Predicting speech quality based on interactivity and delay

Alexander Raake[1], *Katrin Schoenenberg*[1], *Janto Skowronek*[1], *Sebastian Egger*[2]

[1]AIPA, Telekom Innovation Laboratories (T-Labs), TU Berlin, Germany
[2]Telecommunications Research Center, Vienna, Austria
alexander.raake@telekom.de, egger@ftw.at

10.21437/Interspeech.2013-364

Abstract

A new model of speech quality under delay is presented that includes conversational interactivity. It is based on two previously reported narrowband telephony conversation tests involving different delays, with subject-pairs judging overall quality after each conversation. The tests were conducted with different conversation scenarios targeting different levels of interactivity. The instructions given prior to the tests were varied in their emphasis on speed of task completion. Based on the test results, the paper proposes an extension of a widely used conversational speech quality model, the so-called E-model (ITU-T Rec. G.107), to cover the joint effect of interactivity and delay. To this aim, two new parameters are introduced, one of which represents the minimum perceivable delay, and the other expresses in how far users will attribute the delay-effect to the conversational quality of the line. Based on the analysis of the recorded test conversations in terms of its surface structure (turns, speaker activities, etc.), prominent differences and delay-dependencies of a number of conversation parameters were found that characterize the impact of delay on the conversational flow and on perceived quality.

Index Terms: speech quality, conversation analysis, model, QoE, delay

1. Introduction

The impact of delay is two-fold: (1) In case that acoustic coupling between loudspeaker and microphone exists, delay results in echo; (2) echo-free delay limits the communicability of the connection, which may impact the conversation efficiency or even effectiveness, and sometimes is noticed in terms of low speech quality. Having a quantitative measure of the impact of delay on speech quality and conversation performance still is an unsolved question. This is due to the fact that in many cases of everyday telephone conversations, delay only has a negligible effect on speech quality as it is measured, for example, in conversation tests. Instead, delay influences the conversational structure and behavior of the interlocutors [1]. As already shown by [2], the interactivity of the conversation decides on whether and to which extent delay impacts speech quality. As compared to the 60ies, satellite transmission has lost in relevance, since it has mostly been restricted to one side of a connection and to hard-to-reach areas. In the past years, however, speech quality under delay has become an issue once again, since packet-switched fixed and mobile telephone connections can suffer from higher delays than what is known from circuit-switched telephony. Further, with the increasing variety of device-types, and thus factors that may impact quality, developers and service providers need to know what choices to make for the resulting delay when designing a new network. During the past years, bodies such as the ITU-T have intensively discussed how the additional quality degradation $\Delta Q(\Delta T)$ due to a delay increased by ΔT can be bench-marked against, for example, the gain due to a better speech codec that increases delay. In the present study, we propose an extension of the so-called E-model [3] that covers the speech quality impact due to delay considering the interactivity and delay-sensitivity of the interlocutors. To this aim, in Section 2 we briefly review recent studies on speech quality under delay, different studies on surface-structure analysis of telephone conversations, models proposed for delay detectability and speech quality, and the resulting E-model lgorithm. Section 3 briefly summarizes the two previously reported conversation tests conducted by the authors that form the basis for the model proposed in this paper. The details on the design and surface parameter analysis for these studies are presented in [4]. In the present paper, we use the results for deriving the novel speech quality model that includes the minimum perceivable delay and the delay-sensitivity of a conversation as explicit parameters. Section 4 describes the proposed model and evaluates it against the test data. A short discussion and conclusion are given in Section 5.

2. Speech quality & delay: State of the art

To assess conversational quality, conversation tests are the most realistic method [5]. For such tests, the two interlocutors are typically involved into a conversation using conversation scenarios. A variety of scenarios exist, which lead to different degrees of interactivity (here listed from low to high interactivity): Free conversation, where the two partners speak about a predefined topic or topic of their choice (FC); matching pictures or other data (PM); short conversation tests with both subjects having complementary information to be exchanged, yielding controlled but close-to natural interaction (SCT) [6, 7]; modified SCT, with a more structured information exchange (interactive SCT, iSCT) [5]; random number verification, where subjects take turns verifying a pre-defined and slightly differing set of numbers (RNV) [2]; random number exchange, where subjects exchange numbers in turns as fast as possible (RNE) [2].

