applied sciences

MDPI

Article

Perceptual Quality of Audio-Visual Content with Common Video and Audio Degradations

Helard Becerra Martinez [1,*], Andrew Hines [1] and Mylène C. Q. Farias [2]

[1] School of Computer Science, University of College Dublin, Dublin 4, Ireland; andrew.hines@ucd.ie
[2] Department of Electrical Engineering, University of Brasília, Brasília 70.910-900, Brazil; mylene@ieee.org
* Correspondence: helard.becerra@ucd.ie

Abstract: Audio-visual quality assessment remains as a complex research field. A great effort is being made to understand how visual and auditory domains are integrated and processed by humans. In this work, we analyzed and compared the results of three psychophisical experiments that collected quality and content scores given by a pool of subjects. The experiments include diverse content audio-visual material, e.g., Sports, TV Commercials, Interviews, Music, Documentaries and Cartoons, impaired with several visual (bitrate compression, packet-loss, and frame-freezing) and auditory (background noise, echo, clip, chop) distortions. Each experiment explores a particular domain. In Experiment 1, the video component was degraded with visual artifacts, meanwhile, the audio component did not suffer any type of degradation. In Experiment 2, the audio component was degraded while the video component remained untouched. Finally, in Experiment 3 both audio and video components were degraded. As expected, results confirmed a dominance of the visual component in the overall audio-visual quality. However, a detailed analysis showed that, for certain types of audio distortions, the audio component played a more important role in the construction of the overall perceived quality.

Keywords: video quality assessment; audio and video quality; QoE

Citation: Martinez, H.B.; Hines, A.; Farias, M.C.Q. Perceptual Quality of Audio-Visual Content with Common Video and Audio Degradations. *Appl. Sci.* **2021**, *11*, 5813. https://doi.org/10.3390/app11135813

Academic Editor: Cheonshik Kim

Received: 3 May 2021
Accepted: 18 June 2021
Published: 23 June 2021

1. Introduction

As new types of codecs, transmission protocols, and application scenarios evolve, the importance of quality assessment methodologies for the different types of multimedia signals (including audio and video contents) increases. With the introduction of Quality of Experience (QoE) approaches, traditional methods based exclusively on Quality of Service (QoS) measurements are no longer the only way to measure the quality of a signal. In addition to QoS measurements, QoE approaches take into account characteristics of the Human Visual System (HVS) and the Human Auditory System (HAS). In the last decades, QoE approaches have been used to develop several objective quality metrics for digital videos and speech/audio signals [1–3]. However, despite the achievements made in the area of visual quality [4–6], several gaps remain open in the area of multimedia quality. As pointed out by Pinson et al. [7], simultaneously measuring the quality of multimedia contents (e.g., video and audio) is still a pending issue.

Modelling human's perception of audio and video signals individually is a challenging task. When the interaction between audio and video signals is considered, then the level of complexity increases. Part of the difficulty lies in the fact that the cognitive processing method used to interpret the interaction of audio and video stimulus is not yet completely understood [6]. Therefore, understanding how humans perceive, process, and interpret different types of signals is key to develop an accurate audio-visual quality assessment model. To gain some knowledge in this direction, researchers have measured the audio-visual quality of video sequences by performing psychophysical tests with human participants

(also known as perceptual or subjective experiments) [8]. Such experiments can be used to provide a measure of the level of influence that certain (audio and video) distortions have on the audio-visual quality perceived by human users. This information is key in the process of testing and developing objective quality metrics, now more than before, since quality metrics based in Machine Learning (ML) techniques are highly dependent on annotated data for training and testing new models [6,9–11].

Psychophysical experiments are commonly carried out within a controlled environment (e.g., soundproof laboratories), where test stimuli (e.g., audio-visual sequences) is presented to a group of non-expert human participants (or experts depending on the application being studied). With the goal of generating reliable and reproducible results that represent quality as perceived by human users, several international agencies and research organizations have collected and published guidelines and recommendations on how to perform psychophysical experiments [8,12–15]. Although these recommendations are widely adopted and applied in most perceptual studies, they often limit the representation of an authentic user experience, which is a requirement in multimedia applications. For this reason, researchers have created novel methods to deal with this matter. For example, an immersive methodology proposed by Pinson [16] tried to recreate a more natural media consumption environment for a human consumer, thus, allowing to collect human responses that are more realistic and useful for a quality perception analysis.

Quality perception of audio-visual signals has been studied thought several perceptual experiments [17–21]. Early experiments have identified the visual component as the dominant influence in the overall audio-visual quality; yet, it has been argued that this influence does not apply to all types of audio-visual applications. Studies have found the audio component as being very important in applications like video conference [20]. Other studies have confirmed that audio and video interaction depends on human, technological, and contextual factors [9]. In order to have a deeper understanding of this interaction, researchers have proposed new methods to subjectively assess audio-visual quality, more specifically, long-duration audio-visual stimuli [22,23]. However, there is a limited number of experiments that aim to study the overall audio-visual quality in video sequences. In the majority of these experiments, only the video component is processed and the audio component is left unimpaired [24]. Among the few studies that have explored the overall audio-visual quality of sequences with distortions in both audio and video components, we can cite the works of Pinson [19], Becerra [25]. More recently, the study presented by Min et al. [18] who conducted experiments in which the content had compression distortions in both audio and video components. Although their results confirmed the dominance of the video component on the overall quality, the audio component had a considerable impact on quality for some types of media content.

In summary, although several studies explored how users perceive audio-visual quality in common multimedia scenarios, most of these studies ignore the audio component and the effect of audio distortions on the overall audio-visual quality. Additionally, the complex interactions of audio and video components and the different factors influencing the perception of quality have not been properly studied. This work presents a compilation of the results from three psychophysical experiments where participants assessed the overall quality of a diverse set of audio-visual sequences. All three experiments used the immersive methodology, including audio and video impairments. In Experiment 1, artifacts were added to the visual component, while the audio component remained as is. In Experiment 2, the visual component was kept untouched while artifacts were added to the audio component. Finally, in Experiment 3, both video and audio components were impaired by audio and visual artifacts. This paper performs a detailed analysis of the experiments and compares their results, binding conclusions and insights among all three experimental scenarios. Our main goal is to present a wider picture of the audio-visual quality of video sequences and attempt to drive global conclusions.

