Revisiting Videoconferencing QoE: Impact of Network Delay and Resolution as Factors for Social Cue Perceptibility

Conference Paper · June 2023

DOI: 10.1109/QoMEX58391.2023.10178483

Chenyao Diao

7 PUBLICATIONS 67 CITATIONS

SEE PROFILE

Rakesh Rao Ramachandra Rao
Technische Universität Ilmenau

45 PUBLICATIONS 526 CITATIONS

SEE PROFILE

Alexander Raake
Technische Universität Ilmenau

396 PUBLICATIONS 5,932 CITATIONS

SEE PROFILE

Revisiting Videoconferencing QoE: Impact of Network Delay and Resolution as Factors for Social Cue Perceptibility

Chenyao Diao, Luljeta Sinani, Rakesh Rao Ramachandra Rao, Alexander Raake
Audiovisual Technology Group, Technische Universität Ilmenau, Germany
chenyao.diao, luljeta.sinani, rakesh-rao.ramachandra-rao, alexander.raake@tu-ilmenau.de

2023 15th International Conference on Quality of Multimedia Experience (QoMEX) | 979-8-3503-1173-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/QOMEX58391.2023.10178483

***Abstract*—Previous research from well before the Covid-19 pandemic had indicated little effect of delay on integral quality but a measurable one on user behavior, and a significant effect of resolution on quality but not on behavior in a two-party communication scenario. In this paper, we re-investigate the topic, after the times of the Covid-19 pandemic and its frequent and widespread videoconferencing usage. To this aim, we conducted a subjective test involving 23 pairs of participants, employing the Celebrity Name Guessing task. The focus was on impairments that may affect social (resolution) and communication cues (delay). Subjective data in the form of overall conversational quality and task performance satisfaction as well as objective data in the form of task correctness, user motion, and facial expressions were collected in the test. The analysis of the subjective data indicates that perceived conversational quality and performance satisfaction were mainly affected by video resolution, while delay (up to 1000 ms) had no significant impact. Furthermore, the analysis of the objective data shows that there is no impact of resolution and delay on user performance and behavior, in contrast to earlier findings.**

***Index Terms*—quality of experience, videoconferencing, WebRTC, conversational quality**

I. Introduction

In recent years, applications and services enabling videoconferencing (VC) have been widely used for different purposes such as professional meetings, e-learning, and personal socializing. The increase in the usage of videoconferencing tools as a result of the Covid-19 pandemic has resulted in the widespread adoption of a number of now well-known solutions such as Zoom, Jitsi Meet, BigBlueButton, Google Meet, Webex, and Microsoft Teams. Most of the above systems use WebRTC as the underlying technology to realize real-time transmission of audiovisual signals in the browser, without requiring additional software or even server support in-between [1]. Due to the availability of well-configurable open-source WebRTC tools such as Jitsi meet, systematic tests on videoconferencing QoE (Quality of Experience) are enabled.

Typical impairments encountered in communication using VC systems include transmission delay, audiovisual asynchrony, packet loss, jitter, and quality switching (mostly video, seldom also audio). Unlike other factors such as packet-loss or resolution impairments, the assessment of the impact of delay is more challenging, since it is not a directly perceivable media degradation. Over the years, several studies have investigated the impact of these factors on the overall conversational quality and user behavior in typical video conferencing scenarios. For instance, the study in [2] investigates the effects of delay and packet-loss rate on perceived audiovisual and overall conversational quality, and their quality rating results indicate a higher tolerance to delay when compared to packet-loss rate. Furthermore, previous research in [3] investigated the impact of synchronous and asynchronous transmission delays on videoconferencing QoE, suggesting that interaction cues and thus user-behavior such as visual interaction may change in the presence of delay, beyond known changes in conversational patterns [4]. In addition, the authors of [3] argue that participants can hardly identify delay-only conditions as technical degradation, except in the case of highly interactive and less natural conversation tasks such as random number verification. Similarly, the impact of video resolution switching on conversational quality has been investigated in, e.g., [5], [6], but without considering behavior. Further, it has been reported that a reduced video resolution or specific camera configurations are expected to affect the transfer of visual, and non-verbal social cues [7], [8]. With the recent widespread usage of videoconferencing tools, it is important to investigate the relevance of earlier findings regarding today's perception of overall conversational quality and user behavior. Hence, in this paper, we aim to investigate if the general conclusions presented in [3] are still valid and also extend that work by investigating not only network impairments with synchronous transmission delays but including conditions representing video resolution degradation.

