 DOI: 10.1109/QoMEX.2014.6982329

ANALYSIS OF CALL-QUALITY PREDICTION PERFORMANCE FOR SPEECH-ONLY AND AUDIO-VISUAL TELEPHONY

Benjamin Weiss, Sebastian Möller, Benjamin Belmudez, Blazej Lewcio

Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany

ABSTRACT

In this paper, we will compare several approaches for predicting the quality of entire speech-only or audio-visual telephone calls from subjective judgments or predictions of the quality of individual segments of these calls. The comparison will be done using several databases which have been collected in a test paradigm to simulate conversational structures. The subjective judgments for individual conversation segments obtained in these tests, as well as instrumental quality estimations for these segments, are used as the basis for predicting episode-final quality ratings. Although different modeling approaches reach a similar performance, an optimum model is proposed which leads to comparably high prediction accuracy for speech-only and audio-visual telephony. Both the test paradigms as well as the call-quality prediction models are subject to standardization activities of ITU-T Study Group 12, for which this analysis is considered an input.

Index Terms **—** call quality, episodic quality, time-varying quality, temporal integration, recency effect

1. MOTIVATION AND INTRODUCTION

Whereas media are mostly used in *episodes* such as a movie, a TV show, a telephone call, a web browsing session, or alike, multimedia quality is frequently assessed in time frames which are shorter than such episodes. For example, transmitted speech quality is commonly assessed from short samples of approx. 8 s [13], visual or audiovisual quality from stimuli of approx. 10 s [17][18], or web browsing from short-term tasks which can be carried out within less than a minute [20]. Such short stimuli enable efficient testing, they do however not necessarily reflect the full experience which users might get during a typical usage episode. This discrepancy may turn into a problem in case that the media or interaction quality behaves variably over the time of a usage episode, such as during mobile telephony or mobile video streaming over unreliable networks.

In order to address this issue, three approaches have been followed in the past. One possibility is to ask test participants continuously for an instantaneous rating, e.g. using a slider [16][10]. These methods are recommended for obtaining ratings of *instantaneous* quality, but they have also been criticized regarding the cognitive effort which has to be put on the rating task, and which may distract from normal media usage, perhaps leading to ratings which are not ecologically valid. A method to circumvent this problem is to ask test participants to adjust media quality to a predefined level, as proposed e.g. by Borowiak et al. [1].

A second approach is to ask test participants for ratings after an actual usage period of typical length, as it is done e.g. in a standard conversation test [13][14] or in a videotelephony test [19]. Whereas these tests – in case that appropriate usage scenarios are selected – provide ecologically valid results with respect to the entire episode, they do not provide insight into how individual instances of the episode – with potentially time-varying quality – impact the episode-final rating. This is a limitation when service providers seek for weightings of instantaneous bad quality "events" for the episode-final experience of a user.

A third approach is to use simulations of usage instances during the assessment. For example, the structure of a phone conversation can be roughly simulated by playing pre-recorded segments of speech (about 6…10 s long) from a far-end interlocutor to a test participant, asking the participant to respond to questions during short pauses between the played segments, and asking for an episode-final rating after all segments. For each segment, a standard quality rating can be obtained in a standard listening-only test according to ITU-T Rec. P.800 [13], reflecting approximately the *instantaneous* quality of the simulated call. In case that the speech segments are semantically linked in terms of a short "story", the judgment of the test participant is expected to reflect approximately a natural conversation situation, although the test participant remains slightly more passive than s/he would be in a real conversation.

In the past, such "simulated conversational structures" have been used to assess the quality of time-varying speech quality, and a standardized description of the method is available in ETSI TR 101 506 [5]. The method has also been applied to videotelephony calls [1][21] by adjusting the scenarios in order to make meaningful use of both speech and video channels; however, there is no standardized description of that extended methodology yet.

