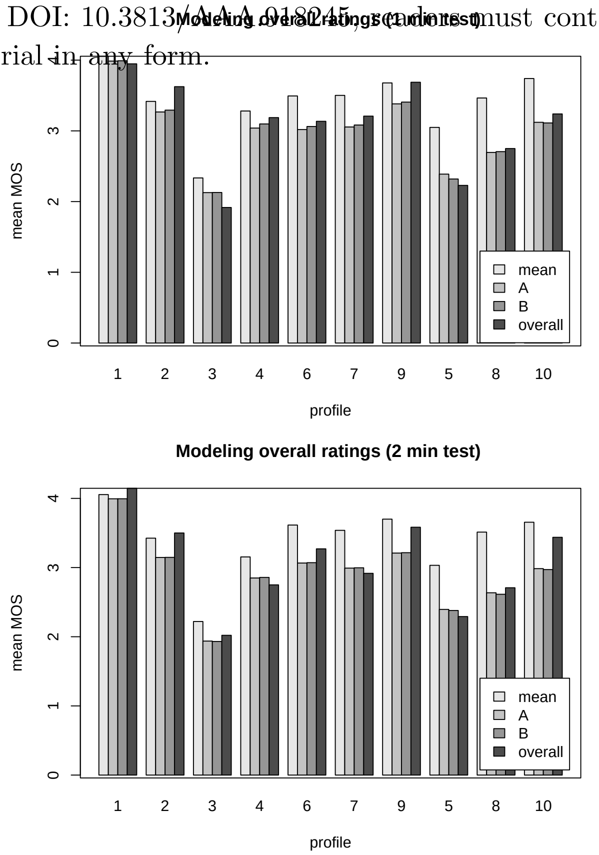

Figure 7. Comparison of MOS values: mean of individual-sample ratings ("mean"), model A, model B, actual dialog-final ratings ("overall").

Table I. Pearson's r and standard error comparing dialog-final ratings to estimated ones (plain mean, model A, model B); experiment 1 & 2.

experiment	means		model A		model B	
	r	E_p	r	E_p	r	E_p
1 min	0.85	0.26	0.94	0.19	0.95	0.19
2 min	0.83	0.28	0.93	0.21	0.93	0.21
both	0.84	0.26	0.94	0.19	0.94	0.20

3.2. Comparison to existing models

In this section, two existing models are used with the dialog ratings obtained in experiments 1 and 2 to compare

[4] Recall that the 2 min test has half of the data compared to the 1 min test.

their performance to the models proposed in the last section. The algorithms are described in detail, as some adjustments were necessary to account for the experimental structure.

The factors incorporated into the models introduced in the last section resemble those of the Rosenbluth's model [20], that has already been mentioned in the introduction. There, ratings of individual short samples are weighted according to their absolute rating and relative to their position within a longer speech sample; see Equation 7 for the calculation of the individual weighting factors and Equation 8 for the averaging part. The relative positions L_n of the mid-point of such short samples range from 0 to 1:

$$W_n = \max[1, 1 + (0.038 + 1.3 \cdot L_n^{0.68}) \cdot (4.3 - MOS_n)^{(0.96+0.61*L_n^{1.2})}] \quad (7)$$

n : sample index
L_n : relative position of sample n

$$\overline{MOS}_{\text{Rsbl}} = \sum_{n=1}^{N} W_n MOS_n / \sum_{n=1}^{N} W_n \quad (8)$$

n : sample index
N : number of samples
W_n : weighting factor

We applied this model to the simulated dialogs of experiments 1 and 2, taking the starting point of a dialog as $L = 0$, and the end of the last short sample as $L = 1$. Table II shows the correlations between the estimates of this model and the dialog-final ratings from experiments 1 and 2. It is not surprising that this model performs only slightly worse compared to the models developed on the basis of the experimental results (Section 3.1), as all take into account the position of the individual samples and weight the samples according to the strength of their impairment. Although the recency effect is included in our models in absolute terms, while [20] uses a relative description, both implementations produce reasonable results.