979-8-3503-1173-0/23/$31.00 ©2023 IEEE

II. Study design

To evaluate conversational quality under different network delay and video resolution conditions, we set up a WebRTC-based videoconferencing system. We conducted a pre-test to determine the levels of delay and resolution to use. Based on detectability and acceptability thresholds [9], three delay conditions (0ms, 500ms, 1000ms) in combination with three resolutions (240p, 480p, 1080p) were selected. We designed a full factorial 3x3 within-subjects experiment resulting in 9 conditions. The condition order was randomized for each test session using a Latin square design.

A. Participants

Twenty-three pairs of subjects took part in the conversation test. In five pairs, one of the subjects was a member of the team, as a result of the unexpected absence of scheduled participants. The members' results were excluded from the final analysis, resulting in the analysis of the data from 41 participants. Each participant has been compensated (with 12 euros) for participation. Sixteen dyads were reported to be familiar with each other, whereas seven dyads were strangers. Participants' age ranged from 22 to 36 years (M = 27.04, SD = 3.58), where 27 were female and 19 were male.

B. Task and Procedure

Considering the aim of the study, it was critical that the employed task stimulates the use of the video channel for possible social cues. For this, we considered the Celebrity Name-Guessing task (CNG), as suggested by ITU-T Rec. P.1305 [9] and as used in [3].

A training session was conducted to familiarize the participants with the dynamics of the task while interacting face-to-face. In addition, the videoconferencing system was introduced to the participants, using a reference condition (1080p with 0 ms delay) to familiarize the participants with the test procedure and questionnaire. After the training session, participants performed nine trials, one for each test condition. Each trial lasted three minutes and each participant was asked to choose up to two celebrity cards for the partner to guess. After each trial, participants were asked to rate *overall conversational quality* ("How would you rate the overall conversational quality of the call?") and *task performance satisfaction* ("How would you rate your performance on this trial?"). Absolute Category Rating (ACR) was used for both ratings with a scale from 1 to 5. Moreover, the participants were asked one further question that helped us track the *task performance*, i.e., "Were you able to guess the celebrity on this trial?" (0-No, none of them; 1-Yes, one of them; 2-Yes, all of them). In addition, to prevent test participants from losing their attention and becoming fatigued, a longer break of five minutes was taken after 5 trials. In total, for each pair, the experiment session lasted around 45 minutes.

C. Experimental Setup

We used a self-hosted instance of Jitsi Meet as the VC platform. VP9 was used as the video codec and the framerate was set to 30 fps. Google Chrome was used to access Jitsi and to support VP9. For the client side, we used two identical sets of hardware components, including Linux-based laptops (Ubuntu 20.04 LTS), Logitech BRIO 4K cameras, MOTU M4 audio interfaces, LG 27" UHD 27UL850 monitors, and Beyerdynamic DT290 headsets.

For the condition control, video resolution changes between conditions were implemented by modifying the Jitsi Meet media configuration file. We employed the Linux kernel module NetEm [10] on both client sides to apply delays to the outgoing traffic. We gathered the real-time RTP and media stream statistic data via the Chrome webrtc-internals tool to verify the changes and monitor the service.

To further analyze user behavior, we recorded audio from both sides' microphones and video from the webcams and screens. We used the V4L2 loopback kernel module to create virtual loopback devices that receive media streams directly from the webcam. Additionally, for the recording, we used FFmpeg to capture and encode audio and video in raw MP4 format with lossless compression.

D. Feature Extraction

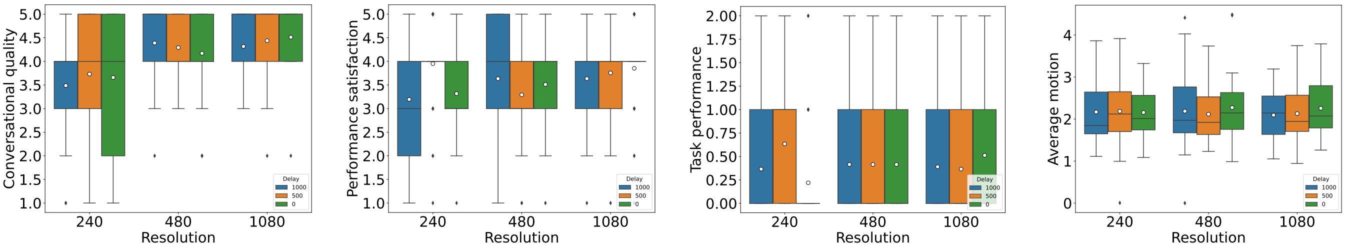
As in [3], a motion feature extracted using the bitstream parser presented in [11] was used, to investigate the differences in user behavior for different test conditions. This feature corresponds to the average motion in the video for a particular test session and is used as a proxy for detecting changes in user movement. In addition to individual movement, we assessed the presence of positive facial expressions. The OpenFace(v.2.2.0) [12] toolkit was used to extract the facial action unit (AU) features from the webcam recordings. We used the average value of the intensity of AU6+AU12 to represent the positive emotion expressions as defined in [13].

