Conversational Subjective Tests Based on Video-telephony Platform

Jose Joskowicz[1]
Universidad de la Republica
Montevideo, Uruguay
josej@fing.edu.uy

Mengying Liu[1]
China Mobile Research Institute
Beijing, China
liumengying@chinamobile.com

Rafael Sotelo
Universidad de Montevideo
Montevideo, Uruguay
rsotelo@um.edu.uy

Alejandra Armendariz
Universidad de la Republica
Montevideo, Uruguay
aarmendariz@fing.edu.uy

Lei Yang
China Mobile Research Institute
Beijing, China
yangleiyj@chinamobile.com

***Abstract*—Video-telephony applications are widely used in offices, education, medical care, social and other fields. This paper introduces a method designed for conversational subjective tests, based on an open-source video-telephony platform, focusing on multimedia quality and interaction experience perceived by users under various network states. The research is carried out as part of the 'Computational model used as a QoE/QoS monitor to assess video-telephony services' (G.CMVTQS) project, which is under study in ITU-T SG12 Q.15. Six different laboratories, from four countries, are collaboratively working on the conversational subjective tests. The designed test bed is described, along with the lab deployment, the simulation of network distortions and the method used for subjective evaluation.**

I. Introduction

With the development and popularization of video-telephony services, video-telephony platforms have been widely used in offices, education, medical care, social and other fields. The demand for video-telephony quality assessment is increasingly prominent. Moreover, the quality of video-telephony services not only directly affects user experience but is also one of the important factors for attracting and maintaining user stickiness.

The user experience of video-telephony applications has attracted lots of attention from the research community. Skowronek et al. provide their readers with an entry point to the field of QoE of telemeetings, by sharing a comprehensive survey of factors and processes, and an overview of relevant state-of-art QoE assessment methods [1]. Scholars also delve into various specific research aspects. Vučić et al. identified that key system-related QoE influence factors can be divided into three categories: media quality, functional support, and usability and service design [2]. Among these categories of QoE influence factors, we focus more on media quality and how it affects user experience. The research on media quality can also be subdivided into several dimensions, such as influence factors [3] [4], perception of user experience [5] [6], dataset construction [7], and objective assessment methods [8] [9].

This paper introduces a method to perform conversational subjective tests based on a video-telephony platform, and it takes a variety of parametric factors that affect media quality into consideration. The research is carried out as part of the 'Computational model used as a QoE/QoS monitor to assess video-telephony services' (G.CMVTQS) project, which is under study in ITU-T SG12 Q.15 [10]. The expected output of this ongoing collaborative project consists of a set of parametric objective quality assessment models that predict the quality of single-channel bidirectional video-telephony calls comprising both audio and video components. In G.CMVTQS project, two types of subjective tests are designed: the audiovisual material subjective test and the conversational subjective test [11]. The audiovisual material subjective test simulates one-way communication, and the conversational subjective test simulates two-way communication. This paper concentrates on the latter. The subjective database generated will be used for the training and validation of a future quality assessment model for video-telephony services.

The experimental setup and test conditions of conversational subjective tests are introduced in this paper. Subjective evaluation is conducted using a video-telephony platform, simulating various network distortions. Related network parameters include delay, jitter, packet loss, and bandwidth. This paper also provides some insights and analysis on commonly-used packet loss recovery mechanisms in video-telephony applications. Moreover, in order to learn the impact of multimedia asynchrony on users' interaction experience during video calls, cases with audio and video asynchrony are included in this test as well.

II. Conversational subjective tests

A. Lab deployment

During conversational subjective tests, two evaluators in different rooms make video calls. The selected

[1]Jose Joskowicz and Mengying Liu are co-first authors.

video-telephony platform for these tests is BigBlueButton, an open-source video conferencing system [12]. The test bed is presented in Fig. 1. Netem [13] and tc commands are applied for the outgoing traffic of the server's network interface, to simulate different network distortions.