2.1. Speech quality impact

For the majority of cases, the effect of delay on speech quality assessed in ITU-T Rec. P.800 and P.805 type conversation tests has been found to be almost negligible, with some noticeable roll-off of quality starting at one-way delay values as high as $Ta > 400\ ms$. Examples taken from Kitawaki & Itoh's and the main author's own work [2, 5] are shown in Fig. 1. As can be seen from the graph, the quality-impact due to delay is rather marginal in case of SCT- or iSCT-based conversation tests. This is in line with more recent findings e.g. described in [8, 9, 10]. The more prominent effect of delay found for RNV in the study

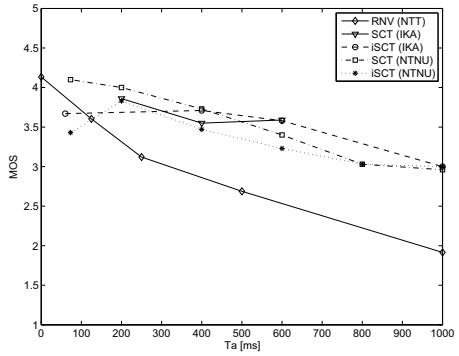


Figure 1: *Speech quality results obtained in some example conversation tests. NTT: Data from [2]; IKA: Data from [5]; NTNU: Data from NTNU, Norway as reproduced in [5].*

by Kitawaki & Itoh [2] has been explained by the authors with the better delay detectability, as will be discussed below. Note that it is this curve which forms the basis for the predictions provided by the so-called E-model [3]. It has been discussed eversince whether this curve actually is realistic, since it apparently was derived by Kitawaki and Itoh with trained subjects.

2.2. Conversational impact

When a two-party telephone conversation is affected by echo-free delay, people may react differently to it depending on the type of conversation and the two partners' individual conversation behavior. The impact of delay on conversations can be addressed using conversation analysis in terms of surface structure. Approaches of this type are described, for example, in [11, 12, 2, 13, 14, 15, 16, 4]. In these studies, the authors have sought parameters that describe the conversation in case of no or low delay, and analyzed how the parameters behave under delay. The ultimate goal is to come to a parameter set that captures the delay-impact on *perceived* quality.

In-depth considerations on different conversation structure parameters for the described tests can be found in [4], showing different sojourn times and state probabilities in conversation states such as mutual silence (M), double talk (D) and single talk (S) with different transitions. In addition, the *speaker alternation rate* (SAR) was proposed by [13] for characterizing conversational interactivity. It quantifies the number of alternations between the talkers per minute. These speaker alternations can be transitions from S_AA to S_BB (single talk of speaker A or B seen from the respective viewpoint of that speaker, A, B) either over mutual silence or over double talk. SAR decreases with increasing delay, as expected due to the increased amount of silences. The authors of this paper have proposed a complementary measure in [4], the SAR corrected by the delay for each alternation that passes by mutual silence, $SARc$. As shown in Fig. 4 and [4], this measure is much less delay-dependent, but like the uncorrected SAR can well account for different types of scenarios.

The following model considerations will focus on $SARc$ only to describe the role of the conversation in delay perception. To capture the users' sensitivity to delay, two additional parameters are used: the *delay sensitivity* sT and the *minimal perceivable delay* mT. Here, sT can be interpreted as the extent to which users may attribute any perceivable delay impairment to the quality in terms of the system. In turn, mT is the delay threshold required for a given conversation to yield perceivable delay. Note that these two parameters are not conversation structure parameters, and will be discussed further in Section 4.

2.3. E-Model

To make direct use of the model developed in this paper for network planning and monitoring, the so-called E-model is used as the modeling basis. It is the tool recommended by ITU-T for network planning [3]. Using parameters known during network planning such as frequency weighted line attenuation values, delay or packet loss rates, or the model-specific degradation introduced by a given speech codec, the model provides predictions of the expected speech quality. In contrast to signal-based models such as PESQ [17] and [18], it allows the conversational situation to be considered. The E-model is based on the assumption that degradations related with technical characteristics of the end-to-end transmission can be transformed onto a perceptual scale, the so-called Transmission Rating scale (R-scale). On this scale, certain types of degradations are considered to be additive in terms of their impact on overall quality R:

$$R = Ro - Is - Id - Ie,eff + A \tag{1}$$

The transmission rating R ranges from 0-100 for narowband (300-3400 Hz), and 0-129 for wideband (50-7000 Hz) [5, 19, 20]. Using an S-shaped relation, it can be transformed to the MOS-scale as used in typical Absolute Category Rating (ACR) tests according to ITU-T P.800. In Eq. 1, the *basic signal-to-noise ratio* Ro refers to the quality due to the signal-to-noise ratio (reflecting line and room noise and the speech levels) The *simultaneous impairment factor* Is covers all degradations that are simultaneous with the transmitted speech such as signal-correlated noise or inadequate speech levels. The *delayed impairment factor* Id covers all degradations delayed to the speech signal, such as pure, echo-free delay and listener and talker echo. The *effective equipment impairment factor* Ie,eff includes the impairment due to coding and coding under packet loss. The *advantage factor* A enables the predictions to be adjusted to the context of use, so that, for example, the higher tolerance for degradations in a mobile context can be covered. For this paper, $A = 0$.