The findings in [16] reveal that instantaneous ratings slowly adapt to changes in quality, with different time constants for large improvements and large degradations. Instead of using a weighted average of short-term ratings to estimate overall quality as in the previous models, the model in [19] averages estimated instantaneous quality ratings. To compare our models to this approach, instantaneous quality has to be estimated from the individual sample ratings. This estimation has to cover the pauses of our simulated dialogs as well; for these pauses, we assume that the quality is equal to the one of the preceding speech sample. The algorithm used here is taken from [19]. Two different time constants $\tau_1 = 9$ s and $\tau_2 = 14.3$ s (taken from [25], because this leads to a better performance) reflect (τ_j), which differs for sudden degradations (τ_1) or improvements (τ_2) in quality (Equation 9):

$$MOS_{t_i} = MOS_{t_{(k+1)}} - (MOS_{t_k} - MOS_{t_{(k+1)}})e^{-\frac{t_i - t_k}{\tau_j}} \quad (9)$$

MOS_{t_i} : estimated instantaneous rating at instance t_i
$t_k; t_{k+1}$: temporal borders of adjacent segments with assumed constant quality
MOS_{t_k} : estimated instantaneous rating at instance t_k
$MOS_{t_{k+1}}$: actual MOS value of the sample at t_{k+1}
τ_j : time constant for exponential decay

Table II. Pearson's r and standard error comparing dialog-final ratings to estimated ones (model Rsbl, model time-avrg; experiment 1 & 2.

experiment	model Rsbl		model time-avrg	
	r	E_p	r	E_p
1 min	0.92	0.24	0.88	0.24
2 min	0.92	0.25	0.89	0.27
both	0.92	0.24	0.88	0.26

From these estimates of instantaneous ratings for the short samples, the integral quality is calculated by averaging:

$$\overline{MOS} = \frac{\sum_{i=1}^{I} MOS_{t_i}}{I} \quad (10)$$

i: instantaneous ratings' index
MOS_{t_i} : estimated instantaneous rating
I : number of instantaneous ratings

In a last step, the recency effect is included considering the amount and position of the last significant degradation. Due to the profiles used in experiments 1 and 2, we have represented this degradation by the minimum of the short samples, if it deviates more than 0.5 (MOS) from the time average:

$$\overline{MOS}_{\text{time-avrg}} = \overline{MOS} + (k \cdot (MOS_{t_m} - \overline{MOS}))e^{-\frac{y}{\tau_3}} \quad (11)$$

$\overline{MOS}_{\text{time-avrg}}$: estimated dialog-final rating
MOS_{t_m} : quality at time t_m of the last degradation in a simulated dialog
$k = 0.7$
$\tau_3 = 30$
y : delay between end of MOS_{t_m} and the end of a call

This model performs only a little better than plain averaging ratings of individual samples, see Table II. However, this does not necessarily imply that the model is inappropriate, as the free parameters of the model have not been adjusted to the experimental data.

3.3. Data modeling with instrumental quality estimates

Both models we have presented in Section 3.1 describe dialog-final quality ratings in a satisfying way. But in order to use these models for prediction, it is important to rely on instrumental methods rather than on perception experiments. As a consequence, we applied both models to quality estimations for individual samples obtained by ITU-T Rec. P.862 [1].[5] Figure 8 shows the estimates in comparison to the auditory ratings, separated for each speaker.[6]

[5] Throughout this paper, we also used the mapping function ITU-T Rec. P.862.1 [26] on the estimates obtained by ITU-T Rec. P.862.

[6] The noticeable points with P.862 scores about 4.5 are those 16 basic samples that have not been processed.

Table III. Pearson's r and standard error comparing dialog-final ratings with model predictions on the basis of P.862 estimates for individual samples (plain means, model estimates); experiment 1 & 2.

exp.	means		model A		model B		model Rsbl	
	r	E_p	r	E_p	r	E_p	r	E_p
1 min	0.75	0.31	0.89	0.24	0.89	0.25	0.87	0.29
2 min	0.70	0.34	0.84	0.29	0.84	0.29	0.84	0.33
both	0.73	0.32	0.87	0.26	0.87	0.26	0.86	0.30

The MOS values of the single samples show a strong linear correlation ($r = 0.93$) with the instrumental quality estimation (P.862), which is in line with other studies (cf. [5]). When using P.862 scores instead of MOS values with the two models presented in (1)–(6), dialog-final judgments are considerably better estimated than by plain averaging of the P.862 scores. Table III shows the results of the correlation analysis, including Rosenbluth's model. A comparison to Table I and Table II shows that the correlation with the dialog-final judgments is lower when relying on estimated than when relying on auditory MOS. Still, the increase in prediction accuracy due to the models compared to the plain mean is comparable for auditory and estimated MOS.

