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

On the influence of involvement on the quality of multiparty conferencing

Janto Skowronek, Falk Schiffner, Alexander Raake

Assessment of IP-based Applications, Telekom Innovation Laboratories,
Technical University of Berlin, Berlin, Germany

janto.skowronek@telekom.de, alexander.raake@telekom.de

Abstract

The present study is a first investigation on the potential influence of conversation and task involvement on quality ratings in the context of multiparty telemeetings. The results show that involvement has indeed a significant impact, and it can be larger than the effect of the tested technical degradations. This study confirms experience from the two-party context and gives an estimate of rating differences between listening and conversation tests.

Index Terms: conferencing, involvement, quality

1. Introduction

Multiparty conferencing and telemeeting systems are continuously gaining more and more importance in professional life and, especially in recent years, also in private life. Despite this trend, knowledge on the quality assessment of such systems is still limited to a certain extend: the ITU-T provides a recommendation on the subjective assessment of such multiparty telemeetings [1], based on a number of published and ITU-T internal studies (e.g. [2, 3, 4, 5, 6]). However, [1] also points out that more experience needs to be built up in order to further improve such assessment methods. The present study intends to contribute to this knowledge build-up by investigating two research questions: First, how far would quality ratings be affected by the involvement of test participants into the conversation? Second, what would be the consequences for the design of a quality assessment test, i.e. the selection of the test paradigm and corresponding tasks that participants are conducting during the test?

Concerning the test paradigm, there are two main options for the assessment of telecommunication quality: listening tests and conversation tests. It is well accepted in the field (e.g. [7]) that conversation tests are more natural than listening tests, while listening tests enable subjects to focus on the stimuli, leading to a higher sensitivity than conversation tests. Thus the task that subjects do during a test, e.g. listening vs. conversation, can influence the quality rating, depending on the amount of distraction from perceiving any degradations caused by the task. For that reason, the reporting of quality ratings in terms of Mean Opinion Scores (MOS) should always clarify, if the MOS ratings are obtained by a listening-only or conversation test. For instance, [8] is standardizing a corresponding nomenclature, and [9] is providing guidance for an appropriate reporting and interpretation of MOS values.

Considering the quality assessment of multiparty teleconferencing systems, it can be assumed that the task dependency on quality ratings exists here as well. However, task dependency might be even more prominent in multiparty tests than in conventional speech quality tests for two reasons.

First, a multiparty conference requires in general a higher cognitive load from participants than a one-to-one call [10]. As humans have a maximum cognitive capacity that needs to be shared between simultaneous tasks (e.g. [11]), test subjects have less cognitive capacity available for the quality rating task due to the extra effort needed for following the multiparty conversation. Furthermore, if the test is designed such that subjects are more involved in the task, e.g. playing an active role in a scenario compared to passively listening to a conversation, subjects experience an additional cognitive load due to the task involvement.

Second, one of the main differentiators between a multiparty conference and a one-to-one call is the special conversational situation, i.e. a group communication via a telecommunication medium. That means test subjects should be presented with stimuli or conversation tasks that sufficiently resemble such a conversational situation [1]. This brings us to the notion of conversational involvement: the more a test subject is feeling to be involved in a conversation, the more the subject's quality reference will be based on the conversational situation, and thus the better the quality rating will reflect the multiparty experience.

However, in practise it is difficult to separate between "conversational involvement" and "task involvement", because we here consider quality assessment tests, in which subjects will be explicitly asked to do tasks, which in turn require from subjects to listen or contribute to a conversation. Therefore, we will here not distinguish between these two aspects, but will investigate involvement from a holistic perspective.

After having discussed the importance of involvement for multiparty quality assessment, we can now for-

10.21437/PQS.2013-25

mulate our research questions more precisely:

1. How large is the difference in quality ratings when test subjects perform assessment tasks with different levels of involvement; what would be the consequences of such differences for selecting an appropriate assessment method?
2. To what extend are test subjects able to differentiate between quality, cognitive load and involvement; what would be the consequences for designing appropriate assessment questionnaires?

Furthermore, a number of interesting side aspects that we wanted to check are:

3. In the context of the different experimental tasks, how did subjects perceive the subtask of actually giving a quality rating?
4. Is there a difference between assessing involvement directly after a conference call and in retrospective after the whole experiment?

To answer these questions, we conducted an experimental study, which will be described and analyzed in the remainder of this paper.

