Quality Assessment of Asymmetric Multiparty Telephone Conferences: a Systematic Method from Technical Degradations to Perceived Impairments

Janto Skowronek, Julian Herlinghaus, Alexander Raake
Assessment of IP-based Applications - Telekom Innovation Laboratories
Berlin University of Technology, Germany
janto.skowronek@telekom.de

Abstract
The present paper addresses multiparty telephone conferences with asymmetric quality degradations. We propose a systematic method that allows to investigate how individual technical degradations can lead to the perception of quality impairments by the different interlocutors in a conference call. By conducting this analysis for a number of degradations, a detailed picture on the complexity of assessing asymmetric conditions is drawn, which in turn verifies the need for such strictly systematic assessment approaches.

Index Terms: multiparty conferencing, quality, asymmetry

1. Introduction
Users often report dissatisfaction with multiparty telephone conference calls [1]. Although standardized quality assessment methods (e.g. [2]) and prediction models (e.g. [3]) exist for one-to-one telephone connections, and although first subjective test methods are standardized for multiparty situations [4], more knowledge is needed on the quality of experience (QoE) of multiparty conferences in order to improve existing systems.

While communication-related aspects were investigated in the past, e.g. [5, 6, 7], the present paper focuses on the technical aspects that are relevant for multiparty conferences. Given the vast amount of possible system architectures and potential impairments, we focus in this work on a SIP-based (Session Initiation Protocol) conference bridge in a Voice-over-Internet-Protocol (VoIP) context. Furthermore, we here assume that degradations occur on the sending side of only one participant; hence we consider multiparty connections with asymmetric degradations.

Towards the goal of better understanding multiparty QoE, the research questions are: How does a technical degradation on one side translate into perceptible impairments for each participant? Are these impairments indeed perceived by non-expert subjects? In the remainder of this paper, a structured technical analysis is applied to answer the first question (Sec. 2), followed by a subjective experiment (Sec. 3) and corresponding statistical analysis to answer the second question (Sec. 4).

2. Structured technical analysis
Asymmetric degradations in a multiparty telephone conferences can lead to very different quality perceptions for each interlocutor. These individual impairments are dependent on a number of factors, such as the type and strength of degradation, the system topology, the number of interlocutors, and the locations in the different end-to-end paths where degradations occur. To decompose the complexity inherent in such an asymmetric conference, we propose a systematic four-step approach to infer potentially perceivable impairments from technical degradations.

Step 1: Description of multiparty situation
We here consider a conference connection between three interlocutors $ILx$ (with $x \in \{1, 2, 3\}$) over a central conferencing bridge. Note that the following analysis and discussion can easily be adapted for more than three interlocutors or for peer-to-peer scenarios. Figure 1 shows, from left to right, the transmission chain from each interlocator as sender to each interlocutor as receiver: every interlocutor is in an environment (dashed rectangle); he or she talks into a terminal device (telephone), which sends the signal via a connection (line) to the conference bridge (solid-line rectangle); the bridge mixes the signals from two interlocutors (mesh of lines) and sends this via a connection (line) to the third interlocutor; on the receive side, every interlocutor is still in the same environment (dashed rectangle) to the third interlocutor; on the receive side, every interlocutor is still in the same environment.

![Figure 1: Diagram showing a) the connections of a three-party conference using a central conference bridge with every interlocutor shown as sender (s) and receiver(r), and b) indication of potential degradations caused by the environment (Des & Der), by the terminal devices (Dds & Ddr), by the connections (Dcs & Dcr) and by signal processing that may be applied at the input (send-side) or output (receive-side) of the conference bridge (Dbx & Dbr). As the degradations can occur for every interlocutor, all variables are functions of interlocutor x, with x ∈ {1, 2, 3}.](image-url)
Table 1: Results of the structured technical analysis showing the translation of technical degradations into perceptible impairments from the perspective of each interlocutor. The degradations are described by the E-Model parameters ($Par$), see Sec. 2 Step 2, and occur at location $D$ defined in Fig. 1). $I_{x,y}$ notes the impairment from the perspective of interlocutor $x$ with regard to interlocutor $y$ ($x, y \in \{1, 2, 3\}$) and the integral view of interlocutor $x$ of the entire conference is referred to as $I_{x,C}$.

