VRIJE UNIVERSITEIT

Personal Quality of Experience: Accurately modelling Quality of Experience for multi-party desktop video-conferencing based on systems, context and user factors

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op dinsdag 10 september 2019 om 09.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Marwin Robert Schmitt

geboren te Berlin, Duitsland

promotor: prof.dr. D.C.A. Bulterman

copromotoren: dr. P.S. Cesar
dr. J.A. Redi

Personal Quality of Experience: Accurately modelling Quality of Experience for multi-party desktop video-conferencing based on systems, context and user factors

Marwin Schmitt

Acknowledgments

Firstly I'd like to thank my promotor Dick Bulterman, without whom this thesis would not have been possible. Seldomly I have met someone who would grasp something so quickly and deliver sharp comments so to the point. The biggest thanks goes to my supervisor Pablo Cesar, who accompanied me through all stages of the PhD. He was not only my advisor in all research matters but also the person I could go to with the all the experiences I was going through during that part of my life. I could not have wished for a person that I could talk more openly and honestly with than Pablo. Thank you very much.

Over the Qualinet project, I became acquainted with Ernestasia Siahaan, a fellow PhD in Quality of Experience. I want to thank Ernestasia for the project we did together and for introducing me to her supervisor Judith Redi. Judith soon helped me substantially with my research and eventually became another supervisor of mine. While for a moment I was worried that the very different "styles" of Pablo and Judith would clash with me being in the middle, it turned out that they were perfectly complementing each other. Large parts of my research were carried out in the EU FP7-Project VConect, in this way I would like to thank the European Commission for making my research possible and enabling me to work together with so many excellent researchers and institutions throughout Europe. I especially would like to thank Ian Kegel, Peter Hughes, Nikolaus Färber, Marian Ursu, Erik Geelhoed, Yaroslav Kryvyi and Manolis Falelakis. During my PhD I had the opportunity to work in International Telecom Union (ITU) within the study group 13, which handles QoE, on the effects of delay. This gave me another point of view to the telecommunication world, the standardization organization. Here I would like to especially thank Gunilla Berndtsson as well as Janto Skowronek.

I also would like to thank the reading committee A van Halteren, Ian Kegel, Max Mühlhäuser, Vanessa Evers and Alexander Raake without whose effort this thesis would not have been possible. In the daily life, one of the most important things is always with whom you actually spend your work day with. I had the great luck of having throughout kind colleagues at the Distributed and Interactive Sys-

tems group, where Jack Jansen, Sergio Cabrero, Steven Pomperton, Fons Kujik, Chen Wang, Kees Blom and Rufael Merkuria. After Pablo, Simon Gunkel is probably the person I worked with the closest, we conducted many studies in the beginning together and I would consider him my closest friend in Amsterdam. While our topics did not overlap very much I enjoyed the fruitful discussions with Thomas Röggola and Jan Willen Kleinrouwler.

Besides the people you meet in your daily life, friends and family are the ones that bring you through tough times and that you can share the good times with. Here special thanks go to fellow PhD "Leidensgenossen", even though, they were working on ever so different topics from me, I could share and discuss the PhD experience with Christoph, Linda and Christina on many occasions. A would like to especially thank here also my friend Mila, who was always a friend during this time. Even though she was physically far, no matter where and what, I could always count on her. Throughout this whole time, I was supported by Ester, with whom I shared my whole experience of life, the good ones and the bad ones. A special thanks for proof reading my thesis and giving me a push when I needed it. I was glad, to always be sure of the support of my siblings Eliu, Luzie, Bill and Tobi. Very fundamental was my brother Bill, who helped me move houses several times during the period thesis was written on. Moreover, he posed together with Eileen for the cover of the thesis. In that regard I also want to thank my sister Luzie and my brother Tobi for the crazy fun remote photo session for my cover. Most fundamental, the strongest "core" support was always that I knew, no matter what would happen, I could count on my father Olaf and my mother Tina.

Abstract

In telecommunication systems the final and most important evaluation is that of the end user. Subjective studies in the area have shown that the opinion of users under the same technical conditions is highly diverse. While rigorous guidelines and narrowly setup studies can alleviate the diversity, results still clearly show that users have different experiences under the same technical conditions. Thus, models that accurately estimate the experience of users in telecommunication systems need to include not only the technical conditions, but also the aspects of the user and the context, in which the telecommunication is used.

This thesis approaches the challenge with the concept of Personal Quality of Experience (QoE). The goal changes from determining an average quality opinion of a specific system configuration, to estimating the experience of a specific user in a specific situation during usage of a system. This Personal QoE should reflect how well the system delivers the service it is providing, taking into account the situational context, the user's behavior and the user's individual traits. Specifically, this thesis, explores within the case of multi-party desktop video-conferencing how different contextual factors, the user's behavior and characteristics can be incorporated in user evaluations, as well as how the accuracy of the results can be improved with these factors.

Within this multi-party setup the thesis explores how the context of multi-party video-conferencing constructs a different quality perspective for each participant. The individual audiovisual streams each participant receives are dependent on the network connection between each participant and can thus be different on each receivers' site. In turn, each user is presented with a different composition of media qualities. This thesis shows that this results in a "contrast effect": lower video qualities are perceived worse the more good video qualities are co-present and vice versa, good video qualities are perceived better the more low video qualities are present.

A further aspect about multi-party conversations is that the conversation dynamics are more complex than in the dyadic case. While in the dyadic case the roles of speaker and listener are symmetrical in group conversations some individuals take often a much more active

role than others. This thesis studies these roles with different levels of delay introduced in the video-conferencing system, which is known to interfere in remote conversations. The results show clearly that more active participants are more impacted by the delay than less active interlocutors, thus showing that user behavior is a factor influencing QoE. The thesis explored user engagement as part of the user factors. Engagement describes the user's current state of involvement with the ongoing conversation and their focus on the task at hand. On the one hand, an engaged user is more concentrated on the activity, whereby they can more easily notice quality changes in the medium that conducts this activity. On the other hand, engagement goes along with a reduced awareness of aspects not directly concerned with the activity. Therefore, the user's perception of quality details could also be reduced. In a study manipulating the video-quality, next to quality ratings also engagement was assessed, it was clearly shown that users who reported a higher engagement, also reported a higher perceived quality. Finally, this thesis showed, with the help of a variance component analysis employing mixed effect models, that building partly individual models for users and groups more than doubled the accuracy of the models. As classical statistical models are unpractical to approach a multitude of features, the Lasso feature selection algorithm was used to construct different models. The models combining up to 12 features, i.e. the system factors and a selection of user and interaction factors, could achieve a performance close to the variance explained with mixed effect models.

Contents

1 Introduction

Since the early beginnings of real-time tele-communication, researchers have conducted subjective studies to determine how users experience the quality of the system (e.g. [29, 90, 116]). The Mean Opinion Score (MOS), an average of user ratings on a standardized scale, is the de-facto standard to quantify the users' perceived quality. However, many studies have shown that user opinions can be highly diverse even though the same technical conditions are set up. Even small changes in the setup of a study can have a strong impact on the outcome [102, 59, 85]. This makes it even more difficult to conduct these kinds of studies and achieve clear results. In order to tackle this problem, extensive guidelines and recommendations for conducting user studies have been drawn up by the community and standardized by International Telecom Union (ITU) (e.g. [72, 73, 69, 71]). Although such standards are necessary to allow comparison among studies, ensure consistent quality between research laboratories and lead to homogeneous study results, they do not address the challenge related to the fact that, in real world usage, different users will have different opinions about the quality under the same technical conditions. Thus, it is very hard for system providers to optimize the quality with ratings obtained from user studies.

In this thesis, we approach this challenge by applying the concept of personal Quality of Experience (QoE). The goal changes from determining an average quality opinion of a specific system configuration, to estimating the experience of a specific user in a specific situation during use of a system. This personal QoE is meant reflect how well the system delivers the service it is providing, taking additionally into account non-system factors, meaning the context specific to the situation, the user behavior and the user's individual traits. To be able to achieve this goal we need to understand how these various non-system factors influence the impact that system factors have on the experience of the user. In turn, user studies, which normally consist of system factors as input factors and user ratings as output factors, need to be extended and incorporate user and usage factors. Taking into account these non-system factors make user studies more challenging, as it is not possible to control non-system factors in the experiment setup and it is often difficult to operationalize and quantify them.

The specific case to be studied to be in this thesis is multi-party

desktop video-conferencing. The thesis will explore how different contextual factors, user behavior and user characteristics can be incorporated in user evaluations, as well as how the accuracy of the results can be improved by using at these factors. Finally, the thesis will address and demonstrate how models for estimating the personal QoE of users can be build.

In the last decade, video-conferencing has established itself as one of the standard tools for tele-communication. Video-mediated group communication has increasingly become part of our daily lives[1]–we use it to catch up with family overseas, do a job interview or watch the latest football match with far away friends. Current Internet connections deliver sufficient bandwidth that allow current end user devices to run group video-conferencing sessions.

However, bad quality is still a reality. Long delays and pixelated videos constantly interrupt our remote conversations. These disturbances are caused by network fluctuations and temporary bottlenecks in the Internet. Video quality is steadily increasing, for example with higher resolutions, and so is the needed bandwidth. Moreover, in recent years, video traffic has constantly increased the share of bandwidth used[2] and desktop multi-party video-conferencing has been the fastest growing area. The current approach for delivering high quality to customers is to push the maximum quality that is possible under the current network conditions. With this approach, substantial resources are spent on minuscule quality differences that users will not even notice. To manage the amount of traffic in a sustainable and ecological manner, we need to change the delivery scheme from the maximum possible quality to the maximum needed quality. But how do we determine what the "needed" quality is?

In order to build systems that make smart decisions, not only based on the available resources but also based on the experience they provide, we first need to understand how users experience different qualities. Past user evaluations of tele-communication technologies, as we discuss in detail in section 2.4, were designed to evaluate telephone connections and are not suitable for video-conferencing services. These tests assess a MOS for a given connection, which could

[1] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html

[2] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html

be used to plan and build a telephone network. However, these assessments do not provide the fine-grained knowledge necessary to optimize ongoing remote conversations. To perform real-time optimizations of specific video-conferencing sessions, it is insufficient to work with averages for all usage situations and user groups. The optimization might improve the experience for one group of users while at the same time deteriorate the experience of another group of users. To apply real-time optimizations, we need to gain a deep understanding of which user characteristics, usage situation or behavior lead to different experiences. This thesis explores how users experience the fluctuating quality of modern group video-conferencing and how we can provide the necessary fine-grained results needed to improve modern video-conferencing systems.

To better understand what users are experiencing we should have a closer look at the video-conferencing case. Remote conversations are the basis of virtually all activities conducted via video-conferencing. The purposes and goals for holding conversations are manifold and unlimited, people might want to discuss a point of view, clarify and solve problems, build and maintain social relationships, or simply relax and have fun. Whether we have a good or bad experience ultimately depends on how the conversation develops, depending on its purpose and fulfillment. However, there are more factors, at a secondary level, that influence our experience, such as: how comfortable we are in the environment, the mood in which we start the conversation, how smooth the interaction goes and, in case of video-mediated conversations, how well the video-conferencing system works. From the perspective of building, running and evaluating video-conferencing systems or services, we are not directly concerned with the actual purpose and course of a conversation from the perspective of meaning and content. We are concerned with enabling the conversation, and thus the whole experience, by providing the system part.

As we have established, the content- and meaning-driven experience of users is outside of our scope. As system designers and evaluators, we are interested in the quality of the experience that the system enables users to have, therefore only concerned with the role of the system. This can be described as the Quality of Service (QoS) approach: the quality of the complete service is an aggregation of the

quality of the technical parameters of the system [165]. In turn, many user tests for this approach have been designed to focus explicitly on the technical parameter that is tested, but do not take into account the actual usage situation. Such tests reveal the boundaries of human perception but tell us little about the participants' experience. This approach showed some shortcomings. For example, while technical improvements went sometimes unnoticed by users, services that provided clearly worse quality were preferred by users and evaluations showed high diversity in users opinions [102].

Past research has evaluated the experience users have when using tele-communication systems. As mentioned before, these studies have shown that different users have a different experience under the same system conditions. Thus, the challenge in determining the impact of system parameters does not lie in designing a better testing scenario or providing more detailed instructions to participants to achieve that all ratings converge. The challenge remains in acknowledging that different users actually have a different experience. Therefore, rather than assessing the MOS of a connection, we should move towards estimating the personal QoE for individual users. The key towards this kind of knowledge lies in understanding the interplay of system factors and non-technical influencing factors. It has been shown (for example with the variance component analysis reported in chapter 6) that the impact of the system factors is strongly dependent on the other influencing factors. Considering the interplay between system and non-technical factors helps to understand the results of evaluations and will eventually allow us to optimize systems or services for the actual ongoing conversation, in the specific situation it takes place and for each participant individually.

The foundation for this approach is the concept of Quality of Experience (QoE), which tries to capture the experience a user has while using a system or service in respect to its expected function [104]. Fig. 1.1 shows the conceptual model that guided the research presented in this thesis. This conceptual model shows the different influencing factors structured by our research approach. System factors are placed on the left side as they provide our starting point and we can accurately control and measure them. On the right side of the model we placed the QoE, as this is the construct we want to investigate and measure. The dark orange arrow from System Factors to

QoE indicates that this is the main relationship we are interested in. Since we have already established that a simple direct relationship between System Factors and QoE is insufficient to quantify QoE in a meaningful way, we placed User Behavior and Factors in the middle, as they mediate the impact of System Factors on QoE. All these elements are embedded in the Context Factors, as the context in which a video-conference takes place will influence user behavior and user factors, as well as the QoE. Part of the system factors, such as the used device, are also part of the context factors.

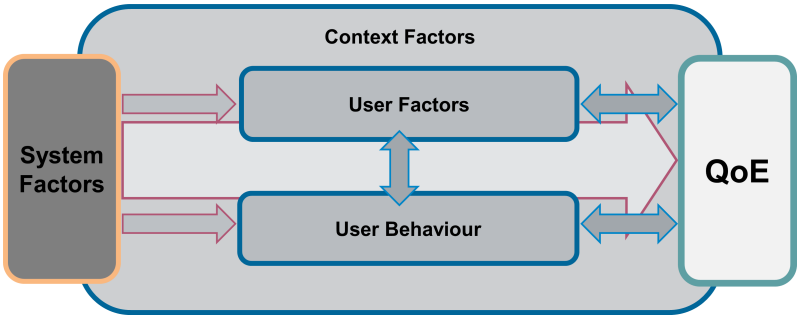


Figure 1.1: Conceptual model of QoE (based on [125, 118, 44, 135])

The need for a new approach of evaluating communication systems is rooted in the new infrastructure for tele-communication, especially video-conferencing. The first telephone services were built on a model of dedicated connection setups. In the beginning, each telephone call would have a dedicated line or assured resources in the infrastructure of the telecom provider. The connection would not change during the call. This way, stable quality was assured, unless influenced by external factors. In this type of ecosystem, evaluating the average quality for a given setup would suffice in order to plan and build a system with a chosen ratio between quality and resources [102].

The video-conferencing solutions adopted by the mass market are over-the-top services, running over the public Internet. In consequence, many components in a video-conferencing solution, such as the network, are out of the control or scope of the video-conferencing provider. As such, video-conferencing solutions are competing with

all existing traffic, resulting in frequent fluctuations in available bandwidth and transmission delay. The providers gain more control over other components, such as the video-conferencing client, in the sense that they can be modified more easily and extended when compared to solutions tightly coupled to the network. Together with this shift in technology, a shift in the conceptual model occurred: while telephone calls are usually paid by the minute, the Internet is "always on" and the video-conferencing service does not result in additional costs. As a logical result, video-conferencing is used for a wide-range of activities: cooking together from distance, watching videos or having small group gatherings. The multi-party scenario is especially resource consuming and prone to quality fluctuations. The available bandwidth has to be divided amongst the connected clients. Moreover, depending on each participant's connection and the employed infrastructure of the service, each participant will have a connection with different quality.

Video-conferencing providers are facing a dramatic rise in complexity of the technology and usage scenarios. On the other hand, the current infrastructure also gives rise to opportunities for optimizing individual connection at runtime. However, the "one-size-fits-all" approach of quality evaluations neither gives insight into the actual experience of participants, nor does it allow the provider to make optimizations.

Insights into the user experience can naturally only be obtained through evaluations with the different users. While there is a large body of research on video-mediated real-time communication (see chapter 2), they provide little insight into the quality aspects of the system or service. Many of the early user evaluations of multi-party video-conferencing did not manipulate technical parameters but conducted comparisons between audiovisual, audio-only and face-to-face meetings (amongst others [147, 163, 114]). They used the turn-taking metric [129], a conversation analysis technique that looks at the timing of utterances in terms of "on-off" speaking patterns to establish that the communication patterns in video-mediated group conferences change due to the introduced delay and the absence of certain cues (such as gaze). Most of the research on QoE in the area of real-time communication is focused on two-party conversations and audio-only connections, leaving us with a gap of knowledge in the

rapidly growing[3] area of multi-party conferencing.

In Fig. 1.2 we have extended the model in Fig. 1.1. We have broken down the factors that research previously addressed, and added the factors that are explored in this thesis. We can see that the majority of previous research has been conducted in the area of two-party studies (in blue) and that multi-party studies were often audio-only conferencing (green dashed).

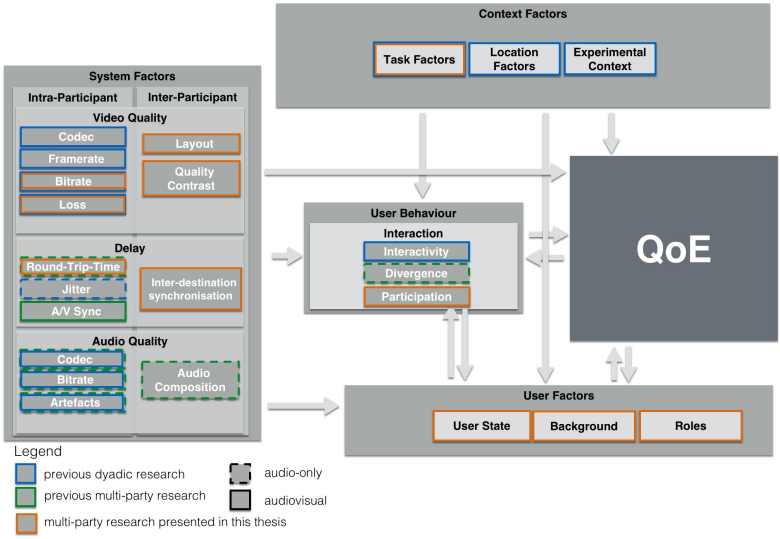


Figure 1.2: Extension of the QoE model in Fig. 1.1 detailing aspects that have been researched in the area of QoE for teleconferencing

1.1 Research Questions

This thesis is concerned with the evaluation of video-conferencing from the end user perspective. Within the larger context of design-

[3] http://blogs.skype.com/2016/01/12/ten-years-of-skype-video-yesterday-today-and-something-new/

ing and developing video-conferencing systems, the ultimate goal will be to improve systems based on the outcome of such evaluations. Desktop video-conferencing is an over-the-top service used in many situations by a large variety of users. The assumptions is that these services can be optimized in real-time by understanding not only the impact of the system factors in general, but also in light of the ongoing situation and current users. The overarching research question can be formulated as **"What is the QoE a particular user has in a video-conferencing session?"**. A multitude of aspects are involved in the factors influencing QoE in multi-party video-conferencing . Therefore, we have to break this question down into subquestions that can be operationalized. Furthermore, the insights we acquire during the evaluations have to be specific and quantitative, so that we can attribute them to specific factors of QoE.

More specifically, we can formulate one subgoal of the overarching question as "What is the impact of a system factor on the QoE in relation to non-technical factors?" In other words, we are interested in the interplay of technical and non-technical factors in QoE.

In order to achieve the required fine-grained control needed to examine such relationships, we opted for controlled laboratory studies as our main methodology. For these studies, we employed a classical experimental design, in which we manipulated a set of factors (e.g. independent variables), and measured the change in other factors (e.g. dependent variables). Non-technical factors, such as the participants' behavior, cannot be controlled like the technical factors of a video-conferencing system. Hence, we treat these factors as co-variates that we measure during the experiment.

We structured our approach based on the three non-technical influencing factors. We first approach an aspect which is novel in the multi-party context: experiencing different video-qualities at the same time. Then we examine the interplay of a system factor and user behavior by looking at delay and conversational roles. Finally, we study the impact of video-quality on QoE in relation to user factors, particularly how we can predict it by including user behavior, background information and the current user state.

1.1.1 Context

In the context of multi-party video-conferencing it is necessary to compose the video stream of each participant into one screen for the overall session. The video quality of each of these streams is dependent on the resources, bandwidth and device capabilities, available at the receiver and sender site. Eventually, all participants may see each other in a different quality, resulting in a composition of different qualities. This unique feature of multi-party video-conferencing is yet unexplored. Meaning that it is currently unknown if and how the differences between the video qualities influence the user experience. This aspect is addressed by the following research question.

- **Research Question 1: What is the effect of the composition of video-streams from different participants in different encoding qualities on the overall perceived video quality?**

1.1.2 User Behavior

Video-conferencing allows us to communicate with other people over distance. The interaction with our communication partners is an integral part of the experience we are having. Video-conferencing, belonging to the field of computer-mediated-communication, is significantly different to other computer systems, since most of the interaction does not happen with the system but through the system. In order to understand how interaction shapes the experience in video-conferencing, we have to look at the dynamics of the ongoing conversation.

Research found that the interactivity (i.e. the pace of a conversation) plays a crucial role in the perception of delay [80, 34, 90]. These works are based on dyadic conversations, which have a different dynamics than small group conversations. Dyadic conversations have a symmetric relationship between the two conversation partners. One participant is always the speaker and the other the addressee. When the speaker changes, the roles are automatically reversed. In small groups, we also have one speaker but multiple listeners, some of whom may not be directly addressed by the current speaker, the "side-listeners". Due to the symmetry in dyadic conversations, both

partners are equally affected by the delay. This does not necessarily hold true for the multi-party situation. If we want to understand the QoE of individuals, we need to examine if the delay affects participants in the same session differently. To advance the knowledge in this aspect, we formulate the following research question.

- **Research Question 2: How does the delay impact the QoE of different participants based on their conversational behavior?**

1.1.3 User Factors

User factors are concerned with background aspects of participants, such as previous experiences with video-conferencing and personal preferences. They also include user state, with factors like the current mood and state of mind of the user. The user state has a reciprocal relationship with QoE, which means that the current user state can influence the quality experience and the other way around.

In this thesis, we are focusing on the user state engagement. In this context we are using engagement [112] to describe the user's current state of involvement with the ongoing conversation. Engagement is closely related to the concept of immersion [97]. Both concepts describe the user's concentration on the task at hand and the awareness of the surrounding environment and time is reduced. On the one hand, an engaged user is more concentrated on the activity, which makes it more likely that the user quickly notices quality changes of the medium that conducts this activity. On the other hand, engagement goes along with reduced awareness of aspects not directly concerned with the activity. Therefore, the user's perception of quality details may be lowered. We use the following research question to guide the research into the engagement aspect of the user state in relation to QoE.

- **Research Question 3: Is the QoE of participants related to their engagement?**

1.2 Contributions

This thesis compiles the results from several experiments that were conducted with the goal of answering the overall research question. Fig. 1.3 provides an overview of the individual factors examined in this thesis and color-coded according to the research questions they address.

1.2.1 Overall Contribution

This thesis contributes to answer to the overall question **"What is the experience of a particular in an ongoing video-conferencing session?"** by advancing the knowledge in the area of the individual experience of users in multi-party video-conferencing sessions.

This thesis fills the gap of research knowledge for multi-party aspects in video-conferencing needed to build QoE estimation models. With a large body of knowledge available on multi-party audio conferencing, this thesis provides insights related to video-quality. It quantifies the influence on QoE that originates from composing different video qualities in the same session. It provides QoE thresholds for video-quality with current HD capable systems.

This thesis further provides the necessary insights to building models that are more accurate by focusing on the individual. The results presented in this thesis show that non-system factors have a higher influence on QoE of participants than system factors. A detailed analysis shows that the impact of delay is significantly different for participants in the same session based on their conversational role. It further shows that more engaged participants report a significantly better QoE than less engaged participants under the same conditions. The thesis paves the way for predicting the individual QoE, for it shows that using a multitude of user and interaction factors can achieve prediction accuracy that significantly exceeds that of models which take only system factors into account.

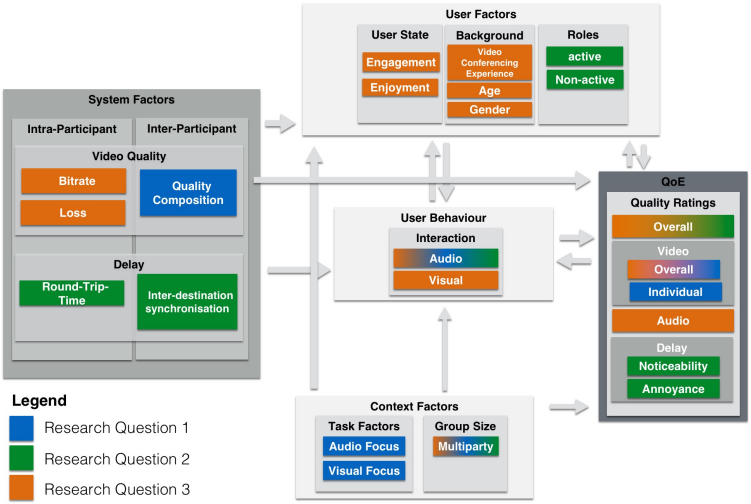


Figure 1.3: Extension of the QoE model in Fig. 1.1 with the specific factors examined in this thesis

1.2.2 Contribution 1: The contrast effect - the composition of different video qualities affects the perceived video quality.

Research question 1 (blue in Fig. 1.3) focuses on whether the composition of the different video stream qualities has an effect on the perceived video quality. To address this question, we conducted an experiment that explored the impact of the system factor *Quality Composition*, which only exists in the context of *multi-party* sessions. In the experiment, participants rated the *overall* and *individual perceived video quality* of two video clips from different *tasks*. The analysis of the results showed that there is a contrast effect for the individual ratings: the ratings for the lower encoded streams were worse the higher encoded streams were co-presented, and vice versa. The higher encoded stream was rated higher the lower encoded streams

were presented. The results also suggest that the *overall perceived video quality* was affected. In the experiment, video-clips from two different *tasks* were employed: one *audio focused* and one *visual focused*. The *visual focused task* was constantly rated higher than the *audio focused task*, suggesting that the kind of interaction has an influence on the perception. Similarly, taking into account the *audio interaction* of the video clip allowed us to the improve the precision of the *audio focused task* but showed no effect on the *visual focused task*.

This subject is described in detail in chapter 3 and is based on the article "The contrast effect: QoE of Mixed Video-Qualities at the same time" submitted to the Springer Journal "User and Quality of Experience".

1.2.3 Contribution 2: QoE-TB: A testbed for interactive multi-party QoE studies.

Research question 1 was concerned with perceived video quality and not specifically QoE. Thus, it was feasible to conduct the corresponding experiment in a passive study. In passive experiments participants are presented short video-clips with content related to video-conferencing and asked to rate them. Passive experiments have the advantage that they need less resources and are easily repeatable. The drawback of passive experiments is that they do not include actual interaction and do not provide the same experience as being in a video-conferencing session. In contrast, in interactive experiments, participants use a video-conferencing system and their ratings reflect the QoE in real world scenarios. As the rest of research questions were concerned with interaction and user factors, a passive methodology was not appropriate. Real-time control and measurement of system parameters is an essential task. Initial feasibility studies with commercial video-conferencing systems showed that these systems do not fulfill the requirements for a video-conferencing system to be used for conducting QoE studies. As a result, QoE-TB, a video-conferencing toolset that was designed especially to run user studies, was developed. In the course of this work, a requirement analysis was conducted, in which researchers in the field were asked about their problems and needs with regard to the systems used in these

studies. QoE-TB fulfills these requirements by providing methods for fine-grained control of the system parameters, facilitating the moderator to listen into the conversation without being present in the session as well as to remotely control the clients, script experiment procedures and recording facilities for all transmitted media on the receiver and sender side. Furthermore, it contains extensive capabilities for logging the experiment steps and it is easily extensible for specific experiment tasks. The player and analyzer components allow playback of recorded sessions, speech segmentation and calculation of speech metrics as well as annotation of the conducted experiments.

This contribution is described in detail in chapter **??** and based on the following articles

- "A Quality of Experience Testbed for Socially Aware Video-Mediated Group Communication" presented in 2013 at the Socially Aware Multimedia Workshop of ACM Multimedia
- ITU-T Contributions C135 - "Evaluation of multi-party audiovisual telemeetings" "Requirements for a QoE Testbed for Audiovisual Telemeetings" presented at the "ITU-T StudyGroup 10: QoS, QoE and Performance" Meeting in December 2013
- ITU-T Contributions C222 "Requirements for a QoE Testbed for Audiovisual Telemeetings" presented at the "ITU-T StudyGroup 10: QoS, QoE and Performance" Meeting in September 2014

1.2.4 Contribution 3: Participants experience delay differently depending on their conversational role.

Research question 2 (green in Fig. 1.3) inquires whether participants perceive *delay* differently depending on their involvement in the conversation. In order to answer this question, interactive experiments were conducted in which groups of five participants had an ad-hoc discussion over a video-conferencing system. The discussion was based on a problem-solving team-building exercise. Participants were asked to rate the *overall quality* of the system and specifically indicate how strongly they *noticed* and were *annoyed* by the delay. One

randomly selected participant was assigned the *role* of the discussion moderator. The analysis of the audio data showed that, in reality, the assigned moderator would not always fulfill this role. Clustering the participants by their *audio interaction*, specifically the speaking time, revealed that there were two groups of participants: active participants (mostly one per group, often but not always the moderator) and non-active participants. The two groups had a different perception of the delay. While the QoE of the active participants was significantly impacted at 500ms of added delay, non-active participants perceived this drop at 1000ms. At this level and higher delay levels, the perception of participants was similar. In a setup, in which only one participant got added delay, meaning that the *inter-destination synchronization* was manipulated, it was found that the participants without additional delay also experienced disturbances in the connection.

This contribution is described in detail in chapter 5 and based on

- "Methods for Evaluating MediaSync in Realtime Communication" Chapter in the Springer "Mediasync: Handbook on Multimedia Synchronization"
- ITU-T Recommendation "P.1305 - Effects of delay in telemeetings"
- "The Influence of Interactivity Patterns on the Quality of Experience in Multi-Party Video Mediated Conversation" presented in 2014 at the Socially Aware Multimedia Workshop at ACM Multimedia
- "Asymmetric Delay in Video Mediated Group Discussions" presented in 2014 at QoMEX

1.2.5 Contribution 4: Engagement influences QoE.

Research question 3 (orange in Fig. 1.3) quantifies the impact of user factors on the QoE, specifically *engagement*. The experiment manipulated the *bitrate* of the video encoding and the *packet loss* in the connection and assessed the perception of the *overall*, *video* and *audio quality*. The experiment was conducted with a visually focused

task, in which each participant had to assemble a Lego model with incomplete instructions (each participant got a specific section of the full set of instructions). The results showed that the QoE of participants did not significantly improve after 1024kbit/s and packet loss had a minor, but significant, impact. The statistical analysis showed that while the effects of the system factors were significant, they only accounted for approximately 30% of the variance in the ratings. A performed variance component analysis revealed that similarities within the user (as the experiment was conducted in a repeated measure design) or session (i.e. the group of participants) accounted for approximately 40% of the variance in the ratings. This proved that a large part of the QoE is neither caused by the system factors, nor is it random. To explore the user and interaction factors that play a role in the formation of the QoE, the engagement of participants was assessed with an established engagement questionnaire. The statistical analysis showed that there was a clear relationship between the QoE and the participants' engagement: higher engaged participants also reported a higher QoE. Many of the other measured user and interaction factors showed weak influences on the QoE. As classical statistical models are unpractical to approach a multitude of features, the Lasso feature selection algorithm was used to construct different models. The model combining up to 12 features, i.e. the system factors and a selection of user and interaction factors, achieves a performance close to the variance explained with mixed effect models. A closer analysis of the interaction also showed that worse video quality lead to participants using less *visual interaction* (measured by motion in the video) and relying more on *audio interaction*.

This contribution is detailed in chapter 6 and is based on

- "Towards individual QoE for multi-party video conferencing" published 2018 in IEEE Transactions in Multimedia
- "1Mbit is enough: video quality in multi-party video-conferencing" presented 2016 at QoMEX.

1.3 Structure of the thesis

The remaining part of this thesis is structured in the following way: In chapter 2 we lay out the related work to QoE in multi-party video-

conferencing, showing the foundations on which this thesis is build and the state-of-the-art research in the area. The chapter compiles and extends the related work from the following chapters.

Chapter 3 details the crowdsourcing study which was conducted to assess the "contrast effect". The analysis of the data shows how presenting different video qualities at the same time, as it occurs in multi-party video-conferencing , can emphasize good or bad video qualities of individual streams.

As a foundation for the following chapters, chapter 4, presents QoE-TB, the video-conferencing toolkit build to conduct interactive studies.

Chapter 5 details the relationship between delay and conversational roles. We describe the conducted studies and the data analysis reveals how active participants are more affected than non-active participants.

In chapter 6 we examine the impact of different video-qualities in an interactive study. We show how the video-quality affects QoE and behavior of participants, detail the relationship between quality perception and delay, and present a model for predicting QoE based on a multitude of factors.

Finally, chapter 7 concludes the thesis with a reflection on the impact of this thesis on assessing the individual QoE in multi-party video-conferencing and with a discussion of the future development of video-conferencing systems.

2 Related Work

This chapter lays out the related work to QoE in multi-party video-conferencing. It begins with an introduction into the fundamentals of QoE, describing the approach and conceptual models of QoE. The next section is concerned with background and methodologies for QoE studies in real-time communication, to this end it gives an overview assessment methodologies for QoE and an introduction to research about the interaction in conversations with focus on turn-taking. It follows an overview of the QoE studies related to QoE invideo-conferencing is given, the focus lies here in the in this thesis examined factors delay and video-quality, and research on user factors and QoE.

This chapter compiles and extends the related work from the chapters 3-6.

2.1 Introduction

For a long time the evaluation of media and communication technologies focused on an paradigm called Quality of Service (QoS). The idea behind QoS is that systems or services have (technological) characteristics or parameters that are essential for their functioning. In turn, the quality of these parameters determines the overall quality of a service. The assumption is that a higher QoS always leads to a better (or equal) experience for the user of the service [103]. However, this has sometimes led to the development of higher service quality that would go completely unnoticed by the user, or services that clearly provided worse quality than their competition, but would be preferred by users [103, 22]. It became clear, that QoS did not capture all essential aspects for the actual users of such systems or services. As a reaction to these shortcomings the notion of Quality of Experience (QoE) was developed, which puts the user and his or her perceptions and experiences in the center. After different definitions of QoE, the definition from the whitepaper on QoE [118] is widely accepted, and was adopted by the ITU [68]. The following definition is the adapted version from the Springer book on QoE which adds more context to the definition: *"Quality of experiencing is the degree of delight or annoyance of a person during the process of experiencing. It results from the person's evaluation of the fulfillment of his or her expectations and needs with respect to the utility (pragmatic and hedonic) in the light of the person's context, personality and current state."* [104]

Experiencing in itself is a low-level cognitive process in which stimuli of the environment are perceived and processed in the brain. The judgment process is a higher level cognitive process which requires a more conscious reflection about the experience [122]. For this reflection process, in order to judge the quality of experience that a system or service provides, a human will compare the current experience to his or her expectations and previous experiences. The opinion of a user is thus formed in a two-staged process, in which first a stimuli (i.e. a perceivable aspect of a system or service) is experienced and then, in a second step, judged against internal quality references [118, 153].

When we are experiencing a system or service it is situated in the

larger context of the ongoing activity, the place we are currently at, the mood we are currently in and the goals we have for our current activity. As a result, when we try to measure QoE or aspects of QoE (for example, the video quality of a streaming service), the judgments we gather will, to some degree, reflect more factors than the system or the particular feature we wanted to test. As a result, one of the profound challenges in QoE is to understand how different factors influence the experience of the user. This is especially challenging since is not completely understood which factors are actually influencing the experience in each situation. To address this challenge the community has developed different conceptual models that try to categorize the different influencing factors. The most common model for QoE [118] divides the influencing factors intro three categories: user, context and system influence factors. It has been shown, however, that the model has short-comings when it comes to interactivity and QoE [32]. Therefore, other approaches to better describe the relation between user behavior, user state and QoE have been investigated [125].

The remainder of this chapter is organized as follows. In order to give the reader an overview of the methodologies and approaches employed in this thesis, section 2.2 introduces different assessment methodologies for QoE. To gain a better understanding of the video-conferencing situation, section 2.3 provides an introduction to the interactions that take place in video-conferencing, with a focus on the turn-taking model. Starting from these foundations, section 2.4 presents an overview of research done for assessing QoE in tele-conferencing systems. Special attention is placed in the system factors that this thesis examines in detail: delay (section 2.4.1) and video quality (section 2.4.2). Finally, section 2.5 discusses works that are related to assessing QoE of individuals by laying out the research done in the field of user factors and QoE.

2.2 Subjective Assessment Methodologies

In this section we give an overview of the existing methodologies for conducting QoE user studies. We first give an overview of standards

and guidelines in the area. Then, we introduce the two major categories of subjective assessments: passive and active tests. We explain each approach and discuss their differences. Finally, we detail crowdsourcing, a recent approach for conducting passive studies over the Internet.

In an effort to standardize testing methodologies between different companies, regulatory bodies and research, many of the research methods in the scientific community have been published as ITU recommendations. This includes standards for audio transmission quality [72], conversational quality [73], time varying speech quality [74], regarding audiovisual systems quality [75, 76, 77] and audiovisual quality in telecommunication [78]. The majority of testing methodologies for realtime tele-communication systems and services were developed for two-party scenarios. However, in recent years substantial work on multi-party scenarios has been conducted, resulting, amongst other developments, in an ITU standard for testing methodologies for multi-party tele-meetings [69, 71].