The remainder of this document is divided as follows. In Section 2, a summary of some relevant multimedia perceptual experiments is presented. In Section 3, a description of the rationale behind the study and the experimental design is presented. In Section 4, visual and audio degradations are presented. In Section 5, experimental details are described. In Sections 6–8, all three experiments are described and their corresponding results are presented. Section 9 presents a comparison and a general discussion on the results from all three experiments. Finally, the overall conclusions of the study are presented in Section 10.

2. Previous Work on Perceptual Quality Assessment

As commented before, perceptual experiments help researchers comprehend how different artifacts affect the perceived quality of media content such as audio, video and audio-visual. The studies listed in this section are presented in Table 1. For visual content, the quality impact of packet-loss, bitrate compression and frame-freezing have been studied throughout a range of perceptual experiments. For example, the effects of frame-freezing and packet-loss over full-length movie clips were reported in a perceptual experiment conducted by Staelens et al. [26]. The experiment collected the responses from 56 non-expert viewers for 80 DVD clips in a home viewing environment. The study concluded that participants were more tolerant towards frame-freezing errors, possibly due to the length of the stimuli and the viewing conditions. In another study, Moorthy et al. [27] included a range of impairments such as video compression, frame-freezing, wireless-channel packet-loss and rate adaptation over a mobile-platform setup. A video-only dataset (with a range of content material) was presented to 30 participants during the experiment. The results concluded that viewers were more tolerant to few longer stalling events than several shorter ones. However, these results varied according to the type of content being displayed. A recent study explored the impact of video quality using several video encoding parameters like bitrate, frame-rate and video resolution in multiparty telemeetings. Perceptual results showed that participants' detection of some visual distortions was influenced by how active they were during the calls. The study indicates that, for this particular scenario, ensuring enough resources to active participants and prioritizing the audio quality are key to the overall audio-visual quality [28].

Table 1. Summary of subjective studies listed in Section 2.

Component	Scope	Distortions	Material	Reference
Video	Video quality assessment in home environment	frame-freezing	full-length movie clips	[26]
		packet-loss		
	Video quality assessment over a mobile-patform	Video compression	HD video clips	[27]
		frame-freezing		
		wireless-channel packet-loss		
		Bitrate adaptation		
	Video quality assessment in multiparty telemeetings	Video compression	WebRTC-based video calls	[28]
		Frame-rate		
		Video resolution		
Audio	Speech intelligibility	Background noise	Speech clips	[29]
		Syntactic complexity		
	Speech quality assessment for VoIP	Background noise	Speech clips	[30]
		Competing speaker		
		Clipping		
		Echo		
		Chopped speech		
	Music and video streaming consumption	Stalling effect	Music clips	[31]
Audio-visual	Audio-visual quality assessment in a controlled environment	Video compression	HD video clips	[25]
		Audio compression		
	Audio-visual quality assessment over different devices.	Video compression	SD video clips	[19]
		Audio compression		
	Audio-visual quality assessment in a controlled environment	Video compression	HD video clips	[18]
		Audio compression		
		Video scaling		
	Audio-visual quality assessment over live music streaming	Video compression	DVD Live music clips	[32]
		Audio compression		
		Network errors		
	Audio-visual quality assessment in a real-time scenario	Bitrate adaptation	HD video clips	[33]
		Network errors		
		Frame-rate		
		Quantization parameters		
		Audio and video packet-loss		

For audio content, a range of studies has focused on investigating the perceived quality effect of acoustic background noise [34,35]. In the study presented by Wendt et al. [29], speech intelligibility was compared across different levels of background noise and syntactic structure complexity. Results suggested that intelligibility was more affected by background noise than syntactic complexity. In another study, Harte et al. [30] explored the effects of several voice over IP (VoIP) distortions. The dataset included speech samples impaired with background noise, competing speaker effects, clipping, echo, and chopped speech, all added in isolation. The authors reported that echo and background noise had a heavier impact on the quality perceived by participants. A recent study presented by Schwind et al. [31] investigated the stalling effect over music streaming. The authors state that music streaming consumption differs from a video streaming scenario since music is usually played in the background. In this study, the authors explored the perceived quality of users while they are performing other tasks. For speech enhancement, researchers have exploited the visual information available (e.g., lip movement, facial expression) to achieve stronger performance. Some relevant audio-visual speech datasets are listed in [36–38].

In terms of audio-visual quality, a number of studies have explored the audio and video interactions and their corresponding contribution to the perceived overall quality. Some of these studies showed that, for certain types of content, the video component has a stronger influence on the audio-visual quality. In contrast, the study presented in [20] reported that for certain communication scenarios (e.g., teleconference calls), the audio component plays a more important role in the overall perceived quality. In general, results in the literature show that the influence of the video and audio components on the perceived quality is related to several context factors (nature of experience, e.g., teleconference, sports events, movies, etc.). In addition, human and technological factors are also determinant to model the human perceived quality [39]. Aiming to study these factors and their influence, researchers have explored different experimental methodologies, new and diverse media content, and new types of distortions. For example, Staelens et al. [22] and Borowiak et al. [23] have tried to capture the participants' attention by using long-duration audio-visual stimuli. Both studies applied alternative methodologies to collect human responses.

Becerra et al. [25] conducted a group of experiments to study the impact of heavy audio and video compression artifacts on quality. Similarly, Pinson et al. [19] performed a study where ten different laboratories ran perceptual experiments, which consisted of presenting sequences with compressed audio and video components, considering different environments and devices. More recently, a study presented by Min et al. [18] conducted an audio-visual perceptual experiment introducing distortions caused by audio and video compression and video compression combined with video scaling. Despite their setup differences, all these three studies agreed on the influence of audio on the perceived audio-visual quality. However, a more specific perceptual study presented by Rodrigues et al. [32] explored the trade-off between audio and video quality over live music streaming sequences. The study applied distortions caused by audio and video compression for a mobile network environment. The study concluded that, for this particular case study, reducing the audio bitrate didn't affect the overall audio-visual quality and, the video bitrate reduction had a great impact on the perceived audio-visual quality. In another study, Demirbilek et al. [33] tried to reproduce a real-time communication scenario and performed a study that included variations of compression and network parameters, including video frame rates, quantization parameters, packet-loss rates (audio and video), noise filters, and compression bitrates (audio and video). Results showed that packet-loss errors (for audio and video) had a greater impact on the perceived audio-visual quality than other types of parameters.

Certainly, there is a need to fully understand the audio-visual quality perception process. The complex interactions between audio and video pose a challenge and demand new methods to conduct perceptual experiments. The Immersive Methodology has shown promising results and offers a reliable method to deal with non-traditional experimental

scenarios (e.g., long-duration stimuli, numerous test conditions, and diverse content). In the following Section, the motivation behind this study and the experimental settings are presented.