Both test rooms should have similar environment conditions and lighting. There should be no distractions around evaluators during test sessions. Also, both test rooms should be equipped with similar test devices. PCs with screen size larger than 12 inches or mobile phones with screen size smaller than or equal to 10 inches can be used as terminal devices. Voice should be played through headsets. External ultra-high-definition cameras are used for video capture with PCs.

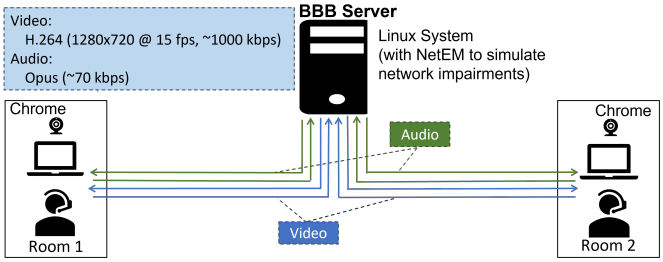


Fig. 1. Test bed of conversational subjective tests

B. Simulation of network distortion

These tests mainly focus on the impact of network distortion on user interaction experience during video calls. Simulated network distortions include delay, jitter, packet loss, bandwidth limitation, and audio and video asynchronism. The parameter specification of active tests is listed in Table I.

Multimedia codec settings such as video resolution, video codec, video frame rate and audio codec are fixed. Delay refers to the Round-Trip Time (RTT) from Bigbluebutton server to clients. Jitter values are set according to added delays. The test values of delay and jitter refer to extra degradations added to the base network. Bandwidth values are determined according to audio and video codec bandwidths. In this test environment, the video codec bit rate is around 1000 kbps for 720p@15fps video streams, and the audio codec bit rate is around 70 kbps. In our case, the bandwidth values cited are the limits we impose to the overall network bandwith.

Audio and video have different robustness regarding packet losses, and they use different mechanisms to protect the transmission quality. Pre-tests were conducted to analyze the impact of various packet loss rates on audio and video qualities in real-time communications (RTC) applications and to find suitable test values for these conversational subjective tests. Pre-tests about audio and video packet loss simulation are introduced in section II-B1.

Asynchronism between audio and video is also a key factor that influences user interaction experience. Discussion on this aspect is introduced in section II-B2. The asynchronism is measured as the time offset between video and audio streams. A negative value indicates that video is ahead of audio, and a positive value indicates that audio is ahead of video.

In G.CMVTQS project, conversational subjective tests are jointly conducted by six participating parties (China Mobile Research Institute - China, Universidad de la República - Uruguay, Universidad de Montevideo - Uruguay, Wuhan University - China, Technische Universität Ilmenau - Germany and Rohde & Schwarz SwissQual AG - Switzerland). In total, 47 different condition combinations are tested in conversational subjective tests. These condition combinations are divided into two groups, one focuses on basic network parameters (packet loss, delay, jitter, bandwidth), and the other focuses on audio and video asynchrony, together with packet loss and bandwidth.

TABLE I
PARAMETER SPECIFICATION OF ACTIVE TESTS

Parameters	Values
Video resolution	1280x720
Video frame rate	15 fps
Video codec	H.264
Audio codec	OPUS
Bandwidth	High = 2.3Mbps, Med = 1.5Mbps, Low = 1Mbps
Delay	Low = 0ms, Med = 200ms, High = 600ms
Jitter	Low = 5% of delay, High = 20% of delay
Packet loss pattern	Random, uniform
Packet loss rate	Audio: None = 0%, Low = 20%, High = 50% Video: None = 0%, Low = 0.5%, High = 3%
Asynchronism between video and audio	None = 0ms Low = -250ms, +250ms High = -500ms, +500ms

1) Multimedia packet loss simulation: When a certain amount of packet losses are added to all outgoing traffics of the server's network interface, it was found that, even with high values of packet loss (up to 25%), there is almost no packet loss in the video stream that reaches the client, and no appreciable video degradation was perceived. The audio packet loss rate detected on the client side is basically consistent with the packet loss rate simulated.