The methodologies employed for assessing perceived quality for realtime tele-communication can be classified in two groups: passive and interactive. Passive tests are conducted by letting users rate the quality of video clips using video-conferencing related content, such as [84, 14, 109]. In contrast, interactive tests use a real video-conferencing setup in which participants interact freely and to varying degrees with each other, such as [141, 14, 50, 136]. Similarly, in the audio domain, listen-only tests and conversation tests are employed [102, p.50ff]. The degree of freedom plays a key role when interpreting the obtained results [63]. Proposed scenarios for interactive tests range from simple number verification tasks, over short scripted scenarios to free conversations [73]. Interactive tests can be used in a general purpose manner, as they use a realistic setup and thus participants experience the test conditions similar to a real world situation. Passive studies provide more consistent results, as they are more easily repeatable than interactive tests, and need less resources. However, they do not provide participants with a realistic experience of the conversation. In turn, passive tests are used to examine one specific aspect or as an initial investigation for previously unexplored aspects. Usually passive and interactive tests correlate with each other, although they can exhibit systematic differences be-

tween them. In a study of listen-only tests and conversation tests it was found that in the listen-only situation, participants rated the quality worse [102, p.129-133]. This is most likely due to the fact that, in the listening-only situation, participants tend to concentrate more on the quality than on the content of the speech. However, often passive tests are used to initially investigate effects which were later confirmed with interactive tests, such as audio-visual quality integration [13], quality perception of a tonal language [24] and improvement of speaker recognition due to spatial audio [121].

In recent years crowdsourcing has become a recurrent methodology for conducting QoE evaluations [56]. In such setups, the test is conducted by participants or *crowdworkers* at home, over a web-platform. These crowdworkers get a small fee for the study, which is usually, like the recruitment, handled by a crowdsourcing provider like Microworkers[1] or Amazon Turk[2]. This methodology has been employed several times for obtaining video quality ratings [23, 3, 57]. Several studies have been conducted to research the methodology, like the influence of video clip length [38], a training phase [42] and fraud detection [58]. These studies have been gathered in recommendation guidelines for QoE assessment in crowdsourcing [55, 56].

This thesis employs passive tests with crowdsourcing methodology to explore the contrasting effect of different video qualities at the same time (chapter 3) and interactive test with free conversations to investigate the relation between QoE, delay and user behavior (chapter 5) and the relation between QoE, video-quality and user factors (chapter 6).

2.3 Interaction in Tele-Conferencing

Video-conferencing is a technology used for having conversations over the distance. To understand the experience participants have during a video-conferencing session it is necessary to understand the conversational interaction. An approach to understand 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 [129]. It describes how we implicitly arrange our conversations by taking turns

[1] www.microworkers.com
[2] www.mturk.com

of connected utterances. An utterance and a turn can be seen as something similar to a word and a sentence in written text (see Fig. 2.1). As a general rule we try to avoid talking at the same time, as it severely hinders understanding. Short overlapping at the end of one turn and the beginning of someone else's turn may occur. This is to some degree culturally dependent, it was found for example to be much more common in Mexican Spanish than in American English [15]. Other overlaps which are intended to occur at the same are the so-called 'back channels', short vocalizations like 'mm-ha', 'okay' or 'yeah' that signal shortly agreement to the speaker and are meant to motivate her or him to continue [131]. In most other cases simultaneous speech is an intentional or unintentional interruption. With an intentional interruption the interrupter tries to take a turn, even though the current speaker has not yielded his or her turn yet. Unintentional interruptions occur most of the time at the beginning of a turn, when two participants start to speak at the same, or nearly the same time[99]. In not formalized conversation settings the organization of turns occurs nearly always implicitly (i.e. without somebody being directly directed to take the next turn). Thus the current speaker takes their turn until they take a pause, often accompanied by other signals, that indicates that they wish to yield the turn to somebody else. An often employed non-verbal signal is to gaze directly at the speaker we expect to speak next [86]. When the current speaker directly prompts another interlocutor to take the next turn we talking about formal handover. Telecommunication inherently alters these pauses due to the delay it introduces. Several studies have investigated the differences in face to face, audio and video-conferencing [114, 147, 115, 172].

This thesis uses different turn-taking metrics (see Fig. 2.1) in the analysis of the interactive studies (sections 5 and 6)

2.4 Evaluation of Video-Conferencing

This section first gives an historical overview of the insights into human factors in video-conferencing and then continues to detail the state-of-the-art for delay and video-quality, the factors under study in this thesis.

Research on the perception and experience during video-

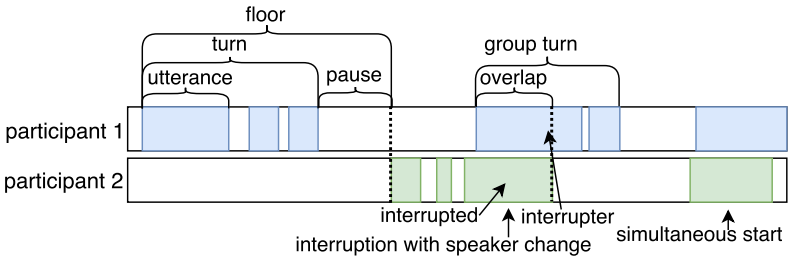


Figure 2.1: Visualization of different speech metrics

conferencing can be drawn from the broad knowledge of audio-only telecommunication, like the telephone or VOIP. Systematic evaluation of quality became of interest in the 90th, when calls were often transferred digitally. The digitalization was more ecological but came with a drop in quality compared to the previously employed systems [102]. The new technologies introduced new distortion sources (e.g. coding and packet loss) and in turn it became more important to understand what still constitutes as a satisfactory or 'good' call quality. During this time systematic work on different quality factors for speech transmission (e.g. bandwidth, frequency, echo, noise, delay) and assessment methodologies was done. The efforts in systematic assessment methodologies led to their standardization in the ITU [72, 73].

Early works regarding the quality of video-conferencing focused on requirements for delay [79] and there was only one study which examined video quality of video-conferencing systems [30]. The majority of the human-factor focus was on understanding the different conversational situations. Several studies examined differences in conversation behavior by comparing face-to-face, audio-and audiovisual conferencing [114, 147, 4, 115] or investigated distributed work aided by video-conferencing [36, 160, 88, 21]. These works established an understanding that the different conversational behavior of participants in tele-conferencing and face-to-face is due to the lack of certain visual cues (e.g. eye contact) and the added delay. Furthermore, it was established that, since the main communication is done over the audio-channel (i.e. speech), high-quality audio is required for a satisfactory conversation [116, 40, 160].

With the addition of a video channel, understanding the combined audiovisual quality became of interest. A study in the area of TV had shown that audio and visual quality are interlinked in quality perception, as degradations in one channel also lead to worse ratings of the other channel [9]. Several works investigated the effect of combined audiovisual quality for video-conferencing [54, 12].

This thesis focuses on exploring the system factors delay and video quality. Research has established that a good audio quality is a necessity for good experience [40, 160, 21, 147], as it is the main information channel, and there is a large body of work on the subject. In the context of video-conferencing, the audio channel takes far less data than the visual channel and it is thus advisable to prioritize the audio channel.

This thesis advances the knowledge about the influnce of system factors in video-conferencing by reassessing the thresholds for perception of delay in multi-party video-conferencing including the first evaluation of asymmetric delays. In doing so, this thesis also provides a novel use of the turn-taking systematic to differentiate the impact delay has on participants in the same session (section 5). It further advances the knowledge about video-quality in multi-party video-conferencing , by approaching the previously unstudied effect of experiencing different video-qualities at the same time (section 3)), establishing the impact of different encoding qualities and their relation to engagement (section 6).

In the following sections, we provide an in-depth overview of related work on delay and video quality.

2.4.1 Effects of Delay

This section introduces aspects related to the perception of delay, the established noticeability thresholds and their relation to conversation dynamics and accustomization to delay. We then detail the findings on the difference between audio-only and audiovisual conferencing. The following sections lay out the research in multi-party conferencing by introducing the early studies which compared remote group and collocated conversations and then provide an overview of the state-of-the-art research on delay and multi-party video-conferencing . Finally, we present findings regarding asymmetric delays between channels (i.e. audio and video-channels) and participants.

Delay is a technical property of the system or service which cannot be perceived directly like audio or video quality. Instead, participants have to infer from the interaction taking place whether there is delay in the connection. Even though technically delay is a physically unavoidable property of tele-communication systems, if it is short enough it can go completely unnoticed by the communication partners. And vice versa, a long enough delay makes some interactions virtually impossible. Thus, in order to develop better system or service , evaluation of delay in video-conferencing is focused on determining thresholds up to which delay either goes unnoticed or becomes unbearable. The corresponding ITU recommendation regarding delay in remote conversations, ITU-T G.114 [67], suggests that 150ms [67] one way delay will go unnoticed by the participants, while around 400ms will result in severe communication problems. This is in line with early works on perception of delay in remote conversations [90]. However more recent works [47, 50, 18, 34] found that many participants still reported acceptable conversational quality with up 600-800ms delay. Two reasons are believed to explain these differences [34]: Firstly, a key influencing factor for the impact of delay is the interactivity of the application (e.g. the pace of the conversation). To this end many, studies employed turn-taking metrics (see section 2.3) to qualify their results [90, 50, 63, 34, 80, 143]. Important operationalizations of turn-taking metrics are the speaker alternation rate [80], a 'speech temperature' metric based on speaker alternation rate, pauses and double talk [50], an unintended interruption rate [34] and a divergence metric to compute the difference in temporal realities of participants [143]. Secondly, while delay may go unnoticed, if participants become aware of delay in the connection, they will often adapt their behavior and over time get used to the delay. With the proliferation of IP based communication, which is more likely to have delay, current users could be more used to delay than the participants in early 1990s [34]. In this context it is also important to notice, that it is possible that conversational problems due to a long delay are not associated to the system. A study [142] found that if not informed about the delay, participants would attribute added delay to characteristics of their conversational partners.

The majority of studies investigating delay in remote conversations investigated audio-only conferencing [90, 50, 63, 34, 123, 80,

151, 47], while fewer works have focused on the video-conferencing scenario [79, 108, 18, 171, 159, 22, 130]. There is not much work that specifically addresses the differences between audio-only and audio-visual applications, however the outcome of the studies suggests that the delay thresholds are roughly the same for both scenarios [70]. Yet, one study comparing audio-only and audiovisual settings found that more delay is tolerated when the video-channel is present [159].

While the majority of studies investigated dyadic scenarios [90, 50, 63, 34, 159, 123, 22, 80, 151, 47, 79, 108, 18, 171, 159, 22, 130] works on evaluating multi-party systems began already in the early 1990s. The early multi-party conferencing system were mostly evaluated by comparing them to face-to-face settings [114, 147, 4, 115, 173]. As these studies wanted to investigate how the conversational situation between the two settings differs, many of these studies employed turn-taking metrics [114, 147, 4, 115]. These studies did usually not compare a number of different delay conditions (as studies investigation thresholds do), but compared the delay the evaluated system delivers under optimal conditions against the face-to-face situation. While each study reported a number of differences between the face-to-face and the remote condition, it is difficult to clearly establish what these differences are. The main reason is that participants in a conversation will adapt their behavior according to the conversational situation, including technical parameters like delay. For example, the introduction of delay and absence of other cues used for turn-taking negotiation, should clearly result in more unintentional interruptions. Noticing these problems, conversation partners will adopt a slower pace and more often formally hand over their turn to an interlocutor [147]. Systematic evaluation of delay in group conferencing began in recent years for audio conferencing [154, 152, 143, 156] and video conferencing [81, 14, 137, 138, 141, 43]. Recently the ITU made a recommendation regarding delay in tele-meetings [70]. Compared to the results of dyadic studies, the results of multi-party delay studies suggest that the perception of delay is more relaxed in the multi-party case, as compared to the dyadic situation [137]. The most likely reason is that participants in the multi-party situation are not always directly involved (either as speaker or addressee) but can also take the role of 'side-listener' [53].

Besides transmission delay, also synchronization aspects (i.e.

asymmetric delays) have been investigated. One aspect is the synchronization of audio and video channel (so called 'lip sync'). Early studies investigated this aspect for video-watching scenarios [158] and the ITU established a standard for Television [65], which recommends a maximum of 90ms audio leading or 185ms video leading. Also studies regarding video-conferencing [141, 14] suggest that it is preferable to have a leading video channel than the other way around. Furthermore, for the multi-party scenario, researchers have investigated asymmetric delays between participants, showing that having only one participant with additional delay disturbs the conversation [137].

2.4.2 Evaluation of Video Quality

This section details the state-of-the-art findings in the perception of video quality. We first detail the impact of various technical properties (such as resolution, encoding, frame-rate, packet-loss etc.). We then move on to explain the effects that result from time-varying quality (such as the 'duration-neglect' effect). Finally, we present the current works on video-quality in video-conferencing.

The systematic study of perceived video quality became of special interest with the rising popularity of streaming videos over the Internet. While previous broadcasting technologies delivered a stable constant quality, streaming videos were of a more variable quality. Furthermore, the traffic that streaming services generate is directly linked to the video-quality and thus of interest for optimization. In turn, a multitude of system factors related to video quality have been investigated.

Regarding the physical size of video, it was found that viewers generally prefer larger images [10, 96]. In video-conferencing it was found that larger video-size improves the feeling of presence [20]. Higher resolution generally leads to a higher perceived quality [84, 41, 119, 10, 164]. In video-conferencing it did not improve task performance but resulted in higher satisfaction [88]. Current high-end systems are on the perceptual boundary of humans, as the difference between full HD and 4K resolution is hardly perceivable on large screen TVs [164]. Besides the resolution, the employed encoder and used bitrate are mainly responsible for determining the spatial video quality. The relation between perceived quality and used encoding bitrate was examined in many studies [84, 175, 166], often with the

goal of developing objective opinion models which could predict perceived quality based on the bitrate [84, 12, 166, 51, 101]. Several works also compared different encoders against each other [84, 175, 149, 181].

Besides the spatial component, the video quality is also dependent on the temporal component, namely the frame-rate, whose impact has been studied in several works [45, 88, 48, 100]. As spatial and temporal quality together comprise the used bandwidth, the trade-off between both has been investigated [82, 179, 100, 19]. These studies have shown that usually users prefer a higher spatial quality compared to temporal quality. Video streaming is often transported with protocols which do not compensate for packet-loss (e.g. RTP over UTP). In turn the impact of packet-loss on video quality was investigated [106, 166, 149, 5]. When packet-loss occurs, part of a video-frame is missing. In turn it was investigated whether the perceived quality suffers more by playing out the distorted frame or by skipping a frame altogether, thus reducing the framerate [162]. Detailed analyses of packet loss have shown that its impact is highly dependent on the type of packet loss and the motion in the video [46]. Also a trade-off between reducing the bandwidth or distortions from packet-loss was investigated [126]. These results can be used in combination with forward-error-correction, in which part of the bandwidth is allocated to transmit redundant data, reducing the probability of actual information loss due to packet-loss [64, 120].

Findings in the area of time varying quality found the 'recency' effect [37, 2, 52]: When participants are asked to rate the quality of a video with segments in different quality, the last presented quality segment had the strongest impact on the overall perceived quality. Further a 'duration neglect' effect was found: The segment with the worst quality had an over proportional impact independently of its duration [52, 37]. Time-varying quality changes got in recent years more attention since the wide-spread adoption of http DASH [167].

Most works evaluating video quality for video-conferencing have been conducted in dyadic settings [54, 18, 33, 13, 161, 130, 12] and employing the Lego® building blocks task [78]. It should be noted that most of these studies use relatively low resolution video (640x480px) and encoding bitrates (maximum 2Mbit) [12]. In today's scenario, higher resolutions (e.g. 720p) are used for videoconfer-

encing, which require higher encoding bitrate. It is unknown whether the results obtained at lower resolutions are applicable to more recent settings.

2.5 User Factors

This section details the research that has been conducted about the interplay of factors related to the user (such as previous experience, demographics, individual preferences and user state) and QoE.

Contrary to assessing the low level of humans' perceptual capabilities, QoE ratings and even perceptual quality ratings usually exhibit a high amount of diversity (i.e. variance). A reason for this diversity is that, the reflection about quality impressions is strongly dependent on personal preferences and individual expectations.

Diversity in QoE perception due to user and context factors has been addressed in several works [59, 85]. It has been shown that user factors can explain more of the variance in user ratings than the system factors [146]. In the context of video watching experiences, social context and demographic factors [178], as well as personality and culture traits [145] have an impact on QoE. In music domain, several works investigated the overall listening experience, a concept that tries to quantify the enjoyment of a listener including all factors (e.g. song, mood, preferences and quality). In this context various user-based factors have been examined [139, 169] and tried to be separated from the audio quality [140]. Previous experiences are known to influence the perception of future experiences [118]. This effect has been studied for Web QoE [150] in which it has been shown that after experiencing bad quality, participants reported a lower QoE even after the quality was back to normal. In relation to this, age has been shown to play a role in QoE, whereby elderly people report more problems in the usage of mobile phones, show more skepticism towards new technology and have a later adoption rate [92] (albeit differences in usage would often disappear after the elderly got more acquainted with the devices [91].

The interplay between user state and QoE has recently also become of more interest [125]. In the context of video streaming, it has been found that participants who are more interested in the video content have a better QoE given the same system factors [117]. Sim-

ilarly, user engagement has been found to play a role in computer mediated human to human interaction [113]. In the context of video watching, it has been found that users of an error free connection reported higher engagement than users with error [26].

Chapter 6 of this thesis shows that the variance in quality ratings in multi-party video-conferencing can be attributed to systematic differences specific to the individual participants and the group. We further detail the relation between QoE and engagement, and show how a multitude of user and behavior factors can be used for a more accurate prediction of the QoE of individual participants.

3

Perceived Video Quality with different Video Qualities of the Same Time

This Chapter details the crowdsourcing study which was conducted to assess the "contrast effect". The analysis of the gathered data shows how presenting different video qualities at the same time, as it occurs in multiparty video-conferencing , can emphasize good or bad qualities of individual streams and investigates the relationship between the perceived quality of the individual streams and the perceived quality of the whole stream.

This chapter is based on

- "The contrast effect: QoE of Mixed Video-Qualities at the same time" submitted to the Springer Journal "User and Quality of Experience".

3.1 Introduction

In our quest for individual QoE we first take a look at the multi-party situation in general. In the multi-party video-conferencing context, we are confronted with multiple video-streams from different sources. In desktop conferencing using an over-the-top architecture in the public Internet, each participant has a different connection to the Internet and a different route to each other participant. In turn, the video-quality of each stream that composes the overall session differs widely. Eventually each participant sees every other participant in a different video-quality depending on their own bandwidth, the bandwidth of the other participants and the architecture of the system in use. As quality perception is highly dependent on the reference that participants implicitly or explicitly choose for their judgment, the question arises if these direct comparison to other streams influences the perception of quality. In other words, does a contrast effect exist when we experience different video qualities at the same time? As this aspect has not been researched before, we conducted an exploratory study, with a passive methodology to keep parameters as constant as possible and with a crowd-sourcing approach to be able to conduct the study in a larger scale.

Besides the aforementioned aspect of contrasting qualities, this chapter deals with the aggregation of perceived quality form individual streams into an overall judgment. We assume that the QoE a participant has with the current session is, in a yet unknown way, composed by the quality of the individual streams. In conceptual models [153] of the *quality formation process*, for multi-party telemeetings, it is theorized that users aggregate the quality from different participant streams into a single judgment. Nevertheless, currently we know little about how this aggregation process works. We know that when users judge quality, it is a relative process in which the user compares what he or she currently experiences against expectations. There is a multitude of factors involved that influence that expectation, of which previous experienced qualities are one of the main factors, serving as a quality reference against which the user will judge. In the multi-party situation, we can have direct examples of different qualities in the ongoing session. In this context, we are interested to know whether this contrast of different qualities alone

influences the quality judgments and the relation or aggregation between these individual scores and the overall judgment.

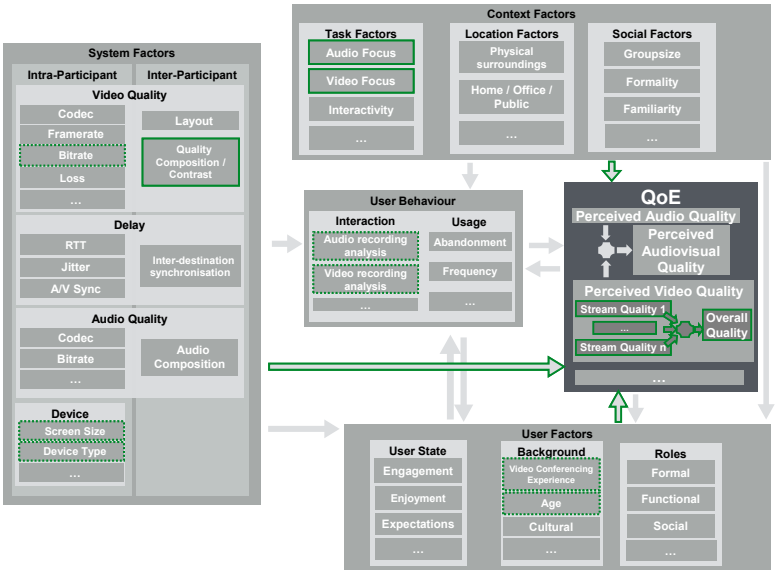


Figure 3.1: Conceptual model of QoE (based on [125]) showing the main influencing factors in multi-party video-conferencing and their relation. The particular factors under study in this work are marked in green and a dotted line indicates that this factor was used as a covariate.

We know from previous works that the perception of video-quality is highly dependent on previously experienced qualities [61, 62, 60]. This effect can further be observed when comparing single-stimulus and dual-stimulus methods. It has been found that dual-stimulus methods, which provide all participants with the same reference frame, yield less variance [174, 95]. The multi-party video-conferencing scenario, with different qualities between participants, adds complexity to the question of how internal reference influences the judgment of quality: different qualities are simultaneously perceived but with different content. The different contents (i.e. the

different video streams) bear many similarities (most of the time all 'head and shoulders' shots) but are still far from the direct perceptual references of dual-stimulus methods. The perceived video quality in multi-party conferencing has only been studied in symmetric quality setups (i.e. the participant perceived each other participant in the same quality) [136, 49, 141]. To our knowledge, QoE in simultaneous mixed quality scenarios, has not been studied in any application scenario.

In this chapter we present a study that investigates the effects of co-present mixed qualities in a multi-party scenario. In Fig. 3.1 we show a conceptual model of the different influencing factors of QoE (based on the QoE and user behavior model [125]) and which factors are taken into account in this work. To contextualize our work, we also add other factors and effects, which are not part of this study. Factors and effects under study have a green border, whereas elements in grey are not considered this time. The model shows that the main system parameters of our study are the individual encoding of the videostreams qualities and their composition. We explore further two well differentiated tasks, which in turn have different speech and video properties. We are interested in the effect that the different screen compositions have on perceived video quality ratings of the individual streams and the overall session, in particular how the individual ratings are aggregated to an overall rating. To conduct this investigation, we presented users with recordings from video conferencing sessions and asked them to rate the video quality. The recordings were taken from two previous interactive laboratory experiments, one of which focused on a conversation, and the other one focused on assembling a Lego model. We encoded each video stream in two different qualities (256kbits and 1024kbits). Due to the large amount of resulting conditions (two video clips with each four participants in all combinations of the two video qualities makes $2*4^2 = 32$ conditions) we opted for a crowd-sourcing approach [55]. The study was split in three campaigns: in one of them we obtained individual ratings only, in another one overall ratings, and in the last one both individual and overall ratings. After we filtered the crowd-sourcing data by reliability criteria (see section 3.2.4 for details) we had ratings from 412 participants giving both kinds of ratings, 178 giving only individual ratings and 180 giving an overall rating. This

resulted in 5,904 ratings that we analyzed to answer the following research questions.

RQ3.1 What is the impact of mixed encoding qualities on the overall perceived quality?

How is the participants' overall impression of the quality of the complete video screen (i.e. containing all 4 streams)? Do low-quality streams have a more severe impact than high-quality streams or the other way around? Our hypothesis is that adding low-quality streams will have a stronger impact on the perceived quality, similar to the influence of bad quality peaks in time varying quality.

RQ3.2 Is the quality perception of an individual stream influenced by the quality of the other streams in the same session?

Our hypothesis is that low-quality streams will be perceived worse when more high-quality streams are co-present, and vice versa, high-quality streams will be perceived better when low-quality streams are co-present, because we are assuming that the streams of other participants will be used as indirect quality references.

Our results show that the ratings obtained from the different campaigns did not significantly differ from each other. The two different video clips from the conversation and the Lego video-conferencing session obtained significantly different ratings. Generally, the Lego clip was rated better and there was less diversity in the ratings among the streams. The overall perceived quality increased the most among an only-low-quality stream composition (i.e. four low quality streams) and a composition with one high-quality stream (i.e. one high quality stream and three low quality streams). For the individual ratings we could observe a contrast effect: lower-quality streams obtained a lower rating based on the number of higher-quality streams co-present - and vice-versa, higher-encoded streams obtained higher ratings based on the number of co-present low-encoded streams. This effect can explain why the difference between only low-quality streams and one high-quality stream in the session is the highest. Comparing the individual ratings of streams with the overall ratings we were

able to see that, except for the case of only low-quality streams, the overall ratings obtained a higher score.

The remainder of this chapter is structured as follows: In section 3.2 we describe the detailed setup of our study and the methodology for the statistical analysis. In section 3.3 we present the analysis of our results and in section 3.4 we discuss how these results can be used for improving predictions in mixed quality video-conferencing scenarios. Finally, we conclude the chapter by showing how our results advance the understanding of perceived video quality for multi-party video conferencing and which steps need to be taken for more accurate QoE predictions.

3.2 Methodology

In this section we lay out the details of the study and the examination of the data we gathered. As we are the first ones to gather individual and overall ratings for different medias, we conducted three campaigns in total: one in which we gathered only the ratings of individual stream, another in which we gathered only the ratings of the session, and one in which both kinds of ratings were gathered at the same time. We start by describing the general design of our study, with the core elements and design decisions. In the next section we provide a detailed explanation of the employed design and procedure. We continue to describe how we prepared the material (i.e. video-clips) for the crowdsourcing this study. We then procede to provide details on the demographics of the crowdworkers and how we filtered our data by reliability criteria before the analysis. Finally, we explain the methodology for the following statistical analysis.

3.2.1 Experiment Design

In previous interactive studies on video quality in video-conferencing we had symmetric quality for all participants (i.e. the streams of participants were treated with exactly the same encoding settings etc). Due to the high amount of possible combinations for asymmetric video quality configurations, interactive studies are not a feasible method for our research questions. Anticipating this, during previous interactive studies, we had asked participants for informed consent

Figure 3.2: Screenshot of the *conversation* video clip with the encodings *hlhl* from upper left (ul) to lower right (lr). Faces blurred for publication.

so that we could use the recorded material in crowdsourcing studies. We selected two 40-second segments from two sessions, which were concerned with different tasks. The length of 40 seconds was chosen as it allows us to keep the study short but still provides enough context for crowdworkers to follow the content [38]. In one clip participants discussed the possibility of using a 'radio device' for rescue when lost at the sea, in the following referred to as the *conversation* task. The task was based on a teambuilding exercise from [17]. In

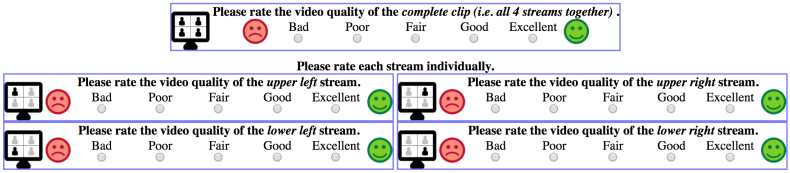


Figure 3.3: Screenshot of the rating scale for the campaign *both*. In the campaigns *overall* only the top question was shown, in the campaign *individual* only the four bottom questions were shown.

the other clip, participants assembled a Lego model, where a small train is nearly finished and the video-conferencing participants are in the process of attaching the chimney, in the following referred to as the *lego* task. The task was based on an ITU recommendation [78]. In this study, we presented each crowdworker once with each of the two clips. Each clip showed 4 streams in a 2x2 layout (see Fig. 3.2), which is the layout also employed in the original interactive study. As the goal of this study was to shed light on the relationship between individual qualities of streams, their composition, and the resulting overall perceived quality, we needed to gather individual ratings of the streams and overall ratings of the session. As we did not know whether assessing these ratings at the same time would influence the results, we ran our study in three different setups, the so-called *campaigns*. The three campaigns were exactly the same except for the amount of ratings we gathered. In the campaign *both* we asked participants for the individual ratings as well as for the overall rating (see Fig. 3.3). In the campaign *overall* we asked only for the overall video quality rating (see Fig. 3.3, upper part). In the campaign *individual* we asked only for the individual ratings (see Fig. 3.3, lower part). A participant could only participate in one campaign and only once.

3.2.2 Preparation of Material

The focus in this study was to investigate the effect of co-presenting different video qualities in the same video-conferencing session. Our original recordings consisted of 4 streams in a 2x2 layout showing one participant each, as it is a common presentation mode in many commercial video-conferencing applications. Thus all participants were presented in the same size. We chose to encode each clip in two different video qualities, *256kbps*, to which we will refer to in the context of our study as *low* quality or in short *l* and with *1024kbps* to which we will refer as *high* quality or in short *h*. The audio was in both videos the same (AAC codec with 10kbps). There were five possibilities of different combinations of the individual stream encodings in one session: the streams can have all the same quality (i.e. all *low* or *high* quality, which we will refer to in a summary notation as *0h4l* or *4h0l* respectively), one stream can be different from the others (one stream *high* quality and the others *low* quality or the other way

Figure 3.4: Workflow for preparing the video material.

around one stream *low* quality and the others *high* quality, in the summary notation *1h3l* or *3h1l* respectively), and two streams *low* and two *high* quality (in summary notation *2h2l*). See also table 3.2 for an overview of this and other factors. However, there are multiple combinations possible to achieve these stream combinations. For example, the combination *2h2l* could be composed by the two upper streams in *high* quality and the two lower streams in *low* quality or the other way around, that is, by having the two upper streams in *low* quality and the the lower streams in *high* quality. To counter balance the effect, we produced and assessed all 16 different possible combinations of streams. In Fig. 3.4 we detail the treatment of the video clips. The original streams had a resolution of 1280x720 pixels encoded in H.264 (the *conversation* streams with 2Mbits the *lego* streams with 4Mbits). The audio was recorded in both cases with the mp3 codec with ca 20kbps per second. We first re-encoded the individual streams with ffmpeg[1]. The four individual streams were then composed to one clip with GStreamer[2] and the final result scaled to 1280x720 pixels and encoded with H264. This results in 16 different *streamcompositions* per videoclip. Each video-stream was always kept at the same position (i.e. the participant who was in the upper left corner was in all configurations in the upper left corner). The screenshot in Fig. 3.2 has an encoding of *hlhl* which is a notation of the short forms of the encodings from upper left to lower right: upper left (ul) stream encoded in high quality, the upper right (ur) stream encoded in low quality, the lower left (ll) stream encoded in high quality and the lower right (lr) stream encoded in high quality. Considering the two different clips we had 32 different stimuli in total.

[1] www.ffmpeg.org
[2] www.gstreamer.org

Table 3.1: Questions and answer options regarding the Internet connection and usage of video services. In cases where we use a shortend label to refer to the option in the following analysis, we added this label in parenthesis after the option.

Question	Options
What is the speed of your Internet connection?	"less than 1Mbit" (slow), "less than 4Mbit" (medium), "less than 12Mbit" (fast), "more than 12Mbit" (ultrafast), "I don't know" (NA)
What type of Internet connection are you using?	"Mobile 3G" (3G), "Mobile 4G" (4G), "DSL" (dsl), "Broadband" (broad), "I don't know" (NA)
How often do you participate in video conferencing / video calls?	"once per day or more", "once per week or more", "once per month or more", "less than once per month", "never"
How often do you watch videos over the internet (e.g. YouTube, Netflix, Facebook or similar)?	"once per day or more", "once per week or more", "once per month or more", "less than once per month", "never"

3.2.3 Procedure

An overview of the procedure is shown in Fig. 3.5. The study consisted of an introduction, gathering of demographical data, a training phase, the two assessments (each with content control questions) and a final page with the information for the crowdworker to get the compensation. On all pages a comment box was displayed so that participants could give feedback. When a potential participant accessed the introduction page we first checked if her or his device met the following requirements: not a mobile device (i.e. tablet or smartphone), a minimum resolution of 1280x720 pixels and a browser able to play html5 videos. Moreover, we checked that the crowdworker had not participated previously in this or one of the other campaigns (checked via the id that the crowdworker provided). If the requirements were not met, the user was redirected to a page explaining that participation was not possible. In the introduction, participants were informed about the purpose of the study and that ratings and

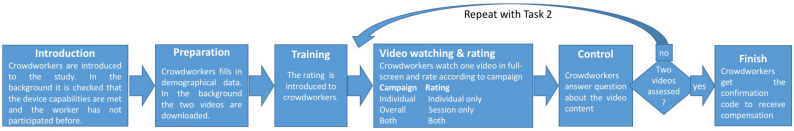


Figure 3.5: The different steps of the crowdsourcing study.

interaction with the page would be saved. In the next step, the participants were asked about demographic information (country of residence, age and gender), questions about the employed machine (laptop, desktop, screen size). We further inquired about their Internet connection (type and speed) and habits about watching videos on the Internet and using video-conferencing (see Table 3.1 for details). During this step the two video clips were completely downloaded in the background. Participants could only move to the next step if all the information was filled in and the videos were completely downloaded. In the training phase participants were shown a screenshot (Fig. 3.2) of the video and a screenshot of the rating scales (Fig. 3.3) with an explanation assuring them that we were only gathering their opinion and there were no right or wrong answers. Furthermore, the crowdworker was informed about the fullscreen mode and the content control questions. On the rating page the videoclip would switch to fullscreen mode once the crowdworker clicked on play. The fullscreen mode ended once the videoclip finished. If the crowdworker ended the fullscreen before the clip had ended, an overlay would appear indicating that the video clip needs to be finished in fullscreen mode in order to complete the study. After the clip had finished playing, the rating scales (see Fig. 3.3) would appear below the video. In the final page, the crowdworker was thanked for participating in the study and the confirmation code, necessary for the crowdsourcing platform, was displayed. The compensation for completing the assessment was 0.35US dollars. Each crowdworker thus rated one randomly chosen clip from each task in random order. The order of the tasks was completely random. For the exact clip chosen for each task, a weighted random choice was implemented to balance the obtained ratings, each clip had a probability of being chosen of *1 - number of ratings for this clip/maximum number of ratings for a clip in this task.*

3.2.4 Participants and Reliability Filtering

The crowdsourcing experiment was conducted over the crowdsourcing platform *Microworkers*[3]. In total 959 crowdworkers finished one of the campaigns, of which 153 did not answer the content questions correctly. We further removed 12 participants because they gave unreasonable ages (e.g. 2 years). In average it took a crowdworker 6.6 minutes to finish the study. We omitted 5 participants which took more than two times the standard deviation longer than the mean duration (sd=6.42min $\rightarrow$ 19.45min) as they likely got distracted with something else during the assessment. We also excluded one participant who reported to be using a smartphone. Furthermore, we employed the reliability filtering suggested by Ribeiro et al. [127] for the campaigns assessing individual and both kinds of ratings. We discarded 31 ratings which had a pearson correlation coefficient smaller than 0.25. For the data from the campaign assessing the overall ratings alone, none of the reliability screenings from Hoßfeld et al. [55] was applicable, as we had only two ratings per subject.

Eventually 739 assessments were left for the statistical analysis. The average age of our participants was 29.4 years (min 18, max 71), 29% of the participants were female, people from 65 different countries participated with the biggest groups being India (20%) and the USA (17%).

3.2.5 Quantitative Analysis

In the analysis we make use of linear regression models [25, p. 161 ff., p.353 ff.] in the form of

$$Y = \beta_0 + \beta_1 X_1 + ...\beta_n X_n + \epsilon$$

We model one dependent variable, the vector Y, through the combination of the independent variables, the vectors $X_1...X_n$ and a random error term ϵ. The coefficients $\beta_0...\beta_n$ are determined in such a way that the sum of squares of the error term is minimized [25, p. 163]. The interaction of two independent variables X_i and X_j (i.e. Y is dependent on the combined state of X_i and X_j) is modeled through

$$Y = \beta_0 + \beta_1 X_1 + ...\beta_i X_i * \beta_j X_j + ...\beta_n X_n + \epsilon$$

[3] www.microworkers.com

Table 3.2 contains an overview of the factors used in the analysis. To assess whether a factor is statistically significant we used a Likelihood Ratio-test (LRT) [25, p. 163] with the factor in question against a model without the factor. A factor was considered statistically significant if the fit of the model with more parameters was better in respect to the added parameters to the model. The null hypothesis is performed with LRT by comparing a model with the factor in question against a model with only an intercept. Because preliminary analyses indicated that our responses had skews or kurtosis in their distribution, we used the bootstrap procedure to obtain the test statistics. The bootstrap procedure makes no assumptions about the population distribution [31]. Given confidence interval are bias corrected and accelerated (BCa) confidence intervals which are more accurate than other estimation methods for skewed data [27]. For the bootstrap we drew on random samples with replacement from the corresponding original data. The LRTs were computed on these bootstrapped datasets and repeated a 1000 times. The resulting bootstrapped statistics were considered significant at $p < 0.05$ when 95% of the computed LRTs were significant at a $p < 0.05$ level. For the performed posthoc tests we bootstrapped a Tukey HSD (with multivariater correction) with 8000 repetitions.

Table 3.2: Factors used in the statistical analysis with used symbol, levels and description

Factor	Symbol	Levels	Description
Independent			
task	T	2 (Lego, conversation)	A video clip from a task related to lego or conversation
stream	SI	4 (ul, ur, ll, lr)	The 4 streams of a clip, upper left=ul, upper right=ur, lower left=ll, lower right=lr
encoding quality	BI	2 (256kbps = low = l, 1024kbps = high = h)	Encoding bitrate of a stream
streams	S	5 (0h4l, 1h3l, 2h2l, 3h1l, 4h0l)	How many high quality and how many low quality streams are in this streamcombination
number of streams	$NS_{h,l}$	0-4 for each encoding quality	Number of low or high quality streams respectively
campaign	C	3 (overall, individual, both)	The three different campaigns
Rating Type	RT	overall rating or mean of individual ratings	Whether the rating was an overall rating or the mean of the individual ratings of this clip
Dependent			
overall rating	RO	5 (bad - excellent)	Rating of the video quality of an entire clip (ITU P.911 [76] 5-point rating scale)
individual rating	RI	5 (bad - excellent)	Rating of an individual stream (ITU P.911 [76] 5-point rating scale)
rating	R	5 (bad - excellent)	Individual and overall ratings

3.3 Results

In this section we present the analysis of the ratings obtained in the crowdsourcing study. The goal is to gain insights about how the perceived video quality is shaped when a session is composed by different video qualities. Thus we ran statistical tests between the ratings users gave and the different combinations of encoding bitrates. Specifically we checked the following:

- Comparison of the different campaigns with different ratings methodologies.
- Analyses of overall (complete video screen) video quality ratings.
- Analyses of individual stream video quality ratings.
- Comparison of overall and individual ratings.
- Analyses of covariates (demographic data).

