Gülnaziye Bingöl, Simone Porcu, Alessandro Floris, Luigi Atzori,

WebRTC-QoE: A dataset of QoE assessment of subjective scores, network impairments, and facial & speech features in *Computer Networks*, Volume 244 (2024), art. 110356.

The publisher's version is available at:
https://doi.org/10.1016/j.comnet.2024.110356

Article Title

WebRTC-QoE: A Dataset of QoE Assessment of Subjective Scores, Network Impairments, and Facial & Speech Features

Authors

Gülnaziye Bingöl[1,2], Simone Porcu[1,2], Alessandro Floris[1,2], Luigi Atzori[1,2]

Affiliations

1. DIEE, University of Cagliari, 09123 Cagliari, Italy
2. CNIT, University of Cagliari, 09123 Cagliari, Italy

Corresponding author(s)

Alessandro Floris (alessandro.floris84@unica.it)

Abstract

In the realm of real-time communications, WebRTC-based multimedia applications are increasingly prevalent as these can be smoothly integrated within Web browsing sessions. The browsing experience is then significantly improved with respect to scenarios where browser add-ons and/or plug-ins are used; still, the end user's Quality of Experience (QoE) in WebRTC sessions may be affected by network impairments, such as delays and losses. Due to the variability in user perceptions under different communications scenarios, comprehending and enhancing the resulting service quality is a complex endeavor. To address this, we present a dataset that provides a comprehensive perspective on the conversational quality of a two-party WebRTC-based audiovisual telemeeting service. This dataset was gathered through subjective evaluations involving 20 subjects across 15 different test conditions (TCs). A specialized system was developed to induce controlled network disruptions such as delay, jitter, and packet loss rate, which adversely affected the communication between the parties. This methodology offered an insight into user perceptions under various network impairments. The dataset encompasses a blend of objective and subjective data including ACR (Absolute Category Rating) subjective scores, webrtc-internals parameters, facial expressions features, and speech features. Consequently, it serves as a substantial contribution to the improvement of WebRTC-based video call systems, offering practical and real-world data that can drive the development of more robust and efficient multimedia communication systems, thereby enhancing the user’s experience.

DOI: 10.1016/j.comnet.2024.110356

Keywords

WebRTC; Quality of Experience; Network Impairments; Facial Expressions; Speech; webrtc-internals.

Specifications Table

Subject	Computer Networks and Communications
Specific subject area	Quality of Experience Assessment of WebRTC-based Audio-Video Communications
Type of data	Table Text
How data were acquired	Subjective tests where the involved subjects have completed a survey to provide feedback on their experience perceived during video calls. The following hardware has been used to conduct the experiments: Hardware: 1 Router Netgear 2.4 GHz; 2 Laptops MSI GE63 Raider RGB 8RE, CPU 8th Gen. Intel Core i7, 16GB RAM, GeForce GTX 1060 6GB GDDR5; 2 Webcams Logitech C920 FHD (30fps@1080p); 2 Headphones Sony WI-C100 Software: Dummynet, NetEm (Network Emulator), PyNetem, WebRTC on Google Chrome Browser version 96, OpenFace 2.2.0 toolkit, OpenSMILE toolkit.
Data format	Raw: .CSV, .TXT
Parameters for data collection	WebRTC-based audiovisual calls were implemented on Google Chrome Browser. 15 Test Conditions (TCs) included combinations of 3 network impairments (delay, jitter, packet loss) to affect communication. 20 subjects participated in the subjective assessment and rated the quality of 15 audiovisual calls implemented with the 15 TCs. The resolution for the video camera was set to 1920x1080 at 30 fps.
Description of data collection	Each of the 20 subjects participated in 15 calls with 15 different TCs, which ended in 2 minutes. The audiovisual calls were recorded. After each call, participants were asked to complete a short survey to evaluate the quality of the call. *Webrtc-internals* dump text files were collected for each participant after every call. Facial expression features were extracted from the recorded video files and reported in a table for each subject and TC. Speech features were extracted from the recorded audio files and reported in a table for each subject and TC.
Data source location	Institution: University of Cagliari City: Cagliari Country: Italy