 DOI: 10.1109/QoMEX.2014.6982329

On the basis of the ratings of the individual segments (obtained with standard subjective methods, such as described in [13] and [18]), researchers have also developed prediction models for the episode-final rating obtained in the simulated conversation test [5][2][9][4]. These models try to describe temporal integration effects beyond a simple average which have been observed in quality assessment of longer episodes, such as the recency effect (the effect that segments at the end of an episode contribute more to the episode-final ratings than segments occurring at the beginning of an episode), the peak-end-rule (the effect that the episode-final rating may be described by a combination of the last and the worst segments in a call), or the primacy effect (the effect that episode-initial segments may have a higher impact on the episode-final rating). A number of such models have been compared on two simulated conversation test databases in [23]. Similar models have also been developed for videotelephony [1] and for long-term video streaming [6].

In order to come up with a single, recommended approach, Study Group 12 of the International Telecommunication Union (ITU-T SG12) has recently defined two work items in its Questions Q.7/12 and Q.9/12. The first item (P.AVQ) is meant to lead to a Recommendation on a subjective methodology for simulated conversation tests which can be applied to speech-only and audio-visual telephony calls. Such a description has already been drafted [7], and the content will be briefly summarized in Section 2. The second item aims at defining a single call-quality model which is again applicable to speech-only and audio-visual calls. Such a model should work not only for subjective ratings obtained for individual call segments, but also for *estimations* of such ratings, using standard recommended prediction models such as the Perceptual Objective Listening Quality Assessment model POLQA [15] for speech, or the models according to ITU-T Rec. J.341 [11] for video.

In Section 3, we will review several modeling approaches which may be candidates for a future call-quality model to be standardized by ITU-T SG12. We will assess the performance of these candidates on a set of speech-only and audio-visual databases which are described in Section 4. Section 5 summarizes the results and discusses implications for the selection of a single model. Section 6 concludes the paper with a proposal for a unique model, to be considered in the standardization work of ITU-T SG12.

2. SUBJECTIVE TEST METHODOLOGY

A typical (speech-only or audiovisual) telephony situation is a dialogue between two persons who both have a certain amount of speaking and listening activity, and where both are interested in the topic of the conversation. Thus, the test situation is divided into parts with listening-only activity and parts with speaking activity, hereby neglecting a timely switch between the activities, and also neglecting double talk. This compromise has to be made in order to guarantee a fixed structure of the conversations which can be used for call-quality modeling later on. In order to ensure the ecological validity of the test, typical content of telephone conversations is chosen (e.g. request for a rental car, discussing the furniture in a room).

A simulated conversation consists of 5 segments of approximately 8…12 seconds, with 4 breaks in between. In each break, the test participant is asked a content-related question referring to the last segment, with 3…5 multiple choice answer options which are presented either on paper or on a computer screen. The position of the correct answer is randomly assigned for each question, and the participants are asked to quickly and orally answer the question before the next segment is automatically played back. After the fifth segment, the participant is asked for an assessment for the overall episode-final quality of the simulated call. Figure 1 shows the schematic structure of a simulated conversation. In a separate session, the individual segments of the simulated conversation are rated following the procedure

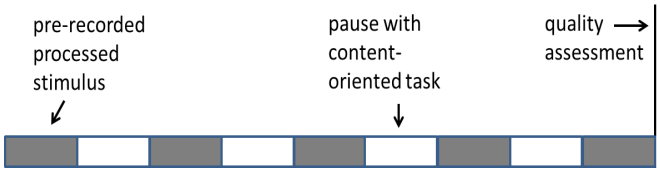

Figure 1: Schematic representation of the simulated conversation structure.

described in ITU-T Rec. P.800 [13].

