Quality Assessment of Audiovisual Communication in Videotelephony: New ITU-T P.940 and P Suppl.31

Mengying Liu, Lei Yang, Vincent Barriac, Yaosi Hu, Jose Joskowicz, Rafael Sotelo
Shijun Zhang, Zhenzhong Chen, Alejandra Armendariz, Martin Wildbaum, Guillermo Fiorina

***Abstract*—With the rapid development of communication technologies, videotelephony services enabled by consumer technology have gradually become an indispensable means of communication in people's daily life, work, and study. It is of crucial importance to perform scientific and accurate assessments for the quality of experience perceived by consumers. Moreover, with the continuous advancement of media codec and transmission technologies, consumers' expectation for service quality have increased, necessitating an expansion and improvement of the videotelephony quality assessment models or tools. This paper presents subjective datasets collaboratively developed by four labs and investigates objective assessment methods. The subjective datasets are constructed from two types of tests: the non-interactive audiovisual material subjective tests and the interactive conversational subjective tests. For non-interactive subjective test dataset, the following factors' impact on video quality are analyzed: codec type, device type, spatial complexity and temporal complexity. For interactive subjective test dataset, the correlations between interaction experience and influencing factors (audio delay, video delay and network transmission degradation) are analyzed. Furthermore, the subjective test datasets were instrumental in the development of the objective assessment model in the recently published ITU-T Recommendation P.940. This paper provides a comparative analysis between the models recommended in P.940 and G.1070 in terms of video quality, interaction experience and overall videotelephony quality.**

***Index Terms*—QoE, QoS, videotelephony, audiovisual communication, conversational quality, interactive experience, subjective evaluation, objective assessment.**

I. Introduction

VIDEOTELEPHONY has gradually evolved from a conceptual application proposed about a century ago into a service enabled by consumer technology that people frequently use in their daily lives. It plays an important role in remote communication with family and friends, online meeting collaboration, and the implementation of remote education and telemedicine [1]. With the widespread deployment of 4G and 5G networks, the convenience and overall experience of using videotelephony have greatly improved. The number of consumers has shown a rapid growth trend in recent years.

However, with the extensive popularization of videotelephony services, their quality problems have gradually emerged. Affected by various factors such as the condition of the transmission network and the limitations of terminal device performance, videotelephony often experiences quality issues such as insufficient clarity, video or audio freezing, and even disconnection during actual use. These quality defects not only reduce consumers' satisfaction level with videotelephony services, but also impede the in-depth application and expansion of videotelephony services in more fields.

Therefore, it is of great importance to perform scientific and accurate quality assessments for videotelephony services. This will help consumer technology developers and service providers quantify the quality of user experience, analyze and locate the root cause of quality issues based on assessment results, and then optimize technical aspects such as network transmission and media coding in a targeted manner. In this way, the quality of experience perceived by consumers and their willingness to continue using the videotelephony service can be effectively ensured.

Subjective evaluation serves as a core methodology in quality of experience (QoE) research, directly reflecting the quality experienced by consumers through the instant feedback of a certain number of subjects. It is crucial to establish standardized and reliable subjective test methods that provide consistent results. However, constrained by the requirement of resource-intensive human participation, subjective evaluation methods are ill-suited for large-scale quality tests or real-time quality monitoring, creating a critical need for objective assessment models developed or calibrated against subjective datasets.

With the continuous expansion of videotelephony services and the growth of the consumer group, this demand for objective quality assessment models tailored for quality monitoring of audiovisual communication has become increasingly urgent. However, within the international standard framework, there was a long-standing gap in standards for objective quality assessment of real-time audiovisual communication. To address this, a study was carried out by members of

Mengying Liu, Lei Yang and Shijun Zhang are with China Mobile Research Institute, Beijing, China (e-mail: liumengying@chinamobile.com; yangleiyj@chinamobile.com; zhangshijun@chinamobile.com).

Vincent Barriac is with Orange Innovation/Networks, Lannion, France (e-mail: vincent.barriac@orange.com).

Yaosi Hu and Zhenzhong Chen are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China (e-mail: ys_hu@whu.edu.cn; zzchen@whu.edu.cn).

Jose Joskowicz, Alejandra Armendariz and Rafael Sotelo are with the Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay (e-mail: josej@fing.edu.uy; aarmendariz@fing.edu.uy; sotelo@fing.edu.uy).

Rafael Sotelo, Martin Wildbaum and Guillermo Fiorina are with the Facultad de Ingeniería, Universidad de Montevideo, Montevideo, Uruguay (e-mail: rsotelo@um.edu.uy; mwildbaum@correo.um.edu.uy; gfiorina@correo.um.edu.uy).

All authors contributed equally to this work.

Corresponding author: Lei Yang.

the ITU Telecommunication Standardization Sector (ITU-T) Study Group 12 (SG12 - Performance, quality of service and quality of experience) [2] and the Video Quality Experts Group (VQEG), with the following objectives:

- To design and specify subjective methods that evaluate the quality experienced by consumers during audio-visual communication using videotelephony services, with a focus on multiple perspective of multimedia quality and interactive experience.
- To perform and validate the proposed subjective test methods, and collect users' opinion scores to construct the subjective quality datasets for the training and validation of objective quality assessment models.
- To develop no-reference computational models that comprehensively consider multimedia quality and interaction experience to evaluate the overall quality perceived by consumers while making video calls.

The achievements of these objectives has supported the development of the recent ITU-T Recommendation P.940 [3] and P Suppl.31 [4]. P.940 proposes a computational model that can be used for assessing the combined effects of network, media stream, and terminal device related parameters on perceived quality. This model provides estimates of multimedia quality, interaction experience and comprehensive videotelephony quality perceived by users. P Suppl.31 introduces subjective evaluation methods that focus on user experience of real-time audiovisual communication mainly affected by system-related factors. This paper presents details of the subjective tests performed following the methods described in P Suppl.31 and analyzes the key influencing factors of user experience quality in multiple perspectives based on the subjective datasets established. It also evaluates the performance of the P.940 model and makes a comparative analysis with the G.1070 model [5] using the same subjective datasets.

The rest of the paper is structured as follows. Section II provides an overview of related works. Section III provides a detailed description of the subjective tests, and analysis of the subjective test datasets. Section IV introduces objective models proposed by ITU-T Recommendations G.1070 and P.940, and presents a comparative analysis based on the subjective datasets. Finally, Section V and VI expose the main conclusions of the paper and future work.

II. Related work

A. Videotelephony technologies

When making a video call, the connection between two users is initially established using Web Real-Time Communication (WebRTC) [6] and Session Initiation Protocol (SIP) [7], with the media coding formats and transmission-related parameters negotiated at this stage.

Common video coding formats currently used for videotelephony services include H.264/AVC and H.265/HEVC. H.264 is well-known for its high compression efficiency and robust network adaptability [8], and has been widely adopted in various videotelephony scenarios and other consumer technology products. As an evolution of H.264, H.265 achieves approximately 50% higher compression efficiency for equal perceptual video quality, requiring less bandwidth and being particularly well-suited for high-definition and ultra-high-definition videos [9].

Opus is one of the common audio coding formats used for videotelephony services. It is an open-source audio coding format standardized by the Internet Engineering Task Force (IETF) as RFC 6716 [10]. Opus has characteristics such as a wide bit rate range, low latency, strong adaptability, and excellent packet-loss resistance performance. It is the default audio coding format in the WebRTC framework and is widely adopted in numerous real-time audiovisual communication applications. The quality comparison between Opus and other audio codecs can be found in the Opus website [11] and relevant publications [12], [13].

Videotelephony services typically use User Datagram Protocol (UDP), Real-time Transport Protocol (RTP) and RTP Control Protocol (RTCP) for the transmission of audio and video data packets. UDP offers the advantages of low latency and high transmission efficiency, but cannot guarantee reliable data transmission [14]. Data losses can be compensated by mechanisms such as Forward Error Correction. RTP is a transmission protocol specially designed for real-time applications. It provides mechanisms such as timestamps and sequence numbers, enabling accurate sequencing and timing of data packets, thereby ensuring the correct decoding of received media [15]. The primary function of RTCP is to monitor and provide feedback on the RTP transmission process [15]. RTCP packets contain statistical information such as packet count, fraction lost, delay and interarrival jitter, which allows the sender to adjust its transmission strategy in a timely manner based on the network conditions.

B. Quality of experience and subjective evaluation methods

Quality of Experience (QoE) is defined as "the degree of delight or annoyance of the user of an application or service" in ITU-T Recommendation P.10/G.100 [16]. The QoE of videotelephony services is perceived as a complex cognitive construct. The variety of influencing factors are generally divided into three main categories: human influence factors (HIFs), system influence factors (SIFs) and context influence factors (CIFs) [17].