III. Experimental Results

Due to the ordinal nature of subjective ratings, we performed the Friedman test followed by the post-hoc paired Wilcoxon signed-rank test with Bonferroni correction to investigate the effect of different levels of test conditions on self-reported overall conversational quality and performance satisfaction. Also, we examined correlations between independent variables, subjective ratings, and objective measures using Spearman's rank correlation. The Kruskal-Wallis test followed by post-hoc Dunn's test with Bonferroni correction was thus performed for the significantly correlated objective components to understand whether the measured user behavior differed based on subjective ratings and test condition levels. The normality distribution assumption data were verified with Shapiro-Wilk's test. Our results are summarized below.

A. Overall Conversational Quality

The distribution of perceived overall conversational quality is presented in Figure 1a. It can be observed that the Mean Opinion Score (MOS) of overall conversational quality increases monotonically with video resolution. The Friedman test showed that resolution has a significant impact on conversational quality ($\chi^2(2) = 24.55, p < .0001$) with delay showing no significant effect ($\chi^2(2) = .94, p = .624$). This is in contrast to the results reported in [] which indicated a significant impact of delay on conversational quality across different resolutions. Following a post-hoc analysis, results showed that participants reported lower conversational quality with 240p than 480p ($z = -.71, p < .001$) and 1080p ($z = -.92, p < .0001$), while no significant difference was found between 1080p and 480p ($z = -.23, p = .298$). Consistent with these results, a positive correlation was found between video resolution and conversational quality ($r(367) = .28, p < .0001$).



(a) Overall conversational quality (b) Performance satisfaction (c) Task performance (d) Average motion

Fig. 1: Results of subjective ratings and objective measures. Subfigures show the distribution of mean scores of the overall conversational quality (1a), performance satisfaction (1b), the number of correct guesses as task performance (1c), and the average motion (1d) across different video resolutions and delays.

B. *Task Performance Satisfaction*

The overall distribution of the performance satisfaction scores is shown in Figure 1b. It was found that highest performance satisfaction was achieved for (240p, 500ms) and (1080p, 0ms) conditions. Performance satisfaction is dependent on whether the participant was able to guess the celebrity correctly or not. Hence, we are of the opinion that the celebrities that were chosen (randomly) for the game for this test condition were either very well known or known to the participants. In addition, further analysis of the data showed that participants had the highest number of correct guesses for these conditions (see Figure 1c), possibly improving performance satisfaction. We found no main effect of delay on self-reported performance satisfaction ($\chi^2(2) = 4.98, p = .083$). We determined a significant effect of video resolution on self-reported performance satisfaction ($\chi^2(2) = 7.72, p = .021$), with post-hoc comparison results revealing that participants who experienced resolution with 1080p reported higher performance satisfaction than 480p ($z = -.33, p = .023$) and 240p ($z = -.35, p = .020$). No significant differences were found between 480p and 240p ($z = .01, p = 1.0$). Accordingly, a positive correlation ($r(367) = .11, p = .043$) between video resolution and self-reported performance satisfaction was found. Also, self-reported performance satisfaction was positively correlated with perceived overall conversational quality ($r(367) = .27, p < .0001$). It is noted that these results may comprise an intrinsic test bias, with subjects mapping perceived quality-differences to performance advantages.

C. *Task Performance*

In addition, a Kruskal-Wallis test showed that task performance (correctness) had a significant effect on performance satisfaction ($\chi^2(2) = 102.72, p < 0.0001$). The post-hoc analysis results using Dunn's test with Bonferroni correction indicated that the self-reported performance satisfaction with two correct guesses was observed to be significantly better than those with one correct guess ($p < .0001$) and no correct guess ($p = .017$). A strong positive correlation between task performance and self-reported performance satisfaction, which was statistically significant ($r(367) = .52, p < .0001$). However, no significant correlation between task performance and self-reported overall conversational quality was found ($r(367) = .03, p = 0.553$). Exploring task performance across test conditions, we found no significant differences regarding different levels of resolution ($\chi^2(2) = 0.33, p = .850$) and delay ($\chi^2(2) = 5.66, p = .059$), using a Friedman test.