2. Experimental study

Twenty test subjects (eleven female, nine male, average age: 29 years, 12 with and 8 without conferencing experience) were invited to an one-hour experiment. The subjects were asked to perform different tasks during a number of telephone conference calls, to answer a questionnaire after each conference call, and to answer a final questionnaire at the end of the experiment. Each task was aiming at a different level of involvement, while the test conduction was specifically designed to maintain an optimal comparability between those tasks. To compare any task effects on quality judgments with the influence of technical system conditions, we applied a balanced design (4th order Greco-Latin square, 5 times repeated) to mix tasks, system conditions and stimulus order.

2.1. Tasks for different levels of involvement

Task 1 (T1) "passive listening": The subject is just listening to recorded telephone conferences. After listening, the subject fills in a quality assessment questionnaire.

Task 2 (T2) "listening with writing minutes": While listening, the subject is asked to write down minutes of the conversation. After listening, the subject fills in a quality assessment questionnaire and a questionnaire about the minutes writing task.

Task 3 (T3) "conversation according to sequential script": The subject participates in a conversation with two other interlocutors using a script. The script triggers a fixed sequence of contributions from the interlocutors by stating who will add what type of information in which order. However, the script gives freedom concerning the exact wording, as it does not contain fully formulated sentences.

Task 4 (T4) "conversation according to scenario description": The subject participates in a conversation with two other interlocutors using a script that provides information in the form of bullet points, tables and pictographs. It allows thus more freedom in the order of contribution and the interaction between people as the T3 script does. However, some underlying structure is given by the way information is distributed among the participants (e.g. one has a question and the others have information concerning that question). In fact, these scenarios are shortened versions of the multiparty test scenarios used in [4].

2.2. Optimal comparability

While the tasks should invoke different levels of involvement, they also needed to be as comparable as possible. For that purpose, a number of measures minimized the variation between the different tasks: First, the T4 scenarios had been designed and tested to provide the same underlying conversational structure [4] while the actual content differs between those scenarios[1]. By removing the open discussion part from the original scenarios, we further increased the comparability. The T3 scripts were based on those same scenarios, and the recordings used for T1 and T2 were made accordingly to the script of T3. Second, in all stimuli, two of the three interlocutors were always the same speakers: The first two authors acted as interlocutors both in the conversation tasks (T3 & T4) and in the recordings for the listening tasks (T1 & T2). Third, the roles of the first two authors in the scenario was always the same (one having a question, one a solution); the third role (having a constraint) was either assumed by the third speaker in case of the listening tasks or by the test subject in case of the conversation tasks. Fourth, by means of exercising beforehand, the first two authors aimed for the same interaction behavior throughout all scenarios and test sessions. Fifth, the first two authors also developed beforehand strategies (e.g. waiting, directly addressing, interrupting) to get more "passive, hesitant" test subjects involved in the conversation in the same way than more "active, talkative" test subjects.

2.3. System conditions

We decided to limit the technical conditions to one type of degradation, but having different levels of that degradation. The disadvantage of this decision is that any results of this study can not be easily generalized for other degra-

[1] Subjects should not have exactly the same content in all tasks in order to avoid any effects due to such repetitions (e.g. learning, annoyance).

dations. One advantage, however, is that we do not introduce additional experimental variables in terms of the perceptual and conversational nature of different degradation types, i.e. different degradation types can lead to different perceptual dimensions [12] and they can lead to different conversational behavior (e.g. effect of speech signal distortions vs. effect of echo or delay). A second advantage is that we can better quantify any effects, as we have multiple data points for that degradation available, without the need to extend the experimental effort for subjects.

Our choice fell on packet loss, as it is the most prominent degradation in today's telecommunication systems. By testing different packet loss rates with our test system, we aimed to cover a rather broad quality range from imperceptible to perceptible and very annoying, while limiting the upper loss rate in order to avoid negative effects on the conversation flow. Eventually, we opted for random packet loss at 0, 5, 10 and 15 %. Note that we deliberately did not use more realistic burst-like packet loss behavior in order to avoid any temporal interaction between bursts of lost packets and possible important parts of any information, e.g. to avoid that whole digits of a telephone number are disturbed in one session but not in another. Furthermore, the packet loss was introduced in receiving direction of test participants; thus test participants heard all other interlocutors with the same degradation.

The test system comprised a central conference bridge using Asterisk and off-the-shelf VoIP telephones (SNOM870), which were connected in a local network (all via one router). The used codec was G.711 A-law and the packetloss simulation was realized with the TC filter and Netem software packages. Note that the telephones apply as packetloss concealment the G.711 Appendix I standard, which is known to show a high quality robustness against packet loss ([13]), explaining why we could go up to 15% loss rate without having extreme distortions.

