IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 15, 2020, accepted May 29, 2020, date of publication June 5, 2020, date of current version June 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3000467

QoE Assessment of Mobile Multiparty Audiovisual Telemeetings

DUNJA VUČIĆ [1], (Member, IEEE), AND LEA SKORIN-KAPOV [2], (Senior Member, IEEE)
[1]Ericsson Nikola Tesla d.d., 10000 Zagreb, Croatia
[2]Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia

Corresponding author: Dunja Vučić (dunja.vucic@ericsson.com)

This work was supported in part by the Croatian Science Foundation under Project IP-2019-04-9793 (Q-MERSIVE).

ABSTRACT The expectations of modern mobile users are increasingly moving towards being able to access demanding services regardless of context or system influence factors, such as network conditions, service topology, and device processing capabilities. Multiparty audiovisual telemeetings are an example of a real-time, delay sensitive, and heavy load service, demanding to run on smartphones that are limited in display size, processing power, and battery capacity. In this paper, we first provide an overview of multiparty audiovisual calls established via mobile devices and key aspects influencing Quality of Experience (QoE). We then report on the results of five user studies conducted over the course of the past 4 years, focused on investigating the impact of video quality in terms of different video encoding parameter configurations (namely bitrate, frame rate, and resolution) on subjective QoE scores for WebRTC-based video calls. We identify lower and upper bounds on video configuration parameters when used in the context of three-party calls. Results have shown that in certain cases it is better to provide constant lower objective video quality than to switch between higher and lower qualities, since participants start to perceive impairments. Finally, we investigate the relationship between objectively measured video quality impairments (blurriness and blockiness) and subjective user scores. Obtained results indicate that the Birnbaum-Saunders distribution for blockiness and the Burr and Gamma distributions for blurriness provide good fits for quality ratings. Gathered results aim to provide input for deriving QoE-aware service adaptation strategies, enabling increased resource allocation efficiency while maintaining acceptable end-user QoE.

INDEX TERMS Mobile multiparty telemeetings, quality of experience, user studies, video encoding parameters, adaptation strategies, blockiness, blurriness.

I. INTRODUCTION

Mobile devices, services, and applications have become an inseparable part of our daily lives, affecting relationships, social norms, and communication and interaction methods. With technological advancements, we are witnessing changes in end users' expectations in terms of service quality and performance, both in private and business contexts. Consequently, mobile operators are in the process of planning and deploying ultra fast and low latency 5G networks, expected to cross new performance thresholds in connectivity speeds, number of connected devices, and possible services. Among those services leveraging network support for heavy load and low latency communication, multiparty audiovisual calls are expected to further gain popularity.

A key challenge to address is meeting end user Quality of Experience (QoE) requirements, as well as making efficient use of network and system resources. To gain insights into how user-perceived service quality can be measured and optimized within available resource constraints and in a given context of use, there is a need to assess the impact of various relevant QoE influence factors.

With respect to standards, comprehensive recommendations are given for subjective quality assessment of multiparty audiovisual calls. Relevant recommendations focus on a specific test modality (audio, video, audiovisual) and paradigm (interactive, non-interactive), with the goal being to test various conditions or parameters, such as codec type, fixed or variable bitrate, frame rate, resolution, noise cancellation, background noise, synchronization or transmission impairments [1].

On the other hand, objective full reference, reduced reference, and no reference metrics [2] enable less expensive and less time consuming quality assessment, however at the cost of certain deviation from actual subjective user perception. In cases of multiparty video calls established via

The associate editor coordinating the review of this manuscript and approving it for publication was Li Minn Ang.

smartphones, full-reference metrics are usually not convenient to measure in real life scenarios. Even if we disregard delay or packet loss per individual stream, and have perfectly ordered streams and synchronized screen recordings from all participants, the issue still remains that participants' previews can be placed in different sized windows with different zooming possibilities (Figure 1). For applications that allow re-arrangement of preview windows on the screen, window positions can be different for each participant. As a consequence, in our studies we focus on no-reference video quality metrics and explore the relation with subjective ratings. Moreover, studies have shown that participants differently rate overall screen quality (encompassing all video streams portrayed simultaneously on the screen) as compared to the mean quality of each individual video stream [3].

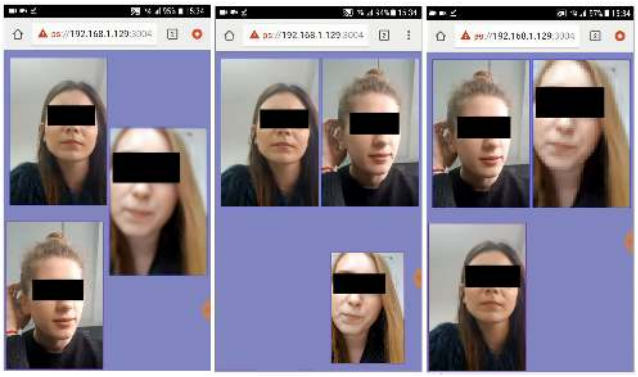

FIGURE 1. An example of different screen layouts portrayed on three different participant devices during a three-way audiovisual call.

Given the complexity of assessing QoE for mobile multiparty video calls, and in understanding the impact of underlying influence factors, the contributions of this paper may be considered as twofold. First, we focus on the challenge of service quality management by investigating how to **adapt video encoding parameters so as to achieve the best possible QoE**, while delivering the service under various network and system resource constraints. Secondly, we study **the relationship between user ratings and distributions of objective video metrics**, namely blurriness and blockiness measured per video frame. Such a relationship can provide valuable insights when aiming to model and estimate QoE based on objectively measurable metrics.

The focus of our research has been on multiparty video calls established via smartphone devices. The employed research methodology is portrayed in Figure 2, highlighting the key steps involved in assessing the impact of selected video parameters (bitrate, resolution, frame rate) on QoE, under various system and network conditions. We report on the results of five different user studies, three of which were conducted in a laboratory environment and two in a home environment, involving a total of 141 participants. Participants took part in three-party video calls using various smartphone devices. Results of the studies are elaborated on in this paper and quantify the impact of both device capabilities and (adjusted) video quality on QoE.

This paper extends our previous works and brings together key results from studies conducted over the past four years and published in our earlier conference papers ([4]–[7]). Conducted measurements were based on the open source WebRTC (Web Real-Time Communication) technology that provides real-time audio and video communication within browsers without the need for plugins or other third-party software [8]. We build on these results with new subjective studies aimed at determining the lower threshold of video bitrate and resolution (per video stream) needed to achieve acceptable QoE. Finally, we use the results of conducted studies to investigate the relationship between subjective user scores (MOS) and objective video metrics.

The remainder of the paper is organized as follows: Section II discusses definitions related to multiparty audiovisual telemeetings. In Section III, we describe the quality assessment aspects and analyze related work on QoE studies. Section IV gives an overview of previously conducted QoE studies under different test conditions, including a short description of the employed test methodology and key results obtained for each study. Section V presents the details of a QoE study which involved the collection of both subjective scores and objective video metrics (blurriness and blockiness). The relationship between subjective and objective metrics are analyzed. Finally, Section VI discusses limitations of the conducted studies, summarizes conclusions, and provides an outlook for future research.

II. MULTIPARTY AUDIOVISUAL TELEMEETINGS

ITU-T Recommendation P.1301 defines terms and methods for the subjective quality assessment of audio and audiovisual multiparty telemeetings [9]. A *telemeeting* is defined as a meeting in which participants are located in at least two different locations and the communication takes place via a telecommunication system. The term *multiparty* indicates the involvement of more than two participants in a telemeeting. With respect to the number of participants and number of participant locations, the following setup can be used: two sites with more than one person at at least one site (*multiparty point-to-point*), more than two sites with one person at each site (*multiparty one-per-site*), and more than two sites with more than one person at at least one site (*multiparty multi-point*).

The term telemeeting presents communication used in conventional business video conferencing scenarios, as well as in more flexible private meetings held in a leisure context [10]. While telemeetings organized in a business context generally have specific objectives and agendas, with a set of tasks that must be completed, telemeetings held in a private/leisure context generally have the primary objective of experiencing a sense of presence or social connection [9]. Due to different objectives corresponding to different meeting contexts, the quality expected by the participants may be different,

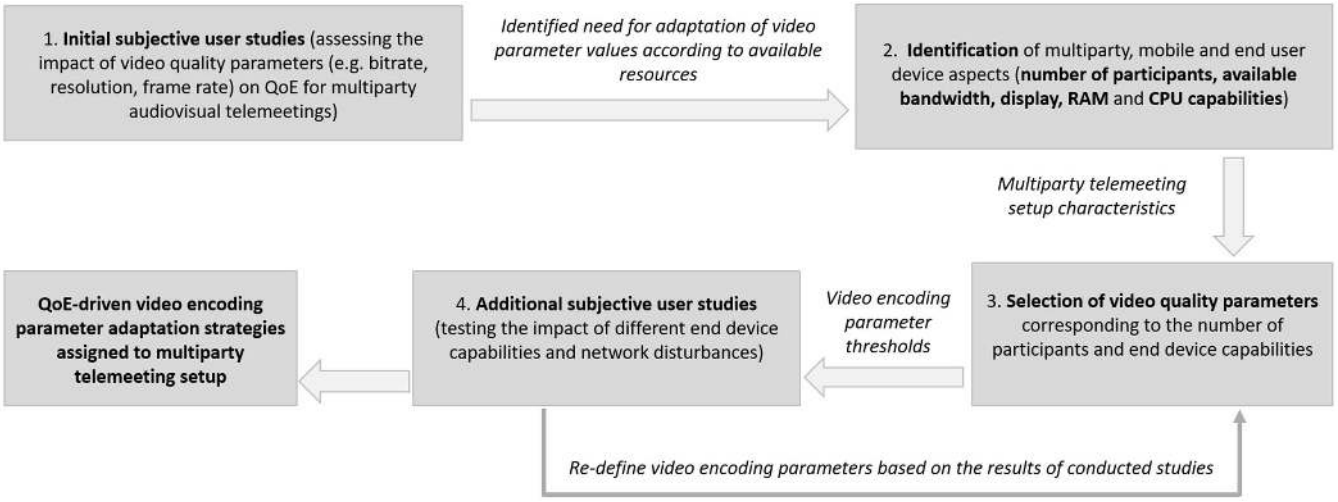

FIGURE 2. Research methodology for deriving QoE-driven video encoding parameter adaptation strategies for multiparty telemeetings.

with participants likely being less critical in the private context [11], [12].

