Quality and User Experience (2018) 3:7
https://doi.org/10.1007/s41233-018-0020-2

RESEARCH ARTICLE

The contrast effect: QoE of mixed video-qualities at the same time

Marwin Schmitt[1] · Dick C. A. Bulterman[1] · Pablo S. Cesar[1]

Received: 13 October 2017 / Published online: 26 May 2018

Abstract
In desktop multi-party video-conferencing videostreams of participants are delivered in different qualities, but we know little about how such composition of the screen affects the quality of experience. Do the different videostreams serve as indirect quality references and the perceived video quality is thus dependent on other streams in the same session? How is the relation between the perceived qualities of each stream and the perceived quality of the overall session? To answer these questions we conducted a crowdsourcing study, in which we gathered over 5000 perceived quality ratings of overall sessions and individual streams. Our results show a contrast effect: high quality streams are rated better when more low quality streams are co-present, and vice versa. In turn, the quality perception of an overall session can increase significantly by exchanging one low quality stream with a high quality one. When comparing the means of individual and overall ratings we can further observe that users are more critical when asked for individual streams than for an overall rating. However, the results show that while contrast effect exists, the effect is not strong enough, to optimize the experience by lowering the quality of other participants.

Keywords Video-conferencing · Video-quality · Multi-party · QoE · Crowd-sourcing · Subjective study · Perceived video-quality

Introduction

Multi-party videoconferencing has become a standard tool in the repertoire of real-time communication toolkits.Small group gatherings connected by video have moved from dedicated business solutions to publicly available services over the Internet. 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 the bandwidth available by him or herself, the bandwidth of the other participants and the architecture of the system in use. Service providers are interested in monitoring and predicting the QoE of participants to understand customer satisfaction and to take optimization and bandwidth allocation decisions. If the bandwidth is limited, it has to be divided between service and users. If these decisions are purely taken on the bandwidth level, we may divide the available bandwidth equally, but provide unequal QoE to the users [34]. In the context of multi-party video-conferencing, decisions have to be taken on how to distribute the bandwidth between the different streams. However, currently no knowledge exists about the effects of a mixed quality setup. In conceptual models [64] of the *quality formation process*, for multi-party tele-meetings, it is theorized that users aggregate the quality from different participant streams into a single judgment. But currently we know little about how this aggregation process works. The objective of this paper is to explore how the different qualities of individual streams are aggregated to an overall quality perception.

We know from previous works that the perception of video-quality is highly dependent on previously experienced qualities [31, 32, 35]. This effect can further be observed when comparing single stimulus and dual stimulus methods. It has been found that dual stimulus methods, which provides all participants with the same reference frame, yield less variance [41, 67]. The multi-party video-conferencing scenario, with different qualities between participants, adds

✉ Marwin Schmitt
schmitt@cwi.nl

Dick C. A. Bulterman
dcap@cwi.nl

Pablo S. Cesar
p.s.cesar@cwi.nl

[1] Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands

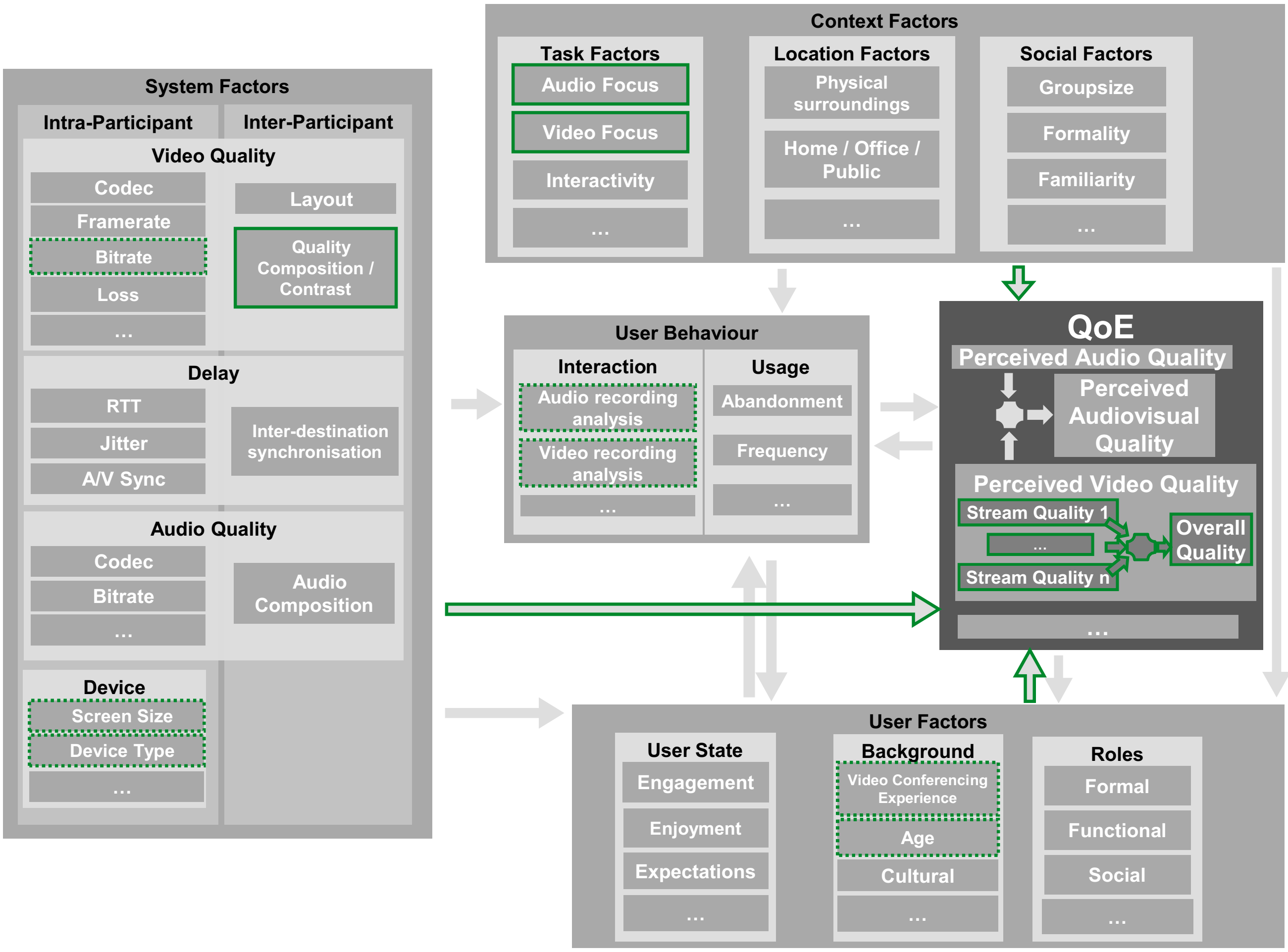


Fig. 1 Conceptual model of QoE (based on [47]) showing main influencing factors in multi-party video-conferencing and their relation. The particular factors under study in this work in green where dotted lines indicate covariates

complexity to the question of how internal reference influences the quality judgment: 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) [24, 52, 58]. To our knowledge, QoE in simultaneous mixed quality scenarios, has not been studied in any application scenario.