Data accessibility	Repository name: WebRTC-QoE: A Dataset of Quality of Experience in Audio-Video Communications Direct URL to data: https://dx.doi.org/10.21227/mb47-hf44
Related research article	Author's name: G. Bingöl, L. Serreli, S. Porcu, A. Floris and L. Atzori. Title: The Impact of Network Impairments on the QoE of WebRTC applications: A Subjective study. Conference: 2022 14th International Conference on Quality of Multimedia Experience (QoMEX), Lippstadt, Germany, 2022, pp. 1-6. DOI: 10.1109/QoMEX55416.2022.9900882. Author's name: G. Bingöl, S. Porcu, A. Floris and L. Atzori. Title: QoE Estimation of WebRTC-based Audiovisual Conversations from Facial Expressions. Conference: 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Dijon, France, 2022, pp. 577-584. DOI: 10.1109/SITIS57111.2022.00092. Author's name: M. Hamidi, G. Bingöl, A. Floris, S. Porcu, and L. Atzori. Title: Analysis of Application-layer Data to Estimate the QoE of WebRTC-based Audiovisual Conversations. Conference: 2023 IEEE Globecom Workshops, Kuala Lumpur, Malaysia, 2023. DOI: 10.1109/GCWkshps58843.2023.10464821. Author's name: G. Bingöl, S. Porcu, A. Floris, and L. Atzori. Title: QoE Estimation of WebRTC-based Audio-visual Conversations from Facial and Speech Features. Journal: ACM Trans. Multimedia Comput. Commun. Appl. 20, 5, Article 130 (January 2024), 23 pages. DOI: 10.1145/3638251.
Related project(s)	This work has been partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, "Sustainable Mobility Center" Centro Nazionale per la Mobilità Sostenibile, CNMS, CN_00000023, and by the Italian Ministry for University and Research under the PON "Ricerca e Innovazione" 2014-2020 (PON R&I) "Azione IV.4 Dottorati e contratti di ricerca su tematiche dell'innovazione" with D.M. 1062 on 10.08.2021.

Value of the Data

1. Why are these data useful?

   In recent years, due to the Covid-19 pandemic, the utilization of videoconferencing tools has experienced a significant increase for various purposes, such as work meetings, e-learning, and social interactions. WebRTC-based applications have emerged among the other videoconferencing tools because of the easy embedding of real-time voice and video communications on Web browsers, which prevents the installation of dedicated tools on the user's device. Since the real-time traffic generated by WebRTC-based tools is subject to distortions that depend on the network connection status, there is a need to assess the impact of common network impairments on the user's Quality of Experience (QoE). Moreover, while asking users to provide feedback on the perceived QoE remains the main approach to collect ground-truth data, alternative unobtrusive methods (i.e., that do not require user feedback) have emerged, which rely on the physiological bases of the user's perceptual and cognitive processes. In this regard, facial expressions and vocal speech characteristics have gathered particular attention because they naturally convey human emotions driven by the user's emotional state. To address these challenges, we have conducted a subjective experiment using WebRTC technology in a controllable environment under varying network impairments to evaluate the QoE and define QoE estimation models based on facial and speech features. The resulting dataset includes data collected from 20 subjects across 15 different Test Conditions (TCs) and includes subjective scores, offering a unique perspective on user perceptions under various TCs, which is crucial for improving video call technology. It also includes application-level data concerning WebRTC sessions' statistics, which were collected with the *webrtc-internals* tool during the audiovisual calls. Finally, the dataset includes features extracted from users' facial expressions and speech, offering a more nuanced understanding of user reactions and emotions. This dataset allows researchers to develop QoE models that can be used to engineer audio-video call technologies that rely on the WebRTC framework. These models may predict the users' quality from all the extracted features, i.e., network and application layer parameters obtained from the WebRTC platform, users' facial expression features, and speech features. These may be treated in isolation or using a proper fusion model. The study's chosen network parameters also allow for deep understanding of how the WebRTC's concealment techniques can provide high-quality real-time video communications.