3. Motivation

As commented before, the complexity of assessing audio-visual quality lies, among other factors, in the little understanding there is on how the auditory and visual stimuli are perceived, and also at what stage, and how, the human perceptual system integrates them. This reflects the gaps between neurophysiology and computer science in terms of quality assessment. Undoubtedly, there is still a great demand for studies that investigate the complexity involving the integration of multimodal stimuli, as stated by Akhtar and Falk in a recent survey [6]. Unfortunately, many studies that assess the audio-visual quality consider impairments only in the video component and leave the audio unimpaired, i.e.; subjects are presented with audio-visual content but with no distortion in the audio component. As for the ones that do include audio distortions, they are usually limited to weak distortions caused by audio compression. Table 2 presents a summary of some of the most important audio-visual databases and their corresponding perceptual experiments. As it can be observed, most of the available audio-visual material is limited in terms of audio distortion. These conditions make it difficult reaching to more significant conclusions in terms of visual and auditory stimuli integration.

Table 2. Audio-visual quality assessment databases and perceptual studies. A: Audio; S: Speech; M: Music; V: Video; A/V: Audio-visual; TT: Telephone transmission; SRC: Source stimuli; HRC: Hypothetical reference circuit; TS: Test sequences; *n*: Number of subjects; BG: Background noise; AC: Audio compression; CS: Competing speaker; Ec: Echo; Ch: Chop speech; Cl: Clipping; ACA: Amplitude compression and amplification; BF: Butterworth filtering; WCN: White and crowd noise; VC: Video compression; TE: Transmission Errors; FF: Frame Freezing; BA: Bitrate adaptation.

DataSet			Test Material					Subjective Experiment			
Year	**Name**	**Focus**	**Media**	**SRC**	**HRC**	**TS**	**Distortion**	**Methodology**	**Scale**	**MOS**	***n***
2010	PLYM [40]	Mobile IP	V: 144 p, 8–15 fps, 7–14 s A: 8 kHz, 16-bits	6	10	60	VC, TE	ACR	9-point	MOSa,v,av	16
2012	VQEG-MM [19]	IPTV	V: 480 p, 30 fps, 10 s A: 48 kHz, 16-bits	10	5	60	AC, VC	ACR	[1–5]	MOSav	35
2012	TUM [41]	IPTV	V: 1080 p, 50 fps, 10 s	5	4	20	VC, TE	ACR	11-point	MOSav	21
2013	Live-Music [42]	Music	V: YouTube	100	5	500	ACA, BF, WCN	MRR	[0–1]	MOSa	60
2013	UnB-AVQ 2013 [25]	IPTV	V: 720 p, 4:2:0, 30 fps A: 16-bit, 48 kHz, 8 s	8	12	72	AC: MPEG-1 Layer 3 VC: H.264	ACR	[1–5]	MOSav	16
2013	VTT [43]	Streaming	V: 1080-720-480 p, 20–30 fps, 10 s	12		125	AC, VC, TE	ACR	[1–5]	MOSa,v,av	24
2016	INRS [44]	IPTV	V: 720 p, 42 s	1	160	160	VC: H.264, TE	ACR	[1–5]	MOSav	30
2016	MMSPG [45]	IPTV	V: 2160-1080-720 p, 24 fps A: Stereo, 5.1 Surround	9	3	27	Display devices	ACR	[1–9]	MOSa,av	20
2017	LIVE Mobile Stall [46]	Mobile IP	V: YouTube, 720-640-360 p	24	26	174	Video stalling	ACR	[1–5]	MOSv	54
2018	UnB-AVQ 2018 [47]	IPTV	V: 720 p, 30 fps, 4:2:0 A: 16-bits, 48 kHz, 37 s	60	12	720	VC: H.264–H.265, TE, FF	Immersive	[1–5]	MOSav	60
2018	UnB-AVQ 2018 [47]	IPTV	V: 720 p, 30 fps, 4:2:0 A: 16-bits, 48 kHz, 35 s	40	20	800	Ch, Cl, CS, Ec, BG Noise	Immersive	[1–5]	MOSav	40
2018	UnB-AVQ 2018 [47]	IPTV	V: 720 p, 30 fps, 4:2:0 A: 16-bits, 48 kHz, 34 s	40	20	800	Ch, Cl, CS, Ec, BG Noise VC: H.264–H.265, TE, FF	Immersive	[1–5]	MOSav	40
2018	LIVE-NFLX-II [24]	IPTV	V: 1080 p, 30 fps, 25 s	15	28	420	TE, BA	ACR	[1–100]	MOSv	65
2020	LIVE-SJTU [18]	IPTV	V: 1080 p, 30 fps, 8 s A: 16-bits, 48 kHz	14	24	336	VC: H.265, Scaling AC: AAC	ACR	[1–5]	MOSav	35

In this study, we aim to use a wider collection of audio and video distortions, which are not commonly included in the available audio-visual databases. Such distortions were selected by the researchers based on previous studies from the literature and a particular interest in studying some specific types of distortions. Three types of visual distortions were selected: video coding, packet loss, and frame freezing. As for the audio component, four types of distortions were selected: background noise, clipping, echo, and chop. For each type of distortion, different levels of degradation were selected by audio and video

experts using empirical criteria. This consisted of examining the audio and video sequences and choosing very clear quality levels, which will then be considered as the experimental test conditions of the study.

It is understood that the combination of some of these audio and video distortions, and the test conditions considered, do not necessarily appear in a real transmission scenario. Therefore, it is important to state that the purpose of using these distortions, and combining them in a perceptual experiment, is to investigate how human participants perceive them and at what level the audio and visual distortions influence their perception of the audio-visual quality. Then, recreating a particular application scenario is out of the scope of this study. It is expected that using diverse audio and video distortions and making available the annotated material will contribute to the study of multimodal quality perception, which is commonly restricted to a single type of audio distortion (i.e., audio compression). Based on these findings, more specific studies can be performed considering a particular use-case scenario, which will include specific types of degradation and content material.

4. Video and Audio Distortions

We used a large set of source sequences for this set of experiments and processed them, introducing video and audio distortions. The selected distortions were based on previous studies that analyzed audio and video distortions. For each type of degradation, we generated a number of distortion levels, which were used to establish the Hypothetical Reference Circuits (HRCs) for each perceptual experiment. This section describes the different types of degradations, detailing the procedure used to treat the source contents and generate the Processed Video Sequences (PVS).