D. *Overall Motion*

We found no significant correlation between overall motion and subjective ratings. Based on the Kruskal-Wallis test, no significant difference ($\chi^2(8) = 3.76, p = .878$) in motion was found for the different test conditions (see Figure 1d), unlike the results reported in [3]. This is again in contrast to the results reported in [3] which showed significant differences in the overall motion in the video for different delay conditions.

E. *Facial Expression*

We assessed correlations between positive facial expressions, task performance, and self-reported data. We determined a positive correlation with task performance ($r(367) = .12, p = .014$). No significant difference in facial expression was found for different levels of resolution ($\chi^2(2) = 1.87, p = .393$) and delay ($\chi^2(2) = 0.91, p = .633$) when analyzed with a Friedman test.

IV. Conclusions

The aim of this paper was to test the reproducibility of the results presented in [3] and also to extend that work. The results showed that there was no impact of delay on the overall conversational quality in contrast to the results reported in [3]. In addition, it was further observed that performance satisfaction was positively correlated with perceived conversational quality. Task performance was not significantly reduced despite network barriers, but it significantly affected users' self-reported performance satisfaction. Similarly, no difference in user behavior in terms of user movement and positive facial expression was found across test conditions. In future work, further social cue analysis in terms of conversational turn-taking and eye-gaze analysis will be conducted, based on the recordings from this and follow-up tests.

Acknowledgment

This work was funded by the Carl-Zeiss-Stiftung under the project: "Co-Presence of Humans and Interactive Companions for Seniors (CO-HUMANICS)".

References

[1] B. Sredojev, D. Samardzija, and D. Posarac. “WebRTC technology overview and signaling solution design and implementation”. In: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2015, pp. 1006–1009.

[2] F. Brauer, M. S. Ehsan and G. Kubin. ”Subjective evaluation of conversational multimedia quality in IP networks”. In: 2008 IEEE 10th Workshop on Multimedia Signal Processing, Cairns, QLD, Australia, 2008, pp. 872-876.

[3] K. Schoenenberg, A. Raake, and P. Lebreton. “Conversational quality and visual interaction of video-telephony under synchronous and asynchronous transmission delay”. In: 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX). 2014, pp. 31–36.

[4] K. Schoenenberg, A. Raake, S. Egger, R. Schatz. ”On interaction behaviour in telephone conversations under transmission delay”. Speech Communication, 63. 2014, pp. 1-14.

[5] D. Vučić and L. Skorin-Kapov. ”QoE evaluation of WebRTC-based Mobile Multiparty Video Calls in Light of Different Video Codec Settings”. In: 2019 15th International Conference on Telecommunications (ConTEL), Graz, Austria, 2019, pp. 1-8.

[6] D. Vučić and L. Skorin-Kapov. ”QoE Assessment of Mobile Multiparty Audiovisual Telemeetings”. In: IEEE Access, vol. 8. 2020, pp. 107669-107684.

[7] J. N. Bailenson. “Nonverbal Overload: A Theoretical Argument for the Causes of Zoom Fatigue”. In: Technology, Mind, and Behavior 2.1. 2021.

[8] G. Fauville, A. Queiroz, M. Luo, J. Hancock, & J. Bailenson. ”Impression Formation From Video Conference Screenshots: The Role of Gaze, Camera Distance, and Angle”. Technology, Mind, and Behavior, 3(1: Spring 2022), 2022, pp. 1-11.

[9] ITU-T Recommendation. P.1305. Effect of delays on telemeeting quality. 2016.

[10] S. Hemminger. “Network emulation with NetEm”. In: Linux Conf Au. 2005.

[11] R.R.R. Rao, S. Göring, P. List, W. Robitza, B. Feiten, and A. Raake ”Bitstream-Based Model Standard for 4K/UHD: ITU-T P.1204.3 — Model Details, Evaluation, Analysis and Open Source Implementation,” 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), Athlone, Ireland, 2020.

[12] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. “Openface 2.0: Facial behavior analysis toolkit”. In: 2018 13th IEEE international conference on automatic face & gesture recognition. 2018, pp. 59–66.

[13] A. Sayette, F. Cohn, M. Wertz, A. Perrott, and J. Parrott. ”A psychometric evaluation of the facial action coding system for assessing spontaneous expression”. Journal of nonverbal behavior 25. 2001, pp.167-185.

[14] K. Gros and N. Chateau. “The impact of listening and conversational situations on speech perceived quality for time-varying impairments”. In: MESAQIN 2002. 2002, pp. 17–19.