The delayed impairment factor can be further decomposed as:

$$Id = Idte + Idle + Idd, \tag{2}$$

where $Idte$, $Idle$, and Idd are the impairment due to talker echo, listener echo and echo-free delay, respectively. The pure delay handling provided by the E-model has been reflected in the thresholds for ideal and no longer acceptable mean one way delay values, with $Ta = 150\ ms$ in the earlier, and $Ta = 400\ ms$ in the latter case [21]. In the E-model, the delay-impairment factor Idd is calculated as follows:

$$Idd = \begin{cases} 0 & \text{for } Ta \leq 100 \text{ ms} \\ 25\left\{(1+X^6)^{1/6}\right. & \\ -3\left(1+[X/3]^6\right)^{1/6}+2\big\} & \text{for } Ta > 100 \text{ ms} \end{cases}, \tag{3}$$

with

$$X = \frac{log_{10}(Ta/100)}{log_{10}\,2}. \tag{4}$$

3. Conversation tests

Two conversation tests have been conducted, one at Telekom Innovation Laboratories/TU Berlin, Germany ("T-Labs"), the

other at the Telecommunications Research Center, Vienna, Austria ("FTW"). In both cases, SCT- and RNV-type tests have been conducted. In the case of FTW, one additional set of conditions was run with the iSCT scenarios [5, 7]. At T-Labs, one test was conducted as a random number verification task, however introducing a competition by instructing the subjects that the fastest conversation-couple would obtain a reward, thus creating an incentive for faster conversations (RNT – random number verification, timed). The reward was written out for the fastest and most correct pair of conversation partners. The tests were conducted following [22, 7]. In the T-Labs tests, Snom 870 IP telephones were used (high echo attenuation), in the FTW tests, headsets with closed earphones. For more details on the test set-up, see [4]. The test settings are shown in Table 1.

	FTW	T-Labs
# subjects	34	48
Mean Age	23.15 (SD=3.36)	30.44 (SD=8.36)
Female/ Male	F: 11 / M: 23	F: 24 / M: 24
Network	VoIP + NetEm	VoIP + NetEm
Codec	G.711	G.711
Scenario	SCT_1, RNV_1	SCT_2, RNV_2, RNT
Delays[ms]	100,200,400,800,1600	100,225,425,825,1625

Table 1: Test settings for FTW and T-Labs conversation tests. See [4] for more details

For prediction algorithm development, the mean quality ratings in terms of Mean Opinion Scores (MOS) have been transformed to the E-model R-scale using the S-shaped transformation given in [3]. The result were R-values for the respective test and scenario $R_{xx,scen}$, with xx indicating the test lab ($xx \in$ [FTW,T-Labs]) and $scen$ indicating the test scenario ($scen \in$ [SCT,iSCT,RNV,RNT]). Subsequently, the R-ratings have been transformed to obtain the estimated impairment values associated with pure delay, using the following equation:

$$\mathbf{Idd}'_{xx,scen} = \mathbf{R}_{xx,G.107} + \mathbf{Idd}_{xx,G.107} - \mathbf{R}_{xx,scen}. \quad (5)$$

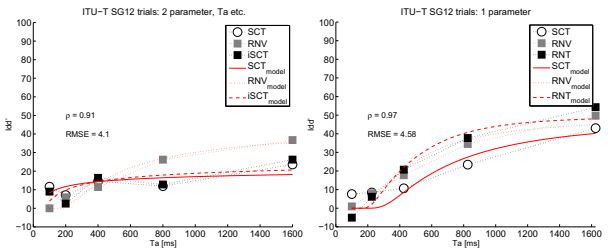
Here, $\mathbf{Idd}'_{xx,scen}$ is the vector for the impairment values for the given lab and scenario to be used for model development, with each entry representing one delay condition. $\mathbf{R}_{xx,G.107}$ is the vector with the predictions provided by the current version of the E-model for the different delay conditions, and $\mathbf{Idd}_{xx,G.107}$ the respective delay-related impairment factor vector. Note that $\mathbf{R}_{xx,G.107}$ reflects additional small degradations due to talker echo (all phone characteristics have correctly been used for determining these values). Hence, the sum of the latter two indicates the different quality values if delay is excluded, but all other aspects such as echo and the line noise settings used in the test are being considered. Subtraction of the MOS-ratings transformed to the R-scale, $\mathbf{R}_{xx,scen}$, leads to a vector of E-model-specific delay-impairment values $\mathbf{Idd}'_{xx,scen}$.

