Quality of Multiparty Telephone Conferences from the Perspective of a Passive Listener

Janto Skowronek, Anne Weigel, Alexander Raake
Assessment of IP-based Applications, Telekom Innovation Laboratories, Technische Universtät Berlin

Introduction

This paper concerns the quality perception of multiparty telephone conferences. Survey studies reported that participants of multiparty telephone conferences often report dissatisfaction[1, 2]. Apparently, telephone conferences do not provide a quality that customers are used to from conventional two-party telephony. The International Telecommunication Union (ITU) has acknowledged two major differentiators between two-party telephony and multiparty telephone conferences [3]: one is having a group conversation, which is a special communicative situation; the other is the possibility of having asymmetric conditions, that is individual connections provide different qualities.

To take both aspects into account, our previous work [4] has focused on the quality assessment of asymmetric conditions by employing conversation tests and by asking for quality ratings concerning the overall conference call and concerning the individual connections of participants. In other words, test subjects rated the quality from the perspective of an active participant. In real-life, however, it is often the case that only a few persons are actively contributing while a number of participants are just listening to the conversation. Conversation tests are less suited for such a use case if they are not specifically designed to put participants into such a passive listener role. Furthermore, it is known in the field that test subjects are less sensitive to quality impairments in conversation tests than in listening-only tests [5, 6], which was also shown for multiparty scenarios [7].

For that reason the present study applied a listening-only test to mimic the perspective of the passive - and therefore - more critical listener. Thus the research goal is to learn more about the quality perception of a multiparty telephone conference from the perspective of a passive listener. Addressing the aspect of asymmetric conditions in particular, the specific research question is to investigate if there is a simple relation between the perceived quality of individual connections and the overall conference call, whereas the study mimics a three-party scenario (conversation between two active speakers, test subject is the third passive listening participant).

Hypotheses

Based on a technical analysis according to [4], it is possible to translate the research question on the quality relation between individual connections and overall call into a set of specific hypotheses. To do this, the following definition of variables will be used:

- Q_{ic}: Overall conference call quality from the perspective of interlocutor i.
- Q_{ij}: Individual connection quality of interlocutor j from the perspective of interlocutor i.
  Noting that in case of asymmetric conditions, individual connections might be impaired or not impaired, the following notation will help to distinguish these two cases:
  - $Q_{ij,0}$: individual connection not impaired
  - $Q_{ij,x}$: individual connection impaired

With these variables the following hypotheses are now formulated, bearing in mind that there are two active speakers, thus two individual connections can be judged.

Hypothesis 0:

In the reference condition, i.e. the technically best condition in the test, all individual connections are perceived as unimpaired. That means, Q_{ic} in that condition defines the highest quality rating for the conference call quality, $Q_{ij,0}$ in that condition defines the highest quality rating for the individual connections.

Hypothesis 1:

The relation between the conference call quality scores and the individual connection quality scores is a simple average. This can be translated into three cases, depending on the condition:

- H1a: Reference condition (if H0 holds): All individual connections are per definition unimpaired and equal, thus $Q_{ic} = Q_{ij,0}$
- H1b: Asymmetric conditions (one connection unimpaired, one connection impaired): $Q_{ic} = (Q_{ij,0} + Q_{ij,x})/2$
- H1c: Symmetric conditions (both connections with same impairment): $Q_{ic} = Q_{ij,x}$

Hypothesis 2:

There is no mutual influence of the individual connections. This can be checked by performing two comparisons:

- H2a: For all asymmetric conditions, the score for $Q_{ij,0}$ is equal to that of the reference condition.
- H2b: $Q_{ij,x}$ of an asymmetric condition = $Q_{ij,x}$ of the corresponding symmetric condition.

Tabelle 1: Description of test conditions.

Condition	Parameters
RefWB	Both speakers with G.722 wideband codec, considered as reference condition
PL-WB	One speaker with 5% random packet loss, G.722 without Packet Loss Concealment (PLC)
NB	One speaker with G.711 narrowband codec
NBsym	Both speakers with G.711
PL-NB	One speaker with 5% random packet loss, both speakers with G.711 without PLC
LE	Listener echo for one speaker, i.e. hear this voice twice (echo level 8% or speech signal, echo delay 245ms)
LEsym	Listener echo for both speakers
REV	Reverb (Small room, Pre-delay 33ms, Decay 82ms)
SCN	Speech correlated noise (pink noise -20dB, gate parameters: threshold -47.1 dB, attack: 33.4ms, hold: 1200ms, release: 799.8ms)