3.4. Discussion

Results of both experiments show a systematic relationship between dialog-final judgments within simulated conversational structures and individual ratings of the short samples the conversations are composed of: In only one case dialog-final judgments are significantly higher than the mean of single ratings for short samples. Dialog-final ratings depend on the position of the individual samples (recency effect), and they are affected strongly by the rating of the sample with minimum quality. Because of these two influencing factors, the "peak-end" rule [10] seems to be more appropriate to describe subjects' behavior than the recency effect.

The dialog-final quality was successfully estimated by two new and one existing [20] models which use quality measures for short samples (5–6 s) as an input. The strong difference between the modeling with MOS values and P.862 scores is surprising. It seems that the error of the instrumental estimates for the short samples adds up with the difference in the MOS values between dialog-final and mean ratings. This results in rather low correlations for model predictions based on P.862 scores. Figure 9 shows that P.862 frequently under-estimates the quality of the simulated (AMR and GSM-HR) samples, and more frequently over-estimates the quality of the processed (real-life) samples. Because the simulated samples are mainly used for low-quality profile positions, the estimation error may accumulate to the modeling error.

With the help of both models, the correlation between the dialog-final and the estimated judgments could be increased considerably compared to the mean, both for subjective MOS values (10% absolute improvement) as well as for instrumental P.862 scores (13% absolute improvement). The models are assumed to be largely independent of the conversational length between 1 and 2 min. However, these models are based on data from profiles with not more than one strong degradation, even though some of the intended constant profiles (1–3) had some unique variation.

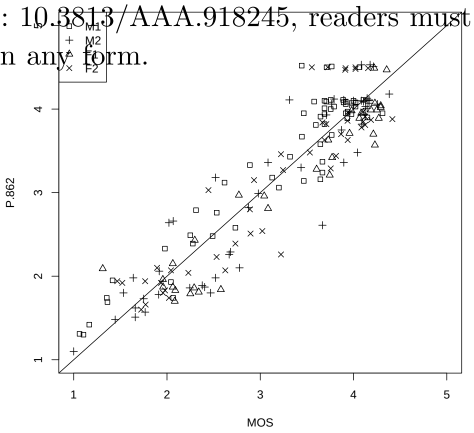

Figure 8. MOS and P.862 values of single samples, separated for each speaker.

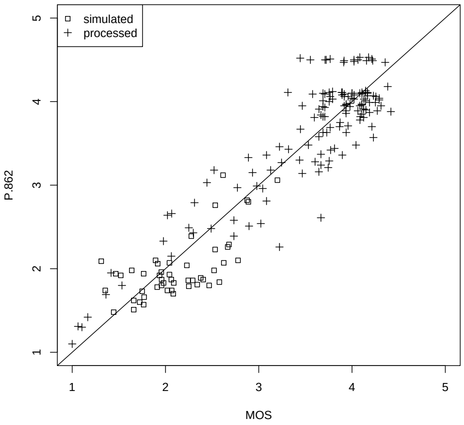

Figure 9. MOS and P.862 values of single samples, separated for degradation type.

4. Independent verification of the models

Using the test procedure proposed in Section 2.3, two additional experiments were carried out at a different laboratory with newly recorded material to evaluate the proposed models. They are briefly described here. With the results of these experiments (cf. [7]), it should be tested if the models are independent of the particular conversational structure – especially the implementation of the recency effect. Furthermore, it was important to get additional results concerning the modeling success based on subjective compared to instrumental measures. As the Rosenbluth model performs similar to our models for the data of experiments 1 and 2, it is included in the comparison.