2. Who can benefit from these data?

   Researchers and developers working on WebRTC-based video call systems, signal processing, and QoE modelling can significantly benefit from this dataset. It provides valuable insights into objective application data from *webrtc-internals* and users' subjective evaluations under various network conditions, which can be crucial for improving user experience and system performance. Indeed, objective QoE models that connect network- and application-related data with the QoE may be derived from these data. Moreover, the dataset provides facial and speech features extracted from the users during the audiovisual calls, which offer a unique opportunity for emotion detection. Investigation of affective behavior and perceived QoE serves as a valuable resource for seeking to enhance WebRTC-based video call systems, offering a wealth of real-world data to guide their efforts.

3. How can these data be used for further insights and development of experiments?

   The dataset can be employed to examine the impact of different network conditions on video calls and their subjective experimental evaluation from user interactions. Based on the results provided by this dataset, novel experiments can be designed to further investigate the impact of additional network impairments or different ranges for the considered network impairments so as to extend the validity of the current solution. The extracted facial and speech features can be used to determine users' emotional states, providing a rich source of data for developing and refining emotion detection algorithms related to the utilization of multimedia services. Further analysis would be required to determine whether detected users' emotional states (e.g., negative emotions) are due to user interactions or to the event of disturbed communication. Moreover, data fusion techniques can be tested on the dataset to derive novel QoE estimation models aimed at predicting the QoE based on a combination of network-related data, application-level data, facial expression features, and speech features.

4. What is the additional value of these data?

   This dataset offers a unique blend of objective data from *webrtc-internals* and subjective experimental results, providing a holistic perspective on video call quality. This includes network parameters, application-level data, and user-perceived quality, which are further enriched by the inclusion of facial expressions and speech features captured during the call, allowing for a more detailed understanding of the user experience. There is no other similar contribution in literature to the best of the authors' knowledge.

Data

1. Subjective_results_dataset.csv

This dataset encompasses subjective evaluation results from 20 subjects (users) who assessed the quality of WebRTC-based audio-video calls under 15 distinct TCs, which included combinations of 3 network impairments (delay, jitter, packet loss) to affect the communication. The single discrete Absolute Category Rating (ACR) scale with five category labels (1-Bad, 2-Poor, 3-Fair, 4-Good, and 5-Excellent) was used by the users to rate the perceived QoE. A total of 300 ACR scores were obtained (20 participants x 15 TCs). The size of this dataset is 4.00 KB. The structure of the table is depicted in Figure 1.

The significance of each column is explained as follows:

- Test Condition (TC): It enumerates the TC numbers, which span from 1 to 15.

• Delay [ms]: It refers to the time it takes for a signal to travel from one point to another and is represented at three levels: 0 ms (no delay), 500 ms (moderate delay), and 1000 ms (significant delay).

• Jitter [ms]: It refers to the variability in delay and is represented at two levels: 0 ms (no jitter) and 500 ms (moderate jitter).

• Packet Loss Rate [%]: It refers to the loss of data packets during transmission and is represented at three levels: 0% (no packet loss), 15% (moderate packet loss), and 30% (significant packet loss).

• User: The users, identified from 1 to 20, participated in 15 video calls and evaluated the quality of these calls on a scale from 1 to 5.

Test Condition (TC)	Delay [ms]	Jitter [ms]	Packet Loss Rate [%]	User1	User2	User3	User4	User5	User6	User7	User8	User9	User10	User11	User12	User13	User14	User15	User16	User17	User18	User19	User20
1	0	0	0	5	4	3	5	3	3	5	5	3	3	3	4	4	4	3	5	3	5	4	4
2	500	0	0	4	3	3	4	3	2	4	5	5	3	1	5	3	5	3	4	3	4	3	5
3	1000	0	0	4	4	3	4	3	3	5	4	5	3	3	5	3	4	3	4	4	4	4	3
4	500	500	0	2	2	3	4	3	3	5	3	4	1	3	4	4	4	4	3	4	4	4	4
5	1000	500	0	3	3	3	4	2	3	4	4	4	3	3	5	4	4	3	3	3	3	3	4
6	0	0	15	3	3	4	3	3	3	4	3	3	2	4	5	3	3	2	4	3	3	3	3
7	500	0	15	3	3	3	3	3	3	3	2	4	2	3	5	3	4	4	2	2	4	3	2
8	1000	0	15	2	1	3	3	3	3	3	2	3	2	3	3	3	3	3	3	2	3	4	2
9	500	500	15	1	3	2	3	3	3	3	1	2	2	2	4	2	2	2	3	3	2	3	2
10	1000	500	15	2	3	3	1	3	2	3	3	2	1	2	5	1	3	4	3	1	3	3	2
11	0	0	30	1	3	2	3	3	3	2	2	3	2	2	4	3	3	3	3	2	3	3	2
12	500	0	30	2	2	2	1	2	2	2	1	3	2	2	5	2	3	2	2	1	2	3	2
13	1000	0	30	1	3	2	3	3	2	2	2	3	1	1	5	2	3	2	2	1	2	3	1
14	500	500	30	1	3	1	3	3	3	1	2	2	2	4	2	2	2	1	1	1	2	3	1
15	1000	500	30	1	3	1	1	3	2	2	1	1	1	1	3	1	1	1	1	1	3	3	1

