Towards context-aware interactive Quality of Experience evaluation for audiovisual multiparty conferencing

Marwin Schmitt[1], *Judith Redi*[12], *Pablo Cesar*[12]

[1] CWI: Centrum Wiskunde & Informatica
[2]Delft University of Technology

m.r.schmitt@cwi.nl, j.a.redi@tudelft.nl, p.s.cesar@cwi.nl

Abstract

In modern video conferencing services, just as in common video delivery, most of the resource optimization is taken care of in the codec layer. Modern codecs like H.264 use detailed perceptual models to optimize the data reduction in way that it is least noticed by us. Already early evaluations of telecommunication systems could establish that there are different thresholds for a good quality depending on the situation. It is further known that subjective quality perceptions vary from user to user. But the space of user and context factors is still largely unexplored. To gain insight in which parameters are key in differentiating quality perception, we need to explore the interaction in different situations while keeping a tight control over the system parameters. In this paper we explore how clustering participants by their interaction or rating behavior can reveal subgroups that show significantly different perception of the QoE delivered by the same videoconferencing system. While for a cluster of users we find video quality to influence other QoE dimensions such as audio, for another cluster this is not the case. We explore whether this effect is due to conversational dynamics (contextual factor) or individual preferences (user factor) and discuss what this would mean for the design of future video-conferencing systems, that want to dynamically adapt to situation and participants.

Index Terms: QoE; multiparty; audiovisual conferencing; contextual factors; human influence factors

1. Introduction

In multimedia delivery systems, the shift from Quality of Service (QoS) to Quality of Experience (QoE) has not only put the user in the center of the system evaluation, but also acknowledged a shift in the telecommunication infrastructure design. In nowadays' multitude of services and providers, end user devices and scenarios in which they are used, system optimization cannot depend only on technological (system) factors anymore. Conceptual models for QoE measurement and optimization include indeed, besides system factors, both contextual and human influence factors [1]–[3]. As we have engineered the systems, their properties are more or less known to us, and set to maximize the trade-off between delivery quality and employment of resources. The challenge lies in correctly identifying and understanding the key elements dominating the vast space of context and user factors. QoE studies have already revealed differences between tasks [4] and user preferences [5], social context[6], personality [7] and more [1].

Figure 1 Screenshot from the trial

In this paper we focus on desktop multi-party conferencing. With modern hardware and internet connections, multiparty videoconferencing has become available for the masses. Multi-party video-conferencing does not only pose new challenges for the system (more resources are needed and inter-destination synchronization is required) but small group conversations follow also a different dynamic than one-to-one conversations [8], which in turn requires re-evaluation of QoE in such scenarios. The attention is divided between multiple streams. Speaker selection is in collocated settings highly correlated to gaze [9] which is not faithfully conveyed in desktop video conferencing.

In the previous work we analyzed a study regarding the influence of video quality impairments due to coding (between 256Kbit and 4Mbit) and packet loss (0% or 0.5% loss) on QoE in visual-focused task for multi-party video-conferencing [10]. A screenshot of the system used in the experiment is shown in Figure 1. The impact of encoding and loss rate on quality had been studied only in two-party scenarios [11][12][4]. We presented how the manipulated system factors (encoding bitrate and packet loss of the video streams) influenced the QoE in terms of perceived video, audio quality and overall quality. With the help of linear mixed effect models we could relate variance in ratings to either system factors or group and user differences. The analysis showed that the perceived audio quality was also slightly affected by the manipulation of the video quality (note the audio quality stayed the same throughout the experiment). This effect that has been shown in previous research [13]. Our analysis showed that the differences in video quality could only explain small parts of the differences in perceived audio quality, but gave indications that the differences are related to either user or group factors. In this paper we extend the analysis of [10] with a detailed analysis of the perceived audio quality.

Specifically our research questions are:

- Are there patterns, not related to the system factors, which can explain the large variations in perceived audio quality ratings?
- If not related to system factors, do these patterns relate to user characteristics or preferences?

- If not related to system factors, are these patterns related to the conversation dynamics?