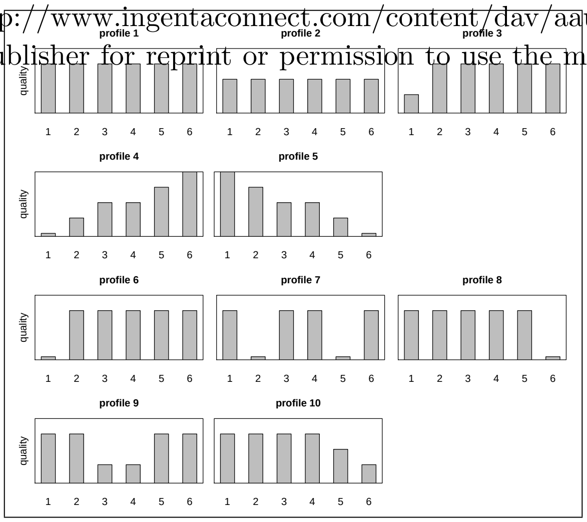

Figure 10. Schematical presentation of different profiles for experiment 3. Pauses indicate subjects' interaction.

4.1. Material and procedure

Experiments 3 and 4 are quite similar to the first and second ones. As before, two sets of profiles were prepared, of one and two minutes duration, respectively. Similar scenarios (dentist, car rental, zoo visit, kitchen purchase) were used as basis for the new material. However, instead of German, the new material was recorded from four British English speakers (2 females, 2 males). All speech samples were degraded only with AMR simulation, including different degrees of frame loss. The samples had a duration of 5 s.

In contrast to Experiments 1 & 2, all samples and interaction breaks were 5 s long. The new 2 min test is in fact the repeated structure of the new 1 min test. Due to the shorter samples, there were six (1 min test) and 12 (2 min test) instead of five samples used.

As in the first experiments, 10 degradation profiles were used. Two of them are new: Profile 3 exhibits a low quality only at the beginning instead of a constant poor quality. Profile 7 was changed to have two bursts instead of one, placed at sample number 2 & 5 (1 min) and 4 & 9 (2 min). In Figure 10 and Figure 11, you can see that strong degradations at the beginning and at the end of a dialog lasted one sample for both experiments, while those in the middle lasted longer (2 or 4 samples for the 1 and 2 min test, respectively).

In the 1 min test, each profile was realized twice. Material was taken from one male and one female speaker. In the 2 min test, every profile was realized once, with half of the dialogs with samples from the other male and other female, respectively. Subjects had to verbally answer the questions regarding the sample just heard and had to mark the answer on a sheet of paper. The final rating of the perceived quality of a simulated dialog was done by pressing one of the buttons on a special terminal. 26 native English-speaking subjects participated in this study. As in the first two experiments, the subjects also rated the short samples which had been presented to them in the profiles individually, resulting in 26 votes per sample.

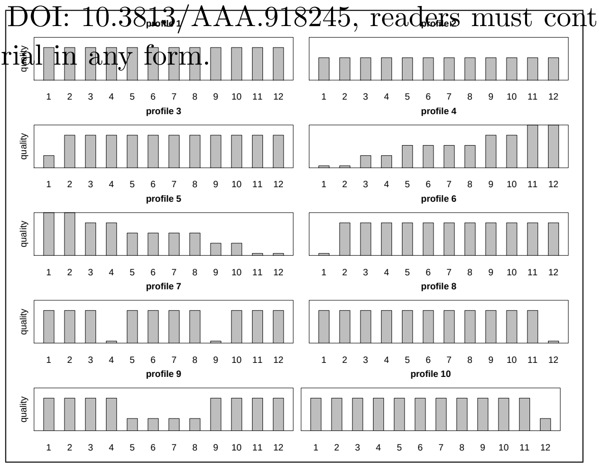

Figure 11. Schematical presentation of different profiles for experiment 4. Pauses indicate subjects' interaction.

Table IV. Pearson's r and standard error comparing dialog-final ratings to estimated ones (plain means, model estimates); experiment 3 & 4.

exp.	means		model A		model B		model Rsbl	
	r	E_p	r	E_p	r	E_p	r	E_p
1 min	0.89	0.32	0.98	0.17	0.97	0.17	0.96	0.24
2 min	0.92	0.26	0.99	0.12	0.99	0.12	0.87	0.47
both	0.90	0.30	0.98	0.15	0.98	0.16	0.93	0.32

Table V. Pearson's r and standard error comparing dialog-final ratings with model predictions on the basis of P.862 estimates for individual samples (plain means, model estimates); experiment 3 & 4.