In this work we present a study that investigates the effects of co-present mixed qualities in a multi-party scenario. In Fig. 1 we show a conceptual model of the different influencing factors of QoE (based on the QoE and user behavior model [47]) 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 but are related to videoconferencing QoE. Factors and effects under study are held in dark blue, in contrast, elements in gray 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 realize 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 of which one focused on a conversation and the other one focused on assembling a Lego model. We encoded each video stream in two different qualities (256 and 1024 kbits). Due to the large amount of resulting conditions (two video clips with each four participants in all combinations of the two video qualities makes $2 \times 4^2 = 32$ conditions) we opted for a crowd-sourcing approach [30].

The study was split in three campaigns: in one we obtained individual ratings only, in one overall ratings and in the last both kind of ratings. After we filtered the crowdsourcing data by reliability criteria (see "Participants and reliability filtering" section for details) we had ratings from 412 participants giving both kind of ratings, 178 that provided only individual ratings and 180 giving an overall rating. This resulted in 5904 ratings that we analyzed to answer the following research questions.

- *Q1* What is the impact of mixed encoding qualities on overall perceived quality? How is the overall impression participants have about 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 impact the perceived quality stronger, similar to the stronger influence of bad quality peaks in time varying quality.
- *Q2* 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 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 between the streams. The overall perceived quality increased the most between 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 obtain 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 the paper is structured as follows: in "Related work" section we detail what research already exist on perceived quality and multi-party video-conferencing and how our findings extent this knowledge. In "Methodology" section we describe the detailed setup of our study and the methodology for the statistical analysis. "Results" section presents the analysis of our results and in "Discussion" section we discuss how these results can be used for improving predictions in mixed quality video-conferencing scenarios. Finally we conclude the paper by laying out where 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.

Related work

We start by presenting studies that have been conducted in the area of QoE in (multi-party) video-conferencing and in which areas our study extents the available knowledge. We then lay out which fundamental principles in the perception of quality have been found and detail the relation and differences to the effects examined in this work. Finally we detail different assessment methodologies used for QoE and how our study adapts these methods.

Subjective studies investigating QoE in video-conferencing

The research community has established a body of work that investigated several of the influencing factors [1] for QoE in video-conferencing. In the main focus are the system factors, with aspects like audio and video quality [6, 14, 40, 43, 62], audiovisual synchronization [58] and delay [51, 59]. Further studies regarding non-technical influencing factors have looked into interaction [20, 47, 59], the user state, such as engagement [52] and contextual factors like different devices [10, 66]. Most works in the area of multi-party conferencing have studied the impact of delay on QoE [10, 51, 55, 58, 62, 63, 65]. One of the main factors determining the impact of delay was found to be the interactivity of the conversation [19, 25, 37, 58]. The interactivity of the conversation is determined by analyzing the conversation in terms of the so called 'turn taking systematic', which describes the implicit organization of conversations (i.e. 'who speaks when') and is largely dependent on timings such as the length of pauses [49]. It was found that the impact of delay is generally perceived less strong in multi-party sessions than in two party sessions [54]. In small group conversations participants may take the role as the side listener in which they are neither addressed nor speaking [4], thus the interactivity is lower. This is in line with the finding that the participant who plays a central role in small group discussion notices delay stronger than less involved participants [55]. Studies investigating the audio and video quality in video-conferencing assessed the QoE for different encoding qualities [39], packet loss [12, 27, 50], resolution [10] and frame-rate [27]. In the area of multi-party video-conferencing one study investigated different bitrates and packet losses [52] and found that encoding each stream with 4mbps did not yield a significant improvement to encoding each stream with 1mbps. Further the impact of different layouts [24] and devices [66] were investigated.

All these studies used a symmetric video quality setup for the different participants, i.e. each participant was encoded in the same way. Our study focuses on the aspect when different video qualities are present in the same session.

Perception of quality

Our study investigates how different qualities shape together a combined quality perception. This is similar to two aspects which have been researched in the area of QoE: how variations in quality over time are combined to a final judgment [21] and how audio and video quality are combined to an overall perceived audiovisual quality [5]. Findings in the area of time varying quality found the 'recency' effect [2, 21]: when participants are asked to rate the quality of a video with segments in different quality the last presented quality had a stronger than average impact. Further a 'duration neglect' effect was found: the segment with the worst quality had an over proportional impact independently of its duration [21, 26]. Such temporal aspects have further been integrated in QoE prediction models [7].

Another form of determining the combined perceived quality from different individual qualities is researched in the context of audiovisual quality. Here it is investigated how the audio and video channel impact the overall experience of the user [5–9, 27, 56]. One of the main findings from this research was that there is a clear interaction effect between the audio and visual channel: if one of the two is kept constant and the other stream varied in quality the perceived quality of the unmodified stream will also be rated worse [5].

In this work we are dealing with the combined quality perception of the streams from multiple participants in the same session, as described in the conceptual model for the quality formation process for multi-party conferencing [64]. This shares the characteristic with audiovisual quality, that the different qualities are presented at the same time, but in contrast the streams of participants contain different content. Time-varying quality deals with unequal quality on different contents, but in contrast to our work at different points in time.

Assessment of video quality in video-conferencing

The majority of testing methodologies were developed for two-party scenarios, however in recent years substantial work on multi-party scenarios has been conducted, resulting amongst other in an ITU standard regarding testing methodologies for tele-meetings [38]. The methodologies employed for assessing perceived video quality for telecommunication can be classified in two methodologies: passive and interactive. Passive tests are conducted by letting users rate the quality of video clips using video conferencing related content, for example [10, 39, 43]. In contrast, interactive tests use a real video-conferencing setup in which participants interact to varying degree freely with each other, for example [10, 25, 57, 58]. Similar in the audio domain listen-only tests and conversation tests are employed [42, p.50 ff.]. Although test situations are always to some degree artificial, 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 consistence results as they are better repeatable than interactive tests and need less resources but do not provide participants with a realistic experience of the conversation. In turn passive tests are used to concentrate on the examination of one specific aspect or as initial investigation for previously unexplored aspects. Usually passive and interactive tests correlate with each other, although they can exhibit systematic differences between 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 [42, p.129–133]. This is most likely due to that participants concentrated in the listening only situation more on the quality than the content of the speech and vice versa in the conversation more on the content. However, often passive tests are used to initially investigate effects which were later confirmed with interactive tests, for example audiovisual quality integration [9], quality perception of a tonal language [15] or improvement of speaker recognition due to spatial audio [44]. As our study aimed to find effects on the perceived video quality and is an initial investigation , the passive evaluation methodology was suitable.

In recent years crowdsourcing has become an often employed methodology for conducting QoE evaluations [28]. In such setups, the test is conducted by participants at home, so called *crowdworkers*, over a webplatform. 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 [3, 13, 29]. Several studies have been conducted that researched the methodology, like the influence of video clip length [22], a training phase [23] and fraud detection [33]. These studies have been gathered in recommendation guidelines for QoE assessment in crowdsourcing [28, 30], after which we designed this study. To our knowledge,

[1] www.microworkers.com.

[2] www.mturk.com

crowdsourcing has not been used in the context of videoconferencing. Further there is no knowledge on assessing the quality of individual parts and the overall quality of a media presentation. Thus, our study is split into three parts, campaigns, in which we test three different rating methodologies: gathering only individual ratings, gathering only overall ratings and gathering both ratings at the same time.

Methodology

In this section we lay out the details of the conducted study and the examination of the gathered data. As we are the first 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. A detailed explanation of the employed design and procedure is provided in the next section. We follow by describing how we prepared the material (i.e. videoclips) for the crowdsourcing this study. Then we are detailing the demographics of the crowdworkers and how we filtered our data by reliability criteria before the analysis. Finally we detail the methodology for the following statistical analysis.