As audio-visual convergent multimedia services, the assessment of QoE generally begins with the evaluation of user-perceived audio and video quality, which is closely related to the physiological characteristics of the human visual system (HVS) and the human auditory system (HAS). These single-modal quality issues influence the subjective evaluation of integrated audiovisual quality through cross-modal perceptual interaction effects [18].

Furthermore, as one of the most important characteristics of videotelephony services, real-time interaction is primarily determined by end-to-end latency, which impacts the communication efficiency, the naturalness of conversation and the interpersonal coordination [19], [20]. Generally, the latency should be kept below human perception threshold to approximate face-to-face interaction. High latency not only introduces perceivable conversation delays, but may also lead to interaction failures such as speech gaps or overlaps [21].

Another important factor affecting user experience and immersion in audiovisual communications is audiovisual synchronization or lip-sync [22]. The differences between audio and video in aspects such as coding processing and network transmission requirements may lead to offsets in their end-to-end delays. Most videotelephony services leverage timestamps or buffer management to achieve audiovisual synchronization at the user-perceptual level. Notably, users exhibit asymmetric thresholds of perceptibility for audio ahead of video and video ahead of audio conditions [23]–[25].

ITU has built a comprehensive subjective evaluation standard framework for multimedia services, encompassing different quality dimensions (image quality, audio quality, video quality, audiovisual quality, etc.), test scenarios (non-interactive and interactive), and test environments (controlled/in-lab and uncontrolled/crowdsourcing). Some studies also dedicate to further improving subjective evaluation methods [26], [27], or exploring appropriate methods for emerging multimedia services and devices [28], [29], so as to keep up with the evolving technological trends. Moreover, the results of subjective experiments are not only core outcomes and important assets of researches related to QoE/QoS evaluation, but also the foundation for analyzing the impact of key influencing factors [30], [31] and developing objective assessment models.

C. Objective assessment methods

Objective assessment methods are categorized into three types based on their dependency on the knowledge of the original signal: full-reference (FR), reduced-reference (RR) and no-reference (NR) [32]. FR methods can precisely quantify media quality through direct comparison with the original signal, with the most commonly-applied approaches including PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index Measure) and VMAF (Video Multimethod Assessment Fusion) [33] for video, and PESQ (Perceptual Evaluation of Speech Quality), POLQA (Perceptual Objective Listening Quality Analysis) for audio [34]. But their practical application scenarios are very limited, being confined only to closed-loop scenarios where the original signal can be fully obtained. RR methods require access to specific features of the source signal, which makes them slightly more applicable than FR methods [35]. However, RR methods are still not suitable for scenarios where the original signal is confidential, such as services involving user-generated content (UGC). NR methods are completely independent of the original signal, offering superior scalability and cost-efficient deployment. [36]

For NR methods, most recent studies have focused on machine learning or deep learning approaches based on video content, such as [37]–[40]. Additionally, some studies have explored QoE estimation approaches using facial features and speech features, based on data collected from audiovisual conversations [41], [42]. Another critical type of approach is parameter-based objective assessment models. Videotelephony services involve the transmission and sharing of users' voice and portrait information. Due to the strict privacy and data protection regulations, parametric-based NR models are more suitable for quality monitoring of videotelephony services.

The well-known parameter-based NR model in the videotelephony industry is the opinion model proposed in ITU-T Recommendation G.1070, which evaluates multimedia quality by integrating audiovisual media quality and audiovisual delay impairment factor [5]. Although some studies investigated its application in quality monitoring [43], it is stated that the application of G.1070 model is limited to QoE/QoS planning. Over the past years, videotelephony services have rapidly evolved toward higher complexity and real-time interactivity, making fine-grained evaluation of user experience a core requirement. Service providers can leverage quality assessment results to identify deployment flaws and optimize resource allocation, thereby continuously enhancing QoS. Against this backdrop, models for quality monitoring can better support dynamic adjustment of resource allocation and codec configurations, enabling rapid quality enhancement and user experience assurance. To address these needs, the research presented in this paper is conducted, leading to the development and publication of ITU-T Recommendation P.940, a computational model used for the monitoring and quality assessment of videotelephony services [3].

Moreover, objective assessment methods can support related research on quality perception-based coding or transmission optimization technologies or strategies, contributing to the improvement of QoS and QoE [44], [45].

III. Subjective videotelephony quality assessment

A. Subjective test scheme

In consideration of the characteristic of real-time interaction in videotelephony services, the subjective quality assessment encompasses two complementary sets of tests: the non-interactive audiovisual material subjective tests and the interactive conversational subjective tests. These two sets of tests are designed with reference to ITU-T Recommendation P.910 [46] and P.920 [47], respectively. The following subsections describe these subjective tests and analyze the key influencing factors, mainly for video quality, audiovisual quality, and interaction experience in terms of latency and synchronization.

1) Non-interactive audiovisual material subjective tests: The non-interactive audiovisual material subjective tests simulate the scenario where users uni-directionally receive audio and video. They focus on the impact of media coding technologies and media playback on quality perceived by consumers.

The experimental design of non-interactive tests is consistent with that described in ITU-T P Suppl.31 [4] and [48] . The source sequences are high-quality videos with a duration of approximately 10 seconds, captured by cameras and microphones commonly used in videotelephony. They emulate typical daily videotelephony scenarios, mainly presenting head-and-shoulders images, and covering four representative scenarios: home, office, restaurant and outdoor. Fig. 1 presents samples of these four scenarios (the images have been blurred for the protection of facial information privacy). The audio and video coding impairments were simulated by utilizing the FFMPEG tool, thus generating test sequences with varying quality levels. The factors investigated include video resolution, video frame rate and video bit rate for video in H.264

(Baseline) and H.265 (Main) codec format, and audio bit rate for stereo audio in Opus codec format, as shown in Table I.

Fig. 1. Four representative scenarios included in the non-interactive audiovisual material subjective tests: home, office, restaurant, outdoor.

TABLE I
PARAMETER INVESTIGATED IN NON-INTERACTIVE AUDIOVISUAL MATERIAL SUBJECTIVE TESTS [4], [48]

	Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**
Video codec	H.264	H.264 H.265	H.264 H.265	H.264 H.265	H.265
Video resolution	240p	480p	720p	1080p	2160p (4K)
Video bit rate (bps)	200, 300k	375, 512, 750k	1, 1.5, 2M	2, 5, 10M	6, 10, 20, 30M
Frame rate (fps)	15, 30	15, 30	15, 30	15, 30, 60	30, 60
Audio codec	Opus	Opus	Opus	Opus	Opus
Audio bit rate (bps)	32, 48k	32, 48, 64k	48, 64, 128k	64, 128k	64, 128, 192k
Audio channels	2 (stereo)	2 (stereo)	2 (stereo)	2 (stereo)	2 (stereo)

These test sequences were played on different types of terminal devices, including mobile phones, PCs and TVs, to collect subjective opinion scores on audio quality, video quality and audiovisual quality using the ACR-HR (Absolute Category Rating – Hidden Reference) method. The speech of test sequences are in English. Non-interactive tests were conducted by China Mobile with native English speakers, and Wuhan University with local Chinese students familiar with English.

2) Interactive conversational subjective tests: The interactive conversational subjective tests aim to emulate the scenario of two users making a video call. The emphasis is on the influence of network impairment factors on quality perceived by users, specifically, the interaction experience and the overall quality during video calls.

The experimental design of interactive tests is the same as that described in ITU-T P Suppl.31 [4] and [49], [50]. Specifically, the video call connection was established between the Chrome browsers of two PCs through the open-source WebRTC-based platform BigBlueButton [51]. During the test process, subjects were grouped in pairs and sat in two separate test rooms with the same environment and devices, as illustrated by Fig. 2. The test environment is quiet, without interference or distraction. The conversation task for subjects was 'name-guessing' game, and they could also have free conversation on topics of interest, provided that there is a good balance between each subject's listening and speaking time. In addition, to enhance subjects' sensitivity to delay and synchronization experience, subjects were required to first conduct a turn-taking number-counting session which lasts about 10 seconds at the start of the 3-minute-long conversation. After each conversation, subjects were invited to provide feedback on the difficulty of interacting with and interrupting their partners, and whether they encountered any other issues during the video calls.

The codec configurations of audio and video streams were fixed, aiming to focus on the impact of network impairment factors including packet loss, delay, jitter and bandwidth. Table II lists the media codec configurations and network impairments simulated by using Netem [52] and tc commands [53] in the basic network environment, which is a wired connection network specifically dedicated to interactive tests.

Fig. 2. Example of test environment in interactive conversational subjective tests.