4.1. Video Degradations

The video component of the source sequences was subject to three types of distortions: video coding, packet-loss, and frame-freezing. These types of degradations were applied to source sequences in Experiments 1 and 3 [48,49]. The processing algorithms used to generate the PVSs are described below.

4.1.1. Coding Artifacts (Compression)

The Advance Video Coding (AVC) H.264/ MPEG-4 and the High Efficiency Video Coding (HEVC) H.265 [50,51] standards were selected to compress the source stimuli. Throughout a visual examination, researchers selected four bitrate values for each coding standard and they labelled it as Low, Medium, High, and Very High quality. The sample sequences used for this analysis were not included in the main experiment. The selection was made by taking into account bitrate values used in previous works and picking four clear quality levels [52,53]. Table 3 presents the bitrate values used for each codec.

Table 3. Compression Bitrate values used for each codec.

	Low	Medium	High	Very High
H.264/AVC	500 kbps	800 kbps	2000 kbps	16,000 kbps
H.265/HEVC	200 kbps	400 kbps	1000 kbps	8000 kbps

4.1.2. Packet-Loss

As a first step, video bit-streams for the corresponding H.264 and H.265 coding standards were generated. To simulate a packet-loss effect, packets from the Network Abstraction Layer (NAL) were discarded from the video bit-stream using the NALTools software, as done in similar studies [54]. The packet-loss ratios used for this study were: 1%, 3%, 5%, 8%, and 10%. Although these values were taken from a real transmission scenario found in video streaming applications [55,56], recreating a particular use case scenario was not the main intention. These packet-loss ratios served to define five clear quality levels that could be distinguished by participants.

4.1.3. Frame-Freezing

For this study, a frame freezing without skipping was selected; that is, pauses of the video did not discard any of the incoming frames. To recreate a frame freezing event, the following parameters were considered: (a) number of events, (b) position in the sequence, and (c) length of the event. For each sequence, a maximum of three freezing events was inserted. These events were inserted at the "start", "middle", and "end" of the sequence. As for the length of the events, they were set to 1, 2, and 4 s.

Five levels of discomfort were established from the combination of these three parameters (number, position, and length of the events). The levels were labelled as "S1", "S2", "S3", "S4", and "S5", scaling from a low annoyance effect (S1) to a high annoyance effect (S5). Table 4 presents the parameter combination for these five annoyance levels.

Table 4. Frame-Freezing parameters. n: Number of events; P[1,2,3]: start, middle, end; L[1,2,3]: 1, 2, 4 s.

	Level	n	P1	P2	P3	L1	L2	L3
Low	S1	1	2			2		
Medium	S2	2	1	3		1	3	
	S3	2	2	3		2	2	
High	S4	3	1	2	3	2	2	3
	S5	3	1	2	3	3	3	2

It is important to mention that, since frame-freezing is only present in video services (Video-on-Demand or YouTube) based on reliable transport mechanisms (e.g., Transmission Control Protocol—TCP), a user of these services does not experience packet-loss distortions, which are common in services based in non-reliable transport mechanisms, such as User Datagram Protocol (UDP). Therefore, since frame-freezing and packet-loss degradations do not appear in the same types of scenarios, in our experiments, they were not inserted simultaneously into a single video sequence.

4.2. Audio Degradations

For this study, the audio component of the source sequences was impaired with four audio degradations: background noise, clipping, echo, and chop. The TCD-VoIP dataset [30] was used as a reference to reproduce this set of distortions. As stated in [30], these four audio degradations are platform-independent, i.e., they are not attached to a particular codec, network or hardware. As commented before, these distortions were selected with the purpose of studying their impact on the audio-visual quality independently of a particular use case scenario.

4.2.1. Background Noise

For background noise distortion, two parameters were considered to generate different test conditions: the type of noise (babble, car, road, and office) and the SNR level associated with the noise (5, 10, 15 dBs). Combining these parameters resulted in four test conditions (Table 5).

4.2.2. Clipping

The clipping effect was generated by multiplying the audio signal by four different amplitude multipliers (11, 15, 25, 55). As a result, four different test conditions were generated for this audio distortion (Table 5). The reference values used as amplitude multiplier were taken from the TCD-VoIP dataset [30].

4.2.3. Echo

To recreate an echo effect, delayed samples were added to the original audio sample. Four test conditions (Table 5) were obtained by varying three parameters: (a) amplitude

percentage of the delayed sample, (b) time delay between the original and the delayed sample, and (c) percentage reduction of the delayed samples.

4.2.4. Chop

A choppy speech effect was generated by discarding samples from an audio signal. Four test conditions (Table 5) were obtained by varying three parameters: (a) length of discarded samples, (b) sample discard frequency and (c) discarded samples treatment.

The audio and video degradations and their corresponding test conditions described in this section were used as the basis to build three audio-visual sub-sets, which compounds the UnB-AV dataset [47]. Each sub-set was used in a different experiment which will be described in the following sections.

Table 5. Audio Degradations Parameters.

Degradation	Conditions	Parameters	Range
Chop	4	Rate Period Mode	1, 2, 5 (chops/s) 0.02, 0.04 (s) previous, zeros
Clip	4	Multiplier	11, 15, 25, 55
Echo	4	Alpha Delay Feedback	0.175, 0.3, 0.5 (%) 25, 100, 140, 180 (ms) 0, 0.8 (%)
Noise	4	Noise type SNR	car, babble, office, road 15, 10, 5 (dB)

5. Immersive Audio-Visual Experiments

As mentioned earlier, in this work, we performed three different perceptual experiments. All three experiments used the immersive experimental methodology described by Pinson et al. [16]. With the goal of capturing the real media-consumption experience, this methodology proposes an experimental environment that is supposed to be as natural as possible. The main goal of this methodology is to encourage the subject's engagement with the test content. All experiments used three main aspects of the immersive methodology:

- *Length of stimuli:* Longer stimuli (30–60 s) is used along with the three perceptual experiments. According to what was reported in [16], this time length is considered sufficient to capture the participants' attention, transmit an entire idea, and still maintain a tolerable test duration.
- *Content diversity:* Present each source content only once at each session to prevent fatigue and content memorization. Therefore, the accuracy of the results is associated with the number of different content sources used in the test. In the immersive methodology, for each HRC, each participant should rate five to ten stimuli, which leads to a good estimate of the participant's opinion about each HRC [16];
- *Input media:* Use audio-visual material and ask participants to rate the overall audio-visual quality and the content. This release participants from the task of separating audio and video quality when presented with audio-visual content.