Experiment design

In previously conducted interactive studies regarding video quality in video-conferencing we had symmetric quality for all participants (i.e. the streams of participant 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 that the recorded material may be used for crowdsourcing studies. We selected two 40 seconds segments from two sessions which were concerned with different tasks. The length of 40 seconds was chosen as it allows to keep the study short but still provides enough context for crowdworkers to be engaged with the content [22]. In one clip participants discuss the possibility of using a ’radio device’ for rescue when lost at sea, in the following referred to as the *conversation* task. The task was based on a teambuilding exercise from [11]. In 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 [45]. In this study, we presented each crowdworker exactly once with each of the two clips. Each clip showed 4 streams in a 2 × 2 layout (see Fig. 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, 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 the overall rating (see Fig. 3). In the campaign *overall* we asked only for the overall video quality rating (see Fig. 3, upper part) and accordingly in the campaign *individual* only for the individual ratings (see Fig. 3, lower part). A participant could only participate in one campaign and only once.

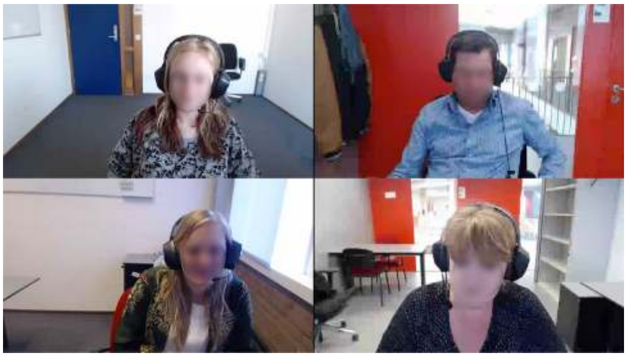
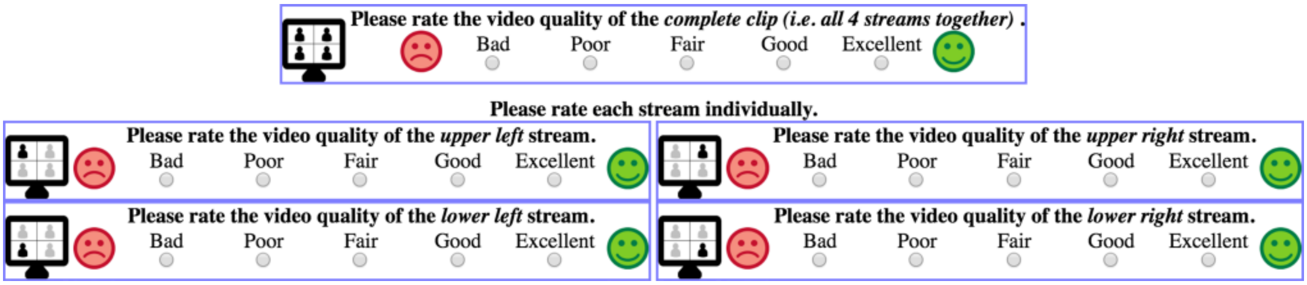


Fig. 2 Screenshot of the *conversation* video clip with the encodings *hlhl* from upper left (ul) to lower right (lr). Faces blurred for publication

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 2 × 2 layout showing one participant each, as it is a common presentation mode in many commercial videoconferencing applications. Thus all participants were presented in the same size. We chose to encode the videostream of each clip in two different video qualities, 256 kbps, to which we will refer to in the context of our study as *low* quality or in short *l* and with 1024 kbps 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 are 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



Please rate the video quality of the *complete clip (i.e. all 4 streams together)* .

Bad ○ Poor ○ Fair ○ Good ○ Excellent ○

Please rate each stream individually.

Please rate the video quality of the *upper left* stream.

Bad ○ Poor ○ Fair ○ Good ○ Excellent ○

Please rate the video quality of the *upper right* stream.

Bad ○ Poor ○ Fair ○ Good ○ Excellent ○

Please rate the video quality of the *lower left* stream.

Bad ○ Poor ○ Fair ○ Good ○ Excellent ○

Please rate the video quality of the *lower right* stream.

Bad ○ Poor ○ Fair ○ Good ○ Excellent ○

Fig. 3 Screenshot of the rating scale for the campaign *both*. In the campaigns *overall* only the top question was shown, in the campaign *individual* on the four bottom questions were shown

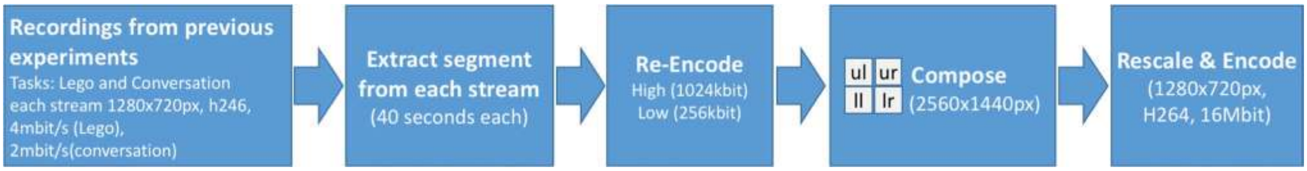


Fig. 4 Workflow for preparing the video material

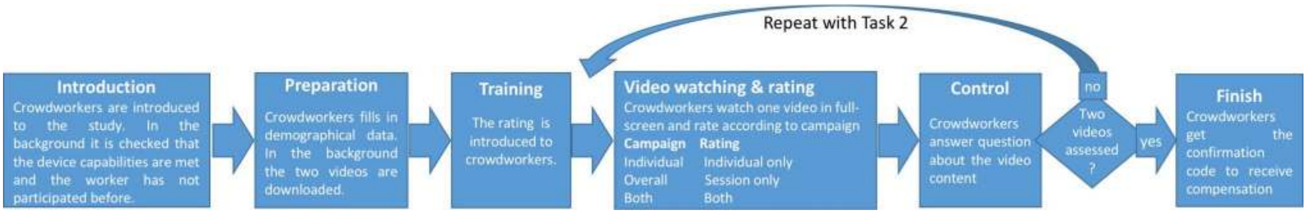

Repeat with Task 2
Introduction
Crowdworkers are introduced to the study. In the background it is checked that the device capabilities are met and the worker has not participated before.
Preparation
Crowdworkers fills in demographical data. In the background the two videos are downloaded.
Training
The rating is introduced to crowdworkers.
Video watching & rating
Crowdworkers watch one video in full-screen and rate according to campaign
Campaign Rating
Individual Individual only
Overall Session only
Both Both
Control
Crowdworkers answer question about the video content
no
Two videos assessed ?
yes
Finish
Crowdworkers get the confirmation code to receive compensation

Fig. 5 The different steps of the crowdsourcing study

can be different from the others (one stream *high* quality and the others *low* quality or the other way 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 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 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. 4 we detail the treatment of the video clips. The original streams had a resolution of 1280 × 720 pixels encoded in H.264 (the *conversation* streams with 2Mbits the *lego* streams with 4 Mbits). The audio was recorded in both cases with the mp3 codec with ca 20 kbps per second.

We first re-encoded the individual streams with ffmpeg.[3] The four individual stream were then composed to one clip with GStreamer[4] and the final result scaled to 1280 × 720 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. 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.

[3] www.ffmpeg.org.

[4] www.gstreamer.org.