Similar to the methodology for speech-only call quality, a simulated videotelephony conversation is constructed from a set of five consecutive segments of short audiovisual clips which are linked by the storyline. Participants are asked to watch each segment and then to verbally answer a question related to the audiovisual content. The question can be related to the auditory content, to the video content, or to the audiovisual one. The segments consist of different speakers (male and female ones) and comprise different scene backgrounds and conversation topics. The clips should involve the video channel by means of visual cues (e.g. showing objects to the camera, pointing dates on a wall calendar behind the person, body gestures), so that the test participants have to pay attention to the video channel as well. At the end of each simulated conversation, test participants rate the audiovisual quality of the entire dialog using a standard overall quality scale, as recommended in [18]. The individual segments are additionally rated in a separate test session according to the procedure of [18].

3. CALL-QUALITY MODELS

The models evaluated here have been selected from and are termed as in [1]. The ETSI [5] and the Weiss model [22]

 DOI: 10.1109/QoMEX.2014.6982329

have been developed specifically based on data collected using the test methodology described in Section 2. As the method of simulating conversational structures does not account for interactivity-degrading effects from echo and delay, models for long-term stimuli without interactive effects can also be applied on the available data. These are the Rosenbluth [9], Clark [4] for audio only, and the Hamberg model [6] for image quality. These three currently represent all related models known to us aiming at predicting end-of-call quality with time-varying transmission characteristics.

The ETSI model calculates a weighted average (Equation 1) of short segments' quality (subjective MOS or instrumental estimates of MOS) with weights *a* of index *i*, by incorporating a recency effect starting at 19 s from the end of the call (2), and a weight of 0.5 otherwise:

$$(1)\quad MOS_{sum} = \sum_{i=1}^{n} a_i MOS_i \Big/ \sum_{i=1}^{n} a_i$$

$$(2)\quad a_i = 0.5(19 - t_i)/19 + 0.5$$

The final call-MOS is the weighted sum minus the impact of the strongest degradation, regardless of its duration, see Eq. (3):

$$(3)\quad MOS_{call} = MOS_{sum} - 0.3\,(\overline{MOS} - \min(MOS_i))$$

The Weiss model is similar to the ETSI model, except for the weighted average, taking into account the difference from the overall mean, see Eq. (4). Accordingly, different weights are used starting from 24sec from the end of the call (Eq. 5) with 0.7 otherwise. The strongest degradation is subtracted just like with the ETSI model, see Eq. (3).

$$(4)\quad MOS_{sum} = 2\sum_{i=1}^{n} (a_i (MOS_i - 0.5\overline{MOS})) \Big/ \sum_{i=1}^{n} a_i$$

$$(5)\quad a_i = 0.3\cos\frac{\pi t_i}{48} + .7$$

The Rosenbluth model takes into account the magnitude and relative position of the short segments when calculating the average, see Eq. (6), with *L* being the position from 1 to 0 relative to the end of the call. The temporal center of each short segment is chosen for applying it on the data presented here:

$$(6)\quad a_i = \max\begin{bmatrix} 1{,}1 + (0.038 + 1.3L_i^{0.68}) \\ \cdot (4.3 - MOS_i)^{(0.96 + 0.61L_i^{1.2})} \end{bmatrix}$$

The Hamberg model incorporates a recency effect in the exponential weighting function (see Eq. (7)), with episode length *T*, and additionally normalized by dividing a_i by the sum of all weights. The final quality of the call is calculated as a weighted sum of the power of the differences between the quality of the short segments and an optimal power as represented in Eq. (8), with the power *p* of 3.05. Here, the magnitude of the impairment is taken into account. The model was developed on the basis of quality scores obtained on the R-scale. The quality scores obtained with this scale can be mapped to the MOS scale using the transformation defined in [6]. The Hamberg model is applied by constant calculating instantaneous ratings from short segments for each second and transforming these to the R-scale for calculation and then back to MOS values. $R_{i,ref}$ can be the local maximum at *i,* or the maximum in each of the data sets to fit, i.e. each profile, or the global maximum. The last possibility was chosen here to cope with different durations and different experienced quality ranges in each experiment, and results in a higher fit than with profile-related maxima.