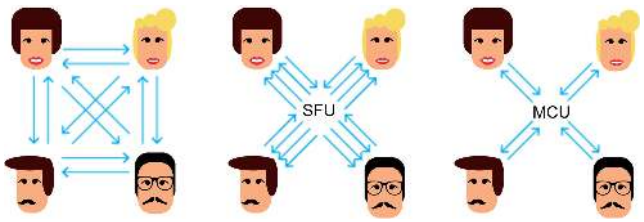

FIGURE 3. Different topologies used during multiparty calls: Peer-to-Peer, Selective Forwarding Unit (SFU), and Multipoint Control Unit (MCU).

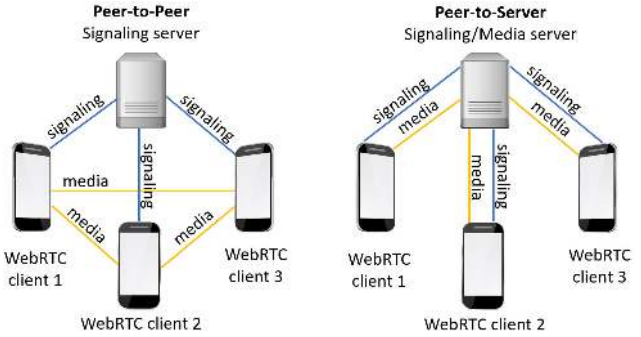

FIGURE 4. Example for a three-party WebRTC Peer-to-Peer and Peer-to-Server architecture.

With respect to technical realization, several connection types are possible for multiparty audiovisual telemeetings, as illustrated in Figure 3. One possibility is a *full-mesh topology*, where in communication with n peers, each peer handles $n-1$ download streams and $n-1$ upload streams (for illustration purposes, we assume each peer to be transmitting one stream). Peer-to-Peer topology is most affordable but requires a high amount of processing power, lacking in older smartphones, and higher capacity in terms of available bandwidth. To release the load on both the end user device resources as well as the network, part of the processing and data transmission burden may be shifted to a centralized media server, albeit with potentially higher operational costs (due to administration, signaling, and media distribution) [13]. A *Selective Forwarding Unit (SFU)* requires peers to upload their own stream, and distributes it to all other connected peers. Each peer handles one upload stream and $n-1$ download streams (for illustration purposes, we assume each peer to be transmitting one stream). Finally, peers connected to a so-called *Multipoint Control Unit (MCU)* generally handle one upload and one download stream, while the MCU is responsible for mixing uploaded streams into a single stream, adapting streams, and distributing to other peers [14].

With respect to video conferencing platforms and technological solutions, WebRTC has become a widely popular technology and framework for developing real-time multimedia communication applications. WebRTC standardization activities are conducted by two groups, the W3C (World Wide Web Consortium) responsible for defining Javascript API interfaces [8], and the IETF (The Internet Engineering Task Force) RTCWEB group responsible for the architecture and protocol requirements [15]. WebRTC includes interfaces built into the browser for capturing and coding of media streams from local devices (e.g., video cameras and microphones), peer connection establishment, and media and data transfer to remote participants. The basic WebRTC architecture includes a server and at least two peers. Each peer loads the application in their local environment (browser). A WebRTC session can be established directly between browsers (P2P), or indirectly via a signaling and media server (Figure 4), with standards and protocols providing mechanisms for connection establishment and NAT (Network Address Translation) traversal. In all cases, a signaling server is required to establish communication between end-user browsers [14].

A WebRTC application flow involves opening peer connections, discovering peers, and support for media streaming.

Captured raw media streams are passed to the encoder for data compression. Where applicable, Forward Error Correction (FEC) mechanisms are used for packetized audio samples and video frames with application specific headers [16]. The WebRTC standard mandates the support for certain audio (Opus and G.711) and video codecs (H.264 and VP8), while other codecs are also commonly supported (e.g., VP9 for video).

Earlier studies have shown that previous deployments of video conferencing applications such as Google+, iChat, and Skype based on peer-to-peer topology had problems providing good quality to their end users and sustaining high-quality multiparty video conferencing services over the ''best-effort'' Internet [17], [18]. As a result, nowadays multiparty applications commonly employ an architecture relying on cloud-based models, such as SFU or MCU. Together with support for quality adaptation, such service architectures have led to improved end user QoE. However, user mobility and variable resource availability still remain a challenge, with operators unable to provide full coverage, thus resulting in users experiencing varying levels of service quality over the course of one session [19].

III. STANDARDIZATION AND RELATED WORK

A. QUALITY ASSESSMENT ASPECTS

Multiparty audiovisual telemeetings are commonly assessed using opinion purpose-designed questionnaires, where participants complete and report ratings for audio-visual and overall quality, AV synchronization and/or interactivity degradation. In terms of time frame, quality may be assessed [20]:

- after stimulus/stimuli presentation, or
- continuously during stimulus/stimuli presentation.

Different rating scales are used to correlate opinions with numerical values, enabling the calculation of arithmetic means (in the case of ordinal scales assuming equal intervals between quality levels). Furthermore, different scales are used depending on the judgment type. If ratings are collected *after* participants have been exposed to stimuli, then it is common to use a discrete 5-point absolute category rating (ACR) scale for quality marked with: 1 ***"Bad"***, 2 ***"Poor"***, 3 ***"Fair"***, 4 ***"Good"***, 5 ***"Excellent"***, while interactivity degradation is marked with: 1 ***"Very annoying"***, 2 ***"Annoying"***, 3 ***"Slightly annoying"***, 4 ***"Perceptible but not annoying"***, 5 ***"Imperceptible"*** [21]. For *instantaneous* judgment, a continuous scale with the same labels is suggested. Participants assess quality by moving a slider during the session, where the slider position corresponds to the currently perceived quality level [22].

In the context of quality assessment for multiparty telemeetings, an important aspect to consider is the *number of participants* and *number of participant locations* [9]. Furthermore, an important aspect of the telemeeting system that has to be considered is communication mode, which may refer to *audio-only, video-only, and audiovisual* mode. Evaluation can further differ in terms of the type of quality dimension. Non-interactive quality may be assessed by listening-, viewing-, or listening-and-viewing-only quality of test stimuli, while conversational/interactive quality is commonly assessed by participants engaged in an actual conversation.

Every experimental design decision impacts the user experience, hence it is important to pay attention to the *type of task* as well. Standards recommend use of the free conversation test task, due to the resemblance to real-life natural conversations, thus enabling participants to keep their focus on the screen.

Finally, with respect to the test methodology, an important consideration is the system set-up. In a multiparty environment, participants may be using heterogeneous devices and access networks. Consequently, they may not only experience different impairments and quality degradation, but may also have different quality expectations.

We note that all studies reported in this paper involved three-party calls with participants located at three different sites. Participants took part in free conversation, and assessed conversational/interactive quality. In User Studies 2 through 5, a symmetric setup was used (with a centralized media server) under controlled conditions.

Having described key characteristics of multiparty telemeetings, and important aspects to consider when conducting subjective studies, we now give an overview of related studies that have investigated the impact of various IFs on QoE.

B. RELATED STUDIES ON QoE FOR MULTIPARTY AUDIOVISUAL TELEMEETINGS

One of the most commonly cited definitions of QoE defines it to be *"the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state"* [23]. The Qualinet White Paper further identifies QoE IFs as *"any characteristic of a* ***user****,* ***system****, service, application, or* ***context*** *whose actual state or setting may have influence on the Quality of Experience for the user"*.

The complexity in assessing and modeling QoE arises not only from the multitude of different system, context, and user IFs [24], but from the difficulty in controlling certain factors during studies. As we discovered during our studies, certain factors may be unintentionally manipulated during the course of tests, hence impacting user ratings (e.g., perceived video quality degraded due to unintentional/unplanned device CPU overuse).

Experiments conducted over Wi-Fi, focusing on mobile video call quality, showed sensitivity to bursty packet losses and long packet delays [25]. De Moor *et al.* evaluated the impact of impaired video (with a 20% packet loss), impaired audio (with restricted CPU usage on the client WebRTC application) and both streams, audio and video with 500 ms delay and 300 ms jitter [26]. Results showed that disturbances in both audio and video had the most negative impact

on overall quality, while video-only impaired scenarios performed somewhat better than audio-only impaired scenarios.

Based on a conducted survey involving 140 participants, Husić *et al.* identified the following seven factors as having the strongest impact on user satisfaction in the case of WebRTC video calls: audio quality, image quality, quality of service, service price, loss of video frames, ease of use, and procedure of accessing web environment [27]. Based on this classification, García *et al.* proposed the following key performance indicators for QoE estimation: call establishment time, end-to-end delay, perceived audio, video, and audiovisual quality [28]. Skowronek *et al.* identified mobility, device and encoding interoperability, ease of use, and additional collaboration possibilities (e.g., exchanging pictures, files, chatting) as the most important aspects for telemeeting services [29].

Studies exploring video quality for telemeeting scenarios in different contexts with combinations of factors such as resolution, encoding bitrate, viewing distance, and up-scaling of video formats found that bitrate and viewing distance were the most significant factors affecting subjective video quality [30]. The efficiency of video compression may be considered in terms of achievable compression ratio with minimal or non-perceivable quality degradation. High compression ratios lead to perceptual spatial or temporal artifacts. Spatial artifacts such as blocking, blurring, ringing, basis pattern effect, and color bleeding can be detected within individual frames, when the video is paused, and with no need to reference adjacent frames. Temporal artifacts such as flickering, jerkiness and floating can be noticed while the video is being played [31].

Jana *et al.* [32] investigated video artifact evaluation for two-way video conversations in stationary and mobile scenarios using the following no-reference spatial metrics: blocking, blurring, and temporal smoothness. Results showed that blocking and blurring are highly correlated when they are caused by packet loss. However, different coding techniques can perform differently in terms of avoiding loss of high frequency components, and thus show less blurring or blocking in different contexts.