Figure 1: Structure of the table *Subjective_results_dataset.csv*.

2. Webrtc_internals_dataset.zip

   This dataset contains text files collected using the *webrtc-internals* tool during the video calls. The zip is organized into 20 distinct folders, each labeled from 'User1' to 'User20'. Each user folder contains 15 text files (.txt), each named following the pattern 'webrtc_internals_dump-TCx_y-z-t.txt'. In this naming convention, 'x' denotes the TC number, ranging from 1 to 15. The 'y-z-t' segment varies with each TC, where 'y' signifies the delay value (0, 500, or 1000), 'z' indicates the jitter value (0 or 500), and 't' represents the packet loss rate (0, 15, or 30). Each text file includes application-level data concerning WebRTC sessions' statistics in a JSON format. A total of 300 *webrtc-internals* dump text files were obtained (20 participants x 15 TCs). The size of this zip dataset is 28.7 MB (396 MB uncompressed).

3. Facial_expression_features_dataset.zip

   This dataset contains facial expression features extracted from the recorded videos with face images using the OpenFace toolkit [1]- [2]. The zip is organized into 20 distinct folders, each labeled from 'User1' to 'User20'. Each user folder contains 15 .csv files, which are named following the pattern 'TCx_y-z-t.csv'. In this naming convention, 'x' denotes the TC, ranging from 1 to 15. The 'y-z-t' segment varies with each TC, where 'y' represents the delay value (0, 500, or 1000), 'z' indicates the jitter value (0 or 500), and 't' represents the packet loss rate (0, 15, or 30). A total of 300 facial expression feature files in .csv format were obtained (20 participants x 15 TCs).

For each face image of each TC, the OpenFace outputs 6 gaze direction features, 280 eye region landmarks and 35 Action Units (AUs). The size of this zip dataset is 547 MB (2.14 GB uncompressed). The structure of the table is represented in Figure 2.

Each column carries a specific significance, which is elaborated as follows:

- frame: the frame number in the context of sequences.
- face_id: the identifier assigned to each face when multiple faces are present.
- timestamp: the elapsed time in seconds during the processing of a video sequence.
- confidence: the level of confidence the tracker has in the current landmark detection estimate.
- success: a face has been detected in the frame and it has been tracked accurately.
- gaze_0_x, gaze_0_y, gaze_0_z: the normalized eye gaze direction vector in world coordinates for eye 0, which is the eye on the left in the image.
- gaze_1_x, gaze_1_y, gaze_1_z: the normalized eye gaze direction vector in world coordinates for eye 1, which is the eye on the right in the image.
- gaze_angle_x, gaze_angle_y: the eye gaze direction, averaged for both eyes and expressed in world coordinates in radians, is converted into a format that is easier to use than gaze vectors.
- eye_lmk_x_0, eye_lmk_x_1, ..., eye_lmk_x55, eye_lmk_y_1, ... eye_lmk_y_55: the pixel coordinates of 2D landmarks in the eye region.
- eye_lmk_X_0, eye_lmk_X_1, ..., eye_lmk_X55, eye_lmk_Y_0, ..., eye_lmk_Z_55: the position of landmarks in the eye region in 3D space, measured in millimeters.
- 17 AUr: detect the activation intensity (from 1 to 5) of a particular facial muscle. These are: *AU01_r, AU02_r, AU04_r, AU05_r, AU06_r, AU07_r, AU09_r, AU10_r, AU12_r, AU14_r, AU15_r, AU17_r, AU20_r, AU23_r, AU25_r, AU26_r, AU45_r.*
- 18 AUc: Identify the activation of a particular muscle and note its presence (0 for absent, 1 for present). These are: *AU01_c, AU02_c, AU04_c, AU05_c, AU06_c, AU07_c, AU09_c,*