2.4. Questionnaires

According to the study goals the target variables we wanted to measure were quality, involvement, and cognitive load. While there are indirect measurement methods known for cognitive load, emotional engagement (reflecting involvement) or task influence, e.g. task completion performance & time or physiological measurements, we opted for using questionnaires. The reason was that we were interested if there are any effects in the context of a typical quality assessment experiment, for which questionnaires about quality are widely accepted. Furthermore, own experience in comparing self-reported measures of cognitive load with an indirect measurement of cognitive load by means of a memory test showed high congruence between the different measurement approaches [10].

There were in total four questionnaires A to D. Most questions had to be answered on a 7-point continuous scale, some questions had to be answered with yes or no, and some questions asked for free text.

Questionnaire A was given to subjects after each call and comprised eight questions: QA1) involvement in conference call; QA2) quality of connection; QA31) own difficulties in speaking or understanding (yes/no); QA32) if the other interlocutors had difficulties in speaking or understanding (yes/no); QA32perc) how sure are you in judging question QA32 (in %); QA33) effort to follow the conference; QA41) difficulty in judging quality; and QA42) influence of conversation on judgment. With these questions we intended to measure a number of different aspects in order to cover the specific research questions and side questions (Sec. 1). While QA1 asked directly for involvement, QA33, QA41 , and QA42 are essentially – even though indirectly – asking for those factors that are potentially influenced by the involvement (Sec. 1), i.e. cognitive load (QA33) and task impact on quality rating (QA41 & QA42). QA2 is in fact the standard conversation/listening opinion scale according to [7], except that we used a 7-point continuous scale instead of a 5-point discrete ACR scale; QA31, QA32, and QA32perc are inspired by the conversation effort scale proposed in [7].

Questionnaire B was given to subjects only after each T2 call ("listening with writing minutes") and after they answered Questionnaire A. This questionnaire was used to get more insights on the potential difference between T1 ("passive listening") and T2. It comprised five questions about the minutes writing task: QB1) have you found all important information (yes/no); QB1perc) how sure are you in judging QB1; QB21) difficulties in writing the minutes (yes/no); QB22) reasons if QB21 was answered with yes; QB3) influence of minutes writing on answers in Questionnaire A.

Questionnaire C essentially resembled the writing minutes task and subjects answered it during every T2 call: it comprised a number of specific questions about the actual content of the call, guiding subjects in their minutes writing task. We did not further analyze the answers of this questionnaire, e.g. if they found all wanted information; but we checked, if they have done the task at all and took a short look at the notes.

Questionnaire D was given to subjects at the end of the experiment. It repeated question QA1 for all four tasks, i.e. it asked for the degree of involvement for the individual tasks in retrospective, after the subjects had experienced all four tasks. By assigning these questions to the individual experimental stimuli (task & packet loss rate), we generated one question QD that can be analyzed in the same way as its counterpart QA1.

Table 1: Repeated measures ANOVA and Sidak Posthoc test ($x\%$ vs. $y\%$) results for the different questions of Questionnaires A and D. The p-values are shown and significant differences are denoted with an * for a significance threshold of $p < 0.05$. Greenhouse-Geisser corrections are used for ANOVAs if sphericity assumptions are violated; Posthoc test results are only considered if the corresponding ANOVAs found significant differences [14].

	QA1	QA2	QA31	QA32	QA32perc	QA33	QA41	QA42	QD
Packet loss									
ANOVA	0.293	0.011*	0.026*	0.101	0.112	0.024*	0.681	0.024*	0.012*
0% vs. 5%	—	0.998	0.585	—	—	0.692	—	0.666	0.315
0% vs. 10%	—	0.236	0.176	—	—	0.006*	—	0.397	0.760
0% vs. 15%	—	0.085	0.003*	—	—	0.023*	—	0.086	0.128
5% vs. 10%	—	0.763	0.982	—	—	1.000	—	1.000	0.072
5% vs. 15%	—	0.049*	0.740	—	—	0.168	—	0.377	0.953
10% vs. 15%	—	0.605	0.982	—	—	0.352	—	0.358	0.036*
Task									
ANOVA	0.000*	0.000*	0.004*	0.248	0.697	0.001*	0.214	0.228	0.000*
T1 vs. T2	0.513	0.310	0.989	—	—	0.139	—	—	0.385
T1 vs. T3	0.001*	0.001*	0.038*	—	—	0.107	—	—	0.000*
T1 vs. T4	0.001*	0.000*	0.520	—	—	0.142	—	—	0.000*
T2 vs. T3	0.001*	0.003*	0.009*	—	—	0.010*	—	—	0.001*
T2 vs. T4	0.001*	0.017*	0.083	—	—	0.011*	—	—	0.001*
T3 vs. T4	0.364	1.000	0.964	—	—	0.990	—	—	0.999