Silva *et al.* conducted experiments measuring user annoyance caused by different strength combinations of blockiness, blurriness, and packet loss intensity. Disturbances were inserted in video sequences characterized by diverse content and displayed to subjects on a 23 inch monitor [33]. Results showed that subjects were able to identify artifacts only when one source of impairment with high strength was present, while they had difficulties identifying low strength artifacts. A higher level of annoyance correlated with more artifacts being included in the experiment and their respective intensities. Subjects reported that blockiness had the strongest impact on "annoyance", and in some cases blurriness masked impairments caused by packet loss.

The possibility to estimate perceivable quality impairments in terms of blockiness and audio distortion using machine learning, and to predict the occurrence of disturbances was investigated in [34]. The authors studied call scenarios with no impairments and with realistic technical impairments (packet loss and delays). Results showed that impairments could be estimated with a high level of accuracy, thus proving the potential of exploiting machine learning models for automated QoE-driven monitoring and estimation of WebRTC performance.

In audiovisual conversational services, the *task* being performed is recognized as a significant QoE IF. Schmitt *et al.* investigated the impact of video quality on the ability to interact in experiments involving a four-party desktop video conference, where participants were given the task of collaboratively building a Lego model. Results showed that subjects with a higher engagement in the task reported a higher QoE [35]. Similar findings were reported in [26], where as a conversation incentive, a *Celebrity name guessing* task was used. The authors concluded that the test task was more engaging than intended, consequently impacting QoE ratings in an unwanted way. In [36], the authors explored the effect of task complexity and/or duration in the case of WebRTC video calls (established over smartphones). Obtained results confirmed that QoE is significantly determined by the task complexity and duration.

While it is clear that a wide range of system, context, and human IFs affect QoE in multiparty audiovisual telemeetings, questions remain as to the level of the impact of particular factors, especially in a mobile context. For example, the question of whether certain impairments cause strong, noticeable or imperceptible quality degradations commonly depends on the particular scenario, context, as well as the individual involved users. In the following sections, we narrow down our focus and study mobile three-party video calls conducted in a leisure context. In terms of QoE IFs, we focus on **device capabilities**, various **video encoding parameters**, and network impairments such as **packet loss**.

IV. OVERVIEW OF SUBJECTIVE QoE STUDIES

A. OVERALL GOALS AND METHODOLOGY

Deploying multiparty video communication solutions on smartphones calls for the need to optimize video encoding parameters due to limited device processing power and dynamic wireless network conditions. Given the mobile device context and corresponding screen sizes, the question arises as to which video quality levels should be maintained during a call so as to achieve acceptable QoE. In other words, increasing video quality beyond a certain threshold will likely not contribute to user perceivable QoE improvement. In cases of variable and limited system and network resource availability, video encoding adaptation strategies may be deployed to downsize traffic by adapting parameters such as bitrate, resolution, and frame rate, so as to optimize end user QoE.

Our main research focus has thus been geared towards deriving QoE-driven service adaptation strategies, based on the adjustment of video encoding parameters in accordance

TABLE 1. Characteristics and capabilities of smartphones used over the course of reported studies.

Parameter	Samsung SIII	Samsung S5	LG G3	Samsung S6	Samsung S7
Chipset	Exynos 4412 Quad	Qualcomm MSM8974AC Snapdragon 801	Qualcomm MSM8974AC Snapdragon 801	Exynos 7420 Octa	Exynos 8890 Octa
CPU	Quad-core 1.4 GHz Cortex-A9	Quad-core 2.5 GHz Krait 400	Quad-core 2.5 GHz Krait 400	Octa-core (4x2.1 GHz Cortex-A57 4x1.5 GHz Cortex-A53)	Octa-core (4x2.3 GHz Mongoose 4x1.6 GHz Cortex-A53)
GPU	Mali-400MP4	Adreno 330	Adreno 330	Mali-T760MP8	Mali-T880 MP12
RAM	1 GB	2 GB	2 GB	3 GB	4 GB
Display size	4.8"	5.1"	5.5"	5.1"	5.1"
Display resolution	720 x 1280 px	1080 x 1920 px	1440 x 2560 px	1440 x 2560 px	1440 x 2560 px

with available resources. Given the wide range of potential test conditions, in this section we report on a number of subjective user studies we have conducted over the course of the past four years, and highlight the main findings of each study. The study goals, set-up, and main findings are summarized in Table 2.

Studies were based on investigating the impact of different parameters such as video bitrate, frame rate, resolution, and smartphone capabilities (Table 1) on QoE. Test setup included three-party symmetric and asymmetric conditions, with the setup involving both natural home and laboratory environments. Communication flows were realized via both the public Internet, and in a controlled local area network. In all cases, the audiovisual telemeetings were realized via the WebRTC paradigm over UDP [37]. With respect to codecs, VP8, G.711 and Opus were used. The VP8 codec is a royalty free codec and is based on two frame types: intraframes and interframes [38]. Intraframes, known as key frames, are decoded without reference to any other frame in a sequence. Interframes are encoded with reference to prior frames, specifically all prior frames up to and including the most recent key frame.

While our initial study (US1) involved asymmetric end user device conditions, we later opted to avoid test design complexity caused by the influence of different devices. Consequently, after the first study, we started to use a symmetric setup, so as to maintain a similar quality of captured and reproduced audio and video at each participant. In all subsequent studies, we therefore preset the same quality per outgoing streams for all participants. To further decrease the potential impact of contextual factors, participants were further not able to customize the layout of the application.

In a real-time videoconferencing system with more than two interlocutors and a telemeeting established via the Internet, we are unable to completely control end-to-end network performance. Therefore, to be able to fully control network conditions and impairments, most studies were conducted in a controlled environment and local area network.

The placement of the camera and microphone on the smartphone in relation to the participants was arbitrary. In all of our tests, participants were free to hold the smartphone in their hand or place it on a stand provided to them at the viewing distance and position they preferred.

With respect to selected study participants, ages ranged from 20 to 65, and all were non-experts in the AV field. All participants had good hearing and viewing abilities (some with corrected vision - glasses or lenses).

It is important to highlight that all measurements were conducted in a **leisure context** between three acquaintances, ensuring a smooth and continuous flow of conversation. The conversations were all conducted using the Croatian language, as this was the native language for all participants. Participants did not use written materials and they were instructed to use natural conversation without any predefined task, trying to retain their attention on the mobile device.

For all studies, participants were located in separated rooms, one person per room, with different acoustics and background noise characteristics as well as video backgrounds and room colors. We performed measurements both during daylight and with artificial lights, avoiding direct light sources on the participants and cameras.

In all user studies, at the beginning of each test session, a preliminary test was carried out aimed to familiarize participants with the task and assessment questionnaire, and to make sure they felt comfortable during the evaluation. Preliminary results were not taken into account.

To prevent participant fatigue, we adhered to relevant standards, which state that the total number of tests must be reasonable and limited [39]. The total time for testing should be balanced with respect to the time spent engaging in the service per test condition. Thus, to prevent fatigue, experiments were limited to a maximum one hour duration, and participants were given 5 minute breaks between each test condition.

Call initiation was not in the focus of the studies, so the conference call was established by the test administrator if needed. After the completion of each test condition, participants were asked to rate overall quality, audio quality, video quality, and AV synchronization using the 5-pt. ACR scale. Even though participants were asked to rate audio quality and synchronization, in-depth insights on types of distortions were not identified. We only focused on the video quality and visual impairments.

In the following subsection, we further present User Studies 1-4, highlighting test methodologies and key results. Study 5 is further discussed in Section V.

TABLE 2. An overview of conducted subjective QoE studies.

User Studies	Participants/ MIN/MAX/AVG age	Research Goal	Number of tests	Topology / Service	End user device	OS / Browser	Manipulated parameters	Findings
US1, 2015, [4]	18 males, 12 females, 29 / 65 / 35	Identify the degree of impact of different smartphone configurations on QoE (with devices differing in CPU, display size, and resolution).	120	P2P / Kurento, Appear.in, Talky, vLine	Samsung S5, Samsung S3, LG G3	Android 4.1.2., 4.4.2. / Chrome 40.0.2214.109	Device capabilities	Tested end user device capabilities have a strong impact on QoE and processing burden should be pushed to a centralized conferencing server.
US2, 2016, [5]	14 males, 13 females, 32 / 65 / 38	Investigate what video resolutions are needed to achieve satisfactory QoE under different bandwidth constraints.	108	SFU / Licode	3 x Samsung S6	Android 6.0.1 / Samsung Internet browser 4.0.10-53	Video resolution, bitrate	While higher video resolutions contribute to better video quality, they also impose higher processing requirements on the system, and lead to congestion under certain bandwidth limitations.
US3, 2017, [6]	16 males, 14 females, 1 fixed user per test group, 33 / 49 / 40	Analyze how GCC handles network packet loss under different video resolutions, bitrates, and frame rate constraints and how packet loss impacts QoE.	120	SFU / Licode	3 x Samsung S6	Android 6.0.1. / Chrome 55.0.2883.91	Video resolution, bitrate, frame rate, packet loss	Performance measurements showed that packet loss caused severe disturbances, and in some cases GCC reduced the video bitrate to nearly zero. The impact of a "lost" video stream on overall QoE was found to differ greatly among participants, which can be attributed to differences in end user expectations.
US4, 2018, [7]	21 males, 6 females, 20 / 29 / 21	Find cause of unexpected disturbances during WebRTC sessions.	72	SFU / Licode	3 x Samsung S6	Android 7.0.0. / Chrome 57.0.2987.132	Video resolution, bitrate, frame rate	Test cases with lower resolution did not overuse CPU, as was the case with test conditions based on higher preset resolution. Lack of send bandwidth occurred more often with test conditions requiring higher bitrates.
US5, 2018	7 males, 20 females, 20 / 25 / 22	Establish a lower threshold for acceptable perceived quality, and investigate the relationship between blurriness / blockiness and QoE	63	SFU / Licode	3 x Samsung S7	Android 7.0.0. / Chrome 63.0.3239.111	Video resolution, bitrate, frame rate	The following video codec parameters are not recommended for a three-party video conference via smartphones: 120x180, 15 fps, 100 kbps; Settings 360x480, 15 fps, 300 kbps represent an upper border case which should be preset for the smartphones with 4GB of RAM in case of a three-party conferences if we want to avoid CPU overuse which participants can detect.