Our approach is to look into patterns of user characteristics (via the assessed ratings) and interaction (via speech patterns). We identify two different groups of users: one for which audio quality is strongly penalized by losses in video quality, and one for which this effect is negligible. The analysis of speech patterns, revealed only small differences in average simultaneous talk length, even though the users penalizing the perceived audio quality reported a strong subjective effect of experiencing more double talk. We discuss the results under the question whether these differences in QoE perception are due to different psychological profiles (i.e. user factor) or due to the actual interaction happening (i.e. context factor). While we see slight indications that the differences are more likely to relate to user individual characteristics, it is clear that more research is needed on the connection how a single system factor influences the holistic QoE and on how conversational problems relate to QoE is needed.

2. Study Setup

To investigate the effect of video quality impairments on audio and visual QoE, we organized videoconferencing sessions involving 4 participants. 7 groups participated, with a total of 28 subjects (18 female, average age: 31.9, sd: 10). To stress video usage, we chose a task that would require visual interaction: building a Lego® model together. Each of the 4 participants received a disassembled Lego model and only part of the instructions to assemble it: Interaction with the other participants was therefore necessary to be able to build the whole model

We chose bitrate and packet loss rate as system factors as they are the main dynamic parameters for video quality. The values (bitrate:256kbps, 1024kbps, 4096kbps; loss: 0%; 0.5% random) were combined into a full factorial design to be prototypical of different home scenarios. Each group experienced 4 of the 6 possible conditions (counterbalanced). After completing one round with one of the 4 conditions (for a duration of 7 minutes), participants filled in a questionnaire including audio and video quality evaluation questions (based on ACR scales) and questions on the conversational dynamics (see Table 1). Further they filled out a questionnaire at the end of the experiment with additional questions regarding the enjoyment of study and task, video-conferencing experience and more (see Table 3).

The layout of the streams was a 2x2 grid with equal sizes for each participant (shown in Figure 1) on a 27" WQHD Screen. The self-view of the participants was always shown in highest quality. Participants were wearing Logitech Creative Soundblaster Xtreme headsets. High range Logitech C920 webcams were used for all participants. The cameras are USB2 devices and thus cannot stream 720p at 24fps video without compression. We used the internal H.264 encoding capabilities of the camera. We evaluated the parameters which can be set over the universal-video-class driver and found that the bitrate would saturate at 5.8Mpbs. We always request the maximum quality from the camera and re-encode it with x.264 in GStreamer to have more fine grained control over the quality and better comparability to previous studies and simulations conducted in our group. We further optimized the encoding for realtime communication (preset: realtime; speedpreset: ultrafast; iframe-interval: 24frames (1second); sliced threads enabled). The audio was encoded with an AMR narrowband codec. Note that only video was impaired, whereas audio was unimpaired. Further details on the experimental setup can be found in [10].

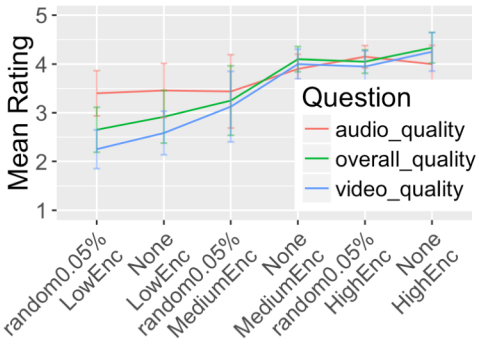


Figure 2 Perceived quality ratings [16]

3. Analysis

The analysis of the impact of bitrate and loss on overall, audio and video perceived quality was presented in [10]. The results showed that at 1Mbit encoded H.264 streams the quality perception seemed to saturate (see Figure 2). Increasing the bitrate to 4Mbit was noticed by only very few participants, whereas lowering it to 256Kbit severely decreased the QoE. The fine details of the Lego pieces and instructions would get lost at these rates. In turn, participants compensated by describing their actions in much more detail. The packet loss at 0.5%, (implying only small momentarily impairments) was disturbing only few participants.

