Towards individual QoE for multi-party video conferencing

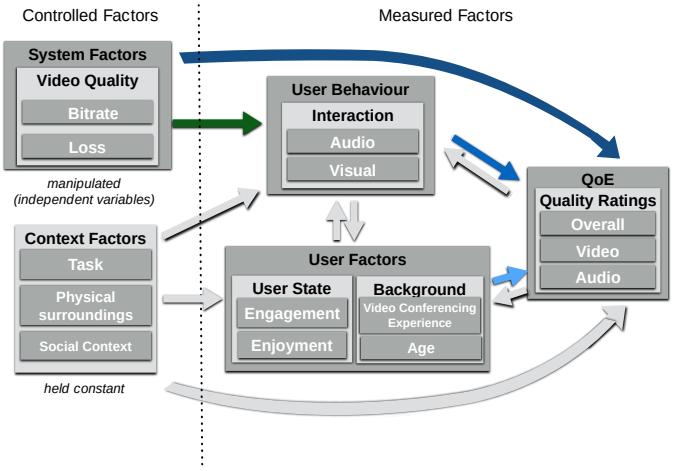
Marwin Schmitt, Judith Redi, Dick Bulterman and Pablo Cesar

***Abstract*— Video-conferencing is becoming an essential part in everyday life. The visual channel allows for interactions which were not possible over audio-only communication systems such as the telephone. However, being a de-facto over-the-top service, the quality of the delivered video-conferencing experience is subject to variations, dependent on network conditions. Video-conferencing systems adapt to network conditions by changing for example encoding bitrate of the video. For this adaptation not to hamper the benefits related to the presence of a video channel in the communication, it needs to be optimized according to a measure of the Quality of Experience (QoE) as perceived by the user. The latter is highly dependent on the ongoing interaction and individual preferences, which have hardly been investigated so far. In this paper, we focus on the impact video quality has on conversations that revolve around objects that are presented over the video channel. To this end we conducted an empirical study where groups of 4 people collaboratively build a Lego® model over a video-conferencing system. We examine the requirements for such a task by showing when the interaction, measured by visual and auditory cues, changes depending on the encoding bitrate and loss. We then explore the impact that prior experience with the technology and affective state have on QoE of participants. We use these factors to construct predictive models which double the accuracy compared to a model based on the system factors alone. We conclude with a discussion of how these factors could be applied in real world scenarios.**

***Index Terms*— Multi-Party video conferencing, Quality of Experience, Over-the-top, subjective quality, quality metrics, user study**

I. Introduction

Video-conferencing has now reached the end consumer market, and is booming. Cisco is reporting a growth in desktop video conferencing[1]; Skype reports that multi-party (group) video conferencing is especially on the rise[2]. Video-conferencing for the consumer is a de-facto over-the-top service, and thus varying quality is unavoidable. To ensure user satisfaction, video-conferencing solutions try to adapt their configuration to changes in the system conditions, e.g. by adapting the encoding quality, resolution and framerate in the event of a decrease of bandwidth availability [1]. The goal is to provide the user with the best Quality of Experience (QoE), given the system constraints, yet using the least amount of necessary resources.



annoyance of a user with a service or system [2]. It is a measure of how system factors such as bitrate, delay, or packet loss impact user experience in a given context, and it is essential to steer resource usage optimization in multimedia systems. In this paper, we are primarily interested in characterizing QoE for videoconferencing systems. We strive to design a QoE model that can automatically predict a user's QoE when using the videoconferencing system, and steer adaptation accordingly.

Whereas models for videoconferencing QoE prediction exist [3], there is room for improvement in their accuracy. Most of these models base their estimations on an analysis of system factors only: for example encoding bitrate, or packet loss. It should be clear instead, from the definition given above, that QoE depends not only on characteristics of the video-conferencing system, but also of the context of usage and the users using it [4]–[6]. This paper sets out to understand how these elements could be integrated into a model for videoconferencing QoE estimation.

Our previous research has shown that, when modeled as random factors, user individual preferences explain as much variance as system factors in video-conferencing QoE ratings [4]. Several researchers have similarly addressed the high diversity of users' opinions [7], [8]. This diversity cannot be merely ascribed to poor experimental design or small sample sizes, as even with large numbers a significant diversity within users' opinions remains [7]. Different users really do have a

M. R. Schmitt, CWI: Centrum Wiskunde & Informatica, 1098XE Amsterdam, Netherlands (e-mail: m.r.schmitt@cwi.nl).

J. A. Redi, TU Delft, Mekelweg 4, 2628 CD Delft, Netherlands (e-mail: j.a.redi@tudelft.nl).

P.S. Cesar, CWI: Centrum Wiskunde & Informatica, 1098XE Amsterdam, Netherlands (e-mail: p.s.cesar @cwi.nl).

D. Bulterman, CWI: Centrum Wiskunde & Informatica, 1098XE Amsterdam, Netherlands (e-mail: Dick.Bulterman@cwi.nl).

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

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

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 QoE of streamed video [5], [9]. In addition, dynamic factors, hereafter referred to as *user state*, which include motivation, engagement and enjoyment, can also influence and be influenced by QoE [10].

Furthermore, it has been shown that in computer mediated communication, the impact of system factors on QoE depends on the ongoing interaction between users [11]–[13]. Hence, interaction should also be accounted for, when modeling videoconferencing QoE. The '*Framework for QoE and User Behavior Modelling'* [14] conceptualizes the reciprocal relationship of QoE, user state (e.g. mood) and user behavior. Both user state and user behavior are as well an input to QoE as 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 not interested participants). A long delay usually leads to a worse QoE with a faster conversation [15]; 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 rudeness of fellow interlocutors [16]. We argue that to be able to steer QoE optimization in videoconferencing systems, it is of essence to clarify these mechanisms first.

In this paper, we set out to assess the impact that system factors, in the context of multi-party video-conferencing, on a) the user behavior, and especially interaction of participants and b) the QoE under consideration of user factors. Our approach to gain insight into this topic is depicted in Fig. 1. The figure is based on the model proposed in [14] and shows the factors examined in this paper and their relation. Our hypothesis is that the context will shape, together with the user and the system factors, 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 [14], 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 *current state* of the user.