The video packet retransmission mechanism for mitigating packet loss was turned off, so that the simulated packet loss rates could be controlled and similar to the final packet loss rates at the application level. Moreover, it is worth noting that due to the embedded forward error correction mechanism, Opus exhibits good packet loss resistance characteristics [54], [55]. Given the differences in packet loss resistance mechanisms and performances between audio and video streams, in order to comprehensively test various degrees of impact that packet loss has on the quality of user experience, different packet loss rates were deliberately set for audio streams and video streams, with the simulated packet loss rate on audio stream being higher. In addition, the asynchronism between video and audio listed in Table II is measured as the difference between video and audio, where a positive value indicates audio ahead of video, and a negative value indicates video ahead of audio.

Interactive tests were performed by four laboratories using the ACR (Absolute Category Rating) method, and subjects were local people who speak local native languages: China Mobile (China, Chinese), Wuhan University (China, Chinese), Universidad de la República (Uruguay, Spanish) and Universidad de Montevideo (Uruguay, Spanish). Each group of subjects engaged in an online audiovisual conversation lasting approximately 3 minutes under each test condition. Their subjective opinion scores on the overall videotelephony

TABLE II
PARAMETERS INVESTIGATED IN INTERACTIVE CONVERSATIONAL SUBJECTIVE TESTS [4]

Parameter	Settings	Values
Video resolution	Fixed	720p
Video codec	Fixed	H.264 (Baseline)
Audio codec	Fixed	Opus
Video frame rate	Fixed	15 fps
Packet loss pattern	Fixed	Random, uniform
Packet loss rate	None	Audio: 0%, Video: 0%
	Low	Audio: 20%, Video: 0.5%
	High	Audio: 50%, Video: 3%
Bandwidth	High	3000kbps
	Med	2300kbps
	Low	1500kbps
Delay	Low	0ms
	Med	200ms
	High	600ms
Jitter	Low	5% of delay
	High	20% of delay
Asynchronism between video and audio	None	0ms
	Low	-250ms, +250ms
	High	-500ms. +500ms

quality, audiovisual interaction delay experience, audiovisual media synchronization experience, video quality and audio quality were collected at the end of each conversation.

B. Subjective test datasets

After completing subjective tests, the method for post-experimental screening of subjects using Pearson linear correlation, which is recommended in ITU-T P.910 [46] was applied, ensuring valid ratings from at least 15 subjects for each test sequence or test condition tested by a laboratory. The final test results of non-interactive audiovisual material subjective tests and interactive conversational subjective tests form two subjective test datasets, respectively.

1) Statistical information of subjects: Subjective tests were jointly conducted by multiple laboratories, which ensured the diversity of subjects and thus enabled the collected data more representative. In general, the subjects were aged between 19 and 60, including those recruited from society as well as students and faculty/staff recruited from universities who were not involved in this study.

Based on the questionnaire collected before the start of subjective test sessions, 63% of the subjects were male and 37% were female. Statistical information related to experience of using video calls is shown in Fig. 3, specifically including average duration of each daily video call, and approximate weekly duration of video calls.

2) Non-interactive subjective test dataset: The non-interactive subjective test dataset includes subjective scores from 446 test sequences tested on mobile phones, 513 test sequences tested on PCs, and 248 test sequences tested on TVs. Audio-only quality, video-only quality and audiovisual quality were tested independently. China Mobile (CMCC) and Wuhan University (WHU) performed non-interactive tests, using the same test sequences and terminal devices that are similar in terms of screen size and resolution. Fig. 4 shows the distributions of MOS values for audio quality, video quality and audiovisual quality rated in non-interactive tests.

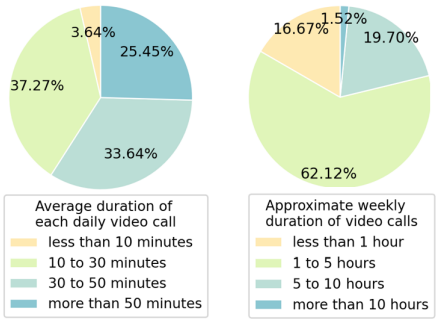


Fig. 3. Statistical information related to subjects' experience with video-telephony services.

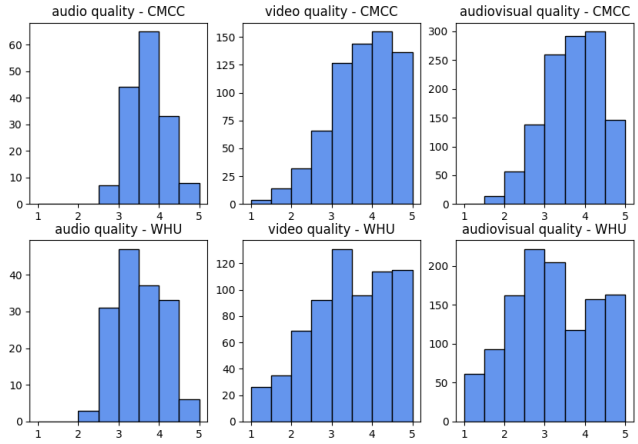


Fig. 4. Distribution of the MOS values of non-interaction audiovisual material subjective tests.

Based on the analysis of audio quality data, it is revealed that in the non-interactive tests, when there are no other degradations such as packet loss, the MOS for audio quality is more influenced by recording scenarios and devices than by audio bit rate. Across test sequences from various recording scenarios, when the audio codec format is Opus and the bit rate ranges from 32kbps to 192kbps as listed in Table I, there are no notable differences in the audio quality perceived by users. Audio quality disparities illustrated in the MOS distributions are primarily attributed to recording scenarios and devices, with key considerations including background noise, the capability of recording devices in terms of sound capture and noise reduction. Modeling of these factors is generally associated with the features of audio streams and devices, and helps to further improve the accuracy of objective assessment for audio quality. However, due to privacy restrictions and protections, it is difficult to obtain such information in real-world videotelephony services.

Regarding the video quality, the Spearman correlation between the MOS values from CMCC and WHU is 0.915, indicating that the test results obtained by two laboratories are consistent, and the test methods and procedures adopted possess high reliability and stability.

It is already a wide recognized consensus that video quality is susceptible to significant influences from parameters such as resolution, frame rate, bit rate and packet loss rate. In

addition, based on the video quality data obtained from non-interactive tests, we analyzed the influence of video codec formats and terminal device types on the video quality perceived by uses, by applying the Mann-Whitney's U test. Table III presents the p-values corresponding to the data subset of H.264(Baseline) relative to that of H.265(Main), as well as p-values between the data subsets of different device types. To ensure a fair comparison, only the test sequences shared between two corresponding data subsets are included. It can be seen from the results shown in Table III that the video codec format has a significant impact on the video quality. The device type generally also has a significant impact on the video quality. Nevertheless, it is worth noting that for videos with a resolution of no less than 720p, there is no significant difference in the video quality perceived by users on mobile phones and personal computers.

TABLE III
p-VALUES FOR ANALYZING THE IMPACT OF DIFFERENT VIDEO CODEC FORMATS AND TERMINAL DEVICE TYPES ON VIDEO QUALITY IN NON-INTERACTIVE TESTS

Perspectives	***p*-value (CMCC data)**	***p*-value (WHU data)**
H.264(Baseline) vs. H.265(Main)	0.0010	0.0001
Mobile phone vs. PC (resolution<720p)	0.0001	0.0069
Mobile phone vs. PC (resolution≥720p)	0.0921	0.4821
Mobile phone vs. TV (resolution≥720p)	1.55e-07	0.0110
PC vs. TV (resolution≥720p)	0.0006	0.0458

Additionally, the non-interactive subjective test dataset further verified the correlation between video quality and temporal information (TI) and spatial information (SI). TI and SI are measures that indicate the number of temporal changes of a video sequence, and the amount of spatial detail in a picture, respectively, as defined in ITU-T Recommendation P.910 [46]. Taking 1080p sequences as an example, we first calculated the average values of MOS for different sequences under the same test condition, by taking both CMCC data and WHU data into consideration. Then, based on the TI and SI of the source sequences, three groups of test sequences were selected to represent the scenario with high-TI and high-SI, the scenario with low-TI and high-SI, and the scenario with low-TI and low-SI. The MOS values of these three groups of test sequences were compared with the corresponding average MOS for different sequences tested under the same test condition. As shown in Fig. 5, when the resolution is 1080p and the video bit rate ranges from 2Mbps to 10Mbps, test sequences in the scenario with high TI and high SI generally exhibit lower video quality than the corresponding average MOS, while test sequences in the scenarios with low-TI generally show higher video quality than the corresponding average MOS, especially when sequences are displayed on TVs.