The previously reported analysis showed that the manipulation of video quality had a small effect on audio quality (see Figure 2), as already noted in literature [14]. We measured the influence with the help of linear mixed effect models, counting bitrate and loss as fixed effects and test group and individual participant as random effects. The group factors is more likely to be related to the (shared) interaction and the user more likely to be related to individual characteristics of the user. The model (m1) was found to predict perceived audio quality best:

(m1) *audio quality ~ bitrate + loss + (bitrate|Group/User)*

For brevity we are reporting the models in the R [15] notation. This related to the classical matrix notation $y = X\beta + Z\gamma + \epsilon$ in the following way:

- y the dependent variable which is to be predicted is on the left hand side of the ~ (here perceived audio quality)
- The fixed effects matrix X is constructed by expression of factors, + denotes a linear combination without interaction, * with (thus here bitrate and loss without interaction)
- The random effects matrix Z is constructed by expression in the form of (slope factor|random factor/subfactor). The

Table 1 Questions regarding perceived conversation dynamics (assessed after each condition using a 5point likert scale). Label in bold. Correlation coefficients of Pearsons's r, significant correlations ($p < 0.05$) in bold.

Question	**Correlation Coefficient of perceived audio quality and Question for:**	
	Cluster1	Cluster2
Because of the quality of the video-conferencing, it was easy for me to interrupt people when I wanted to. (**interruptions**)	**0.67**	0.23
Because of the quality of the video-conferencing, it did not happen that someone else and I started talking at the same time.(**doubletalk**)	**0.62**	0.05
Because of the quality of the video-conferencing, it was easy to keep track of the discussion. (**keep_track**)	**0.66**	**0.35**
The discussion was lively. (**lively_discussion**)	**0.77**	**0.38**

/ denotes nested factors (here slopes are constructed per bitrate for each group and user within that group)

- The to be estimated β, γ and ϵ are omitted in this notation. For further details see [15].

To evaluate the goodness of fit of linear mixed effect models, two kind of R^2 values can be computed [16]: marginal R^2, which quantifies the explained variance due to the fixed factors, and conditional R^2, which quantifies the explained variance considering also the random effects.

The general effect of video bitrate and loss settings on audio quality was weak (marginal R^2 of 8.45%) [10] but the high when considering user and group idiosyncrasies (conditional R^2 value of 73.69%). This big difference led us to investigate, whether there are factors in the groups or users, which would explain the perceived audio quality further.

In particular, the comparison of models with only group or user as random factors in [10] pointed out that most of the ratings variance could be explained by the characteristics of the individual user, rather than the group that each participant belonged to. This motivated us to look first into individual differences in QoE perception.

To capture individual sensitivity to QoE globally, rather than at a session level, we clustered participants according their average audio quality rating across all 4 sessions, using K-means. An elbow-plot revealed that 2 clusters would give the best ratio of explained variance to number of clusters. The resulting two clusters of 13 and 15 participants seemed include participants with a distinctively different sensitivity to audio quality: the first cluster had an average audio quality rating (across all participants and conditions) of 3.13, while for the second group this was 4.2. It is important to note that, except for one group, the two clusters included participants from all groups; i.e., participants to the same session were not necessarily clustered together. Hence, a group or session effect on their ratings is unlikely. An inspection of their audio quality ratings (see Figure 3a) showed that their perception shows a different pattern. To quantify this distinction, we build two new models, using cluster as a further fixed factor, either not (m2) or interacting (m3) with bitrate and loss:

(m2) *audio quality ~ (bitrate+loss) + cluster+(bitrate| Group /User)*

(m3) *audio quality ~ (bitrate+loss)*cluster+(bitrate| Group /User)*

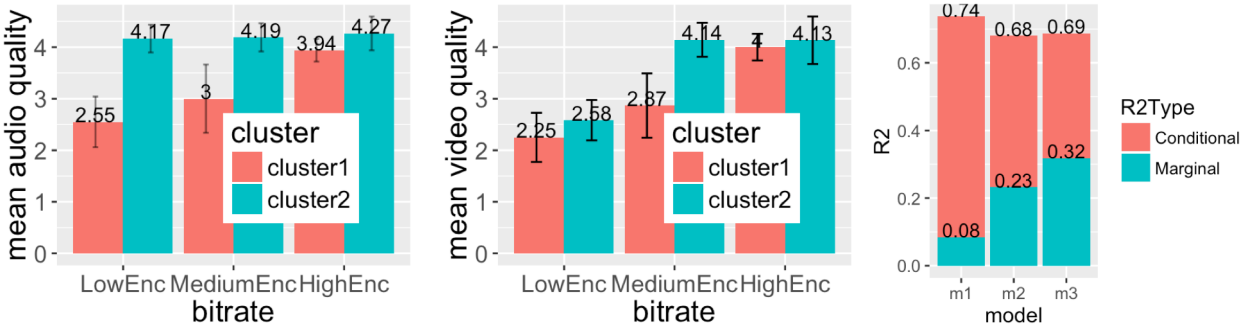
To assess whether adding a factor improves the model significantly we used the Likelihood Ratio Test (LRT) which compares for two models if the improvement of fit of the model with more factors is better in relation to the amount of factors needed to achieve this fit. The model with interaction is the one best fitting the data (LRT(m1, m2)= $\chi^2(1)$=17,66;p<0.001 and