$$(7)\quad a_i = e^{\frac{i-T}{25.9}}$$

$$(8)\quad (R_{ref} - R_{call})^p = 1.38^p \sum_{i=1}^{T} a_i (R_{i,ref} - R_i)^p$$

The Clark model uses a different attempt than the other approaches, as its main aim is to cover the effect of instantaneous user ratings responding exponentially to sudden changes in transmission quality within one episode, see Eq. (9), with *k* and *k*+1 representing the temporal borders of adjacent segments with assumed constant quality, and the constant τ_j being 9sec for degradations and 14.3 s for improvements:

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

The estimated call-MOS represents the time average of these modelled instantaneous ratings, also incorporating a recency effect (Eq. (10)), with the time of the last relevant degradation t_m, the time span *y* between MOS_{tm} and end of the call, and the constant τ_3=30 s.

$$(10)\quad MOS_{call} = \overline{MOS} + 0.7(MOS_{t_m} - \overline{MOS}) e^{-y/\tau_3}$$

As with the Hamberg model, instantaneous rating are calculated with 1Hz., Also, the interactive breaks are taken into account as time for the exponential reaction to quality changes, thus improving the results in Table 1 from r=.806 to r=.866.

4. DATABASES

Ten databases are used for validating, all applying the method of simulating conversational structures [5]. Eight are audio-only, and two use audio-visual stimuli. The first five audio-only databases contain data for narrowband-transmitted speech (300-3400 Hz transmission bandwidth):

1. **G60**, a German test of simulated conversations of 1 min length (10 profiles, participants *N*=24, 40 call-MOS, one for every speaker)
2. **G120**, a corresponding German 2-min test (10 profiles, *N*=24, 20 call-MOS, one for each speaker's gender)

Short segment ratings are collected for both data sets together [23]. Both sets have similar degradation profiles (10) and the same number of short segments (5 for each profile). To test more complex profiles including two strong

 DOI: 10.1109/QoMEX.2014.6982329

degradations especially for the smaller datasets of the 2-min tests, an additional unpublished test with 6 profiles was conducted based on the stimulus material from G60/G120:

3. **G120b**, a German 2nd 2-min test (6 profiles, *N*=18, 12 call-MOS, one for each gender.)

Ten degradation profiles quite similar to G60 and G120 have been used in two English test sets [23], where the 2-min test exhibits double the number (12) of short segments, as this time the duration was similar to the 1min test (6 segments):

4. **E60**, the English 1-min test (10 profiles, *N*=13, 20 call-MOS, one for every speaker)
5. **E120**, the English 2-min test (10 profiles, *N*=13, 10 call-MOS, profiles split in half for gender)

For all this data, there are PESQ values available as instrumental estimates of the short segments.

The other three audio-only databases exhibit also codec switching between narrowband and (super-) wideband, as well as handovers between wireless networks within one of the five short German segments (and POLQA estimates are available, except for S60b) [21]:

6. **S60a**, the 1st 1-min test with codec switching (10 profiles, *N*=13, 10 call-MOS, averaged over four speakers)
7. **S60b**, the 2nd 1-min test with codec switching (10 profiles, *N*=14, 10 call-MOS, averaged over four speakers)
8. **S60c**, the 3rd 1-min test with codec switching (15 profiles, *N*=17, 15 call-MOS, averaged over four speakers)

Although there are individual call-MOS available for four speakers, the averaged MOS are used here, as there are only averaged MOS and POLQA values for the short segments at hand.