To collect data for our investigation, we conducted an empirical study. We manipulated the system factors *encoding bitrate* and *packet loss*, which vary based on network conditions and can be dynamically adjusted during a conferencing session, to impact video quality and QoE in general. We chose not to manipulate context for this specific investigation, fixing the *physical surroundings* and task. We used an ITU-T recommend *task* [17], in which participants cooperatively build a Lego® model together over video-conferencing (see screenshot Fig. 2). We choose this task as it is representative for the common situation [18] in which users show objects to communications partners. The task is often employed in audiovisual communication test [19]–[21] and was adapted by us for a multi-party situation. 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-report 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 [21]–[23].

We specifically focus on the following three research questions:

R1. 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?
R2. How do user factors influence QoE perception?
R3. Does accounting for user and interaction factors on top of the system ones improve a model's accuracy in predicting QoE?

The hypothesis for R1 is that if the video quality is insufficient to perform the task at hand unhindered, 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. For R2 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. Finally, to address R3, we employ the elastic net [24] 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 that 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 paper is structured as following: section II contextualizes the study within other research that has been done in this area. In section III we detail the study setup and data gathering. In section IV we detail the data preparation and methods used in the analysis. In section V we present the analysis of the user behavior, in section VI the analysis of user factors and QoE and in section VII a model for predicting QoE. Section VIII discusses the results and how they would be applied in real world context. Finally section IX concludes the paper.

II. Related Work

The most common model for QoE [2] includes three categories of independent variables: user, context and system influence factors. It has been addressed that the model has some short comings when it comes to interactivity and QoE [11] and an approach to better describe the relation between user behavior, user state and QoE has been presented [14]. Further the quality formation process from [2] has been refined for the multiparty context [25].

A. Impact of system factors on videoconferencing QoE and interaction

Video quality for video conferencing is usually assessed with subjective tests (passive or active) based on which objective video quality metrics are developed. Passive tests are conducted by letting users rate the quality of video clips using video conferencing related content, for example [26], [27], [28]. As

these tests have limited ecological validity, active studies have been proposed, where participants actually interact through the system. The majority of these works has been conducted in two party scenarios [29], [20], [30], [31], [19], and employing the Lego® building blocks [17] task. It should be noted that most of these studies use relatively low resolution video (640x480px [31]) and encoding bitrates (maximum 2Mbit [31]). In today's scenario, higher resolutions (e.g. 720p) are used for videoconferencing, which require higher encoding bitrate. It is unknown whether the results obtained at lower resolutions are applicable to more recent settings.

Looking more in detail at studies on the impact of system factors on videoconferencing QoE, results exist on packet loss, encoding bitrate and delay mostly. Detailed analysis of packet loss have shown that its impact is very dependent on the type of packet lost and the motion in the video [32]. Packet loss in video-conferencing was approached by measurement of system behavior (e.g. [33]) , or simulations (e.g. [34]) but there is only one study which investigated the effect in video-conferencing with an interactive subjective test [20] with a relatively dated setup (CIF, 15fps, theora codec).

The majority of studies investigating QoE in multi-party video-conferencing has looked at delay, as it inherently interferes with turn-taking, the process describing "who speaks when" in a conversation [22]. This may in turn hinder communication, thereby impacting QoE. The effect of delay in multi-party situations was found to be more relaxed than in two party settings [13], [27], further asymmetric settings, like a single participant with high delay [35] or audio/video de-synchronicity [21], have been researched. Faster paced conversations are more susceptible to delay (measured over e.g. the speaker alternation rate [15]) and unintended interruptions disturb the experience severely [36]. In collocated settings, it has been shown that turn taking is also dependent on a number of non-verbal cues like gaze [37] and nodding [38]. The usage of visual cues in video-conferencing is relatively unexplored, one study [30] used manually annotated cues, but did not find significant differences between the two tested resolutions (640x480px vs 320x240px). Another study employed an automatic method to calculate motion of the video and found an impact based on different delay levels [21].

Recently, the impact of encoding bitrate and packet loss on multiparty video-conferencing has also been studied [4]. The study revealed that participants had an 'okay' QoE with low encoding quality (256kbps) and good to excellent experience with medium and high encoding quality (1Mbps and 4Mbps respectively) without significant differences between these two levels. Packet loss seemed to have a marginal effect instead. Interestingly, the results also indicated that (1) although only video quality was manipulated, perceived audio quality was judged as lower in worse video quality conditions and (2) this effect (as well as the overall QoE perception) was strongly varying across users [39]. In fact, a large amount of the variance in the data could be explained by factors other than the system ones [4].

B. *Effect of context and user factors on QoE*

Diversity in QoE perception due to user and context factors has been addressed in several works [7], [8]. It has been shown that user factors can explain more of the variance in user ratings than the system factors [4], [9]. In the context of video watching experiences, social context and demographic factors [5], and personality and culture traits [6] have an impact on QoE. With respect to context, studies have looked at physical surroundings (e.g. in public with a mobile vs at home on a computer) [40] or economic factors [41].

Previous experiences are known to influence the perception of future experiences [2]. This effect has been studied for short time frames for Web QoE [42][43] in which it has been shown that after experiencing bad quality, participants reported a worse QoE even after the quality was back to normal. In relation to this, age has been shown to play a role in QoE, with elderly people reporting more problems in the usage of mobile phones, more skepticism towards new technology and a later adoption rate [44] (albeit differences in usage would often disappeared after elderly got more acquainted with the devices [45]).

The interplay between user state and QoE has recently also become of more interest [14]. 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 [46]. Similarly, user engagement has been found to play a role in computer mediated human to human interaction [47]. In the context of video watching it has been found that users of an error free connection reported higher engagement than users with error [10].

C. *Prediction of videoconferencing QoE*

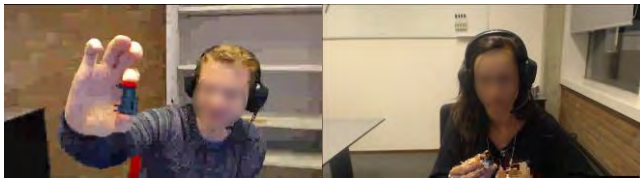
Most work for prediction of QoE in real time communication with automated methods has focused on audio only connections (e.g. the ITU e-model [48]). The ITU has a recommendation for predicting the perceived quality of an audiovisual connection [49], but it requires a large amount of system specific parameters that need to be obtained in user studies for each use case. The method has recently been extended to integrate the encoding resolution and video size automatically [3]. Including user factor besides system factors has been used to improve the accuracy of models predicting QoE for video watching scenarios [6], [50]. To our knowledge there are no models for predicting the QoE in multi-party videoconferencing based on video quality, nor any that consider user factors and interaction for an increased accuracy.