Table 1 Questions and answer options regarding the Internet connection and usage of video services

Question	Options
What is the speed of your Internet connection?	"Less than 1 Mbit" (slow), "less than 4 Mbit" (medium), "less than 12 Mbit" (fast), "more than 12 Mbit" (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"

In some cases we will later use a shorter label which We added in parenthesis after the options participants were shown

Procedure

An overview of the procedure is shown in Fig. 5. The study consisted of an introduction, gathering of demographical data, a training phase, the two assessments (each with content control question) and a final page with the information for the crowdworker to get the compensation. On all pages a comment box was displayed that allowed participants to give feedback. When a potential participant accessed the introduction page it was first checked if her or his device met the following requirements: not a mobile device (i.e. tablet or smartphone), a minimum resolution of 1280 × 720 pixels and a browser able to play html5 videos. Further it was checked that the crowdworker had not participated previously in this or one of the other campaigns (checked via the provided crowdworker id). 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 interaction with the page would be saved. In the next step, the participants were asked about demographical 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 1 for details). During this step also 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 the participants were shown a screenshot (Fig. 2) of the video and a screenshot of the rating scales (Fig. 3) with an explanation assuring them that we are gathering their opinion and there are no right or wrong answers. Further 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 and end the fullscreen mode once the videoclip was finished. If the crowdworker ended the fullscreen before the clip had ended, an overlay would appear that the video clips needs to be finished in fullscreen mode for completing the study. After the clip had finished playing, the rating scales (see Fig. 3) would appear below the video. In the final page the crowdworker was thanked for participating in the study and displaying the confirmation code which was needed for the crowdsourcing platform. The compensation for completing the assessment was 0.35 US cents. 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*.

Participants and reliability filtering

The crowdsourcing experiment was conducted over the crowdsourcing platform *Microworkers*.[5] 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 min to finish the study. We omitted 5 participants which took more than two times the standard deviation longer than the mean duration (sd = 6.42 min → 19.45 min) as they likely got distracted with something else during the assessment. We further excluded one participant who reported to be using a smartphone. We further employed the reliability filtering suggested by Ribeiro et al. [48] 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. [30] was applicable as we had only two ratings per subject.

[5] www.microworkers.com.

Table 2 Factors used in the statistical analysis with used symbol, levels and description

Factor	Symbol	Levels	Description
Independent factors			
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 (256 kbps = low = l, 1024 kbps = high = h)	Encoding bitrate of a video stream. Audio stream and muxing not included.
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 factors			
Overall rating	RO	5 (bad–excellent)	Rating of the video quality of an entire clip (ITU P.911 [60] 5-point rating scale)
Individual rating	RI	5 (bad–excellent)	Rating of an individual stream (ITU P.911 [60] 5-point rating scale)
Rating	R	5 (bad–excellent)	Individual and overall ratings

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%).

Quantitative analysis

In the analysis we make use of linear regression models [16, p. 161 ff., p.353 ff.] in the form of

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

We model one dependent variable, the vector Y, through the combination of the independent variables, the vectors $X_1 \ldots X_n$ and a random error term ϵ. The coefficients $\beta_0 \ldots \beta_n$ are determined in such a way that the sum of squares of the error term is minimized [16, 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

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

Table 2 contains an overview of the factors used in the analysis. To assess whether a factor is statistical significant we are using a Likelihood Ratio-test (LRT) [16, p. 163] with the factor in question against a model without the factor. A factor was considered statistical 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 [18]. Given confidence interval are bias corrected and accelerated (BCa) confidence intervals which are more accurate than other estimation methods for skewed data [17]. For the bootstrap we are drawing random samples with replacement from the corresponding original data. The LRTs are computed on these bootstrapped datasets and repeated a 1000 times. The resulting bootstrapped statistics are considered significant at $p < 0.05$ when 95% of the computed LRTs are significant at a $p < 0.05$ level. For the performed posthoc tests we are bootstrapping a Tukey HSD (with multivariater correction) with 8000 repetitions.

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 with different video qualities. Thus we ran statistical tests between the ratings users gave and the different combinations of encoding bitrates. Specifically we check 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

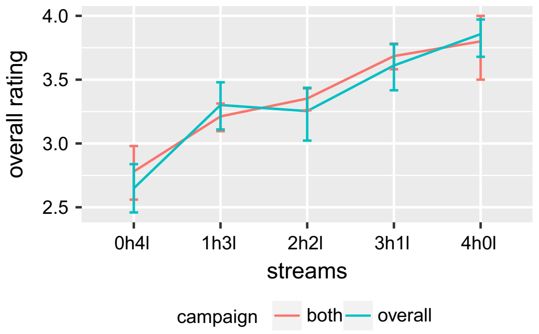


Fig. 6 Line plot comparing the overall quality ratings from the campaigns 'both' and 'overall'

- Comparison of overall and individual ratings
- Analyses of covariates (demographical data)

Campaigns

To gain insight how the quality perception of individual streams and complete session is related we gathered both kinds of ratings. However, we were concern whether assessing these ratings at the same time or separately influences our results. We are not aware that there is any knowledge on this topic. To gain insight into this, 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 is rather small (see Fig. 6) and also for the individual ratings most streams received similar ratings in both campaigns (see Fig. 13, "Appendix"). Bootstrapped LRTs confirmed that there was no significant difference for the overall ratings and only two of the 16 individual streams were received significantly different ratings (see Table 3). We concluded that the different assessment methodologies do not have a significant impact on the ratings. Thus, in the following analyses we will handle the data from the different campaigns together.

$$RO = \beta_0 + \beta_1 S + \beta_2 T + \epsilon \tag{m1}$$

$$RO = \beta_0 + \beta_1 S + \beta_2 T + \beta_3 C + \epsilon \tag{m2}$$

$$RI_{t,q} = \beta_0 + \beta_1 S_{t,q} + \epsilon \tag{m3}$$

$$RI_{t,q} = \beta_0 + \beta_1 S_{t,q} + \beta_2 C_{t,q} + \epsilon \tag{m4}$$

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

Perceived overall quality

We wanted to quantify the impact, that changing the individual streams encoding quality has on the perceived quality of the complete screen (overall quality). As expected a higher combined encoding quality lead also to a higher overall perceived quality (see Fig. 7). We confirmed with bootstrapped LRTs that *streams* and *task* are both significant factors without an interaction effect (see Table 4). We continued with a bootstrapped post-hoc test and marked the groups of different conditions in Fig. 7 with dotted circles.

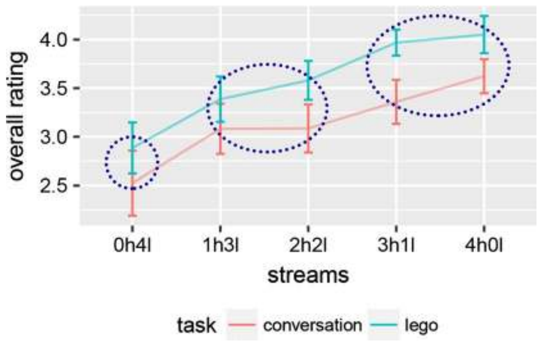


Fig. 7 Line plot of mean overall ratings by streams and task with 95% CIs as errorbars. The dotted circles indicate statistical significant contrast groups determined by a bootstrapped post-hoc test. Conditions grouped by a circle are not significantly different to each other, but to the other groups

Table 4 Bootstrapped Likelihood Ratio Tests for the response variable *overall quality rating*

Model 1	Model 2	Factor under test	*p* value
m5	m6	Streams	< **0.05**
m6	m1	Task	< **0.05**
m1	m7	Interaction between streams and task	> 0.05