exp.	means		model A		model B		model Rsbl	
	r	E_p	r	E_p	r	E_p	r	E_p
1 min	0.89	0.33	0.97	0.19	0.97	0.19	0.95	0.26
2 min	0.88	0.33	0.97	0.18	0.97	0.17	0.87	0.43
both	0.88	0.32	0.97	0.18	0.97	0.18	0.92	0.31

4.2. Results

MOS values of short samples are estimated quite well by P.862 ($r = 0.96$). Subjects' ratings have not been analyzed in detail, as only the performance of the models is of interest here. Dialog-final ratings and the plain mean of single samples (in MOS) correlate with $r = 0.89$ ($E_p = 0.32$, 1 min test) and $r = 0.92$ ($E_p = 0.26$, 2 min test). By applying the models, these correlations increase significantly as expected. Table IV provides the exact results. The models seem to work sufficiently well also for samples with two minima, as estimated quality of the three realizations of profile 7 (with two minima) are within one standard deviation of the modeling results with data from the other profiles. Using P.862 for estimating individual samples' MOS values, the dialog-final judgments are predicted very well by models A and B, cf. Table V.

The relatively low correlation between Rosenbluth model's estimates and dialog-final ratings of experiment 4 are caused by one of the 10 dialogs, that is characterized by one strong degradation at the final position. Its dialog-final rating is $MOS = 3.23$. This dialog is estimated too pessimistically (2.26 with auditory MOS, 2.17 with P.862). The effect does not occur in experiment 3, so this model seems to overestimate the impact of the short dialog-final degradation.

4.3. Discussion

With data from these two independent experiments (3 & 4), both models reliably estimate dialog-final judgments based on quality measures of short samples. This is also true for Rosenbluth's model, with the exception of profile 8 in experiment 4. The confirmation for another group of subjects and the changes in number of turns are a first evidence for the validity of the models within the timespan of dialog durations of 1 min to 2 min. The scalable weighting factors of the first step of the models were used successfully. However, results from one profile with two minima instead of one (profile 7) are not sufficient to generalize the results regarding time-varying quality to further types of degradation profiles. Simulated channel degradations (experiments 3 & 4) limit the error of the modeling results with the P.862 method drastically. However, the simulation is not as realistic as considering real-life mobile or Voice over IP scenarios, as for the majority of conditions in experiments 1 & 2.

5. Conclusion

A method of simulating conversational structures has been presented to assess dialog-final judgments for telephone transmission quality. With this method, data was collected for speech samples mostly transmitted over real mobile networks and analyzed regarding the relationship between dialog-final quality and varying quality of short samples. Two effects were found, a recency effect and the impact of the strongest degradation, which even account for call quality of profiles with continuous quality changes. Two models incorporating these findings describe and predict dialog-final ratings on the basis of assessments of short samples, as well as on the basis of estimates obtained by the instrumental method of ITU-T Rec. P.862. The models are general, so other instrumental methods might also be used in conjunction with them. In contrast to other studies, final ratings are assessed from a simulated dialog, that reduced subjects' concentration on speech quality rating and exhibited the temporal structure of a conversation. Quality of short samples was assessed separately. As both models perform equally well, the simpler model (A) is favored here. The model is independent of the number and duration of samples and interaction breaks, which was confirmed by two additional sets of data. However, the validation is restricted to the profiles used, which are characterized by equal durations for every sample and break. The model itself reflects to some degree the degradation profiles, especially concerning the impact of the minimum, which is modeled in a second modeling step. Considering the fact that it successfully covers ratings with constant quality, constantly increasing quality, and with two minima, we assume that it is robust enough to be generalized. Still, further tests are desirable, especially verifying the model with respect to real conversation with its unique types of degradation.

We compared our predictions to two other models known from the literature. Rosenbluth's model is related to our models and shows only a minimal but systematically lower performance. There is only one profile that is estimated with considerably worse efficiency. It is encouraging though, that this model proves to be reliable for our data, as it was only developed for 2 kinds of degradations (bursts and clipping) and for 1 minute samples. It might be interesting to further compare this one to model A for more complex profiles. In our study, there is only a slight difference between modeling the recency effect with a relative (as in the Rosenbluth model) or with an absolute integration of short samples' temporal distance to the end of a call.