In addition, the immersive methodology poses a specific setup to generate the test stimuli for the experiment. One recommendation is to set the number of stimuli sources (w) in the experiment as an integer multiple of the number of HRCs (y). The combination of each source stimuli and HRC produce a total of ($w \cdot y$) test stimuli. Then, each participant in the experiment rates (w/y) test stimuli for each HRC under study. It is expected that, when all participant's responses are pooled, (n/y) participants have rated each individual test stimuli, where n is the number of human participants.

In this section, we describe the source stimuli, apparatus and physical conditions, experimental procedure, and statistical methods used in the conduction of this set of experiments.

5.1. Source Stimuli

The UnB-AV dataset [47] was used for this study (this dataset is available for download from the site of the University of Brasília at www.ene.unb.br/mylene/databases.html, accessed on 22 June 2021). This dataset contains one-hundred and forty (140) video clips (with accompanying audio) distributed in three sub-sets and used in three different perceptual experiments: Experiment 1 (60 sequences), Experiment 2 (40 sequences) and Experiment 3 (40 sequences). Pristine versions of these video clips were pre-processed to standardize some video and audio characteristics. For the video component, the spatial and temporal resolutions were set to 1280 × 720 (720p) and 30 frames per second (fps), plus the color space configuration was set to 4:2:0. As for the audio component, the sampling frequency and bit-depth were fixed to 48 kHz and 16 bits, respectively. Note that all videos considered for the study had characteristics equal or above the ones set in this pre-processing phase. Video clips were 19 to 68 s long, with an average duration of 36 s. Sample frames of 18 video clips grouped by content genres are presented in Figure 1. Content material was selected based on the recommendations found in [16] and the Final Report on the validation of objective models multimedia quality assessment (phase 1) of the Video Quality Experts Group (VQEG) [57]. A scatter-plot showing the distribution of the spatial and temporal information, defined by Ostaszewska and Kloda [58], for all video clips is presented in Figure 2.

Figure 1. Sample frames of the videos from the UnB-AV Database. The database include different genres contents like: Sports, TV Commercials, Interviews, Music, Documentaries, and Cartoons [47].

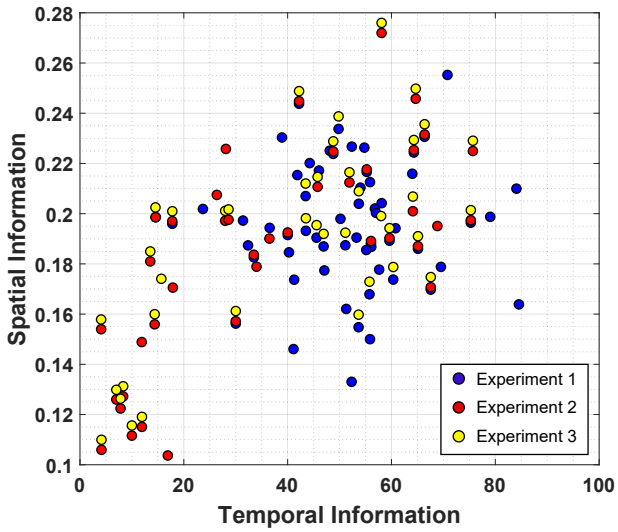


Figure 2. Source videos spatial and temporal information measures.

5.2. Apparatus and Physical Conditions

A recording studio from the Núcleo Multimedia e Internet (NMI) of the Department of Engineering (ENE) of the University of Brasília (UnB) was used to conduct the perceptual experiments. This type of facility guaranteed sound isolation during each experimental session, in which only one participant was admitted. All participants were assigned the same workstation (Table 6). Lighting and viewing conditions were set following the recommendations in ITU-T BT.500.1 [8,59].

Table 6. Detailed specifications of the Experiments 1–3.

Setup			
Monitor	Samsung SyncMaster P2370 Resolution: 1920 × 1080; Pixel-response rate: 2 ms; Contrast ratio: 1000:1; Brightness: 250 cd/m^2		
Earphones	Sennheiser Hd 518 Headfone Impedance: 50 Ohm; Sound Mode: Stereo; Frequency response: 14–26,000 Hz;		
Sound Card	Asus Xonar DGX 5.1		
	Experiment 1	**Experiment 2**	**Experiment 3**
Stimuli			
Source Stimuli	60	40	40
Test Conditions (HRC)	12	20	20
Test Sequences (PVS)	720	800	800
Average length	37 s	35 s	34 s
Scores per PVS	5	2	2
Scores per HRC	300	80	80
Participants			
Total	60	40	42
Male	18	15	16
Female	42	25	26
Age Range	19–36	21–36	20–34

A quality assessment web-based platform, developed by the Grupo de Processamento Digital de Sinais (GPDS), was used for displaying the test clips and collecting the responses from participants. The experiments were carried out with volunteers (mostly graduate students) from the Electrical Engineering and Computer Science Departments from the University of Brasília. No particular expertise was needed in terms of digital video and

audio defects. Although no vision or hearing tests were requested from participants, unimpaired hearing was a pre-requirement. In addition, any participant that wears glasses or contact lenses to watch TV was asked to use them for the experimental session. Details about the participants' genders and ages are presented in Table 6.

5.3. Experimental Procedure

Overall, an entire experimental session was divided into three sub-sessions: Display Session, Training Session, and Main Session. Figure 3 presents an illustration of the sub-sessions used during all three perceptual experiments, which are described next. The same procedure was used in all three experiments.