Further analysis reveals that the impact of TI and SI is also related to video codec configurations such as bit rate, frame rate and resolution. A higher SI indicates more complex intra-frame details, and a higher TI indicates more complex inter-frame motion information. Both situations require relatively higher bit rates to ensure the video quality after encoding and decoding. Therefore, under limited bit rates, sequences with high spatial and temporal information suffer greater information loss due to compression, leading to lower quality than other sequences. Conversely, sequences with low spatial and temporal information have less compression loss, resulting in slightly better quality. Moreover, incorporating features related to spatial and temporal information can help improve the accuracy of parametric evaluation of video quality, especially for sequences with the same parameters but different scenarios. However, due to user privacy protection and data confidentiality regulations, independent third-party quality assessment or monitoring tools cannot access or parse video content information in the media transmission pipeline. Thus, they are unable to directly extract spatiotemporal complexity features from video contents. Future research could explore non-intrusive feature extraction approaches to indirectly extract features related to spatiotemporal complexity.

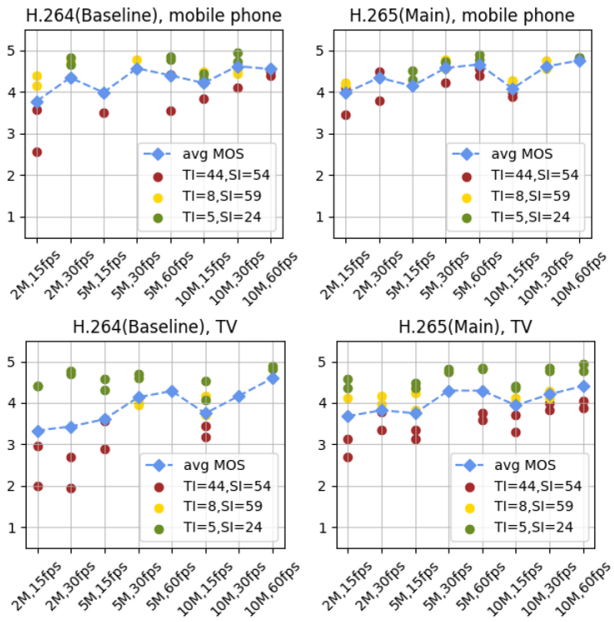


Fig. 5. Comparison between the average MOS across different sequences and MOS values of test sequences in scenarios with high and low levels of TI/SI. The x-axis represents the test conditions of the comination of bit rate (bps) and frame rate for 1080p resolution.

The Spearman correlation between the MOS values of audiovisual quality from CMCC and WHU is 0.905, indicating that the test results are consistent. We also analyzed the impact of different terminal types on perceived audiovisual quality based on the data from the non-interactive subjective test dataset, using Mann-Whitney's U test. Mobile phones and PCs are equipped with headphones as the audio testing device, while TVs use their built-in speakers for audio testing. The results in Table IV show that terminal types have a significant effect on audiovisual quality.

3) Interactive subjective test dataset: The interactive subjective test dataset includes subjective scores for 43 different

TABLE IV
p-VALUES FOR ANALYZING THE IMPACT OF DIFFERENT TERMINAL DEVICE TYPES ON AUDIOVISUAL QUALITY IN NON-INTERACTIVE TESTS

Perspectives	***p*-value (CMCC data)**	***p*-value (WHU data)**
Mobile phone vs. PC (resolution<720p)	1.41e-09	2.15e-05
Mobile phone vs. PC (resolution≥720p)	0.0036	0.0145
Mobile phone vs. TV (resolution≥720p)	1.46e-08	1.37e-08
PC vs. TV (resolution≥720p)	0.0099	0.0013

test conditions collaboratively tested by four laboratories, generating 1151 rows of data. Considering laboratory resources and overall test duration, some conditions were evaluated by multiple laboratories as common conditions, while others were tested by a single laboratory. The participating laboratories are CMCC, WHU, Universidad de la República (UdelaR) and Universidad de Montevideo (UM). All laboratories adhered to active test protocols for setting up test rooms, test devices and platform configurations, ensuring consistency across test environments.

The active test design encompassed network impairments of varying levels, ranging from single-type to multi-type combined impairments. The interactive tests strove to maintain a balanced distribution of test conditions across different network impairment levels and types. Fig. 6 shows the distribution of the opinion scores of all subjects after post-screening.

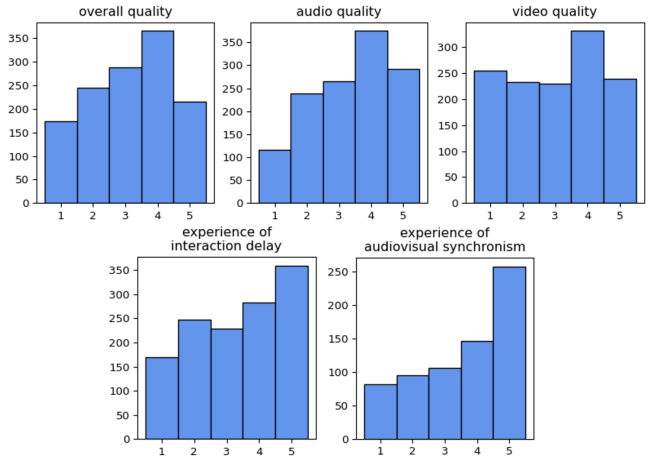


Fig. 6. Distribution of the opinion scores of all subjects in interactive conversational subjective tests.

Based on the interactive subjective test dataset, we analyzed the Spearman correlations between network parameters and quality dimensions, including audio quality (Q_audio), video quality (Q_video), interaction delay experience (Q_delay), interaction synchronization experience (Q_sync) and overall quality (Q_overall). As shown in Fig. 7, the heatmap visually demonstrates a strong positive correlation between overall quality and other quality dimensions.

According to Fig. 7a, audio quality in interactive tests exhibits negative correlations with audio packet loss rate and audio delay, and positive correlations with available bandwidth and audio bit rate. Among these, audio packet loss rate and bit rate are the most strongly correlated parameters with audio quality. Fig. 7b presents that video quality in interactive tests shows negative correlations with video packet loss rate, delay and jitter, and positive correlations with available bandwidth and video bit rate. Under the situation that video codec settings are fixed in the interactive test environment, video packet loss rate, delay and jitter have a more pronounced impact on video quality. These findings align closely with industry consensus on how network impairments affect audio quality and video quality, validating the relationship between key network parameters and user experience.

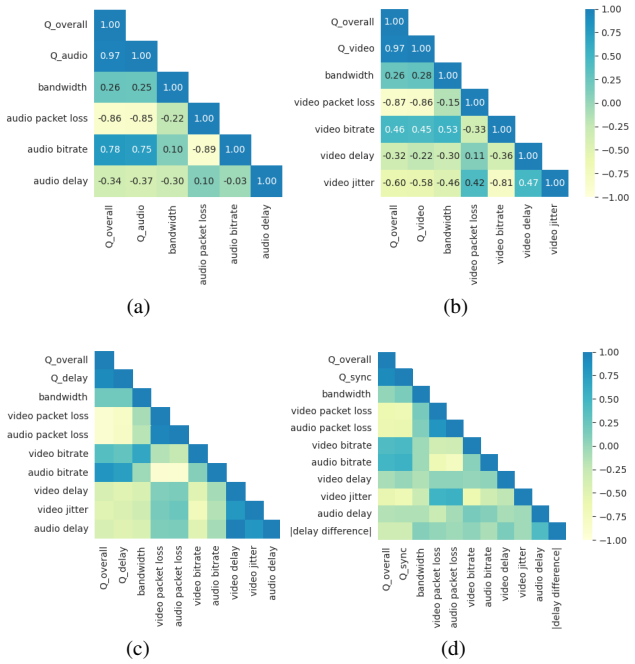
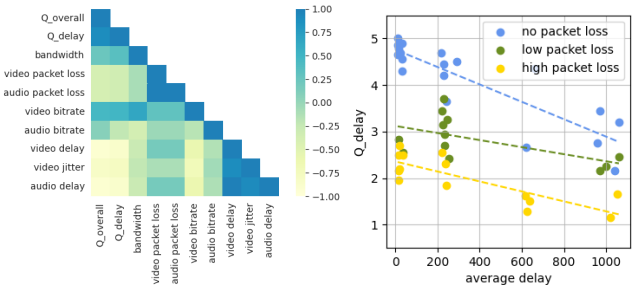


Fig. 7. Spearman correlation analysis of (a) audio quality, (b) video quality, (c) interaction delay experience, and (d) media synchronization experience, respectively, with the overall quality and relevant key network parameters

When test conditions involving packet loss are excluded for analysis, as illustrated by Fig. 8a, the quality of interaction delay experience shows a strong negative correlation with media delays measured at the network level ($SRCC \approx 0.87$). However, when packet loss rates are included in the analysis, as shown in Fig. 7c, the quality of interaction delay experience correlates more strongly with packet loss rates ($SRCC \approx 0.82$) than with delays in the transmission network ($SRCC \approx 0.49$). Fig. 8b reveals that user-perceived interaction delay is not only closely correlated with end-to-end media delays in the network, but also linked to packet loss rates. When packet loss occurs, media data may experience either delayed arrival (due to retransmission mechanism) or severe data loss. A Jitter buffer is designed to absorb the jitter in the arrival time of data packets, mainly by buffering received data packets and reordering them before decoding [56]. If a delayed packet arrives within the processing capacity of the jitter buffer, the playback of the corresponding media content will be delayed as well, because of the extra waiting time in the jitter buffer

and decoder procedures. If the delayed packet exceeds the processing capacity of the jitter buffer, causing buffer overflow, this packet will be discarded, which leads to decoding failure or media distortion such as media freezing. Both scenarios introduce additional latency perceptions at the user level, highlighting the complex mapping between technical-layer delays and user experience on interaction delay.