<table>
<thead>
<tr>
<th>Degradation ($Par$)</th>
<th>$I_{1,C}$</th>
<th>$I_{1,1}$</th>
<th>$I_{1,2}$</th>
<th>$I_{1,3}$</th>
<th>$I_{2,C}$</th>
<th>$I_{2,1}$</th>
<th>$I_{2,2}$</th>
<th>$I_{2,3}$</th>
<th>$I_{3,C}$</th>
<th>$I_{3,1}$</th>
<th>$I_{3,2}$</th>
<th>$I_{3,3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loudness (OLR)</td>
<td>$D_{ds1}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Loudness (OLR)</td>
<td>$D_{ds2}$</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bg. Noise ($Ps, Pr$)</td>
<td>$D_{es1}$</td>
<td>own</td>
<td>nob</td>
<td>0</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bg. Noise ($Ps, Pr$)</td>
<td>$D_{es2}$</td>
<td>one</td>
<td>0</td>
<td>nfe</td>
<td>0</td>
<td>own</td>
<td>0</td>
<td>0</td>
<td>nob</td>
<td>one</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bg. Noise ($Ps, Pr$)</td>
<td>$D_{es3}$</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>nfe</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>nfe</td>
<td>own</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Echo ($TELRI, Ta$)</td>
<td>$D_{es1}$</td>
<td>two</td>
<td>0</td>
<td>del</td>
<td>del</td>
<td>all</td>
<td>del</td>
<td>tec</td>
<td>all</td>
<td>del</td>
<td>tec</td>
<td>tec</td>
</tr>
<tr>
<td>Packet Loss ($Ppl$)</td>
<td>$D_{es1}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Packet Loss ($Ppl$)</td>
<td>$D_{es2}$</td>
<td>one</td>
<td>0</td>
<td>dist</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Packet Loss ($Ppl$)</td>
<td>$D_{es3}$</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>dist</td>
<td>one</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Coding of perception

- 0 = normal (no impairment)
- del = delayed interlocutor
- own = hear something concerning myself
- att = attenuated voice
- tec = hear own voice (talker echo)
- nob = noise in own background
- nfe = background noise from far end
- one = hear something concerning one interlocutor
- del = delayed interlocutor
- tec = hear own voice (talker echo)
- nfe = background noise from far end
- one = hear something concerning one interlocutor
- tec = hear own voice (talker echo)
- nfe = background noise from far end
- one = hear something concerning one interlocutor
- tec = hear own voice (talker echo)
- nfe = background noise from far end
- all = hear something concerning all

Step 2: Degradations to be considered and points in which they degrade speech signals

We here limit the degradations to loudness loss, background noise, echo, and packet loss, and we focus on degradations that occur at the send-side. Note that we define the degradations by the transmission parameters used in the E-Model [3] and that the background for the following argumentation stems from [8].

Loudness loss can be described by the parameter $OLR$ (Overall Loudness Rating). Although loudness loss can occur at various points in the transmission chain, we consider here a case in which it is caused by the terminal device in sending direction, hence $D_{ds}$.

Background noise is reflected by the parameters $Ps$ (send side room noise) and $Pr$ (receive side room noise). Since background noise can be transmitted to the other participants and since it can interfere with the signals a participant can hear from the others, it is both a send- and receive-side degradation, hence $D_{es}$. One

Echo is caused by a delayed feedback, usually stemming from the terminal device when counter measures such as echo cancelation are missing or insufficiently working. Echo is a send-side degradation $D_{es}$ since it affects the speech signal in the send side microphone: the participant’s voice is overlaid with the feedback signal from the loudspeaker. As echo requires also some delay, we here assume a one-way delay of 75 ms on the connection for both send and receive direction ($\pm$150 ms roundtrip delay). Hence, $D_{ds}$ is actually the level of the echo signal, described by the parameter $TELRI$ (Talker Echo Loudness Rating); the delays are connection degradations $D_{es}$ and $D_{es}$, corresponding to the parameter $Ta$ (Absolute one-way delay).

Packet loss is considered here as random packet loss that occurs on the send-side connection to the conference bridge, hence $D_{es}$. Note that packet loss can also occur on the receive-side and it can show different distributions (e.g. bursts), the corresponding parameter is $Ppl$ (Packet loss probability).

Step 3: Analysis of affected paths and assumed perception by conference participants

For this step, we assume that the degradation occurs for $IL1$ and infer what speech signals all participants would hear. This analysis can best be done by following the signal paths that a degradation takes in Figure 1. Note that the following discussions consider only the perceptions of $IL1$ and $IL2$, because $IL3$ would perceive the same situation as $IL2$.