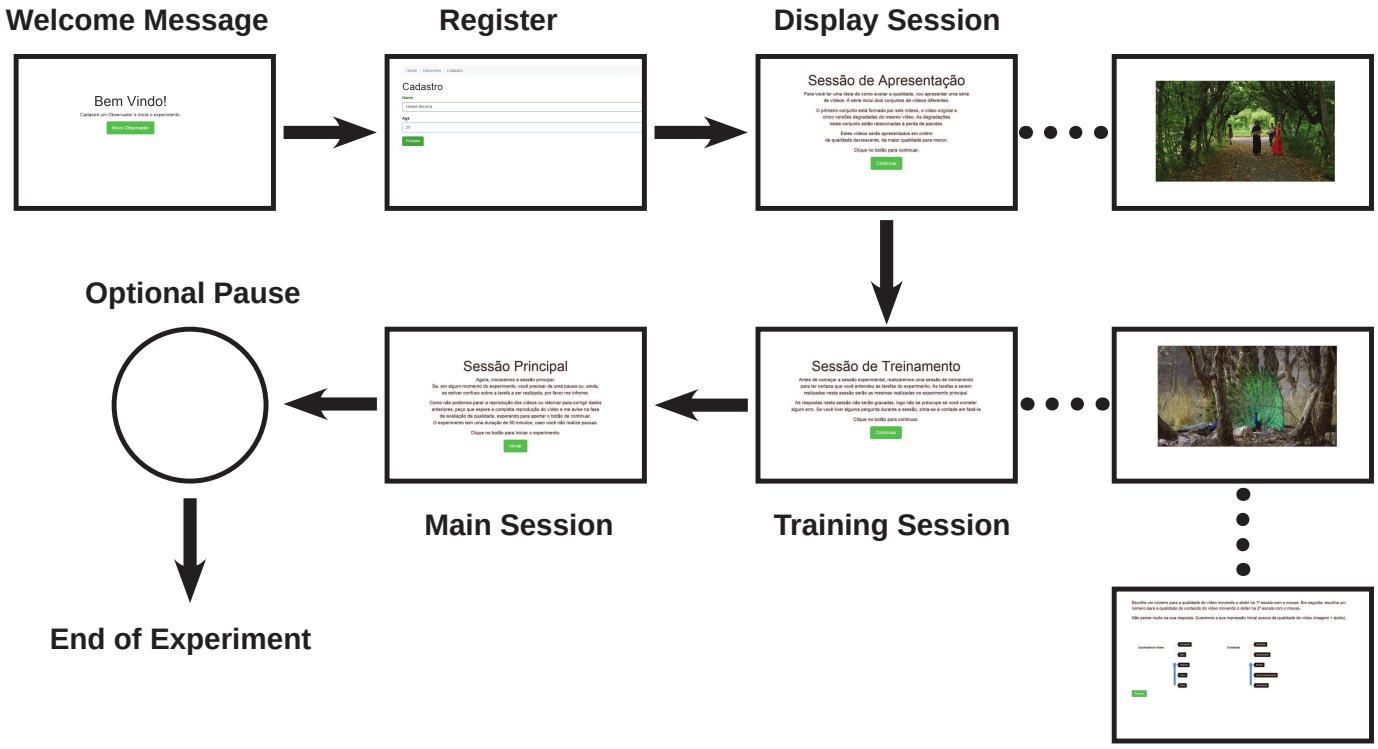


Figure 3. Sub-sessions of the perceptual experiments.

- *Display Session*
  In this session, participants observed a set of original video clips and their distorted versions (processed video sequences—PVS). The session had the objective of familiarizing the participants with the quality range of the sequences in the experimental session. In this session, an original source stimuli sequence and a corresponding PVS were presented to the participant. This procedure was repeated for each HRC of all experiments. Once the display session was over, the experimenter asked the participant if he/she have perceived the differences between the sequences.
- *Training Session*
  In the training session, the participant performed a demonstration run of the main session. For this session, the objective is to train subjects on how to insert their responses using the quality assessment interface. After each test stimuli was presented, participants were asked to rate the overall audio-visual quality using a five-point Absolute Category Rating (ACR) scale, ranging from 1 to 5. The quality scale was labelled (in Portuguese) as "Excellent", "Good", "Fair", "Poor", and "Bad" (Figure 4). A second question is related to the personal opinion of the participant regarding the video clip content. A second five-point ACR scale is presented to the participant with the following labels: "Intriguing", "Interesting", "Neutral", "Uninteresting",

and "Boring" (Figure 4). The labels for the second scale were inspired by the speech experiment reported in [16].

- *Main Session*
  In this session, the actual experimental task was executed. A number of video clips from the entire stimuli pool were presented to the participants in a randomised fashion. No repeated content was allowed; that is, no two videos had the same (source) content. In each session, participants assessed five PVSs corresponding to each HRC, with each PVS being rated by approximately five participants. To avoid fatigue, a break was introduced in the middle of the experiment. Overall, the entire experimental session lasted, on average, 50 min.

Figure 4. ACR Quality and Content scales.

5.4. Statistical Analysis Methods

The scores given by human participants to any test stimuli are called perceptual scores. The perceptual scores from all participants are averaged for each PVS, resulting in a mean opinion score (MOS). For the three experiments, quality and content scores associated with each PVS were gathered. The scores were averaged according to the type of HRC.

The quality scores were processed to generate the Mean Quality Score (MQS) per-HRC, given by:

$$\mathrm{MQS}_{\mathrm{HRC(j)}} = \frac{1}{n} \cdot \sum_{i=0}^{n} S_j(i), \tag{1}$$

where $S_j(i)$ is the score given by the ith subject for the jth element of the set of m HRCs and n is the total number of subjects. In other words, $\mathrm{MQS}_{\mathrm{HRC(j)}}$ gives the average quality score for the j-th HRC, measured across all subjects and content originals.

Similarly, the Mean Content Score (MCS) per-HRC is obtained by taking the average of the content scores given by all subjects:

$$\mathrm{MCS}_{\mathrm{HRC(j)}} = \frac{1}{n} \cdot \sum_{i=0}^{n} CS_j(i), \tag{2}$$

where $CS_j(i)$ is the content score given by the i-th subject to the j-th HRC test sequence, with $j = \{1, 2, \ldots, m\}$.

6. Perceptual Experiment 1 (Video-Only)

In this first experiment, volunteers were presented with a set of audio-visual video clips, and they were asked to rate them based on their perceived quality and content. Three types of distortions were added to these video clips: compression distortions generated by coding the videos at different bitrate levels (and codec algorithms), packet losses transmission distortions generated by deleting video bitstream packets, and frame-freezing transmission distortions generated by deleting and repeating frames. Distortions affected only the video

component; meanwhile, the audio component remained untouched. The experiment focused on analysing the impact of visual distortions on the perceived audio-visual quality.