Experimental Study

To validate the hypotheses, we analyzed a listening-only test, which is described in more detail in [8]. Twenty-four subjects (12 female, 12 male, age: 19 – 52 years, on average 31 years) participated in the test. As stimuli we used 32 different 40s-long excerpts from recorded conversations of a (yet unpublished) conversation test study, whereas one speaker (a confederate in that conversation test) is present in all excerpts used here. By means of signal processing using the ProTools software and an ITU toolbox [9], we realized nine technical conditions, whereas each speaker could be processed separately to realize symmetric and asymmetric conditions. Table 1 provides a description of the technical parameters. Per asymmetric condition, each technical impairment was applied four times, twice on the confederate (the one speaker present in all recordings), twice on the other speakers. Per symmetric condition, each technical impairment was applied twice, since the confederate and the other speakers are affected simultaneously. The reference condition was applied four times in order to have roughly 10% of the stimuli covered with the reference condition. The test design was a within-subject design with randomized presentation order of stimuli across subjects.

The subjects rated both the conference call quality and the quality of the two individual connections (i.e. speakers). In order to trigger test subjects to judge each of the two different levels of quality (call vs. connection) as conceptually two separate items, we decided to use two different scales. The conference call was rated with a 5-point absolute category rating (ACR) scale according to [10]; the individual connections were rated with an extended and continuous scale according to [11]. This should on the one hand avoid that subjects simply give to all questions the same rating, on the other hand it should avoid that subjects first rate the individual connections and then form a "visual" mean of those ratings ("visual" in terms of the position of the marks on the paper questionnaire). To enable a proper analysis of results, a transformation to the ACR-scale according to [12, 13] was applied to the collected data.

Results

The two panels of Figure 1 show the errorbar plots (mean and 95% confidence interval) for the nine conditions, whereas the reference condition is repeated in each panel for better visual comparison.

Furthermore the Figure shows – based on the hypotheses – also the expected relation between the ratings of Q_{ic}, $Q_{ij,0}$, and $Q_{ij,x}$, which are described first before the actual results are presented. Both, the dashed and dotted lines of Figure 1 use the actual ratings of the impaired connections $Q_{ij,x}$.

Then, the dotted lines show the expected values of Q_{ic} and $Q_{ij,0}$ if both hypotheses H1 and H2 hold. The lines reflect the assumption that each technically unimpaired connection $Q_{ij,0}$ is rated equally as the quality scores of the reference condition (H2), and they reflect that the conference call quality Q_{ic} is the arithmetic mean of the individual connection quality scores $Q_{ij,0}$ and $Q_{ij,x}$ (H1). A special case is the condition NB-PL, in which the "unimpaired" connection $Q_{ij,0}$ is actually a narrowband (NB) connection and not the wideband reference (RefWB). Thus, the expected value is actually that of the impaired connection of the NB connection ($Q_{ij,x}$).

The dashed lines show the expected values of Q_{ic} if only H1 holds. The lines assume that Q_{ic} is the arithmetic mean of $Q_{ij,0}$ and $Q_{ij,x}$ (H1), but they do not require any theoretical expectations concerning $Q_{ij,x}$ as they use the actual ratings of $Q_{ij,x}$.

Now, with the visualization of the expected values, the hypotheses can be validated. H0 is confirmed because Q_{ic} and $Q_{ij,0}$ of the reference condition RefWB reached the maximum quality scores in the test. H1 is only confirmed for the conditions RefWB (i.e. H1a), NB (i.e. H1b), NBsym (i.e. H1c), LEsym (i.e. H1c), as the dotted line lies inside the confidence intervals for Q_{ic}. For the other conditions PL-WB, PL-NB, LE, REV, and SCN, the hypothesis H1 (i.e. H1b) is not confirmed, as the dotted line lies outside the confidence intervals for Q_{ic}. H2a is only confirmed for the conditions NB and PL-NB, as the dotted line lies inside the confidence intervals for $Q_{ij,0}$. For the other conditions PL-WB, LE, REV, and SCN, H2a is not confirmed, as the dotted line