3.3.1 Campaigns

To gain insight on how the quality perception of individual streams and complete session were related, we gathered both kinds of ratings. However, we were concerned that assessing these ratings at the same time or separately could influence our results. To our knowledge, no previous research on this topic has been conducted so far. To gain insight into this aspect, we conducted three different campaigns: *overall* (obtained only ratings of the overall clip), *individual* (obtained only ratings of the individual streams) and *both* (both ratings at once). The difference between the campaigns for the overall ratings was rather small (see Fig. 3.6). Moreover, also for the individual ratings most streams received similar ratings in both campaigns (see Fig. 3.7). Bootstrapped LRTs confirmed that there was no significant difference for the overall ratings and only two out of the 16 individual streams received significantly different ratings (see table 3.3). We concluded that the different assessment methodologies do not have a significant impact on the ratings. Thus, the following analyses are based on the data from the different campaigns together.

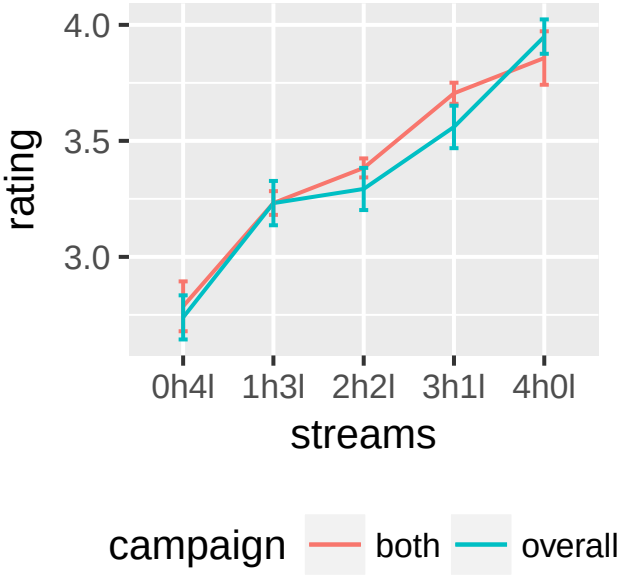


Figure 3.6: Line plot comparing the overall quality ratings from the campaigns 'both' and 'overall'

$$RO = \beta_0 + \beta_1 S + \beta_2 T + \epsilon \tag{3.1}$$
$$RO = \beta_0 + \beta_1 S + \beta_2 T + \beta_3 C + \epsilon \tag{3.2}$$

$$RI_{t,q} = \beta_0 + \beta_1 S_{t,q} + \epsilon \tag{3.3}$$
$$RI_{t,q} = \beta_0 + \beta_1 S_{t,q} + \beta_2 C_{t,q} + \epsilon \tag{3.4}$$

where $t \in T$ and $q \in BI$

3.3.2 Perceived Overall Quality

We wanted to quantify the impact, that changing the individual stream encoding quality has on the perceived quality of the complete screen (overall quality). As expected, a higher combined encoding quality led also to a higher overall perceived quality (see Fig.

Table 3.3: P-values of bootstrapped Likelihood Ratio Tests for the campaigns

Model 1	Model 2	Factor under test	p-value
3.1 3.3	3.2 3.4	campaigns campaigns	>0.05 >0.05 except (t = *conversation*, q = *low quality*, SI = *lr*) <0.05 and (t = *Lego*, q = *high quality*, SI = *ll*) <0.05

3.8). We confirmed with bootstrapped LRTs that *streams* and *task* are both significant factors without an interaction effect (see table 3.4). We continued with a bootstrapped post-hoc test and marked the groups of different conditions in Fig. 3.8 with dotted circles.

It is noticeable that the *Lego* task received constantly higher ratings than the *conversation* task. Furthermore, we can see, in Fig. 3.8, that the impact of going from only low quality streams to one high quality stream (*0h4l* to *1h3l*) had a much stronger impact than the other way around (*4h0l* to *3h1l*).

Table 3.4: Bootstrapped Likelihood Ratio Tests for the response variable *overall quality rating*

Model 1	Model 2	Factor under test	p-value
3.5	3.6	streams	<0.05
3.6	3.1	task	<0.05
3.1	3.7	interaction between streams and task	>0.05

$$RO = \beta + \epsilon \tag{3.5}$$

$$RO = \beta_0 + \beta_1 S + \epsilon \tag{3.6}$$

$$RO = \beta_0 + \beta_1 S + \beta_2 T + \beta_3 S * T + \epsilon \tag{3.7}$$

3.3.3 Perceived Quality of Individual Streams

In this section we are examining how participants rated the quality of individual streams regarding the stream encoding, the task and the composition of the whole screen (i.e. co-presence of other encodings).

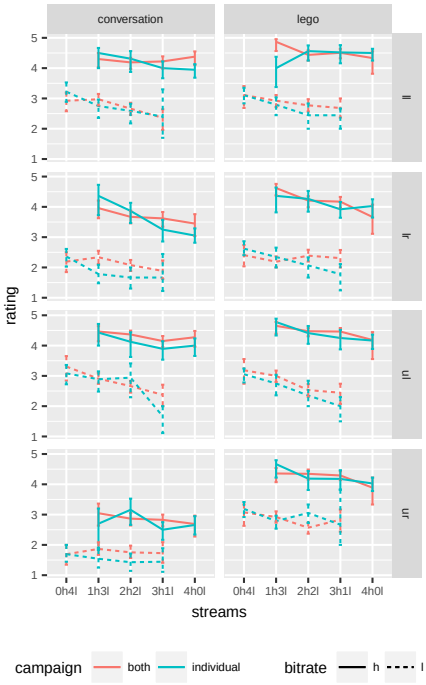


Figure 3.7: Line plot comparing the individual quality ratings from the campaigns 'both' and 'individual'

The quality of *high* and *low* encoded streams was perceived clearly different (see table 3.5) with an average difference of circa 1.5 points between them (see Fig. 3.9a). Like with the overall ratings, there is a statistical and clearly visible difference between the tasks, but no interaction between stream encoding and task (see table 3.5 and Fig. 3.9b respectively). The pattern of the overall ratings is also here present: the *lego* task was generally rated higher than the *conversation* task. We now turn to the effect of the composition of the complete screen, i.e. the co-presence of other encodings, on the quality perception of individual streams. There was a clear trend that low quality encoded streams got rated worse the more they were co-present with other high quality streams and vice versa the high

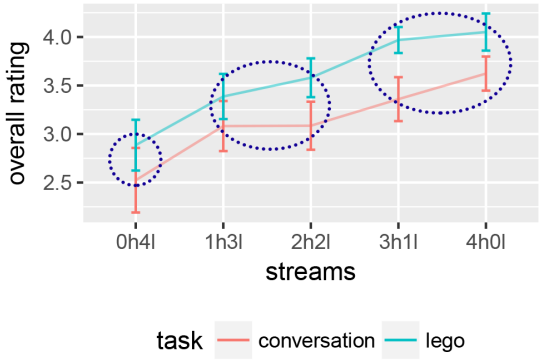
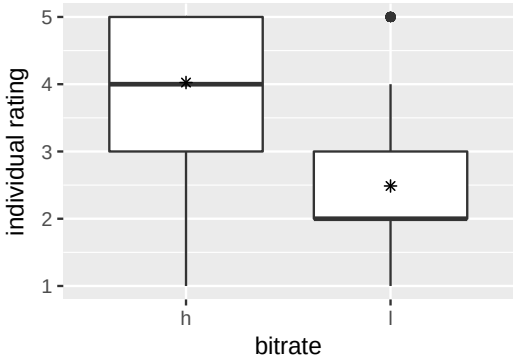
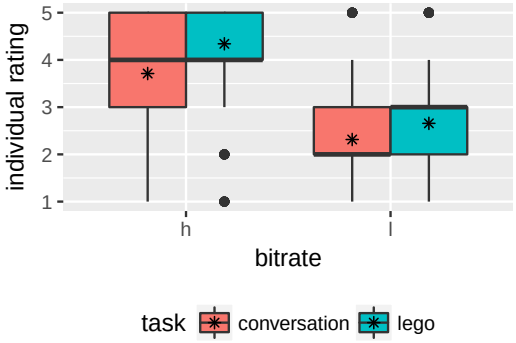


Figure 3.8: Line plot of mean overall ratings by streams and task with 95% CIs as errorbars. Conditions which are significantly different from other conditions are grouped with blue dots.

quality streams got rated better the more low quality streams were co-present (see Fig. 3.10). As indicated by the inverted slopes of both encodings, we statistically confirmed that the number of streams is a significant factor in interaction with the encoding quality (see table 3.5). We continued with a bootstrapped post-hoc test to assess which conditions were significantly different from each other and marked them with dotted circles in Fig. 3.10. For the *high* quality streams there were three groups, while for the *low* quality streams there were only two, indicating that the effect is slightly weaker for the *low* quality streams (see Fig. 3.10). The fact that *low* and *high* quality ratings were also decreasing seems to indicate that the more low-quality streams are present, the better a high-quality stream looks, and vice versa, the more high-quality streams are present, the worse a low-quality streams looks.



(a) Boxplot of the ratings high and low bitrate with mean marked as *



(b) Boxplot of the ratings high and low bitrate by task with mean marked as *

Figure 3.9: Individual stream ratings for high (h = 1024kbps) and low (l = 256kbps) streams

Table 3.5: Bootstrapped Likelihood Ratio Tests for the response variable *individual stream quality rating*

Model 1	Model 2	Factor under test	p-value
3.8	3.9)	stream encoding	**<0.05**
3.9	3.10	task	**<0.05**
3.10	3.11	interaction between stream encoding and task	>0.05
3.10	3.12	interaction number of streams and stream encoding	**<0.05**

$$RI = \beta + \epsilon \tag{3.8}$$
$$RI = \beta_0 + \beta_1 BI + \epsilon \tag{3.9}$$
$$RI = \beta_0 + \beta_1 BI + \beta_2 T + \epsilon \tag{3.10}$$
$$RI = \beta_0 + \beta_1 BI + \beta_2 T + \beta_3 BI * T + \epsilon \tag{3.11}$$
$$RI = \beta_0 + \beta_1 BI + \beta_2 T + \beta_3 NS_h + \beta_4 NS_l + \epsilon \tag{3.12}$$

3.3.4 Overall versus Individual Ratings

In this section we are comparing the ratings of the overall clip and the ratings of the individual streams (the factor *ratingtype*). There is a trend that the overall ratings are higher than the individual ratings (see Fig. 3.11). A bootstrapped LRT, comparing a model with *streams* and *task* against a model with additionally *ratingtype* as explanatory variables, confirmed that the *ratingtype* was a significant factor (p of LRT(3.13, 3.14) <0.05). It is noticeable that there is a significant bump of higher ratings in the *1h3l* case for the overall ratings, while the mean of the individual ratings displays a linear behavior (see Fig. 3.11). The reason is found in the individual differences in quality perception of the individual streams.

$$R = \beta_0 + \beta_1 S + \beta_2 T + \epsilon \tag{3.13}$$
$$R = \beta_0 + \beta_1 S + \beta_2 T + \beta_3 RT + \epsilon \tag{3.14}$$

The contrast effect, described in the previous section 3.3.3, was present for most individual streams (see Fig. 3.12). However for some

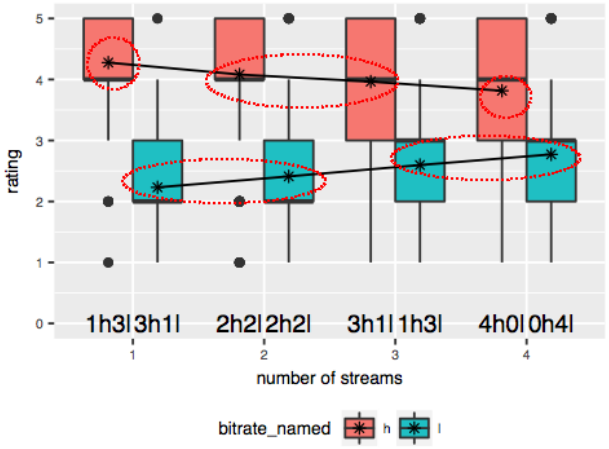


Figure 3.10: Box plots with additional means and a line between them of overall rating per streams. The dotted circles indicate statistical significant contrast groups determined by a bootstrapped post-hoc test. Note that the x-axis shows the number of high or low quality streams in that sessions respectively. Hence marker 1 represents the streamcomposition *3h1l* for low quality streams and *1h3l* for high quality streams, as is additionally indicated at the x-axis. Conditions which are significantly different from other conditions are grouped with blue dots.

streams, nearly no change was visible, for example, the low-encoded upper-right stream of the conversation task (purple dotted line on the left in Fig. 3.12). We can further observe that the streams of the *conversation* task were not only lower rated in average, but also that the variation between streams was much higher than in the *lego* task. This variation also shows that each stream had a different baseline that holds for both encoding bitrates (e.g. the upper right (ur - purple) stream in the *conversation* task and the lower right (lr - green) stream in the *lego* task are the lowest rated streams in both bitrates). Thus we can see that by building simply the mean of the individual scores, in the variation between different qualities, participants and tasks, the contrast effect was not visible anymore. But when we look at the overall streams and the individual streams separated by quality, it is clear that it influenced both the individual as well as the overall quality perception.

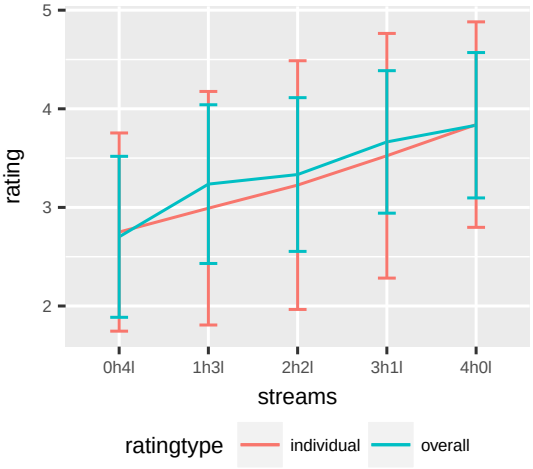


Figure 3.11: Mean of individual ratings and overall ratings with standard deviation as errorbars

To gain insight into why the streams of the different participants were rated so differently, we analyzed the behavior of the participants in the clips. We extracted speech metrics in the form of on-off patterns from the clip and computed the percentage of time participants had spoken in the conversation. A boostrapped LRT for individual quality (3.12 with additionally *percentage of speaking time*) showed that there was an improvement in the fit of the model. We further compared this model against a model including the interaction between *task* and *percentage of speaking time*, which revealed that there was improvement for the *conversation* task, but not for the *Lego* task. For the conversation task we can see that there is a trend of higher ratings with more talking time, while for the lego task no such effect appears (Fig. 3.13). We further extracted the Spatial Activity (also called Spatial Information a measurement of the spatial complexity based on the standard deviation in frames) and Temporal Activity (also called Temporal Information measurement based on the differences between frames) of the videos (see [148, 75]). We added these models to the model for the individual qualities (3.12). A bootstrapped LRT showed that Spatial Activity improved the fit

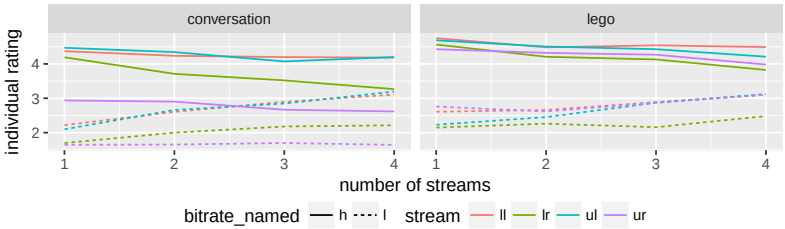


Figure 3.12: Rating of each stream by encoding bitrate and number of streams in the same quality. It should be noted, that each stream is encoded by position (see Table 3.2) which always corresponds to the same participant per task, thus the stream *ul* for the *lego* task does not show the same participant as the *ul* stream of the *conversation* task.

of the model for the conversation task, but not with the Lego task. Temporal Activity did not improve either of the two tasks.

3.3.5 Covariates

We tested whether the gathered background information had an influence on the ratings by using a bootstrapped LRT with the models for overall and individual perceived quality (3.1 and 3.12 respectively) against the model extended by the factor in question. We could not find a significant difference in ratings given by male or female participants (factor **gender**) for either individual or overall perceived quality. For the factor **age** there was a weak effect for the individual quality ratings (3.12), however when checking for influential data points, this effect was due to only two participants over 65, thus we opted for not drawing any conclusion about the relation of age and quality ratings. The kind of **device** participants reported (laptop or desktop) did not have a significant impact on the ratings. However, the **display size** participants reported did have a significant improvement for the models of individual and overall ratings: participants with a larger display gave worse ratings. This roughly follows previous research which found that larger display result in worse ratings [66]. However, when we checked for influential data points, the effect was depending on 12 participants with display sizes of 27 inches

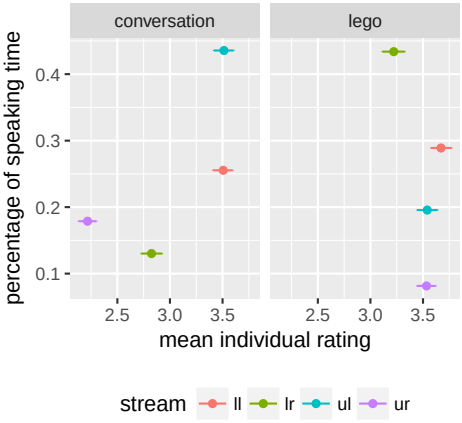


Figure 3.13: Amount of speaking time in percent against average rating (individual ratings) with 95% CIs

or larger. Due to the sparsity of this data we do not draw further conclusions about display size at this point. Moreover, these participants also reported to have fast Internet connections, which is also related to having a better quality.

One of the main factors in determining a participant's perception of quality are his or her previous experiences. However, it is very difficult to assess to which quality participants are accommodated to and what kind of fluctuations they commonly encounter in daily life. Thus, besides asking participants about the frequency in which they watch videos over the Internet and use video-conferencing, we also asked participants about the type and speed of their Internet connection. The assumptions is that the quality of the videos they watch over the Internet is related to their Internet connection.

In fact, neither the frequency of **video-conferencing or Internet video usage** improved the fit of the models 3.1 or 3.12. However, including the type of **Internet connection or the speed participants** reported, both improved the fit of the model 3.1. Participants who reported a better connection gave worse ratings (see in Fig. 3.14 and Fig. 3.15). This supports the theory that a better Internet connection leads to a higher baseline on expectations of video quality.

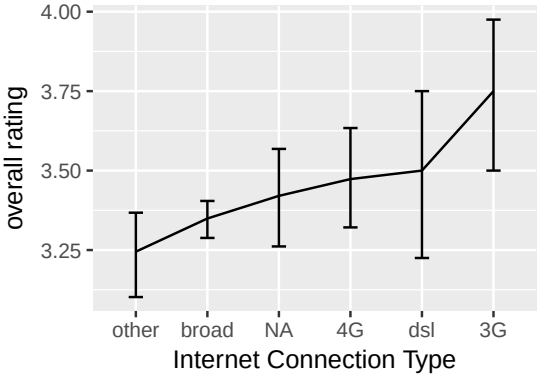


Figure 3.14: Mean of ratings per Internet connection type with 95% confidence intervals.

However, it is not easy to get accurate information from participants (32% reported *NA* or *other* in at least one of the two questions). Moreover, whereas the average of *slow* Internet connections falls out of this pattern, the variance is here also the highest. The worst rating was given by participants who did not know about their Internet connection speed. We further analyzed the time participants took to download the videoclips for the experiment. They were significantly correlated with the reported Internet speed (pearson correlation coefficient of -0.31, i.e. higher reported speed was linked to shorter download times). However, this more objective measurement of Internet speed did not significantly improve the fit of the models.

3.4 Discussion

The main findings from the analysis were:

RQ3.1 The change in the overall perceived video quality from only low-quality streams (*0h4l*) as copared to having one high-quality stream (*1h3l*) was greater than the other way around (from only high quality streams (*4h0l*) to one low quality stream *3h1l*)).

RQ3.2 The individual ratings for high and low quality were affected by

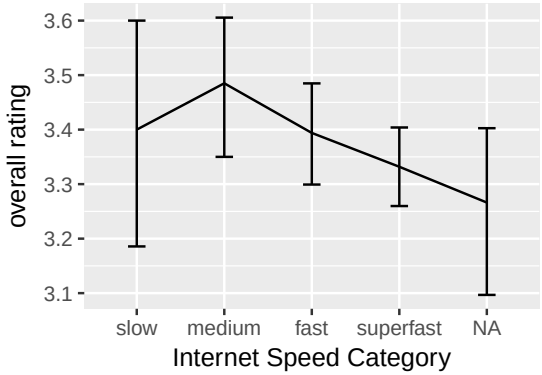


Figure 3.15: Mean of ratings per Internet connection speed with 95% confidence intervals.

the co-presented streams: high-quality streams were perceived better the more low-quality streams were present and vice versa, low-quality streams were perceived as worse when more high-quality streams were present.

From our data, we could conclude that co-presenting different video qualities significantly affects the perceived video quality. It shows that the composition, or co-presentation, of multi-party video-conferencing and the encoding quality are interacting with each other. We will be able to improve the accuracy of QoE estimation models for multi-party video-conferencing by taking such effects into account.

The cases *0h4l* $\rightarrow$ *1h3l* and *4h0l* $\rightarrow$ *3h1l*, when the composition changes from an 'all-the-same' to a mixed quality condition, were of special interest to us. It might have been the case that the break in these setups interrupts the experience so strongly that it is not advisable to actually go to a mixed-quality composition. However, our data shows that this is not the case. While the contrast effect has a significant impact, it is not strong enough to minimize the benefits from having one stream in better quality. This means that we can be sure that if we follow a 'best effort' approach of optimizing each stream individually, a better quality for one stream will never result in a worse overall QoE.

However, the large variation between different streams in the same session, indicates that distributing the available bandwidth between participants can be done best by taking the current interaction into account. Some combinations of three low-quality streams and one high-quality stream (*1h3l*) were rated higher than other combinations of two high quality (*2h2l*). For the *conversation* task, combinations of three low-quality streams and one high-quality stream (*1h3l*) were also higher than three high-quality streams (*3h1l*). In the *conversation* task a large part of the variance could be explained by taking into account how much participants spoke. However, the more visually focussed *lego* task did not follow this pattern. This shows, that we are missing interaction models for cases when the interaction has a different focus than only conversing.

Furthermore, we could consistently observe that the quality of the Lego task was consistently higher rated than the conversation task. The Lego task should be more demanding for the visual quality as the Lego models have small details and are of more interest. Intuitively, we would expect from such properties more critical user ratings. Besides the different content, the video clips had also a different pre-processing, while both clips were encoded in the same manner for this clip, the Lego clips were recorded with 4 Mbps while the Conversation task was only recorded with 2 Mbps. At this moment, the reasons for the different ratings of the two tasks remains unclear.

3.5 Conclusion

In this chapter we presented our exploratory research about how QoE is affected by different video qualities in the same multi-party videoconferencing session. We investigated perceived video quality with a passive crowdsourcing study. By employing different campaigns we established that asking about the perceived quality of individual streams and the overall session quality at the same time does not significantly affect the ratings of crowdworkers. This reduces the effort that has to be made in future studies about mixed quality.

We showed that a contrast effect from presenting different qualities at the same time exists: lower-encoded streams get rated worse the more high-encoded streams are present and vice versa, high-encoded streams are perceived better the more low-encoded streams

are present. Furthermore, we showed that the activity of the session, roles of participants and individual differences between the participants, plays a significant role in determining the final perceived quality. From this we can conclude that a model for estimating the overall QoE in a multiparty session will need to take the screen composition, including the different encodings, into account. Beyond the influence factors analyzed in this work, individual factors, most likely related to the activity and role within the session, are more powerful influencing factors, and will need to be taken into account for accurate estimation models. Even though our study employed a static layout, providing each video-session participant with an equal amount of space, the differences in ratings between them are strong.

Our findings showed that in multi-party video-conferencing a contrast effect on the perception of video quality exists. As an initial investigation on whether such an effect exists, we fixed several factors that differ in real world video conferencing setups and need to be investigated to fully understand the impact of mixed video qualities on QoE. The main steps that need to be taken is to conduct interactive tests and explore further factors and setups. The interactive test in real video-conferencing sessions is needed to confirm whether the contrast effect is perceived when the user is participating in a video conversation. During passive evaluations, as used in this study, it is possible that in their ratings, participants pay no or little attention to the actual content of the material, whereby they detect quality differences that would otherwise go unnoticed [102, p.129-133]. A similar effect could be shown in an interactive study in which participants, who reported a higher engagement in the task, reported also a higher QoE [134]. A further challenge in conducting an interactive study is that usually variance in the ratings is higher. These variances can be accounted for by including moderating factors such as interaction (e.g. speaking time or speaker alternation rate), user state (e.g. engagement or mood) and user aspects (e.g. familiarity with video-conferencing) in the study. However, to include such factors, the study needs to have a large enough sample size, as these factors are usually covariates of a study, as their variance and range is hardly known and controllable. Such an interactive study should further give insights on how the contrast effect is moderated by interaction. Our study showed that in one of the two recorded sessions the speaking

time was a good predictor for the perceived quality. It is our assumption, that if users are participating themselves in the video conference, the role of this moderating factor increases, as they are more engaged in the conversation. Another factor that is substantially different in a interactive study is the length of the video-material. Both video clips used in this study had a length of 40 seconds, which is longer than the often employed 5-15 seconds clips, but much shorter than a typical stimulus length in an interactive test (5-10 min), and thus could have an influence on the results. Other important factors that need to examined are the number of participants and the layout of the videostreams. This would be for one, keeping the layout constant, like in this study, but varying the number of participants. Our study indicated that the strength of the contrast effect, depends on the number of streams in different quality. Inferencing this pattern further would mean that with more streams, a higher contrast is possible. However, the weight of this stream for the overall quality would be reduced. This would mean the individual perceived quality of a single stream is more strongly affected but it might not show a stronger effect for the overall perceived quality of the session. Furthermore, it is possible that a stream gets more attention depending on its position in the layout. The upper left position would–at least in the western society–be a natural starting point as documents start in that position. Our study showed large differences between the streams, however these can also be due to participant shown in this stream as in our case each position was always linked to the same participant. Using a dynamic layout (e.g. the 'speaker-big, thumbnails for others' like for example Google Hangout employs) on the other hand provides substantial changes in the perception of the contrast from the spatial to the temporal domain. As they are not presented at the same time, the user cannot make a simultaneous comparison of both qualities. However the contrast should be stronger, as they are presented in a larger part of the screen. If a significant difference in the quality perception between these two methods exist, this could guide layout decisions for video-conferencing systems.

4

QoE-TB: A tool for conducting subjective QoE studies

This chapter presents QoE-TB, the video-conferencing toolkit build to conduct the interactive studies. We present the requirements gathered for a video-conferencing toolkit aimed at conducting subjective studies. Then QoE-TB is introduced in its design and implementation of QoE-TB, detailing the main components: a multi-party video-conferencing client, an ObserverControl component and Player and Analyzer Toolset.

This chapter is based on the articles

- "A Quality of Experience Testbed for Socially Aware Video-Mediated Group Communication" presented in 2013 at the Socially Aware Multimedia Workshop of ACM Multimedia
- ITU-T Contributions C135 - "Evaluation of multi-party audiovisual telemeetings" "Requirements for a QoE Testbed for Audiovisual Telemeetings" presented at the "ITU-T StudyGroup 10: QoS, QoE and Performance" Meeting in December 2013
- ITU-T Contributions C222 "Requirements for a QoE Testbed for Audiovisual Telemeetings" presented at the "ITU-T StudyGroup 10: QoS, QoE and Performance" Meeting in September 2014

4.1 Introduction

If we want to understand the QoE that an individual has in a multi-party video-conferencing sessionwe need to understand the impact of system factors on the individual. Such knowledge is obtained through extensive user trials under diverse, but controlled, conditions. We further need to be able to record the ongoing interaction in a complete manner. While technological advancements have lead to a wide range of available communication solutions, subjective evaluations that assess the quality of communication are sparse. Assessing QoE requires conducting subjective tests for different and varied communication conditions, which need an infrastructure with some particular features: controllable, recordable, extensible, and dynamic. Unfortunately, none of the publicly available solutions provide the flexibility and level of control, which is required to extensively investigate the influence of network and media parameters on the QoE. We investigated how such experiments can be done with Google's Hangout but we ran into several problems. The control and manipulation of the technical aspects are only indirectly possible through simulating network conditions. If we are to investigate asymmetric network conditions this requires an extensive infrastructure. Monitoring the experiment sessions becomes also problematic. In standard video-conferencing software, the experiment conductor cannot be hidden, which influences the trial. Solutions for recording the media streams in the original and degraded version are either accompanied by quality reduction (which does not allow reasoning about the original perceived quality), or require expensive specialized hardware.

The here described solution, QoE-TB, tries to fill a current gap: the lack of an adequate testbed for controlled experiments, which allows obtaining conclusive results regarding QoE in video-mediated group communication. By offering an end-to-end solution where the conditions can be controlled and manipulated, the testbed aims at facilitating the execution of such subjective tests for specific conditions.

The remainder of this chapter is structured in the following way: Section 4.2 describes the requirements we derived to design QoE-TB and section 4.3 describes the design and implementation of QoE-TB. We then discuss the implications which follow from the design

decisions in section 4.4 and finally conclude this chapter with section 4.5, in which we reflect about QoE-TB usage during the course of this thesis.

4.2 Requirements

In chapter 2 we detailed conceptual models of QoE and how QoE is parted into the following different aspects: influencing factors, user behavior/interaction and the resulting QoE. In Fig. 4.1 we show how these aspects map to the different phases of our experiment. The influencing factors become our independent variables of the experiment. In order to gain different experiment conditions, we want to manipulate one or more of these variables. The other factors, which we believe to have an impact, we try to keep as constant as possible or to monitor as good as possible. For the user and context variables this is achieved through the experiment design, but we can monitor the system factors to control whether they were actually constant. As we need the actual interaction of the experiments for the detailed analysis later on, these have to be recorded during the experiment session. Finally we measure, through assessment, the impact of our independent variables on the dependent variables, the QoE of the participant. All these aspects provide us data for the analysis of the experiment. The Independent variables give us the scope of what we are investigating, while the dependent variables give us the impact on the QoE. The recorded session allows for a qualitative analysis to understand our data better.

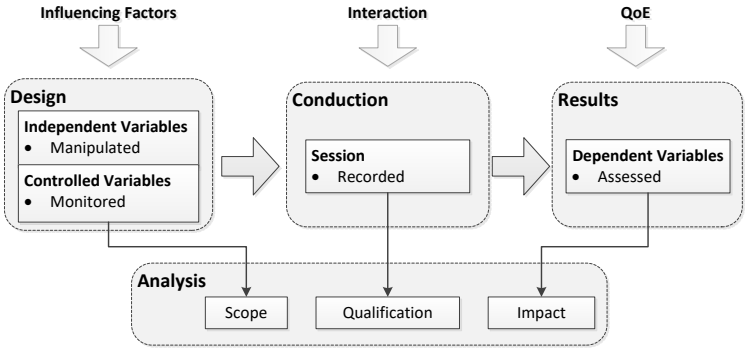


Figure 4.1: Conceptual model to experiment mapping

Based on this process we derived the requirements for our testbed. We distinguish between must-have and optional requirements. The former are necessary to conduct experiments, while the latter are not necessary, but often make the experiments practically more feasible.

R1. Direct application layer QoS parameter manipulation in real-time and for each participant individually

As we detailed in the previous section there is a complex interplay of the system factors. Therefore, we need to be able to control the parameters which directly influence the QoE. Modern communication networks have a fluctuating performance, so we need to be able to modify these characteristics during the runtime of the experiment. And to investigate the asymmetric connections of different participants we need to be able to manipulate them individually for each participant.

R2. Monitoring of the application layer QoS parameters

To assure that the controlled parameters are constant we need to monitor the system status. Even in controlled environments it is possible that fluctuations occur, for example in the delay which is affected by the whole media processing pipeline.

R3. Recording of the transmitted and received media from each participant's perspective

If we are to analyze the actual interaction between the participants, we need to record the media in the original version at the transmitter, as well as the degraded version at the receiver side. To make asymmetric investigations we need to record the media from each individual participant.

R4. Optional: Integration capabilities for activities

Integration of the activity into the testbed has several benefits. First, measurements of the task scores can be easily synchronized and processed with the other data obtained from the experiment. Second, using external components (e.g. a pen-and-paper version) requires more context switch and can lead to a different interaction.

R5. Optional: Questionnaires for subjective assessment

We tested paper-based, web-based and into-the-testbed integrated questionnaires. Integrating the questionnaires into the testbed has many advantages. The questionnaire can be made an intrinsic part of the experiment, so that completing the questionnaire can trigger the next step of the experiment. Furthermore, it is easy to dynamically integrate aspects of the session at hand, e.g. questions about specific participants or based on the completion of the task.

R6. The testbed should facilitate the experiment conductor with a live monitoring capability with the possibility to interact with the participants

Live monitoring is essential, so that problems that might arise during the experiments can be identified. For example, in our trials we had a failing microphone or a case in which a screensaver caused confusion. In other cases, the questionnaires had to be clarified. Also the activity of the experiment can take an unplanned course and interaction is needed. While this is not strictly necessary to conduct experiments, it highly reduces the risk of failed experiment sessions.

R7. Optional: Experiment Progress integration

Integration of the experiment status (set parameters, conditions, conducted questionnaires) assures the development of the

experiment.

4.3 QoE-TB

In this section we present our developed testbed. We first give an overview of the different components and employed technologies. We then describe each component in more detail while explaining how we could satisfy the requirements from section 4.2.

An overview of the components can be seen in 4.2: the Video Conferencing MultiClient (VCMC), the ObserverControl Client and the Session Player and Analyzer. All these components are usually configured for one study with an *ExperimentDefinition*. While QoE-TB can be used on the fly with the default GUI and manual control of the different parameters, an ExperimentDefinition provides a central point to configure a QoE-TB for a specific study. This typically includes the definition of an Experiment Script which defines the different conditions to be tested in this study, a specific layout for the client with (if desired) integrated tasks, and the definition of questionnaires. The ExperimentDefinition takes also care that the log files have the correct markings for the different conditions which simplifies the export of speech patterns and questionnaire results for statistical analysis.

We implemented the media processing pipelines of our testbed using GStreamer, a flexible, open-source toolkit with source-filter-sink based architecture. While GStreamer is implemented in C, we implemented the not-so performance critical components in the more lightweight programming language Python. This gives us a flexible platform, which is easily extensible and customizable. We implemented the GUI with Gtk, as it is the recommended toolkit for GStreamer and fulfilled our needs.

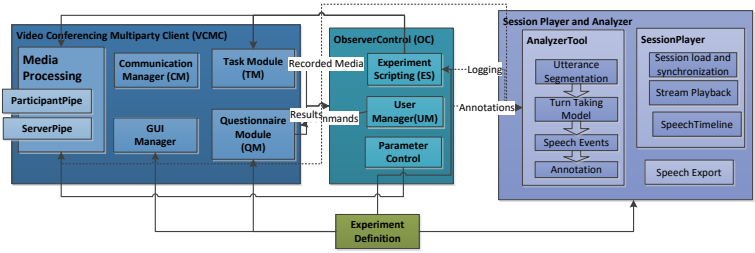


Figure 4.2: The different components of QoE-TB. The solid lines arrows mark logical communication between the components with a flow in direction of the arrow. The dashed lines going into the analyzer, mark the offline processing, i.e. the analyzer accesses the recorded data from the experiment. The double sided arrow shape symbolizes the audio/video streams between the clients.

4.3.1 Client

The clients are full-featured multiparty-video conferencing applications which are directly connected with each other. A description of the individual components follows.

GUI

Fig. 4.3 shows a screenshot of the GUI as used in some of our trials. The client shows the other participants in a square layout, the user themself in the upper left corner, and the task component below the user.

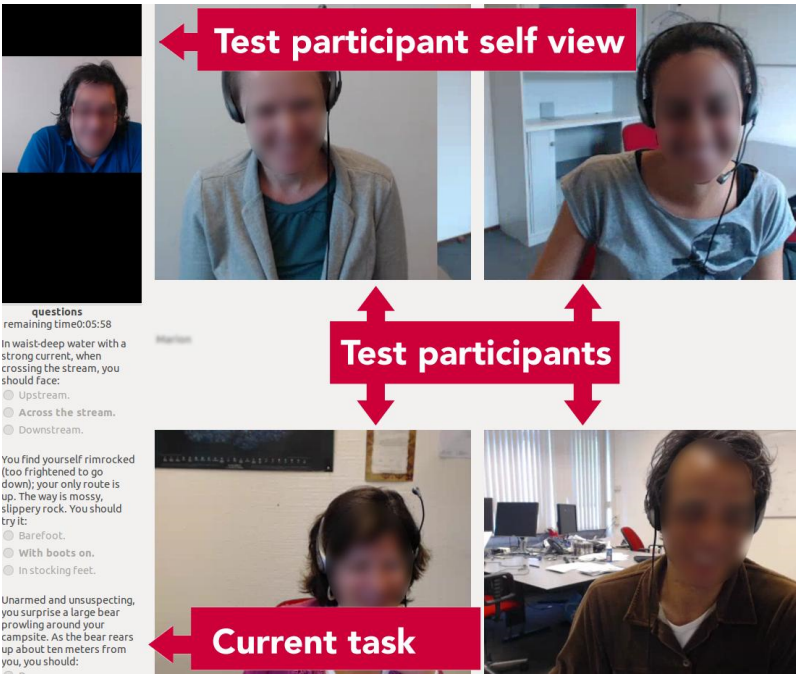


Figure 4.3: GUI of the Client

We employed an M-V-VM pattern [157] to keep our GUI easily customizable. Gtk allows us to use the Glade Builder, a WYSWYG GUI Builder which creates XML files that can be dynamically loaded. We implemented small binding facilities that allowed us to customize the layout without changing the underlying source code. Through these bindings the in Glade can be specified where to place the different video streams, which properties to display (e.g. in Fig. 4.3 we show the name of the participants above the streams), bind to commands of the model and place the task component. In some cases it is necessary to make GUI interactions that are beyond the capabilities of bindings, in which case a view-model is employed, which encapsulates GUI related issues without the need to touch the underlying model. The communication between view-model and view is done via an Observer pattern for which we implemented a small observable base class for view-models. The view needs to contain one

widget for the observer video, which will be normally hidden but can be made visible, fulfilling requirement R6.