The audio-visual (AV) version of this method results in 1.5 min long German video calls, displaying a person's head and shoulders in front of a fixed background, and during gesturing also arms and hands [1]. Sometimes, dialog-relevant material, like a calendar or brochure, is also displayed by the talker. Five short segments are used to form 15 different degradation profiles, 11 similar for audio and video, and four asymmetric combinations of different audio and video profiles from the symmetric 11. That data set is defined as follows:

9. **AV90a**, the 1st 90-sec test (15 profiles, *N*=23, 15 call-MOS, averaged over four speakers)

The call-MOS values were obtained for 4 speakers (2 males and 2 females), i.e four different dialogues, resulting in 60 call-MOS values. Due to testing limitations, the short segment ratings are averaged over two original clips, i.e. there is one MOS for each quality condition, neglecting semantic content. Similarly, the instrumental estimates are averaged for each condition and over all speakers. This approach differs from the one described in [23] where the segments constituent of the dialogues were individually assessed. It can be expected that the perceived quality of the short samples varies due to the semantic content which may, in turn, impact the modeling accuracy.

A repetition of this procedure with degradations due to wireless networks and switching between speech and video codecs resulted in the final AV-wireless data set [21]:

10. **AV90b**, 2nd 90-sec test (13 profiles, *N*=18, 13 call-MOS, averaged over four speakers)

For the two AV databases, instrumental estimates of the short segments are available for audio (POLQA) and video channels (J.341), combined to an audio-visual MOS estimate as described below.

5. ANALYSIS

For the validation of the different models on the 10 databases, no normalization of the subjective judgments was conducted, as the types of degradations, codecs, number and types of profiles, and number of participants varied too much to decide on an appropriate normalization procedure. Thus, all call-MOS values were taken into account, eight exhibiting audio-only quality, two representing audio-visual quality. Pearson's correlation coefficient *r*, and the root-mean-square error (RMSE) are used as performance indicators, see Table 1. The scatter plot for the best performing model is found in Figure 1.

Table 1: Pearson correlation coefficients and RMSE for call-MOS predictions on the basis of subjective judgments of the individual segments.

Model	*r*	RMSE
Average	.825	.478
ETSI	.914	.266
Weiss	.918	.261
Rosenbluth	.909	.282
Clark	.878	.336
Hamberg	.880	.778

A replication of the modelling, using estimates of short segments' quality (PESQ, P.OLQA, J.341), results in lower fits in general, but in a similar ranking of the models (see Table 2 and Figure 2). Estimates of audio-visual quality are calculated using Equation 11 from [1] for the AV90a and the related Equation 12 from [20] for AV90b, each found for the subjective data and thus applied on instrumental estimates:

$$MOS_{AV} = 0.114 MOS_A + 0.242 MOS_V + 0.103 MOS_A MOS_V + 0.887 \quad (11)$$

$$MOS_{AV} = 0.147 MOS_A + 0.262 MOS_V + 0.09 MOS_A MOS_V + 0.837 \quad (12)$$

As the coefficients are quite similar, a generic integration function might be provided with more audio-visual data to validate.

 DOI: 10.1109/QoMEX.2014.6982329

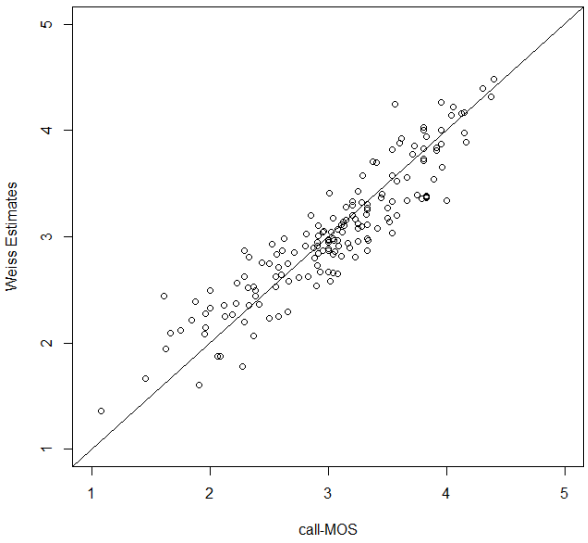

Figure 2: Scatter plot of subjective call-MOS and estimations using the Weiss model with subjective MOS as an input.