3. Analysis and results

To identify significant impacts of tasks and system quality, we computed repeated-measures ANOVAs and Sidak PostHoc tests [14] for every question and visualized the directions of effects, if any, by means of errorbar plots (mean and 95% confidence interval).

Table 1 shows the ANOVA and Posthoc tests for the Questionnaire A and D, Figure 1 the corresponding errorbar plots. Note that for the Questionnaire B we did not find any significant differences nor did we find any trends from which we could obtain any insights; hence we skip the detailed results of that questionnaire.

Concerning the involvement question QA1, significant differences were between listening and conversation tasks (T1 & T2 vs. T3 & T4), but there were no significant differences between the two listening tasks T1 & T2 or between the two conversation tasks T3 & T4. In contrast to our intensions, the writing minutes task T2 did not increase the feeling of participants of being part of the call compared to just listening to the call (T1), and the more structured conversation task (T3) did not decrease that feeling compared to the more open conversation task (T4). As expected, there was no significant influence of the packet loss rate.

Concerning the quality question QA2, decreasing system quality (higher packet loss rate) resulted in decreasing quality ratings, although differences were only significant between 0% and 15%; and the mean scores varied between 3.600 for 0% and 2.625 for 15%. In contrast to our intensions, the selection of the packet loss rates was too conservative, as it did not cover the whole quality range. Interestingly, the quality ratings were also influenced by the tasks, as we found significant differences between the listening tasks (T1 & T2) and the conversation tasks (T3 & T4), with mean scores ranging from 2.225 for T1 to 3.905 for T4. Hence subjects were more critical in the listening tasks, and the difference between the tasks was 1.680 points on the used 7-point scale and thus substantially larger than the difference between the 0% and 15% packetloss, which was 0.975 on the 7-point scale.

Question QA31 shows an increasing difficulty to communicate with increasing packet loss rate, though the differences are significant only between 5% and 15%. In addition, QA31 shows a similar task dependency than Q2, grouping listening and conversation tasks. QA32 and QA32perc do not show any significant differences, neither for packet loss rate nor for the tasks. Apparently, the perception of having difficulties in conversing/listening shows a task and condition dependency that is rather consistent with the quality judgments.

The cognitive load question QA33 shows only a task dependency (significant difference T2 vs. T3, trend of grouping listening and conversation tasks), but no influence of the packet loss rate. Thus, self-reported cognitive load appears to be more dependent on the tasks than on the tested technical conditions.

The effort question QA33 showed the similar general difference between the listening and conversation tasks. However, the direction of that effect was unexpected. The ratings for the conversation tasks were more positive, meaning less effort, than for the listening tasks. We assumed that the conversation tasks require more concentration than listening only tasks, but apparently participants can follow a conversation easier if they take ac-