B. USER STUDIES

1) USER STUDY 1

Methodology: Our initial research focused on studying the impact of different smartphone configurations (differing in terms of CPU, display size, and resolution) on QoE [4]. Tests were run using available WebRTC-based conferencing applications available on the market, and involved the following set-ups: (1) all three participants in the group had the same smartphone configuration, (2) each participant in the group had a different smartphone configuration. Tests were conducted in a natural environment on mobile phones over both a Wi-Fi and commercial mobile network.

Video calls were set up using WebRTC applications running on the Internet and commercial network, and using the Kurento Media Server[1] installed in a local network. In certain cases, when one or more participants are located behind a firewall/NAT, additional traversal mechanisms are needed. Such mechanisms include use of ICE (Interactive Connectivity Establishment), STUN (Session Traversal Utilities for NAT), and TURN (Traversal Using Relays around NAT) [40]–[42] servers.

Overall 120 tests were performed. Tests included 30 participants with an average age of 35, while the youngest participant was 29 and the oldest 65 years old.

Results: Results showed the impact of different device factors on user QoE, and imply minimum smartphone requirements for three-party video conferencing as being a 2.5 GHz processor and 2GB RAM. For test purposes, the following three WebRTC based applications were used that were available on the market at the time: Appear.in, Talky and vLine. In addition to the fact that the high-end smartphones available at the time of the study were not so powerful, one of the main reasons for low performance was the fact that video stream quality was not dynamically adapted to device capabilities and network conditions. Further results showed that tested devices that had the same amount of RAM but different resolutions resulted with comparable QoE scores. The reported study provides insights with respect to the high processing capabilities needed by mobile devices to meet the CPU requirements imposed by video conferencing services, and discussed the potential of pushing the processing burden to a centralized conferencing server. Hence, with the following studies, we pushed our experiments to the centralized open source media server **Licode**.[2] Studies 2-5 relied on the use of a central server which helps to reduce the load on the end user devices.

2) USER STUDY 2

Methodology: Our second study was conducted in a controlled lab environment, with all streams transmitted via the Licode media server connected via a local network, and using (at that time modern) 3 GB smartphones.

Licode is a platform based on WebRTC technology and enables a user to create, initialize, and publish a stream when connected to a room. The Licode architecture is based on two components, a client API *Erizo*, responsible for signaling and handling connections to virtual meeting rooms and streams in web applications, and a video conference management

[1] http://www.kurento.org/

[2] http://lynckia.com/licode/

API *Nuve* responsible for room management and user access control. Hence, we were able to set video parameters (same for all clients) using Licode (resolution and bandwidth, with set bandwidth in this case referring to target encoding bitrate).

We were not able to collect more detailed session data, since calls were established using the Samsung Internet browser, which did not offer access to webrtc internal logs. Twenty-seven participants grouped into groups of three participants per call took part in the study, and were instructed to interact with each other in a natural and leisure manner. The average participant age was 38, while the youngest participant was 32 and the oldest 65 years old. All participants used the same device (Samsung Galaxy S6) and were connected via a local network. The test schedule consisted of 12 testing conditions and 108 conducted tests. Testing conditions included all combinations of three different video resolutions (960 × 480, 640 × 480, 480 × 320), and different bitrate constraints (300 kbps, 600 kbps, 1200 kbps, and 50000 kbps), as portrayed in Figure 5.

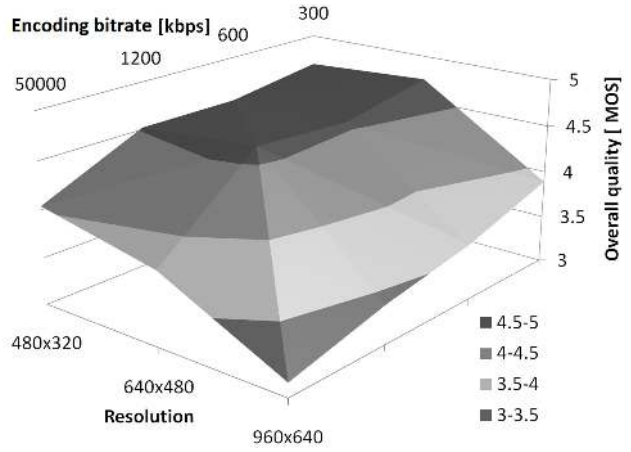

FIGURE 5. Overall quality for each combination of encoding bitrate and resolution settings (User Study 2) [5].

Results: We focused on examining **which video resolutions are needed for achieving a satisfactory QoE**, and how encoding bitrate limitations impact the QoE in the context of three-party mobile video telemeetings. We observed that the highest streamed resolution and bitrate yielded the lowest MOS results for all test cases, as shown in Figure 5. We attribute these findings to insufficient smartphone processing power. Nowadays, streaming at a resolution of 960 × 640 in the context of video calls is generally considered unnecessary. Even a resolution of 640 × 480 will be often reduced due to CPU overuse, and as such may be considered unnecessary for smartphones. With respect to bitrate limitation (referring to the target output video bitrate sent by each participant), 600 kbps is also a rate which will often be reduced by the Google Congestion Control (GCC) algorithm in the context of a three-party mobile telemeeting [7]. The GCC algorithm, implemented in browsers and utilized by WebRTC to provide congestion control for real-time communications over UDP, at the time our tests were conducted implemented both a delay-based controller (on the receiver side) and a loss-based controller, which is run on the sender side in response to feedback from the receiver. According to the packet loss values, the following adaptation decisions are taken [43]:

- if 2-10% of the packets have been lost the sender rate will be kept unchanged.
- if more than 10% of the packets have been lost the rate will be decreased.
- if less than 2% of the packets have been lost, then the rate will be increased.

3) USER STUDY 3

Methodology: With our third study, we shifted from using the Samsung browser to using Chrome, as this provided the opportunity to access the *webrtc-internals* tool implemented within Chrome [6]. *Webrtc-internals* is an internal functionality for collecting statistics about ongoing WebRTC sessions [44]. To obtain statistics, a session has to be opened in the Chrome browser, and while in that session, another tab has to be open with the following URL: chrome://webrtc-internals.

Thirty participants were involved in this study, with an average age of 40, while the youngest participant was 33 and the oldest 49 years old. All participants were connected in a local network with the same device, Samsung S6. The test schedule consisted of 12 testing conditions, leading to a total of 120 conducted tests.

FIGURE 6. Testbed set-up over a LAN connection (User Study 3).

Based on conclusions drawn from our previous studies, we decided to exclude test scenarios with bitrates higher than 600 kbps from further investigation, as such high bitrates are not needed to improve user perceived quality. However, we included various frame rates and packet loss duration into consideration. Packet loss was artificially inserted in the experiments with the hardware-based Albedo Net.Storm[3] impairment generator, as portrayed in Figure 6. We performed

[3]http://www.albedotelecom.com/pages/emulation/src/netstorm.php

tests in which two video resolutions were altered (640 × 480, 480 × 320), under different bitrate constraints (300 kbps and 600 kbps), with one of two frame rates assigned (15 fps and 20 fps). The 8 test conditions are portrayed in Figure 7.

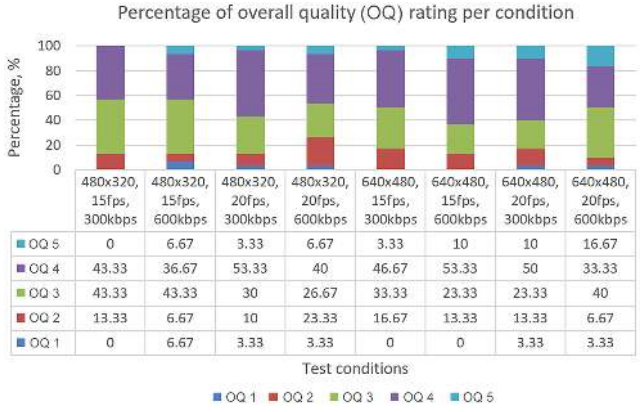

FIGURE 7. Distribution of ratings per test condition for overall quality (User Study 3) [6].

Since packet loss can have significant impact on user perceived quality, we wanted to obtain insights into the performance of multiparty telemeetings under short term, but severe, packet loss. We further wanted to investigate the behaviour of the Google Congestion Control algorithm, implemented in the Google Chrome browser. In response to packet loss measurements, the GCC algorithm (running on the end user devices) adapts resolution, frame rate, and bitrate, according to the packet loss value [45]. As a consequence, actual streamed values start to differ from the preconfigured ones (as configured for each test scenario using the Licode media server).

Results: Ten seconds of inserted bursty packet loss caused 25 to 50 seconds of video conversation with lower quality, after which the service managed to restore values to those that were preconfigured. In some cases, we observed that stream quality was never restored to the initial settings, but continued to stream at the reduced quality level.

In 8% of test scenarios, after the packet loss disturbance was inserted, the video stream from one participant was completely lost until the end of the session, while the audio stream managed to recover. As an interesting finding, even for the leisure context, we found that temporary loss of a video stream did not have a very significant impact on the overall reported user perceived quality, as we expected.

Only in one test case, configured to a resolution of 480 × 320, 15 fps, and 300 kbps, a rating of 1 (or "Bad") was not given for any of the rated variables (audio quality, video quality, overall quality, and AV synchronization). We note that the test condition corresponding to a resolution of 480 × 320, 15 fps, 600 kbps had the highest overall number of bad ratings (when combining all rated variables) Figure 7.

The lowest recorded streaming quality (resulting from activation of the GCC algorithm) corresponded to a resolution of 240 × 160, with frame rate 1 fps, and bitrate 15 kbps. In some cases, bitrates with values around 30 kbps lasted for approximately 30 seconds, which may likely be considered too long in the case of a 3-minute long conversation (as we used for testing). Therefore, our goal for User Study 4 was to determine the lowest acceptable video quality, i.e., the minimum streaming video configuration parameters needed to maintain acceptable user perceived service quality.

4) USER STUDY 4

Methodology: Twenty-seven participants took part in the fourth study, with an average age of 21, while the youngest participant was 20 and the oldest 29 years old. The test schedule consisted of eight testing conditions, with 72 tests performed.

In this study, we used the same setup as in User Study 3 (test conditions are portrayed in Figure 8), with one exception: the Net.Storm impairment generator was excluded from the setup. All participants used the same device, Samsung S6, with the same encoding parameter settings as in User Study 3, so as to compare ratings with and without packet loss. However, instead of the stable and controlled conditions, stream quality adaptation occurred in every test scenario. We thus wanted to further explore what caused quality degradation [7].