In the video streams received by clients, there are duplicated RTP packets and out-of-order RTP packets. In audio streams received by clients, RTP packets are in order, and some are lost as the result of packet loss simulation. It was found that BigBlueButton has a retransmission mechanism, to protect video quality. If a video packet is lost, the client feeds back relevant information to the server through the NACK (Negative Acknowledgement) message in the RTCP packet. After receiving the NACK message, the server generates a retransmission request accordingly for packet retransmission processing. On the client side, this retransmitted packet arrives out-of-order, but it is still within what the jitter buffer can handle. Those out-of-order packets are reordered before being decoded, so no quality loss is perceived in the client's reconstructed video.

Besides, in order to understand the packet loss recovery strategies of common video conferencing platforms, the performance of Webex, Microsoft Teams, Google Meet and Zoom are analyzed, with packet loss simulation in the downstream network of clients. With up to 20% of packet

loss, there are no perceptible video degradation in all four applications. For Webex, the video starts to show degradations with packet loss higher than 20%. With packet loss up to 30%, video in Microsoft Teams begins to have degradations, and video resolution of Google Meet starts to decrease. For Zoom, it is noticeable that the video resolution begins to drop from 25% of packet loss, but without showing artifacts. It begins to show important artifacts around 50% of packet loss.

The network trace analysis results show that Webex uses a strategy similar to BigBlueButton, with RTP packets out of order and RTCP messages sent very often. Microsoft Teams and Google Meet use the strategy called out-of-band FEC. In this case, except for two typical audio and video streams, there is a third independent RTP stream sent from server to client, with 'RFC2198' (RTP Payload for Redundant Audio Data). For Zoom, the data packets cannot be interpreted since it doesn't use standard RTP. Overall, our analysis on these video conferencing platforms are basically consistent with previous study published by Nisticò et al. [14], which provides more insights comparing different video conferencing tools. Retransmission or redundancy of media packets is very common in current video telephony applications. Thus, even with high packet loss rate in the transmission network, the application on client side actually perceives very little packet loss, and there is no noticeable video degradation.

However, the conversation subjective database aims at including different levels of network degradations perceived by clients, and their corresponding user opinion scores will be used for training of quality assessment models. Therefore, for this project, it was decided to disable video retransmission in BigBlueButton, by skipping the retransmission request generation when the server receives NACK messages. For future quality assessment model, packet loss rates at application level will be part of the model inputs.

With video retransmission disabled, the video packet loss rate detected on the client side is close to the packet loss rate set in simulation. And the video shows degradations as loss rate increases. With 0.5% of video packet loss, video degradations are observable as 'freezing'. With more than 3% of video packet loss, the video scene sometimes severely freezes.

Audio streams use OPUS codec. The original transmission mechanism provided by BBB was used without modifications. Preliminary test results show that there is very little degradation with 10% of audio packet loss. The audio degradation become noticeable when audio packet loss rate is around 30%, and annoying when around 50%.

Audio decoders embed mechanisms for missing audio frames reconstruction, but robustness of video streams against packet loss is rather based on retransmission or redundant data. Considering that audio and video streams have different robustness against packet loss, packet loss simulation are applied to audio and video streams independently, with different test values, as illustrated in Table I. Test values of audio and video packet loss rates are determined through preliminary tests that combines packet loss with other types of network degradation. Medium and High levels of loss rates correspond to noticeable but not annoying impairment and clearly degraded quality respectively.

2) Multimedia asynchronism simulation: In order to better understand the impact of asynchronism between audio and video on user interaction experience in video-telephony scenario, the conversational subjective test includes simulation of multimedia asynchronism, by adding different delays to audio and video streams.

