The Influence of Interactivity Patterns on the Quality of Experience in Multi-party Video-mediated Conversations under Symmetric Delay Conditions

Marwin Schmitt
CWI: Centrum Wiskunde & Informatika
The Netherlands
schmitt@cwi.nl

Simon Gunkel
CWI: Centrum Wiskunde & Informatika
The Netherlands
gunkel@cwi.nl

Pablo Cesar
CWI: Centrum Wiskunde & Informatika
The Netherlands
p.s.cesar@cwi.nl

Dick Bulterman
FX Palo Alto Laboratory
California, USA
dick.bulterman@fxpal.com

ABSTRACT

As commercial, off-the-shelf, services enable people to easily connect with friends and relatives, video-mediated communication is filtering into our daily activities. With the proliferation of broadband and powerful devices, multi-party gatherings are becoming a reality in home environments. With the technical infrastructure in place and has been accepted by a large user base, researchers and system designers are concentrating on understanding and optimizing the Quality of Experience (QoE) for participants. Theoretical foundations for QoE have identified three crucial factors for understanding the impact on the individual's perception: system, context, and user. While most of the current research tends to focus on the system factors (delay, bandwidth, resolution), in this paper we offer a more complete analysis that takes into consideration context and user factors. In particular, we investigate the influence of delay (constant system factor) in the QoE of multi-party conversations. Regarding the context, we extend the typical one-to-one condition to explore conversations between small groups (up to five people). In terms of user factors, we take into account conversation analysis, turn-taking and role-theory, for better understanding the impact of different user profiles. Our investigation allows us to report a detailed analysis on how delay influences the QoE, concluding that the actual interactivity pattern of each participant in the conversation results on different noticeability thresholds of delays. Such results have a direct impact on how we should design and construct video-communication services for multi-party conversations, where user activity should be considered as a prime adaptation and optimization parameter.

Categories and Subject Descriptors

H.4.3 [**Communications Applications**]: Computer conferencing, teleconferencing, and videoconferencing

General Terms

Measurement, Experimentation, Human Factors

Keywords

QoE, video-mediated communication, multiparty video-conferencing, delay, user study.

1. INTRODUCTION

Video-Conferencing is lately moving from the office to the home, where broadband bandwidth is widely available and devices can now easily join a session. More recent advances are enabling the next logical step: video-mediated group conversations. This paper explores this use case, analyzing the influence of the interactivity of each of the participants in the conversation. Unlike previous research that typically concentrates on system factors, we take into consideration the context and user factors. In particular, we study delay on the QoE of small gatherings.

The majority of previous research on QoE for remote communication has concentrated on dyadic use cases. For audio-only dyadic communication this has been extensively investigated and 150ms have been established as an industry standard for an acceptable delay [6]. This simple model of optimizing towards a minimal delay is not sufficient for the reality of the internet infrastructure. Currently, the situation is different, since video communication providers operate in the internet, where a multitude of uncontrollable, unknown and unforeseeable network problems can arise. To appropriately configure this complex infrastructure, under varying network conditions, a meticulous and comprehensive study of the different factors affecting the QoE is required. Previous research [11] has indicated that system factors alone are not sufficient to understand the QoE of an individual in a specific situation. This paper aims to advance this understanding by comparing different contexts and different interactivity patterns. In the past, the turn-taking model has been used in dyadic conversations to estimate the interactivity of a conversation [5]. However, this approach is not applicable anymore to multi-party conversations, since a different perception for each participant might arise depending on his/her involvement. We already reported about the investigation of asymmetric delay conditions and general differences between asymmetric, symmetric and dyadic setups

Table 2 System Configuration

System Setup	Desktop PCs (Core i7, 16GB Ram, SSD) Webcam (Logitech HD C920) Headset (Creative Soundblaster Xtreme) Video: 640x480px, 30fps, H264 Audio: Speex Network: Local Gigabit LAN, UDP, RTP
Conditions	•0ms-delay (avg = 75ms, sd = 31ms) •500ms-delay (avg = 564ms, sd = 34ms) •1000ms-delay (avg=1065ms, sd = 39ms) •2000ms-delay (avg = 2058ms, sd= 57ms).

[12]. In this paper we make a more fine-grained analysis user interactivity taking into account role-theory as well.

