4th International Workshop on Perceptual Quality of Systems (PQS 2013)
2-4 September 2013, Vienna, Austria

A Management Conversational Quality Predictor

Jan Holub[1], Peter Pocta[2], Jean-Yves Monfort[3], Joachim Pomy[4]

[1]Department of Measurement, 13138, FEE, Czech Technical University, Prague, Czech Republic
[2]Department of Telecommunications and Multimedia, FEE, University of Zilina, Zilina Slovakia
[3]JYM C.I.S., Pleumeur-Bodou, France
[4]Telecommunications & Int'l Standards, Bensheim, Germany

holubjan@fel.cvut.cz, pocta@fel.uniza.sk, jeanyves.monfort.6@orange.fr, consultant@joachimpomy.de

Abstract

In a majority of modern networks traditional end-to-end transmission planning is no longer possible, but popular field testing in technologies such as UMTS, NGN and LTE, typically reveals only one quality component of the QoS. This paper describes an operational quality estimator developed within ETSI STF 436 project called Management Conversational Quality Predictor (MCQP) which can combine results from field trials with other impairments such as one-way delay, echo loss, type of conversation, etc. Potential areas for an extension of the designed model are outlined in a conclusion of the paper.

Index Terms: E-model, Conversational quality, Field testing, Quality predictor

1. Introduction

The need for a conversational quality estimator for technical management level has been identified from different stakeholders because the one-way quality is not the quality really experienced by the users involved in a conversation. The following principles have been assumed during the development phase of the estimator:

- Principle 1: to provide a decision support tool for the management level.
- Principle 2: to hide parameters which are not needed by transmission planners these days or not accessible/monitored and which may create confusions for technical managers instead of helping them. In other words, many parameters are either not known to the technical decision makers, or they could have a wide range of values (e.g. the real terminal quality, the user's speech level, the local and distant noise levels).

Today, the current E-model, defined in the ITU-T Recommendation G.107 [1], is rarely used to support decisions before changes are implemented in a network. Management is asking how much impact deployment of a new technology will have on a quality perceived by the user. So, a tool described in this paper entitled Management Conversational Quality Predictor (MCQP) implements the parameters effectively impacted by these new technologies. Instead of providing instructions for many parameters, most of which finally are left at their default values, it is better to hide these parameters inside the tool, and make only most important network parameters available, such as delay, talker echo, listening quality and interaction level. In order to compare the developed approach with the E-model, several graphs are provided as a result of the project, comparing subjective results, the new predictor outputs and the E-model values for a number of variable parameters. If the technical managers are currently using E-model, they will be able to use these graphs to move to the new predictor without losing the historical evolution of the networks.

To reach the goal, a set of conversational tests according to ITU-T P.805 [2] in English and Czech were run. The tool is based on these subjective test results. The test scenarios are defined in order to create several conversational interactivity levels between the two subjects involved in each conversation. To take into account the interactivity between the talkers, a new parameter called Talker Alternation Rate (TAR) is introduced. In some previous works (e.g. by F. Hammer in [10]), Speaker Alternation Rate (SAR) was used to denote the same, however, SAR can make confusion having alternative meaning of Specific Absorption Rate. Therefore, authors of this contribution propose to use TAR instead.

The remainder of this paper is structured as follows. Section 2 describes an experiment design. Section 3 presents subjective test results. Section 4 analyses the test result and defines the MCQP comparing its performance with E-model. The results are discussed in Section 5. Section 6 concludes the paper and suggests some areas for future research.

2. Experiment Design

The subjective conversation tests are covering the following characteristics:

- Different coders: 3 coders, G.711 [3] A-law, G.729AB [4] (@ 8kbit/s), AMR-NB [5] (@ 12.2kbit/s)
- Different delay values: 3 values, 100, 300, 600 ms one-way delay
- Different echo situations: 2 situations, weak echo, strong echo, TELR= 46dB, 32dB
- Different conversational scenarios: 3 levels of interactivity i.e. different categories. The exact test scenarios can be found in Annexes B and C of ETSI TR 103 121 [6] and are mostly based on the scenarios defined in ITU-T P.805 [2].