(a) (b)

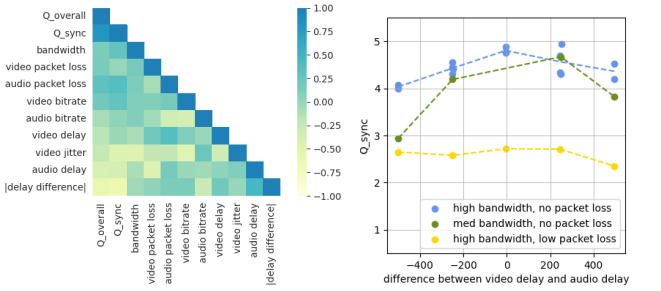
Fig. 8. (a) Spearman correlation analysis of interaction delay experience with the overall quality and relevant key network parameters (test conditions involving packet loss are excluded). (b) Analysis of the relationship between interaction delay experience and average delay of audio and video streams under different levels of packet loss.

In audiovisual communication, media synchronization can be categorized into two scenarios: audio ahead of video, and video ahead of audio. In interactive tests, we simulated media asynchronism by introducing different network transmission delays to audio and video streams. The asynchronism is quantified using the delay difference (video delay - audio delay), where positive values indicate audio ahead of video, and negative values indicate video ahead of audio.

Fig. 9a shows that when test conditions involving packet loss are excluded for analysis, the quality of user experience in terms of media synchronization correlates closely with the absolute difference between video delay and audio delay ($SRCC$ ≈0.64). In contrast, Fig. 7d demonstrates that when packet losses are taken into consideration, media synchronization experience exhibits a higher correlation with packet loss rates ($SRCC$ ≈0.63) than with the absolute delay difference ($SRCC$ ≈0.32). Fig. 9b further analyzes the relationship between media synchronization experience and the asynchronism (measured as the delay difference), under different network conditions. Under high and medium bandwidth without packet loss, the interactive subjective test dataset confirms that asynchronism impacts user experience asymmetrically between the audio-ahead-of-video and the video-ahead-of-audio scenarios. In cases where packet loss occurs, media freezes or distortion become the primary drivers of experience evaluation. For example, if video freezes due to packet loss but audio continues playing with lower quality because of better packet loss robustness, users will perceive extreme video lag. Such audiovisual mismatch or asynchronism attenuates the actual effect of technical-layer delay differences on media synchronization experience.

IV. Objective Videotelephony Quality Assessment

In videotelephony and videoconferencing services, both the audio and video content involve users privacy. Due to privacy



(a) (b)

Fig. 9. (a) Spearman correlation analysis of media synchronization experience with the overall quality and relevant key network parameters (test conditions involving packet loss are excluded). (b) Analysis of the relationship between media synchronization experience and the difference between video delay and audio delay, under different network conditions

protection and data encryption requirements, the optimal solution for real-world quality monitoring of these services is to use no-reference objective quality assessment models based on parametric features.

Prior to the publication of ITU-T Recommendation P.940 [3], G.1070 model was sometimes utilized to address needs for objective quality assessment and monitoring, primarily due to the absence of dedicated standard models designed specifically for quality monitoring. G.1070 proposed an opinion model that estimates the videotelephony quality for QoE/QoS, but its scope is limited to network planners, not to quality monitoring [5]. P.940 model can be applied to quality monitoring scenarios, and provides not only overall videotelephony quality but also fundamental perspectives of user experience quality. It can help testers, monitors and service providers to pinpoint weak links in user experience, analyze causes of quality degradation and support optimizations for QoE/QoS. It is necessary to acknowledge the shared limitations of the G.1070 and P.940 models. Both are applicable to head-and-shoulder scenes and recommended parameter ranges, while their generalization to other scenarios involving different content types and codecs requires further research and validation. Furthermore, although video content complexity information is beneficial for video quality assessment, such information is difficult to be obtained due to regulations on user privacy and data protection. Therefore, neither model has incorporated video content complexity features into quality assessment at this stage.

The following sections briefly introduce the well-known G.1070 model, describe the new P.940 model, and provide a comparative analysis of their performance.

A. ITU-T G.1070 model

The G.1070 model is composed of audio and video quality estimation functions, and a multimedia quality integration function.

1) Audio and video quality estimation functions: The speech quality S_q is evaluated using the E-model defined in ITU-T Recommendation G.107 [57] for narrowband speech, and the E-model defined in ITU-T Recommendation G.107.1

[58] for wideband speech. The video quality V_q is evaluated as follows:

$$V_q = 1 + I_{coding} \exp\left(-\frac{Ppl_V}{D_{PplV}}\right) \tag{1}$$

where I_{coding} is the basic video quality affected by coding distortion, and D_{PplV} is the packet loss robustness factor.

The parameters used for the calculation of basic video quality affected by coding I_{coding} are video frame rate Fr_v and video bit rate Br_v. The calculation is expressed as:

$$I_{coding} = I_{O_{fr}} \exp\left\{-\frac{(\ln(Fr_V) - \ln(O_{fr}))^2}{2D_{FrV}^2}\right\} \tag{2}$$

where

$$O_{fr} = v_1 + v_2\, Br_V, \quad 1 \le O_{fr} \le 30 \tag{3}$$

$$I_{Ofr} = v_3 - \frac{v_3}{1 + (\frac{Br_V}{v_4})^{v_5})} \tag{4}$$

$$D_{FrV} = v_6 + v_7 Br_V, \quad 0 < D_{FrV} \tag{5}$$

The calculation of the packet loss robustness factor D_{PplV} also relies on video frame rate Fr_v and video bit rate Br_v, and is expressed as:

$$D_{PplV} = v_{10} + v_{11} \exp(-\frac{Fr_v}{v_8}) + v_{12} \exp(-\frac{Br_v}{v_9}), 0 < D_{PplV} \tag{6}$$

G.1070 recommends different values for constants v_1 to v_{12} depending on video codec format, video resolution, key frame interval and device display size.

2) Multimedia quality integration function: The overall multimedia quality MM_q is calculated by taking both audiovisual quality and multimedia delays into consideration. MM_q is expressed as:

$$MM_q = m_1 MM_{SV} + m_2 MM_T + m_3 MM_{SV} MM_T + m_4 \tag{7}$$

where MM_{SV} represents the audiovisual quality, and MM_T represents the audiovisual delay impairment factor, which is calculated using speech delay T_S and video delay T_V. MM_{SV} and MM_T are expressed as follows:

$$MM_{SV} = m_5 S_q + m_6 V_q + m_7 S_q V_q + m_8 \tag{8}$$

$$MM_T = \max\{AD + MS, 1\} \tag{9}$$

$$AD = m_9(T_S + T_V) + m_{10} \tag{10}$$

$$MS = \min\{m_{11}(T_S - T_V) + m_{12}, 0\}, \quad if \quad T_S \ge T_V \tag{11}$$

$$MS = \min\{m_{13}(T_V - T_S) + m_{14}, 0\}, \quad if \quad T_S < T_V \tag{12}$$

The recommended values for constants m_1 to m_{14} are different depending on the video display size and conversational task.

B. ITU-T P.940 model

The new ITU-T Recommendation P.940 proposes a computational model that can be used for estimates of multimedia quality, interaction experience and comprehensive videotelephony quality perceived by users. P.940 model can be used for quality monitoring of videotelephony services, or be used by service providers to self-test to further improve the QoE and QoS. The model layout is shown in Fig. 10. The input parameters are categorized into five groups: I.11 primarily includes audio coding quality-related parameters, I.12 focuses on video coding quality-related parameters, I.13 consists of terminal-related factors, I.14 covers network transmission impairment parameters (packet loss rates and interarrival jitter), and I.15 involves audiovisual interaction information (audio delay and video delay). The assessing blocks and outputs are described in the following subsections.