III. Study Design

Our investigation starts from a user study aimed at quantifying the impact of system and user factors on interaction and, in turn, QoE. We designed the study to resemble a current multi-party desktop video-conferencing at home, and especially in a scenario where video usage would be core to support the communication. In the following we detail the setup of the experiment by first explaining how we designed the visual 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).

A. *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 [5]. We



(a) (b)

Fig. 2 Screenshot from (a) low quality (256kbit) video stream of participant showing object into the camera and (b) high quality video (4Mbit) with distortion

adapted the ITU-T P.920 building blocks task [17], [20], [21] to a multiparty situation (and in particular, to parties of 4). Each participant was supplied with an unassembled Lego® locomotive[3], and only part of the instructions to build it. Other participants had complementary parts of the instructions: so, 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) to make the task more demanding for the video quality. Based on pre-trials with colleagues and experience from previous experiments, we opted for having, per each group of participants, four rounds of seven minutes each, each round covering a different experimental condition (i.e., combination of dependent variables). Together with introduction, questionnaires and debriefing this would make for 2 hour sessions with each group.

B. 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 receivers monitor [51]. The parameters encoding and network can dynamically change during a session and thus are of more interest for optimization, in comparison to 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 1). As the perceived video quality might be influenced by 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 were 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 as *bitrate* and *loss*). We used a H.264 coded optimized for real time communication. The detailed configuration can be seen in Table 1.

We choose three encoding bitrates (with a for real-time communication configured version of H.264, see Table 1) 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. We further decided on two packet loss levels: (1) ***no packet loss***, typical for a wired connection and (2) **0.5% random packet loss**, a common scenario for a slightly impaired wireless network [28]. The screenshot in Fig. 2a shows the low encoding quality and Fig. 2b shows a screenshot

Table 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%)	

with high encoding quality with packet loss (of which the seen effect would mostly only last for a fraction of a second).

C. 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 in 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 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 [52]. 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 condition possible, for anchoring purposes. The experimental task then began, structured in for 7-minutes rounds 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 finished it. Between each condition we asked if a pause was needed. After the four rounds, a final questionnaire was administrated and participants were gathered again for a debriefing.

D. Apparatus

Each of the participants performed the task in a separate room with similar lighting and background conditions. A computer, display, webcam and headset (see Table 1 for detail) were provided. For the experimental task, we used the video conferencing client presented QoE-TB [51]. The software employs GStreamer for the 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.

[3] Exact model: Lego® item 6060873

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. The encoding bitrate can be set in the camera up to 20mbps, but tests showed that not more than 5.8mbps would be delivered in practice. The video was always captured in highest quality and then re-encoded with GStreamer x264 to have more control of the exact settings (see Table 1).

E. 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 [17]). We further asked six questions based on a questionnaire developed for engagement in computer usage [53]. After the experimental task was finished, one post experiment questionnaire regarding demographical information and enjoyment of the task was administered (questions shown in Table 2, except age and gender). All items (condition and post experiment) were assessed on a 5-point ACR scale [17]. In addition, throughout the experiment, the audiovisual streams of all sessions were captured on the sending and receiving sides.

IV. Data preparation and analysis

To answer the research questions presented in the introduction, we employ different techniques. First, we use descriptive models to analyze the relationship between the independent variables that we controlled (*bitrate* and *loss*) and interaction cues we extracted from the audiovisual streams. We then proceed 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 use feature selection with a machine learning approach. In the following, 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 apply in our analysis.

A. 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 [54]. Hence, we looked for indicators of these changes.

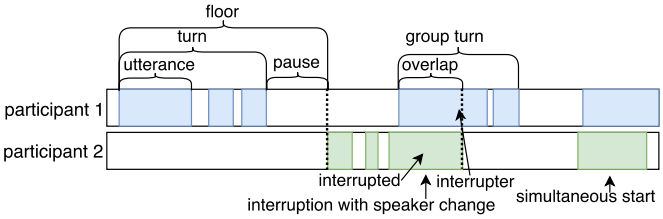


Fig. 3 Visualization of speech patterns

We used the data as received at the client side to include the system delay in the calculations. From the recorded audio we extracted chunks of speaking/not-speaking blocks with the help of the Adintool from the Julius voice recognition software[4]. The tool outputs blocks (start/end times) of voice activity. The blocks correspond to single utterances, which can then be investigated singularly or in groups to better understand speech patterns. The metrics are 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 [22], [55]):

- **Turn:** A sequence of blocks from a single speaker with less than 200ms pauses between the blocks (similar to a sentence, except that we don't 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
- **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

For each conversation element except for speaker alternation rate, 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

Table 2 - Questions of the post experiment questionnaire. The 'label'column indicates the abbreviation that will be 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

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

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 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 degradations 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 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 we are using Temporal Activity (TA, sometimes also referred to as Temporal Information - TI). TA is recommended by the ITU [56] to quantify the amount of movement present in videos, e.g. when comparing the performance of different encoders [57]. 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 [23] in presence of delay. TA is defined as the total change in luminance between one frame and the previous. For frame t_n:

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

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 participant movement. We computed TA for every participant and round by means of the mitsu video analytics toolset[5].

We further used the publicly available face recognition software OpenFace[6] [58] to quantify changes in distance of participants from the screen. OpenFace estimates the head position in 3D, rotation of the face and recognizes facial action units. We use 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 more away from the camera.

B. Statistical analysis

To investigate the influence of system on interaction and in turn on quality of experience we make use of linear mixed effect models (LMEs) [59]. Linear models are the simplest types of models that can be used to explain data; for Occam's razor principle, we prefer to employ those over non-linear ones to avoid overfitting. 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.