The factor under test was considered to have a statistical significant influence at a level of $p < 0.05$ (marked in bold)

Table 3 *p* values of bootstrapped Likelihood Ratio Tests for the campaigns

Model 1	Model 2	Factor under test	*p* value
m1	m2	Campaigns	> 0.05
m3	m4	Campaigns	> 0.05 except (t = *conversation*, q = *low quality*, SI = *lr*) < 0.05 and (t = *Lego*, q = *high quality*, SI = *ll*) < 0.05

Table 5 Bootstrapped Likelihood Ratio Tests for the response variable *individual stream quality rating*

Model 1	Model 2	Factor under test	p value
m8	m9	Stream encoding	$<$ **0.05**
m9	m10	Task	$<$ **0.05**
m10	m11	Interaction between stream encoding and task	$>$ 0.05
m10	m12	Interaction number of streams and stream encoding	$<$ **0.05**

The factor under test was considered to have a statistical significant influence at a level of $p < 0.05$ (marked in bold)

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

$$RO = \beta + \epsilon \tag{m5}$$

$$RO = \beta_0 + \beta_1 S + \epsilon \tag{m6}$$

$$RO = \beta_0 + \beta_1 S + \beta_2 T + \beta_3 S * T + \epsilon \tag{m7}$$

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).

The quality of *high* and *low* encoded streams was perceived clearly different (see Table 5) with an average difference of circa 1.5 points between them (see Fig. 8a). Like with the overall ratings there is a statistical and clearly visible and difference between the tasks but not interaction between stream encoding and task (see Table 5 and Fig. 8b 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 are co-present with other high quality streams and vice versa the high quality streams got rated better the more low quality streams were co-present (see Fig. 9). 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 5). We continued with a bootstrapped post-hoc test to assess which conditions are significantly different from each other and marked them with dotted circles in Fig. 9. For the *high* quality streams there were three groups, while for the *low* quality streams there were only two, indicating

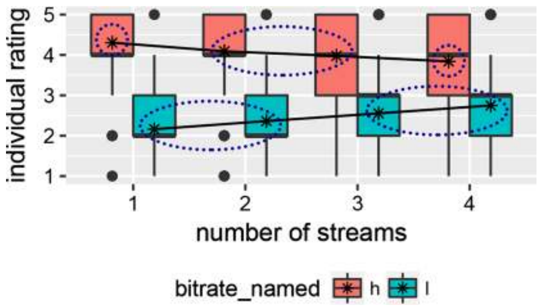
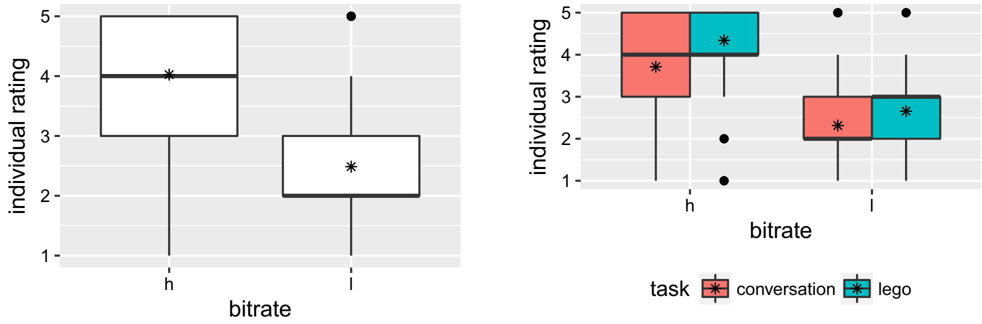


Fig. 9 Box plots with additional means and a line between them of overall rating per streams. The dotted circles indicate statistical significant contrast groups determined by a bootstrapped post-hoc test. Conditions grouped by a circle are not significantly different to each other, but to the other groups. 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



(a) **(b)**

Fig. 8 Individual stream ratings for high (h = 1024kbps) and low (l = 256kbps) streams. **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

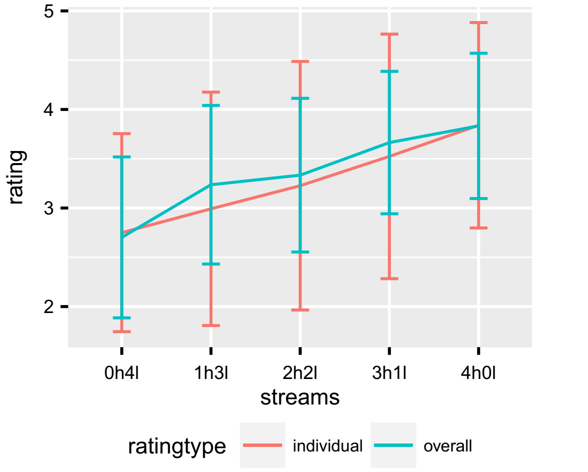


Fig. 10 Mean of individual ratings and overall ratings with standard deviation as errorbars

that the effect is slightly weaker for the *low* quality streams (see Fig. 9). That as well *low* and *high* quality ratings were decreasing seems to indicate that as more low quality streams around the better a high quality stream looks and vice versa the more high quality streams around the worse a low quality streams looks.

$$RI = \beta + \epsilon \tag{m8}$$

$$RI = \beta_0 + \beta_1 BI + \epsilon \tag{m9}$$

$$RI = \beta_0 + \beta_1 BI + \beta_2 T + \epsilon \tag{m10}$$

$$RI = \beta_0 + \beta_1 BI + \beta_2 T + \beta_3 BI * T + \epsilon \tag{m11}$$

$$RI = \beta_0 + \beta_1 BI + \beta_2 T + \beta_3 NS_h + \beta_4 NS_l + \epsilon \tag{m12}$$

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 were higher than the individual ratings (see Fig. 10). 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 (m13, m14) < 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. 10). The reason is found in the individual differences in quality perception of the individual streams.

$$R = \beta_0 + \beta_1 S + \beta_2 T + \epsilon \tag{m13}$$

$$R = \beta_0 + \beta_1 S + \beta_2 T + \beta_3 RT + \epsilon \tag{m14}$$

The contrast effect, described in the previous "Perceived quality of individual streams" section, was present for most individual streams (see Fig. 11). However for some 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. 11). We can further observe that the streams of the *conversation* task were not only lower rated in average but the 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

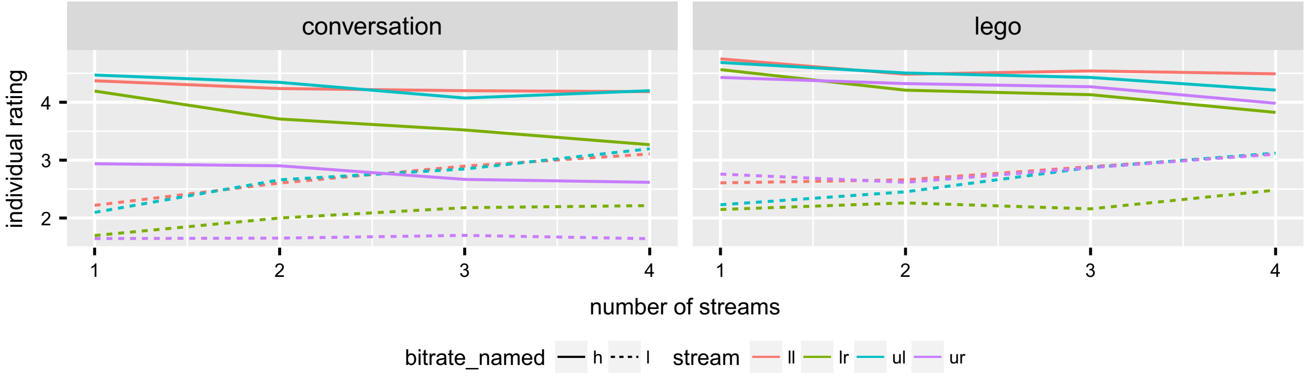


Fig. 11 Rating of each stream by encoding bitrate and number of streams in the same quality. It is to note, that each stream is encoded by position (see Table 2) which corresponds to always the same participant per task, thus the stream *ul* for the *lego* task does not show the same participant as the *ul* stream of the *conversation* task