Table 2 Model m3 p-values of pairwise comparison for each cluster between bitrate and loss (row 3-4) and between the clusters for each level of bitrate and loss (row 6)

cluster	Encoding			Loss	
	Low-High	Low – Medium	Medium - High	None –0.5%	
cluster1	<0.01	0.18	0.67	0.99	
cluster2	0.92	0.99	0.92	0.93	
	Low	Medium	High	None	0.5%
Cluster1-Cluster2	<0.001	0.13	0.69	<0.001	<0.001

LRT(m2, m3)= $\chi^2(3)$=9.32;p<0.05). The marginal R^2 values (variance explained by the fixed effects bitrate, loss and cluster) and conditional R^2 values (variance explained considering within Group and User differences) are plotted in Figure 3c. Our initial model m1 had only 8.45% explained variance by fixed effects. The addition of the cluster fixed factor in interaction with the other two improves this to 32% of explained variance. Hence, there seems to be a strong influence of individual sensitivity on audio quality. We checked the contrasts with the R lsmeans [17] package. The overall difference between the two clusters is significant ($p < 0.001$) and further paired comparisons (shown in Table 2) reveal that this difference gets stronger as the quality degradations get stronger. More specifically, participants in cluster 1 perceived the degradation in video quality (as to influence the audio quality as well. Participants in cluster 2, equally large, scored audio quality significantly higher, and independent on the video degradation. It was interesting, at this point, to check whether the participants in cluster 2, rating audio quality higher, were more positive in judging all aspects of QoE, i.e., generally more tolerant to impairments. We added participant cluster as a factor to models predicting perceived video quality and overall quality of [10] (analogue to model m3 for perceived audio quality). The clustering factors showed to improve the models significantly for perceived video quality and overall quality ($\chi^2(4) = 9.76$; $p < 0.05$ and $\chi^2(4) = 15.44$; $p < 0.01$, respectively). In the plot of perceived video quality in Figure 3b we can observe the trend that cluster1 participants also here rated the quality more critical than cluster2 participants.

To investigate whether the perceived audio quality is related to the (perceived) conversation dynamics, we correlated (with Pearson's r) the perceived audio quality ratings of the participants in each cluster with the questions regarding conversation dynamics (see Table 1). We can see the general pattern that for cluster1 the perceived audio quality is strongly correlating with all questions regarding conversation dynamics. For participants in cluster2 there is only weak correlations with,



(a) (b) (c)

Figure 3 (a) Perceived audio quality by cluster (b) perceived video quality by cluster (c) marginal and conditional R^2 values of models

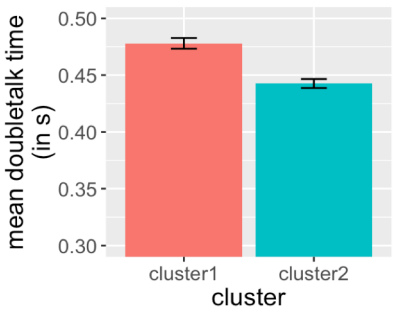


Figure 4 double talk time by cluster with standard error

liveliness of the discussion and keeping track of the conversation.

We can see that for participants, who rated the audio quality worse with worse video quality, similar effects seemed to happen with the conversation dynamics

Thus we proceeded to investigate difference in the actual speech interaction of the participants by segmenting the audio data in on-off speaking patterns [18]. We found a difference in the average time participants were involved in two or more people speaking at the same time (hereon double talk). The plot of mean double talk time in Figure 4 shows that participants in cluster1 were slightly longer involved in double talk. A generalized linear model with gamma link function revealed that the difference is statistically significant ($\chi^2(1) = 32.70$; $p < 0.001$).