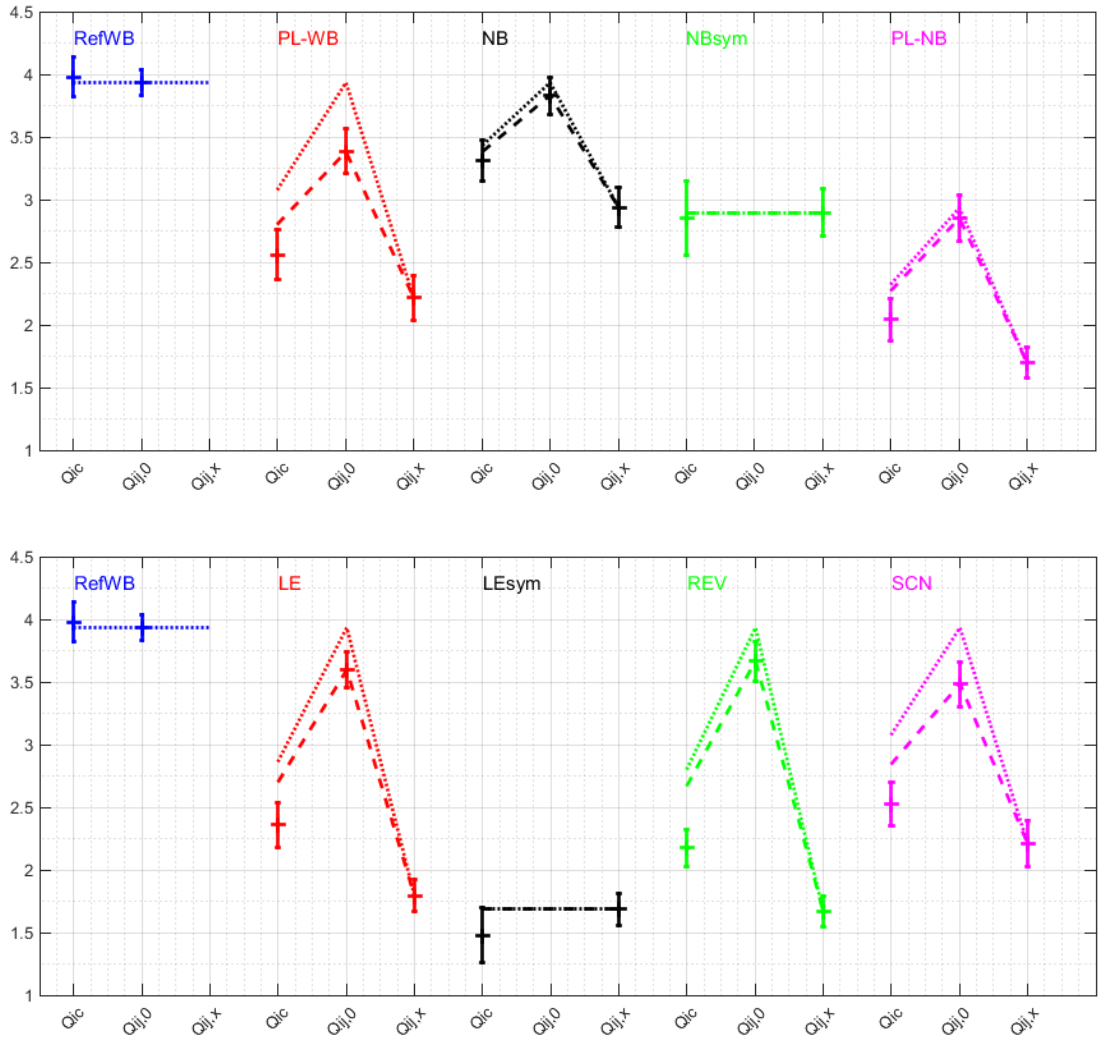

Abbildung 1: Test results for the nine conditions shown in two panels; the results of the reference condition RefWB are repeated for better visual comparison. The two panels show errorbar plots (mean and 95% confidence intervals of the obtained quality ratings for the conference call quality score Q_{ic}, the individual connection quality score of an unimpaired connection $Q_{ij,0}$, and the individual connection quality score of an impaired connection $Q_{ij,x}$. Furthermore the dotted line shows the expected values for Q_{ic} and $Q_{ij,0}$ in case that hypotheses H1 and H2 are confirmed, i.e. Q_{ic} is a simple mean of $Q_{ij,0}$ and $Q_{ij,x}$ and there is no mutual influence of the individual connections. The dashed line shows the expected values for Q_{ic} in case that only hypothesis H1 is confirmed.

lies outside the confidence intervals for $Q_{ij,0}$. H2b is confirmed, as there are no significant differences (ANOVA with PostHoc tests) between NB & NBsym and LE & LEsym. Given that H2a is not confirmed for all conditions, H1 needs to be re-evaluated by looking at the dashed lines, which, in contrast to the dotted line, do not require a confirmed hypothesis H2. However, even with the correction for H2, the results for H1 do not change: for each condition, the dashed line lies either inside or outside the confidence intervals for Q_{ic}, as does the dotted line.

Discussion

The confirmation of H0 proves a successful experimental manipulation, that is, the condition with the best technical quality in the test was also rated as the best condition. The fact the hypotheses H1 and H2 are confirmed for some conditions but are rejected for other conditions shows that there is no simple relation between the individual connection quality scores and the conference call quality score.