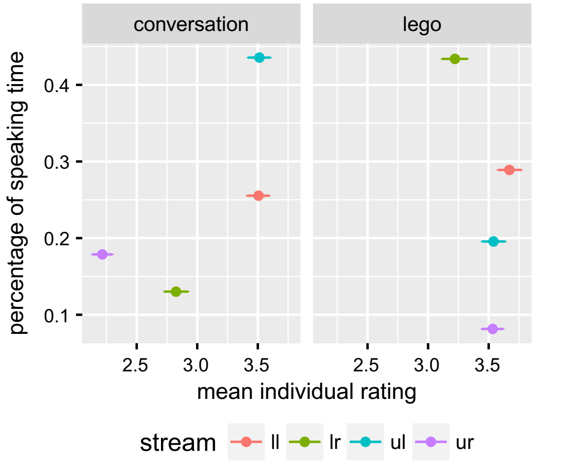


Fig. 12 Amount of speaking time in percent against average rating (individual ratings) with 95% CIs

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 as well individual as also overall quality perception.

To gain insight 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 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 (m12 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 improvement was 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. 12). 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 [36, 61]). We added these models to the model for the individual qualities (m12). A bootstrapped LRT showed that Spatial Activity improved the fit of the model for the conversation task, but not with the Lego task. Temporal Activity did not improve either of the two tasks.

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 (m1 and m12 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 (m12), however when checking for influential data points, this effect was due to only two participant over 65, thus we opted to draw currently no conclusions 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. 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 [46]. However, when we checked for influential data points, the effect was depending on 12 participants with display sizes of 27 inches or larger. Due to the sparsity of this data we are not drawing further conclusions about display size at this point and further 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 quality perception of a participants are his or her previous experience. 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 including the frequency of video-conferencing or Internet video usage improved the fit of the models m1 or m12. However, including the type of Internet connection or the speed participants reported, both improved the fit of the model m1. Participants who reported a better connection gave worse ratings (see in Figs. 14, 15, "Appendix"). This supports the theory that a better Internet connection leads to a higher baseline on expectations of video quality. However, it is not easy to get accurate information from participants (32% reported *NA* or *other* in at least one of the two questions). Also the average of *slow* Internet connections falls out of this pattern, however, the variance is here also the highest.

The worst rating gave 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.

Discussion

The main findings from the performed analysis were:

- Q1) The change in overall perceived video quality from only low quality streams (*0h4l*) 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*)).
- Q2) The individual ratings for high and low quality were affected by the co-presented streams: high quality streams were perceived better the more low quality streams were around and vice versa low quality streams worse with more high quality streams present.
- Q3) The overall quality ratings were constantly higher than the mean of the individual quality ratings. The mean of the individual quality ratings followed a very strict linear pattern while the overall ratings were more curved.

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* → *1h3l* and *4h0l* → *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 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 as strong as that it would minimize the benefits from having one stream in better quality. This means, 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 even higher than three high quality streams (*3h1l*). In the *conversation* task a large portion of the variance was explainable 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.

Further 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 the visual channel plays a bigger role in the interaction. 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 is unclear.

Conclusion

In this paper we present our exploratory research about how QoE is affected by different video qualities in the same multi-party video-conferencing 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 presented and, vice versa, high encoded streams are perceived better the more low encoded streams are present. Further we could show that the activity of the session, the roles of participants and the individual differences between the participants, played 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 into account , including the different encodings. Beyond the influence factors analyzed in this work, individual factors, most likely related to the activity and role within the session, are strong influencing factors, and will need to be taken into account for accurate estimation models. Even though our study employed a static layout, which provided each video-session participant with an equal amount of space, the differences in ratings between them were strong.

Future work

Our findings show that in multi-party video-conferencing a contrasting effect on the perception of video quality exists. As initial investigation whether such an effect exists, we fixed several factors that differ in real world video conferencing setups, but further factors need to be investigated to fully understand the impact of mixed video qualities on QoE.

The main steps that need to be taken are to conduct interactive tests and explore further factors and setups. The interactive tests in real video-conferencing sessions will show how the contrasting effect is perceived when users are participating in a video conversation. During passive evaluations, as used in this study, it is possible that the participants (of the passive study) pay little or no attention to the actual content of the material and, thus, detect quality differences that otherwise go unnoticed [42, p.129–133]. A similar effect was shown in an interactive study in which participants with a higher engagement in the task reported also a higher QoE [53]. A further challenge in conducting an interactive study is that usually variances in the ratings are 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) into the study. However to include such factors, studies need to have a high number of samples, as these factors are usually covariates of a study, in such that their characteristics in terms of variance and range are hardly known and controllable.

Further, such an interactive study should give insights into how the contrasting effect is moderated by interaction. Our study showed that in one of the two recorded sessions the speaking time was a good moderating variable for the perceived quality. We assume that if users participate directly in the video conference themselves the role of this moderating factor increases as they are more engaged in the conversation. Another factor which is substantially different in interactive studies 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, the length could have an influence on the results.

Further important factors that need to be examined are the number of participants and the layout of the video streams. This would be for one keeping the layout constant, like in this study, but varying the number of participants . Our study indicates 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 contrasting effect is possible. However, this would reduce thweight of this stream for the overall quality. This would mean that the individual perceived quality of a single stream would be stronger affected but might not show a stronger effect for the overall perceived quality of the session.

Further , it is possible that a stream will get more attention depending on its position in the layout. This wold be the upper left position, at least in the western society where the view falls first according to the reading direction. Our study showed large differences between the streams, however, these can also be due to individual appreciation of the shown video participants as in our case each position was always linked to the same video participant.

Using a dynamic layout (e.g. the 'speaker-big, thumbnails for others' like for example Google Hangout employs) on the other hand provide 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. In case that a significant difference in the quality perception between these two methods exists, this could guide layout decisions for video-conferencing systems.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix

See Figs.13, 14 and 15.