Sixty (60) source stimuli (out of the 140 stimuli pool described in Section 5) were considered for the experiment. Their video components were processed, generating PVSs with different types of visual impairments. The PVSs were organized in a group of ten (10) different HRCs, as described in Table 7, with no HRC having simultaneously packet-loss and frame-freezing distortions. For this reason, two test scenarios were considered for Experiment 1: a coding-packetloss scenario (HRC1 to HRC5) and a coding-freezing scenario (HRC6 to HRC10). In order to help participants recognize the range of quality in the experiment, two anchor (ANC) conditions were included. These ANCs corresponded to video clips compressed at extremely high bitrate levels, with no packet-loss or frame freezing distortions. These twelve (12) test conditions (10 HRCs plus 2 ANCs) were replicated for all sixty (60) source stimuli, resulting in seven hundred and twenty (720) test stimuli. As mentioned before, for each experimental session, participants saw the content corresponding to each original sequence only once. That is, each participant was presented with only 60 test stimuli, out of the 720 available, all corresponding to different source contents. These 720 sequences compound the first part of the UnB-AV dataset [47]. The sub-set is labelled as UnB-AV-Experiment-1, and it is available for download along with the entire UnB-AV dataset.

Table 7. HRC and ANC corresponding parameters used in Experiment 1.

HRC	Codec	Bitrate (kbps)	Packet-Loss Rate (PLR)	Freezing Pauses		
				Number	Length	Position
$HRC1_{E1}$	H.264	500	10%	-	-	-
$HRC2_{E1}$	H.265	400	8%	-	-	-
$HRC3_{E1}$	H.264	2000	5%	-	-	-
$HRC4_{E1}$	H.265	1000	3%	-	-	-
$HRC5_{E1}$	H.265	8000	1%	-	-	-
$HRC6_{E1}$	H.265	200	-	3	3, 3, 2	1, 2, 3
$HRC7_{E1}$	H.264	800	-	3	2, 2, 3	1, 2, 3
$HRC8_{E1}$	H.265	1000	-	2	2, 2	2, 3
$HRC9_{E1}$	H.264	2000	-	2	1, 3	1, 3
$HRC10_{E1}$	H.264	16,000	-	1	2	2
$ANC1_{E1}$	H.264	64,000	-	-	-	-
$ANC2_{E1}$	H.265	32,000	-	-	-	-

Figure 5 presents the quality scores given by participants for each of the HRCs. For each HRC, the MQS_{HRC} value is represented by the larger dot in the middle. It can be observed that for most test conditions, the responses are consistent, i.e., the spread of points is small. Moreover, test conditions with higher quality scores (e.g., ANC1, ANC2, and HRC10) presented a higher agreement among participants. In general, it can be observed that participants used the entire scale presented to them (1 to 5).

Figure 6a,b shows the Mean Quality Score (MQS_{HRC}) and Mean Content Score (MCS_{HRC}), respectively, for Experiment 1. For each HRC, a confidence interval of 95% is included, plus all HRCs are grouped according to Frame-Freezing and Packet-Loss degradations. A combination of a bitrate value (BR) plus a packet-loss ratio (PLR) or frame-freezing configuration (number, length and position) is assigned to each HRC (see Table 7). The MQS_{HRC} values (presented in Figure 6a) fell between 1.92 and 4.5, with no evidence of a scale saturation. As expected, MQS_{HRC} increases with the strength of the bitrate, packet-loss, and frame-freezing degradations. That is, participants were able to distinguish between different degradation levels inserted in the visual component.

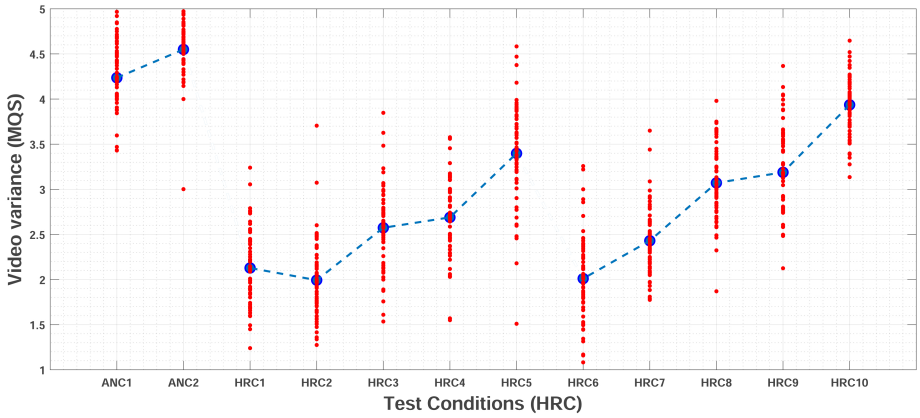
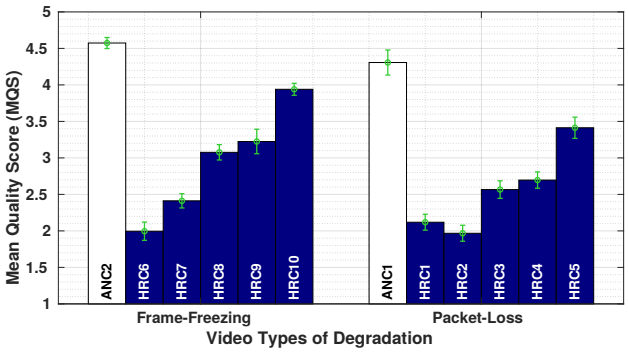
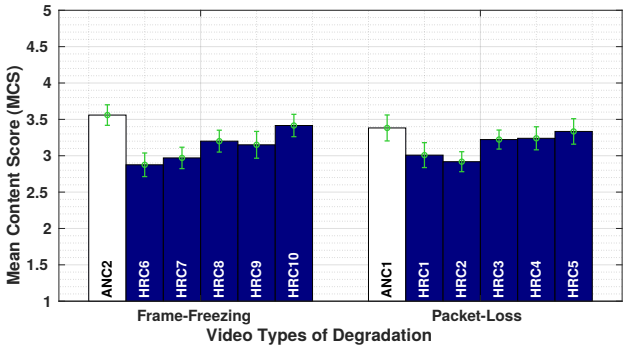


Figure 5. Mean Quality Score (MQS), and its respective spread of scores, for the different Hypothetical Reference Circuit (HRC) in Experiment 1.



(**a**) Mean Quality Score—MQS



(**b**) Mean Content Score—MCS

Figure 6. Mean Quality Score (MQS_{HRC}) and Mean Content Score (MCS_{HRC}) for the different Hypothetical Reference Circuit (HRC) in Experiment 1.