Communication Manager

The CommunicationManager provides signaling mechanisms, such as sending commands, negotiating the connections, feedback and chat messages. We implemented the communication via commands, which can be easily bound to the GUI or use them in scripts for the experiment. We implemented this command layer on top of XMPP. In our cases, XMPP was sufficient, but the communication delay (1-2 seconds) may be too long for scenarios in which the media should be changed very frequently, so this abstraction makes this change simple.

GUI Manager

The GUI Manager loads with Glade, Gtk GUI Builder, constructed XML files and connects them to the functionalities VCMC. To achieve this the GUI Manager expects that the file contains placeholder widgets that later serve to display the streams of the participants. This can be enumerated placeholder for participants or an area where all participants will be added (and the space is dynamically allocated to them), a selfview and a conductor placeholder. The conductor placeholder will be hidden by default from QoE-TB. Further the GUI can contain custom widgets for tasks specific to this study.

Media Processing

The system is designed so it runs in a controlled environment at the moment we transmit data using the Real-time Transport Protocol (RTP) over the User Datagram Protocol (UDP). Fig. 4.4 shows a simplified version of a sending and a receiving pipeline for the video stream. Besides the normal elements for capturing, encoding, and transmission, we added elements for monitoring and controlling the network and media. To keep track of the temporal aspects, we log the delay of every frame. To do this, we directly insert a barcode into the video, which we crop-out at the receiving side before presenting the video to the user (compare Fig. 4.3 and Fig. 4.6). By directly

inserting the timestamps into the video we measure the delay of the whole processing pipeline, instead of only of the network delay. For the complete "mouth-to-ear" delay we need to also consider the delay of capturing and rendering equipment, which can be assumed to be static and can be measured using external tools [83].

In the media processing component, during the construction of the pipelines, certain manipulation capabilities are implemented. These parameter controls register themselve at the ObserverControl module which in turn displays control options for them. The parameters Resolution/Frequency and Frame-rate/Sample-rate can be manipulated directly at the corresponding capturing elements (with respect to the capabilities of the devices). For the other parameter, we use the following:

- **Distortion** We can control distortion by inserting available filters from GStreamer (e.g. blur) or changing the codec settings. The easy extensible plugin architecture of GStreamer makes it easy to develop and integrate custom, more complex distortion patterns.

- **Delay** The minimum delay our system achieves, in the ideal conditions of our local network, is in average 70ms with a 25ms standard deviation. We can add delay by increasing buffers on the sending and the receiving side.

- **jitter** We keep the network delay constant by employing a jitterbuffer. We can add jitter by adjusting the buffer on the receiving side.

- **Interstream (Audio/Video) Synchronization** We can achieve audio/video (de)synchronization by manipulating the delay buffers in audio and video streams separately.

- **Inter-participant Synchronization** Since there is a separate pipeline for every participant we can achieve basic (de)synchronization by setting different delays for each participant. Since we have synchronized clocks and the capturing timestamps more complex synchronization algorithm can be built on top of this.

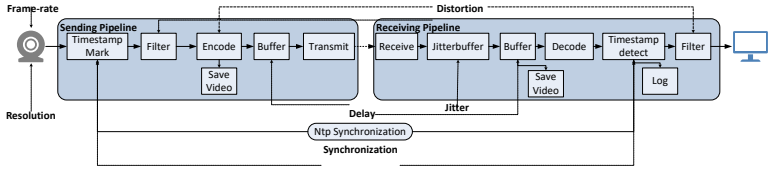


Figure 4.4: Simplified sender and receiver pipeline showing where parameter manipulation takes place.

Since GStreamer allows us to modify these parameters during runtime, we have shown that our testbed fulfills requirement R1. We save the captured and received media streams in each pipeline thus fulfilling requirement R3. As we already described, we log the delay of every frame at each client, so that we are able to monitor delay, jitter and synchronization. The distortion of the media is available through the recorded media streams. While the Resolution/Frequency and Frame-rate/Sample-rate are usually constant they are also monitored through delay and the recorded media, thus fulfilling requirement R2.

Questionnaire Module

The Questionnaire Module is responsible for managing questionnaires: loading them from a configuration file, saving the data, constructing the GUI, and sending feedback to the observer. We created a lightweight library that allows us to rapidly create a questionnaire with the standard elements Likert (type), option lists, dropdown lists, comboboxes and free texts. The questionnaires can be easily defined over XML. It is easily extensible for more complex questionnaires. The results are saved locally and transmitted back to the observer for live control. This component fulfills our optional requirement R5.

Task Integration

The integration of components specifically designed for an activity can be done in two ways. Either a task specific view-model is created, through which, as previously described, the GUI can be customized to integrate the task. Or the task can be implemented as

a separate Gtk widget, which can be integrated in the GUI of the client.The infrastructure of the client can be used for remote commands (later available in experiment scripting), synchronized logging and data transmission/collection. This component fulfills our optional requirement R4.

4.3.2 ObserverControl

The experiment conductor (using the ObserverControl Client) is usually not shown to the other participants, not to influence the trial, but can dynamically join the conversion, if necessary, to give feedback or additional instructions. The ObserverControl Client is an extension of the normal client, thus it also contains all the modules available at the client, but they are left out of the diagram not to overload it. The ObserverControl Client is composed of a normal client with a customized GUI and the ObserverControl Window shown in Fig. 4.5. The ObserverControl Window can, similarly to the Client, be customized for the specific experiment. The upper left part shows the commands for this experiment. These include showing or hiding the observer, showing or hiding different questionnaires, setting the delay and assigning roles. Normally these commands are not to be executed manually, as they are performed by a script, but due to unforeseeable events during an experiment it might be necessary to execute them manually. Furthermore, below the observer can make annotations with a small comment or just mark a point-of-interest without comment.

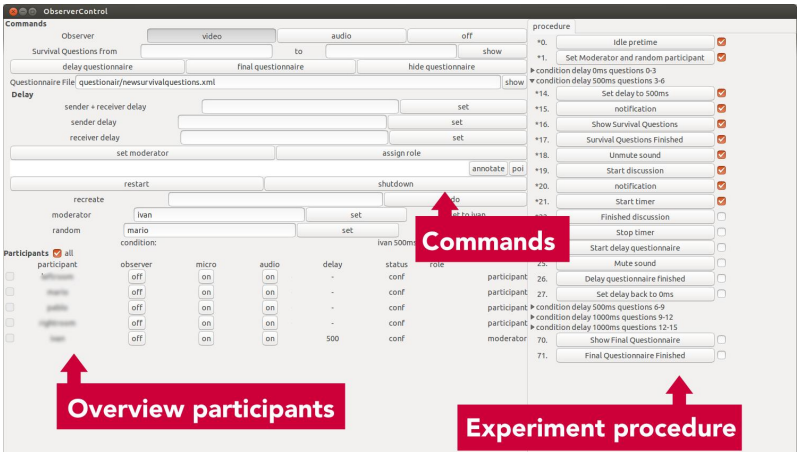


Figure 4.5: ObserverControl Window

Experiment Scripting

The different steps of the experiment can be scripted based on the status of the system. E.g. to automatically show a questionnaire after a task is finished, set new conditions after all questionnaires are filled out and so forth. Each individual step is logged and available for data analysis. The script can be seen on the right side in Fig. 4.5. The experiment consists of steps, which can invoke a command and can have a condition. When the condition is satisfied automatically the next step is executed. When the condition is satisfied the checkbox on the right is checked. It is also possible to include steps which serve only as a checklist item for the experiment conductor (e.g. as a reminder to give specific instructions) and have to be manually checked once the experiment should go to the next step. The button of each experiment step allows the manual execution of this step, if for some reason some steps should be skipped or repeated. All steps and their satisfaction are synchronized logged and are available later in the analysis.

User Management

The lower left part of GUI in Fig. 4.5 shows the user management. Here we can see the status of each client, manipulate some properties directly, like turning on or off the sound, microphone, video or the showing of the observer. To apply commands or experiment steps only to specific participants the checkboxes on the left side can be used to select a subset of the participants.

4.3.3 Session Player and Analyzer

The analyzer tool, shown in Fig. 4.6, is used for the data preparation and analysis after the experiment. It can play the recorded media streams synchronized as they were rendered during the experiment. The experiment steps and annotations are available as bookmarks, shown in the lower left corner, so we can easily step to the corresponding events in the recorded streams. Finally, the speech pattern data can be exported for the statistical analysis.

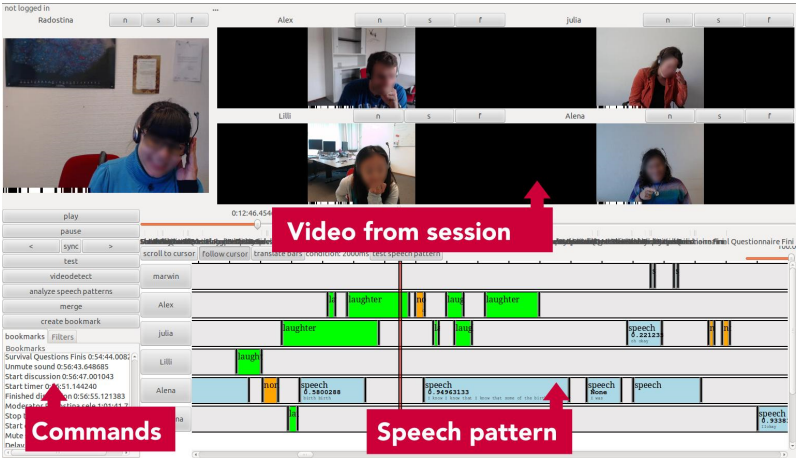


Figure 4.6: Analyzer Window

Session Player

The SessionPlayer enables to playback recorded sessions from an experiment. Via the logs and timestamps the streams are synchronized as they were during the session. The main components of the SessionPlayer are a module to load and save (annotations and speech processing) sessions, media components to playback the recorded media and the time (seen in the lower part of Fig. 4.6), which displays the extracted on off patterns.

Speech Analysis

For the investigation of effects towards speech patterns we use a four-stepped approach.

Utterance Segmentation First, we segment the data in "on-off" patterns of sound activity. To do this, we support two different voice-activity detectors: the Adintool from the open-source speech recognition software Julius, and the reference implementation of the generic sound activity detector of the ITU. The speech analyzer allows comparing different sensitivity settings.

Turn-Taking In the next step we filter out on patterns which are too short for speech (15ms) and fill silent gaps in the speech (e.g. resulting from stop-consonants) and constructing the turns.

Events As detailed in section 2.3 some speech patterns events are particularly interesting to assess whether conversational problems exist. This is especially the occurrence of simultaneous speech. In this case, it is particularly interesting whether people start to speak at nearly the same time, or if a speaker change occurred after the simultaneous talk.

Annotation These turns have to be manually categorized in terms of whether they are really speech, or non-verbal utterances like "uhm" or laughter. Furthermore, we can annotate them with tags based on the purpose we can observe. For example, simultaneous speech can be a backchannel mechanism to show attention or agreement, which is intentionally simultaneous in comparison false-starts which are often an accident. The turns and events can further be annotated with comments or self-defined

tags. The speech of segments can also be transcribed. We investigated automatic speech recognition based on Google's Web Speech but the results were not satisfying. It is possible to manually add events like non-verbal communication like gesturing.

4.4 Discussion

Even though an argument can be made that conducted research and tools are separate entities, of which tools are merely interchangeable means to an end, in reality the two are interwoven. As the availability of technologies and tools available hast first made certain kind of research possible. The range in which way tools facilitate science ranges from game changing new measurement technologies that enable approaches not possible before, to tools which simplify necessary work in the scientific process and hence enable to focus more on the actual scientific aspects. QoE-TB mainly focuses on the aspect of providing support for the scientific process but also integrates some innovative approaches to improve control and precision in system parameter control and measurement. The embedding of visually encoded timestamps within the video frame itself allows to have precise information about the real delay of the employed system. For the complete "glass-to-glass" (camera to (remote) screen) delay ,this measurement is missing the delays for capturing and displaying screen. However, these systems work independent of CPU or network load and are thus quite stable. There are further measurement tools [83] for glass-to-glass delay, which employ a similar visual encoding of timestamps, but by means of displaying and capturing it with an external device also include camera and monitor delays. In turn, these systems use the actual video-conferencing system for their measurement and cannot be employed during an actual user test. However, together these systems can provide the complete glass-to-glass delay. Furthermore, the timestamps can be used in the analytical phase of an experiment for post-processing. Several objective metrics use a full-reference approach, thus they require the original and the transmitted frame as input. However several distortion factors (e.g. lower fps, packet-loss, jitter, CPU load) will result in an unequal amount of frames in the videos on sender and receiver side. The timestamps allow to correctly map sender and receiver

frames onto each other and assess which frames were lost. There are further aspects which are specific requirements for conducting user studies, for which alternative solutions come with drawbacks: the integration of experiment task material and the experiment conductor. In the first trials we experienced that providing task material on paper notes took away much attention from the visual channel. It is essential that the conductor can listen to the experiment, first of all to assure that everything is working as expected, answer questions by the participants if needed but also to observe the interaction actually taking place during the experiment. Furthermore, the aspects of integrating questionnaires, scripting of the experiment and logging of the experiment steps take away error sources, and help to save time in conducting and post-processing the data.

4.5 Conclusion

QoE-TB was successfully used in several trials (see chapters 5 and 6) showing that it is capable of simulating diverse but fine-grained system conditions. The easy extension of specific elements for tasks helped to foster desired behavior that we wanted to study more closely, like the integration of UI elements for quiz-questions, which could only be controlled by the moderator of the session. The extensive recording and analysis capabilities made the insights into the interplay of interaction and system factors possible.

5

Conversation & Delay

This chapter describes interactive studies conducted regarding delay in multi-party video-conferencing . In these studies symmetric (all participants with delay) and asymmetric delay conditions were researched (one participant having delay). Further differences in perception of participants based on speaker behavior were investigated by clustering participants by their percental speaking time. The conducted data analysis reveals that active participants are stronger affected by delay than non-active participants.

This chapter is based on

- "Methods for Evaluating MediaSync in Realtime Communication" Chapter in the Springer "Mediasync: Handbook on Multimedia Synchronization"
- ITU-T Recommendation "P.1305 - Effects of delay in telemeetings"
- "The Influence of Interactivity Patterns on the Quality of Experience in Multi-Party Video Mediated Conversation" presented in 2014 at the Socially Aware Multimedia Workshop at ACM Multimedia
- "Asymmetric Delay in Video Mediated Group Discussions" presented in 2014 at QoMEX
- "Mitigating Problems in Video-mediated Group Discussions: Towards Conversation Aware Video-conferencing Systems" presented in 2014 at the Understanding and Modeling Multiparty, Multimodal Interactions at the ACM International Conference on Multimodal Interaction

5.1 Introduction

In the introduction of this thesis (see section 1) we have pointed out that the experience of video-conferencing users is diverse. We have subsequently argued that modern evaluation of multimedia systems needs to be able to estimate the personal QoE of a specific user instead of a Mean Opinion Score (MOS). In chapter 3 we have shown, how in one multi-party video-conferencing session the perceived video quality differs from user to user, depending on the complex interplay of the different video qualities of all participants.

However, obtaining an accurate and personal QoE is much more complex than taking into account detailed and specific system parameters for each user individually. As the conceptual frameworks for QoE [118, 104] have pointed out, non-system factors play a crucial role in the QoE of multimedia users. Many of these factors cannot be assessed in passive studies (as conducted in chapter 3), hence it is necessary to conduct interactive studies. In the previous chapter **??** we detailed the testbed that we specifically developed to conduct laboratory multi-party video-conferencing studies. In this chapter we employed this testbed, in order to start taking into account behavioral factors of multi-party video-conferencing sessions and work towards the goal of predicting the personal QoE of individual participants.

The two factors we are examining in this chapter are the verbal interaction and the delay of the system. Works on evaluating remote conversation systems (e.g. the telephone) already reported that delay interferred with conversation [90]. These studies showed how the timings of our conversational turns and utterances were altered by the delay in the system (e.g. [90, 147]). It was established that faster paced conversations would suffer more from the delay than more calmly paced ones [143, 80, 50].

In this chapter we are using the verbal interaction not to classify the conversation as a whole, but to differentiate the experience of users within the same session. By classifying participants based on their participation in the conversation we could show that their QoE is affected at different delay thresholds.

Conversations heavily depend on the timing of utterances to implicitly organize the conversation and to manage who speaks when (so called turn-taking, see section 2.3). Delay of a real-time sys-

tem or service cannot be perceived directly. We only notice it, because it alters the timings of utterance in a conversation that we rely on in turn-taking. Small group conversations follow the same implicit turn-taking process as dyadic conversations, but the situation is more complex. While within dyads the roles are clear, speaker and addressee alternate turns, in the case of group conversations, participants can also become side-listeners [8] who may not want, or may not be expected to answer to the current speaker. On the other hand, the number of participants who may want to take the next turn are greater. Therefore, in order to determine who speaks next, group conversations depend more on non-verbal cues (e.g. gaze) than dyadic conversations [110].

Due to the increased complexity of the conversational situation in the multi-party context, we assume that delay impacts multi-party video-conferencing sessions differently than in dyadic video-conferencing sessions. In this chapter we are addressing research question 2 of this thesis: **How does the delay impact the QoE of different participants based on their conversational behavior?**. We approach this general research question with four more specific research sub-questions that we wanted to answer with this study.

RQ5.1 What are the lower (just-noticeable) and upper (not-acceptable) boundaries for delay in small-group video-mediated discussions?

While it can be argued that due to the symmetric nature of dyadic conversations delay may be perceived in the same way by both participants, this is no longer true for the participants in the multi-party case. We thus want to investigate whether participants in the same session are differently impacted by delay, depending on their conversational behavior. We formulated the following question for our research.

RQ5.2 What influence has the conversation role and interactivity patterns on the perception of delay?

The multi-party case differs further from the dyadic case as it has a more complex system setup. Due to the heterogeneous structure of the Internet, asymmetric delays between participants are likely

to occur. Depending on each participant's Internet connection, the route between participants and architecture of the employed video-conferencing system, delays can widely differ between participants. We thus want to quantify the impact of asymmetric delays in multi-party video-conferencing . Our assumptions is that the contrast of such an asymmetric situation is the strongest when only one participant is on a different delay level to all other participants, and is thus the best starting point for our investigation. Accordingly, we formulate the following research questions for either one participant having added delay or only one participant not having added delay. Concretely, we are looking to answer the following research questions.

RQ5.3 How is the QoE of the whole group affected by one participant having delay? Is there a difference between the participants with a higher delay as compared to the ones without added delay?

RQ5.4 How is the QoE of a group with all people having high delay different from a group with only one participant having a high delay?

To gain insight into these questions we conducted a 59-participant study on the effects of symmetric and asymmetric delay in five-people group discussions. We first conducted the symmetric trials with one way delay ranging from 150ms to 2150ms, to establish the boundaries between when delay would go unnoticed and when a conversation would break down (research question RQ5.1). The exercise we used was a "surviving in the wilderness discussion" scenario, similar to the "desert survival problem" [94], since this is a commonly used scenario in small-group communications research. Our version was modified from the original ranking items task to a quiz-style scenario based on a team-building exercise [17]. To reinforce the effect of a central role emerging, one random participant was assigned the role of moderator. Together with the obtained speech patterns this would allow us to investigate the research question RQ5.2. For the asymmetric sessions, we added delay to only one participant (research question RQ5.3). We added a delay of 1000ms to this participant because, as the symmetric study had shown, at this level of delay communication was still possible but it iwas clearly perceived as very

disruptive. Since we had observed differences between participants based on their conversational role, we added or withheld the delay from the moderator and one randomly chosen participant. We further compared the results of the asymmetric setup with the results of the symmetric case (research question RQ5.4).

The results of the symmetric study showed that, in general, the delay at 1000ms was perceived as strongly disruptive. Even with 2000ms participant still managed to complete the tasks, but often switched to a more formal turn-taking mechanism, such as the moderator calling out each participant about their opinion. The central role in a conversation was best determined by clustering participants by their percental speaking time in "active" and "non-active" participants. The active participants already reported a strong decrease in their QoE at 500ms while the no-active participants experienced this drop at a 1000ms delay. For the asymmetric case, the data showed that already one participant with delay had a significant impact on the conversation, while one participant without delay did not alleviate the situation. Further the sessions with one participant having delay, were not differently rated depending on whether this participant was the moderator or not. Also participants with delay did not report a statistically significant different experience as compared to participants without delay in the same session. This highlights the complexity of the relation between delay and conversation, and the difficulty for participants to pin-point the exact cause of their problems.

The remainder of this chapter is structured as follows: in section 5.2 we detail the setup of the study, scenario, participants, conditions and gathered data. We then show our analysis of the results in section 5.3, assessing the effects of delay in symmetric and asymmetric cases, qualifying the results by speech patterns and comparing the ratings from the symmetric and asymmetric studies. In section 5.4 we discuss the thresholds, the differences between active and non-active participants, the perception in the asymmetric case and the differences to the symmetric case, and we compare the results obtained in this study to results from dyadic studies. Finally, we summarize the findings of this chapter in section 5.5.

5.2 Methodology

5.2.1 Participants

The study was conducted with 59 participants. We conducted all sessions with groups of five people, except one session. One participant did not show up and we were not able to find a replacement in such short notice. We recruited 39 participants via social media and flyers in universities and institutes, who mainly consisted of students and researchers. We recruited the other 20 participants using a recruitment company to complement our demographic with a group of different age and background. The experiment was conducted in English, in which all participants were fluent. 20 participants were assigned to the asymmetric condition and 39 to the symmetric condition. All participants in the asymmetric condition were recruited from universities or institutes. Their average age was 32.7 years (Stdev 10.6, min 20 max 60), and 33 of the participants were female. The average age of the participants recruited from university and institutes was 26.9 years and the average age of the participants recruited via the company was 44.1 years.

5.2.2 Scenario

Our scenario was a consensus-based decision-making task in a moderated small group discussion. The task of our participants was a quiz-style scenario. The participants were given different questions related to survival in the wilderness and they had to consensually pick together one answer from a list of options. The task was based on the team building exercise from [17], a variation of the well known "desert survival problem" [94]. One participant was asked to be the moderator, to submit the final group answers and move the discussion along in order to keep a 10-minute length constraint per round. The order of the quiz-questions did not change in the experiment but the order of the delay did. After each round we assessed subjective feedback via questionnaires. Upon arrival of the participants we had an introduction round, in which we explained our research and the experiment. Afterwards all participants were seated in separate rooms with a running video-conferencing system.

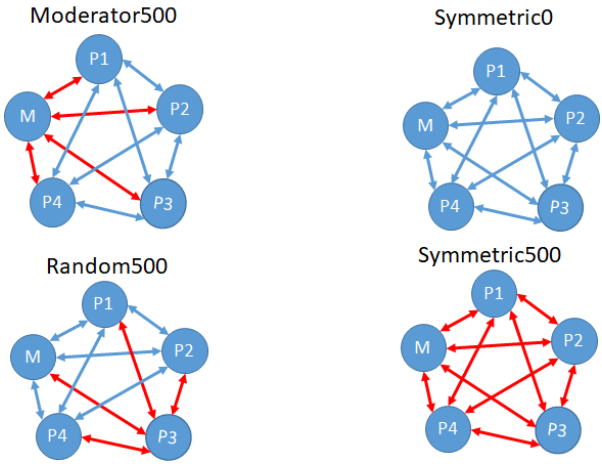


Figure 5.1: Illustration of the different conditions. Red connections are conditions with added delay. In this example the participant P3 was randomly chosen to have added delay.

5.2.3 Conditions

In the symmetric conditions, we tested delays up to 2000ms one-way delay. In the asymmetric case, we decided to add delay to only one participant, as this should be the most different case to the symmetric situation. To reduce the number of test conditions in the asymmetric case, we tested only up to 1000ms one-way added delay. In each session one randomly chosen participant and the moderator (also randomly chosen) got an additional delay (both directions: sending and receiving delay) of 500ms or 1000ms.

Table 5.1 shows the different delay conditions in detail. The GroupConditions denote the maximum delay present in the tested group (e.g. Group500 means in the asymmetric case one participant with a 500ms added delay and in the symmetric case all participants with a 500ms added delay). Fig. 5.1 illustrates the different conditions.

Table 5.1: Delay Conditions

GroupCondition	Asymmetric	Symmetric
Group0	Symmetric0: No participant had an added delay. The base delay was ˜150ms.	
Group500	Random500 / Moderator500: The randomly assigned participant or the moderator respectively had a 500ms added delay (i.e.˜650ms)	Symmetric500: All participants had an added delay of 500ms (i.e. 650ms)
Group1000	Random1000 / Moderator1000: The randomly assigned participant or the moderator respectively had 1000ms added delay (i.e.˜1150ms)	Symmetric1000: All participants had an added delay of 1000ms (i.e. 1150ms)
-	-	Symmetric2000: All participants had an added delay of 2000ms (i.e. 1150ms)

5.2.4 Procedure

Before the beginning of the actual experiment, we had an introduction round to shortly get to know each other and introduce our research. We then seated each participant in separate rooms, where one of the test-system was already setup and running (see sections 5.2.5 and 5.2.6). For each group we used the delay conditions in randomized order. In each condition, participants had to answer three questions, first individually and then together in a 10-minute group discussion. After each condition, the participants answered a questionnaire about their experience during the previous round. After all conditions, participants had to answer an additional questionnaire assessing demographical data, such as age and previous usage of tele-communication systems. We concluded with a discussion of the experiment in a semi-structured group interview.

5.2.5 Testsystem

We used the testbed described in chapter 4. It is a video-communication system designed to conduct tests in a controlled environment. The delay was added by increasing buffers in the media-

Table 5.2: System Configuration

System	Desktop PCs (Core i7, 16GB Ram, SSD)
Webcam	Logitech HD C920
Headset	Creative Soundblaster Xtreme
Video	640x480px, 30fps, 2mpbs H.264
Audio	Speex
Network	Local Gigabit LAN, UDP, RTP

processing pipeline. This approach directly manipulates the system parameters in software instead of using network simulators. The clocks of the machines were synchronized every 15 seconds with an NTP server at the institute. The delay was measured by inserting timestamps at the sender side and reading them out at the receiver side. As we used a configuration with 30fps, this approach has a measurement accuracy of ca. 33ms. All data was recorded on the sending and receiving sides. The system hides the experiment conductor, but gives him or her the ability to interact with the participants if assistance is needed. The configuration of the client interface can be seen in Fig. 5.2. As the figure shows, each participant had an image of him/herself in the upper left corner and an equal representation of the other four participants as the main view. In the lower left corner the questions of this round were presented. The moderator had controls enabled to select and submit the chosen answers.

5.2.6 Apparatus

As we wanted to simulate a home situation we used Desktop PCs (Core i7, 16GB Ram, SSD) with a webcam (Logitech HD C920) and headset (Creative Soundblaster Xtreme 3D). We transmitted the videos in SD Quality (640x480px, 30fps, H264) and the audio was encoded with Speex. The computers were connected over a Gigabit LAN connection and RTP over UDP was used as transportation protocol.

5.2.7 Data

We collected questionnaire data from each participant in each delay condition. Each questionnaire included 15 items, with a nine point

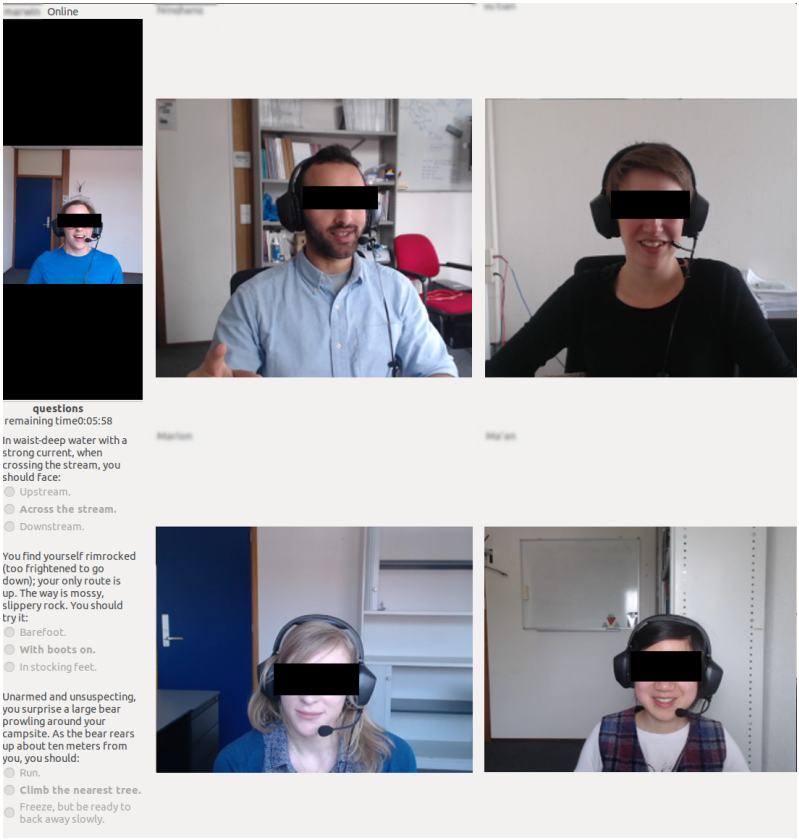


Figure 5.2: Screenshot of the client from the trial

Table 5.3: Perceived quality 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

Table 5.4: Perceived satisfaction about task questions and labels

Label	Question	Scale
satisfaction _discussion	I am satisfied about the course of the discussions in our team.	Completely Disagree <-> Completely Agree
satisfaction _outcome	I am satisfied with the quality of the outcome of our team.	Completely Disagree <-> Completely Agree
contribution	To what extent do you feel that you have contributed to the team's final out-come?	Not at all <-> Very Much

likert-type scale. The final questionnaire at the end of the session included questions about the background and the experience of the participant. As objective data, we measured question scores, from the individual and group results.

The questionnaire contained three items to investigate the perceived quality (table 5.3) and three questions asking participants about their satisfaction with the discussion (table 5.4). For the analysis, the ratings were adjusted so that always a higher value meant a better perception, i.e. higher quality, less annoyance or less noticeable delay. The three questions were meant to complement each other.

5.3 Results

The responses were normally distributed, with respect skewness and kurtosis below 2. We used ANOVA to compare the goodness of a fitting a linear model with our data, to assess whether there was general effect of our independent variable delay on the dependent variables (see tables 5.3 and 5.4). We performed a pairwise difference test with the pairwise student's t-test to see which conditions were significantly different.

5.3.1 Main Effect of Delay

Symmetric Study

We investigated the general trend that with higher delay the perception of quality is worse. The responses to the questions regarding perceived quality (see table 5.3) are plotted in Fig. 5.3. For all items, a lower score means a worse perception, i.e. less quality, more annoyance or that the delay was more noticeable. We performed ANOVA by modeling the responses as a linear function of the delay condition, with the user as a within subject factor and the group as a between subject factor. We compared the fit of our data to this linear function, to see if the differences in the delay conditions were statistically valid.

The analysis revealed that the influence of delay on the quality question was statistically significant ($p < 0.05$), as was the influence of delay on annoyance ($p < 0.05$). 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 statistically significant ($p < 0.05$).

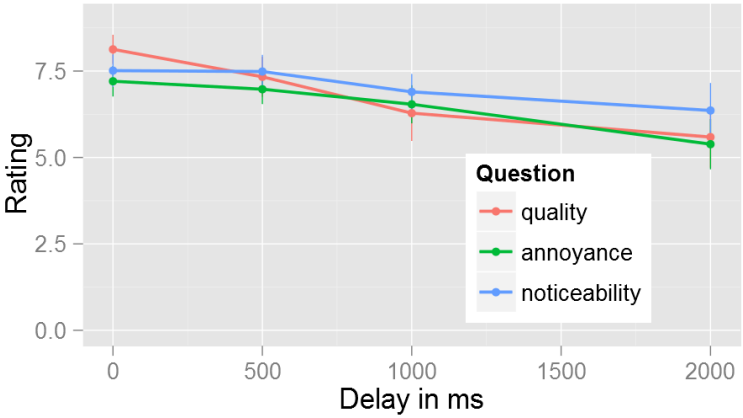


Figure 5.3: Average Questionaire Quality Ratings with 95% confidence intervals

Asymmetric Study

We concentrate on the three quality items in the asymmetric case, their averages per condition are displayed in Fig. 5.4. The error bars in this and the following Figures represent the 95% confidence intervals.

The performed ANOVA showed that the delay is an influencing factor for all three items with p = 0.00852 for quality, 0.01336 for annoyance and 0.00052 for noticeability.

We performed a pairwise t-test to see whether these differences were perceptible. The noticeable differences were between symmetric0 and Moderator1000 (p-value = 0.035) and Random1000 (p-value = 0.0165). Random500 and Moderator1000 were different (p-value = 0.012). Moderator500 and Random1000 were also different (p-value = 0.023).

In other words the difference between no delay and one of the participants having 500ms delay was not perceptible but the difference to 1000ms was. The difference to the 500ms delay cases towards the 1000ms cases was perceptible in some cases. For annoyance and noticeability the difference was perceptible between the 0ms and 500ms

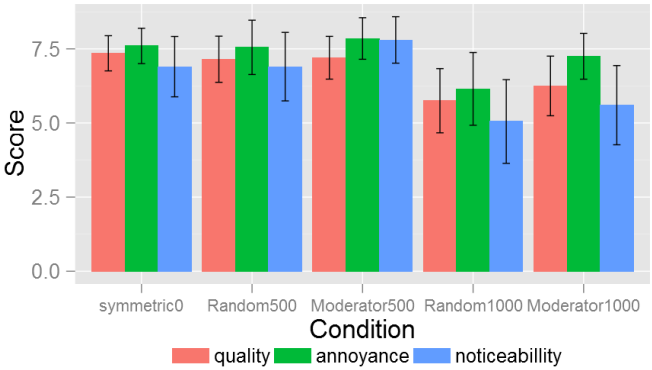


Figure 5.4: Average Responses to Quality Questions

and the Random1000 condition ($p < 0.05$), and indications for differences to the Moderator1000 case ($p < 0.15$).

As we did not find significant differences between the Moderator and Random cases, we merged the Random500 with the Moderator500 and the Random1000 with the Moderator1000 condition, as shown in Fig. 5.5.

The t-test between the different conditions showed that for all three variables, the difference between Group0 and Group500 is not significant ($p > 0.05$), but between Group500 and Group1000 the degradation in QoE is perceptible ($p < 0.05$).

We further compared how (in these conditions) the perception of participants with delay differed from participants without delay. We did not find significant differences between the perception for any of the three variables, Fig. 5.6 depicts the responses for quality.

5.3.2 Qualification by Speech Patterns

We further hypothesized that a concrete speech pattern would influence perception. While the approach in previous research was to build an interactivity metric for the whole conversation [50, 143, 171, 80], we used speech patterns to group the participants. We clustered our participants by speech patterns using k-means into two groups: active and non-active participants.

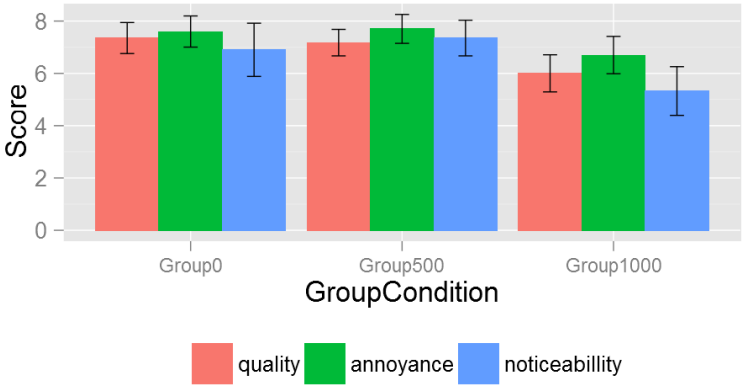


Figure 5.5: Average responses to quality questions by group conditions

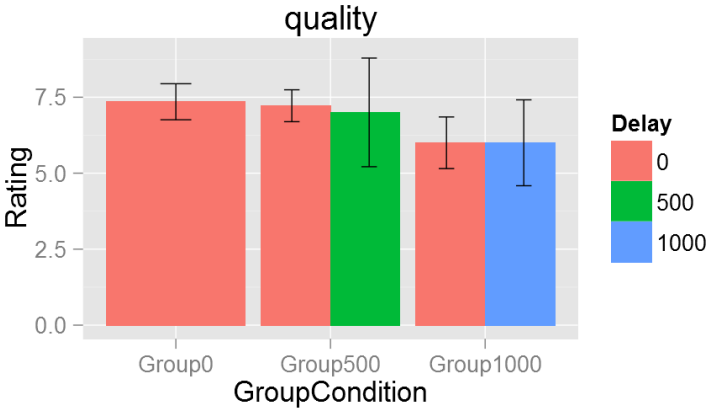


Figure 5.6: Responses for Quality by GroupCondition and Delay

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 gained the most explanation of variance with two clusters.

Symmetric Study

Fig. 5.7 shows the results for the three questionnaires. 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 followed a similar pattern. Active participants had a significant ($p = 0.025$) rise in annoyance between 0ms and 500ms while for non-active participants the difference was insignificant. 1000ms was the statistically significant ($p = 0.009$) difference for non-active participants, being nearly the same as for active participants. Interestingly the difference between 1000ms and 2000ms was strongly noticeable for non-active participants ($p = 0.0003$) but not for active participants ($p=0.15$).

Noticeability was generally less affected by delay. Both groups started with a similar perception at 0ms, going minimally up for non-active participants and slightly down for the active one, but for both groups the difference was not significant. Due to the large variance the difference became 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 happened between 500ms and 1000ms ($p= 0.048$) but

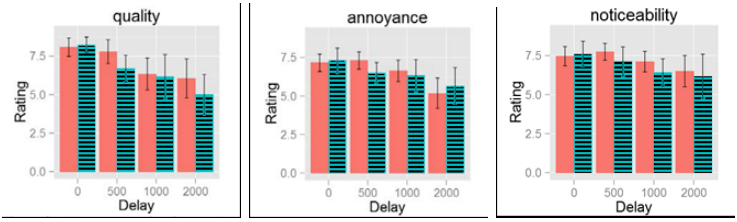


Figure 5.7: Average questionaire results clustered by blocks percentage duration of active participants (blue striped), and non-active participants (red solid) with 95% confidence intervals.

not between 0ms and 1000ms ($p = 0.116$).