$OLR$: the attenuated speech signal of $IL1$ is transmitted to $IL2$ and $IL3$, the other speech signals are at normal level: $IL1$ hears $IL2$ and $IL3$ at normal levels; $IL2$ hears $IL1$ at a low level and $IL3$ at a normal level.

$Ps, Pr$: $IL1$ hears the normal speech signals of $IL2$ and $IL3$ and in addition the background noise in his or her environment: $IL2$ hears the voice of $IL1$ plus the transmitted background noise of $IL1$ plus the speech signal from $IL3$ without additional background noise.

$TELRI$, $Ta$: the speech signal at the receive-side of $IL1$, which contains the mixed signals of $IL2$ and $IL3$, is fed back into the send-side of $IL1$. $IL2$ hears this signal in addition to the signals from $IL1$ and $IL3$ and $IL2$ hears the direct signal from $IL1$, his/her own voice (talker echo), the direct signal from $IL3$ and the feedback from $IL3$ (listener echo).

$Ppl$: the degraded speech signal of $IL1$ is transmitted to $IL2$ and $IL3$, the other signals are normal: $IL2$ hears degraded speech from $IL1$ and normal speech from $IL3$.

Step 4: Comprehensive view for all participants

By exchanging all interlocutors in Step 3, one can complement the picture for all participants and summarize it in a convenient format such as Table 1. Note that all considerations are referring to a certain reference condition reflecting a normal VoIP connection that is also noted in Table 1 bottom row. In terms of transmission parameters, this reference condition corresponds to the default parameter values defined in [3].
3. Experiment

To investigate how subjects actually perceive the impairments identified in Sec. 2, we conducted a three-party conversation test. The technical setup reflected the situation assumed in the previous section. It comprised off-the-shelf VoIP-telephones (SNOM870) connected via a local router to a conference bridge (Asterisk) running on a Linux laptop. Using an Asterisk plugin, the speech signals were routed via the Jack Audio Connection Kid (JACK) to an audio processing software (PureData). All connections used the G.711 A-law codec and the minimum one-way delay of this test system was 150 ms (reference condition). The background noise consisting of nearly stationary street noise (Ps = Pr = 65 dB SPL) was played back via loudspeakers placed in the test rooms (the room noise floor of 44 dB SPL could be neglected compared to the explicitly introduced background noise). Random packet loss (Ppl = 10%) was realized by applying the TC Filter and Netem softwares to manipulate the packet streams. All signal levels were calibrated with an head and torso simulator (HATS).

We invited 17 groups of 3 subjects (age 16-69, 34 female, 13 male, 28 with & 19 without multiparty conferencing experience); in 4 cases a colleague complemented the groups. The subjects held conference calls using shortened versions of the Three-party Conversation Test Scenarios described in [9]. The sessions comprised an informal call to get to know each other and to get used to the multiparty situation, a training call to practise the scenarios, and 10 calls with the test conditions. All connections used the G.711 A-law codec and the minimum one-way delay of this test system was 150 ms (reference condition). The background noise consisting of nearly stationary street noise (Ps = Pr = 65 dB SPL) was played back via loudspeakers placed in the test rooms (the room noise floor of 44 dB SPL could be neglected compared to the explicitly introduced background noise). Random packet loss (Ppl = 10%) was realized by applying the TC Filter and Netem softwares to manipulate the packet streams. All signal levels were calibrated with an head and torso simulator (HATS).

We invited 17 groups of 3 subjects (age 16-69, 34 female, 13 male, 28 with & 19 without multiparty conferencing experience); in 4 cases a colleague complemented the groups. The subjects held conference calls using shortened versions of the Three-party Conversation Test Scenarios described in [9]. The sessions comprised an informal call to get to know each other and to get used to the multiparty situation, a training call to practise the scenarios, and 10 calls with the test conditions.

The questionnaire was inspired by previous work [7] as well as standard assessment methods [2]. It comprised 5-point scales for 7 questions, presented in the following order:
- **OvQual**: “What was your personal intuitive overall impression of this teleconference?”
- **Conc**: “It required (very much... very little) concentration to follow the conference.”
- **ConQual**(...) : “The quality of the own connection was (bad... excellent).”
- **ConQual**(...) : “The quality of the connection of Mr./Ms. X was (bad... excellent).”
- **ConQual**(...) : “The quality of the connection of Mr./Ms. Y was (bad... excellent).”
- **ComEff** : “It required (very much... very little) effort to communicate with the other participants.”
- **Accept** : “I would find it (very unacceptable... very acceptable) to attend a telephone conference with such a system.”