Looking at the results of H1 more closely, symmetric conditions, i.e. both individual connection have the same impairment, as well as the reference condition follow the expectations. In case of asymmetric conditions, however, the picture is not clear: one asymmetric condition follows the expectations (NB), while the others do not. Nevertheless, in those cases not following the expectations, the real conference call quality is systematically lower then the expected mean. That means, there appears to be a stronger influence of the worst connection in the overall quality judgment.

Looking at the results of H2 more closely, in case of the

asymmetric conditions, the picture is not clear: two of five asymmetric conditions follow the expectations (NB and PL-NB), while the three other conditions do not follow the expectations (LE, REV, and SCN). It was not possible to find a clear reason why some technical conditions show the mutual influence and others not. We can only hypothesize that the bandwidth might play a role, as the mutual influence is not visible for the narrowband impairments; or the narrowband impairments are special in the sense that narrowband telephony is that what people are still most used to in real life; or the three other condition are special in the sense that they might require more listening effort or the like, even though we designed the stimuli such that speech intelligibility was not affected.

Another aspect to check is the question whether the mutual influence of the individual connections is linked with the fact that the arithmetic mean of individual connection quality scores is not reflecting the conference call quality. If this would be the case, then only in those cases in which the mutual influence is visible (H2 rejected), the conference call quality deviated from the mean (H1 rejected), while in those cases in which no mutual influence is visible (H2 confirmed), the conference call quality does not deviate from the mean (H1 confirmed). In almost all cases, this link between H1 and H2 – either both confirmed or both rejected – is indeed visible. However, the condition PL-NB violates this, since H2 is confirmed but not H1.

Conclusions

To answer the research question, the results show that there is not a simple relation such as a mean operation between the conference call quality Q_{ic} and the individual connection quality Q_{ij}, because the relation apparently depends on the actual condition.

For that reason, the next steps are to investigate other functions to better explain this relation. Furthermore, a future comparison with conversation test results would allow to verify whether this relation would differ between passive listeners and active participants. This would also provide more insights on the impact of the communicative situation on multiparty quality perception.

Literatur

[1] Yankelovich, N., Walker, W., Roberts, P., Wessler, M., Kaplan, J., Provino, J., "Meeting Central: Making Distributed Meetings More Effective", Proceedings of CSCW 2004, pp 419-442, ACM Press, 2004.

[2] Skowronek, J., "Internetumfrage zur Wahrnehmung heutiger Telefonkonferenzen im geschäftlichen Umfeld", In: Fortschritte der Akustik (DAGA2012) - 38. Deutsche Jahrestagung für Akustik, pp 899-900, Deutsche Gesellschaft für Akustik, 2012.

[3] ITU-T, "Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual telemeetings", International Telecommunication Union, 2012.

[4] Skowronek, J., Herlinghaus, J., Raake, A., "Quality Assessment of Asymmetric Multiparty Telephone Conferences: A Systematic Method from Technical Degradations to Perceived Impairments", Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), pp 2604 - 2608, International Speech Communication Association, 2013.

[5] Möller, S., "Assessment and Prediction of Speech Quality in Telecommunications", Kluwer Academic Publishers, 2000.

[6] ITU-T, "Recommendation P.805 - Subjective evaluation of conversational quality", International Telecommunication Union, 2007.

[7] Skowronek, J., Schiffner, F., Raake, A., "On the influence of involvement on the quality of multiparty conferencing", 4th International Workshop on Perceptual Quality of Systems (PQS 2013), pp 141-146, International Speech Communication Association, 2013.

[8] Weigel, A., "Beurteilung der Qualität von Telefonkonferenzen mit asymmetrischen Verbindungen aus der Perspektive eines passiven Zuhörers", Bachelorthesis, Technische Universität Berlin, 2014.

[9] ITU-T, "Recommendation G.191 - Software tools for speech and audio coding standardization", International Telecommunication Union, 2010.

[10] ITU-T, "Recommendation P.800 - Methods for objective and subjective assessment of quality", International Telecommunication Union, 1996.

[11] ITU-T, "Recommendation P.851 - Subjective quality evaluation of telephone services based on spoken dialogue systems", International Telecommunication Union, 2003.

[12] Köstermann, F., Guse, D., Wältermann, M., Möller, S., "Comparison between the Discrete ACR Scale and an Extended Continuous Scale for the Quality Assessment of Transmitted Speech", To appear in: Fortschritte der Akustik (DAGA2015) - 41. Jahrestagung für Akustik, Deutsche Gesellschaft für Akustik, 2015.

[13] Wältermann, M., Raake, A., Möller, S., "Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech", ITU-T Contribution COM 12 – C 39 – E, International Telecommunication Union, Geneva, February 2009.