Clustering by speaking time

Based on the assumption that interaction is an important factor for perception, we clustered the participants with kmeans by their percentile-part of the conversation. We had used this clustering into "active" and "non-active" participants in the symmetric delay study as it revealed big perceptional differences between these groups. This resulted in two groups in which both the randomly selected participant and the moderator were active participants and two groups in which one of them was active and the other one non-active. In none of groups both were non-active.

The responses for quality of this clustering are shown in Fig. 5.8. The difference between both clustered groups in the asymmetric case were not as clear as in the symmetric study, but we report them here as they follow the same trend. In the Group0 condition, there were strong indications that active participants had a different perception than non-active participants ($p < 0.063$). Differences in perception between the rounds were a trend for active participants with $p = 0.157$ between the conditions Group0 and Group500 and $p = 0.134$ between Group500 and Group1000. For non-active participants the difference was noticeable between Group500 and Group1000 with $p < 0.05$ and not perceptible between Group0 and Group500 ($p=0.39$).

Annoyance was not significant in any of the cases. Active participants could more easily distinguish noticeability between conditions Group500 and Group1000 with $p < 0.05$, whereas for non-active par-

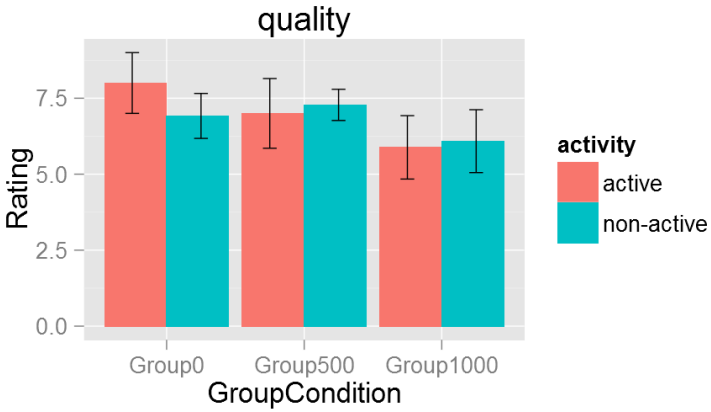


Figure 5.8: Responses to Quality by GroupCondition and Activity

ticipants it was less clear ($p = 0.133$).

5.3.3 Comparison between Symmetric and Asymmetric Study

A t-test comparison between of the overall quality scores between the symmetric and asymmetric studies, restricted to the base condition in which no participant had delay, showed that participants had a different perception in both experiments ($p < 0.05$). However this differences dissappeared when we only compared participants that were recruited through the same method ($p = 0.34$).

While our data on the whole set of participants showed that in the symmetric-delay case the perception of active participants was significantly different from non-active participants ($p < 0.05$), this pattern was less significant in the subset of participants recruited from university and institutes. We, thus, report these findings as trends. For the participants recruited from the universities,the t-test revealed a p-value of 0.1558 for the symmetric case and a p-value of 0.1572 for the asymmetric.The perception for non-active participants in both cases is not significantly different. At a 1000ms case active

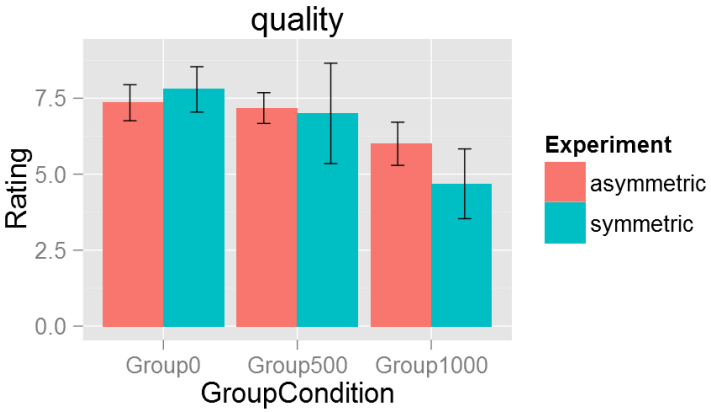


Figure 5.9: Responses to Quality by GroupCondition for asymmetric and symmetric conditions

and not active participants perceived the condition similarly.

Further comparison between the symmetric and asymmetric conditions (Fig. 5.9) showed no statistically significant difference in condition Group500. However, in the case of condition Group1000, they were just above the significant confidence ($p = 0.0508$).

We compared people with delay in the asymmetric case with people in the symmetric case with the same added delay. The trend we found was that even for participants with delay in the asymmetric case, the delay was less annoying than for participants from the symmetric study. However, due to the small number of participants that had delay in the asymmetric case, we have a low statistical confidence (p-value of 0.13), thus we can only report it as a trend. Active participants, however, noticed the difference between a group with no delay or a group with 500ms added delay ($p = 0.03171$).

The sample of participants with added delay in the connection was too low for a statistically significant result. As we had an asymmetric setup at most one participant had delay in the connection, giving us only 4 observations for a complete group. When we comparedthe perception of active people in the symmetric and asymmetric conditions,

we did not find evidence that this was different.

5.3.4 Subjective and Objective Performance

Subjective Performance

The scoring of the questions regarding experienced satisfaction with task and discussion (see table 5.4) went down as more delay was added (see Fig. 5.10). The boxplot with the three responses per delay condition (Fig. 5.10), shows that responses of satisfaction_discussion and satisfaction_outcome are very similar. They have a pearson correlation of 0.85 ($p < 0.05$) and it is quite likely that both questions measure the same underlying principle. We can thus average satisfaction˙discussion and satisfaction˙outcome with the label satisfaction. The responses to satisfaction˙discussion are not normal distributed, while the responses to the other two questions are (in respect to kurtosis and skewness below 2). Neither did the composite variable satisfaction follow a normal distribution. We thus used the Wilcox signed rank test to see whether there was a significant difference between our conditions. This revealed that differences were not significantly perceptible between the 0ms and 500ms condition (p-value = 0.149) and between 1000ms and 2000ms (p-value = 0.412). For all other conditions we had $p < 0.001$. For the contribution question, for which the responses are normal distributed with respect to kurtosis and skewness below 2, we use ANOVA, comparing the fit of a linear function as within subject design and Group as a blocking factor to see if we have an influence of delay. This is the case ($p < 0.01$) and a pairwise comparison of the different conditions revealed that the conditions with 1000ms and 2000ms were significantly worse rated that the conditions with 500ms or 0ms added delay. In other words, the level of satisfaction people had with the discussion was higher if they had a delay of 500ms or less compared to a delay of 1000ms and more.

Objective Performance

We investigated whether the added delay had an influence on the number of correct answers (survival˙score˙group). In Fig. 5.11 we plotted the group scores per round color coded for the different de-

Figure 5.10: Responses to Satisfaction Questions

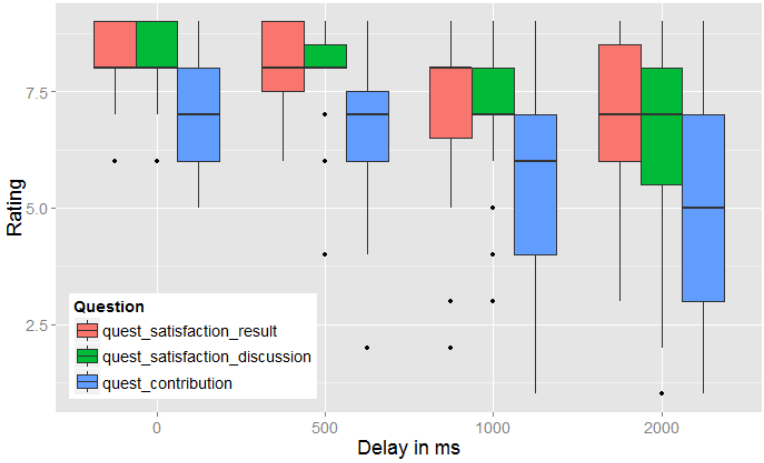
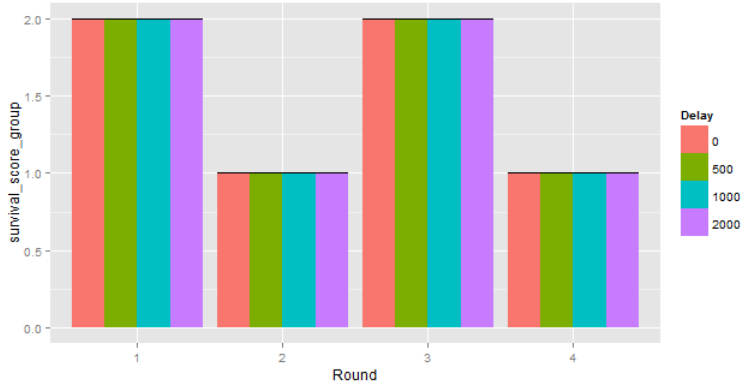


Figure 5.11: Group Survival Scores per Round



lay conditions. In each round, participants achieved the same group score, independently from the delay. In round 1 and 3 all group answered two questions correct and in round 2 and 4 all groups answered 1 question correct. Thus we can assume that the questions in round 2 and 4 were more difficult than the questions in round 1 and 3.

Even though we had chosen a task in which discussion is an integral part in forming the solution, we could not find a statistical significant influence in the task performance. On the other hand, the satisfaction with the discussion and the feeling of having contributed was significantly less with higher delay.

5.4 Discussion

In this section we discuss the results we presented in the previous section. We first take up the thresholds we obtained in our statistical analysis and illustrate concrete effects on participants with excerpts from the post experiment discussion and details of the effect of different communication styles. We then delve deeper into the different perception of delay for active and non-active participants and for delayed and non-delayed participants within the same asymmetric delay group. Finally we draw comparisons between the asymmetric, symmetric and delay conditions and draw upon the available literature to compare multi-party delay studies with dyadic delay studies.

5.4.1 Thresholds

Our data from the generalized case (see Fig. 5.3), suggests that noticeable quality degradation sets in between 500ms and 1000ms delay. This is a higher delay than reported in dyadic studies [159, 171] and similar to the multi-party study from Berndtsson et al. [14].

Contrary to disturbances in audio-and video-streams, participants cannot directly observe delay. It is only indirectly perceivable due to e.g. longer pauses and more simultanous 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, which 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 it still seemed 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 had forced participants to employ additional explicit organization mechanisms. Instead of somebody taking a turn by simply speaking, one participant 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.

Communication Styles

Our moderator was instructed not to make the final decision by himself. Instead, the moderator was requested to make sure that everybody's opinion was heard, to move the discussion along if necessary, and to fill in the group answer in the form. As this description was intentionally very broad, different styles of moderation emerged in interplay with the other participants. In six of the eight groups the

moderator adapted at some point a systematic approach to hear the opinion of everybody. While in all groups the moderator directly inquired the opinion of particularly silent participants in some situations. Only in these groups the moderator started the discussion of a question in a structured way. The moderator proceeded either by asking each participant which answer they had chosen, or by asking who had chosen a specific answer. In four of these groups an organic discussion without an explicitly enforced structure proceeded as the different options were debated. Two moderators maintained almost throughout the whole session an explicit turn-taking scheme. These moderators showed an assertive behavior in situations when somebody spoke "out of turn" by interrupting them with a comment like "*I will come back to you in a moment.*". Three moderators started the session with an organization scheme while three others adopted it at a later stage. In three groups, the moderator explicitly mentioned that the delay was so strong that he/she would adopt a strategy to handle it. In one of these groups, the moderator had already employed this scheme in most occasions. Of the groups which adopted a communication strategy later, only one changed back to free discussion when there was less delay.

While the results from the task scores showed that participants could still communicate all necessary information to solve the task, their subjective satisfaction with the discussion and the completion of the task was severely impacted. A likely explanation is that the explicit organization schemes especially reduce the amount of non-task focus communication, as it was found in a study comparing push-to-talk with free talking audio communication [35]. While this more social talk is not necessary for the completion of the task, it fulfills important roles for the feeling of working together as a group.

From the comments and behavior in the sessions and debriefing discussions we found indications that low and high moderated groups noticed the delay, but due to different interactions.

In three of our groups, after several incidents of unintended interruptions, the moderator announced that they were going to switch to an explicit turn-taking scheme with calling out names. In one group with many active participants, a lot of double talk occurred in lower delay conditions. One participant mentioned that she guessed a delay was taking place in that moment, since laughter in reaction to jokes

came particularly late.

However, it was also mentioned at least three times (once during the debriefing) that participants noticed particularly long delays in strongly moderated groups.

We could not statistically confirm that there were more long pauses in strongly moderated groups. We noticed that in less strongly moderated groups more questions which did not directly address a specific participant were asked (e.g. "*Who chose answer one?*"). In such situations it is natural for a longer pause to occur in the conversation. Such pauses are often perceived as natural part of the conversation, different from the long pause when one participant is asked a direct question.

Communication Problems

We present here two examples, both in the 2000ms condition, from a more freely conversing group and a strongly moderated group. Three participants in the more freely conversing group attempted to say something, a simultaneous talk start occured, all participants turned silent, a long pause, all participants started again roughly at the same time, pause, all three started again and bursted out in laughter after they realized it had happened again. After this happnned, the moderator decided to call out names. In the stronger organized group:
P1 (Moderator): "*P2*"
P2: "*Can you hear me?*"
P1: "*Yes.*"
As with a delay of 2000ms, the other participants did not hear the answer for another 2 seconds, two of them decided to answer after roughly 3 seconds.
P3: "*Yes*"
P4: "*We can hear you.*"
After P2 hears P1, she/he started to talk, but was interrupted shortly after by P3 and P4. He/She was confused and annoyed and asked again: "*Can you hear me?*" Then the three people answered at the same time, after which P1 presented her/his reasoning (for the discussion).
Both incidents were mentioned in the discussion afterwards. People in the first group said they found it funny since this was an experiment. However, in real life, this would have probably been the point

where they would have stopped the connection. P1 reported to be very annoyed by this incident: "*I was already at the top of my annoyance level, I was like "Hello can you hear me?*"
In the debriefing participants reported that, the delay was more problematic in situations where they wanted to discuss things in more detail.
P5: "*if you just vote it was okay - it was more problematic when we needed to brainstorm*"
They reported that the moderator had a more crucial role with higher delay.
P6: [discussing higher delay conditions] *for us it was easy . . . but you* [to moderator] *you needed to keep control.*"
Besides calling out the participants' names, two moderators also employed the strategy of assessing who chose what by asking participants to raise their hands. In one session, this was attempted by two participants but was not adopted by the others. In the group where the moderator used this method particularly often, he/she also stated that over the longer time until people rose their hand the delay was noticeable. In general, the video was reported as a helpful addition in group conferencing. It was mentioned as particularly better than audio-only conferencing.
P7: [. . .] *in telephone conferences there is only noise* [everybody speaking over each other] *and silence. . . you really need a strong moderator.*
It was also an easy way to assess the opinion of particularly silent participants.
P5: [. . .] *P8 didn't say much but it was always easy to see if she was agreeing and following along or had a different opinion.*

5.4.2 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. In most groups, we could observe, 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, which resulted in 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.

Our presumption of the differences between the two groups is that the experience has an influence in the perception that accounts for less strong differences for active and passive participants. However, as we performed controlled experiments and not a long-term longitudinal study, we only have little insights into the previous experience of the participants. We asked the participants about the frequency in which they use various communication mediums, but even though the younger group had more previous experience, the differences were not statistically significant. The correlations between previous experience and perception were also not statistically significant.

Some of our participants reported in the discussion that the overall quality and the delay were never as bad as they had experienced it during some of their Skype sessions. This might indicate that besides the frequency, more data about the actual experience participants had in Skype before is necessary.

As our clustering by speaking time showed, the perception of a group with a delayed active speaker is not much different than the perception of a group with a non-active participant having delay. This highlights that the difficulties which arise due to delay in the conversation are not easy to explain by the users.

5.4.3 Perception of asymmetric delay within the group

The data reported by our participants did not show a significant difference for participants with delay compared to participants without delay in one session up to 500ms.

This was also reflected by our participants in the discussion. While some people noticed delay in the connection, they did not attribute it to others. Only very few people reported that they felt they were delayed in respect to others. More people reported that they had the feeling of being delayed, as compared to people who stated that they had the feeling somebody else was delayed.

Participants did not get the feeling that only communication with

the delayed person was problematic, but attributed it to a more general group discussion feeling. The comments reflected that people notice sometimes problematic instances, but do not necessarily concentrate on the details, e.g. with whom this problem occurs. TAs a participant stated when discussing the experience of delay in a different session:

[P1]: "*There were some awkward moments when you wanted to say something and someone else wanted to say something ... you have to kind of sync it ... but I think I'm used to it*"

There were some reports where people could identify that they or somebody else had some particular delay. Afterwards, we asked whether they had noticed particular instances in which they had the feeling that the delay was particularly noticeable.

[P2 - moderator]: "*At some point I realized I said something and I had to wait for quite a while that there was a delay ... I just realized that once.*"

[P3]: "*I noticed after a while you took longer to ask the second question.*"

This leads to the interpretation that observing other people in the same discussion having a communication problem also reduces the QoE for those participants who are not directly having the communication problem.

5.4.4 Perception of delay between symmetric and asymmetric condition groups

Our study confirms that a delay up to 500ms is barely perceivable in a video-mediated group discussion. The perception of a group in which one participant has a delay is not much different from a group in which all people have a delay. In the case of 1000ms delay, the QoE of the groups with only one participant delayed was significantly better than in a group in which all participants had a high delay.

But our analysis showed that the variance of perception of people with a delay was higher than the ones without a delay. In turn, we could not statistically confirm that the perception of somebody with a delay in a group without delay is better than the perception of participants in a round with all people having delay.

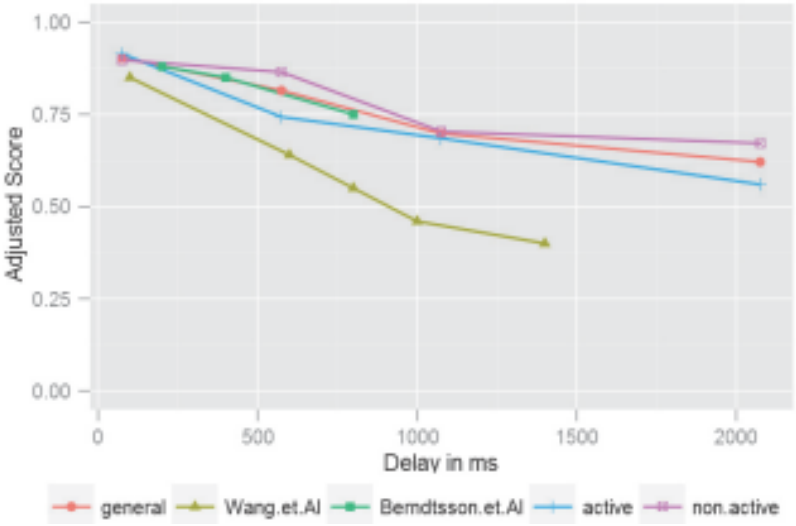


Figure 5.12: Comparison with Berndtsson et al [14], Wang et al [171]

5.4.5 Comparison dyadic and multi-party conversation

Our studies showed that a delay of up a 500ms added delay was barely perceivable by not so active participants and in most cases with up to 1000ms added delay a normal conversation could still be sustained.

These results are lower than the findings reported from previous research in dyadic communication [171, 159]. Our findings are similar to the results from [14] which supports our results. We plotted our results from the symmetric conditions together with the results from Wang et al. [171] and Berndtsson et al. [14] in Fig. 5.12. Since these studies differ in their setup and scenario this comparison is not meant to be a head-to-toe comparison, but to show general trends. In all three studies, the same question was used to investigate the perceived quality (in Wang et al. [171] a Chinese translation), only with different scales (5 point and 9 point). We adjusted the scores to a scale from 0 to 1 between the minimum and maximum possible in the corresponding study. The results from the study from Wang et Al.

[171] are qualified by the average length of talkspurts in a sentence, we accordingly computed the average length of turns in our experiment (7.9s) and compared only the results with this average talkspurts length. It shows that the perceived quality in our multi-party study and the study performed by Berndtsson et al. [14] degrade much slower than the dyadic study by Wang et al. [171].

The main differences, besides the number of participants, is that Wang et al. [171] employed a scripted scenario. This is likely to also yield more sensitive thresholds. The study by Tam et al. [159], employed an unscripted dyadic conversation, and found strongly noticeable negative impacts at 500ms, suggesting a more relaxed threshold for unscripted conversation but stronger than for group communication. However, as they administered different questions and had a system conveying eye-gaze faithfully, the comparison is even more difficult (thus we did not plot these results in Fig. 5.12). Further Geelhoed et al. [43] reported that in their multi-party study, a delay of over 1000ms (one-way delay) has a surprisingly small negative impact and people could still have a normal conversation. Also this study is hard to compare since it used different questions and a high-end video-system (life-sized displays for every participant, faithful eye-gaze).

5.5 Conclusion

In this chapter we made a step towards better understanding the QoE of individual participants in multi-party video-conferencing sessions, by uncovering differences in perception of delay based on conversational roles. To this end we conducted a QoE evaluation of delay in a five-party scenario with off-the-shelf end-user hardware. We found that the degradation of quality perception was strongly influenced between a 500ms and 1000ms added delay.

We described how we designed a scenario that allowed us to gain insights into how a role influences the perception. We then used 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. The outcome of the task results suggest that participants can still communicate all necessary information. The subjective ratings and feedback reveal that the communication gets severely disturbed and the satisfaction is lowered. Users adapt to video-communication situation by employing explicit turn-taking schemes.

We further investigated the impact of asymmetric delay, where different participants have a different delay. The presence of a delay for one participant has a strong negative impact on the whole group experience. At a 500ms-added delay, the experience is similar to the symmetric case and noticeable by active participants. With 1000ms the disruption is less intense than in the symmetric case, but similarly perceived by all participants in the group.

The results show 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 aspects participants prioritize in situations where the resources are limited in demanding multi-point scenarios. The comparatively strong impact of asymmetric delay, indicates that models that aim to describe the QoE of individual participants in one multi-party session, should incorporate, at least, the delay factor and the system properties of all participants.

6

Engagement & Video Quality

In this chapter we examine the impact of different video-qualities in an interactive study. We analyze the impact of video-quality on QoE and show with a variance component analysis that user and group specific aspects explain more of the variance in ratings than video quality. By using cues extracted from the video streams, we further show that participants adapt their behavior to accommodate bad video quality. We then examine user factors in detail. We show that the user state factors engagement and enjoyment are particular useful to understand the QoE of participants and detail the impact of different user background factors, such as age or previous experience with video-conferencing. Finally we use the factors available to us, to construct a predictive model using a feature selection algorithm for predicting QoE. We compare the impact of different influencing factors on the performance of the model.

This chapter is based on

- "Towards individual QoE for multi-party video conferencing" published 2018 in IEEE Transactions in Multimedia
- "1Mbit is enough: video quality in multi-party video-conferencing" presented 2016 at QoMEX.

6.1 Introduction

In the previous chapters of this thesis we have looked at two non-system factors: context (how the composition of different video-qualities has to be considered with multiple streams) and behavioral (how conversational roles influence the perception of delay). In this chapter we turn towards the remaining influence factor, the user factors. Several researchers have addressed the high diversity of users' opinions [57, 85]. This diversity cannot be merely ascribed to poor experimental design or small sample sizes, since even with large numbers a significant diversity within users' opinions remains [57]. Different users have a different experience with the same system factors: individual differences with respect to demographics, personality, and cultural background, for example, have been shown to play a role in the QoE of streamed video [177, 146]. In addition, dynamic factors, hereafter referred to as user state, which include motivation, engagement and enjoyment, can also influence and be influenced by QoE [26].

With user factors we describe thus individual aspects of users that influence their experience of the quality of a multimedia system. We distinguish two different kinds of user factors. One belonging to the category of "user state factors", which are aspects that describe the current physical and psychological state of a person, e.g. current emotions like happyness or anger, physical aspects like being tired or state of minds like concentrated. The other kind of user factor are "user background factors", which describe aspects related to the history of a user that might make him or her appreciate the quality of system differently, e.g. previous experiences of the multimedia system, age or gender.

User state factors are volatile and as such they can change during a video conferencing session, while user background factors evolve in the long term. User state factors are seen as input and output factors for QoE [118], e.g. an annoyed user might be more severely affected by quality than a happy user, but quality degradations might also make users become annoyed. User background factors, on the other hand, are seen as only influencing factors (i.e. input factors) for QoE (even though QoE might be seen as long-term influence factor of some background factors like previous experiences).

In this chapter we examine how user state factors as well as user background factors can influence the impact that system factors have on QoE. To this end, we conducted a study investigating the QoE of participants, but we additonally assessed the engagement of participants and gathered demographical background data from the participants. We show how each individual's engagement impacts the QoE. To come closer to the goal of personal QoE we then present a QoE prediction model in which we include, besides the system factors, a multitude of non-system factors, such as the engagement as user state, user background and behavioral factors.

Engagement describes how interested and involved the participant is with the currently performed activity. We explored the user state engagement as it is relatively neutral from an emotional perspective: users can be engaged in a task for many reasons, such that their liking a particular activity, or their just wanting to finish it quickly. For other user states, like anger and happyness, we would hypothesize that they work mainly as priming effects, i.e. an angry user will react in a more negative manner to quality problems while a happy user will be more forgiving. However, with engagement we found that the amplification and softening of the negative impact on QoE are also plausible hypotheses. More engaged users might notice quality degradations more, since they are more concentrated in the task and thus more disturbed by them or, on the contrary, they could notice degradations less since their focus is more on the task and less on the quality.

To gain insight into user background factors we gathered different demographical factors from our participants (e.g. age, gender, frequency of usage of video-conferencing). By including user state, user background and interaction factors, in this chapter we approach personal QoE with a QoE model that can automatically predict a user's QoE when using the videoconferencing system. Whereas models for videoconferencing QoE prediction exist [11], there is room for improving their accuracy. Most of these models base their estimations on an analysis of system factors only: for example encoding bitrate, or packet loss. This chapter sets out to understand how these elements could be integrated into a more complete model for videoconferencing QoE estimation.

Furthermore, it has been shown that in computer-mediated com-

munication, the impact of system factors on QoE depends on the ongoing interaction between users [32, 141].Hence, interaction should also be accounted for, when modeling videoconferencing QoE. The 'Framework for QoE and User Behavior Modelling' [125] conceptualizes the reciprocal relationship of QoE, user state (e.g. mood) and user behavior. Both user state and user behavior are both an input to QoE and an output. Take the example of a brainstorming session over video-conferencing. It could have a fast paced interaction due to the excitement of participants (or a rather slow one due to a lack of interest in the participants). A long delay usually leads to a worse QoE with a faster conversation [80]; this might in turn cause frustration and break the initial excitement, eventually leading to the abandonment of the current service in favor of another (e.g., email). On the other hand, for other users the effect could be different: some people may find the disrupted interaction still as natural, and some people might only attribute it to the rudeness of fellow interlocutors [142]. We argue that to be able to steer QoE optimization in video-conferencing systems, it is essential to clarify first these mechanisms.

In this chapter, we set out to assess the impact that system factors, in the context of multi-party video-conferencing, have on a) the user behavior, and especially the participant's interaction, and b) how user factors moderate the impact of the system factors on QoE. To gain insights into this topic, we depict our appraoch in Fig. 6.1. The figure is based on the model proposed in [125] and shows the factors examined in this chapter and their relation. Our hypothesis is that the context, together with the user and the system factors, will shape the user behavior. The user behavior describes how the users interact with the system and, through the system, with each other. Interaction depends on the task at hand and the current state of the user (e.g. depending on engagement). In addition, and differently from [125], we consider also the possibility that user behavior can be influenced by system factors. Finally, user, context, and system factors, along with user behavior, will influence the users' QoE, which in turn will influence user behavior and the current state of the user.

To collect data for our investigation, we conducted an empirical study with an interactive subjective test on visual quality in a HD multiparty video conferencing system. We manipulated the system factors encoding bitrate and packet loss, which varied based on

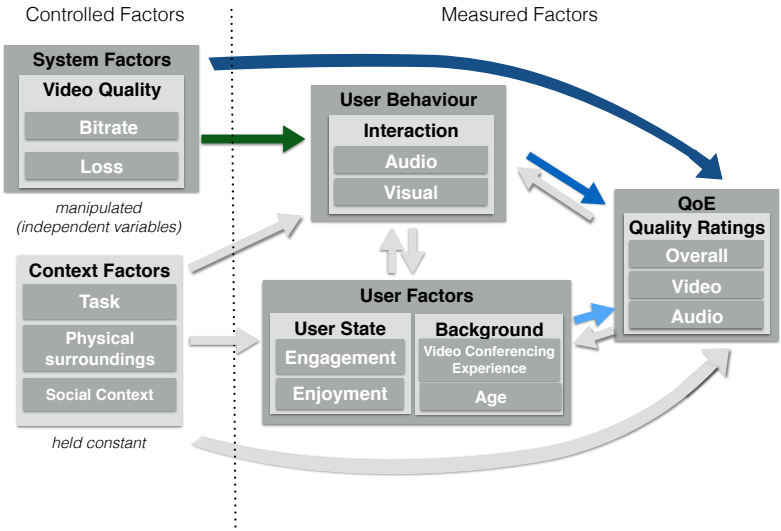


Figure 6.1: Conceptual model of the factors examined in this chapter

network conditions and could be dynamically adjusted during a conferencing session to impact video quality and QoE in general. We chose not to manipulate context, fixing the physical surroundings and task. We used an ITU-T recommend task [78], in which participants cooperatively had to build a Lego® model together over video-conferencing (see screenshot in Fig. 6.4). We chose this task as it is representative for the common situation [89] in which users show objects to communications partners. The task is often employed in audiovisual communication test [13, 18, 141] and was adapted by us for a multi-party situation. We used a desktop-based 4-way video-conferencing scenario with WQHD screens and 720p video-streams encoded in H.264. We tested the effect of encoding the video streams to bitrate levels fitting the bandwidth delivered by different internet access technologies typical for domestic households: broadband, DSL and mobile. We injected packet-loss in the videostreams, simulating typical slightly impaired wireless networks. We recruited always groups of participants which were familiar with each other, i.e. friends or family which shapes the social context of our study. We had participants self-reporting their Quality of Experience, as well as personal information covering both demographics and current state (and especially enjoyment and engagement). Finally, we quantified audiovisual interaction by analyzing both the audio and video feeds of the experimental sessions to understand speech patterns and user activity [129, 128, 141]

This chapter is concerned with research question 3 layed out in the introduction of the thesis (see 1.1.3): **Research Question 3: Is the QoE of participants related to their engagement?**. Besides the user state as the main interest of this study, we further follow up on the user background and user behavior factors. The following specific research questions are answered in this chapter:

RQ6.1 How do user factors influence QoE perception?

RQ6.2 How does a change in video quality (as caused by a decrease in encoding bitrate and/or an increase in packet loss) impact interaction, and in turn QoE?

RQ6.3 Does accounting for user state, user background and user behavior factors on top of the system ones improve a model's accuracy in predicting QoE?

For RQ6.1 we investigate how demographical factors and prior experience with video conferencing, as well as the current state of the user represented by engagement and enjoyment, influence QoE. The hypothesis for RQ6.2 is that if the video quality is insufficient to properly perform the task at hand, users will adapt their behavior to accommodate for the bad quality. Thus, we examine how different visual and conversational interaction cues are affected by the system factors. Finally, to address RQ6.3, we employ the elastic net [180] feature selection algorithm to determine, from all the factors listed above, which are the most relevant to be included in a predictive model for individual videoconferencing QoE. We eventually show, linear models with including a subset of our user and interaction features, more than doubled the accuracy of prediction compared to relying on system factors alone.

The remainder of this chapter is structured as follows: In section 6.2 we detail the study setup and data gathering, data preparation and methods used in the analysis. In section 6.3 we present the analysis of the impact of the system factors on QoE, user behavior, user factors and develop a model for predicting QoE. Section 6.4 discusses the results and how they would be applied in a real life context. Finally section 6.5 concludes the chapter.

6.2 Methodology

Our investigation starts from a user study aimed at quantifying the impact of system and user factors on interaction and, in turn, on QoE. We designed the study to resemble a current multi-party desktop video-conferencing at home, especially focusing on a scenario where video usage would be core to support the communication. In the following sub-sections we detail the setup of the experiment by explaining: how we designed the visually-focused scenario (experiment task); which system factors we manipulated (independent variables); how we administered the conditions (experiment design and protocol); which measures we obtained (dependent variables and covariates); and how we realized the setup technically (apparatus).

6.2.1 Experimental task

We focused the task around the common scenario that video conferencing participants often use the video channel to show objects that are the current topic of the conversation (Zhu, Heynderickx, and Redi 2015). We adapted the ITU-T P.920 building blocks task [78, 18, 141] to a multiparty situation (and in particular, to parties of 4). Each participant was supplied with an disassembled Lego® locomotive[1] (see Fig. 6.2 for an assembled version), and only part of the instructions to build it. Other participants had complementary parts of the instructions, whereby to complete the model, participants had to communicate and share their part of the instructions. Compared to the original ITU scenario, our model included smaller pieces (smallest ca. 5mmx5mmx2mm) in order to make the task more demanding for the video quality. Based on pretrials with colleagues and the experience from previous experiments, we opted for having four rounds of seven minutes each, per each group of participants, each round covering a different experimental condition (i.e., combination of dependent variables). Together with introduction, questionnaires and debriefing this made for 2-hour sessions with each group.

6.2.2 Independent variables

The main technical components determining the video quality in a video conferencing application are the capturing quality of the senders' webcam, the encoding quality, the network capacity (bandwidth and packet loss) and the receiver's monitor. The encoding and network parameters can dynamically change during a session, whereby they are more interesting to optimize than webcam and monitor, which are usually fixed. As a result, we decided to have the same monitor and webcams for all participants (see more detail in table 6.1). As the perceived video quality might be influenced by the size and layout of the video streams, we also decided to have a fixed party size of 4 taking part in the task and to show the video streams of all 4 participants in the same size in a 2x2 layout (thus, including self-view). We choose instead to manipulate encoding bitrate and loss rate as independent variables (often referred in short

[1] Exact model: Lego® item 6060873

as bitrate and loss). We used H.264 for real time communication. The detailed configuration can be seen in table 6.1.

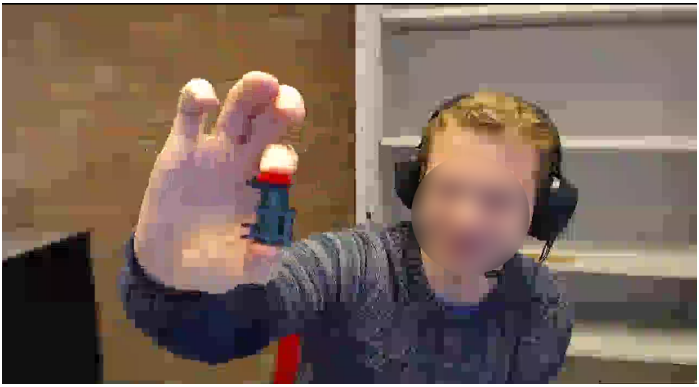
We thus designed the study to investigate the effect on QoE of two independent variables, bitrate and packet loss rate, set at typical levels of domestic environments. With respect to the encoding bitrate of the video stream (thus excluding the bitrate for audio-stream encoding and mixing), we envisioned three condition that represent common internet connections: "low encoding" (256kbs up and 768kbs down), similar to mobile or slow xDSL connections; "medium encoding" (1Mbps up and 3Mbps down) representative of a typical xDSL connection and "high encoding" (4Mbps up and 12Mbps down) for broadband-like TV cable connections. Each bitrate level was further combined with one of two levels of packet loss, i.e. (1) no packet loss, as would occur on a wired connection and (2) 0.5% packet loss, likely to occur over an impaired wireless network [109]. This resulted in a full factorial design with 6 conditions. The screenshot in Fig. 6.3a shows the low encoding quality and Fig. 6.3b shows a screenshot with high encoding quality with packet loss (of which the seen effect would mostly only last for a fraction of a second).

Figure 6.2: Picture of the finished Lego model

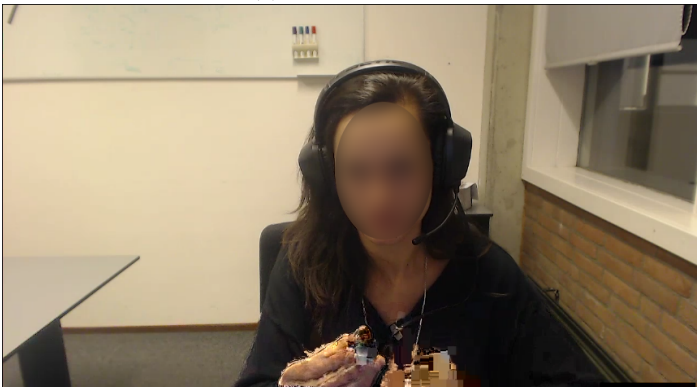
6.2.3 Experimental design and protocol

With 3 bitrate values and 2 loss rates, we had a full factorial design with 6 conditions. To not risk fatigue, we decided on a mixed blocked design. 28 people participated in the experiment (18 female, average age: 31.9, sd: 10), thus we had 7 groups of 4 participants each. Each group assessed 4 of the 6 conditions in a counterbalanced order, hence each condition was rated by at least 16 participants.

Upon arrival, participants were briefed about the purpose of the study, after which they gave written consent for data gathering. Each



(a) low video quality



(b) high video quality

Figure 6.3: Screenshot from (a) low quality (256kbit) video stream of participant showing object into the camera and (b) high quality video (4Mbit) with distortion

participant was then led to a separate experimental room, and seated at a distance of 68cm from the monitor to be used for the experiment, as recommended by ITU-T P.913 [77]. The video-conferencing software was started remotely by the experimenter. In the beginning, the experimenter was present in the video conferencing to ensure that the system was working properly (e.g. adjusting the volume), and that the participants understood the experimental task. In this respect, a brief training session was also run where participants familiarized with the best and worst possible condition, for anchoring purposes. The experimental task then began, structured in four rounds of 7-minutes each with a different condition. The participants were informed beforehand that after 7 minutes the system would automatically display a questionnaire (see section III.E) and the next round would begin when all participants had filled it. Between each condition we asked participants if a pause was needed. After the four rounds, a final questionnaire was administrated and the participants were gathered again for a debriefing.