According to ITU-R BT.1359 [15], acceptability thresholds of classical TV contexts were close to 90 ms when video is delayed, and 180 ms when audio is delayed. Saidi et al. presented a subjective audiovisual quality assessment experiment, and results show that the same dissymmetry applies for video-telephony contexts, but with larger acceptability thresholds, rather at least 150 ms and 250 ms respectively [16]. The maximum time offset between audio and video tested in [16] is 400 ms, either video ahead of audio or audio ahead of video. And the corresponding perceptual experience is between 'perceptible but not annoying' and 'slightly annoying'.

Referring to existing studies and combined with preliminary test results, the level of multimedia asynchronism simulation in conversational subjective tests, measured as the difference between video delay and audio delay, can be -500 ms, -250 ms, 0 ms, +250 ms and +500 ms. The user experiences of low and high levels of multimedia asynchronism are 'perceptible but not annoying' and 'slightly annoying' respectively. Specific values of added audio and video delays are listed in Table II. Cases of low asynchrony when both audio and video have delays are also included, in order to learn the impact on perceptual experience.

TABLE II
AUDIO AND VIDEO DELAYS FOR ASYNCHRONISM SIMULATION

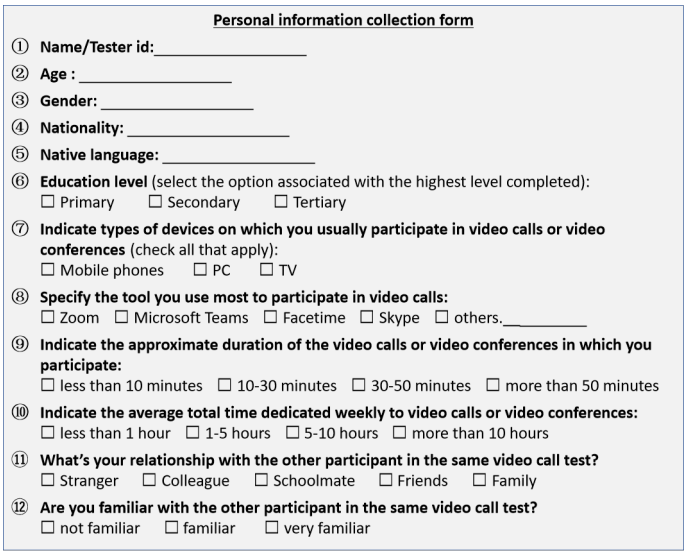
Asynchronism between audio and video	video delay	Audio delay
+500 ms	500 ms	0 ms
+250 ms	250 ms	0 ms
+250 ms	600 ms	350 ms
-250 ms	0 ms	250 ms
-250 ms	350 ms	600 ms
-500 ms	0 ms	500 ms

C. Subjective evaluation

The subjective evaluation procedure is designed according to ITU-T Recommendation P.920 [17], and subjective quality is evaluated using the absolute category rating (ACR) method. Evaluators are 'non-experts' in the field of quality assessment.

Similar to most subjective tests in the field of multimedia, a training session is available at the beginning to help evaluators familiarize themselves with the test operation and the range of quality covered. Before tests, evaluators' personal information is collected for statistical purposes. Evaluators' hearing and vision are also pre-checked. The personal information to be collected includes basic information

(age, gender, nationality, principal language, education level), user habits (commonly-used device types and tools when making video calls in daily life, approximate duration of most video calls, average total time dedicated weekly to video calls) and information about test partner (relationship and familiarity with the other evaluator in the same video-telephony test).The personal information collection form is as shown in Fig. 2.



Personal information collection form

① **Name/Tester id:**_____

② **Age :** _____

③ **Gender:** _____

④ **Nationality:** _____

⑤ **Native language:** _____

⑥ **Education level** (select the option associated with the highest level completed):
☐ Primary ☐ Secondary ☐ Tertiary

⑦ **Indicate types of devices on which you usually participate in video calls or video conferences** (check all that apply):
☐ Mobile phones ☐ PC ☐ TV

⑧ **Specify the tool you use most to participate in video calls:**
☐ Zoom ☐ Microsoft Teams ☐ Facetime ☐ Skype ☐ others._____