This paper aims to investigate the following novel research questions, regarding delay in multi-party video-mediated conversations:

- Context factors: Where are the lower (just-noticeable) and upper (not-acceptable) boundaries for delay in small-group video-mediated discussions?
- User factors: What influences have conversation roles and interactivity patterns on the perception of delay?

2. RELATED WORK

Theoretical models [9, 11] established that QoE is shaped mainly by three aspects: the system, the user and the context. From the system side, we want to investigate delay, since it is an inherent factor of remote-communication. The dyadic case has been investigated for unscripted scenarios [14] and scripted scenarios [15]. The multi-party scenario has been evaluated for the high-end halo system [4] and for scenarios that use unconventional settings (TV screen, several cameras) [2]. To our knowledge, there has been no investigation of delay effects in an unscripted multi-party video-mediated conversation. From a technical perspective, several studies evaluate realistic network conditions. For example, when connecting two computers between New York and Hong Kong, the round trip delay is up to 776ms for Google+, and 1467ms for Skype. Other systems, like Mebeam [8], have even higher one-way delay of up to 2770ms on average. Contrary to the system, the context is still under research as to which factors should be considered and which their impact is. So far research in video-mediated communication focused mainly on workplace scenarios and high-end systems (especially in its early cases) [4]. This, however, has recently shifted towards the home environment, exploring new scenario and setups for connecting families [1]. One approach is to use scripted conversations, i.e. the participants are told beforehand exactly what to say (e.g. alternated counting, number verification or a script for a service encounter [7]). These tasks reduce variability and influences that might occur in a natural conversation. The revealed lower boundaries are more sensitive, but this might not reflect the actual threshold in a conversation. For unscripted scenarios, the minimal requirement is to provide a topic for conversation [4], e.g. favorite food. Often goal-oriented tasks are employed, like a decision-making process [7].

An approach to look at the organization of conversations (who speaks when, and how do we manage not to speak all at the same time) is the turn-taking model [10]. It describes how we implicitly arrange our conversations by taking turns of connected utterances. Speech metrics have been used to qualify interactivity of a conversation [5] or the differences of speech synchronization [13].

Table 1 Questions and labels

label	Question	Scale
quality	What is your opinion of the connection you have just been using?'	Bad <-> Excellent
annoyance	To what extent where you annoyed by delay in the connection?	No annoyance <-> severe annoyance
noticeability	'How noticeable did you perceive the delay in the connection?	Not at all <-> Very much

3. METHODOLOGY

Our study of multi-party video-mediated discussions under delay investigated asymmetric and symmetric delay conditions. We already reported about the asymmetric case which used a slightly different setup (different delay conditions and quiz questions) thus this paper focuses on the symmetric case. The study followed an experimental design with randomized conditions. In our study participated 39 (20 female, 19 males, with an average age of 36 years (min 20 years, max 60 years)). All groups were of mixed gender, in two groups 3 and 2 participants respectively knew each other beforehand. The experiment was conducted in English, in which all participants were fluent. Participants were in groups of 5, except one group with 4 participants, as one participant did not show up and it was not possible to find a replacement in time. All participants were seated in separate rooms after an introduction round in which we explained our research and the experiment.

The task of our participants was a quiz style question-select answer scenario. The participants had to discuss together the best answer to questions about surviving in the wilderness. The task is based on the team building exercise from [3]. One participant was asked to be the moderator, to submit the final group answers and move the discussion along to keep the 10 minutes time constraint per round. The order of the quiz-questions did not change in the experiment but the order of the delay. Each round of questions was in total 8 times discussed, twice under each condition. After each round we assessed subjective feedback via questionnaires, in this paper we examining our questions related to perceived quality, shown in Table 1. The questions all used a nine-point likert-like scale.

To conduct the experiment and set the desired delay conditions, we used the VMC-TB, presented in [11], the exact technical configuration and test conditions can be found in Table 2.

4. RESULTS

The responses were normally distributed, with respect skewness and kurtosis below 2.