6.2.4 Apparatus

Each of the participants performed the task in a separate room with similar lighting and background conditions. An identical setup of computer, display, webcam and headset (see table 6.1 for details) was provided in every room. For the experimental task we used QoE-TB (see chapter 4). The software employs GStreamer for media handling and transports them with RTP over UDP as the transport protocol.

To realize the packet loss, RTP packets were dropped on the sender's buffer, thus all participants saw the same distortions. The employed webcams (Logitech C920) compressed the captured video in H.264 in the camera as the USB2 link could not transfer raw video for resolutions above VGA. Although the encoding bitrate can be set in the camera up to 20mbps, tests showed that in practice not more than 5.8mbps would be delivered. The video was always captured in the highest quality and then re-encoded with GStreamer x264 to have more control of the settings (see table 6.1). The experiment conductor monitored the session, but was only visible and audible during the introduction. The screenshot in Fig. 6.4 shows the timecode embedded in the video which was cropped in the video of participants (8px).

Table 6.1: System Setup

Hardware	Model Nuc 5i5ryh: Core i5u, 8GB Ram, SSD		
	Displays	Dell 27" 2560 x 1440 (WQHD)	
	Headsets	Creative Soundblaster Xtreme	
	Webcams	Logitech C920	
Fixed System Parameters	Resolution	1280x720 – per participant	
	Framerate	24 fps	
	Encoding	H264 (x264) with Tune zero-latency, ultrafast speed-preset, GOP size 24, no b-frames, sliced threads encoding	
	Audio	AMR encoded	
	Delay	One-way ca. 120 ms	
Conditions	Encoding Bitrate		
	LowEnc: 256kbps	MediumEnc: 1024kbps	HighEnc: 4096kbps
	Loss		
	None (0%)	Random (0.5%)	

Figure 6.4: Screenshot from Experiment

6.2.5 Dependent variables

After each condition, the participants filled in a questionnaire about the experience they just had. Five questions were directly related to the quality: three ITU questions regarding overall, audio and video quality, one question inquiring annoyance and one question assessing how well they could see facial expressions ('How well did you see facial expressions of other people?' on a scale from 'very well' to 'not at all'. For the other items see [78]). We further asked six questions based on a questionnaire developed for engagement in computer usage [111]. After finishing the experimental task, one post experiment questionnaire regarding demographic information and enjoyment of the task was administered (questions shown in table 6.2, except age and gender). All items (condition and post experiment) were assessed on a 5-point ACR scale. In addition, throughout the experiment, the audiovisual streams of all sessions were captured on the sending and receiving sides.

Table 6.2: Questions of the post experiment questionnaire. The 'label' column indicates the abbreviation used to indicate the dependent variable in the following analysis

Question	Scale left/right	label
I enjoyed participating in this study	Not at All / Very Much	enjoyment
I liked the task of playing with Lego.	Not at All / Very Much	likelego
I am very experienced in using video-conferencing systems.	Very unexperienced / Very experienced	priorexp

6.2.6 Data preparation and analysis

To answer the research questions, we employed different techniques. First, we used descriptive models to analyze the relationship between the independent variables that we controlled (bitrate and loss) and the interaction cues we extracted from the audiovisual streams. We then proceeded with descriptive models to analyze how user factors alter the impact from system factors on QoE. To combine the interaction, user and system factors into a predictive model, we used a machine-learning feature-selection approach. In the following sections, we first describe how we quantified interaction from both the audio and video feeds of the experimental sessions. Then, we introduce the statistical and learning methods that we applied in our analysis.

Interaction cue extraction

To quantify interaction, we extracted several indicators from both the audio and the video streams.

Audio stream analysis. The analysis of the audio recordings aimed at better understanding (changes in) speech patterns among the participants, looking at turn-taking, overlapping speech, and pause length. Previous research has shown that, for example, delay alters the natural communication patterns (Kitawaki and Itoh 1991). Hence, we looked for indicators of these changes.

We used the data as received at the client side to include the

system delay in the analysis. From the recorded audio we extracted chunks of speaking/not-speaking blocks with the help of the Adintool from the Julius voice-recognition software[2]. The tool outputs timespans (i.e. start/end times) of voice activity. The blocks correspond to single utterances, which can then be investigated independently or in groups to better understand speech patterns. The metrics were calculated for each participant separately, based on his/her temporal reality, i.e. based on how the audio arrived with the delay at that participant. Due to this, metrics such as pause duration may have slightly different values across participants, which would not exist if all participants were collocated.

Per participant and round, we identified a number of elements in the conversation (adapted from (Sacks, Schegloff, and Jefferson 1974; Sellen 1995)):

- Turn: A sequence of blocks from a single speaker with less than 200ms pauses between the blocks (similar to a sentence, except that we do not necessarily speak in complete sentence structures).
- Pause: moments on which no participant is speaking.
- Floor: part of a conversation held by the same speaker. A floor starts when a participant begins speaking alone and ends when the next speaker starts with an utterance.
- Overlap: moment in which two or more participants speak simultaneously. It is detected as an overlap in parts of two or more blocks. The person who started to speak first is referenced to as the interrupted, the other one as the interrupter.
- Group turn: a turn containing an overlap.
- Uninterrupted turns: turns without overlap.
- Speaker-alternation rate: frequency of change in speakers holding the floor.
- Simultaneous start: an overlap within the first 200ms of the turns.

[2]https://julius.osdn.jp/juliusbook/en/adintool.html

- Interruption with speaker change: change in speaker after an overlap occurred, e.g. A starts speaking, B starts speaking, A stops speaking while B continues.

Except for speaker-alternation rate, for each conversation element we recorded the number of occurrences (count per minute), duration (mean length in seconds), percentage per participant over the total number of occurrences or total duration of that condition (percentage count and percentage duration). For speaker-alternation rate, we computed the occurrences of speaker changes per minute. For double talk metrics (overlap, group turn, simultaneous start) we also counted how many times a participant was interrupted or interrupting from the perspective of each participant (e.g. with a high delay both participants could get the impression that they were interrupted).

Video stream analysis. To better investigate the impact of video quality on interaction (one of the focuses of our study), we analyzed video streams. For the video analysis we used the unimpaired video streams (sender side), to limit the impact that degradation may have in the computation of the indicators described hereafter.

A preliminary inspection of the video feeds revealed changes in posture and movement of participants depending on the quality conditions. Here, we focused on two constructs, which should relate to visual interaction: movement of participants and distance to the screen. More movement is related to the showing of objects to the camera and moving closer to the screen is often performed by a user so that he/she can see details better.

To quantify the movement of participants used Temporal Activity (TA, sometimes also referred to as Temporal Information - TI). TA is recommended by the ITU [75] to quantify the amount of movement present in videos, e.g. when comparing the performance of different encoders [181]. In our use case, TA provides an interesting tool to quantify the amount of physical activity of a participant: having a fixed camera position and fixed background, changes in TA must come from the movements performed by the participant. Previous research has shown an increase of TA [128] in presence of delay. TA is defined as the total change in luminance between one frame and the previous. For frame t_n:

$$TA(t_n) = rms[F(t_n) - T(t_{n-1})]$$

where rms is the root mean square function (over all pixels in the frame) and $F(t_n)$ is the luminance only video frame at time t_n. Since in our setting, the background is fixed for all participants, TA will be mostly influenced by movements of the participant. Hence, a drop in TA will indicate a decrease in the movement of the participant. We computed TA for every participant and round by means of the mitsu video analytic toolset[3].

We further used the publicly available face recognition software OpenFace[4] [144] to quantify changes in distance of participants from the screen. OpenFace estimates the head position in 3D, the rotation of the face and it also recognizes facial action units. We used the estimated distance to the camera (in mm) as a measure of the distance of the participant from the screen (as the camera is always mounted on top of the screen). The values are expressed in mm to the screen, thus a higher value means that the participants are farther away from the camera.

Statistical analysis

To fully understand the impact that bitrate and loss have on QoE ratings and interaction, it is important to untangle these aspects from individual idiosyncrasies of participants and the group dynamics in a specific session. To do so, we resorted to linear mixed effect models (LMEs) [6], which extend linear models (such as ANOVAs) by introducing the concept of random factors. Linear models are the simplest types of models that can be used to explain data; following the Occam's razor principle, we prefer to employ those over non-linear ones to avoid over-fitting. Moreover, linear models have high interpretability and allow the quantification of the effect of the independent variables (factors) on the dependent one (in our case, QoE measures), which is highly desirable in this exploratory phase.

Linear models assume that dependent variables (in our case the QoE scores) can be modeled by a linear combination of the levels of the independent variables of interest, the so-called fixed effects (here: bitrate and loss). In a linear mixed model, the concept of 'random effect' is introduced to explain the systematic impact that unobserved

[3] http://vq.kt.agh.edu.pl/index.html
[4] https://cmusatyalab.github.io/openface/

variables, uncorrelated with the fixed effects, may have on the dependent variable. This is especially useful to model data obtained from within-subjects designs (such as ours), where observations (ratings) cannot be considered to be independent (as they would be in linear models), since they are expressed by the same user or within the same session. For example, ratings from the same user may be correlated due to unobserved factors (e.g. mood, prior experience, personal preferences [178]). LMEs model these correlations in the data by accounting for the so-called random factors, on top of the fixed ones (i.e., the manipulated independent variables, such as bitrate in our case). An individual offset (i.e. intercept) or slope (i.e. coefficient) is built in the model for each level of the random factor(s) (e.g., for each subject). This allows to explore the differences in the random factors in more detail, and to explain a larger part of the data variance, thereby making the effect of the fixed factors stand out more. LMEs are commonly employed in the field of psychology for user studies because they allow to investigate the effect of a factor while accounting for individual differences. Compared to a traditional repeated measure ANOVA, LMEs allow to better model the mixed experiment design of repeated measure and between subject design. With LMEs, random factors can be modeled in a nested manner. In this experiment the repeated measures from participants nested within groups of participants. LMEs can further handle unequal number of samples per condition. Many QoE models employ (transformed) linear models as they are interested in only predicting an average perceived quality rating (Mean Opinion Score). In this work the focus is exactly in exploring these individual aspects. A visualization of this is proposed in Fig. 6.5, where we model group as random factor per bitrate.

In formal terms, a linear mixed model predicts the dependent variable y based on the following format:

$$y = X\beta + Z\gamma + \epsilon \tag{6.1}$$

Where y is the vector of our responses (different quality ratings) of length n = 112 (7 groups, 4 participants per group, 4 ratings per participant); X is the design matrix for fixed effects, so with a maximum size of n x 11 (3 bitrate levels + 2 loss levels + 6 interactions); β are the coefficients of the fixed effects; Z is the design matrix for the

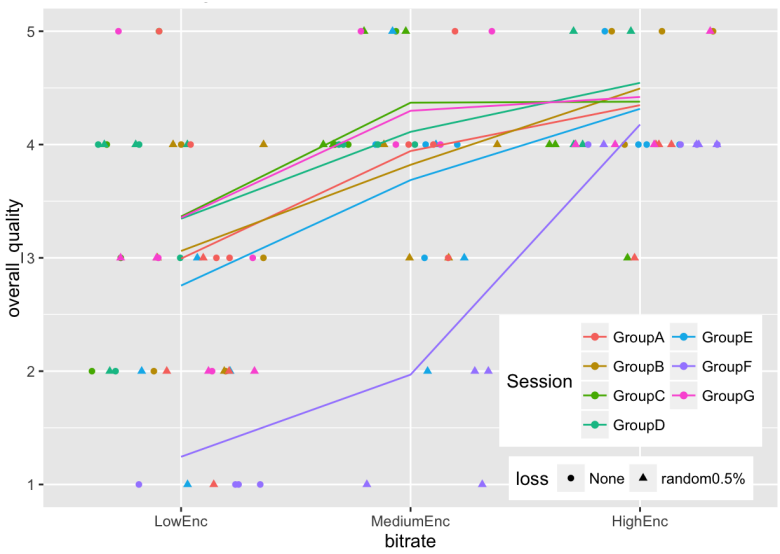


Figure 6.5: Ratings and fitted Model 2 per Session

random effects, with a maximum size of n x 224 (28 users + interactions of 28 users with 7 groups); γ are the coefficients of the random effects and ϵ is the vector of the residuals of length n. We specify then the model 6.1 in different ways to test the impact of fixed and random effects on QoE. For brevity, we denote the models using the R notation:

$$y \sim bitrate + loss + bitrate : loss + (bitrate|Group/User) \quad (6.2)$$

with the specific notation

- bitrate:loss denoting the interaction between bitrate and loss
- (bitrate|Group/User) being short for (bitrate|Group) + (bitrate|Group:User), where in turn (bitrate|Group) denotes a random effect of the group per bitrate and (bitrate|Group:User) denotes a random effect of the user per group per bitrate

In analyzing our data, we consider the influence of two random factors:

User Having repeated measures from (a random sample of) participants, we may expect them to use the rating scale in different ways, or to have different quality preferences (or idiosyncrasies).

Group The specific group interaction and conversation (again randomly sampled from all possible interactions and conversations) may influence the experience of all members of this group, and thereby their ratings.

To assess whether a factor in our model has a significant impact we used the likelihood ratio test (LRT) [25], which detects whether a model with the factor in question has a significantly better fit than the same model but without the factor, in comparison to the additional parameters used. Having established that a factor has a significant impact on the dependent variable, we further investigated it in detail, clustering participants based on the factor with k-Means [98]. With this clustering we reducte the number of values of factor, which helps to visualize and understand the effect of the factor better. The number of clusters was determined with an elbow plot [87].

To evaluate the goodness of the fit of the different models and the relation between our independent variables, effects of group and user, we computed the marginal R^2 (variance explained by fixed effects, higher is better), conditional R^2 (variance explained by fixed effects and random effects, higher is better) and AIC (measurement of goodness of fit in relation too number of parameters used; where smaller is better [1]) in Fig. 6.8. The values were obtained as described in [107].

Predictive model

The LMEs rely on the availability of the users' self-reported ratings, which we obtained in the post analysis of an experiment, but not in real life scenarios. In this work we examined how well prediction would work if instead we included engagement, demographics and interaction cues in our models. The challenge here is that these factors might be correlated while many statistical models assume that all factors are independent (i.e. absence of multicollinearity).

Methods that include regularization have been known to help with correlated features (i.e., the factors we feed into the model) [39]. The basic idea of regularization is to introduce a penalty term in the cost function that drives the model parameter optimization, yielding better generalization and limiting over-fitting [39]. In this work, we made use of the Elastic Net [180], which uses a combination of L1 and L2 regularization. The L1 regularization term includes the sum of the absolute value of the model coefficients to the cost function. This ensures that coefficients for unimportant features will be set to 0, thereby performing feature selection. The L2 regularization term (sum of the square of the coefficients), makes the cost function strictly convex, also allowing the selection of correlated features.

To evaluate the performance of our models we employed the coefficient of determination (R^2), which quantifies the proportion of variance explained by the model compared to the total variance in the data. This is the most commonly used method to evaluate goodness of fit in statistical modeling. To assess how correlated the finally selected factors q are, we used the variance inflation factor (VIF), a statistical diagnostic method to check the severity of multicollinearity of fixed factors of a model [93]. Every factor j = 1, . . . , q is modeled as a linear combination of the other q-1 factors. The VIF is defined over the resulting i.e., the coefficient of determination for factor j of the model as:

$$VIF_j = \frac{1}{R_j^2} \tag{6.3}$$

Perfect independent variables that show no signs of correlation would have a 0 VIF (as a rule of thumb, VIF should be below 10 [93]).

6.3 Analysis

6.3.1 System Factors

Our basic assumptions were that a higher bitrate leads to higher or equal quality ratings, while higher packet loss leads to a lower or equal rating. Fig. 6.6 shows average scores with 95% confidence intervals for the three dependent variables (overall, audio and video quality)

in the six experimental conditions, ordered according to the expected perceived quality.

As our data set is not big enough to build random slopes per both bitrate and loss, we started our analysis by comparing a model with random slope per bitrate (M1 in table 6.3 with a model with random slopes per loss ($y \sim bitrate * loss + (loss|Group/User)$). The model with random slopes per bitrate performed better on all three dependent variables (conditional $R^2 = 0.785$/AIC=287, for overall quality in M1, $R^2 = 0.591$/AIC=309 for the same dependent variable, with the other model; similar results were obtained for video and audio quality), thus we choose to model the random slopes per bitrate.

We investigated the impact of the effects (fixed: bitrate and loss, random: group and user) on QoE by comparing, via a likelihood ratio test (LRT)[25], different versions of the same model, herewith called full and reduced models. The reduced models include only a subset of the effects considered by the full models; if the LRT returns a p-value smaller than 0.05, this indicates that the reduced model has a significantly worse performance than the full one, and hence the omitted effect has a significant influence on the dependent variable. For example, to test whether an interaction of bitrate and loss has a significant impact on QoE, we performed the LRT between the full model with interaction (M1) against the model without interaction (M2 in table 6.3).

The LRT shows that for here is no significant interaction between bitrate and loss for all our assessed quality ratings ($p > 0.05$ for overall, video and audio quality). Based on this finding, we proceed to test the different effects from M2.

The p-values resulting from the LRT between M2, and its reduced versions investigating the effect of loss, bitrate, user and group on QoE are shown in table 6.3. Loss is a significant factor for all dependent variables except audio quality (see p-values for M2B in table 6.3). Bitrate, on the other hand, is a main effect for all dependent variables (see M2L). The User effect impacts the ratings of overall quality and audio quality (M3). Group, on the other hand, impacts all dependent variables (M4). Upon closer inspection of the data, we observed that there was one group showing a markedly different behavior. Fig. 6.5 plots the fit of individual groups from M1 (see table 6.3) as lines, the raw ratings as points (jittered on the x-axis) against

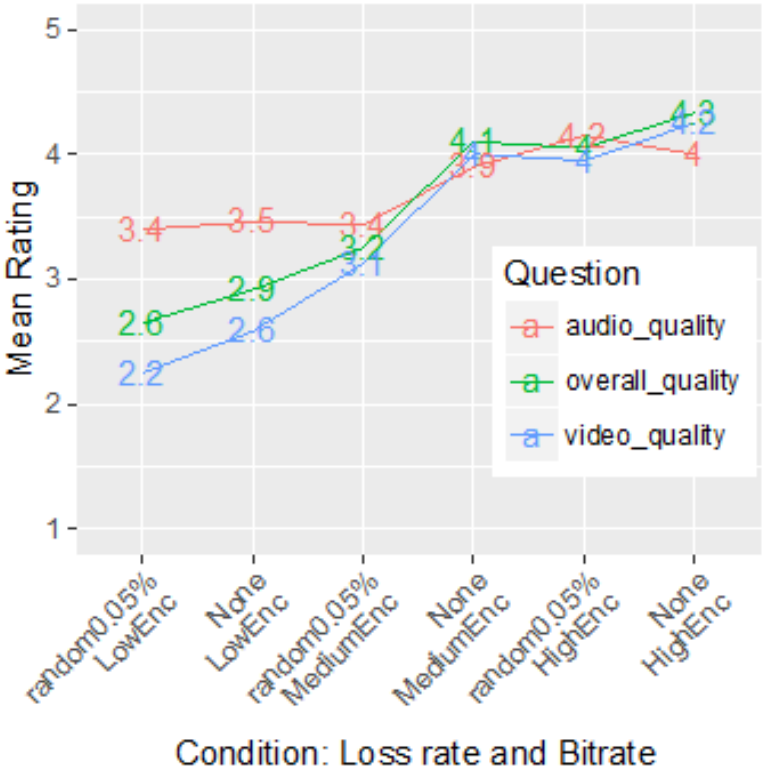


Figure 6.6: Mean ratings with 95% confidence intervals per condition

Table 6.3: P-values from the likelihood ratio test between full model and reduced model (unfiltered dataset/filtered dataset), testing the impact of removing a fixed or random effect on goodness of fit. A p-value < 0.05 indicates that the full model is a significantly better fit than the reduced model, thereby indicating that the influence of tested effect on QoE is significant. We did not find obvious violations of normal distribution and homoscedasticity of the residuals of the presented models.

Effect tested	interaction between bitrate and loss	loss	bitrate	User	Session
Full model	M1) ~ bitrate + loss + loss:bitrate + (bitrate| Group / User)	M2) ~ bitrate + loss + (bitrate | Group / User)	M2) ~ bitrate + loss + (bitrate | Group / User)	M2) ~ bitrate + loss + (bitrate | Group / User)	M2) ~ bitrate + loss + (bitrate | Group/User)
Reduced model	M2) ~ bitrate + loss + (bitrate | Group / User)	M2B) ~ bitrate + (bitrate|Group / User)	M2L) ~ loss + (bitrate|Group / User)	M3) ~ bitrate + loss + (bitrate|Group)	M4) ~ bitrate + loss + (bitrate|User)
overall quality	0.689/0.656	0.006/0.009	0/0.001	0.004/0.012	0.038/0.999
video quality	0.206/0.274	0.011/0.005	0/0.001	0.148/0.325	0.006/0.301
audio quality	0.165/0.091	0.722/0.976	0.03/0.052	0.001/0.005	0.003/1

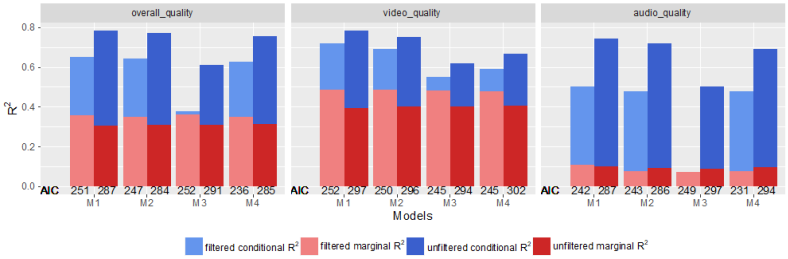


Figure 6.7: Marginal R^2, Conditional R^2 and AIC per Model

the different bitrate levels: it is clear that ratings from Group F are somewhat anomalous. As we have no indication that these are measurement errors, we performed our analysis again, on all data except those related to this anomalous group (hereon we refer to this set of data as the "filtered" set). The results for the effects bitrate, loss and user are unaltered, as shown in table 6.3. However, leaving out the group factor (M4) did not result anymore in significantly worse fit.

In Fig. 6.7, we plotted a bar-diagram for each dependent variable. Each bar shows the marginal R^2 (red) and the conditional R^2 (blue) of a model, for both unfiltered (darker colors) and filtered data (lighter colors). Additionally, we noted the AIC of each model above the label. Comparing the filtered with the unfiltered dataset, we observe that the proportion of variance explained by the fixed factors (red) is generally higher in the filtered data. While the total explained variance is higher for the unfiltered dataset, the AIC indicates that the models perform better on the filtered dataset in comparison to how many parameters they need for explaining the variance. The perceived audio quality is generally poorly explained by the fixed effects, even though the total explained variance is quite high. Furthermore, the difference in conditional R^2 between the model including group as a random effect (M3) and the model including user as a random effect (M4) is large, showing that the audio quality was perceived very differently for participants at a base level. This is particularly obvious for the filtered dataset in which including group as random effect does not lead to any more explained variance. Video quality has the largest marginal R^2 indicating that user and group factors

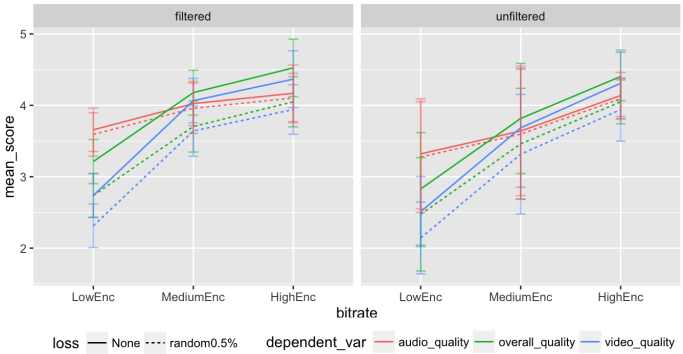


Figure 6.8: Fitted Model 2

play the smallest role for these assessments. For video quality we can see that using user as a random effect instead of group leads to little improvement. Taking into account also the interaction of group and user (M2) gives for video quality the largest gain in R^2, indicating that group interaction as well as individual idiosyncrasies play an important role. For overall quality, in all models fixed effects explain less variance than they do for video quality. Nevertheless, a similar conditional R^2 is achieved.

To precisely quantify the impact on QoE of the fixed effects, we performed a post-hoc analysis of the individual conditions using multivariate t-distribution adjustment for multiple comparisons. Table 6.4 shows the p-values of the pairwise comparisons of the QoE assessments per each pair of levels of each fixed effect. The data shows that for nearly all our variables there is a clear difference between the low bitrate condition and higher bitrates, but users cannot differentiate between medium and high bitrate encoded streams. For audio quality, only the difference between low and high quality is significant, and only in the filtered dataset. We have plotted the mean values and 95% confidence intervals in Fig. 6.8. As we can see a lot of the variance in the unfiltered model was due to Group F.

Finally, we looked into what could explain the different results for Group F. We found obvious differences in this group ratings in another two covariates: reported level of engagement and reported level of enjoyment. In the boxplot in Fig. 6.9 it is noticeable that

Table 6.4: P-values by fixed factors (unfiltered dataset/filtered dataset)

question	LowEnc-HighEnc	LowEnc - MediumEnc	MediumEnc - HighEnc	None – random0.5%
overall_quality	0/0	0/0	0.22/0.25	0.01/0
video_quality	0/0	0.01/0	0.3/0.38	0.02/0.01
audio_quality	0.07/0.02	0.28/0.07	0.43/0.77	0.75/0.66

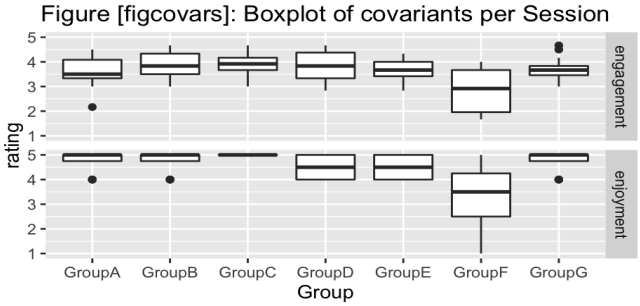


Figure 6.9: Boxplots of covariates per Group

the group was less engaged and enjoyed the session less. An ANOVA with the different Groups as fixed factor showed that for engagement (which was assessed after each round) the ratings from Group F were lower than those of the other groups ($F(6,105)=6.533$, $p<0.001$). For enjoyment (assessed at the end of the experiment session) the results are not as clear but still show a negative trend ($F(6,21)=2.1538$, $p=0.08949$, contrast for Group F being different $p=0.0135$). We concluded that it is necessary to extend our models by taking user factors and behavior into account (see sections 6.3.3 and 6.3.2 respectively).

6.3.2 User Behavior Analysis

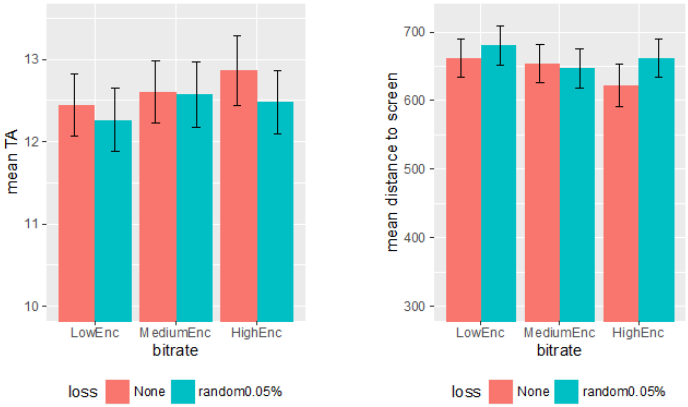
In this section we investigate whether users adapt their behavior in presence of impairments in the video feed. Specifically, we hypothesize that the interaction in presence of highly impaired video (low encoding condition) will be different than when video is provided at

higher bitrates. As we have a task that involves showing objects into the camera we further hypothesize that participants will use the video channel less when more impairments are present. This could, in turn, lead to an increased speech activity to compensate.

For the analysis we use LMEs (See section 6.2.6), modeling the interaction cues as dependent variables and the system factors as fixed factors. As interaction is highly personal but also dependent on the other group members, we are including User and Group as random factors.

Visual Interaction

As detailed in section 6.2.6, we use Temporal Activity (TA) and distance of participant to screen (DTS) as indicators for visual interaction. As these metrics are calculated per frame, but our system factors are on a-per-round granularity, we averaged TA and DTS per round. We first analyzed the impact of system factors on TA. As Fig. 6.10a shows, the less impaired the video is (higher bitrate, lower loss), the more participants move. LRT confirms that even though difference in TA between bitrate conditions is small, it is significant (0.29 points TA difference between low encoding and high encoding). In more detail, the contrasts show that the difference between low and high encoding is significant (p=0.02) and so is the one between no and 0.5% packet loss ($p = 0.03$). Including interaction between bitrate and loss does not provide a significantly better fit ($p = 0.36$), nor does adding Group or User as random factors p˜1). Fig. 6.10b shows the impact of bitrate and loss on the average distance participants kept from the screen (DTS). Participants are closer to the camera with better quality (for both bitrate and loss). Both bitrate and loss have a significant effect ($p < 0.05$ in both cases), also with interaction ($p = 0.03$). Neither including Group as a random factor ($p = 0.65$), nor including random slopes per participant ($p = 0.98$) improved the fit. The contrast showed that the distance to the screen (DTS) was significantly smaller for high encoding than for low and medium encoding ($p < 0.01$ and $p = 0.01$ respectively) and the contrast between conditions with and that without packet loss was significant ($p < 0.01$). In other words, with more impairments, participants moved less and were further away from the screen, which indicates that they interacted less visually.

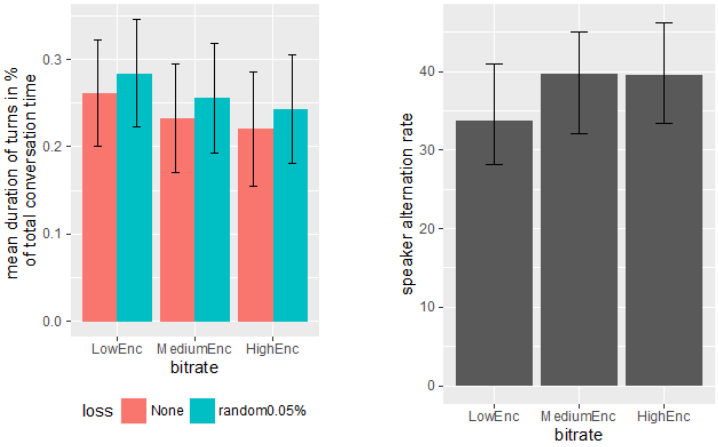


(a) (b)

Figure 6.10: (a) Mean Temporal Activity (TA) by bitrate and loss with 95% confidence intervals (b) Mean distance of participants from the screen, as impacted by bitrate and loss with 95% confidence intervals

Speech Patterns

As already done for the visual cues, we averaged the speech metrics per round. Perhaps due to the fact that they could rely less on the visual channel, participants seemed to speak more in the low bitrate condition. The LRT test showed a significant effect of bitrate and loss, as well as their interaction, on turns percentage duration (each $p < 0.05$). As can be seen inFig. 6.11a, participants speak for longer time in the low bitrate condition and more with packet loss in that condition while there is no significant difference in the medium and high encodings in both bitrate conditions. The conversation also gets slower, as can be seen from the significant lower speaker alternation rate in the low encoding condition (Fig. 6.11b). For speaker alternation rate, there is a significant effect of bitrate ($p < 0.01$) but not of loss ($p > 0.05$). Furthermore, participants speak significantly more at the same time in the low bitrate condition. There is a significant effect of bitrate on group turns duration ($p < 0.05$) but no effect of packet loss ($p > 0.05$). Again here the differences are between the



(a) (b)

Figure 6.11: (a) Percentage of turns by bitrate and loss (b) mean speaker alternation rate by bitrate. Each with 95% confidence intervals

low encoding condition and the higher ones. We thus corroborate our hypothesis that 1Mbit per second is sufficient to enable the task without hampering interaction.

6.3.3 QoE User Factor Analysis

In the previous section 6.3.3 we showed that system factors (i.e. bitrate and loss) had a significant impact, but there was large variance in the responses. We employed linear mixed effect models to take the specific user and the group into account. The models, also revealed a strong effect of User and Group factors (e.g. overall quality had 30% explained variance by system factors, but 79% explained variance by system factor when combined with User and Group factors). In other words, different users were affected by system factors differently, and also by the group they carried out the experiment with. This motivated us to look into how user factors affect QoE. Specifically, we now move to investigate the impact of static factors, such as demographics and previous experiences, and of dynamic factors, such as

engagement determining the current state of the user.

Prior Experience and Age

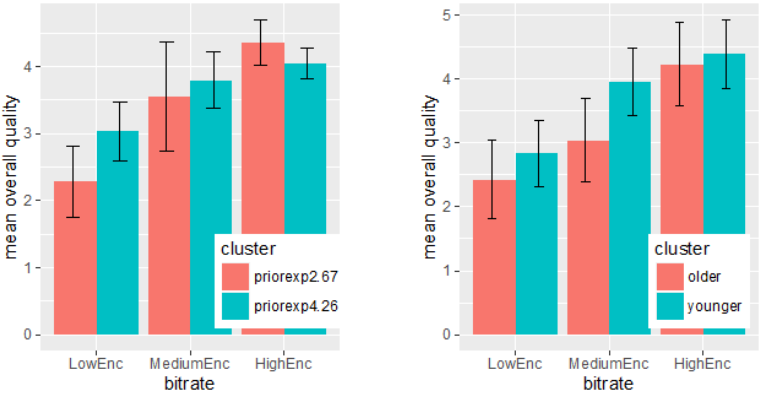
Both prior experiences [62] and age [124] have been hypothesized to influence QoE, and research in computer-human interaction with elderly users has suggested that there might be a relation between these two factors [91]. Different age groups may be used to different media technologies, and be more or less acquainted with different types of impairments. For example, coding artifacts are a typical problem of digital media over the internet, which is nowadays the preferred way to consume video content, but rarely appear in analog TV or DVD content, to which senior people may be more accustomed.

To investigate whether these factors play a role in QoE, we include them, individually, as covariates in our models for each dependent variable (overall, video and audio quality, annoyance and recognition of facial expressions). We then check through LRT whether the addition of each covariate is significantly beneficial to the goodness of fit of the model, as compared to the basic LME with only bitrate and loss.

Our analysis shows that overall quality and recognition of facial expressions (label facial) are significantly affected by prior experiences (label priorexp, each $p < 0.05$), whereas including age as a covariate only results in a better fit for overall quality.

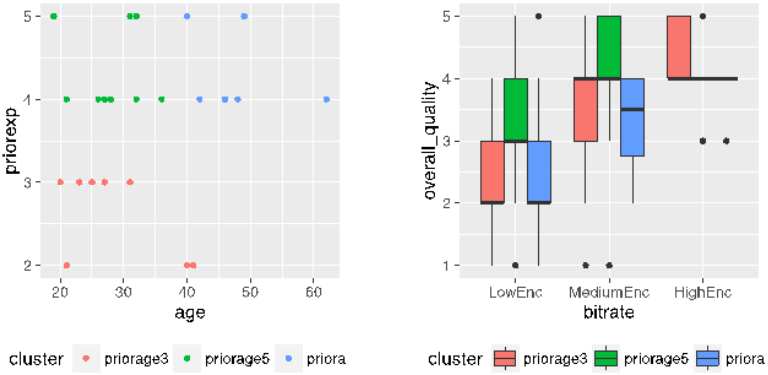
To understand the effects of prior experience on QoE we clustered the participants in two groups, based on how they rated their prior experience (priorexp) with videoconferencing (one less experienced group with mean ˜2.67 (9 participants) and the other with ˜4.26 (19 participants)). The plot of overall quality ratings for both groups in Fig. 6.12a shows that the less experienced group penalizes worse quality much more. The more experienced group gives lower ratings for the best quality: the pattern suggests that more experienced participants are less affected by quality changes. We also clustered participants according to age, into two groups with averages of ˜25 years (9 participants) and ˜44 years (19 participants), respectively. Fig. 6.12b shows that the older age group scores QoE lower than the younger group ($p < 0.05$).

Interestingly, the LRT also revealed that adding both factors and their interaction to the model was beneficial. By adding the com-



(a) (b)

Figure 6.12: Mean overall quality ratings with 95% confidence intervals by (a) prior experience groups (b) age groups



(a) (b)

Figure 6.13: Clustering by age and prior experience. a) (left) clusters by both factors. b) (right) overall quality ratings by clusters

bined age*priorexp factor to the LME model, we obtained a better performing model than those including just one of the factors (each $p < 0.05$). To understand the impact of this term, we performed a clustering on both factors. In preparation for this, we scaled age to 1-5 not to give it more weight than prior experience ratings. We obtained three clusters (suggested by an elbow plot), shown in Fig. 6.13a. We found a young and experienced group (green), an older and experienced group (blue) and an inexperienced group (red). In Fig. 6.13b we can see that the younger and more experienced group (green) is, indeed, more relaxed than the other two groups: the less experienced younger participants and the older participants independent of prior experience.

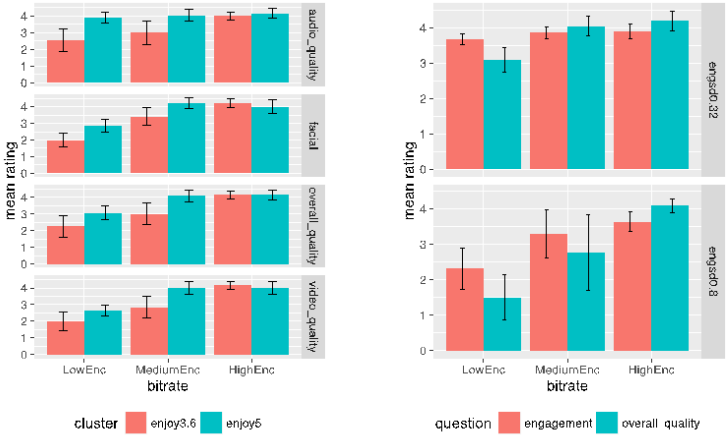
Current state of the user

To estimate the user current state, we assessed engagement during the experiment. We further asked participants about enjoyment of the study and the Lego® task.

Engagement and enjoyment have both been linked to QoE [26], both as influencing factors and influenced variables. In this work we investigated them as influencing factors. We used enjoyment as a measure for how comfortable participants were in the context of participating in this study (as measured by a question at the end of the whole experiment). Engagement is used as a proxy of flow, immersion in a task: it has been shown that impairments can disturb this flow [26], and a flow interruption can hamper QoE.