LMEs extend classical linear models, to adapt them to repeated measures experimental designs (such as ours, where subjects were exposed to multiple conditions). In repeated measures setups, groups of data may not be fully independent from each other. For example, QoE measures coming from the same participant may be similarly biased depending on the participant's individual preferences [60]. 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 (see e.g. the analysis of different groups in [4]), 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 repeated measure / between subject experiment design. In LMEs random factors can be modeled in a nested manner, here repeated measures from participants nested within groups, and can 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.

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

$$y = X\beta + Z\gamma + \epsilon \quad (2)$$

where X is the design matrix for fixed factors with the corresponding coefficients β, Z is the design matrix for the random factors with corresponding coefficients γ, and ϵ represents the residuals. It is assumed that the random effects are independent and distributed as $N(0, \tau^2)$, the errors are independent and distributed as $N(0, \sigma^2)$, and the random effects and errors are independent. This construction has the advantage of allowing an explanation of variance due to individual (or group) differences (random factor matrix Z), making the effect of the fixed factors more significant in turn. In our scenario this construction is particularly appealing, as we have repeated measures for both single participants and groups (due to the mixed block design). Variance in the data may be due to both individual preferences and group interactions, as we showed in our previous work [4]. Hence, to be able to quantify correctly influences of user, group interaction and system factors on QoE, we adopt LMEs for our analysis.

To assess whether a factor in our model has a significant impact we are using the likelihood ratio test (LRT) [61] which detects whether a model with the factor in question has a significant better fit than the same model but without the factor, in comparison to the additional parameters used. Having

[5] http://vq.kt.agh.edu.pl/index.html

[6] https://cmusatyalab.github.io/openface/

established that a factor has a significant impact on the dependent variable, we further investigate it in detail, clustering participants based on the factor with k-Means [62]. The fewer groups help to visualize and understand the effects of the factor better. The number of clusters was determined with an elbow plot [63].

C. Predictive model

The LMEs employed in previous work [4] rely on the availability of self-reported ratings of the users, which are available to us in the post analysis of an experiment, but not in real life scenarios. In this work we examine how well prediction would work if instead we include engagement, demographics and interaction cues in our models. A 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) [64]. 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 overfitting [64]. In this work, we make use of the Elastic Net [24], 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 employ 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 use the variance inflation factor (VIF), a statistical diagnostic method to check the severity of multicollinearity of fixed factors of a model [65]. Every factor $j = 1, \ldots, q$ is modeled as a linear combination of the other $q-1$ factors. The VIF is defined over the resulting R_j^2, i.e., the coefficient of determination for factor j of the model as:

$$VIF_j = \frac{1}{1-R_j^2} \quad (3)$$

Perfect independent variables that show no signs of correlation would have a 0 VIF (as a rule of thumb, VIF should be below 10 [65]).

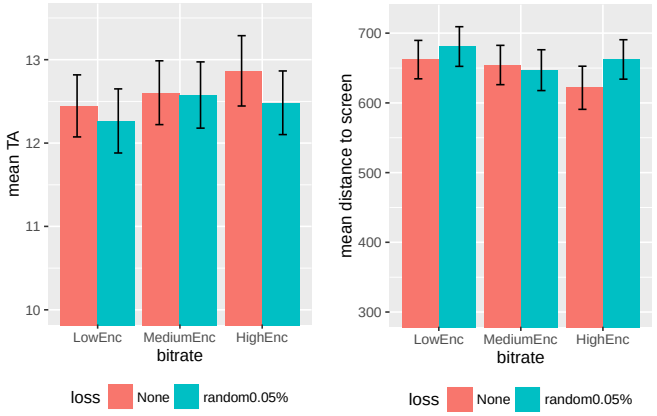


Fig. 4 (a) Mean Temporal Activity (TA) by bitrate and loss with 95% confidence intervals (b) Mean distance of participants from the screen, as impacted by bitrate and loss with 95% confidence intervals

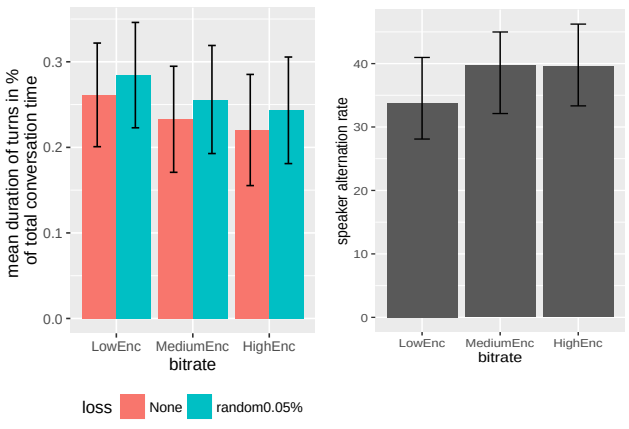
V. User Behavior Analysis

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. In turn this could lead to an increased speech activity to compensate.

For the analysis we use an LME (See section IV.B), 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.

A. Visual Interaction

As detailed in section IV.A we are using 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 are averaging *TA* and *DTS* per round. We first analyze the impact of system factors on *TA*. In Fig. 4a we can observe that the less impaired is the video (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). More in 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. 4b 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), with also interaction (p =0.03). Neither including *Group* as a random factor improved the fit (p = 0.65) nor did including *random slopes* per participant (p = 0.98). 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 without *packet loss* was significant ($p < 0.01$). In other words, with more impairments, participants moved less and were further away from the screen – both indicate that they interacted less visually.



(a) (b)

Fig. 5 (a) Percentage of turns by bitrate and loss (b) mean speaker alternation rate by bitrate. Each with 95% confidence intervals