⑨ **Indicate the approximate duration of the video calls or video conferences in which you participate:**
☐ less than 10 minutes ☐ 10-30 minutes ☐ 30-50 minutes ☐ more than 50 minutes

⑩ **Indicate the average total time dedicated weekly to video calls or video conferences:**
☐ less than 1 hour ☐ 1-5 hours ☐ 5-10 hours ☐ more than 10 hours

⑪ **What's your relationship with the other participant in the same video call test?**
☐ Stranger ☐ Colleague ☐ Schoolmate ☐ Friends ☐ Family

⑫ **Are you familiar with the other participant in the same video call test?**
☐ not familiar ☐ familiar ☐ very familiar

Fig. 2. Personal information collection form

There are no limitations of the conversational topic, but the content of the conversation must involve the interaction between evaluators. And there should be a good balance between each evaluator's listening time and talking time. In our case, we suggest the evaluators to play 'name-guessing' game during test session. They could also discuss on their own topics if they prefer. This task specification is more realistic in terms of conversation, but it will diminish the sensitivity of the evaluators to delay.

In order to improve evaluators' sensitivity to delay and asynchronism, and to learn the impact on user interaction experience, we suggest to add a 'number-counting' part before 'name-guessing' game. In the 'number-counting' part, two evaluators in the same call take turns counting numbers in order, and use hand gestures to show corresponding number at the same time. Fig. 3 shows the participants performing the 'number-counting', at the start of a session, during the preliminary tests. The 'number-counting' part takes only a few seconds to get a brief impression on delay and asynchronism especially when there are such degradations in this conversation session. Evaluators will then concentrate on the 'name-guessing' game or free speaking for the rest time. Each conversation lasts for three to five minutes.

After the conversation ends, evaluators are asked to rate their opinion scores on the following quality dimensions: the overall quality of this video call, perceptual experience on delay, perceptual experience on asynchronism between image and sound, audio quality and video quality of this video call. Ratings of overall quality, audio quality and video quality use five-grade scale with corresponding quality labels listed in III. Ratings of perceptual experience on delay and asynchronism also use five-grade scale but with reduction labels listed in III.

Fig. 3. Preliminary test session

TABLE III
FIVE-GRADE SCALE WITH CORRESPONDING QUALITY LABEL AND REDUCTION LABEL

Score	Quality label	Reduction label
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

In addition, actual values of key parameters are monitored during each test session. Key parameters include but not limited to frame rate, bit rate, RTT, jitter, packet loss rate, bandwidth of audio and video received by the client side. These values will be extracted from test records such as network traces and specific log files and will be part of the subjective database together with evaluators' opinion scores. RTT and jitter can be obtained by printing the RTCP packet information before being encrypted by SRTCP, and the calculation method refers to RFC 3550 [18].

In order to facilitate the conduction of subjective tests, a script program was developed. For each session, this script automatically adjusts the simulated network conditions (i.e. Netem commands) and starts capturing different kinds of logs (network traces, BBB logs, etc.). When a conversation session needs to be finished, it sends a pop-up window to each participant's test device, reminding them to rate opinion scores of perceived quality, then stops the network degradation simulation and automatically collects all the logs. Rating of opinion scores are performed using web forms on smartphones.

The conversational subjective tests will be jointly conducted by six participating parties. Each party will test 15 conditions with 20 evaluators. Some of these 15 conditions are common

conditions that will be tested by multiple parties. These common conditions are shared as benchmarks to align databases from all parties in the near future.

III. Future Work

The G.CMVTQS project is still under study. Conversational subjective tests introduced in this paper will soon be conducted in the following months, by six laboratories located in different countries. Results of subjective tests from multiple parties will be analyzed and collaboratively form a complete subjective database for training and validation of quality assessment model for video-telephony services.

IV. Conclusion

The conversational subjective test introduced here provides a clear design for subjectively evaluating user interaction experience in video-telephony scenarios, which helps to understand the impact of various network distortions on quality of experience. The collaboration between six laboratories will produce a useful subjective database for the development of new QoE models for video-telephony services. Also, this paper analyzed how video-telephony applications take advantages of packet retransmission and redundant data to fight against packet loss in transmission network and protect multimedia quality perceived by users.