10.21437/PQS.2013-26

- 54 conditions in English, 18 conditions in Czech, in total 72 conditions
- 48 votes per condition (equals to 3456 votes in total)
- The equivalent of a reference terminal - real-time adaptation to ES 202 737 [7] in send and receive direction, with diffuse field correction as per ITU-T Recommendation P.57 [8]
- Different languages - The majority of tests are conducted in English language and a limited number of tests are in Czech language.

3. Subjective Test Results

The subjective conversational tests have been performed on 24 English and 8 Czech native talker pairs. The test environment conformed to ITU-T P.800 requirements. A proprietary DSP-based real-time network simulator has been used. Its terminals have been calibrated on professional Head and Torso Simulator. Details can be found in [6].

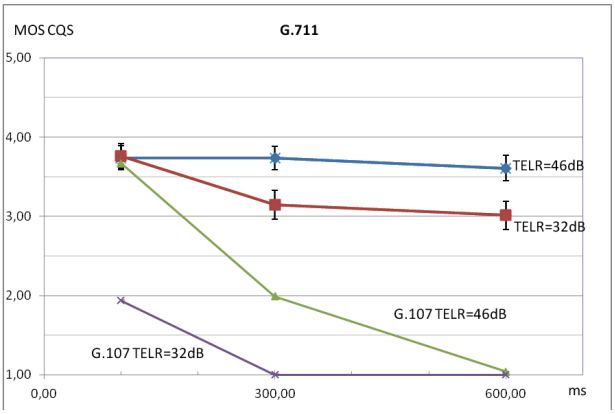

Figure 1: *Subjective test results example: G.711 coder and two tested TELR values (32dB, 46dB) including CI95% uncertainty intervals. Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600ms, the connecting lines are shown for informative purposes only [6].*

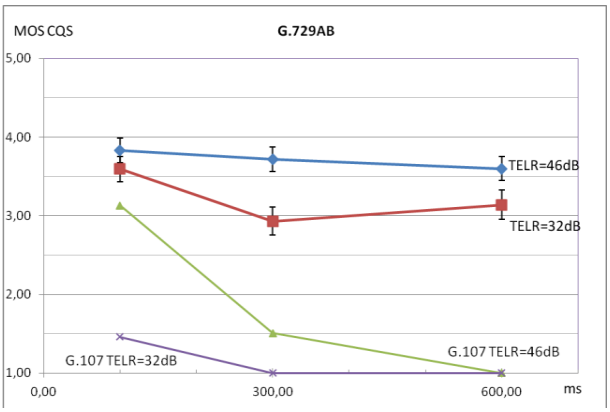

Figure 2: *Subjective test results example: G.729AB coder and two tested TELR values (32dB, 46dB) including CI95% uncertainty intervals. Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600ms, the connecting lines are shown for informative purposes only [6].*

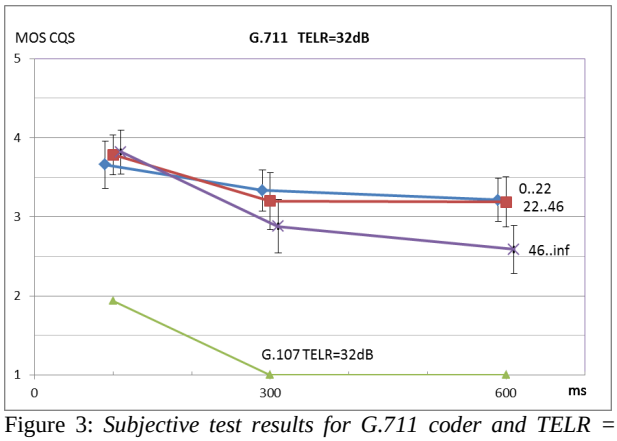

Figure 3: *Subjective test results for G.711 coder and TELR = 32dB, split for 3 different interactivity levels based on TAR analysis, including CI95% uncertainty intervals. Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600ms, the connecting lines are shown for informative purposes only [6].*