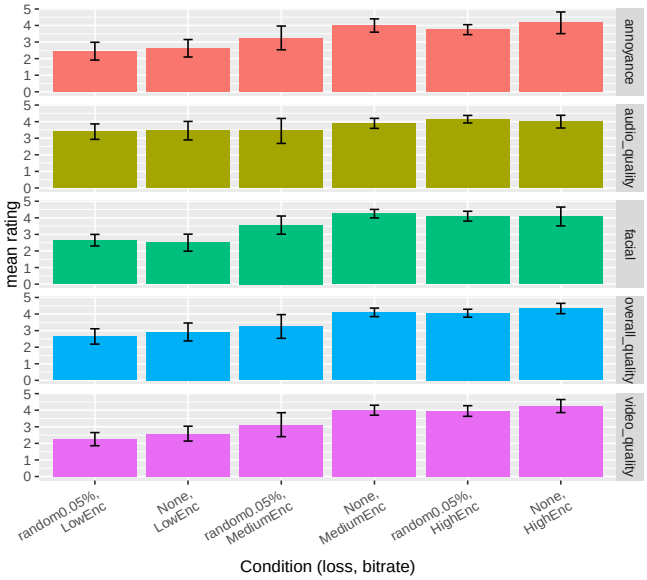


Fig. 6. Average ratings of the 5 QoE questions with 95% confidence intervals

B. *Speech Patterns*

As already done for the visual cues, we average the speech metrics per round. Perhaps due to the fact that they can rely less on the visual channel, participants seem 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 in Fig. 5a, 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. 5b). For *speaker alternation rate* there is a significant effect of *bitrate* ($p < 0.01$) but not of *loss* ($p > 0.05$). Further, participants are speaking 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 *low encoding* condition and the higher ones, we thus further corroborate our hypothesis that 1Mbit per second is sufficient to enable the task without hampering interaction.

VI. QoE User Factor Analysis

In previous work [4] we had investigated the impact of system factors on *audio*, *video* and *overall quality* as observed from the experiment reported in Section III. Fig. 6 shows average scores with 95% confidence intervals for the five questions inquiring about QoE (*overall*, *audio* and *video quality*, *annoyance* by video quality and recognition of *facial* expressions, see also section III.E) in the six experimental conditions, ordered according to the expected perceived quality. We can see in Fig. 6 that the ratings have a large variance, suggesting that factors other than *bitrate* and *loss* could have influence on QoE. To look deeper into this, we employed an LME (see Section IV.B) modeling *bitrate* and *loss* as fixed factors and *User* and *Group* as random factors. The p-values for the obtained contrasts [66] for each dependent variable are listed Table 3. We found a significant impact of the system factors, with the *high* and *medium encoding* obtaining significantly higher ratings (except for audio quality) than *low encoding*. However, the *high encoding* (4Mbit) did not increase the QoE significantly when compared to *medium encoding* (1Mbit). *Packet loss* also had an effect, albeit smaller. 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 the User and Group factors). In other words, different users were affected by system factors differently, and the group they carried out the experiment with also mattered. This further motivate us to look into how user factors affect QoE. Specifically, we will now 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.

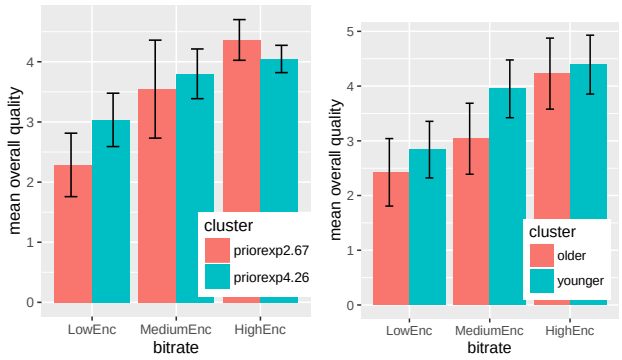
Table 3 Significant differences between Conditions for QoE questions

Question	LowEnc-HighEnc	LowEnc - MediumEnc	MediumEnc - HighEnc	None – random0.5%
Overall quality	**>0.01**	**>0.01**	0.22	**0.01**
video quality	**>0.01**	**0.01**	0.3	**0.02**
Audio quality	0.07	0.28	0.43	0.75
Annoyance	**>0.01**	**0.01**	0.3	**0.03**
Facial	**>0.01**	**>0.01**	0.62	0.18

A. *Prior Experience and Age*

Both prior experiences [43] and age [67] 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 [45]. 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*,



(a) (b)

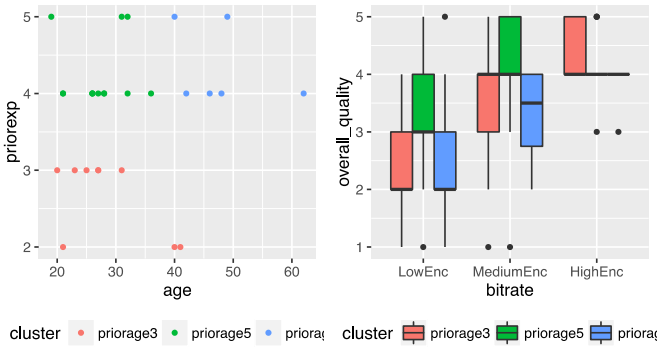
Fig. 7 Mean overall quality ratings with 95% confidence intervals by (a) prior experience groups (b) age groups

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.7a 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 more 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. 7b shows that the older age group scores QoE lower than the younger group ($p < 0.05$).

Interestingly, the LRT also revealed adding both factors and their interaction to the model was beneficial. By adding the combined *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. 8a. We found a young and experienced group (green), an older and experienced group (blue) and an unexperienced group (red). In Fig. 8b we can see that the younger and more experienced group (green) is indeed more relaxed then the other two groups; the less experienced younger participants and the older participants independent of prior experience.



(a) (b)

Fig. 8 Clustering by age and prior experience. a) (left) clusters by both factors. b) (right) overall quality ratings by clusters

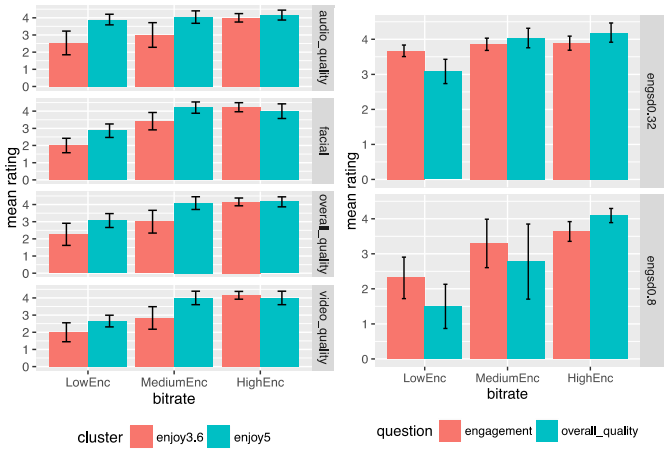
B. *Current state of the user*

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 [10], both as influencing factors and influenced variables. In this work we investigated them as influencing factors. We use *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 with flow, immersion in a task: it has been shown that impairments can disturb this flow [10], 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 however 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. 9a, 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 III.E). A reliability analysis revealed an excellent 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



(a) (b)

Fig. 9 (a) Significantly affected QoE ratings by enjoyment groups (b) Engagement and overall quality ratings by sd engagement clustering. Each with 95% confidence intervals.