We further investigated whether there were significant differences in the questions we assessed at the end of the experiment by comparing them with an unpaired Wilcoxon rank sum test. The questions and p-values are plotted in Table 3. The test showed a significant difference in enjoyment of the study (mean rating 4.07 for cluster 1 compared to 4.87 for cluster2) and in the rating of their own video quality which was shown unimpaired during the whole experiment (mean rating 3.85 for cluster 1 compared to 4.67 for cluster2).

4. Discussion

The analysis of the perceived audio quality showed that users could be differentiated into two groups: the first group more sensitive to impairments in videos, to the point that their audio quality evaluations would be impacted by video distortions as well, and the second group generally less bothered by impairments in video, and judging audio quality as high independent on the video bitrate. In regards to our first research question, the high difference between random and fixed factors, was indeed an indication that further patterns could be found, which were not related to the system factors. Our next two research questions, whether these impacts are related to user characteristics or interaction behavior could not be solved conclusively. Further analysis showed that the first group perceived also to have more conversation problems with degrading video quality. The relation between conversation problems and audio quality is not yet extensively explored. It is known that the interactivity (e.g. amount speaker changes) aggravates the effect of delay [19][20][21]. There are also indications that in cases of double talk higher audio quality improves the experience [22]. We have some indications that participants in this cluster were indeed more involved in double talk. However, further research is needed to properly characterize the relationship between conversational dynamics and quality perception.

We see two main lines of interpretation:

(1) We have two distinct user groups that have their origin in individual differences, idiosyncratic preferences or focus. Participants in cluster1 could have generally a more holistic experience (as they rate also video and overall quality more critical), or they could be used to a better quality (as they rate the unimpaired self-view lower). The lower enjoyment allows no interpretation of cause or effect.

(2) The interaction (e.g. involvement in double talk) was really different for both groups. The difficulty with the automated analysis of double talk is that it is hard to differentiate between double talk that is usually not perceived negatively (e.g. so called backchannels (e.g. a confirming "mhm") or laughing at the same time) and real interruptions (intentional or unintentional). The latter is usually longer than the former which gives indication that the difference in mean double talk time could be important. As this is an exploratory result, the study design does only allow limited inferences of the cause, further studies would be needed to investigate this in detail.

If we presume that user factors are important, services that can gather long-term information about the users (e.g. social networks) would be able to create better services, personalizing delivery strategies. On the other hand, if the actual interaction is the key factor, such smart services need advanced capabilities for automated conversational dynamics analysis. The challenges in here further lie in the realtime processing data which would be additionally noisier than in laboratory experiments. Besides clarification of the interaction vs preferences question, further research towards the interaction with additional system factors, delay and audio quality, would provide insights how to optimize video-conferencing systems. The relation is needed to fine tune delay vs quality (for buffer and loss handling strategies) and trade-off between audio and video quality (from a bandwidth perspective if would be easier to deliver unimpaired audio quality, but if users have a bad impression of the audio quality due to bad video quality it might not improve the QoE after all).

Here, as in most current research, insights about different user groups or roles are obtained with the help extensive post-processing and analysis. It is a challenging task for future work, to turn this knowledge into mechanisms for actual videoconferencing systems, where such inferences have to be made in real-time.

Table 3 P-values of Wilcoxon Rank Sum Test for the final questionnaire on 5 point likert-like scale (end labels in parenthesis)

Question	**p-value**
In enjoyed participating in this study (**enjoyment; Not at all –very much**)	<0.01
I liked the task of playing with Lego. **(likelego; Not at all –very much)**	0.63
How would you rate the quality of your own video? **(ownvideo; bad –excellent)**	<0.01
I noticed delay in the connection and it was: **(delay;very annoying – imperceptible)**	0.1
Did you have problems determining which participant was speaking? **(problemsspeacking; Never-very often)**	0.1
I am very experienced in using video-conferencing systems. **(priorexp; Very unexperienced-Very experienced)**	0.37
Age	0.61

References

[1] D. Geerts, K. De Moor, I. Ketyko, A. Jacobs, J. Van den Bergh, W. Joseph, L. Martens, and L. De Marez, "Linking an integrated framework with appropriate methods for measuring QoE," in *QoMEX'10*, 2010, pp. 158–163.