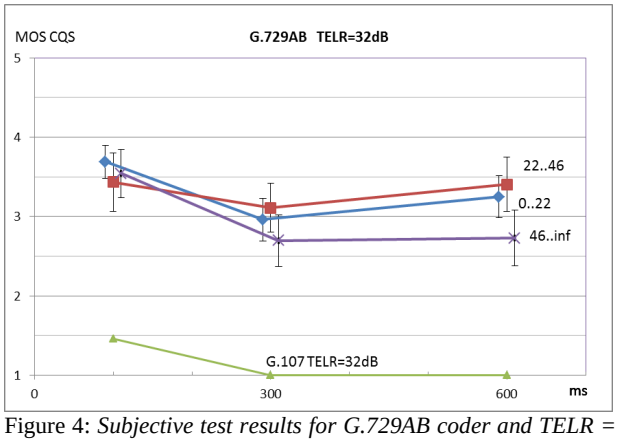

Figure 4: *Subjective test results for G.729AB coder and TELR = 32dB, split for 3 different interactivity levels based on TAR analysis, including CI95% uncertainty intervals. Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600ms, the connecting lines are shown for informative purposes only [6].*

As outlined above, the graphs are derived to show the differences between the E-model and the new approach. In fact, two different values of MOS-CQ are obtained for each combination of input parameters (codec, delay, echo level, etc.):

- The E-model (G.107) output, recalculated from R to MOS scale (referred further as „E-model“)
- MOS-CQS as obtained by subjective tests with appropriate 95% confidence intervals (CI95%)

Results obtained for other coder, TELR and TAR combinations can be found in [6].

4. Result Analysis and Model Design

For low echo condition of TELR=46dB, the subjective sensitivity to delay is significantly lower than as predicted by E-model. The typical difference between MOC-CQS for 100ms and 600ms is for low echo condition approximately 0.5 MOS. For coders deploying higher perceptual compression (G.729AB) affecting the listening quality the MOS-CQS becomes for stronger echo (TELR=32dB) non-monotonic with new local minima located (in our case) at 300ms. Similar effects have been reported by previous experiments by various labs [9].

The MCQP model was designed and trained based on polynomial fit of subjective test data. Only English data have been used for the training, but the model is valid for Czech data, too. Its input variable values are:

- end-to-end delay
- the talker echo (TELR)
- Talker Alternation Rate (TAR)
- Coder used (affecting listening quality)

It should be noted the model is quite simple as the number of parameters is currently limited. Further subjective data are needed to properly consider other important parameters e.g. effect of background noise or other possible impairments.

However, for the given set of subjective data it achieves significantly higher correspondence with conversational subjective data than the E-model in the context of different call types. It also considers the influence of call interactivity and distorted echo that is not considered by E-model at all.

The following analyses have been performed and are reported in Table1:

a) Pearson correlation coefficient *R* between MOS-CQS and E-model output. This analysis shows the differences between existing standardized estimator and subjective test results.
b) *RMSE** against E-model (root mean squared error with suppressed influence of subjective testing uncertainty). This analysis shows the differences between the nearest CI95% interval border and the standardized E-model result (zero if the E-model output is located within the CI95% interval).
c) Pearson correlation coefficient *R* between MOS-CQS and the developed predictor MCQP output. This analysis shows the difference between the developed predictor and subjective test results.
d) *RMSE** against the developed predictor MCQP (root mean squared error with suppressed influence of subjective testing uncertainty). This analysis shows the differences between the nearest CI95% interval border and the developed predictor MCQP (zero if the MCQP output is located within the CI95% interval).

Table 1 *Result analysis overview [6]*.

	MOS-CQS versus E-model	MOS-CQS versus MCQP
R	0,546	0,911
RMSE	1,984	0,148
*RMSE**	1,722	0,029

4.1 Language comparison

The comparison of results of tests performed in Czech language and in English language clearly indicate insignificant (0.2 MOS in average) systematic offset causing Czech testers being virtually more demanding (more critical), however, the reason of this systematic offset is not clear. It can be caused e.g. by slightly lower average TAR for Czech tests (32,6) than for English tests (34,4) or by different age distribution or by some other unknown reason.