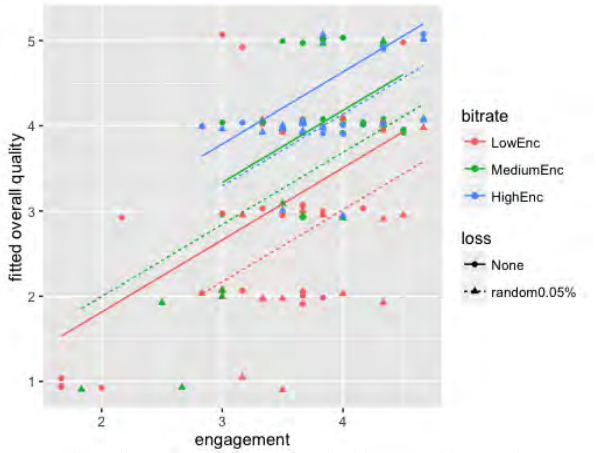


Fig. 10 Overall quality against engagement by bitrate and loss. Points represent the individual ratings (jitter on y for better visibility), lines represent the fitted model

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 [7], [39] 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 mean 0.3) whereas the other showed more variance (7 participants, sd mean .8). We can see in Fig. 9b that with different *engagement* in the same condition the experience of participants with more fluctuation in their engagement is significantly worse than their more engaged counterparts. We checked the contrasts to confirm that the ratings of both groups are statistically different ($p < 0.05$ except in the *high encoding* condition). Further 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* we show in Fig. 10 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*.

Interrestingly, *Engagement* explains a lot of the variance that we formerly had attributed to the random factors *User* and *Group* [4]. In Fig. 11 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 [4] as the one best explaining

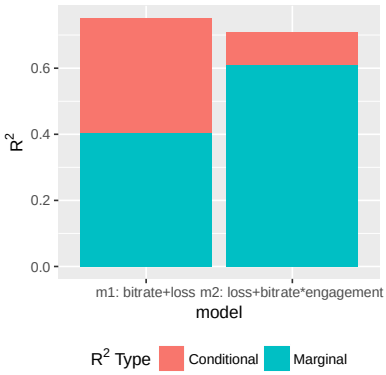


Fig. 11 Comparison of Marginal and Conditional R^2 of modeling overall quality with system factors alone (m1) or with engagement (m2). The random factors for both modes is (bitrate|User/Group).

variance in our data bsed 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*: *overall quality* ~ *bitrate+loss+(bitrate|User/Group)*). We introduce here model *m2*, which additionally includes *engagement* as a fixed factor, interacting with *bitrate* and *loss* (*m2*: *overall quality* ~ *(bitrate+loss)*engagement+(bitrate|User/Group)*). As we can see in Fig. 11, *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 the variance not explained by *bitrate* and *loss* in *m1*, and which westill followed a systematic within individual *Users* and *Groups,* can be for a large part explained by Engagement of the user with the conversation.

VII. A MODEL FOR PREDICTING VIDEOCONFERENCING QOE

So far, we have detailed how system factors influence the interaction of participants and the users current state, 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 IV.C, we will be using an elastic net to model QoE, as its properties fit our scenario well (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 (see IV.A), 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 4. We tested with the VIF (see section IV.C)

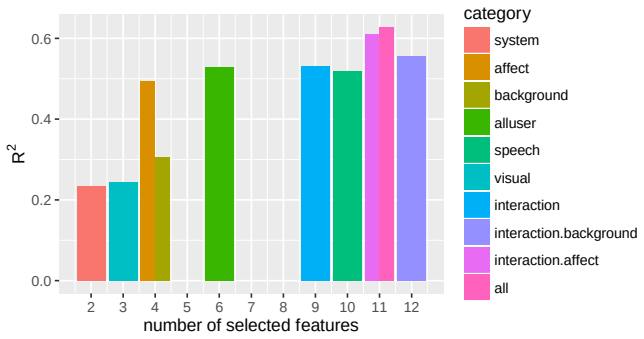


Fig. 12 R^2 and number of selected features for the QoE prediction model when fed with the different feature sets

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. 12 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 4), 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 4). The combination of interaction and user factors (*interaction* + *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.

VIII. Discussion

In this paper, we analyzed how bitrate and packet loss impact interaction and engagement in videoconferencing, and 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. Video conferencing shows its added value best in these situations, compared to audio-only solutions (e.g. telephone conference), 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.

With respect to interaction, we showed that low encoding (256kbit) had a significant impact on movement patterns of users as well as speech patterns, with respect to higher bitrates. 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 by talking more, as proven by an increase in the length of speaking turns.

Table 4 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

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 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 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 the home. While bandwidth is steadily increasing[7] so is the variation in them. In the foreseeable future users will be at locations in which no high speed connection is possible. Our study showed that the video quality is in such cases 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 [57]. 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. One 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 a valuable step into making video conferencing a tool that is available everywhere.

As detailed by conceptual models of the quality formation process [25], [2] the past experiences form a feedback loop influencing future QoE perceptions. While the effect has been studied in smaller scale [42], [43], 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 [68]. 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 degradations of streaming media over the internet, which we introduced in this study, are only 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 participants with similar level of experience suggests that the extent to which participants have dealt with degradations in the past plays a fundamental role in how their QoE is affected. Specifically, it would seem that the more participants are used to a certain type of artifact, the less this affects QoE, following an habituation process [68]. Of course, 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 further found that the QoE was influenced by engagement and enjoyment. While the main experience of users will be 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 [39] 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 [2], 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 affective state. Yet, to be useful within QoE control cycles, insights about user affect must correctly represent the current state of the individual user, yet measure it unobtrusively. 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

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

from objective data such as behavior, social cues or sensory data [69] are currently being developed by the Affective Computing community. For example, engagement can nowadays be inferred from physiological measurements (e.g. GSR or EEG) [70]. Audio and video cues [71], 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 [72]. Systems able to detect in real time mood and engagements 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.