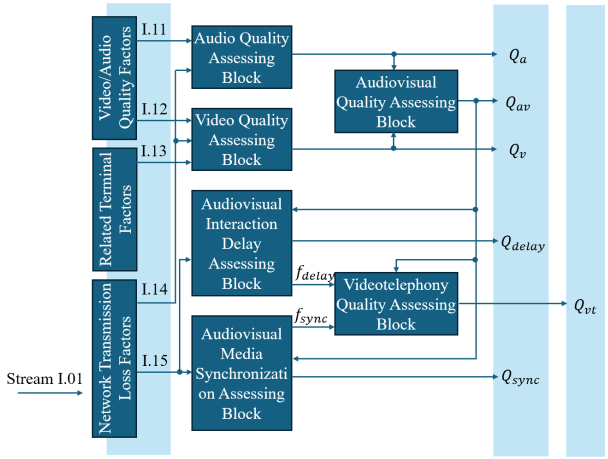


Fig. 10. The model layout of the computational model proposed by ITU-T Recommendation P.940.

1) Audio and video quality assessing blocks: The P.940 model adopts the fullband E-model defined by ITU-T Recommendation G.107.2 [59] as a provisional method for audio quality Q_a assessment. For the Opus codec format, it specifically uses the equipment impairment factor and the packet loss robustness factor derived in [60].

The video quality Q_v is calculated as follows:

$$Q_v = 1 + Q_{basic} f_{network} \tag{13}$$

where Q_{basic} represents the basic video quality affected by video codec impairment and terminal display, and $f_{network}$ represents the impact of network impairment on user perceived video quality.

The calculation of basic video quality Q_{basic} consists of a preliminary evaluation of the video images and the impact of video frame rate on video smoothness. The input parameters are video bit rate Br_v, video frame rate Fr_v, video resolution width R_w and height R_h, and screen resolution width R_{sw} and height R_{sh} of the terminal device. Q_{basic} is expressed as:

$$Q_{basic} = \left(1 + v_1 - \frac{v_1}{1 + (f_1 f_2)^{v_2}}\right) \cdot f_3 \tag{14}$$

$$f_1 = v_3 \ln \frac{Br_v}{Fr_v} \tag{15}$$

$$f_2 = 1 - \exp\left(\frac{v_7}{Scale}\right), \quad if \quad R_w R_h < R_{sw} R_{sh} \tag{16}$$

$$f_2 = 1 - \exp\left(v_8 Scale\right), \quad if \quad R_w R_h \geq R_{sw} R_{sh} \tag{17}$$

$$f_3 = 1 - \exp\left(v_9 Fr_v\right) \tag{18}$$

The video quality estimation module of P.940 model takes video resolution as part of the input parameters, and introduces the $Scale$ factor to indicate the scaling relationship between video resolution and screen resolution. Moreover, the calculation of $Scale$ differs slightly depending on whether the video orientation (landscape or portrait) matches the screen orientation. If the orientations are the same, then:

$$Scale = \frac{\sqrt{R_{sw}^2 + R_{sh}^2}}{\sqrt{R_w^2 + R_h^2}} \tag{19}$$

else:

$$Scale = \frac{\sqrt{R_{sh}^2 + \left(\frac{R_{sh}^2}{R_{sw}}\right)^2}}{\sqrt{R_w^2 + R_h^2}} \tag{20}$$

P.940 recommends different values for coefficients v_1 to v_9 depending on the terminal device type and video codec format.

The network impairment factor $f_{network}$ represents the impact of network impairment on video quality, mainly video packet loss rate Plr_v and video interarrival jitter jit_v. It is calculated as follows:

$$f_{network} = v_{10} f_{pl} + (1 - v_{10}) f_{jit} \tag{21}$$

$$f_{pl} = \frac{1}{1 + \left(v_{11} Plr_v\right)^{v_{12}}} \tag{22}$$

$$f_{jit} = 1 - \exp\left(\frac{v_{13}}{\left(jit_v/100\right)^{v_{14}}}\right) \tag{23}$$

The P.940 model used the similar integration function (8) for audiovisual quality (Q_{av}), with different coefficients depending on the terminal device type.

2) Videotelephony quality assessing block: The P.940 model comprehensively takes audiovisual quality and user's interaction experience into consideration when evaluating the overall videotelephony quality. Among them, the user's interaction experience mainly include two aspects: one is the interaction delay experience, and the other is the media synchronization experience. In the P.940 model, Q_{delay} and Q_{sync} reflect the user experience of interaction delay and media synchronization, respectively, during video calls, using 5-grade scale. f_{delay} and f_{sync} are factors that represent the impact of interaction delay and media synchronization, which are bounded between 0 and 1.

The audiovisual interaction delay quality (Q_{delay}) is calculated as:

$$Q_{delay} = (w_1 + w_2 Q_{av}) - w_3 f_{delay} \tag{24}$$

$$f_{delay} = w_4 \left(\frac{T_v + T_a}{2}\right)^2 \tag{25}$$

The audiovisual media synchronization quality (Q_{sync}) is calculated as:

$$Q_{sync} = (w_5 + w_6 Q_{av}) - w_7 f_{sync} \tag{26}$$

Due to the asymmetric relationship between user experience of media synchronization and the difference between video delay and audio delay, the calculation of f_{sync} falls into two scenarios: if $T_v \geq T_a$, function (27) should be used; else, function (28) should be used.

$$f_{sync} = \left(\frac{w_{10}}{1 + \exp(w_8(Q_{av} + w_9))}\right)(T_v - T_a)^2 \tag{27}$$

$$f_{sync} = \left(\frac{w_{13}}{1 + \exp(w_{11}(Q_{av} + w_{12}))}\right)(T_a - T_v)^2 \tag{28}$$

Then the overall videotelephony quality (Q_{vt}) is expressed as:

$$Q_{vt} = n_1 Q_{av} - (n_2 f_{delay} + n_3 f_{sync}) + n_4 \tag{29}$$

C. Comparison between P.940 model and G.1070 model

Both G.1070 and P.940 models employ tailored E-models for audio quality assessment based on audio codec information. Therefore, the performance comparison between the two models focuses on three key dimensions: video quality, interaction experience and overall videotelephony quality.

1) Assessment of video quality: Regarding video quality assessment, both G.1070 and P.940 models defined their applicable ranges, primarily considering device screen, video codec and video resolution. Since P.940 model is trained on the subjective test datasets described in this paper, the comparative analysis ensures fairness and comprehensiveness through two approaches: one is directly comparing G.1070 and P.940 models using data within the common applicable range of both models; the other is comparing P.940 model and fine-tuned G.1070 model using all video quality data in the subjective test datasets.

The common applicable range of G.1070 and P.940 models meets the following conditions: 1) the terminal types are limited to mobile phones or TVs, 2) the video codec format is H.264(Baseline), 3) the video resolution ranges from 480p (VGA) to 1080p, and 4) the frame rate does not exceed 30 fps. Based on subjective test datasets that meet these conditions, a comparison of video quality assessment performance between the G.1070 and P.940 models is presented in Table V. Based on the common applicable range and standard-recommended coefficients for each model, the video quality assessment accuracy of the P.940 model is better than that of the G.1070 model, demonstrating a higher Pearson Linear Correlation Coefficient (PLCC) and a lower Root-Mean-Square Error (RMSE) compared to the G.1070 model.

To conduct a comprehensive comparison of the video quality assessment methods proposed by G.1070 and P.940 using all subjective video quality data, we applied least squares method to fine-tune the coefficients for G.1070 model. Table V presents the comparison results between the fine-tuned video quality assessment in G.1070 model and the standard-recommended video quality assessment in P.940 model, across video quality data in the subjective test datasets. It is shown that after fine-tuning the video quality assessment coefficients in the G.1070 model, it achieves lower performance metrics than the new P.940 model, across both non-interactive and interactive test datasets, and remarkably lower if compared only with the interactive test dataset.

TABLE V
PERFORMANCE ANALYSIS OF VIDEO QUALITY ASSESSMENT

Validation on data within the common applicable range		
Model	**PLCC**	**RMSE**
G.1070 model-V_q	0.6796	0.9032
P.940 model-Q_v	0.7356	0.5193
Validation on both non-interactive and interactive test datasets		
Model	**PLCC**	**RMSE**
fine-tuned G.1070 model-V_q	0.8541	0.4720
P.940 model-Q_v	0.8630	0.4582
Validation on interactive test dataset		
Model	**PLCC**	**RMSE**
fine-tuned G.1070 model-V_q	0.8903	0.5380
P.940 model-Q_v	0.9543	0.4889

2) Assessment of interaction experience: Both G.1070 model and P.940 model incorporate the impact of audio and video delays on the quality of video calls, but their perspectives and approaches differ. G.1070 defines an audio-visual delay impairment factor MM_t to quantify quality degradation caused by audio-visual delay and synchronization [5], while P.940 focuses on the quality perceived by users and provides a separate assessment for interactive delay experience Q_{delay} and media synchronization experience Q_{sync}.