Our first hypothesis for affective factors was that participants who enjoyed the experiment more had a higher QoE. For enjoyment this was the case. Adding enjoyment to our LME, in a similar manner as we had done with priorexp and age showed that enjoyment as covariate improved the models for overall, video, audio quality and recognition of facial expressions (each $p < 0.05$). Even though the variance in enjoyment ratings was relatively low (mean 4.5, sd.88), the trend that participants with a higher enjoyment gave better ratings is visible in Fig. 6.14a, in which we plotted two groups (mean 5 and 3.6, each group 12 participants) with the four affected dependent variables.

Engagement was assessed with a six item questionnaire in each round (see section 6.2.5). A reliability analysis revealed an excellent



(a) (b)

Figure 6.14: Significantly affected QoE ratings by enjoyment groups (b) Engagement and overall quality ratings by sd engagement clustering. Each with 95% confidence intervals. s

consistency between the items with a raw Cronbach's alpha of 0.79. We thus computed a combined engagement score per participant. We first checked whether the system factors (bitrate and loss) had a direct effect on engagement. Analogue to how we proceeded with the QoE ratings, we tested if a significant effect exists via a LRT with mixed models and engagement as the dependent variable. The LRT showed that there is no significant effect of bitrate and loss on engagement ($p = 0.29$).

We wanted to understand if this holds for all users. Similar to investigating user subgroups in [57] we examined the engagement ratings from each user in more detail. We looked at how constant the engagement of users was throughout the experiment by taking the standard deviation (sd) of the engagement ratings they expressed after each round. A higher standard deviation of the ratings would indicate higher fluctuations of engagement, possibly due the changes in bitrate and loss. K-means identified two clusters of users, the largest of which had smaller fluctuations in engagement (21 participants, sd

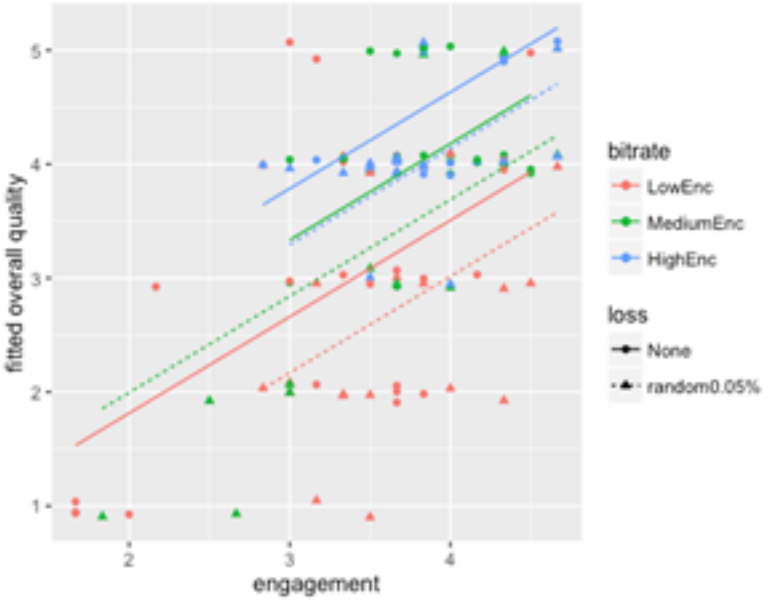


Figure 6.15: Overall quality against engagement by bitrate and loss. Points represent the individual ratings (jitter on y for better visibility), lines represent the fitted model.

mean 0.3) whereas the other showed more variance (7 participants, sd mean .8). We can see in Fig. 6.14b, that the experience of the two groups differs: participants with more fluctuation in their engagement report a lower QoE, even if their engagement is currently the same as in the other group. We checked the contrasts to confirm that the ratings of both groups are statistically different ($p < 0.05$ except in the high encoding condition). Furthermore, the contrasts between bitrates show that for the less engaged participants the difference between medium and high encoding was rated significantly different, while this was not the case for the more engaged group.

Turning now to the relationship between QoE and engagement, we continued to include engagement as covariate to bitrate and loss for modeling QoE. For audio, video and overall quality, engagement proved to be a significant covariate ($p < 0.05$). To visualize the effect engagement has on the overall quality, in Fig. 6.15 we show how the overall quality changes with engagement in the fitted model that contains engagement as covariate. As we can see, a one-point higher engagement yields around 0.5 points higher overall quality.

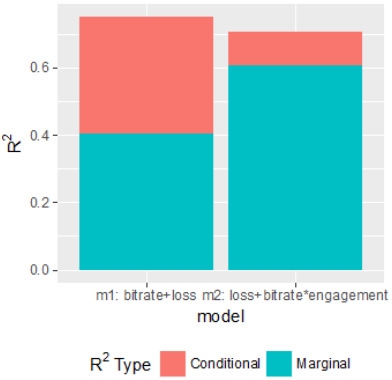


Figure 6.16: Comparison of Marginal and Conditional R^2 of modeling overall quality with system factors alone (m1) or with engagement (m2). The random factors for both modes is (bitrate|User/Group).

Interestingly, engagement explains a lot of the variance that we formerly had attributed to the random factors User and Group. In Fig. 6.16 we visualize the Marginal R^2 (variance explained by fixed factors alone) and Conditional R^2 (variance explained by including the random factors) of a model without (m1) and a model with engagement (m2) for overall quality. Model m1 was identified in section 6.3.1 as the one best explaining variance in our data based only on system factors. The model includes bitrate and loss without interaction and a

random slope per bitrate for the random factors User and Group (m1: $overallquality \sim bitrate + loss + (bitrate|User/Group)$). We introduce here model m2, which additionally includes engagement as a fixed factor, interacting with bitrate and loss (m2: $overallquality \sim (bitrate + loss) * engagement + (bitrate|User/Group)$). As we can see in Fig. 6.16, m1 explains ca. 40% of the variance with the fixed factors (blue part of the leftmost bar) but reaches ca. 75% explained variance including the random factors (fullleftmost bar). m2 explains ca. 60% of the variance with fixed factors (blue part of the rightmost bar). The portion of variance now explained by random factors (individual and group differences) is now much smaller. This suggests that indeed the user engagement is a significant factor for forming QoE in video-conferencing.

6.3.4 A model for predicting videoconferencing QoE

So far we have detailed how system factors influence the interaction of participants and the current state of the users, and how user factors (e.g. prior experience and engagement) influence QoE. The analysis in the previous sections, however, focused on single factors and explanatory statistical models. In this section we are testing how well QoE (specifically overall quality) can be predicted by including our non system factors. It is also of interest to understand which, among the many factors we considered, is most relevant for the prediction.

As detailed in section 6.2.6, we will be using an elastic net to model QoE, as its properties fit well our scenario (handling of correlated variables, feature selection). To investigate the contribution of each type of factor, we divide them into different categories based on our previous analysis: visual cues and speech patterns, background (age and priorexp) and current user state (engagement and enjoyment).

We ran the elastic net algorithm with 10-fold cross-validation with different values of the regularization parameters and selected the model with the lowest Root Mean Squared Error (RMSE). The results for the models accounting for different factors categories, including the input factors, the finally selected factors and model performance (R^2 and RMSE) are shown in table 6.5. We tested with

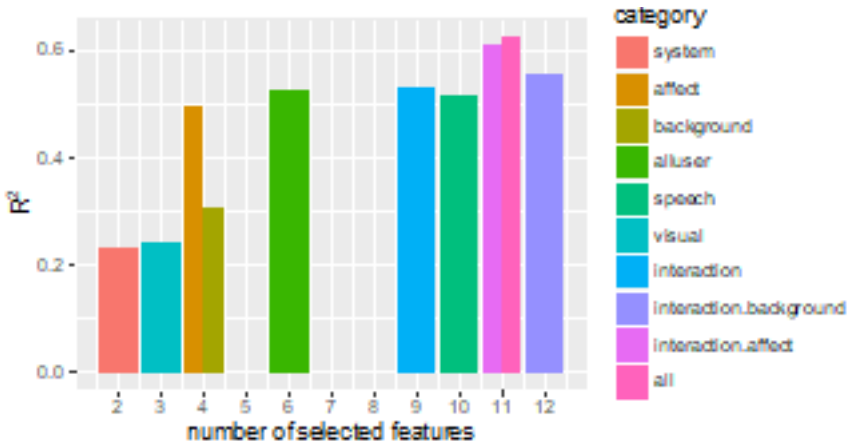


Figure 6.17: R^2 and number of selected features for the QoE prediction model when fed with the different feature sets.

the VIF (see section 6.2.6) that the selected features did not exhibit a too high degree of collinearity, and this was not the case: all were under 10.

The model based on system factors alone performed at best a R^2 value of ca .23. In other words: even though we could clearly show that there is a significant effect of our system factors on QoE, a model predicting an individual's QoE using only the system factors still performs poorly. We can see in Fig. 6.17 that all models perform better than the system factors alone but there are substantial differences on how much the different factors considered improve prediction accuracy. The inclusion of user factors proved to be beneficial in all cases; when adding all factors (user model in table 6.5), the model performed best. It should be noted that the model based on solely current state factors yielded just slightly lower accuracy than the model including all user factor categories.

Interaction cues improve the model compared to a model including only system factors substantially (R^2 of 0.53). Speech features improve the model more, while the visual information yields only little improvement (compare visual, speech and interaction model in table 6.5). The combination of interaction and user factors (interac-

tion + background and interaction + current state) performed better than interaction or user factors alone (interaction and user). The final model including all features achieves an R^2 of 0.63. and outperforms all other models. We can see that if we want to predict the QoE of an individual, system factors alone do not provide sufficient information; especially including the dynamic factors current state and interaction more than doubled our prediction accuracy.

6.4 Discussion

In this chapter, we analyzed how bitrate and packet-loss impact interaction and engagement in videoconferencing. In doing so we showed that when combined with information on the user background, current state and behavior, they can predict QoE with relatively high accuracy. We used a scenario in which video usage was particularly stressed, with the conversation focusing around objects at hand. Compared to audio-only solutions (e.g. telephone conference), video conferencing shows best its added value in these situations, as the object of conversation can simply be shown. Thus, although a number of scenarios exist where videoconferencing is used without a strong visual focus, we wanted to investigate video quality in a scenario in which the visual channel actually played an important role in the conversation. Because of this setup, it is important to note that the results of this study are likely to be more sensitive compared to situations with no direct use of the visual channel.

The quality perception was clearly impacted by the low quality (256kbps), but users did not seem to appreciate the differences between medium (1mpbs) and high quality (4mbps). Low quality video, with visible encoding and loss artifacts, was rated by users mostly as poor quality (mean unfiltered/filtered 2.33/2.52). For overall quality users tended more towards medium quality(mean unfiltered/filtered 2.65/2.97). Both medium and high bitrate delivered a good quality of experience with most users giving good or excellent ratings (overall quality 3.95/4.29). Except for our anomalous group (1.5 – 4.0), the difference between both conditions was small (0.33). The video quality was again rated slightly lower (3.86 - 4.15). The influence of packet loss, while significant, was rather small (average 0.367/ 0.425 difference between a loss and no loss condition). Surprisingly, there

Table 6.5: Input features, selected features, and diagnostics R^2 and RMSE for the constructed models. Speech features include: speaker alternation rate, pauses (count, % duration, duration), utterances, turns, floors and group turns (duration, count per min, % count, % duration) uninterrupted turns (count per min), simultaneous starts (count per min, interrupted/ interrupter count per min,) interruption with speaker change (count per min), overlaps interrupter/interrupted (count per min)

category	features	Selected features	R^2	rmse
system	bitrate, loss rate	bitrate, loss rate	0.23	0.39
background	system + priorexp, age	bitrate, loss rate, priorexp, age	0.31	0.35
current state	system + engagement,, enjoyment	bitrate, loss rate, engagement, enjoyment	0.50	0.26
user	System + background + current state	bitrate, loss rate, priorexp, enjoyment, age, engagement	0.53	0.24
visual	system + Temporal Activity (TA) and distance to screen (mean, sd)	bitrate, loss rate, mean TA	0.24	0.39
speech	system + see caption	bitrate, loss rate, pauses (count, % duration), floors (count per min), simultaneous starts (interrupter count per min), overlaps (duration, interrupted count per min), group turns (duration), blocks (% count)	0.52	0.25
interaction	System + visual + speech	bitrate, loss rate, mean TA + pauses (count, % duration), floors (count per min), simultaneous starts (interrupter count per min), overlaps (duration), group turns (duration)	0.53	0.24
Interaction + background	System + visual + speech + background	bitrate, loss rate, mean TA, pauses (count, % duration), floors (count per min), simultaneous starts (interrupter count per min), overlaps (duration), utterances (duration, % count), priorexp, age	0.56	0.23
Interaction + current state	System + visual + speech + current state	bitrate, loss rate, mean TA, pauses (count, % duration), floors (% count, count per min), simultaneous starts (interrupter count per min), group turns (duration), utterances (% count), engagement, enjoyment	0.61	0.20
all	System + background + current state + visual + speech	bitrate, loss rate, mean TA, pauses (count, % duration), floors (% count, count per min), simultaneous starts (interrupter count per min), utterances (% count), engagement, age	0.63	0.19

was no interaction between packet loss and encoding bitrate. The impact of packet loss was also so small, that it did not impact the audio quality ratings negatively. Our results showed that the perceived audio quality was clearly affected due to the impairment of the video quality. This confirms studies on cross-modal effects of audio and visual quality [176].

The initial analysis of our complete dataset showed that the differences between groups play a big role (compare R^2 values between models M2 and, M3 and M4 on unfiltered dataset). On closer observation this was one group who seemed to have a very different experience than the other groups . Only for video quality, leaving out the user factor (model M3) did not lead to a significant worse fit, suggesting that the impression of the video quality was the most consistent rated within groups from our questions. The impact of bitrate and loss was the strongest on video quality (highest marginal R^2). It is interesting that while the overall quality was quite consistently rated a bit higher than the video quality, the variance was mostly due to user factors and not on a group level. This suggests that the impact of video quality on the overall experience is, even in visual challenging scenarios, a personal preference.

Regarding interaction, we showed that, in comparison to higher bitrates, low encoding (256kbit) had a significant impact on movement patterns of users as well as on speech patterns. We showed that at this lowest quality level the interaction of our participants was affected: the visual channel was not sufficient for the details of the Lego® model and thus participants compensated this by talking more, as proven by an increase in the length of speaking turns.

All participants made heavy use of the video screen to show Lego parts and instructions. In the case of the lowest encoding bitrate, this interaction was hampered. We also observed comments from participants during the study confirming this. In one situation a participant that asked to look at the screen to see how the current step was, was answered (without looking up) 'that doesn't work anyway'. Sometimes participants requested repeatedly to hold the piece or instruction longer and closer to the camera. We conclude that the threshold to enable the visual interaction, without breaking the flow of the interaction, lies between 256kbit and 1Mbit (for 720p H.264 video). If the video quality is below this threshold, users can still

perform the task; however, they have to adapt their behavior. In our case that meant that participants spoke more and made less use of the visual channel. It was also the point in which QoE ratings were severely impacted. This might be the point where, in real life, users will look for alternatives to the current session: reschedule in the hope that the network quality will be better another time or change service altogether. To prevent this, given that video-conferencing is in most cases an over-the-top service, and disruptions due to bad network conditions that cannot be controlled by the videoconferencing provider, system providers may look into implementing tools to support users in their task. For example, we could imagine that in such cases a specialized 'present object' option, which takes a high quality picture that is transmitted additionally to the video stream, could easily improve the interaction. The network conditions were designed by typical conditions that we can find at home. While bandwidth is steadily increasing[5] in average, more and more different Internet access connection types are employed, increasing the diversity of available bandwidth. In the foreseeable future, users will be at locations in which no high-speed connection is possible. Our study showed that in such cases the video quality is not sufficient to support interaction that is visually focused on objects with small details – the very point where video conferencing excels over audio conferencing. H.265 has shown a reduced bitrate consumption, up to 50%, for providing similar perceived quality [181]. Although these measurements were not done with settings specialized for real-time conferencing, they highlight that we have not currently reached the limit of compression. This is an essential part for the future of video-conferencing systems. On the one hand, it raises the quality we can achieve for high-end conferencing connections (i.e. in connection with the more and more widespread 4K resolution screens) but on the other hand, it also raises the quality available for low-bandwidth connections. The latter is of special interest for video-conferencing to get the status of an 'always available' communication medium, even in remote locations with limited data access. This can be a valuable step into making video conferencing a tool that is available everywhere.

As detailed by conceptual models of the quality formation pro-

[5]http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html

cess [153, 118], the past experiences form a feedback loop influencing future QoE perceptions. While the effect has been studied in smaller scale [62, 150], long-term aspects are unclear. We hypothesized that age and previous experiences are related. Our data showed that young experienced participants gave higher quality ratings than the other groups. This may be related to habituation and sensitization [28]. This dual-process describes how we adapt over time to a stimulus: habituation if our reaction weakens (e.g. because the stimulus is repeatedly perceived as negligible), sensitization if the opposite happens. The typical quality degradation of streaming media over the internet, which we introduced in this study, has become common in the last two decades. Young participants grew up with this kind of artifacts, while older users are possibly more acquainted with previous audiovisual media (TV, DVD), which had no coding impairments or highly fluctuating quality. Our finding that QoE was less affected by system factors for younger participants than for older ones with similar level of experience suggests that the extent to which participants have dealt with degradation in the past plays a fundamental role in how their QoE is affected. Specifically, it seems that the more participants are used to a certain type of artifact, the less this affects QoE, following an habituation process [28]. This would need more investigation in the context of QoE, also accounting for quality fluctuations. However, if confirmed, this result may be a game-changer in quality optimization for future generations of users.

We also found that the QoE was influenced by engagement and enjoyment. While the main experience of users is shaped by the conversation they are having, they might notice good quality (and be delighted) or bad quality (and be annoyed by it). In our study we captured how engaged participants were into the session of building the Lego® model. For the majority of participants, the effect of bitrate and loss on engagement was not significant, but still participants with higher engagement reported a higher QoE. Even though the interaction had to change in presence of low bitrate, for the majority of participants this did not disrupt their flow, or at least their engagement. For a subgroup of participants, instead, engagement was influenced by bitrate and loss: they also reported a much stronger degradation in QoE. This goes along with our previous finding that for some participants even the audio quality seemed to be impaired

in presence of video impairments. For some users bad video quality seems to break the experience holistically, also affecting their current affective state.

These findings highlight the complex role that affective states play in QoE. At this point we cannot infer a clear cause-effect relationship between engagement and QoE, and it is possible that they are reciprocally interlinked. This is also mentioned in the Qualinet white paper [118], where affect is both an influencing factor of QoE, and influenced by it. By the inclusion of engagement in our models, we could improve their accuracy. We also explicitly found engagement to explain a large portion of individual and group differences in our models (see section VI.B). Hence, it is of core importance that objective measurements of QoE are enriched with information on the user's affective state. Yet, to be useful within QoE control cycles, insights about user affect must correctly represent the current state of the individual user, while measuring it in an unobtrusive manner. In our study, users were explicitly self-reporting their engagement level at the end of the sessions. Whereas self-report can be easily employed in user studies, it is not suited for real world systems. More promising solutions, inferring affective states from objective data such as behavior, social cues or sensory data [155] are currently being developed by the Affective Computing community. For example, engagement can nowadays be inferred from physiological measurements (e.g. GSR or EEG) [170]. Audio and video cues [105], which are anyway captured in videoconferencing, can improve accuracy in affective state prediction, also providing much finer granularity when the processing is done in real time, as done, for example, in [16]. Systems able to detect in real time mood and engagement changes, and correlate them with quality changes, could potentially much better understand whether the QoE is currently impacted by the network problems and act upon it. On the other hand, this would pose privacy concerns that are yet to be addressed.

Finally we trained a linear model to predict individual videoconferencing QoE based on all the investigated factors. The final model, including all factors, achieved an R^2 of .63, which is a considerable improvement over the R^2 value of .23 with only system factors. By only using system factors and interaction indicators (which can all be obtained automatically, without the need for the user to self-report their

state), our model already achieved an R^2 of .53. The improvement due to the addition of visual interaction features to the audio interaction ones was very small, but it is likely due to the relative coarse off-the-shelve metrics we used. Prediction could be improved with video analysis tools specialized for the video conferencing scenario. An application of such tools could be to estimate the importance of the video to the user in the current situation. The best performance improvement of our models was achieved by including current state features (R^2 .52).

6.5 Conclusion

In this chapter we presented an extensive analysis of a study that investigated the impact of video impairments on videoconferencing QoE. We specifically focused on a scenario where users are dealing with a conversation task that requires both audio and visual interaction, and video usage is particularly stressed.

We investigated different encoding and packet loss settings that could be typical situations at home. Participants that have to use the low bitrate encoding (up 256kbs/down 756kbs) still have an okay (poor-fair) experience. If the connection has also packet loss, similar to an impaired wireless connection, the ratings tend to be poorer. We would conclude that 256kbs delivers an overall satisfactory experience, but participants will not be impressed by the video and often have a poor impression of it. The difference in quality perception between 1024kbs and 4096kbs is rather small (too small to have a significant impact in our study) and most people will give it a good or even excellent rating. Users on broadband connection thus rarely will have a better experience than users with a good DSL connection and we can conclude that 1Mbps is enough as encode bitrate for HD streams in video-conferencing. The impact of packet loss was noticeable but rather small in our experiment. In the high bitrate cases, it seems thus clear that it should be recommended to assign more bandwidth to forward error correction (FEC) than icnreasing the encoding bitrate more. In the low bitrate case both approaches seem to be viable. The bandwidth usage of FEC and quality improvements between 256kbs and 1024kbs would need to be studied in detail to give concrete advise. Packet loss is noticed but is acceptable to most

participants, so lossy protocols like UDP should still be the favorable choice for video-conferencing.

In this context of conversation focused around objects, we could clearly see that a video feed encoding bitrate of 256kbit was interfering with the user experience. It manifested in an interaction that was of slower pace and shifted focus from the video to the audio channel. We observed how impairments affected the QoE of young, experienced participant significantly less than the of other participants. We hypothesize that the reason behind this is more exposure with video degradation which lessens the effect on the experience. Furthermore, the QoE of more engaged participants was higher than that of the less engaged participants. This indicates how once a system has enabled users to engage in an interaction, participants will be quite forgiving about quality degradation, until it brings them out of the flow. With this data we tested predictive models, and by including all the examined factors we doubled the accuracy of our models. This research shows that if we want to accurately estimate the QoE of participants, knowing the system factors alone does not suffice. It is necessary to know the users and understand what they are doing to build systems that can actually balance the quality for the current situation.

7 Conclusions

This chapter concludes the thesis with a reflection on the impact of this thesis on assessing individual QoE in multi-party video-conferencing and discussing the future development of video-conferencing systems.

With this thesis, we aimed to understand the QoE of individual participants. The need for this personal QoE is twofold:

- From a user centric view
  Video-conferencing is a technology that enables small groups to communicate over distance. This is an incredibly broad purpose, since people communicate for many different reasons, in different situations and circumstances. Due to this vast variety of users and usage situations, estimating the QoE of users, while taking only the system factors into account, tells us little about the actual experience of the users.

- From a technological point of view
  Video-conferencing is a de-facto over-the-top service and as such, it is subject to network fluctuations common in the Internet. However, this also allows the sector to quickly adopt new technologies, make real-time adjustments and provide optimization for each individual user. If video-conferencing systems want to provide the optimal quality, they need to take into account the ongoing activity, the context of the conversation and the individual preferences of the user.

This thesis has approached this challenge based on a series of studies that looked at the QoE of users regarding not only system factors, but also the interplay of user behavior, user state and background. The thesis clearly showed that system factors alone are an insufficient source to explain the quality of experience in video-conferencing accurately. The current situation, the ongoing interaction and characteristics of each user are the factors that, in addition to the system quality, determine the QoE of the user. Considering these factors will allow for optimization of the QoE for the individual user. In order to have effective QoE optimization schemes, this thesis proposes the concept of personal QoE. The assumption behind personal QoE, which we have corroborated through the different studies in this thesis, is that being as specific as possible leads to the most accurate estimation of QoE.

The chapters in this thesis successively showed how non-system factors can be taken into account to build personal QoE. In chapter 3 we investigated a specific property of the multi-party scenario,

namely, different video-qualities presented at the same time. In particular, we investigated whether the different video qualities have a contrasting effect on each other. Since such effects have not yet been investigated, our goal was to test whether the effect generally existed. We chose a passive evaluation setup, which produces less realistic results for video-conferencing but more stable results. Implementing the study with a crowdsourcing approach allowed us to gather a large amount of ratings. In section 4, we then moved to interactive studies that looked at the impact of delay in multi-party video-conferencing . Previous works focus on the interactivity of a conversation, for example by measuring how fast the speakers change and the according effect on the perception of delay. To gain better insights into the differences among participants, we used the speaking time to cluster participants in active and non-active participants, who perceived the delay impact to different degrees. In the next chapter, we investigated the impact of video-quality on QoE regarding user factors. We focused especially on engagement with the given task. In doing so, we showed that less engaged participants noticed the quality degradations more strongly. Instead of clustering participants into user groups, we wanted to obtain more accurate estimations of the QoE for individual users. Therefore, we used a feature selection algorithm to build predictive models that estimated the QoE using user interaction, state and background additionally to system factors.

In the following sections, we take a closer look at the research questions we set out to answer at the beginning of this thesis.

7.1 Research Questions

In the following paragraphs, we show how each research question successively brought us closer to a personal QoE by exploring the non-technical influencing factors of QoE in relation to system factors: context, behavior and user. We started by showing how we examined the Context of multi-party video-conferencing in which the video-quality of each stream has to be considered in relation to the video-quality of the other streams. We then took the behavior of users into consideration by showing how users in the same session experienced delay differently, based on their conversational behavior. Then we showed how the current user state, namely the engagement

of individual participants, influenced the perception of video quality. Finally, we presented a model that utilizes a multitude of system, behavior and user factors to accurately predict the QoE of an individual.

Research Question 1: "What is the effect of the composition of video-streams from different participants in different encoding qualities on the overall perceived video quality?"

In the multi-party video-conferencing case, users are confronted with video-streams from different sources, which are likely to differ in quality. Due to the large amount of conditions that result from the possible combinations in a multi-party scenario, we opted for a crowd-sourcing study. Although not assessing the QoE that participants actually have under realistic conditions, this approach allowed us to concentrate on this research question. In chapter 3, we showed that such contrast effects exists: low video qualities are perceived worse depending on the number of high-quality streams co-presented. Vice-versa, high video qualities are perceived better depending on the number of low-quality streams co-presented. This is an important step towards being able to estimate QoE in realistic conditions. However, there is still work to be done to incorporate these findings into usable models. As a first step, the effect has to be evaluated in interactive scenarios. Furthermore, to understand the effect in realistic situations, the relation with user behavior (for which we already provided some initial findings), has to be examined further. Lastly, the work has to be extended for different layouts, such as the Google Hangout style "large speaker image and thumbnails".

Research Question 2: "How does the delay impact the QoE of different participants based on their conversational behavior?"

The previous section discussed the aspect of video conferencing related to experiencing different video qualities at the same time. Another aspect of video-conferencing is that the conversation structure of small groups differs from the conversation structure of dyads. As delay directly interferes with the turn-taking systematic, we conducted an interactive study regarding delay in multi-party video-conferencing . We employed a discussion scenario in which one of the participants had a moderating role. This setup fosters the emergence of a central conversational role, which we detected by clustering

participants by their percental speaking time. We showed that active participants were negatively impacted earlier by the delay than non-active participants. Hence, we are now able to differentiate the impact of delay on QoE for participants in the same session, based on their conversational behavior.

Research Question 3: "Is the QoE of participants related to their engagement?"

After exploring contextual and behavioral properties of multi-party video-conferencing , we focused on user factors. QoE tries to capture the appreciation (or annoyance) a user has with a system or service ,. It is a rather holistic approach that explicitly aims to capture the relation between system factors and user-specific factors. The previous research question focused on a factor that directly interfered with the ability to converse. With this research question, we wanted to investigate the relation of a factor that is not directly involved in the communication process. We chose engagement (how deeply participants are immersed in the current task). We hypothesized that engagement would play a role in determining the QoE, however, it was unclear in which way. More engaged participants are more annoyed than less engaged participants, as they feel more disrupted in the conversation. However, the opposite effect was also possible. By being more focused on the conversation, more engaged participants potentially pay less attention to the technical aspects.

To examine the relationship between engagement and QoE, we conducted an interactive study in which we manipulated the video-quality. Besides the QoE ratings, participants also reported their engagement through an engagement questionnaire. The results showed that higher engaged participants also reported a higher QoE. This was true for the lowest video quality we presented, even though we detected that, in this condition, participants adapted their behavior to the low video-quality.

We showed that the engagement explains many of the differences in QoE that users report in the same situation. Thus, we know that engagement, as part of the current state of users, is an important factor in determining the individual QoE. To operationalize this factor in real world settings, the engagement would have to be determined with automatic methods, following some of the approaches reported by the social signal processing community [168, 7, 132].

Overall Research Question: "What is the QoE a particular user has in a video-conferencing session?"

In order to estimate the QoE of an individual, we showed that multiple factors have to be considered. We approached this by building a model including all data from the experiment previously described. This included the system factors, behavioral data extracted from the audio and video streams, the reported engagement and further user aspects, such as age and reported proficiency with video-conferencing systems. As these are too many features for classical statistical models, we used a feature selection algorithm to reduce the number of features and compared the performance of different models depending on which kind of features were used. The best model, including 13 different features, had an accuracy of slightly over 60%. The model thus achieved an accuracy close to the ca. 70% calculated with the variance component analysis. While different features (depending on what is available to the system or service) should be considered for a model to be implemented in real world settings, this thesis shows that non-technical factors can and should be used to construct models that accurately predict the QoE of individual users.

7.2 Future Work

In the course of this thesis, we have conducted a variety and number experiments, gathered data, observed conversations and talked to participants about their experiences with video-conferencing. It has become clear that the fluctuating quality provided by over-the-top services, together with often complex system setups cause troubles for users. Video-conferencing is connected with a feeling of uncertainty that results from these problems. It is no surprise that the greeting in video-conferencing has moved from "Hello" to "Can you see me?".

In the interviews with participants, it was often hard for them to specify what kind of quality degradations they typically encountered. But they often remembered vividly particular instances, how the resolution was simply too low to understand the expression of job interview partners, how a boring meeting made them realize which patterns a video codec distorts or the relief they felt when a high delay was gone.

Studies on QoE are usually conducted with an experimental de-

sign, meaning that they have a clear set of controlled parameters, which are set to specific levels (independent variables). User responses are gathered (dependent variables) and various other involved factors are measured (covariates). This approach enables us to quantify the relationship between independent and dependent variables, and refine the results with the help of the covariates. It also allows us to observe closely how participants cope with specific problems, such as changing their conversational behavior in case of long delay. However, what is missing is complementary research that explores the situational context in which such conversations are embedded in the real world.

This aspect is not about designing more or better scenarios for the tasks during these studies, it is about the meaning a real conversation has for users. Participants of research studies are relatively indifferent to the conversation itself. Most of the time they have a completely external motivation for participating: they are compensated for their efforts. And still, even in this artificial context, we found that the degree to which participants were engaged in the task had a significant impact on how they experienced quality.

In this thesis, we showed that user behavior and individual user factors contribute to a larger part of the QoE ratings than the system factors. We quantified the relationship between some behavioral factors or user state factors and QoE. While this allowed us to differentiate the QoE each user has and work towards prediction models that are more accurate, it also shows the limits of what we can understand about the QoE of users from such studies. In the laboratory environment, we cannot capture the moments when the conversation was of particular importance to them and it was disrupted by the quality of the system they were using. Now, one could argue that it is obvious that if users care more about the conversations they in turn care more about the quality. But why should we, as evaluators, be concerned with this? The answer is that exactly in these instances system quality is of crucial importance to the users. These are the moments they remember when they reflect on their actual experiences and these are the ones we should try to improve.

The ITU and QoE community has build a comprehensive stack of rules meant to help practitioners avoid mistakes and keep results comparable between laboratories. However, many of such studies

provide results that are on the level of perceptional quality. They obtain just-noticeable differences for various levels of a system's parameters. This kind of data can guide the development of encoding algorithm and help find optimal configurations for a given bandwidth. Nevertheless, they are also limited to this area; they do not tell us how to construct the next generation of video-conferencing systems, as they give little insights on how we can improve the experience of users in critical situations.

To really improve the current video-conferencing systems we need to focus on two types of situations: the ones where the quality is extremely bad and the ones where ample resources are available and we can deliver exceptional quality. Those peak moments will stay in the mind of the user. Both situations have particular challenges associated with them.

Most of the time, bad system quality is caused by factors that are out of the control of the video-conferencing provider, such as bad network quality. We probably need new approaches, which bring new technologies into current systems. As an over-the-top service, if there is a high delay, the video-conferencing service provider can usually do nothing. But researchers on social signal processing, for example, are working on predicting who will speak next [133]. Such guesses are based on gaze patterns and preparation of the participant by drawing breath. The service could indicate in the conversation that a participant wants to speak before he or she actually starts talking.

To focus on situations in which we can provide exceptional quality for users, we need a better understanding of how users accustom to quality levels. It is clear that previous experiences are strongly linked to our expectations and thus to the impression of quality we eventually have. When we look at delay studies conducted in the 90s and in recent years, we notice that current users seem to be less bothered by delay [34]. Comparing studies investigating video-qualities show the opposite effect, older studies with far worse video quality than current systems, achieve, in their respectively highest conditions, similar excellent ratings than todays system [84, 14, 109]. The reasons for this are most likely the improved video quality in many areas (such as HD ready, full HD and ultra HD in the video market) and the fact that users are generally exposed to higher level of video quality. If a service provider wants to establish itself as a high quality

provider, it has to provide outstanding quality. The quality needs to elicit a "wow"-moment from users instead of a simple "no problems"-description. In the discussion with participants from our studies, we could see that the expectations and the impressions with the tested systems had a broad range from "wow, this was quality like on TV" to "yeah, that's how my Skype calls usually go". Understanding when a user will be impressed or disappointed with a certain quality is linked to understanding the base level of quality he or she is used to.

In this thesis we have shown that the current situation and individual preferences play a crucial role in shaping QoE. Understanding the situational context is currently only operationalized with audio cues, and limited to for dynamic Goolge Hangout style layouts and the insertion of comfort noise (gentle noise inserted during silences to reassure that the connection is not lost). A deeper level of conversation analysis has to be used to understand how serious the impact of delay currently is. Visual cue extraction is, to our knowledge, not utilized in any current video-communication system, even though the importance of the video channel differs greatly between and in conversations. In some instances, the video channel is hardly ever looked at because participants are involved in other activities that need their visual focus. On the other hand, there are moments when things are presented directly into the camera, getting the full attention of every participant.

To achieve an understanding of the QoE an individual has, it is necessary to take into account the individual preferences, the current state of mind and idiosyncrasies of the user besides the situational context and the ongoing behavior. This seems to be an opportunity for platform operators that do not only provide video-conferencing services but also other services, such as social networks and video-streaming services. Both Facebook and Google have integrated video-conferencing services and a vast knowledge of activities and preferences of their users. Based on the experiences we made in improving our models with the small amount of data we gathered from our trials, this approach could work for building accurate models that give a very detailed insights into the users' QoE.

However, if we have the integration of situational context, behavioral cues and individual aspects, conferencing systems will still face situations in which the network will experience fluctuations under

which no satisfactory quality can be provided. In many of the situations we observed during our trial, it could be imagined for the system to provide explicit mechanisms that would alleviate the task at hand. Many participants stated that when they experience severe quality degradations for some time, they often reschedule their conversation if possible. The explicit turn-taking mechanisms we observed could be aided with explicit support in the tool, with the possibility to hand over turns and indicate the wish to talk. Similarly, in the Lego task, participants struggled to make out the small details of the Lego model. Here, an "object presentation mode" that would show the object presented to the user in high quality would provide the needed details, even if with a long delay and out of sync with the audio.

The last years have shown that video-conferencing is a technology that has established itself in people's daily lives. People use the existing infrastructure with their computers and mobile devices via the Internet, in a wide variety of situations and manners. In this thesis, we have shown multiple approaches on how we can better understand the actual QoE people are having in these situations and pointed out how technology can be improved with the help of these insights, with the ultimate goal of being able to have trouble-free conversations over distance.

Bibliography

[1] Hirotogu Akaike. “Information Theory and an Extension of the Maximum Likelihood Principle”. In: *Selected Papers of Hirotugu Akaike*. Ed. by Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa. Springer Series in Statistics. Springer New York, 1998, pp. 199–213. ISBN: 978-1-4612-7248-9 978-1-4612-1694-0. DOI: 10.1007/978-1-4612-1694-0_15.

[2] R. Aldridge et al. “Recency effect in the subjective assessment of digitally-coded television pictures”. In: *Fifth International Conference on Image Processing and its Applications, 1995*. Fifth International Conference on Image Processing and its Applications, 1995. 1995, pp. 336–339. DOI: 10.1049/cp:19950676.

[3] Lamine Amour et al. “An Open Source Platform for Perceived Video Quality Evaluation”. In: *Proceedings of the 11th ACM Symposium on QoS and Security for Wireless and Mobile Networks*. Q2SWinet ’15. New York, NY, USA: ACM, 2015, pp. 139–140. ISBN: 978-1-4503-3757-1. DOI: 10.1145/2815317.2815344.

[4] Anne H. Anderson et al. “Impact of video-mediated communication on simulated service encounters”. In: *Interacting with Computers* 8.2 (1996), pp. 193–206. ISSN: 0953-5438. DOI: 10.1016/0953-5438(96)01025-9.

[5] Savvas Argyropoulos et al. “No-reference bit stream model for video quality assessment of h. 264/AVC video based on packet loss visibility”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. 2011, pp. 1169–1172.

[6] Douglas Bates et al. “Fitting Linear Mixed-Effects Models Using lme4”. In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48. DOI: 10.18637/jss.v067.i01.

[7] Roman Bednarik, Shahram Eivazi, and Michal Hradis. “Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement”. In: *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*. ACM, 2012, p. 10.

[8] Steven A. Beebe and John T. Masterson. *Communicating in Small Groups: Principles and Practices*. Longman, 1997. 388 pp. ISBN: 978-0-673-98080-9.

[9] John G. Beerends and Frank E. De Caluwe. “The influence of video quality on perceived audio quality and vice versa”. In: *Journal of the Audio Engineering Society* 47.5 (1999), pp. 355–362.