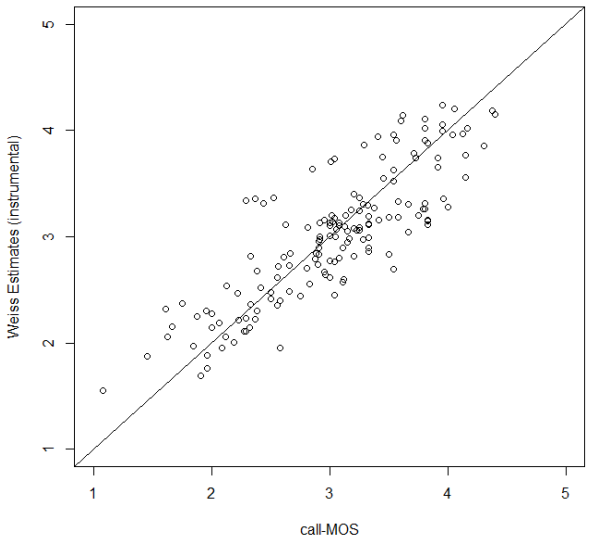

Figure 3: Scatter plot of subjective call-MOS and estimations using the Weiss model using instrumental estimates of MOS as an input.

Table 2: Pearson correlation coefficients and RMSE for call-MOS predictions on the basis of instrumental predictions for the individual segments.

Model	*R*	RMSE
Average	.788	.506
ETSI	.843	.364
Weiss	.846	.362
Rosenbluth	.844	.370
Clark	.804	.430
Hamberg	.824	.803

All models considerably improved the estimates compared to a simple averaging of the short segment quality ratings or estimates. The best performing models are the Weiss, Rosenbluth and ETSI model. However, only default parameter values have been applied, as no parameter optimization was attempted. Therefore, improvements have to be expected, especially for the Hamberg model [1,20]; e.g. setting one of the constants from 1.38 to 1 results in a comparable prediction error (.295 instead of .778). Such optimization is expected to reduce the RMSE to a degree reflecting the ranking of the correlation coefficients. An issue related to this is the reduction of correlation by pooling data sets: As, e.g. for the two AV data sets (90a, 90b), there are apparently different offsets, as the individual correlations are significantly higher for the models than for the pooled data. Especially more AV data is needed to find optimal coefficients.

All well-performing models incorporate effects of the strength of degradation and the position of the short segments; thus, it seems that these two effects are relevant both for speech-only and audio-visual calls.

6. CONCLUSIONS FOR STANDARDIZATION

The simulated conversational test structures work well both in the speech-only and in the audiovisual case; however, it remains to be analyzed how frequently the test participants pay more attention to the audio or to the video channel. This could be analyzed with the help of gaze trackers.

As the data sets available a dominated by audio-only conditions, it cannot be excluded that audio-only and audio-visual stimuli will require different coefficients, even if a single modelling approach will prove to be best for both scenarios. Within the process of standardization, validating the model candidate with call quality ratings from real conversations is also planned.

It remains to be seen for which other episodic experiences such a temporal integration model might work. One obviously interesting application would be video streaming (with a movie serving as an episode), another would be audio streaming (with a podcast of audio story serving as an episode). A third application might be web browsing, where different browsing activities might be integrated to an episode. A fourth one might be gaming.

It will also be interesting to compare models for the temporal integration within one usage episode with models for the temporal integration over multiple episodes; such models are not yet available, but first data which has been collected suggests that the mechanisms might be different in both cases [22][24].

 DOI: 10.1109/QoMEX.2014.6982329

7. REFERENCES

[1] B. Belmudez, B. Lewcio, S. Möller, Call quality prediction for audiovisual time-varying impairments using simulated conversational structures, *Acta Acustica united with Acustica* 99, 792-805, 2012.