[2] S. Moller, K.-P. Engelbrecht, C. Kuhnel, I. Wechsung, and B. Weiss, "A taxonomy of quality of service and quality of experience of multimodal human-machine interaction," in *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, 2009, pp. 7–12.

[3] M. Schmitt, S. Gunkel, P. Cesar, and P. Hughes, "A QoE Testbed for Socially-aware Video-mediated Group Communication," in *Proc. of the 2nd International Workshop on Socially-aware Multimedia*, New York, NY, USA, 2013, pp. 37–42.

[4] B. Belmudez, S. Moeller, B. Lewcio, A. Raake, and A. Mehmood, "Audio and video channel impact on perceived audio-visual quality in different interactive contexts," in *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on*, 2009, pp. 1–5.

[5] D. Strohmeier, S. Jumisko-Pyykkö, and K. Kunze, "Open profiling of quality: a mixed method approach to understanding multimodal quality perception," *Adv Multimed.*, vol. 2010, p. 3:1–3:17, Jan. 2010.

[6] Y. Zhu, I. Heynderickx, and J. A. Redi, "Understanding the role of social context and user factors in video Quality of Experience," *Comput. Hum. Behav.*, vol. 49, pp. 412–426, Aug. 2015.

[7] S. C. Guntuku, M. J. Scott, H. Yang, G. Ghinea, and W. Lin, "The CP-QAE-I: A video dataset for exploring the effect of personality and culture on perceived quality and affect in multimedia," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–7.

[8] P. M. Aoki, M. H. Szymanski, L. Plurkowski, J. D. Thornton, A. Woodruff, and W. Yi, "Where's the 'party' in 'multi-party'?: analyzing the structure of small-group sociable talk," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, New York, NY, USA, 2006, pp. 393–402.

[9] R. Ishii, K. Otsuka, S. Kumano, M. Matsuda, and J. Yamato, "Predicting Next Speaker and Timing from Gaze Transition Patterns in Multi-party Meetings," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, New York, NY, USA, 2013, pp. 79–86.

[10] M. Schmit, J. Redi, and P. Cesar, "1Mbit is enough: Video Quality and Individual Idiosyncrasies in Multiparty HD Video-Conferencing," in *Proceedings of the 8th International Conference on Quality of Multimedia Experience (QoMEX 2016)*, 2016.

[11] S. Egger, M. Ries, and P. Reichl, "Quality-of-experience beyond MOS: experiences with a holistic user test methodology for interactive video services," in *21st ITC Specialist Seminar on Multimedia Applications-Traffic, Performance and QoE*, 2010, pp. 13–18.

[12] B. Belmudez and S. Möller, "Audiovisual quality integration for interactive communications," *EURASIP J. Audio Speech Music Process.*, vol. 2013, no. 1, pp. 1–23, Nov. 2013.

[13] J. G. Beerends, D. Caluwe, and F. E, "The Influence of Video Quality on Perceived Audio Quality and Vice Versa," *J. Audio Eng. Soc.*, vol. 47, no. 5, pp. 355–362, May 1999.

[14] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, "Perceptual-based quality assessment for audio–visual services: A survey," *Signal Process. Image Commun.*, vol. 25, no. 7, pp. 482–501, 2010.

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[16] S. Nakagawa and H. Schielzeth, "A general and simple method for obtaining R2 from generalized linear mixed-effects models," *Methods Ecol. Evol.*, vol. 4, no. 2, pp. 133–142, Feb. 2013.

[17] R. V. Lenth, "Least-Squares Means: The R Package lsmeans," *J. Stat. Softw.*, vol. 69, no. 1, pp. 1–33, 2016.

[18] H. Sacks, E. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

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

[20] K. Schoenenberg, A. Raake, S. Egger, and R. Schatz, "On interaction behaviour in telephone conversations under transmission delay," *Speech Commun.*, vol. 63–64, pp. 1–14, Sep. 2014.

[21] F. Hammer, P. Reichl, and A. Raake, "The well-tempered conversation: interactivity, delay and perceptual VoIP quality," in *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, 2005, vol. 1, pp. 244–249.

[22] P. 131. ITU-T RECOMMENDATION, "ITU-P.1312 - Method for the measurement of the communication effectiveness of multiparty telemeetings using task performance." 25-Apr-2016.