The calculation of the audio-visual delay impairment factor in the G.1070 model exclusively requires audio delay and video delay as input parameters. However, user perception of interactive latency and media synchronization is not solely determined by end-to-end transmission delays. It is also affected by factors such as packet loss rates as analyzed in III-B3. Due to differences in the mechanisms and performance of audio and video streams under weak network conditions, the situation and impact of out-of-order packets and lost packets differ between audio and video streams. This discrepancy may lead to different increased media processing delays, such as media decoding delays and jitter-buffer delays, at the terminal side, which are difficult to measure in real-world videotelephony services. Moreover, the clarity and fluency of video images and speech also affect users' subjective perception and judgment of media synchronization issue. Therefore, in addition to audio and video delays, the P.940 model also takes audiovisual quality as one of its input parameters for interaction delay and media synchronization assessing blocks, addressing the impact of media processing and media display at the terminal sides, enabling a more accurate assessment of users' perceived quality in interactive experiences.

In addition to the quality of experience, the interaction delay and media synchronization assessing blocks of the P.940 model also output two impact factors, f_{delay} and f_{sync}. These two factors serve a similar purpose to AD (absolute audio-visual delay) and MS (audio-visual media synchronization) in the G.1070 model. Although the calculation approaches for f_{delay} and f_{sync} differ from AD and MS, their results are highly correlated ($|SRCC|$>0.95), indicating a strong similarity in how they vary with audio and video delays.

3) Assessment of overall videotelephony quality: The G.1070 model and the P.940 model exhibit some differences in their formulas for integrating audiovisual media quality and the impact of interaction experience. The G.1070 model uses a binomial formula as defined by (7) to fit the relationship between the audiovisual quality, the delay impairment factor and the overall quality, while the P.940 model employs a subtractive formula that adjusts a mapping of audiovisual quality by overlaying delay and synchronization effects, as presented by (29).

Under the condition that the P.940 model uses standard-recommended coefficients, and the G.1070 model employs fine-tuned coefficients for its video quality assessment function along with standard-recommended coefficients for the integration function, we compared these two models' performance in evaluating the overall videotelephony quality based on the interactive subjective test dataset, as listed in Table VI. Additionally, we also applied least squares coefficient fine-tuning to the integration function in the G.1070 model, with the optimized performance metrics also presented in Table VI.

TABLE VI
PERFORMANCE ANALYSIS OF OVERALL VIDEOTELEPHONY QUALITY ASSESSMENT

Model	**PLCC**	**RMSE**
P.940 model-Q_{vt}	0.9639	0.2924
G.1070 model-MM_q-partially finetune[a]	0.9244	0.5042
G.1070 model-MM_q-finetuned[b]	0.9561	0.3124

[a] This version of G.1070 model employs fine-tuned coefficients for video quality assessment and standard-recommended coefficients for integration.
[b] This version of G.1070 model employs fine-tuned coefficients for video quality assessment and fine-tuned coefficients for integration as well.

Table VI shows that with fine-tuned coefficients for video quality assessment and standard-recommended coefficients for integration, the performance of G.1070 model on interactive test dataset is lower than that of the P.940 model. This is mainly because the integration coefficients recommended by G.1070 are primarily for small-screen devices, including the coefficients for audiovisual quality calculation. In contrast, P.940 model recommends different groups of coefficients for audiovisual quality integration depending on the device type. After fine-tuning all coefficients in the integration function of G.1070 model using the same interactive test dataset as the P.940 model, the performance metrics of overall videotelephony quality assessment approximate those of the P.940 model, but are still lower than the new P.940 model.

V. CONCLUSION

This paper presents a cross-lab study on videotelephony quality assessment conducted by members of the ITU-T SG12 and VQEG, aiming to address the needs for quality assessment and monitoring in videotelephony services enabled by consumer technology. This study was a collaborative project undertaken by five participating parties across three continents, which led to the development and publication of Recommendation ITU-T P.940 and its supplement P Suppl.31.

Subjective tests incorporated diverse cultural backgrounds and languages, including 1207 non-interactive test sequences

and 576 interactive test conversations, with over 120 subjective evaluators. The construction of subjective datasets encompassing multiple perspectives of experience quality perceived by consumers, together with the development of objective assessment models, has provided critical support for the formulation of ITU-T Recommendation P.940. Analysis of non-interactive subjective test dataset validated that video codec and device type significantly impact consumer-perceived video quality. The analysis of interactive subjective test dataset reveals that consumers' subjective perception of interaction latency and media synchronization is modulated not only by end-to-end transmission delays but also by factors such as network impairments.

Furthermore, a comparative analysis of G.1070 (QoE/QoS planning) model and P.940 (QoE/QoS monitoring) model was conducted on video quality, interaction experience and overall videotelephony quality, using the same subjective datasets. Overall, P.940 model demonstrates better performance than G.1070 model and has a broader applicability range. Even after fine-tuning G.1070 model coefficients using the same subjective datasets, its performance improved but still slightly inferior to P.940 model.

VI. Future Work

Video sequences under identical test conditions can exhibit significant video quality differences, depending on the temporal and spatial complexity of video scenes. Features related to spatiotemporal complexity will improve the accuracy of video quality assessment, but they are difficult to be obtained based on video contents due to privacy and data protection restrictions. Further work will focus on: 1) developing privacy-compliant methods to extract features related to spatiotemporal complexity, 2) investigating bitstream-based no-reference assessment by incorporating media features to improve assessment accuracy, and 3) integrating extended multimedia functions, such as screen sharing, to build a comprehensive QoE assessment framework for videotelephony and videoconferencing services.

Additionally, this study was conducted for one-to-one videotelephony scenarios. In one-to-many videotelephony scenarios, the P.940 model can also be applied to evaluate the videotelephony quality between a single consumer and another consumer in the same video call. However, the evaluation of overall videotelephony quality in one-to-many scenarios requires integrating the individual videotelephony quality between each pair of interacting consumers, which is also a direction worthy of further research.

References

[1] J. Joskowicz "Video Conferencing Technologies: Past, Present and Future", in *TechRxiv*. December 04, 2023.

[2] ITU-T, Study Group 12 - Performance, QoS and QoE. Accessed: May 20, 2025. Available: https://www.itu.int/en/ITU-T/studygroups/2025-2028/12/Pages/default.aspx

[3] ITU-T, "Computational model used for the monitoring and quality assessment of videotelephony services", *Recommendation P.940*, March 2025.

[4] ITU-T, "Subjective quality evaluation of audiovisual communication in videotelephony services", *Supplement to ITU-T P-series Recommendations P Suppl.31*, January 2025.

[5] ITU-T, "Opinion model for video-telephony applications", *Recommendation G.1070*, June 2018.

[6] W3C Editor's Draft, "WebRTC: Real-Time Communication in Browsers", February 2025. Available: https://w3c.github.io/webrtc-pc/.

[7] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session Initiation Protocol", *RFC 3261*, June 2002. Available: https://www.rfc-editor.org/info/rfc3261.

[8] T. Wiegand, G. J. Sullivan, G. Bjontegaard and A. Luthra, "Overview of the H.264/AVC video coding standard", in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576, July 2003.

[9] G. J. Sullivan, J.-R. Ohm, W.-J. Han and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard", in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649-1668, December 2012.

[10] JM. Valin, K. Vos, T. Terriberry, "Definition of the Opus Audio Codec", *RFC 6716*, September 2012. Available: https://www.rfc-editor.org/info/rfc6716.

[11] Opus Interactive Audio Codec. Available: https://opus-codec.org/.

[12] A. Rämö, H. Toukomaa, "Voice quality characterization of IETF opus codec", in *Proc. Interspeech 2011*, 2541-2544, 2011.

[13] M. Maruschke, O. Jokisch, M. Meszaros, F. Trojahn, and M. Hoffmann, "Quality assessment of two full-band audio codecs supporting real-time communication", in *Proc. Int. Conf. Speech Comput*. Cham, Switzerland : Springer, August 2016, pp. 571–579.

[14] J. Postel, "User Datagram Protocol", *STD 6*, *RFC 768*, August 1980. Available: https://www.rfc-editor.org/info/rfc768.

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

[16] ITU-T, "Vocabulary for performance, quality of service and quality of experience", *Recommendation P.10/G.100*, November 2017.