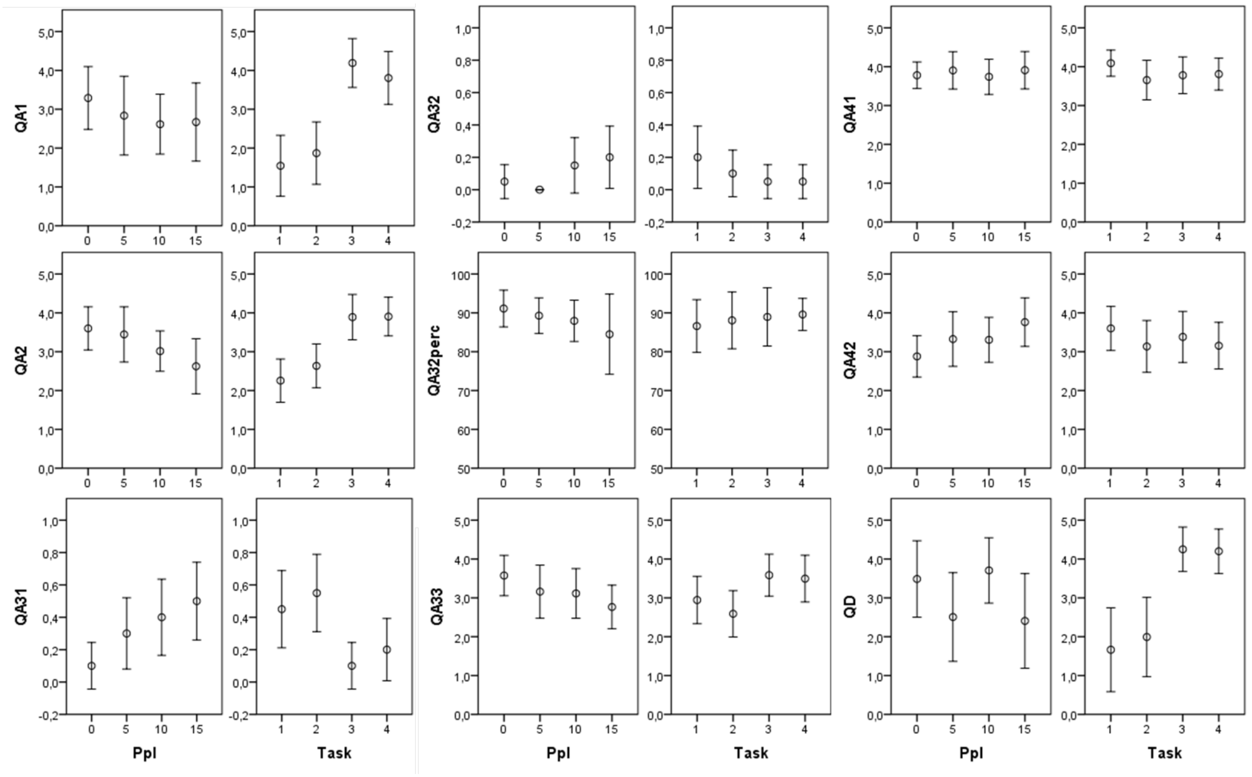

Figure 1: Errorbar plots for the questions of Questionnaires A and D. Shown are the mean values and 95% confidence intervals for the different packet loss rates Ppl and the four tasks.

tively part in it compared to if they are just listening to it. Concerning packet loss, QA33 ratings decrease (means higher effort) with increasing packet loss rate, with significant differences again only between 5% and 15%.

Comparing the results between involvement QA1, quality QA2 and cognitive load QA33, we see that all questions show a similar task dependency, while quality and cognitive load also show a dependency of packet loss rate. To obtain quantitative measures of the similarity in those three variables, we computed the Pearson correlation coefficients and conducted a repeated-measures ANOVAs with QA1, QA2, and QA33 as input variables. The repeated measures ANOVA did not reveal a significant difference between those three questions, and the correlation coefficients ρ have medium values ($\rho(QA1, QA2) = 0.43$, $\rho(QA1, QA33) = 0.36$, $\rho(QA2, QA33) = 0.32$). This means there is some moderate similarity of the questions (correlations), and the differences between those questions are - at least for the present data - not significant (ANOVA).

Continuing with questions QA41 and QA42, no significant differences can be observed (QA41) or can be localized (QA42, ANOVA: significant, Posthoc: no pair significant). In fact the ratings were almost constant, suggesting that these questions are not informative.

Looking at the results for QD, the "retrospective" involvement question, we see the same task dependency as for QA1, the "immediate" involvement question. Furthermore, we now also see a dependency of the technical condition, which is, however, not systematic: there appears to be the ranking 0%, then 10%, then 5%, and then 15%. We also conducted a paired-samples t-tests with QA1 and QD as input variables and checked the Pearson correlation coefficient. There were no significant differences between QA1 and QD and correlation was with a value of $\rho = 0.75$ rather large, though not close to one. Thus the task dependency does not differ between the "immediate" and the "retrospective" involvement question, except for some unexplained peculiarity concerning the influence of the 10% packet loss rate on the retrospective answers.

4. Discussion

After analyzing the collected data, we can now link the results to the detailed research questions.

1. How large is the difference in quality ratings when test subjects perform assessment tasks with different levels of involvement; what would be the consequences of such differences for selecting an appropriate assessment method?

Answer: In the present data, the difference of quality ratings (QA2) between the tasks was up to 1.680 points on the used 7-point scale; a value which we conclude to be of substantial size. Especially the fact that this difference is larger than the difference of quality ratings due to the technical conditions (0.975 on the 7-point scale), emphasizes the role of the task when it comes to assessment tasks with medium or small differences between technical conditions. Concerning the choice of an appropriate assessment method, this result confirms the widely used approaches for two-party settings: Use listening-only tests, if a high sensitivity of subjects is required, use conversational tests, if sensitivity is not an issue but other aspects such as naturalness. However, this results also clearly show that a comparison between listening-only tests and conversational tests are hardly possible or at least can only be interpreted with great care.