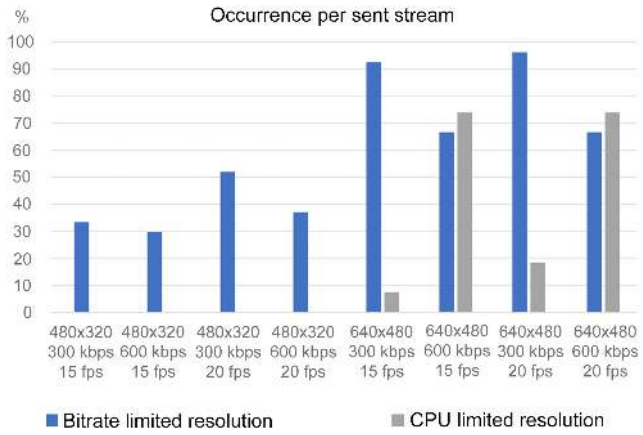

FIGURE 8. Occurrence of resolution degradation due to observed bitrate or CPU limitation per sent stream. Encoding bitrate and CPU limitations were determined based on collected webrtc-internals data (User Study 4).

Results: The primary cause of degradation depended on video resolution and/or video bitrate. Resolution lowered due to CPU overload occurred only in test cases with a predefined resolution of 640 × 480, in particular for test cases with a predefined encoding bitrate of 600 kbps. On the other hand, we observed resolution degradations in all cases where the target encoding bitrate was limited using the Licode settings (Figure 8). Reduction of the video resolution 640 × 480 lasted significantly longer as compared to a resolution of 480 × 320. All test cases where the resolution was set to 640 × 480 managed to hold this preset resolution for less than 20% of the session time, while adaptation took place at two levels: 480 × 360 and 320 × 240. Test cases with a set resolution of

480 × 320 managed to maintain the preset resolution for more than 65% of the session duration, with resolution adapted only once to 360 × 240.

By bitrate and CPU limitation, we refer to values taken from the *webRTC-internals* dump. Due to the *bitrate limitation*, the default resolution was degraded with the most frequent occurrence in the test corresponding to a resolution of 640 × 480, 300 kbps, 20 fps, where 96.29% of streams were adapted. Decreased resolution due to *CPU limitation* appeared most often in the test case corresponding to a resolution of 640 × 480, 600 kbps, 20 fps, where 74.07% of streams were adapted.

These results provided us with insights on resource requirements, and how to exploit these insights to retain acceptable QoE. Our measurements showed that higher video resolutions and bitrate contribute to better video quality in a three-party call only up to a certain point. When the conference is held using a 3GB RAM smartphone, a resolution of 640 × 480 has a strong impact on CPU utilization, especially in the case of higher video bitrates, and as such should be avoided.

Following this series of user studies used to collect a large number of subjective ratings under various conditions, and obtain insights into both session and stream quality, what remained was to investigate the potential of utilizing objective video metrics to infer subjectively perceived quality. In our next study, we therefore included the analysis of screen recordings and the relationship between subjective ratings and objective video quality impairments.

V. USER STUDY FOCUSED ON RELATIONSHIP BETWEEN SUBJECTIVE VQ RATINGS AND OBJECTIVE VIDEO QUALITY IMPAIRMENTS

User Study 5: The goal of this study was twofold: (1) to detect how participants will respond to lower video quality, and attempt to establish a lower threshold for acceptable perceived quality, and (2) investigate how blurriness and blockiness impact QoE and to what extent objective metric values correlate with subjective scores.

A. METHODOLOGY

Measurements involving interactive three-party audiovisual conversations carried out in a leisure context were conducted in a controlled laboratory environment (one participant per site) over a Wi-Fi network, and with symmetric device conditions. In the experiments, video resolution, bitrate, and frame rate were predefined using settings on the Licode server. Licode was installed in a local network on a computer with Intel Core i5 Processor, 2.6 GHz, 8 GB RAM and Ubuntu 14.04 LTS (Figure 9). Participants took part in the call using Samsung Galaxy S7 smartphones with 4 GB of RAM. During each call, the smartphone screen was recorded using the DU recorder application.[4] To monitor video quality

FIGURE 9. Testbed set-up over a LAN connection (User Study 5).

and service performance, WebRTC session-related data was collected via *webrtc-internals*.

The test schedule consisted of 7 testing conditions, with videos encoded with the VP8 video codec, and resolutions, bitrate, and frame rate set according to Table 3. Each test condition was evaluated by 9 groups, leading to a total of 63 performed tests[5]).

TABLE 3. Test conditions used in User Study 5.

Test conditions	Video resolution	Frame rate	bitrate
Test case 1 (TC1)	180x240	15 fps	200 kbps
Test case 2 (TC2)	360x480	15 fps	300 kbps
Test case 3 (TC3)	240x360	15 fps	150 kbps
Test case 4 (TC4)	120x180	15 fps	100 kbps
Test case 5 (TC5)	240x360	15 fps	200 kbps
Test case 6 (TC5)	120x180	20 fps	200 kbps
Test case 7 (TC7)	240x360	20 fps	300 kbps

The setup was symmetrical for all participants within each group. Established video telemeetings lasted for two minutes per test session and were initiated through a WebRTC application within the Google Chrome 63.0.3239.111 browser.

B. PARTICIPANTS

Twenty-seven participants (20 female and 7 male) took part in the study on a voluntary basis, with an average age of 22 years (min age 20, max. age 23). Participants were divided into nine groups, formed based on acquaintances. All participants were students, non-experts in the AV field, and had previous experience with applications such as Skype, Viber, and WhatsApp.

C. RESULTS

WebRTC Internals Data and MOS Values: To check the actual *sent* and *received* video qualities, and to be sure that participants were in fact rating the preset quality levels

[4]https://du-recorder.en.uptodown.com/android

[5]We discarded the data from one group due to erroneous measurements or incomplete responses.

TABLE 4. MOS ratings and WebRTC internals statistics of mean values per test condition.

Test case	TC1	TC2	TC3	TC4	TC5	TC6	TC7
Obtained default resolution	100%	86.28 %	100%	100%	100%	100%	100%
Obtained frame rate and +/-1	96.22%	93.84 %	94.99%	97.06%	96.72%	91.67%	88.12%
AVG frame rate	14.91	14.82	14.85	14.91	14.92	19.78	19.53
MOS Audio quality	3.67	3.83	3.21	3.17	3.67	3.46	3.50
MOS Video quality	3.50	3.75	3.17	2.33	3.67	3.13	3.58
MOS AV synchronization	3.46	3.63	3.17	2.83	3.63	3.29	3.50
MOS Overall quality	3.63	3.75	3.17	2.83	3.63	3.38	3.50

(as opposed to some dynamically adapted levels) we analyzed *webRTC-internals* data. We observed that resolution adaptation occurred only in TC2 within 6 video streams due to CPU overuse (Table 4). In those cases, resolution was decreased to 270 × 360, and lasted at this level for an average of 50.45% of the session time. Within all test cases, packet loss was very low (around 0.001%). Only in TC7, within one group, packet loss yielded 0.96%. Hence, adaptation quality based on the packet loss caused by the GCC algorithm was not triggered in any test case.

If we want to avoid CPU overuse which participants can detect, we conclude that video settings used in TC2 may be preset as an upper bound in terms of resolution, frame rate, and bitrate, when used in the context of three-party conference calls established using smartphones with processing capabilities comparable to those tested (4GB of RAM). On the other hand, while participants provided the highest average quality ratings for TC2, we see that only a slight decrease in average ratings is observed in the case of TC5, albeit TC5 involved resolution set to 240 × 360, the same frame rate, and 200 kbps bitrate (rather than 300 kbps as used in TC2). It is thus worth considering whether the significant increase in resources (from TC5 to TC2) is worth the only slight gain in perceived quality.

We further conclude that the test case with the lowest video quality (TC4: 120 × 180 resolution, 15 fps, 100 kbps) is not a recommendable settings for a three-party video conference, with subjective ratings giving an average of 3.17 for audio quality, 2.33 for video quality, and 2.83 for both synchronization and overall quality. We observed that the cause of such low ratings is not actually the resolution, but rather insufficient bitrate. TC6, which had the same resolution, but a slightly higher frame rate (20 fps) and higher available bitrate (200 kbps), resulted with a video MOS of 3.13 and overall MOS 3.38.

Objective Video Quality Metrics (Blurriness and Blockiness): Digital video systems can add edges (e.g., blocking) or reduce edges (e.g., blurring). Blocking distortion can be introduced by coding and/or transmission errors (when the video encoder is not able to process the whole stream) [46]. Video blurriness can occur during high movement video capture or when the amount of available network bandwidth is not sufficient to transmit the video stream. To analyze objective video quality, we used the MSU Video Quality Measurement Tool (VQMT) Professional Version 10.2.[6]

[6]https://www.compression.ru/

Participants were asked to report whether or not they experienced blurriness and blockiness, and whether or not they noticed any video freezes. While 66.67% participants responded that they noticed blurriness in the test case with the lowest video quality and lowest ratings (TC4), in the objectively highest video quality test case (TC2), blurriness was observed by 50% of all participants. The least number of participants reported having noticed blurriness in TC5 (240 × 360, 15 fps, 200 kbps) with a share of 45.83% (Table 5). Participants reported blockiness in test cases where insufficient bitrate was preset. Blockiness was reported in TC1 only by 8.33% participants, while in TC4 and TC7 by 37.5%.

TABLE 5. Percentage of participants reporting disturbances.

Test case	Blurriness	Blockiness	Freezes
TC1	62.50%	8.33%	4.17%
TC2	50.00%	4.17%	0.00%
TC3	50.00%	29.17%	45.83%
TC4	66.67%	37.50%	25.00%
TC5	45.83%	16.67%	8.33%
TC6	62.50%	25.00%	8.33%
TC7	41.67%	37.50%	29.17%

Based on our results, it turns out that short video freezes did not have a significant impact on reported perceived quality. In fact, only in six sessions (out of a total of 58 sessions), two participants reported having noticed a video freeze. In all other sessions where video was reported as being frozen, this was noticed by only one participant from the session. TC2 is the only scenario where participants did not report any freezes. In the other cases, 4.17-29.17% of participants reported freezes.