4. Results and Model

The test results are shown in Figure 2. As can be seen from the graph, for the same conversation scenarios, the two tests differ in the respective results, as already indicated in [4].

For modeling of the data, the E-model formula shown in Equation (3) has been modified as follows:

$$Idd = \begin{cases} 0 & Ta \le mT \\ 25\left\{(1+X^{6\cdot sT})^{1/(6\cdot sT)}\right. & \\ \left. -3\left(1+[X/3]^{6\cdot sT}\right)^{1/(6\cdot sT)}+2\right\} & Ta > mT \end{cases}, \quad (6)$$



(a) (FTW) (b) (T-Labs)

Figure 2: *Transformed results from conversation tests, shown in terms of impairment values Idd. Also shown are the model predictions including model performance indicators for the two tests, obtained with model #1 (see Tab. 2 and text).*

with

$$X = \frac{log_{10}(Ta/mT)}{log_{10}\,2}. \quad (7)$$

Here, the fixed delay perception threshold of 100 ms has been replaced by the parameter mT (minimum perceivable delay, see above), and the exponent of 6 is multiplied by the delay sensitivity sT. With these choices, the resulting model is identical with the current version of the E-model, when $mT = 100$ ms and $sT = 1$. This is an important property, since the conservative quality values estimated by the E-model are the basis for lower-bound delay planning standards such as [21]. In turn, when choosing different settings for mT and sT, the model predictions are well in line with the test results shown in Fig 2. Here, three different models have been considered: (1) For Model #1 with two free parameters, both mT and sT are used as coefficients in a least-square fitting process using Eqs. (6) and (7). (2) For Model #2, with only one free parameter, mT is set to a fixed, scenario-specific value obtained from the fitting results for model #1. (3) Model #3 uses only sT as free parameter, with $mT = 100$ ms.

The resulting fitting-values are listed in Tab. 2. Note that in addition to results for the two tests conducted at FTW and T-Labs, respective fitting results are also provided for the test results shown in Fig. 1. The MOS-results were transformed to the R-scale as described above, but here the delay impairment ratings Idd were calculated by simply subtracting the R-value for a given condition from the maximum R-value obtained in the respective test. As can be seen from the table, $sT = 1$ can be considered as an upper bound for the delay-sensitivity sT: As expected, the current version of the E-model provides conservative predictions that correspond to a worst-case interaction scenario.

		#1, 2 p.		#2, 1 p.		#3, 1 p.
	Lab.	mT(ms)	sT	mT(ms)	sT	sT
SCT	T-Labs	175.5	0.78	150	0.61	0.42
	FTW	27.7	0.23	150	0.29	0.27
	NTNU	165.7	0.43	150	0.39	0.30
iSCT	FTW	58.9	0.25	150	0.32	0.28
	NTNU	175.3	0.30	150	0.28	0.24
RNV	T-Labs	117.5	0.84	100	0.69	0.69
	FTW	138.3	0.50	100	0.40	0.40
	NTT	52.5	0.62	80	0.89	1.20
RNT	T-Labs	114.9	1.24	100	0.98	0.98

Table 2: Curve fitting results for different cases. The results for model #2 are shown in Fig. 2.