As last step 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 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 that 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).

IX. Conclusion

In this paper 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.

In this context, we could clearly see that a video feed encoding bitrate of 256kbit was interfering with 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 affecting 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 degradations which lessens the effect on the experience. Further the QoE of more engaged participants was higher than that of the less engaged participants. It indicates how once a system has enabled users to engage in an interaction, participants will be quite forgiving about quality degradations, until it brings them out of the flow. With this data we tested predictive models and including all the examined factors did double 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.

X. Bibliography

[1] X. Zhang, Y. Xu, H. Hu, Y. Liu, Z. Guo, and Y. Wang, "Modeling and Analysis of Skype Video Calls: Rate Control and Video Quality," *IEEE Trans Multimed.*, vol. 15, no. 6, pp. 1446–1457, Oct. 2013.

[2] 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." Mar-2013.

[3] B. Belmudez, *Audiovisual Quality Assessment and Prediction for Videotelephony*. Cham: Springer International Publishing, 2015.

[4] M. Schmitt, J. Redi, P. Cesar, and D. Bulterman, "1Mbps is enough: Video quality and individual idiosyncrasies in multiparty HD video-conferencing," in *Proc. 8th QoMEX*, 2016, pp. 1–6.

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

[6] M. J. Scott, S. C. Guntuku, W. Lin, and G. Ghinea, "Do Personality and Culture Influence Perceived Video Quality and Enjoyment?," *IEEE Trans Multimed.*, vol. 18, no. 9, pp. 1796–1807, Sep. 2016.

[7] T. Hoβfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!," in *Proc. 3rd QoMEX*, 2011, pp. 131–136.

[8] E. Karapanos, J.-B. Martens, and M. Hassenzahl, "Accounting for diversity in subjective judgments," in *Proc. CHI*, New York, NY, USA, 2009, pp. 639–648.

[9] M. J. Scott, S. C. Guntuku, Y. Huan, W. Lin, and G. Ghinea, "Modelling Human Factors in Perceptual Multimedia Quality: On The Role of Personality and Culture," in *Proc. 23rd MM*, New York, NY, USA, 2015, pp. 481–490.

[10] K. De Moor, F. Mazza, I. Hupont, M. Ríos Quintero, T. Mäki, and M. Varela, "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.

[11] S. Egger, P. Reichl, and K. Schoenenberg, "Quality of Experience and Interactivity," in *Quality of Experience*, S. Möller and A. Raake, Eds. Springer International Publishing, 2014, pp. 149–161.

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

[13] M. Schmitt, S. Gunkel, P. Cesar, and D. Bulterman, "The Influence of Interactivity Patterns on the Quality of Experience in Multi-party Video-mediated Conversations Under Symmetric Delay Conditions," in *Proc. 3rd SAM*, New York, NY, USA, 2014, pp. 13–16.

[14] P. Reichl, S. Egger, S. Möller, K. Kilkki, M. Fiedler, T. Hossfeld, C. Tsiaras, A. Asrese, "Towards a comprehensive framework for QOE and user behavior modelling," in *Proc. 7th QoMEX*, 2015, pp. 1–6.

[15] J. Issing and N. Farber, "Conversational quality as a function of delay and interactivity," in *Proc. 20th SoftCOM*, 2012, pp. 1–5.

[16] K. Schoenenberg, A. Raake, and J. Koeppe, "Why are you so slow? – Misattribution of transmission delay to attributes of the conversation partner at the far-end," *Int. J. Hum.-Comput. Stud.*, vol. 72, no. 5, pp. 477–487, May 2014.

[17] ITU-T RECOMMENDATION, "ITU-R P.920 - Interactive test methods for audiovisual communications," 2000.

[18] D. S. Kirk, A. Sellen, and X. Cao, "Home video communication: mediating 'closeness,'" in *Proc. CSCW*, New York, NY, USA, 2010, pp. 135–144.

[19] B. Belmudez, S. Moeller, B. Lewcio, A. Raake, and A. Mehmood, "Audio and video channel impact on perceived audio-visual quality in different interactive contexts," in *Proc. MMSP*, 2009, pp. 1–5.

[20] F. Brauer, M. S. Ehsan, and G. Kubin, "Subjective evaluation of conversational multimedia quality in IP networks," in *Proc.10th MMSP*, 2008, pp. 872–876.

[21] 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, pp. 31–36.

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

[23] P. Romaniak, L. Janowski, M. Leszczuk, and Z. Papir, "Perceptual quality assessment for H.264/AVC compression," in *Proc. CCNC*, 2012, pp. 597–602.

[24] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[25] J. Skowronek and A. Raake, "Conceptual Model of Multiparty Conferencing and Telemeeting Quality," in *Proc. 7th QoMEX*, 2015.

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

[27] G. Berndtsson, M. Folkesson, and V. Kulyk, "Subjective quality assessment of video conferences and telemeetings," in *Proc. 19th PV*, 2012, pp. 25–30.

[28] M. Ndiaye, M. C. Larabi, H. Saadane, G. L. Lay, C. Perrine, C. Quinquis and L. Gros, "Subjective assessment of the perceived quality of video calling services over a real LTE/4G network," in *Proc. 7th QoMEX*, 2015, pp. 1–6.

[29] T. Hayashi, K. Yamagishi, T. Tominaga, and A. Takahashi, "Multimedia Quality Integration Function for Videophone Services," in *Proc GLOBECOM*, 2007, pp. 2735–2739.

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

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

[32] J. Greengrass, J. Evans, and A. C. Begen, "Not all packets are equal, part 2: The impact of network packet loss on video quality," *Internet Comput. IEEE*, vol. 13, no. 2, pp. 74–82, 2009.

[33] O. Boyaci, A. G. Forte, and H. Schulzrinne, "Performance of Video-Chat Applications under Congestion," in *Proc. 11th ISM*, 2009, pp. 213–218.

[34] P. Calyam, M. Haffner, E. Ekici, and C.-G. Lee, "Measuring Interaction QoE in Internet Videoconferencing," in *Real-Time Mobile Multimedia Services*, D. Krishnaswamy, T. Pfeifer, and D. Raz, Eds. Springer Berlin Heidelberg, 2007, pp. 14–25.