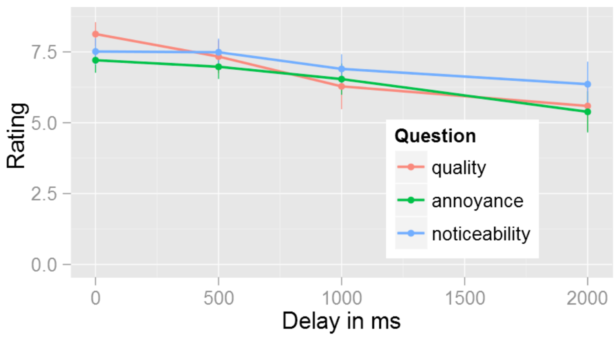

Figure 1 - Average Questionaire Quality Ratings with 95% confidence intervals

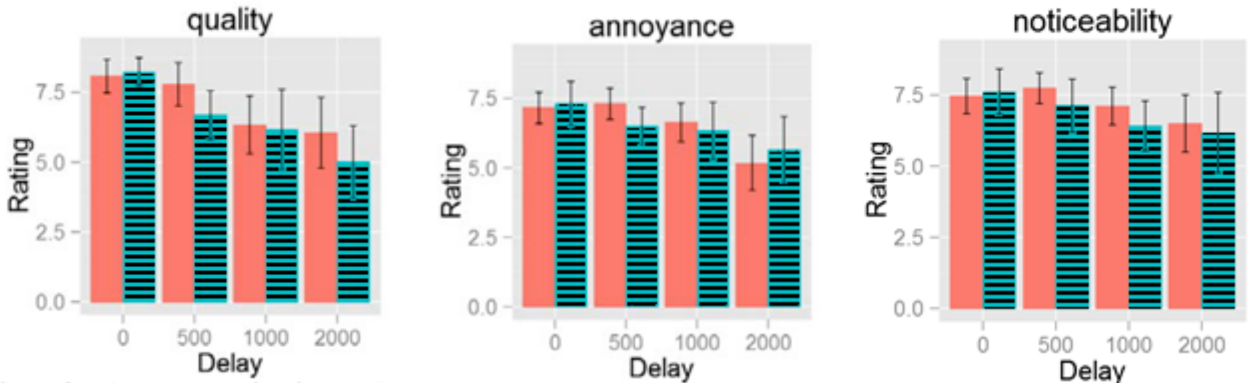

Figure 2 - Average questionaire results clustered by blocks percentage duration of active participants (red solid), and non-active participants (blue striped) with 95% confidence intervals.

We used ANOVA to compare the goodness of a fitting a linear model with our data, to assess whether there was general effect our independent variable delay towards the dependent variables (see Table 1). We performed a pairwise difference test with the pairwise student's t-test to see which tests are significantly different. We clustered out participants by speech patterns using k-means into two groups: active and inactive participants. For the differences between the active and non-active groups (see Section 4.1) the Mann-Whitney U test.

We asked participants to rate the quality of the connection, how annoyed they were by the delay and how much they noticed the delay. The responses to these items are shown in Figure 1. For all items, a lower score means a worse perception, i.e. less quality, more annoyance or the delay was more noticeable.

The analysis revealed that the influence of delay on the quality question was statistically significant ($p < 0.05$). Influence of delay on annoyance was statistically significant ($p < 0.05$). The influence of delay on noticeability was just below the significance confidence of 0.05 ($p = 0.052$). Thus, for the noticeability, we performed a pair-wise comparison of the conditions using a one-tailed pair-wise T-Test. This revealed that the noticeability of delay between 0ms and 500ms is nearly identical ($p = 0.402$), but there is a significant difference between 500ms and 1000ms ($p = 0.018$) and no statistical differences between 1000ms and 2000ms ($p = 0.099$). The differences between 0ms->1000ms, 0ms->2000ms and 500ms->2000ms are also statistical significant ($p < 0.05$).

4.1 Qualification by Speech Patterns

We further hypothesize that a concrete speech pattern will influence perception. While the approach in previous research was to build an interactivity metric for the whole conversation [5, 13, 15], we use speech patterns to group the participants. We clustered our participants by speech patterns using k-means into two groups: active and non-active participants.

We divided participants by the amount of speaking time, when compared to the total speaking time of the group. From the automatically generated speech pattern data we computed the percental amount of speaking time each participant had in each round. For the clustering process we offset this value by the standard deviation of all our samples and the deviation of the group of the participant. We then used the k-means algorithm to perform the classification. The elbow-criterion was used to determine that we gain the most explanation of variance with two clusters.