Blurriness and Blockiness Per Test Case: Based on results obtained during the video conferences, we wanted to further investigate the relationship between objective no-reference video metrics, namely blurriness and blockiness, and subjective user ratings, whereby better objective video quality is achieved by higher measured values of blockiness and blurriness. A summary of results is given in Table 6 and Figure 10.

If we compare TC1 (180 × 240, 15 fps, 200 kbps) and TC7 (240 × 360, 20 fps, 300 kbps), we observe that MOS was higher in TC1 than TC7, for all rated quality dimensions except for video quality (which was only slightly lower). In terms of determining video codec configuration parameters, it may thus be possible to save 100 kbps, avoid possible

TABLE 6. Mean values of video impairments and rated video quality.

Test case	Blurriness median/mean/StDev	Blockiness median/mean/StDev	MOS Video quality
TC1	6.10 / 6.14 / 0.34	36.90 / 37.61 / 5.14	3.5
TC2	6.29 / 6.38 / 0.45	39.41 / 39.98 / 4.84	3.75
TC3	6.72 / 6.68 / 0.61	38.35 / 38.98 / 6.21	3.17
TC4	6.40 / 6.45 / 0.53	36.27 / 36.58 / 4.53	2.33
TC5	6.61 / 6.60 / 0.68	38.44 / 39.04 / 5.18	3.67
TC6	6.29 / 6.32 / 0.43	35.75 / 36.16 / 3.83	3.13
TC7	6.52 / 6.54 / 0.88	38.48 / 39.63 / 7.9	3.58

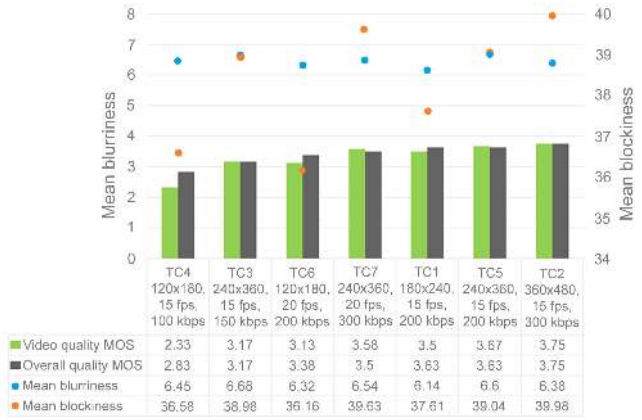

FIGURE 10. Mean values of blurriness and blockiness with associated video and overall quality MOS scores per each test condition.

CPU overuse, and still obtain a higher average score for overall quality. TC1 was preset with a lower resolution then TC7, which participants noticed, but did not have a significant impact when rating other aspects.

Distributions of Blurriness and Blockiness Values for Different Subjective Video Quality Ratings: With the summary statistics, a wide range of values overlapped across different user ratings. Thus, to gain better insights and to visualise data and performance indicators, we used histograms to measure how frequently values appear in our data sets. The histograms for user video quality (VQ) ratings of 1 and 5 have a notably different spread and correlated frequency of values compared to VQ 3 or 4, since video quality was rated as ''Bad'' in only 2.64% cases, and as ''Excellent'' in only 7.93% cases.

The following histograms show the blockiness and blurriness values from all test scenarios associated with corresponding video quality ratings (Figure 11 and Figure 12). We split data into 20 bins for blockiness values and 10 bins for blurriness values. We chose a different number of bins in order to show underlying patterns and data trend. Each bin contains the frequency of occurrences of values in the data set that are contained within that bin. On the graphs, we can observe shifted distributions to the right per higher VQ rating for both blockiness (Figure 11) and blurriness (Figure 12), which correlates to better quality.

Comparing blockiness and blurriness graphs, blurriness values are more inconsistent and spread due to the camera movement and participants moving around, which impacted

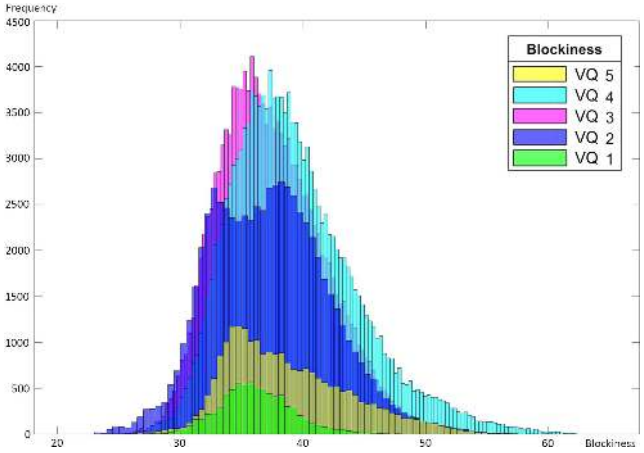

FIGURE 11. Frequency of blockiness values per frame for video quality (VQ) user ratings.

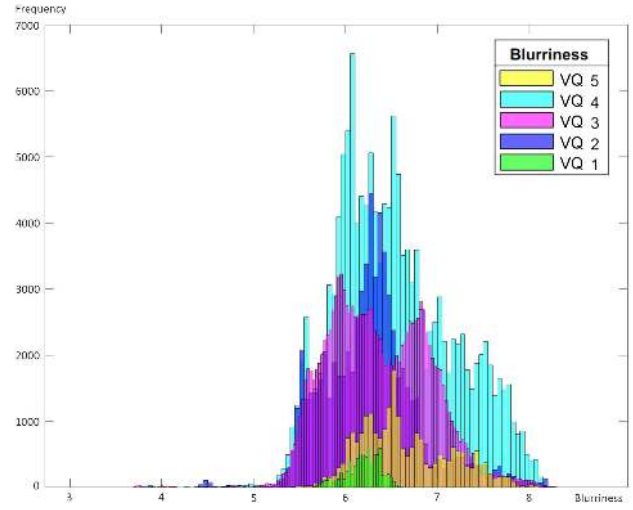

FIGURE 12. Frequency of blurriness values per frame for video quality (VQ) user ratings.

the blurriness. Thus, to better describe sample data we fitted blockiness and blurriness values to common distributions using MATLAB R2018b.[7]

We evaluated (based on log likelihood values and probability plots) that the best fit for blockiness is the Birnbaum-Saunders distribution (Table 7). Birnbaum-Saunders distribution is defined with the beta (scale) parameter and gamma (shape) parameter. Since video quality was most often rated with ''Fair'' or ''Good'', for those two ratings we have the largest value set, and consequently the largest value span. Therefore, fitted distributions with respective probability plots have the longest tales (Figure 13).

Mean values of fitted data samples are in ascending order from video quality user rating VQ 1 to VQ 3, while VQ 5 value is placed between VQ 3 and VQ 4. One of the possible reasons could be due to the significantly smaller

[7]https://www.mathworks.com/

TABLE 7. Measured blockiness values per frame per video quality user ratings with fitted Birnbaum-Saunders distribution.

Blockiness per VQ user rating	1	2	3	4	5
Mean	36.1858	37.282	37.488	39.254	38.744
Variance	8.910	23.663	18.8266	26.151	24.651
Parameter beta estimate	36.063	36.968	37.240	38.925	38.430
Parameter beta Std. Err.	0.03389	0.01840	0.01165	0.01329	0.03202
Parameter gamma estimate	0.08242	0.13021	0.11555	0.13001	0.12789
Parameter gamma Std. Err.	0.00066	0.00035	0.00022	0.00024	0.00059

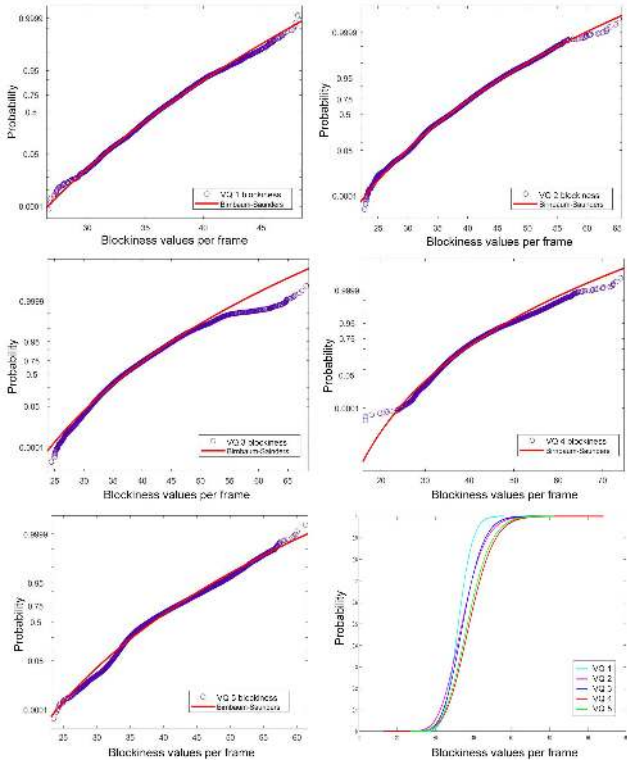

FIGURE 13. Probability plots for Birnbaum-Saunders distribution for blockiness values per frame per video quality user rating.

number of sample data inputs. The highest yielded blockiness measured value for VQ 1 was 48.15, VQ 2: 65.35, VQ 3: 68.15, VQ 4: 74.11, while for VQ 5 it was 61.28. While we can observe a positive trend, we do not observe high consistency, partly because of a large difference between sample set sizes. This trend could also be due to the peak level of annoyance experienced by users at a certain blockiness level, which could later settle to a slightly better QoE beyond this blockiness level owing to saturation effects related to user QoE.

For the blurriness values (measured per frame) values from sessions where video quality was rated with ''Bad'' and ''Poor'' were fitted to a Burr distribution, while values corresponding to sessions rated as ''Fair'', ''Good'', or ''Excellent'' were fitted to a Gamma distribution. A Burr distribution is defined with three parameters: alpha-scale parameter, c-first shape parameter, and k-second shape parameter. Gamma distributions is defined with a-shape parameter and b-scale parameter.