The three ConQual questions reflect the technical quality that subjects perceived and are specific to individual connections from interlocutor x to y. Conc and ComEff address the communication-related aspects subjects experienced during the calls and OvQual and Accept address the overall quality.

4. Results: Perception of impairments

**Data analysis**

For every conversation, judgements of three test subjects were obtained, whereby individual subjects may have different perceptions of the same conversation, and the judgements concern the quality of both the overall conference and of individual connections. To deal with the complexity inherent in this data, we followed again a strictly systematic approach by grouping the data before running the intended statistical tests. By applying formal selection criteria based on the cases defined in Tab. 1, we obtained sets for judgments concerning the whole conference (here called “Level 1 Sets”) and sets of judgements concerning individual connections (“Level 2 Sets”), Tab. 2 shows the formal selection criteria and interpretations of these sets.

To test if the degradations actually influenced the judgments, we ran non-parametric Mann-Whitney tests (see e.g. [10]) between the reference condition and the individual degradations. Six comparisons between Level 1 Sets were possible for the measures OvQual, Conc, ConQual, and Accept; 17 comparisons between Level 2 Sets were possible for the measure ConQual.

**Results**

Loudness differences influenced the overall judgements (No. 1-4) and only those of individual connections, i.e. from one interlocutor to another, which physically differed (No. 6 yes, No. 7 no). The judgment of the own connection was influenced as well (No. 5).

Background noise at the far end side did not influence any judgment (No. 8-14). The reason is that the transmitted noise was rather low and therefore hardly perceptible for subjects. Background noise in the own environment was not considered when assessing the quality of individual connections (No. 19-21), but it affected the overall judgements (No. 15-18).

The own echo was clearly perceived (No. 32) and influenced the overall quality (No. 28, 31), but unexpectedly not the communication related variables (No. 29, 30). The results for listener echo (hearing an interlocutor’s voice double) and for hearing an interlocutor delayed are contrary to our expectations. In case an interlocutor experiences an echo of his/her own voice (talker echo), the echo of the interlocutor, whose voice is also fed back, had no influence (No. 34), although we calibrated the system such that listener echo could be easily perceived. Furthermore, in the same case of talker echo, the delay of the interlocutor, whose voice is not fed back, is perceived (No. 34), while the delay was not perceived in the case when an interlocutor could not perceive any echo at all (No. 22-27).

Packet loss was well reflected in the overall judgements (No. 35-38). The own connection was not judged differently (No. 39) and an interlocutors’ connection was perceived as being different when there was packet loss on that link (No. 41). However, the other connection without packet loss was also judged differently (No. 40).

In summary, the degradations were often judged in the way the structured technical analysis suggested, but some detailed findings were contrary to these expectations.

5. Discussion and conclusion

Asymmetric impairments in a multiparty setting lead to a quite complex situation: Even in case of only one degradation source, the translation into perceptible impairments is multiplied, and different degradations can lead to quite different perceptions, e.g. packet loss (a single voice is distorted) vs. echo (own voice audible & another voice double & some delay).

Verifying whether possible impairments are actually perceived is even more complex as not all impairments were rated as expected. Apparently, some degradations have to be above a certain threshold, e.g. the level of transmitted background noise, before they are included in quality ratings. Hence, some threshold-based decision step is needed, verifying if a degra-
To draw solid conclusions. To achieve more stable results, further research is needed to complement the presented first larger set of results (17 user groups with an overall of 47 participants). It is obvious that due to the studied asymmetric degradation scenarios, a given conversation leads to different perceptual effects for the participants. As a consequence, the number of cases per condition set according to Tables 1 and 2 is reduced.

To conclude, the main purpose of the present paper is to propose a systematic method that decomposes the complexity of asymmetric multiparty conferencing quality assessment. By a stepwise analysis, this approach allowed to draw a detailed picture on the relation between technical degradations and corresponding perception of impairments. While the complexity of the observed results verified the need for such a systematic approach, future work will be necessary to extend this method a) to include the discussed thresholding operation, b) to collect more data for stable modelling, and c) to include other degradation paradigms such as receive-side degradations.
6. References