AU10_c, AU12_c, AU14_c, AU15_c, AU17_c, AU20_c, AU23_c, AU25_c, AU26_c, AU28_c, AU45_c.

frame	face_id	timestamp	confidence	success	gaze_0_x	gaze_0_y	gaze_0_z	gaze_1_x	gaze_1_y	gaze_1_z	gaze_angle_x	gaze_angle_y	eye_lmk_x_0
1	0	0	0.98	1	0.03007	0.25479	-0.96653	-0.24332	0.29676	-0.92344	-0.112	0.284	200.5
2	0	0.033	0.98	1	0.05283	0.27474	-0.96007	-0.22821	0.31841	-0.92007	-0.093	0.306	199.6
3	0	0.067	0.98	1	0.03933	0.25538	-0.96604	-0.23015	0.30003	-0.92575	-0.101	0.286	200.1
4	0	0.1	0.98	1	0.03716	0.27321	-0.96124	-0.22745	0.30569	-0.92457	-0.101	0.298	199.6
5	0	0.133	0.98	1	0.0257	0.26128	-0.96492	-0.23761	0.29634	-0.92505	-0.112	0.287	199.7
6	0	0.167	0.98	1	0.02692	0.27787	-0.96024	-0.23021	0.30899	-0.92278	-0.108	0.302	199.9
7	0	0.2	0.98	1	0.02061	0.28162	-0.95931	-0.22071	0.30792	-0.92546	-0.106	0.303	200.1
8	0	0.233	0.98	1	0.05926	0.27283	-0.96024	-0.21265	0.31459	-0.9251	-0.081	0.302	201
9	0	0.267	0.98	1	0.05682	0.28355	-0.95727	-0.23266	0.32327	-0.91726	-0.094	0.313	201.6
10	0	0.3	0.98	1	0.05097	0.27245	-0.96082	-0.24073	0.30939	-0.91996	-0.101	0.3	202

Figure 2: Structure (a part) of the csv tables included into the *Facial_expression_features_dataset.*

4. Speech_features_dataset.csv

This dataset is a robust assembly of speech features extracted from the recorded audio files using the OpenSMILE toolkit [3]- [4]. The dataset is further enhanced with the integration of ACR scores, which were assigned by each subject for every TC. These scores are embedded within the speech features of each subject. The dataset is exhaustive, encompassing a total of 1,911,900 speech features (calculated as 15 TCs x 20 subjects x 6373 speech features per subject). The total size of this dataset is 18.9 MB. The structure of the table is represented in Figure 3.

Each column carries a specific significance, which is elaborated as follows:

- file: It presents the audio files from which the speech features were extracted.
- Users: The users, identified from 1 to 20, participated in 15 video calls and evaluated the quality of these calls on a scale from 1 to 5.
- Test Condition (TC): It enumerates the TC numbers, which span from 1 to 15.
- Delay [ms]: It refers to the time it takes for a signal to travel from one point to another and is represented at three levels: 0 ms (no delay), 500 ms (moderate delay), and 1000 ms (significant delay).
- Jitter [ms]: It refers to the variability in delay and is represented at two levels: 0 ms (no jitter) and 500 ms (moderate jitter).
- Packet Loss Rate [%]: It refers to the loss of data packets during transmission and is represented at three levels: 0% (no packet loss), 15% (moderate packet loss), and 30% (significant packet loss).
- OpenSmile Speech Features Columns: It presents a detailed view of the speech features. Specifically, from column G through to column IKI, it enumerates 6373 distinct speech features for each user under each TC of each processed audio file in .wav.