Figure 2 shows the results for the three questionnaire items. We performed a pairwise comparison of different delay conditions for active and non-active participants. Active participants have a significant drop in the perception between 0ms and 500ms ($p = 0.014$), but not between the other conditions ($p > 0.05$). For non-active participants only the difference between 500ms and 1000ms is statistically significant ($p = 0.003$, for other conditions $p > 0.05$). The comparison of the differences between active and non-active participants showed that there are indications that the perception of quality is different at 500ms ($p = 0.013$), but very similar at the other conditions ($p > 0.1$).

For annoyance, the results follow a similar pattern. Active participants have a significant ($p = 0.025$) rise in annoyance between 0ms and 500ms while for non-active participants the difference is insignificant. 1000ms is the statistical significant ($p = 0.009$) difference for non-active participants, being now nearly the same as for active participants. Interestingly the difference between 1000ms and 2000ms is strongly noticeable for non-active participants ($p = 0.0003$) but not for active participants ($p=0.15$).

Noticeability is generally less affective by delay. Both groups start with a similar perception at 0ms, going minimal up for non-active participants and slightly for the active one, but for both groups the difference is not significant. Due to the large variance the difference becomes noticeable for active participants between 0ms and 1000ms ($p= 0.034$) and between 0ms and 2000ms ($p= 0.004$) for non-active participants. Interestingly the difference for non-active participants happens between 500ms and 1000ms ($p= 0.048$) but not between 0ms and 1000ms ($p = 0.116$).

5. DISCUSSION

5.1 Thresholds

Our data from the generalized case (see Figure 1), suggests that noticeable quality degradation sets in between 500ms and 1000ms delay. This is a higher delay than reported in dyadic studies [14, 15] and similar to the study from Berndtsson et al. [2].

Contrary to disturbances in audio- and video-streams participants cannot observe a delay directly. It is only indirectly perceivable due to e.g. longer pauses and more double talk. Even though participants in our experiment were aware that the connection might be delayed, it was still difficult for some participants to assess whether it was a technical problem. The variance of the perception of delay was thus very high between participants, as was also revealed in the debriefing discussions:

[P3]: "*It wasn't noticeable for me.*"

[P4]: "*I was already on the top of my annoyance level. I AM LIKE, HELLO I AM TALKING HERE, CAN ANYBODY HEAR ME IN THIS PLACE! WHAT IS HAPPENING? And I was sometimes asking, are you hearing me, and everybody was just looking?* "

And in a different group:

[P1]: *The delay wasn't very annoying; I didn't even notice the delay really. I just noticed it because people were saying, there is a delay.*

[P2 asks]*: oh, really?*

[P1]*: yes I didn't notice it. I just thought people were thinking.*

[P2]: *I was very annoyed by it... It was like 4-5 seconds. We were like 5 sentences (ahead) and then you came. "grgh"*

For a listener, who was not directly involved this seemed to be easier detectable:

[P6]: "*Sometimes people would interrupt each other and you would notice that it wasn't intentional since they were completely unaware of what the other one said"*

Although the participants gave relatively good ratings, after exploring the recordings, we observed that the delay forces participants to employ additional explicit organization mechanisms. Instead of somebody taking a turn by simply speaking, another participants would hand the turn explicitly (verbally) over to another participant. The change of conversation structure and the comments of our participants suggests that with a one-way delay between 1000ms and 2000ms, a conversation without additional explicit organizing mechanisms is not possible.

5.2 Active and non-active participants

The variance we could observe in our responses and the highly different perception reported in the debriefing, suggests that there are other factors at play. We could observe in most group, that the participant assigned as the moderator took over the leading role. In some groups, a particularly shy person was chosen by chance as moderator, causing another participant to take over this role. This classification usually results in one or two active participants. In our data analysis we showed that by using the amount of speaking time, we found two groups with distinct perceptions. For the active participants the noticeable degradations did already occur in between 0ms and 500ms. While for normal participants the difference was still between 500ms and 1000ms.

The results show that to understand the impact of delay the interaction in the context has to be considered in detail. The effects of delay are more present for participants directly involved in the interaction. The more passive listening roles do not occur in dyadic conversations. In the debriefing participants reflected upon that the moderator had a more difficult role:

[P5]: [topic was higher delay conditions] *for us it was easy, but you* [to moderator] *you needed to keep contro*l.