References

[1] J. Skowronek, A. Raake, G. H. Berndtsson et al., "Quality of Experience in Telemeetings and Videoconferencing: A Comprehensive Survey", IEEE Access, vol. 10, pp. 63885-63931, 2022

[2] D. Vučić, S. Barakovic, L. Skorin-Kapov, "Survey on user perceived system factors influencing the QoE of audiovisual calls on smartphones", Multimedia Tools and Applications, Nov. 2022

[3] I. Saidi, L. Zhang, V. Barriac, O. Déforges, "Laboratory and crowdsourcing studies of lip sync effect on the audio-video quality assessment for videoconferencing application", IEEE International Conference on Image Processing (ICIP), pp. 3207-3211, 2019.

[4] F. Schiffner, S. Möller. "Diving into perceptual space: quality relevant dimensions for video telephony.", 8th International Conference on Quality of Multimedia Experience (QoMEX), 2016.

[5] M. Matulin, S. Mrvelj, B. Abramović, T. Šoštarić, M. Čejvan, "User quality of experience comparison between Skype, Microsoft Teams and Zoom videoconferencing tools", International Conference on Future Access Enablers for Ubiquitous and Intelligent Infrastructures (FABULOUS), pp. 299-307, May. 2021.

[6] D. Pal, V. Vanijja, S. Patra, "Online learning during COVID-19: students' perception of multimedia quality", Proceedings of the 11th International Conference on Advances in Information Technology (IAIT), Article No: 27, pp.1-6, Jul. 2020.

[7] R. P. Spang, J. -N. Voigt-Antons and S. Möller, "The Story time Dataset: Simulated Videotelephony Clips for Quality Perception Research," 2022 14th International Conference on Quality of Multimedia Experience (QoMEX), pp. 1-6, 2022

[8] P. Pornpongtechavanich, T. Daengsi, "Video telephony - quality of experience: a simple QoE model to assess video calls using subjective approach", Multimedia Tools and Applications 78, 31987–32006, Jul. 2019.

[9] L. Zhang,I. Saidi, S. Tian, V. Barriac, O. Déforges, "Overview of full-reference video quality metrics and their performance evaluations for videoconferencing application", Journal of Electronic Imaging, 28(2), 023001, 2019.

[10] ITU-T Study Group 12: Performance, QoS and QoE, Q15: Parametric and E-model-based planning, prediction and monitoring of conversational speech and audio-visual quality. Accessed: Apr. 21, 2022. [Online]. Available: https://www.itu.int/net4/ITU-T/lists/q-text.aspx?Group=12&Period=17&QNo=15&Lang=en

[11] M. Liu, J. Joskowicz, R. Sotelo, Y. Hu, Z. Chen and L. Yang, "Subjective Quality Assessment of One-to-One Video-Telephony Services," 2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1-6, 2022

[12] BigBlueButton — Open Source Virtual Classroom Software. Available at: https://bigbluebutton.org.

[13] NetEm — Network Emulator. Available at: https://wiki.linuxfoundation.org/networking/netem.

[14] A. Nisticò, D. Markudova, M. Trevisan, M. Meo and G. Carofiglio, "A comparative study of RTC applications", 2020 IEEE International Symposium on Multimedia (ISM), pp. 1-8, 2020

[15] ITU-R Recommendation BT.1359. Relative timing of sound and vision for broadcasting, 1998

[16] I. Saidi, L. Zhang, V. Barriac, O. Deforges, "Interactive vs. non-interactive subjective evaluation of ip network impairments on audiovisual quality in videoconferencing context", International Conference on Quality of Multimedia Experience (QoMEX), pp. 1-6, Jun. 2016.

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

[18] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, DOI 10.17487/RFC3550, July 2003, https://www.rfc-editor.org/info/rfc3550.