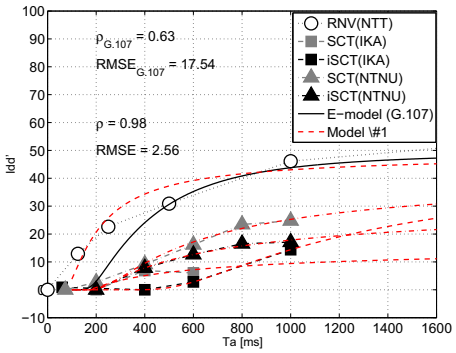


Figure 3: *Speech quality results and predictions for tests from literature (cf. Fig. 1). NTT: Data from [2]; IKA: Data from [5]; NTNU: Data from NTNU, Norway as reproduced in [5].*

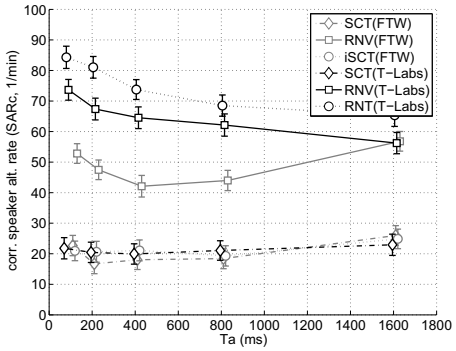


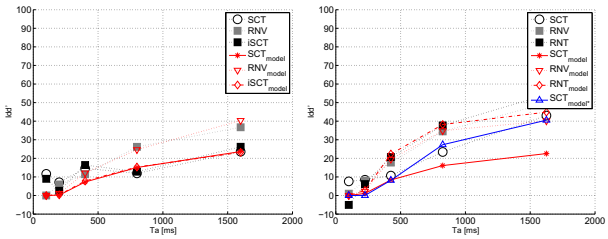
Figure 4: *Speaker alternation rate as a function of delay.*

The respective prediction results for the data given in Fig. 1 with Model #1 are shown in Fig. 3, indicating a much better prediction accuracy ($\rho = 0.98$, $RMSE = 2.56$) than obtained with the current E-model (G.107: $\rho = 0.63$, $RMSE = 17.54$). Similar results can be obtained using the simpler Model #3. From Figs. 2 and 3 it is obvious, that the proposed capturing of delay-sensitivity enables planning predictions with a higher accuracy.

4.1. Relating sT with conversation parameters

In a second step, the new parameters sT and mT were related with conversation structure descriptors. The presentation focuses on $SARc$. As shown in Fig. 4, $SARc$ shows only a limited delay-dependency, which reflects the inherent adaptation of conversation behavior of users to the given delay, without the misleading multiple inclusion of delay when calculating SAR. Thus, $SARc$ is a good start for capturing the effect due to conversational interactivity.

The modeling procedure was as follows: (a) $SARc$ obtained for every conversation type and test have been averaged over the different delay settings to obtain a scenario-test-specific value $mSARc$. (b) For each such setting, the values sT and mT have been plotted over the $mSARc$ values. (c) Using least-square curve-fitting with respectively chosen functions, a relationship between $mSARc$ and sT and mT according to Fig. 4 was determined for the different models. (d) It is assumed that the relation obtained for the *mean* values $mSARc$ will also hold for the values obtained for a given conversation i at a given delay setting Ta_i. The respective mapping between



(a) FTW (b) T-Labs

Figure 5: *Predictions with Eq. (6) using the mappings of $SARc$ to sT and mT based on actually measured $SARc$-values.*

$SARc$ and sT and mT was then used to directly calculate the required values for Equation (6).

The approach leads to highly accurate predictions, see Fig. 5. For this paper, only the results for model #1 are shown. The mapping functions used are shown in Eqs. (8) and (9).

$$mT = 436.02 - 71.56 \cdot \log(16.76 + SARc) \quad (8)$$

$$sT = 0.246 + 0.02 \cdot \exp(0.053 \cdot SARc) \quad (9)$$

Note that for the SCT-tests conducted at T-Labs, subjects obviously were much more sensitive to delay than what the $SARc$-values imply, and what was obtained in the FTW tests. In [4], it was discussed that this is likely due to the mixed test with RNT, RNV and SCT conditions in conjunction with the more strict test instructions for the T-Labs test. Hence, test participants were obviously much more pointed to the delay impairment, and included it more strongly in their judgments. As a consequence, in addition to the *noticeability* of delay in terms of interactivity observable at conversation structure level, the *attention paid* to the delay impairment and its attribution to quality of the system become important. In terms of the model, in this case the delay-sensitivity parameter sT obviously is not only dependent on $SARc$, but also on a non-measurable delay-attention parameter we here refer to as aT. To adjust the model predictions to the the quality ratings for the case T-Labs/SCT, a modified derivation of sT can be used, for example: $sT = f_2(SARc, aT) = f_1(SARc) \cdot aT$, with $aT = 2.54$ (blue curve in Fig. 5). Obviously, people's sensitivity to delay was by a factor 2.5 higher in the case SCT(T-Labs) than found for the FTW test (SCT, iSCT).