And it was noticed that if you are not so active due to the video-stream it was still possible to be part of the conversation:

P6: *P7 didn't say much but it was always easy to see if* [he/she] *was agreeing and following along or had a different opinion.*

6. CONCLUSION

We have reported on the first user experience evaluation of a five party scenario with off-the-shelf end-user hardware. We found that the degradation of quality perception was strongly noticeable between 500ms and 1000ms. We described how we designed a scenario that allows us to gain insight into role based perception. We provided a novel approach to use turn-taking data to gain insights into the differences in experiencing delay for individual participants in one session. In this setting we were able to classify our participants, based on their actual interaction, into active and non-active participants. The analysis showed that more active participants already perceive the quality degradation between 0ms and 500ms while for non-active participants this drop is between 500ms and 1000ms. We observed that communication is possible even with high delays of over 2000ms, but the implicit conversation organization is replaced by an explicit one.

The result shows that even though the QoE of active participants suffers under high delay conditions, the overall average QoE might still be satisfactory. These findings give us indications on which participants to prioritize in situations where the resources are limited in demanding multi-point scenarios.

7. ACKNOWLDGEMENT

This research has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2011-287760 (Vconect project).

8. REFERENCES

[1] Ames, M.G., Go, J., Kaye, J.J. and Spasojevic, M. 2010. Making love in the network closet: the benefits and work of family videochat. *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (2010), 145–154.

[2] Berndtsson, G., Folkesson, M. and Kulyk, V. 2012. Subjective quality assessment of video conferences and telemeetings. *Packet Video Workshop (PV), 2012 19th International* (2012), 25–30.

[3] Biech, E. 2007. *The Pfeiffer book of successful team-building tools: Best of the annuals*. Pfeiffer.

[4] Geelhoed, E., Parker, A., Williams, D.J. and Groen, M. 2009. *Effects of Latency on Telepresence*. HP labs technical report: HPL-2009-120 http://www. hpl. hp. com/techreports/2009/HPL-2009-120. html.

[5] Hammer, F., Reichl, P. and Raake, A. 2005. The well-tempered conversation: interactivity, delay and perceptual VoIP quality. *Communications, 2005. ICC 2005. 2005 IEEE International Conference on* (2005), 244–249.

[6] ITU-T RECOMMENDATION 2003. ITU-R G.114 - One-way transmission time. (2003).

[7] ITU-T RECOMMENDATION, P. 130. 2013. ITU-P.1301 - Subjective quality evaluation of audio and audiovisual multiparty telemeetings.

[8] Lu, Y., Zhao, Y., Kuipers, F. and Van Mieghem, P. 2010. Measurement study of multi-party video conferencing. *Proceedings of the 9th IFIP TC 6 international conference on Networking* (Berlin, Heidelberg, 2010), 96–108.

[9] Patrick Le Callet, Andrew Perkis and Sebastian Möller eds. 2013. Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003). Version 1.2.

[10] Sacks, H., Schegloff, E. and Jefferson, G. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*. 50.5 4 (1974), 696–735.

[11] Schmitt, M., Gunkel, S., Cesar, P. and Hughes, P. 2013. A QoE Testbed for Socially-aware Video-mediated Group Communication. *Proc. of the 2nd International Workshop on Socially-aware Multimedia* (2013), 37–42.

[12] Schmitt, M., Gunkel, S., Pablo, C. and Bulterman, D. 2014. Asymmetric Delay in Video-Mediated Group Discussions. *To appear in 6th International Workshop on Quality of Multimedia Experience (QoMEX), 2014* (2014).

[13] Schoenenberg, K., Raake, A., Egger, S. and Schatz, R. 2014. On interaction behaviour in telephone conversations under transmission delay. *Speech Communication*. 63–64, (Sep. 2014), 1–14.

[14] Tam, J., Carter, E., Kiesler, S. and Hodgins, J. 2012. Video increases the perception of naturalness during remote interactions with latency. *Proc. of CHI'12* (New York, NY, USA, 2012), 2045–2050.

[15] Wang, J., Yang, F., Xie, Z. and Wan, S. 2010. Evaluation on perceptual audiovisual delay using average talkspurts and delay. *Image and Signal Processing (CISP), 201 3rd International Congress on* (2010), 125–128.