- ACR Score: The ACR scale ranges from 1, representing the lowest quality, to 5, indicating the highest quality evaluated by the subjects.

file	Users	Test Condition (TC)	Delay [ms]	Jitter [ms]	Packet Loss Rate [%]	audspec_lengthL1norm_sma_range	audspec_lengthL1norm_sma_maxPos	audspec_lengthL1norm_sma_minPos
./VideoTests/User1/TC1/0-0-0.wav	1	1	0	0	0	8.502644	0.8216545	0.8642898
./VideoTests/User1/TC2/500-0-0.wav	1	2	500	0	0	6.1701984	0.43159622	0.09011079
./VideoTests/User1/TC3/1000-0-0.wav	1	3	1000	0	0	6.9106674	0.84612226	0.87550265
./VideoTests/User1/TC4/500-500-0.wav	1	4	500	500	0	6.316081	0.11070989	0.5025852
./VideoTests/User1/TC5/1000-500-0.wav	1	5	1000	500	0	6.3915243	0.20196964	0.73500204
./VideoTests/User1/TC6/0-0-15.wav	1	6	0	0	15	7.963232	0.277144	0.32646698
./VideoTests/User1/TC7/500-0-15.wav	1	7	500	0	15	6.3506064	0.30160034	0.50660646
./VideoTests/User1/TC8/1000-0-15.wav	1	8	1000	0	15	6.3289537	0.8336479	0.35141566
./VideoTests/User1/TC9/500-500-15.wav	1	9	500	500	15	5.686478	0.44021338	0.22938038
./VideoTests/User1/TC10/1000-500-15.wav	1	10	1000	500	15	5.97985	0.40287238	0.25990972
./VideoTests/User1/TC11/0-0-30.wav	1	11	0	0	30	5.73369	0.18473533	0.99893314
./VideoTests/User1/TC12/500-0-30.wav	1	12	500	0	30	6.191719	0.079852276	0.21936807
./VideoTests/User1/TC13/1000-0-30.wav	1	13	1000	0	30	6.1544757	0.5493086	0.3885899
./VideoTests/User1/TC14/500-500-30.wav	1	14	500	500	30	5.810598	0.13787444	0.49610177
./VideoTests/User1/TC15/1000-500-30.wav	1	15	1000	500	30	4.0455604	0.10042433	0.7821782

Figure 3: Structure (a part) of the table *Speech_features_dataset.csv.*

Experimental Design, Materials, and Methods

We have designed and conducted a subjective quality assessment during which test participants had to talk with a conversation partner using a WebRTC-based application. During the talk, the face and the speech of the participants were recorded to collect a data stream for extracting facial expressions features and speech features. Application-level data were also collected using the *webrtc-internals* tool. The objective of the experimental test was to study the QoE of audiovisual calls through a WebRTC-based application when the network is affected by combinations of three network impairments, i.e., delay, jitter, and packet loss rate (PLR). We selected a total of 15 TCs, which are summarized in Table 1. To generate the jitter, we used the uniform pattern, i.e., the distribution of packet delays was equally weighted. The packet loss was generated using a random approach, i.e., each packet loss was independent from the other.

The network parameters considered were used to offer varying quality levels to the test participants. Initially, the selected network parameter values may seem unconventional. For example, one-way delays exceeding 400 ms or high PLR values (e.g., 5%) are unacceptable in a VoIP scenario. However, WebRTC and Google Chrome employ concealment techniques, including packet retransmissions, Forward Error Correction (FEC), and the Google Congestion Control (GCC) algorithm. These techniques enhance the robustness of real-time video streaming applications against network distortions. Therefore, to provide different quality levels during the assessment, we strategically combined network parameters to counteract the robustness provided by these concealment techniques. Indeed, the values of the network impairments selected for the assessment, and summarized in Table 1, are the results of several tests during which 4 experts in QoE assessment evaluated the impact of the applied network distortions on audiovisual quality.