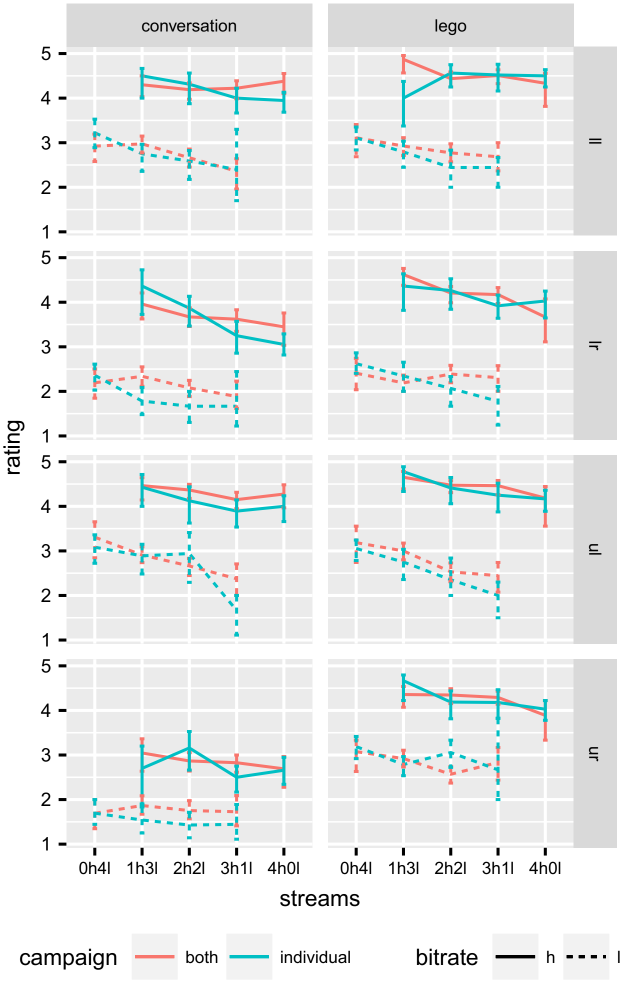


Fig. 13 Line plot comparing the individual quality ratings from the campaigns ‘both’ and ‘individual’

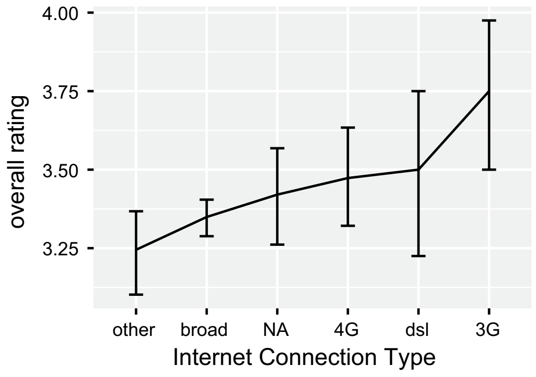


Fig. 14 Mean of ratings per Internet connection type with 95% confidence intervals

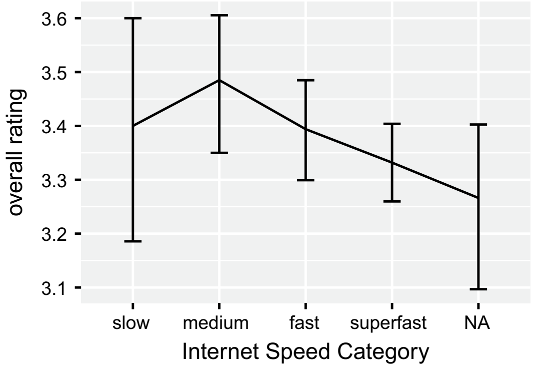


Fig. 15 Mean of ratings per Internet connection speed with 95% confidence intervals

References

1. 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, March 2013. Lausanne, Switzerland, Version 1.2
2. Aldridge R, Davidoff J, Ghanbari M, Hands D, Pearson D (1995) Recency effect in the subjective assessment of digitally-coded television pictures. In: Fifth international conference on image processing and its applications, pp 336–339
3. Amour L, Sami S, Hoceini S, Mellouk A (2015) 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, pp 139–140, New York, NY, USA. ACM
4. Beebe SA, Masterson JT (1997) Communicating in small groups: principles and practices. Longman
5. Beerends JG, Caluwe D, Frank E (1999) The influence of video quality on perceived audio quality and vice versa. J Audio Eng Soc 47(5):355–362
6. Belmudez B (2015) Audiovisual quality assessment and prediction for videotelephony. T-Labs series in telecommunication services. Springer, Cham
7. Belmudez B, Lewcio B, Möller S (2013) Call quality prediction for audiovisual time-varying impairments using simulated conversational structures. Acta Acust United Acust 99(5):792–805
8. Belmudez B, Möller S (2013) Audiovisual quality integration for interactive communications. EURASIP J Audio Speech Music Process 2013(1):1–23
9. Belmudez B, Moeller S, Lewcio B, Raake A, Mehmood A (2009) Audio and video channel impact on perceived audio-visual quality in different interactive contexts. In: Proceedings of the MMSP, pp 1–5. IEEE
10. Berndtsson G, Folkesson M, Kulyk V (2012) Subjective quality assessment of video conferences and telemeetings. In: Proceedings of the 19th PV, pp 25–30. Cited by 0000
11. Biech E (2007) The Pfeiffer book of successful team-building tools: best of the annuals. Pfeiffer, Santa Ana 00000
12. Brauer F, Ehsan MS, Kubin G (2008) Subjective evaluation of conversational multimedia quality in IP networks. In: Proceedings of the 10th MMSP, pp 872–876