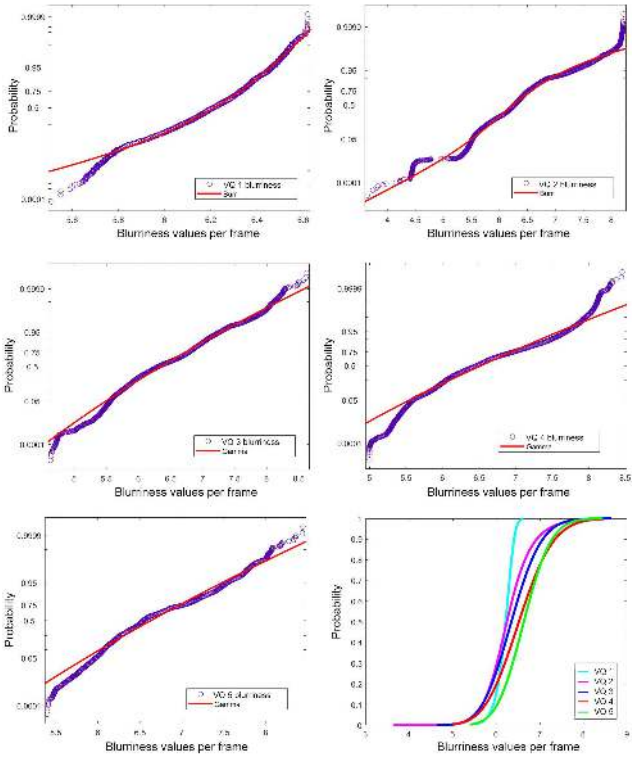

FIGURE 14. Probability plots for Burr and Gamma distributions for blurriness values per frame per video quality user rating.

Figure 14 shows probability plots for blurriness values rated with VQ 1 to 5, collected during sessions across seven different test cases. Results are summarized in Table 8. Mean values of fitted data samples for blurriness ascend in order from VQ 1: 6.18 to VQ 5: 6.64. The highest yielded blurriness measured value for VQ 1 was 6.62, VQ 2: 8.22, VQ 3: 8.46, VQ 4: 8.63, and VQ 5: 8.44.

The blurriness probability plot with fitted distributions shows some shift to the right, where better quality values correspond to higher QoE. However, to obtain more precise results, further testing should be done, with an adapted methodology to achieve more stable session performance.

Due to the similar and overlapping blurriness and blockiness values, it is difficult to correlate specific levels of blurriness and blockiness with user ratings. However, participants did notice the changes in objective video quality and rated them accordingly (Figure 15). In test cases with ''tighter'' bitrate (enough for lower motion) for a chosen resolution, video quality scores correlated better with overall quality than audio quality scores. In test cases with assigned higher bitrates (TC1, TC6, TC7), audio quality scores correlated better with overall quality scores.

TABLE 8. Distribution of blurriness values per frame and per video quality rating level, with fitted Burr distribution for VQ 1 and VQ 2 user ratings, and Gamma distribution for VQ 3, VQ 4, and VQ 5 user ratings.

Blurriness per VQ user rating	1	2	3	4	5
Mean	6.18757	6.28235	6.36187	6.53864	6.6438
Variance	0.03831	0.2781	0.29597	0.38999	0.24336
Parameter alpha estimate	6.6488	6.15611	-	-	-
Parameter alpha Std. Err.	0.04859	0.00485	-	-	-
Parameter c estimate	40.9839	24.1488	-	-	-
Parameter c Std. Err.	0.62142	0.13621	-	-	-
Parameter k estimate	11.4123	0.79168	-	-	-
Parameter k Std. Err.	2.88852	0.00941	-	-	-
Parameter a estimate	-	-	136.744	109.626	181.378
Parameter a Std. Err.	-	-	0.52379	0.40842	1.67562
Parameter b estimate	-	-	0.04652	0.05964	0.03662
Parameter b Std. Err.	-	-	0.00017	0.00022	0.00033

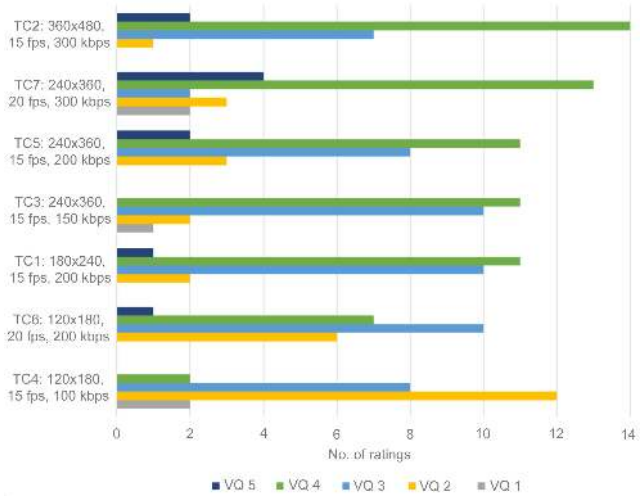

FIGURE 15. Number of occurrences of participant VQ ratings for each level of the used 5 pt. rating scale (User Study 5).

D. STUDY OBSERVATIONS

Over the course of our research studies, we noticed several issues which need to be considered. First of all, we noticed that during the course of conducted test sessions, some participants became more restless (even though the whole evaluation process lasted for a maximum of 45 minutes), thus causing additional movement and potential impact on both perceived quality and objective metrics. An increase in movement should be taken into consideration when defining target bitrates, as more dynamic scenes will likely require higher bitrates to achieve satisfactory QoE.

When testing on small screen sizes, especially in multiparty conference calls where the preview window of each participant is relatively small, even if the objective video parameters are preset to significantly different values, quality assessment on a 5-pt. scale can produce similar results since it can be difficult to distinguish small perceived differences and relate them to the five ratings at the end of the session.

We further noticed that sometimes participants were not able to distinguish impairments, for instance reporting blockiness in cases when it was fairly low, instead of a blurriness which was higher than average. This may possibly be attributed to the small preview size and short term of disturbances. Additionally, we noticed that participants engaged in the conversation can miss to detect short video freezes if: 1) the participant is not an active user, 2) audio quality is unimpaired, or 3) when the participant is staying still during the session.

Usually during a conversation, focus is on the active speaker. Hence, in a multiparty setup, the center of an eye gaze is commonly on the talking participant, while other participants in the group are outside the point of fixation. During our studies, all conducted in a leisure context, we noticed that occasional video impairments did not significantly impact overall perceived quality (however, we note that participants were only engaged in conversation, and were not focused on presenting to each other any particular visual cues). Thus, a key issue is to ensure enough resources to the active participant, prioritizing audio quality over video quality.

Obtained results serve as input for specifying a QoE-driven video encoding adaptation strategy, to be triggered in light of system and/or network resource limitations. A key factor to consider should be the number of participants, since even if enough network (bandwidth) resources are available, device processing capabilities can present a major bottleneck. Even though we did not explicitly focus on the impact of different numbers of participants, in a mobile multi-user environment, only one added stream can make a great difference. The target is to find the optimal resolution to bitrate ratio, which depends on processing capabilities of each display device, movement of the camera or participant, and the speed of the network connection. Our studies have shown that it is better to provide constant lower objective video quality than to switch back and forth between higher and lower qualities (due to CPU overuse or lack of bitrate).

Nowadays, and especially in light of the ongoing Covid-19 pandemic, we are witnessing a drastic increase in the use of videoconferencing tools, for purposes such as e-learning, meetings, and social gatherings [47]. Among various popular tools which have seen a high increase in customer use, such as Zoom and Microsoft Teams, included are also numerous applications based on WebRTC technology (e.g., Google Hangouts Meet, BlueJeans, Lifesize, Slack) [48]. While in the scope of the studies conducted in this paper we have focused on a lab setup using WebRTC, we note that the findings are applicable in a wider context, i.e., across any type of mobile multiparty videoconferencing technology.

VI. CONCLUSIONS AND OUTLOOK

Managing multiparty audiovisual telemeeting services so as to optimize end user QoE requires an understanding of the relationship between QoE and underlying influence factors, as well as an understanding of the potential interactions between IFs. We have thus given an overview of various IFs that should be considered. Given the wide range of system, context, and human factors, we have narrowed our study scope to focus on three-party video calls established via mobile devices. A key challenge faced by multiparty mobile telemeeting providers lies in configuring the video and audio encoding parameters so as to maximize participant perceived quality while meeting resource availability constraints. Our studies have focused primarily on the impact of video encoding parameters, device capabilities, and network impairments on overall QoE.

We bring together the results of 5 user studies conducted over the course of 4 years, with a total of 141 participants having taken part in the studies. We have summarized key findings in terms of acceptable video codec configuration parameters, and have provided insights with respect to the impacts of various network disturbances. Moreover, we have explored the relationship between subjective ratings and distributions of objective metrics, namely blockiness and blurriness.

In our ongoing work, we will conduct more extensive studies to further explore to what extent QoE-related user ratings may be inferred from objective metrics, including both audio and video. Moreover, we are working towards specifying a QoE-aware adaptation strategy that can be utilized by service providers to dynamically adapt both video and audio codec settings to dynamic network conditions end user device capabilities, once again focusing on a mobile context.

REFERENCES

[1] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T Rec. P.910, 2008.

[2] Y. Fang, W. Lin, and S. Winkler, "Review of existing objective QoE methodologies," in *Multimedia Quality of Experience (QoE): Current Status and Future Requirements*, vol. 29. London, U.K.: Wiley, 2015.

[3] M. Schmitt, D. C. A. Bulterman, and P. S. Cesar, "The contrast effect: QoE of mixed video-qualities at the same time," *Qual. User Exper.*, vol. 3, no. 1, p. 7, Dec. 2018.

[4] D. Vucic and L. Skorin-Kapov, "The impact of mobile device factors on QoE for multi-party video conferencing via WebRTC," in *Proc. 13th Int. Conf. Telecommun. (ConTEL)*, Jul. 2015, pp. 1–8.

[5] D. Vučić, L. Skorin-Kapov, and M. Sužnjević, "The impact of bandwidth limitations and video resolution size on QoE for WebRTC-based mobile multi-party video conferencing," *Screen*, vol. 18, p. 19, Aug. 2016.

[6] D. Vucic and L. Skorin-Kapov, "The impact of packet loss and Google congestion control on QoE for WebRTC-based mobile multiparty audiovisual telemeetings," in *MultiMedia Modeling* (Lecture Notes in Computer Science), vol. 11295, I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W. H. Cheng, and S. Vrochidis, Eds. Cham, Switzerland: Springer, 2019, pp. 459–470.