[10] B. Belmudez and S. Moller. “An Approach for Modeling the Effects of Video Resolution and Size on the Perceived Visual Quality”. In: IEEE, 2011, pp. 464–469. ISBN: 978-1-4577-2015-4 978-0-7695-4589-9. DOI: 10.1109/ISM.2011.82.

[11] Benjamin Belmudez. *Audiovisual Quality Assessment and Prediction for Videotelephony*. T-Labs Series in Telecommunication Services. Cham: Springer International Publishing, 2015. ISBN: 978-3-319-14165-7 978-3-319-14166-4.

[12] Benjamin Belmudez and Sebastian Möller. “Audiovisual quality integration for interactive communications”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2013.1 (2013), pp. 1–23. ISSN: 1687-4722. DOI: 10.1186/1687-4722-2013-24.

[13] Benjamin Belmudez et al. “Audio and video channel impact on perceived audio-visual quality in different interactive contexts”. In: *Proc. MMSP*. IEEE, 2009, pp. 1–5.

[14] Gunilla Berndtsson, Mats Folkesson, and Valentin Kulyk. “Subjective quality assessment of video conferences and telemeetings”. In: *Proc. 19th PV*. 2012, pp. 25–30.

[15] Anne Berry. *Spanish and American Turn-Taking Styles: A Comparative Study*. 1994, p180–90.

[16] Abhishek Bhattacharya, Wanmin Wu, and Zhenyu Yang. "Quality of experience evaluation of voice communication systems using affect-based approach". In: *Proc. ACM MM*. MM '11. New York, NY, USA: ACM, 2011, pp. 929–932. ISBN: 978-1-4503-0616-4. DOI: 10.1145/2072298.2071905.

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

[18] F. Brauer, M.S. Ehsan, and G. Kubin. "Subjective evaluation of conversational multimedia quality in IP networks". In: *Proc.10th MMSP*. 2008 IEEE 10th Workshop on Multimedia Signal Processing. 2008, pp. 872–876. DOI: 10.1109/MMSP.2008.4665196.

[19] Sheliey Buchinger and Helmut Hlavacs. "Subjective Quality of Mobile MPEG-4 Videos with Different Frame Rates". In: *J. Mob. Multimed.* 1.4 (2005), pp. 327–341. ISSN: 1550-4646.

[20] William A. S. Buxton. "Telepresence: Integrating shared task and person spaces. Paper presented at". In: *the Proceedings of Graphics Interface*. 1992.

[21] John A Campbell. "Participation in videoconferenced meetings: user disposition and meeting context". In: *Information & Management* 34.6 (1998), pp. 329–338. ISSN: 0378-7206. DOI: 10.1016/S0378-7206(98)00073-1.

[22] Maria-Dolores Cano and Fernando Cerdan. "Subjective QoE analysis of VoIP applications in a wireless campus environment". In: *Telecommunication Systems* 49.1 (2012), pp. 5–15. ISSN: 1018-4864, 1572-9451. DOI: 10.1007/s11235-010-9348-5.

[23] Kuan-Ta Chen et al. "Quadrant of euphoria: a crowdsourcing platform for QoE assessment". In: *IEEE Network* 24.2 (2010), pp. 28–35. ISSN: 0890-8044. DOI: 10.1109/MNET.2010.5430141.

[24] Therdpong Daengsi, Kiattisak Yochanang, and Pongpisit Wuttidittachotti. "A study of perceptual VoIP quality evaluation with thai users and codec selection using voice quality-Bandwidth tradeoff analysis". In: *ICT Convergence (ICTC), 2013 International Conference on*. IEEE, 2013, pp. 691–696.

[25] A.C. Davison. *Statistical Models*. 1 edition. Cambridge University Press, 2008. 738 pp. ISBN: 978-0-521-73449-3.

[26] Katrien De Moor et al. "Chamber QoE: a multi-instrumental approach to explore affective aspects in relation to quality of experience". In: *Proc. SPIE 9014, Human Vision and Electronic Imaging XIX*. 2014. DOI: 10.1117/12.2042243.

[27] Thomas J. DiCiccio and Bradley Efron. "Bootstrap Confidence Intervals". In: *Statistical Science* 11.3 (1996), pp. 189–212. ISSN: 0883-4237.

[28] Michael P. Domjan. *The Principles of Learning and Behavior*. 7 edition. Stamford, CT: Cengage Learning, 2014. 448 pp. ISBN: 978-1-285-08856-3.

[29] James P. Duncanson. "The average telephone call is better than the average telephone call". In: *The Public Opinion Quarterly* 33.1 (1969), pp. 112–116.

[30] James P. Duncanson and Arthur D. Williams. "Video Conferencing: Reactions of Users". In: *Human Factors* 15.5 (1973), pp. 471–485. ISSN: 0018-7208. DOI: 10.1177/001872087301500504.

[31] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[32] Sebastian Egger, Peter Reichl, and Katrin Schoenenberg. "Quality of Experience and Interactivity". In: *Quality of Experience*. Ed. by Sebastian Möller and Alexander Raake. T-Labs Series in Telecommunication Services. Springer International Publishing, 2014, pp. 149–161. ISBN: 978-3-319-02680-0 978-3-319-02681-7. DOI: 10.1007/978-3-319-02681-7_11.

[33] Sebastian Egger, Michal Ries, and Peter 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.

[34] Sebastian Egger, Raimund Schatz, and Stefan Scherer. "It takes two to tango-assessing the impact of delay on conversational interactivity on perceived speech quality". In: *Proc. 11th ISCA*. 2010.

[35] Raquel Fernández et al. "Interaction in task-oriented human-human dialogue: The effects of different turn-taking policies". In: *Spoken Language Technology Workshop, 2006. IEEE*. 2006, pp. 206–209.

[36] Robert S. Fish et al. "Evaluating video as a technology for informal communication". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '92. New York, NY, USA: ACM, 1992, pp. 37–48. ISBN: 0-89791-513-5. DOI: 10.1145/142750.142755.

[37] Barbara L. Fredrickson. "Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions". In: *Cognition & Emotion* 14.4 (2000), pp. 577–606.

[38] P. Fröhlich et al. "QoE in 10 seconds: Are short video clip lengths sufficient for Quality of Experience assessment?" In: *2012 Fourth International Workshop on Quality of Multimedia Experience*. 2012 Fourth International Workshop on Quality of Multimedia Experience. 2012, pp. 242–247. DOI: 10.1109/QoMEX.2012.6263851.

[39] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin, 2001.

[40] Stephen Gale. "Human aspects of interactive multimedia communication". In: *Interacting with computers* 2.2 (1990), pp. 175–189.

[41] Marie-Neige Garcia and Alexander Raake. "Quality impact of video format and scaling in the context of IPTV". In: *Third International Workshop on Perceptual Quality of Systems (PQS)*. 2010, pp. 119–124.

[42] Bruno Gardlo, Sebastian Egger, and Tobias Hossfeld. "Do Scale-Design and Training Matter for Video QoE Assessments through Crowdsourcing?" In: CrowdMM. Brisbane: ACM, 2015.

[43] Erik Geelhoed et al. *Effects of Latency on Telepresence*. HP labs technical report: HPL-2009-120 http://www. hpl. hp. com/techreports/2009/HPL-2009-120. html, 2009.

[44] D. Geerts et al. "Linking an integrated framework with appropriate methods for measuring QoE". In: *QoMEX'10*. 2010, pp. 158–163.

[45] G. Ghinea and J. P. Thomas. "QoS impact on user perception and understanding of multimedia video clips". In: *Proceedings of the sixth ACM international conference on Multimedia*. MULTIMEDIA '98. New York, NY, USA: ACM, 1998, pp. 49–54. ISBN: 0-201-30990-4. DOI: 10.1145/290747.290754.

[46] Jason Greengrass, John Evans, and Ali C. Begen. "Not all packets are equal, part 2: The impact of network packet loss on video quality". In: *Internet Computing, IEEE* 13.2 (2009), pp. 74–82.

[47] Marie Guéguin et al. "On the Evaluation of the Conversational Speech Quality in Telecommunications". In: *EURASIP Journal on Advances in Signal Processing* 2008.1 (2008), p. 185248. ISSN: 1687-6180. DOI: 10.1155/2008/185248.

[48] Stephen R. Gulliver and Gheorghita Ghinea. "Defining user perception of distributed multimedia quality". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)* 2.4 (2006), pp. 241–257.

[49] Simon NB Gunkel, Marwin Schmitt, and Pablo Cesar. "A QoE study of different stream and layout configurations in video conferencing under limited network conditions". In: *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[50] Florian Hammer, Peter Reichl, and Alexander Raake. "The well-tempered conversation: interactivity, delay and perceptual VoIP quality". In: *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*. Vol. 1. 2005, pp. 244–249.

[51] D.S. Hands. "A Basic Multimedia Quality Model". In: *IEEE Transactions on Multimedia* 6.6 (2004), pp. 806–816. ISSN: 1520-9210. DOI: 10.1109/TMM.2004.837233.

[52] David S. Hands and S. E. Avons. "Recency and duration neglect in subjective assessment of television picture quality". In: *Applied Cognitive Psychology* 15.6 (2001), pp. 639–657. ISSN: 1099-0720. DOI: 10.1002/acp.731.

[53] A. Paul Hare. "Types of Roles in Small Groups A Bit of History and a Current Perspective". In: *Small Group Research* 25.3 (1994), pp. 433–448. ISSN: 1046-4964, 1552-8278. DOI: 10.1177/1046496494253005.

[54] Takanori Hayashi et al. "Multimedia quality integration function for videophone services". In: *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE*. IEEE, 2007, pp. 2735–2739.

[55] T. Hoßfeld et al. "Assessing effect sizes of influence factors towards a QoE model for HTTP adaptive streaming". In: *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*. 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX). 2014, pp. 111–116. DOI: 10.1109/QoMEX.2014.6982305.

[56] T. Hoßfeld et al. "Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing". In: *IEEE Transactions on Multimedia* 16.2 (2014), pp. 541–558. ISSN: 1520-9210. DOI: 10.1109/TMM.2013.2291663.

[57] T. Hoßfeld et al. "Quantification of YouTube QoE via Crowdsourcing". In: *2011 IEEE International Symposium on Multimedia (ISM)*. 2011 IEEE International Symposium on Multimedia (ISM). 2011, pp. 494–499. DOI: 10.1109/ISM.2011.87.

[58] Tobias Hoßfeld and Christian Keimel. "Crowdsourcing in QoE Evaluation". In: *Quality of Experience*. Ed. by Sebastian Möller and Alexander Raake. T-Labs Series in Telecommunication Services. Springer International Publishing, 2014, pp. 315–327. ISBN: 978-3-319-02680-0 978-3-319-02681-7. DOI: 10.1007/978-3-319-02681-7_21.

[59] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. "SOS: The MOS is not enough!" In: *Proc. 3rd QoMEX*. IEEE, 2011, pp. 131–136.

[60] Tobias Hoßfeld et al. "Initial delay vs. interruptions: between the devil and the deep blue sea". In: *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. 2012, pp. 1–6.

[61] Tobias Hoßfeld et al. "Pippi Longstocking Calculus for Temporal Stimuli Pattern on YouTube QoE: 1+1=3 and 1·4≠4·1". In: *Proceedings of the 5th Workshop on Mobile Video*. MoVid '13. New York, NY, USA: ACM, 2013, pp. 37–42. ISBN: 978-1-4503-1893-8. DOI: 10.1145/2457413.2457422.

[62] Tobias Hoßfeld et al. "The Memory Effect and Its Implications on Web QoE Modeling". In: *Proc. 23rd ITC*. ITC '11. San Francisco, California: International Teletraffic Congress, 2011, pp. 103–110. ISBN: 978-0-9836283-0-9.

[63] Jan Holub and Ondrej Tomiska. "Delay Effect on Conversational Quality in Telecommunication Networks: Do We Mind?" In: *Wireless Technology*. Ed. by Steven Powell and J.P. Shim. Vol. 44. Lecture Notes in Electrical Engineering. Springer US, 2009, pp. 91–98. ISBN: 978-0-387-71787-6.

[64] Te-Yuan Huang et al. "Could Skype be more satisfying? a QoE-centric study of the FEC mechanism in an internet-scale VoIP system". In: *IEEE Network* 24.2 (2010).

[65] ITU-R. *BT.1359-1 : Relative timing of sound and vision for broadcasting*. 1998.

[66] ITU-T. *G.1070 Opinion model for video-telephony applications*. 2012.

[67] ITU-T. *G.114 - One-way transmission time*. 2003.

[68] ITU-T. *P.10/G.100 Vocabulary for performance and quality of service*. 2017.

[69] ITU-T. *P.1301 - Subjective quality evaluation of audio and audiovisual multiparty telemeetings*. 2013.

[70] ITU-T. *P.1305 - Effect of delays on telemeeting quality*. 2016.

[71] ITU-T. *P.1312 - Method for the measurement of the communication effectiveness of multiparty telemeetings using task performance*. 2016.

[72] ITU-T. *P.800 : Methods for subjective determination of transmission quality*. 1996.

[73] ITU-T. *P.805 - Subjective evaluation of conversational quality*. 2007.

[74] ITU-T. *P.880 Continuous evaluation of time-varying speech quality*. 2004.

[75] ITU-T. *P.910 - Subjective video quality assessment methods for multimedia applications*. 1995.

[76] ITU-T. *P.911 - Subjective audiovisual quality assessment methods for multimedia applications*. 1998.

[77] ITU-T. *P.913: Subjective video quality assessment methods for multimedia applications*. 1999.

[78] ITU-T. *P.920 - Interactive test methods for audiovisual communications*. 2000.

[79] Satoru Iai, Takaaki Kurita, and Nobuhiko Kitawaki. "Quality requirements for multimedia communication services and terminals-interaction of speech and video delays". In: *Global Telecommunications Conference, 1993, including a Communications Theory Mini-Conference. Technical Program Conference Record, IEEE in Houston. GLOBECOM'93., IEEE*. 1993, pp. 394–398.

[80] Jochen Issing and Nikolaus Farber. "Conversational quality as a function of delay and interactivity". In: *Proc. 20th SoftCOM*. 2012, pp. 1–5.

[81] Y. Ito and S. Tasaka. "QRPp1-1: User-level QoS assessment of a multipoint-to-multipoint TV conferencing application over IP networks". In: *Global Telecommunications Conference, 2006. GLOBECOM'06. IEEE*. 2006, pp. 1–6.

[82] Lucjan Janowski and Piotr Romaniak. "QoE as a Function of Frame Rate and Resolution Changes". In: *Future Multimedia Networking*. Ed. by Sherali Zeadally et al. Lecture Notes in Computer Science 6157. Springer Berlin Heidelberg, 2010, pp. 34–45. ISBN: 978-3-642-13788-4 978-3-642-13789-1.

[83] Jack Jansen and Dick.C.A. Bulterman. "User-Centric Video Delay Measurements". In: NOSSDAV. 2013.

[84] Coleen Jones and D. J. Atkinson. "Development of opinion-based audiovisual quality models for desktop video-teleconferencing". In: *Proc. 6th IWQoS*. IEEE, 1998, pp. 196–203.

[85] Evangelos Karapanos, Jean-Bernard Martens, and Marc Hassenzahl. "Accounting for diversity in subjective judgments". In: *Proc. CHI*. CHI '09. New York, NY, USA: ACM, 2009, pp. 639–648. ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1518801.

[86] Adam Kendon. "Some functions of gaze-direction in social interaction". In: *Acta Psychologica* 26 (1967), pp. 22–63. ISSN: 0001-6918. DOI: 10.1016/0001-6918(67)90005-4.

[87] David J. Ketchen and Christopher L. Shook. "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique". In: *Strategic Management Journal* 17.6 (1996), pp. 441–458. ISSN: 1097-0266. DOI: 10 . 1002 / (SICI) 1097 - 0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G.

[88] Jonathan K. Kies, Robert C. Williges, and Mary Beth Rosson. "Evaluating desktop video conferencing for distance learning". In: *Computers & Education* 28.2 (1997), pp. 79–91. ISSN: 0360-1315. DOI: 10.1016/S0360-1315(97)00004-3.

[89] David S. Kirk, Abigail Sellen, and Xiang Cao. "Home video communication: mediating 'closeness'". In: *Proc. CSCW*. CSCW '10. New York, NY, USA: ACM, 2010, pp. 135–144. ISBN: 978-1-60558-795-0. DOI: 10.1145/1718918.1718945.

[90] Nobuhiko Kitawaki and Kenzo Itoh. "Pure delay effects on speech quality in telecommunications". In: *Selected Areas in Communications, IEEE Journal on* 9.4 (1991), pp. 586–593.

[91] Masatomo Kobayashi et al. "Elderly user evaluation of mobile touchscreen interactions". In: *IFIP Conference on Human-Computer Interaction*. Springer, 2011, pp. 83–99.

[92] Sri Kurniawan. "Older people and mobile phones: A multi-method investigation". In: *International Journal of Human-Computer Studies*. Mobile human-computer interaction 66.12 (2008), pp. 889–901. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2008.03.002.

[93] Michael Kutner, Christopher Nachtsheim, and John Neter. *Applied Linear Regression Models- 4th Edition with Student CD*. 4 edition. Boston; New York: McGraw-Hill Education, 2004. 701 pp. ISBN: 978-0-07-301466-1.

[94] J. Clayton Lafferty, Alonzo William Pond, and Human Synergistics. *The Desert Survival Situation: Problem : a Group Decision Making Experience for Examining and Increasing Individual and Team Effectiveness*. Human Synergistics, 1974. 1 p.

[95] Jong-Seok Lee, Lutz Goldmann, and Touradj Ebrahimi. "A new analysis method for paired comparison and its application to 3D quality assessment". In: *Proceedings of the 19th ACM international conference on Multimedia*. MM '11. New York, NY, USA: ACM, 2011, pp. 1281–1284. ISBN: 978-1-4503-0616-4. DOI: 10.1145/2072298.2071994.

[96] Matthew Lombard et al. "The role of screen size in viewer responses to television fare". In: *Communication Reports* 10.1 (1997), pp. 95–106. ISSN: 0893-4215, 1745-1043. DOI: 10.1080/08934219709367663.

[97] D. Lottridge, M. Chignell, and M. Yasumura. "Identifying Emotion through Implicit and Explicit Measures: Cultural Differences, Cognitive Load, and Immersion". In: *Affective Computing, IEEE Transactions on* 99 (2011), pp. 1–1.

[98] James MacQueen and others. "Some methods for classification and analysis of multivariate observations". In: *Proc. 5th Berkeley Symp. on Math. Statist. and Prob.* Vol. 1. Oakland, CA, USA., 1967, pp. 281–297.

[99] Lilla Magyari and J. P. de Ruiter. "Prediction of Turn-Ends Based on Anticipation of Upcoming Words". In: *Frontiers in Psychology* 3 (2012). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2012.00376.

[100] John D. McCarthy, M. Angela Sasse, and Dimitrios Miras. "Sharp or smooth?: comparing the effects of quantization vs. frame rate for streamed video". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 535–542.

[101] Karan Mitra, Arkady Zaslavsky, and Christer AAhlund. "QoE modelling, measurement and prediction: A review". In: *arXiv preprint arXiv:1410.6952* (2014).

[102] Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Boston, MA: Springer US, 2000. ISBN: 978-1-4419-4989-9 978-1-4757-3117-0. DOI: 10.1007/978-1-4757-3117-0.

[103] Sebastian Möller and Alexander Raake. "Motivation and Introduction". In: *Quality of Experience*. T-Labs Series in Telecommunication Services. Springer, Cham, 2014, pp. 3–9. ISBN: 978-3-319-02680-0 978-3-319-02681-7. DOI: 10.1007/978-3-319-02681-7_1.

[104] Sebastian Möller and Alexander Raake, eds. *Quality of Experience*. T-Labs Series in Telecommunication Services. Cham: Springer International Publishing, 2014. ISBN: 978-3-319-02680-0 978-3-319-02681-7. DOI: 10.1007/978-3-319-02681-7.

[105] Wenxuan Mou, Hatice Gunes, and Ioannis Patras. "Alone Versus In-a-group: A Comparative Analysis of Facial Affect Recognition". In: *Proc. ACM MM*. MM '16. New York, NY, USA: ACM, 2016, pp. 521–525. ISBN: 978-1-4503-3603-1. DOI: 10.1145/2964284.2967276.

[106] Mu Mu et al. "Visibility of individual packet loss on H. 264 encoded video stream: a user study on the impact of packet loss on perceived video quality". In: *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2009, pp. 725302–725302.

[107] Shinichi Nakagawa and Holger Schielzeth. "A general and simple method for obtaining R2 from generalized linear mixed-effects models". In: *Methods in Ecology and Evolution* 4.2 (2013), pp. 133–142. ISSN: 2041-210X. DOI: 10.1111/j.2041-210x.2012.00261.x.

[108] Kaoru Nakazono, Yuji Nagashima, and Mina Terauchi. "Evaluation of Effect of Delay on Sign Video Communication". In: *Computers Helping People with Special Needs*. Ed. by Klaus Miesenberger et al. Vol. 4061. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2006, pp. 659–666. ISBN: 978-3-540-36020-9.

[109] M. Ndiaye et al. "Subjective assessment of the perceived quality of video calling services over a real LTE/4G network". In: *Proc. 7th QoMEX*. 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX). 2015, pp. 1–6. DOI: 10.1109/QoMEX.2015.7148096.

[110] David G. Novick, Brian Hansen, and Karen Ward. "Coordinating turn-taking with gaze". In: *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. Vol. 3. 1996, pp. 1888–1891.

[111] Heather L. O'Brien and Elaine G. Toms. "The development and evaluation of a survey to measure user engagement". In: *Journal of the American Society for Information Science and Technology* 61.1 (2010), pp. 50–69. ISSN: 15322882. DOI: 10.1002/asi.21229.

[112] Heather L. O'Brien and Elaine G. Toms. "What is user engagement? A conceptual framework for defining user engagement with technology". In: *Journal of the American Society for Information Science and Technology* 59.6 (2008), pp. 938–955. ISSN: 15322882, 15322890. DOI: 10.1002/asi.20801.

[113] Heather O'Brien and Paul Cairns, eds. *Why Engagement Matters*. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-27444-7 978-3-319-27446-1.

[114] Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. "Conversations over video conferences: an evaluation of the spoken aspects of video-mediated communication". In: *Hum.-Comput. Interact.* 8.4 (1993), pp. 389–428. ISSN: 0737-0024. DOI: 10.1207/s15327051hci0804_4.

[115] Claire O'Malley et al. "Comparison of face-to-face and video-mediated interaction". In: *Interacting with Computers* 8.2 (1996), pp. 177–192. ISSN: 0953-5438. DOI: 10.1016/0953-5438(96)01027-2.

[116] Robert B. Ochsman and Alphonse Chapanis. "The effects of 10 communication modes on the behavior of teams during co-operative problem-solving". In: *International Journal of Man-Machine Studies* 6.5 (1974), pp. 579–619.

[117] Joana Palhais, Rui S. Cruz, and Mário S. Nunes. "Quality of Experience Assessment in Internet TV". In: *SpringerLink*. Springer Berlin Heidelberg, pp. 261–274. DOI: 10.1007/978-3-642-30422-4_19.

[118] Patrick Le Callet, Andrew Perkis, and Sebastian Möller, eds. *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*. 2013.

[119] Stéphane Péchard et al. "From SD to HD television: effects of H.264 distortions versus display size on quality of experience". In: *International Conference on Image Processing*. Atlanta, United States, 2006, pp. 409–412. DOI: 10.1109/ICIP.2006.312480.

[120] Rohit Puri et al. "Forward error correction (FEC) codes based multiple description coding for internet video streaming and multicast". In: *Signal Processing: Image Communication*. Packet Video Communications 16.8 (2001), pp. 745–762. ISSN: 0923-5965. DOI: 10.1016/S0923-5965(01)00005-4.

[121] A. Raake and C. Schlegel. "Auditory assessment of conversational speech quality of traditional and spatialized teleconferences". In: *ITG Conference on Voice Communication [8. ITG-Fachtagung]*. ITG Conference on Voice Communication [8. ITG-Fachtagung]. 2008, pp. 1–4.

[122] Alexander Raake and Sebastian Egger. "Quality and Quality of Experience". In: *Quality of Experience*. T-Labs Series in Telecommunication Services. Springer, Cham, 2014, pp. 11–33. ISBN: 978-3-319-02680-0 978-3-319-02681-7. DOI: 10.1007/978-3-319-02681-7_2.

[123] Alexander Raake et al. "Predicting speech quality based on interactivity and delay." In: *INTERSPEECH*. 2013, pp. 1384–1388.

[124] Judith A. Redi et al. "How Passive Image Viewers Became Active Multimedia Users". In: *Visual Signal Quality Assessment*. Ed. by Chenwei Deng et al. Springer International Publishing, 2015, pp. 31–72. ISBN: 978-3-319-10367-9 978-3-319-10368-6. DOI: 10.1007/978-3-319-10368-6_2.

[125] P. Reichl et al. "Towards a comprehensive framework for QOE and user behavior modelling". In: *Proc. 7th QoMEX*. 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX). 2015, pp. 1–6. DOI: 10.1109/QoMEX.2015.7148138.

[126] Ulrich Reiter, Jari Korhonen, and Junyong You. "Comparing apples and oranges: assessment of the relative video quality in the presence of different types of distortions". In: *EURASIP Journal on Image and Video Processing* 2011.1 (2011), pp. 1–10.

[127] Flávio Ribeiro et al. "Crowdmos: An approach for crowdsourcing mean opinion score studies". In: *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2416–2419.

[128] P. Romaniak et al. "Perceptual quality assessment for H.264/AVC compression". In: *Proc. CCNC*. 2012 IEEE Consumer Communications and Networking Conference (CCNC). 2012, pp. 597–602. DOI: 10.1109/CCNC.2012.6181021.

[129] Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. "A Simplest Systematics for the Organization of Turn-Taking for Conversation". In: *Language* 50.4 (1974), pp. 696–735.

[130] I. Saidi et al. "Interactive vs. non-interactive subjective evaluation of IP network impairments on audiovisual quality in videoconferencing context". In: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX). 2016, pp. 1–6. DOI: 10.1109/QoMEX.2016.7498947.

[131] Emanuel Schegloff. "Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences". In: *Analyzing discourse: Text and talk, Georgetown University Roundtable on Languages and Linguistics*. 1982, pp. 71–93.

[132] Stefan Scherer et al. "Spotting Laughter in Natural Multiparty Conversations: A Comparison of Automatic Online and Offline Approaches Using Audiovisual Data". In: *ACM Trans. Interact. Intell. Syst.* 2.1 (2012), 4:1–4:31. ISSN: 2160-6455. DOI: 10.1145/2133366.2133370.

[133] David Schlangen. "From reaction to prediction: Experiments with computational models of turn-taking". In: *Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking* (2006).

[134] M. R. Schmitt et al. "Towards individual QoE for multi-party video conferencing". In: *IEEE Transactions on Multimedia* PP.99 (2017), pp. 1–1. ISSN: 1520-9210. DOI: 10.1109/TMM.2017.2777466.

[135] M. Schmitt, S. Gunkel, and P. Cesar. "A Quality of Experience Testbed for Video-Mediated Group Communication". In: *2013 IEEE International Symposium on Multimedia (ISM)*. 2013 IEEE International Symposium on Multimedia (ISM). 2013, pp. 514–515. DOI: 10.1109/ISM.2013.102.

[136] M. Schmitt et al. "1Mbps is enough: Video quality and individual idiosyncrasies in multiparty HD video-conferencing". In: *Proc. 8th QoMEX*. QoMEX. IEEE, 2016, pp. 1–6. DOI: 10.1109/QoMEX.2016.7498961.

[137] M. Schmitt et al. "Asymmetric delay in video-mediated group discussions". In: *Proc. 6th QoMEX*. Proc. 6th QoMEX. 2014, pp. 19–24. DOI: 10.1109/QoMEX.2014.6982280.

[138] Marwin Schmitt et al. "Mitigating Problems in Video-mediated Group Discussions: Towards Conversation Aware Video-conferencing Systems". In: *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*. UM3I '14. New York, NY, USA: ACM, 2014, pp. 39–44. ISBN: 978-1-4503-0652-2. DOI: 10.1145/2666242.2666247.

[139] Michael Schoeffler and Jürgen Herre. "About the different types of listeners for rating the overall listening experience". In: *ICMC*. 2014.

[140] Michael Schoeffler and Jürgen Herre. "The relationship between basic audio quality and overall listening experience". In: *The Journal of the Acoustical Society of America* 140.3 (2016), p. 2101. ISSN: 1520-8524. DOI: 10.1121/1.4963078.

[141] K. Schoenenberg, A. Raake, and P. Lebreton. "Conversational quality and visual interaction of video-telephony under synchronous and asynchronous transmission delay". In: *Proc. 6th QoMEX*. 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX). 2014, pp. 31–36. DOI: 10.1109/QoMEX.2014.6982282.

[142] Katrin Schoenenberg, Alexander Raake, and Judith Koeppe. "Why are you so slow? – Misattribution of transmission delay to attributes of the conversation partner at the far-end". In: *International Journal of Human-Computer Studies* 72.5 (2014), pp. 477–487. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2014.02.004.

[143] Katrin Schoenenberg et al. "On interaction behaviour in telephone conversations under transmission delay". In: *Speech Communication* 63–64 (2014), pp. 1–14. ISSN: 0167-6393. DOI: 10.1016/j.specom.2014.04.005.

[144] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering". In: *Proc. CVPR*. 2015, pp. 815–823.

[145] M. J. Scott et al. "Do Personality and Culture Influence Perceived Video Quality and Enjoyment?" In: *IEEE Trans. Multimedia* 18.9 (2016), pp. 1796–1807. ISSN: 1520-9210. DOI: 10.1109/TMM.2016.2574623.

[146] Michael James Scott et al. "Modelling Human Factors in Perceptual Multimedia Quality: On The Role of Personality and Culture". In: *Proc. 23rd MM*. MM '15. New York, NY, USA: ACM, 2015, pp. 481–490. ISBN: 978-1-4503-3459-4. DOI: 10.1145/2733373.2806254.

[147] Abigail J. Sellen. "Remote conversations: the effects of mediating talk with technology". In: *Hum.-Comput. Interact.* 10.4 (1995), pp. 401–444. ISSN: 0737-0024. DOI: 10.1207/s15327051hci1004_2.

[148] Kalpana Seshadrinathan and Alan C. Bovik. "Temporal hysteresis model of time varying subjective video quality". In: *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. 2011, pp. 1153–1156.

[149] Kalpana Seshadrinathan et al. “Study of Subjective and Objective Quality Assessment of Video”. In: *IEEE Transactions on Image Processing* 19.6 (2010), pp. 1427–1441. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2010.2042111.

[150] Junaid Shaikh et al. “Back to normal? Impact of temporally increasing network disturbances on QoE”. In: *Proc. GC Wkshps*. IEEE, 2013, pp. 1186–1191.

[151] Ana Paula Couto da Silva et al. “Quality assessment of interactive voice applications”. In: *Comput. Netw.* 52.6 (2008), pp. 1179–1192. ISSN: 1389-1286. DOI: 10.1016/j.comnet.2008.01.002.

[152] Janto Skowronek, Julian Herlinghaus, and Alexander Raake. “Quality assessment of asymmetric multiparty telephone conferences: a systematic method from technical degradations to perceived impairments.” In: *INTERSPEECH*. 2013, pp. 2604–2608.

[153] Janto Skowronek and Alexander Raake. “Conceptual Model of Multiparty Conferencing and Telemeeting Quality”. In: *Proc. 7th QoMEX*. 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX). IEEE, 2015.

[154] Janto Skowronek and Alexander Raake. “Investigating the Effect of Number of Interlocutors on the Quality of Experience for Multi-Party Audio Conferencing.” In: *INTERSPEECH*. 2011, pp. 829–832.

[155] Mohammad Soleymani and Maja Pantic. “Human-centered implicit tagging: Overview and perspectives”. In: *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 3304–3309.

[156] M. Soloducha et al. “Testing conversational quality of VoIP with different terminals and degradations”. In: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). 2017, pp. 1–3. DOI: 10.1109/QoMEX.2017.7965639.

[157] Erik Sorensen and M. Mikailesc. “Model-View-ViewModel (MVVM) Design Pattern using Windows Presentation Foundation (WPF) Technology”. In: *MegaByte Journal* 9.4 (2010), pp. 1–19.

[158] Ralf Steinmetz. “Human perception of jitter and media synchronization”. In: *Selected Areas in Communications, IEEE Journal on* 14.1 (1996), pp. 61–72.

[159] Jennifer Tam et al. "Video increases the perception of naturalness during remote interactions with latency". In: *Proc. of CHI'12*. CHI. CHI EA '12. New York, NY, USA: ACM, 2012, pp. 2045–2050. ISBN: 978-1-4503-1016-1. DOI: 10.1145/2223656.2223750.

[160] John C. Tang. *Why Do Users Like Video? Studies of Multimedia-Supported Collaboration*. Mountain View, CA, USA: Sun Microsystems, Inc., 1992.

[161] S. Tasaka and N. Misaki. "Customization of interactive services for QoE enhancement in audio-video transmission over bandwidth guaranteed IP networks". In: *2010 IEEE International Conference on Multimedia and Expo (ICME)*. 2010 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2010, pp. 546–551. ISBN: 978-1-4244-7491-2. DOI: 10.1109/ICME.2010.5582594.

[162] Shuji Tasaka et al. "The effectiveness of a QoE-based video output scheme for audio-video ip transmission". In: *Proceedings of the 16th ACM international conference on Multimedia*. MM '08. New York, NY, USA: ACM, 2008, pp. 259–268. ISBN: 978-1-60558-303-7. DOI: 10.1145/1459359.1459395.

[163] L. Ten Bosch, N. Oostdijk, and J. de Ruiter. "Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues". In: *Text, Speech and Dialogue*. 2004, pp. 563–570.

[164] Glenn Van Wallendael et al. "Perceptual quality of 4K-resolution video content compared to HD". In: IEEE, 2016, pp. 1–6. ISBN: 978-1-5090-0354-9. DOI: 10.1109/QoMEX.2016.7498935.

[165] Martín Varela, Lea Skorin-Kapov, and Touradj Ebrahimi. "Quality of Service Versus Quality of Experience". In: *Quality of Experience: Advanced Concepts, Applications and Methods*. Ed. by Sebastian Möller and Alexander Raake. Cham: Springer International Publishing, 2014, pp. 85–96. ISBN: 978-3-319-02681-7. DOI: 10.1007/978-3-319-02681-7_6.

[166] Mukundan Venkataraman, Mainak Chatterjee, and Siddhartha Chattopadhyay. "Evaluating quality of experience for streaming video in real time". In: *Proceedings of the 28th IEEE conference on Global telecommunications*. GLOBECOM'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 4279–4284. ISBN: 978-1-4244-4147-1.

[167] Bjørn J. Villa et al. "Investigating Quality of Experience in the context of adaptive video streaming: findings from an experimental user study". In: *Akademika forlag Stavanger, Norway* (2013).

[168] Oriol Vinyals, Dan Bohus, and Rich Caruana. "Learning Speaker, Addressee and Overlap Detection Models from Multimodal Streams". In: *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. ICMI '12. New York, NY, USA: ACM, 2012, pp. 417–424. ISBN: 978-1-4503-1467-1. DOI: 10.1145/2388676.2388770.

[169] Tim Walton and Michael Evans. "The role of human influence factors on overall listening experience". In: *Quality and User Experience* 3.1 (2018), p. 1. ISSN: 2366-0139, 2366-0147. DOI: 10.1007/s41233-017-0015-4.

[170] Hao-Chuan Wang and Chien-Tung Lai. "Kinect-taped Communication: Using Motion Sensing to Study Gesture Use and Similarity in Face-to-face and Computer-mediated Brainstorming". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. New York, NY, USA: ACM, 2014, pp. 3205–3214. ISBN: 978-1-4503-2473-1. DOI: 10.1145/2556288.2557060.

[171] Jian Wang et al. "Evaluation on perceptual audiovisual delay using average talkspurts and delay". In: *Image and Signal Processing (CISP), 201 3rd International Congress on*. Vol. 1. 2010, pp. 125–128.

[172] David J. Wheatley and Santosh Basapur. "A comparative evaluation of TV video telephony with webcam and face to face communication". In: *Proceedings of the seventh european conference on European interactive television conference*. EuroITV '09. New York, NY, USA: ACM, 2009, pp. 1–8. ISBN: 978-1-60558-340-2. DOI: 10.1145/1542084.1542086.

[173] Leslie A. Whitaker, Jennifer Hohne, and Deborah P. Birkmire-Peters. "Assessing cognitive workload metrics for evaluating telecommunication tasks". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 41. 1997, pp. 325–329.

[174] S. Winkler. "On the properties of subjective ratings in video quality experiments". In: *International Workshop on Quality of Multimedia Experience, 2009. QoMEx 2009*. International Workshop on Quality of Multimedia Experience, 2009. QoMEx 2009. 2009, pp. 139–144. DOI: 10.1109/QOMEX.2009.5246961.

[175] Stefan Winkler and Christof Faller. "Perceived audiovisual quality of low-bitrate multimedia content". In: *Multimedia, IEEE Transactions on* 8.5 (2006), pp. 973–980.

[176] J. You et al. “Perceptual-based quality assessment for audio–visual services: A survey”. In: *Signal Processing: Image Communication* 25.7 (2010), pp. 482–501.

[177] Yi Zhu, Alan Hanjalic, and Judith A. Redi. “QoE Prediction for Enriched Assessment of Individual Video Viewing Experience”. In: *Proceedings of the 2016 ACM on Multimedia Conference*. MM ’16. New York, NY, USA: ACM, 2016, pp. 801–810. ISBN: 978-1-4503-3603-1. DOI: 10.1145/2964284.2964330.

[178] Yi Zhu, Ingrid Heynderickx, and Judith A. Redi. “Understanding the role of social context and user factors in video Quality of Experience”. In: *Computers in Human Behavior* 49 (2015), pp. 412–426. ISSN: 07475632. DOI: 10.1016/j.chb.2015.02.054.

[179] T. Zinner et al. “Impact of frame rate and resolution on objective QoE metrics”. In: *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*. 2010 Second International Workshop on Quality of Multimedia Experience (QoMEX). 2010, pp. 29–34. DOI: 10.1109/QOMEX.2010.5518277.

[180] Hui Zou and Trevor Hastie. “Regularization and Variable Selection via the Elastic Net”. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320. ISSN: 1369-7412.

[181] Martin Řeřábek and Touradj Ebrahimi. “Comparison of compression efficiency between HEVC/H. 265 and VP9 based on subjective assessments”. In: *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2014, 92170U–92170U.