13. Chen K-T, Chang C-J, Chen-Chi W, Chang Y-C, Lei C-L (2010) Quadrant of euphoria: a crowdsourcing platform for QoE assessment. IEEE Netw 24(2):28–35 00063
14. Chen S, Chu C-Y, Yeh S-L, Chu H-H, Huang P (2014) Modeling the QoE of rate changes in Skype/SILK VoIP calls. IEEE/ACM Trans Netw 22(6):1781–1793
15. Daengsi T, Yochanang K, Wuttidittachotti P (2013) A study of perceptual VoIP quality evaluation with thai users and codec selection using voice quality-Bandwidth tradeoff analysis. In: 2013 international conference on ICT convergence (ICTC), pp 691–696. IEEE
16. Davison AC (2008) Statistical models, 1st edn. Cambridge University Press, Cambridge
17. DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. Stat Sci 11(3):189–212
18. Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC Press, Boca Raton
19. Egger S, Schatz R, Schoenenberg K, Raake A, Kubin G (2012) Same but different?—Using speech signal features for comparing conversational VoIP quality studies. In: 2012 IEEE international conference on communications (ICC), pp 1320–1324
20. Egger S, Reichl P (2009) A nod says more than thousand uhmm's: towards a framework for measuring audio-visual interactivity. In: Proceedings of the COST298 conference The Good, the Bad and the Challenging, Copenhagen, Denmark
21. Fredrickson BL (2000) Extracting meaning from past affective experiences: the importance of peaks, ends, and specific emotions. Cognit Emot 14(4):577–606
22. Fröhlich, P, Egger, S, Schatz, R, Mühlegger M, Masuch K, Gardlo B (2012) 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, pp 242–247
23. Gardlo B, Egger S, Hossfeld T (2015) Do scale-design and training matter for video QoE assessments through crowdsourcing?. ACM, Brisbane ACM
24. Gunkel SNB, Schmitt M, Cesar P (2015) A QoE study of different stream and layout configurations in video conferencing under limited network conditions. ResearchGate
25. Hammer F, Reichl P, Raake A (2005) The well-tempered conversation: interactivity, delay and perceptual VoIP quality. In: 2005 IEEE international conference on communications, 2005, ICC 2005, vol 1, pp 244–249. Cited by 0029
26. Hands DS, Avons SE (2001) Recency and duration neglect in subjective assessment of television picture quality. Appl Cognit Psychol 15(6):639–657
27. Hayashi T, Yamagishi K, Tominaga T, Takahashi A (2007). Multimedia quality integration function for videophone services. In: Proceedings of the GLOBECOM, pp 2735–2739
28. Hoßfeld T, Keimel C, Hirth M, Gardlo B, Habigt J, Diepold K, Tran-Gia P (2014) Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. IEEE Trans Multimed 16(2):541–558
29. Hoßfeld T, Seufert M, Hirth M, Zinner T, Tran-Gia P, Schatz R (2011) Quantification of YouTube QoE via crowdsourcing. In: 2011 IEEE international symposium on multimedia (ISM), pp. 494–499. 00066
30. Hoßfeld T, Seufert M, Sieber C, Zinner T (2014) 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), pp 111–116
31. Hoßfeld T, Biedermann S, Schatz R, Platzer A, Egger S, Fiedler M (2011) The Memory effect and its implications on web QoE modeling. In: Proceedings of the 23rd ITC, ITC '11, pp 103–110, San Francisco, California. International Teletraffic Congress
32. Hoßfeld T, Egger S, Schatz R, Fiedler M, Masuch K, Lorentzen C (2012) Initial delay vs. interruptions: between the devil and the deep blue sea. In: 2012 fourth international workshop on quality of multimedia experience (QoMEX), pp 1–6. Cited by 0013
33. Hoßfeld T, Keimel C (2014) Crowdsourcing in QoE evaluation. In: Möller S, Raake A (eds) Quality of experience, T-Labs series in telecommunication services. Springer, Berlin, pp 315–327. https://doi.org/10.1007/978-3-319-02681-7_21
34. Hoßfeld T, Skorin-Kapov L, Heegaard PE, Varela M (2017) Definition of QoE fairness in shared systems. IEEE Commun Lett 21(1):184–187
35. Hoßfeld T, Strohmeier D, Raake A, Schatz R (2013) 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, pp 37–42, New York, NY, USA. ACM
36. ITU-T. ITU-T Recommendation P.910 - Subjective video quality assessment methods for multimedia applications. 1995. 00000
37. ITU-T. P.1305 : Effect of delays on the telemeeting quality, 2016
38. P.1301 ITU-T RECOMMENDATION. ITU-P.1301 - Subjective quality evaluation of audio and audiovisual multiparty telemeetings, February 2013
39. Jones C, Atkinson DJ (1998) Development of opinion-based audiovisual quality models for desktop video-teleconferencing. In: Proceedings of the 6th IWQoS, pp 196–203. IEEE
40. Kim SJ, Chae CB, Lee JS (2012) Quality perception of coding artifacts and packet loss in networked video communications. In: 2012 IEEE Globecom workshops (GC Wkshps), pp 1357–1361
41. Lee J-S, Goldmann L, Ebrahimi T(2011) 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, pp 1284–1284, New York, NY, USA. ACM
42. Möller S (2000) Assessment and prediction of speech quality in telecommunications. Springer, Boston. https://doi.org/10.1007/978-1-4757-3117-0
43. Ndiaye M, Larabi MC, Saadane H, Le Lay G, Perrine C, Quinquis C, Gros L (2015) Subjective assessment of the perceived quality of video calling services over a real LTE/4g network. In: Proceedings of the 7th QoMEX, pp 1–6
44. Raake A, Schlegel C (2008) Auditory assessment of conversational speech quality of traditional and spatialized teleconferences. In: ITG conference on voice communication [8. ITG-Fachtagung], pp 1–4
45. ITU-T RECOMMENDATION. ITU-R P.920-Interactive test methods for audiovisual communications (2000)
46. ITU-T RECOMMENDATION. ITU-T G.1070 Opinion model for video-telephony applications, July 2012
47. Reichl P, Egger S, Möller S, Kilkki K, Fiedler M, Hossfeld T, Tsiaras C, Asrese A (2015) Towards a comprehensive framework for QOE and user behavior modelling. In: Proceedings of the 7th QoMEX, pp 1–6
48. Ribeiro Flávio, Florêncio Dinei, Zhang Cha, Seltzer Michael (2011) Crowdmos: An approach for crowdsourcing mean opinion score studies. In: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2416–2419. IEEE
49. Sacks H, Schegloff E, Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversation. Language 50(4):696–735 10741
50. Saidi I, Hang Lu, Barriac V, Deforges O (2016) 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), pp 1–6
51. Schmitt M, Gunkel S, Cesar P, Bulterman D (2014) Asymmetric delay in video-mediated group discussions. In: Proceedings of the 6th QoMEX, pp 19–24

52. Schmitt M, Redi J, Cesar P, Bulterman D (2016) 1mbps is enough: video quality and individual idiosyncrasies in multiparty HD video-conferencing. In: Proceedings of the 8th QoMEX, pp 1–6. IEEE
53. Schmitt MR, Redi J, Bulterman D, Cesar P (2017) Towards individual QoE for multi-party video conferencing. IEEE Trans Multimed PP(99):1–1
54. Schmitt M, Gunkel S, Cesar P, Bulterman D (2014) Asymmetric delay in video-mediated group discussions. In: 2014 sixth international workshop on quality of multimedia experience (QoMEX), pp 19–24. IEEE
55. Schmitt M, Gunkel S, Cesar P, Bulterman D (2014) The influence of interactivity patterns on the quality of experience in multi-party video-mediated conversations under symmetric delay conditions. In: Proceedings of the 3rd SAM, SAM '14, pp 13–16, New York, NY, USA. ACM
56. Schmitt M, Redi J, Cesar P (2016) Towards context-aware interactive quality of experience evaluation for audiovisual multiparty conferencing. In: Proceedings of the 5th PQS, Berlin, pp 64–68
57. Schmitt M, Redi J, Cesar P, Bulterman D(2016) 1Mbps is enough: video quality and individual idiosyncrasies in multiparty HD video-conferencing. In: 2016 eighth international conference on quality of multimedia experience (QoMEX), pp 1–6. IEEE
58. Schoenenberg K, Raake A, Lebreton P (September 2014) Conversational quality and visual interaction of video-telephony under synchronous and asynchronous transmission delay. In: Proceeding of the 6th QoMEX, pp 31–36
59. Schoenenberg K, Raake A, Egger S, Schatz R (2014) On interaction behaviour in telephone conversations under transmission delay. Speech Commun 63–64:1–14 00002
60. International Telecommunication Union Telecommunication Standardization Sector. *ITU-T Recommendation P.911—Subjective audiovisual quality assessment methods for multimedia applications*. International Telecommunication Union, 1998. 00000 Cited by 0000
61. Seshadrinathan K, Bovik AC (2011) Temporal hysteresis model of time varying subjective video quality. In: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1153–1156
62. Skowronek J, Herlinghaus J, Raake A (2013) Quality assessment of asymmetric multiparty telephone conferences: a systematic method from technical degradations to perceived impairments. In: INTERSPEECH, pp 2604–2608
63. Skowronek J, Raake A (2011) Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing. In: INTERSPEECH, pp 829–832
64. Skowronek J, Raake A (2015) conceptual model of multiparty conferencing and telemeeting quality. In: Proceedings of the 7th QoMEX. IEEE
65. Skowronek J, Raake A, Hoeldtke K, Geier M (2011) Speech recordings for systematic assessment of multi-party conferencing. In: Proceedings of forum acusticum, pp 111–116
66. Vucic D, Skorin-Kapov L (2015) The impact of mobile device factors on QoE for multi-party video conferencing via WebRTC. In: 2015 13th international conference on telecommunications (Contel), pp 1–8
67. Winkler S (2009) On the properties of subjective ratings in video quality experiments. In: International workshop on quality of multimedia experience, 2009. QoMEx, pp 139–144

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.