2. To what extend are test subjects able to differentiate between quality, cognitive load and involvement; what would be the consequences for designing appropriate assessment questionnaires?

Answer: It seems that subjects are able to differentiate between these aspects, but apparently they can not be perfectly separated as we observed moderate correlations. Concerning questionnaire design, this means on the one hand that none of these three questions should be easily omitted. On the other hand, this also means that one would collect - at least partially - redundant data.

3. In how far is it possible to design test tasks such that arbitrary levels of involvement can be achieved, i.e. are the tasks equidistantly distributed across the involvement scale?

Answer: The data showed that there is a large step between the listening tasks and the conversation tasks. Our attempts to generate a listening tasks triggering more involvement by giving an active task while listening (T2) and to generate a conversation task triggering less involvement by sequentially fixing the order of contributions (T3) were not strong enough against the apparently fundamental difference between listening and conversation tasks. It is likely that also other attempts to better distribute tasks across the involvement scale will fail, hence this research question must be answered with a ”No” for the time being.

4. Is there a difference between assessing involvement directly after a conference call and in retrospective after the whole experiment?

Answer: The comparison between QA1 and QD shows that there is hardly any difference, but there is the possibility to obtain noisy data (see the unsystematic packet loss dependency), if this question is asked in retrospective. Hence, asking “immediately” for involvement appears to be the better option.

5. Conclusion

This study showed for the multiparty scenario that the involvement of subjects in the experimental task and conversation has a significant impact on quality ratings that can be even larger than the tested technical conditions. Conceptually the results confirm the experience from two-party scenarios, but they also give a quantitative idea of the deviation that can be expected between a multiparty conversation or listening-only test.

6. Acknowledgements

The authors would like to thank Katrin Schönenberg and Judith Köppe for their valuable support during the initial preparation phase of this study.

7. References

[1] ITU-T, “Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual multiparty telemeetings”, International Telecommunication Union, 2012.

[2] Berndtsson G, Folkesson M, Kulyk V, “Subjective quality assessment of video conferences and telemeetings”, 19th International Packet Video Workshop, 2012.

[3] Hoeldtke K, Raake A, “Conversation analysis of multiparty conferencing and its relation to perceived quality”, IEEE International Conference on Communications (ICC), doi: 10.1109/icc.2011.5963021, 2011.

[4] Raake, A., Schlegel, C., Hoeldtke, K., Geier, M., Ahrens, J., “Listening and Converstional Quality of Spatial Audio Conferencing”, 129th AES Convention, Tokyo, 2010.

[5] Gros, L., Gontran, F., “A study on tasks for assessing audiovisual quality of videoconferencing systems in multipoint conversation tests”, Contribution C259, ITU-T Study Group 12 Meeting, International Telecommunication Union, October 2011.

[6] Skowronek, J., Schoenenberg, K., Berndtsson, G., Folkesson, M., Raake, A., “Initial insights for P.AMT based on reviewing existing recommendations”, Contribution C284, ITU-T Study Group 12 Meeting, International Telecommunication Union, October 2011.

[7] ITU-T, “Recommendation P.800 - Methods for subjective determination of transmission quality.”, International Telecommunication Union, 1996.

[8] ITU-T, “Recommendation P.800.1 - Mean Opinion Score (MOS) terminology.”, International Telecommunication Union, 2006.

[9] ITU-T, “Recommendation P.800.2 - Mean Opinion Score (MOS) interpretation and reporting.”, International Telecommunication Union, 2013.

[10] Skowronek, J., Raake, A., “Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing”, Interspeech 2011, 829-832, Florence, 2011.

[11] Schnotz, W., Kuerschner, C., “A Reconsideration of Cognitive Load Theory”, Educational Psychology Review, Vol. 19, No. 4, 469-508, 2007.

[12] Wältermann, M., “Dimension-based Quality modeling of Transmitted Speech”, T-Labs Series in Telecommunication Services, Springer, 2013.

[13] ITU-T, “Recommendation G.113 - Transmission impairments due to speech processing”, International Telecommunication Union, 2007.

[14] Field, A., “Discovering statistics using SPSS”, 3rd Edition, SAGE publications, 2009.