[35] M. Schmitt, S. Gunkel, P. Cesar, and D. Bulterman, "Asymmetric delay in video-mediated group discussions," in *Proc. 6th QoMEX*, 2014, pp. 19–24.

[36] S. Egger, R. Schatz, and S. Scherer, "It takes two to tango-assessing the impact of delay on conversational interactivity on perceived speech quality," in *Proc. 11th ISCA*, 2010.

[37] Y. Yamakata, A. Miyazawa, A. Hashimoto, T. Funatomi, and M. Minoh, "A Method for Detecting Gaze-required Action While Cooking for Assisting Video Communication," in *Proc. UbiComp*, New York, NY, USA, 2014, pp. 577–582.

[38] S. Egger and P. Reichl, "A Nod Says More than Thousand Uhmm's: Towards a Framework for Measuring Audio-Visual Interactivity," in *Proc."The Good, the Bad and the Challenging ", COST298 conference, Copenhagen, Denmark*, 2009.

[39] M. Schmitt, J. Redi, and P. Cesar, "Towards context-aware interactive Quality of Experience evaluation for audiovisual multiparty conferencing," in *Proc. 5th PQS*, Berlin, 2016, pp. 64–68.

[40] M. T. Diallo, N. Marechal, and H. Afifi, "A Hybrid Contextual User Perception Model for Streamed Video Quality Assessment," in *Proc. ISM*, 2013, pp. 518–519.

[41] K. Kilkki, "Quality of experience in communications ecosystem," *J. Univers. Comput. Sci.*, vol. 14, no. 5, pp. 615–624, 2008.

[42] J. Shaikh, M. Fiedler, P. Paul, S. Egger, and F. Guyard, "Back to normal? Impact of temporally increasing network disturbances on QoE," in *Proc. GC Wkshps*, 2013, pp. 1186–1191.

[43] T. Hoßfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler, "The Memory Effect and Its Implications on Web QoE Modeling," in *Proc. 23rd ITC*, San Francisco, California, 2011, pp. 103–110.

[44] S. Kurniawan, "Older people and mobile phones: A multi-method investigation," *Int. J. Hum.-Comput. Stud.*, vol. 66, no. 12, pp. 889–901, Dec. 2008.

[45] M. Kobayashi, A. Hiyama, T. Miura, C. Asakawa, M. Hirose, and T. Ifukube, "Elderly User Evaluation of Mobile Touchscreen Interactions," in *SpringerLink*, Springer Berlin Heidelberg, pp. 83–99.

[46] J. Palhais, R. S. Cruz, and M. S. Nunes, "Quality of Experience Assessment in Internet TV," in *SpringerLink*, Springer Berlin Heidelberg, pp. 261–274.

[47] H. O'Brien and P. Cairns, Eds., *Why Engagement Matters*. Cham: Springer International Publishing, 2016.

[48] J. A. Bergstra and C. A. Middelburg, "ITU-T Recommendation G.107 : The E-Model, a computational model for use in transmission planning," 2003.

[49] ITU-T RECOMMENDATION, "ITU-T G.1070 Opinion model for video-telephony applications." Jul-2012.

[50] J. Song, F. Yang, Y. Zhou, S. Wan, and H. R. Wu, "QoE Evaluation of Multimedia Services Based on Audiovisual Quality and User Interest," *IEEE Trans Multimed.*, vol. 18, no. 3, pp. 444–457, Mar. 2016.

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

[52] P. ITU-T RECOMMENDATION, "ITU.P913: Subjective video quality assessment methods for multimedia applications," 1999.

[53] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 1, pp. 50–69, Jan. 2010.

[54] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *Sel. Areas Commun. IEEE J. On*, vol. 9, no. 4, pp. 586–593, 1991.

[55] A. J. Sellen, "Remote conversations: the effects of mediating talk with technology," *Hum-Comput Interact*, vol. 10, no. 4, pp. 401–444, Dec. 1995.

[56] ITU-T, "ITU-T Recommendation P.910 - Subjective video quality assessment methods for multimedia applications," 1995.

[57] M. Řeřábek and T. Ebrahimi, "Comparison of compression efficiency between HEVC/H. 265 and VP9 based on subjective assessments," in *SPIE Optical Engineering+ Applications*, 2014, p. 92170U–92170U.

[58] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.

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

[60] Y. Zhu, I. Heynderickx, A. Hanjalic, and J. A. Redi, "Towards a comprehensive model for predicting the quality of individual visual experience," 2015, p. 93940A.

[61] A. C. Davison, *Statistical Models*, 1 edition. Cambridge University Press, 2008.

[62] J. MacQueen and others, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. on Math. Statist. and Prob.*, 1967, vol. 1, pp. 281–297.

[63] D. J. Ketchen and C. L. Shook, "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique," *Strateg. Manag. J.*, vol. 17, no. 6, pp. 441–458, Jun. 1996.

[64] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin, 2001.

[65] M. Kutner, C. Nachtsheim, and J. Neter, *Applied Linear Regression Models- 4th Edition with Student CD*, 4 edition. Boston; New York: McGraw-Hill Education, 2004.

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

[67] J. A. Redi, Y. Zhu, H. de Ridder, and I. Heynderickx, "How Passive Image Viewers Became Active Multimedia Users," in *Visual Signal Quality Assessment*, C. Deng, L. Ma, W. Lin, and K. N. Ngan, Eds. Springer International Publishing, 2015, pp. 31–72.

[68] M. P. Domjan, *The Principles of Learning and Behavior*, 7 edition. Stamford, CT: Cengage Learning, 2014.

[69] M. Soleymani and M. Pantic, "Human-centered implicit tagging: Overview and perspectives," in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, 2012, pp. 3304–3309.

[70] C. Wang, E. N. Geelhoed, P. P. Stenton, and P. Cesar, "Sensing a live audience," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, 2014, pp. 1909–1912.

[71] W. Mou, H. Gunes, and I. Patras, "Alone Versus In-a-group: A Comparative Analysis of Facial Affect Recognition," in *Proc. ACM MM*, New York, NY, USA, 2016, pp. 521–525.

[72] A. Bhattacharya, W. Wu, and Z. Yang, "Quality of experience evaluation of voice communication systems using affect-based approach," in *Proc. ACM MM*, New York, NY, USA, 2011, pp. 929–932.