The experiment involved 20 Italian participants (11 females and 9 males) aged 23-36 years with normal or corrected-to-normal vision. The assessment was conducted at the University of Cagliari’s Department of Electrical and Electronic Engineering. Participants were placed in separate rooms and wore headphones to prevent external disturbances. Before the assessment, participants were briefed that the network conditions would vary for each call session and trained on the rating methods. Specifically, they had to participate in three training call sessions, which were set to give participants an overview of the levels of impairments involved in the experiment. Particularly, these three call sessions included no network impairments, moderate network impairments, and the worst network impairments (as for TC15). The aim of the training session was to make participants aware of the possible quality levels they could encounter during the assessment so as to provide conscious scores. During the assessment the 15 TCs were applied in random order to the participants to avoid influencing the rating process.

The task involved a “Who am I?” celebrity name-guessing game, which consists in guessing the celebrity chosen by the partner by asking, in turn, yes/no questions to the partner. They were also instructed to complete a brief Web-based questionnaire after each call to provide feedback on their perceived quality. They had to use the 5-level labels (1-Bad, 2-Poor, 3-Fair, 4-Good, and 5-Excellent) ACR scale to rate the overall perceived QoE.

Test participants were provided with a laptop to make the WebRTC-based video calls using the Google Chrome browser. An access point was used to create a dedicated wireless network communication between the two laptops, which were placed in two different rooms. The dedicated network avoided undesirable network distortion provided by the Internet.

Network protocols were tested and simulated using Dummynet [5], while NetEm [6] software was used to emulate the considered network impairments. PyNetem was used to establish a Web server for dynamic HTTP requests. The iPerf tool was used to measure the introduced network impairments and verify they respected the desired range. The system was set to randomly introduce one of the combinations of network impairments listed in Table 1 for each WebRTC session. After each session (a 2-minute video call), a text file from *webrtc-internals*-dump file was automatically saved. W*ebrtc-internals* is a feature of the Chrome browser that provides real-time statistics for active WebRTC sessions, which can be visualized in real-time as well as downloaded at the end of the session.

During the audio-visual conversations, we captured the participants’ facial expressions, resulting in a total of 300 videos, recorded in full HD quality (1920x1080) at 30 fps. Subsequently, we utilized the OpenFace toolkit [1]- [2] to extract 300 facial expression data files in .csv format from these recorded facial images. In addition, we employed the OpenSmile toolkit [3]- [4] to extract approximately 1,911,900 speech features from the participants’ recorded speech. Both toolkits were integrated into a Python script written in Python 3.9 for the extraction process to collect the features into a .csv format file.

Table 1 Test Conditions.

TC	Delay (ms)	Jitter (ms)	PLR (%)
1	0	0	0
2	500	0	0
3	1000	0	0
4	0	500	0
5	500	500	0
6	1000	0	15
7	0	0	15
8	500	0	15
9	1000	500	15

10	0	500	15
11	500	0	30
12	1000	0	30
13	0	0	30
14	500	500	30
15	1000	500	30

We would like to highlight that there are other parameters that could be considered to have a better view of all factors affecting QoE in communication networks, such as network stability and bandwidth variations as well the distribution and the types of errors. Additionally, the pool of profiles of the subjects involved in the tests should be further extended in terms of cultural context and preferences to ensure a robust and diverse data collection so as to build more reliable models. We believe that all these aspects should be considered in future studies to further advance in the understanding and modelling of perceived quality in video-calls.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, "Sustainable Mobility Center" Centro Nazionale per la Mobilità Sostenibile, CNMS, CN_00000023, and by the Italian Ministry for University and Research under the PON “Ricerca e Innovazione” 2014-2020 (PON R&I) “Azione IV.4 Dottorati e contratti di ricerca su tematiche dell’innovazione” with D.M. 1062 on 10.08.2021.

References

[1] T. Baltrusaitis, A. Zadeh and Y. C. Lim and L. -P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit,," in *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an,*, China, 2018, pp. 59-66, doi: 10.1109/FG.2018.00019., 2018.

[2] T. Baltrusaitis, M. Mahmoud and and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection.," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Vol. 6. IEEE, 2015*, Ljubljana, Slovenia, doi: 10.1109/FG.2015.7284869, 2015.

[3] Florian Eyben, Felix Weninger, Florian Gross and a. B. Schuller., "Recent Developments in OpenSMILE, the," in *In Proc. of the 21st ACM Int. Conf. on Multimedia. ACM, 835–838.*, Barcelona, Spain, doi: 10.1145/2502081.2502224, 2013.