5. Discussion

As follows from Table 1, new estimator MCQP outperforms the E-model for the given set of test conditions in *R, RMSE* and *RMSE** parameters. However, due to the following significant differences, the estimations provided by the E-model and MCQP can not be directly compared:

The E-model is a complex model taking into account a lot of different parameters and due to its rather pessimistic results (in particular linked with high delay figures) it delivers safe predictions during network planning phase or to guarantee a high quality e.g. for business calls. However, its results are questionable to use during the operational phase and are impaired by a lack of fundamental inputs like interactivity (characterized by TAR). Also the amount of distortion in echo caused by multiple coding of the echo signal is not reflected (only TELR and echo delays are considered).

The MCQP model, on the contrary, provides a good match with MOS-CQS (conversation quality) because one of the major innovations of MCQP is to take into account the interactivity between the talkers and to introduce the new parameter TAR which is very important for the overall quality of speech conversations. At this stage it is rather simple and considers only a limited set of parameters covered by the subjective tests performed within the project. The results provided by the developed model (see for instance the graphs presented in the ETSI TR 103 121 [6]) and a reference implementation give the opportunity to technicians and managers to determine the expected quality of communications, taking into account delay, talker echo, listening quality and TAR. However, to be wider applicable, it should be extended towards other parameters such as noise (effects of noisy environments and of noise cancellation), and bandwidth (considering wideband and super wideband speech).

6. Conclusions

The developed model applies in particular for new IP-based networks where the end-to-end delay may be high and could be seen as a model dedicated to NGN and new mobile networks (e.g. UMTS and LTE).The model has been published as ETSI TR 103 121 [6] and is available on http://www.etsi.org/standards

One of the major innovations of MCQP is to take into account the interactivity between the talkers and to introduce a new parameter TAR which is very important for the overall quality of speech conversations. The graphs made available in this paper give the opportunity to technicians and managers to determine the expected quality of communications, taking into account delay, talker echo, listening quality and TAR.

It should be also noted here that the approach presented in this paper is based on the assumption that one major application

area of the functions described here will be the inclusion of MOS-LQO values derived by drive testing (where no background noise is present) into the dimension of the E-Model. However, since background noise is an important factor in quality perception by users, it is advisable to extend the present approach by background noise aspects. The approach applies only for narrowband speech and should be extended at least for wideband and possibly to higher bandwidths. The model can be also expanded in order to become applicable to dynamic situations, considering IP-related impairments and impairments related to the radio links.

7. References

[1] ITU-T Recommendation G.107: The E-model: a computational model for use in transmission planning.

[2] ITU-T Recommendation P.805: Subjective evaluation of conversational quality.

[3] ITU-T Recommendation G.711: Pulse code modulation (PCM) of voice frequencies.

[4] ITU-T Recommendation G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic code-excited linear-prediction (CS-ACELP).

[5] 3GPP TS 26 071: Mandatory speech CODEC speech processing functions; AMR speech Codec; General description".

[6] ETSI TR 103 121: Speech and multimedia Transmission Quality (STQ); Adaptation of the ETSI QoS Model to better consider results from field testing.

[7] ETSI ES 202 737: Speech and multimedia Transmission Quality (STQ); Transmission requirements for narrowband VoIP terminals (handset and headset) from a QoS perspective as perceived by the user.

[8] ITU-T Recommendation P.57: Artificial ears.

[9] Holub, J. - Tomíška, O.: Non-monotonicity in Perceived Quality of Delayed Talker Echo. In Measurement of Speech, Audio and Video Quality in Networks. Prague: Czech Technical University, 2007, p. 67-68. ISBN 978-80-01-03734-8.

[10] F. Hammer: "Quality Aspects of Packet-Based Interactive Speech Communication", Ph.D. Thesis. TU Graz 2006