The impact of semantics or phonetics was not considered here. In real conversations, it can be assumed, that more informative words suffer much more from quality degradations than non-informative ones (cf. [12]). Further linguistic aspects of speech have shown to carry an influence on quality, cf. [17]. Because of this, it might be desirable to include these factors in the modeling process as well. Further experimental studies are necessary to justify such steps.

Acknowledgment

The authors would like to thank T-Mobile Deutschland GmbH and E-Plus, Germany, for their generous contribution to the German study. Furthermore, we gratefully acknowledge Ericsson for permitting the access to their investigation in English for this publication. We like to acknowledge the support of the entire ETSI STQ mobile group for carrying out this investigation. Finally, we like to thank two unnamed reviewers for their helpful comments.

References

[1] ITU-T Rec. P.862: Methods for objective and subjective assessment of quality. 2001.

[2] ITU-T Rec. P.800: Methods for subjective determination of transmission quality. 1996.

[3] ITU-T Rec. P.800.1: Mean Opinion Score (MOS) terminology. 2006.

[4] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier: PESQ, the new ITU standard for objective measurement of perceived speech quality, part II — psychoacoustic model. Journal of the Audio Engineering Society **50** (2002) 765–778.

[5] ITU-T Rec. P.862.3: Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2. 2005.

[6] P. Gray, R. Massara, M. Hollier: An experimental investigation of the accumulation of perceived error in time-varying speech distortions. Audio Engineering Society, 103rd Convention, New York, 1997.

[7] ETSI TR 102 506: Speech Processing, Transmission and Quality Aspects (STQ); Estimating Speech Quality per Call. 2007.

[8] B. B. Murdock: The serial position effect of free recall. Journal of Verbal Learning and Verbal Behaviour **64** (1962) 482–488.

[9] N. Cowan: On short and long auditory stores. Psychological Bulletin **96** (1984) 341–370.

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

[11] S. Kuwano, S. Namba: Continuous judgement of level-fluctuation sounds and the relationship between overall loudness and instantaneous loudness. Psychological Research **47** (1985) 27–73.

[12] U. Jekosch: Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung. Diploma Thesis. Universität Essen, 2000.

[13] M. Hansen, B. Kollmeier: Continuous assessment of time-varying speech quality. JASA **106** (1999) 2888–2899.

[14] ITU-T Rec. P.810: Modulated Noise Reference Unit (MNRU). 1996.

[15] ITU-T Rec. P.880: Continuous evaluation of time varying speech quality. 2004.

[16] L. Gros, N. Chateau: Instantaneous and overall judgements for time-varying speech quality: Assessments and relationships. Acta Acustica united with Acustica **87** (2001) 367–377.

[17] A. Raake: Speech Quality of VOIP: Assessment and Prediction. Wiley, Chichester, 2006.

[18] ITU-T Rec. P.107: The E-Model, a Computational Model for Use in Transmission Planning. 2005.

[19] A. Clark: Modeling the effect of burst packet loss and recency on subjective voice quality. Internet Telephony Workshop (IPtel 2001), New York, 2001.

[20] ITU-T Delayed Contribution D.064: Testing the quality of connections having time varying impairments. Source: AT&T, USA (J. H. Rosenbluth), 1998.

[21] L. Gros, N. Chateau, S. Busson: Effect of context on the subjective assessment of time-varying speech quality: Listening / conversation, laboratory / real environment. Acta Acustica united with Acustica **90** (2004) 1037–1051.

[22] M. Guéguin, V. Gautier-Turbin, L. Gros, V. Barriac, R. Le Bouquin-Jeannès, G. Faucon: Study of the relationship between subjective conversational quality, and talking, listening and interaction qualities: towards an objective model of the conversational quality. Measurement of Speech and Audio Quality in Networks, Prague, 2005.

[23] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, V. Barriac: On the evaluation of the conversational speech quality in telecommunications. EURASIP Journal on Advances in Signal Processing **8** (2008).

[24] ITU-T Rec. P.830: Subjective performance assessment of telephone-band and wideband digital codecs. 1996.

[25] L. Gros: Evaluation Subjective de la Qualité Vocale Fluctuante. Dissertation. Université de la Méditerran´e Aix-Marseille II, Equipe d'acceuil, France Telecom R&D, F-Lannion, 2001.

[26] ITU-T Rec. P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO. 2003.