[4] Florian Eyben, Martin Wöllmer and a. B. Schuller, "Opensmile: The Munich Versatile and Fast Open-Source," in *In Proc. of the 18th ACM Int. Conf. on Multimedia. ACM, 1459–1462*, Firenze, Italy, doi: 10.1145/1873951.1874246, 2010.

[5] L. Rizzo, "Dummynet: a simple approach to the evaluation of network," in *ACM SIGCOMM Computer Communication Review, vol. 27*, New York, NY, USA, doi: 10.1145/251007.251012, 1997.

[6] S. Hemminger, "Network emulation with NetEm," 2005.

Biographies

Gülnaziye Bingöl earned her B.Sc. degree in Electrical and Electronic Engineering from Omer Halisdemir University in Turkey in June 2019. In 2018, she also had the opportunity to study as a visiting student at the University of Oradea and worked as a summer intern at the University of Cagliari. She further advanced her education by completing her Master's Degree in Internet Engineering at the University of Cagliari in 2021, achieving a full score of 110/110. Currently, Gulnaziye is dedicated to her Ph.D. studies in the Department of Electrical and Electronic Engineering (DIEE/UdR CNIT) at the University of Cagliari. Her research interests revolve around Quality of Experience (QoE), Green Multimedia Services, WebRTC, Data Science, Machine Learning, and Human Behaviour.

Simone Porcu is an Assistant Professor at the University of Cagliari (Italy), where he received his M. Sc. degree in Computer Science and his Ph. D. in Electronic and Computer Engineering respectively in 2017 and 2021. He also obtained the Doctor Europaeus title for his research period abroad at the Technische Universitat Berlin (September 2018-March 2019). He is a member of the Net4U Laboratory (https://sites.unica.it/net4u/) of the Department of Electrical and Electronic Engineering (DIEE) at the University of Cagliari from 2017, a member of the Italian University Consortium for Telecommunications (CNIT) from 2018, and a member of the IEEE from 2018. He is Associate Editor for IEEE COMSOC MMTC Communications - Frontiers and participated as Guest Editor for the MDPI Electronics journal. His main research interests include Quality of Experience (QoE), multimedia communications, artificial intelligence, and metaverse applications.

Alessandro Floris is an Assistant Professor at the University of Cagliari (Italy), where he received his M. Sc. degree in Electronic Engineering and his Ph. D. in Electronic and Computer Engineering respectively in 2011 and 2017. He is qualified for the position of Associate Professor according to the Italian law. He is a member of the Net4U Laboratory (https://sites.unica.it/net4u/) of the Department of Electrical and Electronic Engineering (DIEE) at the University of Cagliari from 2011, a member of the Italian University Consortium for Telecommunications (CNIT) from 2011, and a member of the IEEE from 2014. He is Associate Editor for IEEE Access and for IEEE COMSOC MMTC Communications - Frontiers. His main research interests include Quality of Experience (QoE) based assessment of multimedia systems, QoE-based network and application management approaches, sustainable multimedia communication, and Internet of Things (IoT) applications for Smart Cities.

Luigi Atzori (PhD, 2000) is professor of Telecommunications at the University of Cagliari, where he leads the activities of the Net4U laboratory (Network for Humans) with around 20 affiliated researchers. Since 2018, he has been the coordinator of the master's degree course in Internet Technology Engineering at the University of Cagliari. His research interests fall in the area of Internet of Things (IoT), with particular reference to the design of effective algorithms for the realization of social networks among connected devices to develop the Social IoT paradigm. His works on the IoT and the social IoT have received a great impact with more than 20K citations. His interests also fall in the area of Quality of Experience (QoE), with particular application to the management of services and resources in new generation networks for multimedia communications. Lately, he also applies the study of QoE to IoT services. He serves regularly as referee for several international and national funding programs, in the organizing committee of international conferences, and as associate and guest editor in several international journals (Ad Hoc Networks, IEEE Open Journal of the Communications Society, IEEE Communications Magazine, etc.). He is co-founder of two spinoffs in the areas of IoT solutions for people mobility.