5. Discussion and conclusion

The paper has shown that speech quality under delay can be predicted using a modified version of the ITU-T-recommended network planning tool, the E-model. By including two new parameters, namely the minimum perceivable delay mT and the delay sensitivity sT, high correlations with existing conversation test results could be obtained. The new parameters enable the explicit inclusion of delay-sensitivity for specific conversation types. The next step will target the update of Rec. G.107 based on a set of three to four defined conversation classes with respective values of mT and sT. For quality monitoring, a first model was proposed that includes a mapping of the speaker alternation rate $SARc$ obtained from the actual conversations to sT and mT. Here, the attention of users to delay that cannot be captured by conversation structure parameters needs to be considered in addition (aT). In future work, further conversation tests with both listening- and conversation-degradations will be analyzed, targeting improved planning and monitoring models.

6. References

[1] R. M. Krauss and P. D. Bricker, "Effects of transmission delay and access delay on the efficiency of verbal communication," *J. Acoust. Soc. Am.*, vol. 41, no. 2, pp. 286–292, 1966.

[2] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE J. on Selected Areas in Communications*, vol. 9, no. 4, pp. 586–93, 1991.

[3] ITU–T Rec. G.107, *The E–Model, a Computational Model for Use in Transmission Planning*, International Telecommunication Union, CH–Geneva, 2009.

[4] S. Egger, R. Schatz, K. Schoenenberg, A. Raake, and G. Kubin, "Same but different? - using speech signal features for comparing conversational VoIP quality studies," in *IEEE ICC 2012 - Communication QoS, Reliability and Modeling Symposium (ICC'12 CQRM)*, Jun. 2012.

[5] A. Raake, *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: John Wiley & Sons Ltd, 2006.

[6] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. USA–Boston: Kluwer Academic Publishers, 2000.

[7] I.-T. R. P.805, *Subjective Evaluation of Conversational Quality*, International Telecommunication Union, CH–Geneva, April 2007.

[8] M. Guéguin, R. L. Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," *EURASIP J. Adv. Signal Process*, vol. 2008, pp. 1–15, 2008.

[9] J. Holub, M. Kastner, and O. Tomiska, "Delay effect on conversational quality in telecommunication networks: Do we mind?" in *Wireless Telecommunications Symposium, Pomonia, California, USA*, 2007.

[10] J. Issing and N. Farber, "Conversational quality as a function of delay and interactivity," in *Software, Telecommunications and Computer Networks (SoftCOM), 2012 20th International Conference on*, 2012, pp. 1–5.

[11] P. Brady, "A technique for investigating on-off patterns of speech," *Bell System Technical Journal*, vol. 44, no. 1, pp. 1–22, 1965.

[12] ——, "A statistical analysis of on-off patterns in 16 conversations," *Bell System Technical Journal*, vol. 47, no. 1, pp. 73–91, 1968.

[13] F. Hammer, P. Reichl, and A. Raake, "Elements of interactivity in telephone conversations," *In: Proc. Int. Conf. Spoken Language Processing (ICSLP 2004), KR–Jeju Island*, 2004.

[14] ——, "The well-tempered conversation: Interactivity, delay and perceptual VoIP quality," *In: Proc. IEEE International Conference on Communications (ICC 2005), KR–Seoul*, 2005.

[15] S. Egger, R. Schatz, and S. Scherer, "It takes two to tango – assessing the impact of delay on conversational interactivity on perceived speech quality," *In: Proc. INTERSPEECH 2010*, 2010.

[16] S. Egger, R. Schatz, K. Hoeldtke, A. Raake, and G. Kubin, "Same but different? - surface parameter based comparison of conversational speech quality studies," *Subm. to INTERSPEECH 2011*, 2011.

[17] ITU–T Rec. P.862, *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, International Telecommunication Union, CH–Geneva, February 2001.

[18] ITU–T Rec. P.863, *Perceptual objective listening quality assessment (POLQA)*, International Telecommunication Union, CH–Geneva, 2011.

[19] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, "Impairment factor framework for wideband speech codecs," *IEEE Trans. Audio Speech and Language*, vol. 14, no. 6, pp. 1969–1976, 2006.

[20] ITU–T Rec. G.107.1, *Wideband E-model*, International Telecommunication Union, CH–Geneva, 2011.

[21] ITU–T Rec. G.114, *One–Way Transmission Time*, International Telecommunication Union, CH–Geneva, May 2000.

[22] ITU–T Rec. P.800, *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, CH–Geneva, June 1996.