[7] D. Vucic and L. Skorin-Kapov, "QoE evaluation of WebRTC-based mobile multiparty video calls in light of different video codec settings," in *Proc. 15th Int. Conf. Telecommun. (ConTEL)*, Jul. 2019, pp. 1–8.

[8] W3C. (2019). *WebRTC 1.0: Real-Time Communication Between Browsers*. [Online]. Available: https://www.w3.org/TR/webrtc/

[9] *Subjective Quality Evaluation of Audio and Audiovisual Multiparty Telemeetings*, document ITU-T Rec. P.1301, 2017.

[10] J. Skowronek, "Quality of experience of multiparty conferencing and telemeeting systems," Ph.D. dissertation, Telekom Innov. Lab., Technical Univ. Berlin, Berlin, Germany, 2017.

[11] S. Junuzovic, K. Inkpen, R. Hegde, and Z. Zhang, "Towards ideal window layouts for multi-party, gaze-aware desktop videoconferencing," in *Proc. Graph. Interface*. London, U.K.: Canadian Human-Computer Communications Society, 2011, pp. 119–126.

[12] K. Wac, S. Ickin, J.-H. Hong, L. Janowski, M. Fiedler, and A. K. Dey, "Studying the experience of mobile applications used in different contexts of daily life," in *Proc. 1st ACM SIGCOMM workshop Meas. up Stack (W-MUST)*, 2011, pp. 7–12.

[13] *Different WebRTC Architectures*. Accessed: Jun. 8, 2020. [Online]. Available: https://www.callstats.io/blog/webrtc-architectures-explained-in-5-minutes-or-less

[14] S. Loreto and S. P. Romano, *Real-Time Communication With WebRTC: Peer-to-Peer in the Browser*. Newton, MA, USA: O'Reilly Media, 2014.

[15] B. Leiba and M. Kucherawy. *RtcWeb Status Pages*. Accessed: Jun. 8, 2020. [Online]. Available: https://tools.ietf.org/wg/rtcweb/charters

[16] Uberti. *WebRTC Forward Error Correction Requirements*. Accessed: Jun. 8, 2020. [Online]. Available: https://tools.ietf.org/html/draft-ietf-rtcweb-fec-10

[17] Y. Lu, Y. Zhao, F. Kuipers, and P. Van Mieghem, "Measurement study of multi-party video conferencing," in *Proc. Int. Conf. Res. Netw. (NETWORKING)*, in Lecture Notes in Computer Science, vol. 6091, M. Crovella, L. M. Feeney, D. Rubenstein, and S. V. Raghavan, Eds. Berlin, Germany: Springer, 2010, pp. 96–108.

[18] Y. Xu, C. Yu, J. Li, and Y. Liu, "Video telephony for end-consumers: Measurement study of Google+, iChat, and Skype," in *Proc. ACM Conf. Internet Meas. Conf. (IMC)*, 2012, pp. 371–384.

[19] M. Moulay and V. Mancuso, "Experimental performance evaluation of WebRTC video services over mobile networks," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 541–546.

[20] *Multimedia Quality of Service and Performance—Generic and User-Related Aspects*, document ITU-T Rec. G.1011, 2015.

[21] *Methods for Objective and Subjective Assessment of Speech and Video Quality*, document ITU-T Rec. P.800.2, 2016.

[22] B. Weiss, D. Guse, S. Möller, A. Raake, A. Borowiak, and U. Reiter, "Temporal development of quality of experience," in *Quality of Experience* (Advanced Concepts, Applications and Methods), S. Möller and A. Raake, Eds. Berlin, Germany: Springer, 2014, ch. 10.

[23] P. Le Callet, S. Möller, and A. Perkins, "Qualinet white paper on definitions of quality of experience (2012) version 1.2," in *Proc. Eur. Netw. Qual. Express Multimedia Syst. Services (COST Action IC)*, 2013, pp. 1–23.

[24] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, "Factors influencing quality of experience," in *Quality of Experience* (T-Labs Series in Telecommunication Services), S. Möller and A. Raake, Eds. Cham, Switzerland: Springer, 2014, pp. 55–72.

[25] C. Yu, Y. Xu, B. Liu, and Y. Liu, "'Can you SEE me now?' A measurement study of mobile video calls," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun.*, Apr./May 2014, pp. 1456–1464.

[26] K. De Moor, S. Arndt, D. Ammar, J.-N. Voigt-Antons, A. Perkis, and P. E. Heegaard, "Exploring diverse measures for evaluating QoE in the context of WebRTC," in *Proc. 9th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2017, pp. 1–3.

[27] J. B. Husic, S. Barakovic, and A. Veispahic, "What factors influence the quality of experience for WebRTC video calls?" in *Proc. 40th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2017, pp. 428–433.

[28] B. García, M. Gallego, F. Gortázar, and A. Bertolino, "Understanding and estimating quality of experience in WebRTC applications," *Computing*, vol. 101, no. 11, pp. 1585–1607, Nov. 2019.

[29] J. Skowronek, K. Schönenberg, and G. Berndtsson, "Multimedia conferencing and telemeetings," in *Quality of Experience* (Advanced Concepts, Applications and Methods), S. Moller and A. Raake, Eds. Berlin, Germany: Springer, 2014, ch. 15.

[30] G. Berndtsson, M. Folkesson, and V. Kulyk, "Subjective quality assessment of video conferences and telemeetings," in *Proc. 19th Int. Packet Video Workshop (PV)*, May 2012, pp. 25–30.

[31] K. Zeng, T. Zhao, A. Rehman, and Z. Wang, "Characterizing perceptual artifacts in compressed video streams," *Proc. SPIE*, vol. 9014, Feb. 2014, Art. no. 90140Q.

[32] S. Jana, A. Chan, A. Pande, and P. Mohapatra, "QoE prediction model for mobile video telephony," *Multimedia Tools Appl.*, vol. 75, no. 13, pp. 7957–7980, Jul. 2016.
[33] A. F. D. Silva and C. Q. Mylène, "Perceptual strengths of video impairments that combine blockiness, blurriness, and packet-loss artifacts," *Electron. Imag.*, vol. 2018, no. 12, pp. 1–234, 2018.
[34] D. Ammar, K. D. Moor, L. Skorin-Kapov, M. Fiedler, and P. E. Heegaard, "Exploring the usefulness of machine learning in the context of WebRTC performance estimation," in *Proc. IEEE 44th Conf. Local Comput. Netw. (LCN)*, Oct. 2019, pp. 406–413.
[35] M. Schmitt, J. Redi, D. Bulterman, and P. S. Cesar, "Towards individual QoE for multiparty videoconferencing," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1781–1795, Jul. 2018.
[36] J. B. Husic, E. Alagic, S. Barakovic, and M. Mrkaja, "The influence of task complexity and duration when testing QoE in WebRTC," in *Proc. 18th Int. Symp. INFOTEH-JAHORINA (INFOTEH)*, Mar. 2019, pp. 1–6.
[37] H. Alvestrand, *Overview: Real Time Protocols for Browser-Based Applications*, document draft-ietf-rtcweb-overview-19, 2017.
[38] J. Bankoski, P. Wilkins, and Y. Xu, *Vp8 Data Format and Decoding Guide*, document RFC 6386, 2011.
[39] *Interactive Test Methods for Audiovisual Communications*, document ITU-T Rec. P.920, 2000.
[40] *Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal*, document Request for Comments 8445, 2018.
[41] *Datagram Transport Layer Security (DTLS) as Transport for Session Traversal Utilities for NAT (STUN)*, document Request for Comments 7350, 2014.
[42] *Traversal Using Relays Around NAT (TURN): Relay Extensions to Session Traversal Utilities for NAT (STUN)*, document Request for Comments 5766, 2010.
[43] B. Jansen, T. Goodwin, V. Gupta, F. Kuipers, and G. Zussman, "Performance evaluation of WebRTC-based video conferencing," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 45, no. 3, pp. 56–68, Mar. 2018.
[44] D. Ammar, K. De Moor, M. Xie, M. Fiedler, and P. Heegaard, "Video QoE killer and performance statistics in WebRTC-based video communication," in *Proc. IEEE 6th Int. Conf. Commun. Electron. (ICCE)*, Jul. 2016, pp. 429–436.
[45] G. Carlucci, L. De Cicco, S. Holmer, and S. Mascolo, "Congestion control for Web real-time communication," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 2629–2642, Oct. 2017.
[46] *Objective Perceptual Multimedia Video Quality Measurement of HDTV for Digital Cable Television in the Presence of a Full Reference*, document ITT Rec. J.341, 2016.
[47] T. Favale, F. Soro, M. Trevisan, I. Drago, and M. Mellia, "Campus traffic and e-Learning during COVID-19 pandemic," *Comput. Netw.*, vol. 176, Jul. 2020, Art. no. 107290.
[48] R. Jain. (2020). *COVID-19's Long-Term Impact on Remote Work and Learning*. [Online]. Available: https://www.nojitter.com/technology-trends/covid-19s-long-term-impact-remote-work-and-learning

DUNJA VU I (Member, IEEE) is currently pursuing the Ph.D. degree with the Faculty of Electrical Engineering and Computing, University of Zagreb, in the domain of quality of experience modeling and management for multiparty audiovisual calls. She worked in various companies as a Software Developer, with a focus on mobile networks and applications. She also worked as a Sales and Technical Support Representative for measuring systems and equipment. She is currently with Ericsson Nikola Tesla d.d., Zagreb, Croatia, and working as a System Developer for 4G/5G telecom networks and products.

LEA SKORIN-KAPOV (Senior Member, IEEE) is an Associate Professor with the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, and the Head of the Multimedia Quality of Experience Research Laboratory (MUEXlab). She has published over 100 scientific articles. Her research interests include quality of experience (QoE) modeling of multimedia applications, QoE monitoring of encrypted video traffic, cross-layer negotiation and management of QoS/QoE, and resource allocation and optimization mechanisms. She was a Senior Research Engineer and the Project Manager of the Research and Development Center, Ericsson Nikola Tesla d.d., Croatia. She serves on the Editorial Boards of the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT and *Multimedia Systems* journal (Springer), and has served as a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and *ACM Transactions on Multimedia Computing, Communications, and Applications*.