[17] ITU-T, "Taxonomy of telemeetings from a quality of experience perspective", *Recommendation G.1092*, October 2023.

[18] H. Becerra Martinez, A. Hines and M. C. Q. Farias, "Perceptual Quality of Audio-Visual Content with Common Video and Audio Degradations", in *Applied Science*, 11(13), 5813, 2021.

[19] ITU-T, "Effect of delays on telemeeting quality", *Recommendation P.1305*, July 2016.

[20] C. Diao, S. A. Arboleda and A. Raake, "Effects of Delay on Nonverbal Behavior and Interpersonal Coordination in Video Conferencing", in *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, West Lafayette, IN, USA, 2024.

[21] D. W. Edwards, "Impacts of Telecommunications Latency on the Timing of Speaker Transitions", in *Speech Communication*, 103226, ISSN 0167-6393, 2025.

[22] N. Staelens, J. De Meulenaere, L. Bleumers et al., "Assessing the importance of audio/video synchronization for simultaneous translation of video sequences", in *Multimedia Systems*, vol. 18, pp. 445-457, November 2012.

[23] I. Saidi, L. Zhang, V. Barriac and O. Deforges, "Interactive vs. non-interactive subjective evaluation of IP network impairments on audio-visual quality in videoconferencing context", in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Lisbon, 2016.

[24] F. Braun, R. R. Ramachandra Rao, W. Robitza and A. Raake, "Automatic Audiovisual Asynchrony Measurement for Quality Assessment of Videoconferencing", in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, Ghent, Belgium, 2023.

[25] ITU-R, "Relative timing of sound and vision for broadcasting", *Recommendation BT.1359-1*, November 1998.

[26] B. Naderi and R. Cutler, "A Crowdsourcing Approach to Video Quality Assessment", in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2810-2814, Seoul, Korea, Republic of, 2024.

[27] Y. Zhang et al., "Subjective Panoramic Video Quality Assessment Database for Coding Applications", in *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 461-473, June 2018.

[28] J. Gutiérrez et al., "Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360° Videos: ITU-T Rec. P.919", in *IEEE Transactions on Multimedia*, vol. 24, pp. 3087-3100, 2022.

[29] Z. Zhang et al., "Quality-of-Experience Evaluation for Digital Twins in 6G Network Environments", in *IEEE Transactions on Broadcasting*, vol. 70, no. 3, pp. 995-1007, September 2024.

[30] M. R. dos Santos, A. P. Batista, R. L. Rosa, M. Saadi, D. C. Melgarejo and D. Z. Rodríguez, "AsQM: Audio Streaming Quality Metric Based on Network Impairments and User Preferences", in *IEEE Transactions on Consumer Electronics*, vol. 69, no. 3, pp. 408-420, August 2023.

[31] J. Nightingale, Q. Wang, C. Grecos and S. Goma, "The impact of network impairment on quality of experience (QoE) in H.265/HEVC video streaming", in *IEEE Transactions on Consumer Electronics*, vol. 60, no. 2, pp. 242-250, May 2014.

[32] A. Takahashi, D. Hands and V. Barriac, "Standardization activities in the ITU for a QoE assessment of IPTV", in *IEEE Communications Magazine*, vol. 46, no. 2, pp. 78-84, February 2008.

[33] M. Orduna, C. Díaz, L. Muñoz, P. Pérez, I. Benito and N. García, "Video Multimethod Assessment Fusion (VMAF) on 360VR Contents", in *IEEE Transactions on Consumer Electronics*, vol. 66, no. 1, pp. 22-31, February 2020.

[34] B. Garcia, F. Gortázar, M. Gallego and A. Hines, "Assessment of QoE for video and audio in WebRTC applications using full-reference models", in *Electronics 2020*, 9(3), 462, March 2020.

[35] C. T. E. R. Hewage and M. G. Martini, "Reduced-reference quality assessment for 3D video compression and transmission", in *IEEE Transactions on Consumer Electronics*, vol. 57, no. 3, pp. 1185-1193, August 2011.

[36] Z. Wang and A. C. Bovik, "Reduced- and No-Reference Image Quality Assessment", in *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29-40, November 2011.

[37] A. De Decker, J. De Cock, P. Lambert and G. V. Wallendael, "No-Reference VMAF: A Deep Neural Network-Based Approach to Blind Video Quality Assessment", in *IEEE Transactions on Broadcasting*, vol. 70, no. 3, pp. 844-861, September 2024.

[38] W. Shen et al., "A Blind Video Quality Assessment Method via Spatiotemporal Pyramid Attention", in *IEEE Transactions on Broadcasting*, vol. 70, no. 1, pp. 251-264, March 2024.

[39] S. Jiang, Q. Sang, Z. Hu and L. Liu, "Self-Supervised Representation Learning for Video Quality Assessment", in *IEEE Transactions on Broadcasting*, vol. 69, no. 1, pp. 118-129, March 2023.

[40] W. Xian et al., "Spatiotemporal Feature Hierarchy-Based Blind Prediction of Natural Video Quality via Transfer Learning", in *IEEE Transactions on Broadcasting*, vol. 69, no. 1, pp. 130-143, March 2023.

[41] G. Bingöl, S. Porcu, A. Floris and L. Atzori, "QoE Estimation of WebRTC-based Audiovisual Conversations from Facial Expressions", in *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 577-584, Dijon, France, 2022.

[42] G. Bingöl, S. Porcu, A. Floris and L. Atzori, "QoE Estimation of WebRTC-based Audio-visual Conversations from Facial and Speech Features", in *ACM Transactions on Multimedia Computing, Communications and Applications*, Volume 20, Issue 5, Article No.: 130, Pages 1 - 23, January 2024.

[43] T. Liu, N. Narvekar, B. Wang, et al., "Real-time video quality monitoring", in *EURASIP Journal on Advances in Signal Processing*, 122 (2011), 2011.

[44] C. Udora, J. Adhuran and A. Fernando, "A Quality-of-Experience-Aware Framework for Versatile Video Coding-Based Video Transmission", in *IEEE Transactions on Consumer Electronics*, vol. 69, no. 2, pp. 205-216, May 2023.

[45] H. Yuan, Q. Wang, Q. Liu, J. Huo and P. Li, "Hybrid Distortion-Based Rate-Distortion Optimization and Rate Control for H.265/HEVC", in *IEEE Transactions on Consumer Electronics*, vol. 67, no. 2, pp. 97-106, May 2021.

[46] ITU-T, "Subjective video quality assessment methods for multimedia applications", *Recommendation P.910*, October 2023.

[47] ITU-T, "Interactive test methods for audiovisual communications", *Recommendation P.920*, May. 2000.

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

[49] J. Joskowicz, M. Liu, R. Sotelo, A. Armendariz and L. Yang, "Conversational Subjective Tests Based on Video-telephony Platform", in *2023 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Beijing, China, 2023.

[50] A. Armendariz, J. Joskowicz, R. Sotelo and M. Liu, "A Test Bed for Subjective Multimedia Quality Evaluation in Videoconferencing Systems", in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, pp. 1-4, 2024.

[51] BigBlueButton, Virtual Classroom Software. Available: https://bigbluebutton.org.

[52] The Linux Foundation, Netem - Network Emulator. Available: https://wiki.linuxfoundation.org/networking/netem.

[53] M. Kerrisk, Traffic Control in the Linux Kernel. Available: https://man7.org/linux/man-pages/man8/tc.8.html.

[54] A. Pcjić, P. M. Stanić and S. Pletl, "Analysis of packet loss prediction effects on the objective quality measures of Opus codec", in *2014 IEEE 12th International Symposium on Intelligent Systems and Informatics (SISY)*, Subotica, Serbia, 2014.

[55] P. Orosz, T. Skopkó, Z. Nagy and T. Lukovics, "Performance Analysis of the Opus Codec in VoIP Environment Using QoE Evaluation", in *International Conference on Systems and Networks Communications (ICSNC)*, October 2013.

[56] Y. Cinar, P. Pocta, D. Chambers and H. Melvin, "Improved Jitter Buffer Management for WebRTC", in *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Vol. 17, No. 1, Article 30, April, 2021.

[57] ITU-T, "The E-model: a computational model for use in transmission planning", *Recommendation G.107*, June 2015.

[58] ITU-T, "Wideband E-model", *Recommendation G.107.1*, June 2019.

[59] ITU-T, "Fullband E-model", *Recommendation G.107.2*, March 2023.

[60] M. Al-Ahmadi, P. Pocta and H. Melvin, "Instrumental Estimation of E-model Equipment Impairment Factor Parameters for Super-wideband Opus Codec", in *2019 30th Irish Signals and Systems Conference (ISSC)*, Maynooth, Ireland, 2019.