[2] J. Berger, A. Hellenbart, R. Ullmann, B. Weiss, S. Möller, J. Gustafsson, G. Heikkilä, "Estimation of 'Quality per Call' in Modelled Telephone Conversations", in: *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, Las Vegas, 4809-4812, 2008.

[3] A. Borowiak, U. Reiter, P. Svensson, "Evaluation of audio quality requirements over extended periods of time using long duration audiovisual content", in: *Advances in Multimedia Information Processing. PCM*, Lecture Notes in Computer Science, vol. 7674, pp. 10–20, 2012.

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

[5] ETSI TR 102 506, *Speech Processing, Transmission and Quality Aspects (STQ); Estimating Speech Quality per Call*, European Telecommunications Standards Institute, Sophia Antipolis, 2007.

[6] R. Hamberg, H. de Ridder, "Time-varying image quality: Modeling the relation between instantaneous and overall quality", *SMPTE Motion Image Journal* 108, 802–811, 1999.

[7] ITU-T Contr. COM 12-340, *Methodology for the Assessment of Audiovisual Quality for Simulated Video Calls*, ITU-T Study Group 12 Meeting, Int. Telecomm. Union, Geneva, 2012.

[8] ITU-T Contr. COM 12-109, *Proposal for a Text for Recommendation P.ACQ: Subjective Method for Simulated Conversation Tests Addressing Speech and Audio-visual Call Quality*, Source: Deutsche Telekom AG (S. Möller, B. Weiss), ITU-T Study Group12 Meeting, Int. Telecomm. Union, Geneva, 2013.

[9] ITU-T Delayed Contr. D.064, *Testing the Quality of Connections Having Time Varying Impairments*, Source: AT&T, USA (J. H. Rosenbluth), ITU-T Study Group 12 Meeting, Int. Telecomm. Union, Geneva, 1998.

[10] ITU-R Rec. BT.500-7, *Methodology for the Subjective Assessment of the Quality of Television Pictures*, Int. Telecomm. Union, Geneva, 1996.

[11] ITU-T Rec. G.107, *The E-Model, a computational model for use in transmission planning*, Int. Telecomm. Union, Geneva, 2005.

[12] ITU-T Rec. J.341, *Objective Perceptual Multimedia Video Quality Measurement of HDTV for Digital Cable Television in the Presence of a Full Reference*, Int. Telecomm. Union, Geneva, 2011.

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

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

[15] ITU-T Rec. P.863, *Perceptual Objective Listening Quality Assessment*, Int. Telecomm. Union, Geneva, 2011.

[16] ITU-T Rec. P.880, *Continuous Evaluation of Time Varying Speech Quality*, Int. Telecomm. Union, Geneva, 2004.

[17] ITU-T Rec. P.910, *Subjective Video Quality Assessment Methods for Multimedia Applications*, Int. Telecomm. Union, Geneva, 2008.

[18] ITU-T Rec. P.911, *Subjective Audiovisual Quality Assessment Methods for Multimedia Applications*, Int. Telecomm. Union, Geneva, 1998.

[19] ITU-T Rec. P.920, *Interactive Test Methods for Audiovisual Communications*, Int. Telecomm. Union, Geneva, 2000.

[20] ITU-T Rec. P.1501, *Subjective Testing Methodology for Web Browsing*, Int. Telecomm. Union, Geneva, 2014.

[21] B. Lewcio, *Management of Speech and Video Telephony Quality in Heterogeneous Wireless Networks*, Springer, Berlin, 2013.

[22] S. Möller, C. Bang, T. Tamme, M. Vaalgamaa, B. Weiss, "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.

[23] B. Weiss, S. Möller, A. Raake, J. Berger, R. Ullmann, "Modeling call quality for time-varying transmission characteristics using simulated conversational structures", *Acta Acustica united with Acustica* 95(6), 1140–1151, 2009.

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