By observing the Frame-Freezing HRC group in Figure 6a, we notice that HRC8 and HRC9 presented similar results in terms of MQS_{HRC}. These values can be explained by revising the corresponding HRC parameters in Table 7. In terms of bitrate, studies show that a video encoded with H.264 at a certain bitrate has approximately the same quality as a video encoded with H.265 using half bitrate [53,60]. Then, for the particular case of HRC8 (Codec = H.265 and BR = 1000 kbps) and HRC9 (Codec = H.264 and BR = 2000 kbps), a certain quality equivalence is expected. In terms of frame-freezing, both HRCs had the same number of pauses; then, it can be inferred that the small MQS_{HRC} difference between HRC8 and HRC9 is due to the position (P) and the length (L) of the pause events. For HRC8, the pauses were inserted at positions '2' and '3', both with a duration of 2 s. Meanwhile, HRC9 had pauses inserted at positions '1' and '3', with a duration of 1 and 3 s, respectively. The slightly better MQS_{HRC} obtained by HRC9 (in comparison to HRC8) can be explained

with the results reported in [61], which state that short pauses at the beginning of the video playout (initial loading) are less annoying than the pauses occurring in the middle of the playout (stalling).

With regard to the Packet-Loss group in Figure 6a, HRC3 and HRC4 presented similar results in terms of MQS_{HRC}. As in the previous case, it can be inferred that the small MQS_{HRC} difference is due to the Packet-Loss ratio (PLR) since a certain equivalence is expected in terms of bitrate (HRC3: Codec = H.264 and BR = 2000 kbps; HRC4: Codec = H.265 and BR = 1000 kbps). Studies have shown that H.265 is more sensitive to packet-loses than H.264 [62,63]. For this case, HRC3 and HRC4 have different PLRs, 5% and 3% , respectively. Then, the MQS_{HRC} of a higher PLR (5% with H.264 for HRC3) is slightly smaller than the one of a lower PLR (3% with H.265 for HRC4).

A comparison across groups shows that the packet-loss has a stronger effect on the perceived quality than the frame-freezing. This can be verified by observing the MQS_{HRC} values from HRC3 versus HRC9 and HRC4 versus HRC8. For these two cases, the coding parameters are exactly the same. The main difference in MQS_{HRC} values is due to the packet-loss and frame-freezing parameters. Then, we conclude that pauses during the video playout were less annoying to the participants than severe visual distortions caused by packet losses.

Figure 6b presents the MCS_{HRC} values for each HRC, corresponding to both frame-freezing and packet-loss groups. It is clear that the range of MCS_{HRC} is much smaller than the range of MQS_{HRC}, which fluctuates around '3' ("Neutral" Content). The small differences between the HRCs and ANCs (anchor sequences without distortions) show that participants did not perceive great differences, in terms of content, between degraded and original sequences. There are, however, small variations in the MCS_{HRC} values that somehow follow the MQS_{HRC} behaviour, at a smaller range. This suggests that there is a certain correspondence between quality and content, which is in agreement with the results obtained in previous studies [64,65].

7. Perceptual Experiment 2 (Audio-Only)

For this experiment, the perceived quality responses were collected for a set of audio-visual video clips degraded with audio distortions only. The experiment had the goal of recreating four common streaming audio degradations from the TCD-VoIP dataset [30]: Background noise, Clipping, Chop, and Echo. These degradations were inserted into the audio components of a set of audio-visual sequences. The goal was to analyze the effect of such degradations on the perceived audio-visual quality.

Forty (40) source stimuli (out of the entire 140 stimuli pool) were considered to build an audio-visual dataset, replicating the sequence processing method used in the TCD-VoIP dataset that was taken as a reference for this experiment. For each degradation type, four (4) single test conditions were selected and presented as a particular HRC. The selection of these test conditions was empirical, aiming to cover the quality range of the TCD-VoIP dataset. This resulted in sixteen (16) HRCs organized according to the type of degradation, as described in Table 8. Additionally, four test conditions without degradations were used as anchors (ANCs) to help participants establish the range of quality used in the experiment. All twenty (20) test conditions (16 HRCs plus 4 ANCs) were replicated for all forty (40) source stimuli, resulting in eight hundred (800) test stimuli. These 800 sequences compound the second part of the UnB-AV dataset [47]. The sub-set is labelled as UnB-AV-Experiment-2, and it is available for download along with the entire UnB-AV dataset.

Figure 7 presents the average quality scores obtained for each HRC (i.e., the MQS_{HRC}), along with the individual scores given by each participant of Experiment 2. Results are organized according to the type of audio distortion. For each HRC, the MQS_{HRC} value is represented by the larger dot in the middle. It can be observed that, contrary to what was observed in Experiment 1, most of the responses gathered were disperse along the quality scale. Moreover, there is more agreement among participants for quality scores given to the anchors conditions (ANC1, ANC2, ANC3 and ANC4). This suggests that there was

less agreement regarding the different audio distortions when compared to the agreement observed in Experiment 1 for the different visual distortions.

In Figure 8a,b, the Mean Quality Score (MQS_{HRC}) and Mean Content Score (MCS_{HRC}) are presented. For each HRC, a confidence interval of 95% is included, plus all HRCs are grouped according to Background Noise, Chop, Clip, and Echo distortions. Each HRC is assigned with a combination of parameters according to the type of distortion (see Table 8). The MQS_{HRC} values are between 1.5 and 3.9, depending on the types of distortions and their degradation levels. With the exception of Echo distortions, the MQS_{HRC} range occupied only about 30% of the scale. This means that participants had difficulties distinguishing the quality levels for different types of audio distortions. Naturally, this effect was different for different types of audio distortions.

Table 8. HRC and ANC corresponding parameters used in Experiment 2.

BG Noise	**Noise**	**SNR (dB)**	
$HRC1_{E2}$	car	15	
$HRC2_{E2}$	babble	10	
$HRC3_{E2}$	office	10	
$HRC4_{E2}$	road	5	
$ANC1_{E2}$	-	-	
Chop	**Period (s)**	**Rate (Chops/s)**	**Mode**
$HRC5_{E2}$	0.02	1	previous
$HRC6_{E2}$	0.02	2	zeros
$HRC7_{E2}$	0.04	2	previous
$HRC8_{E2}$	0.02	5	zeros
$ANC2_{E2}$	-	-	-
Clipping		**Multiplier**	
$HRC9_{E2}$		11	
$HRC10_{E2}$		15	
$HRC11_{E2}$		25	
$HRC12_{E2}$		55	
$ANC3_{E2}$		-	
Echo	**Alpha (%)**	**Delay (ms)**	**Feedback (%)**
$HRC13_{E2}$	0.5	25	0
$HRC14_{E2}$	0.3	100	0
$HRC15_{E2}$	0.175	140	0.8
$HRC16_{E2}$	0.3	180	0.8
$ANC4_{E2}$	-	-	-

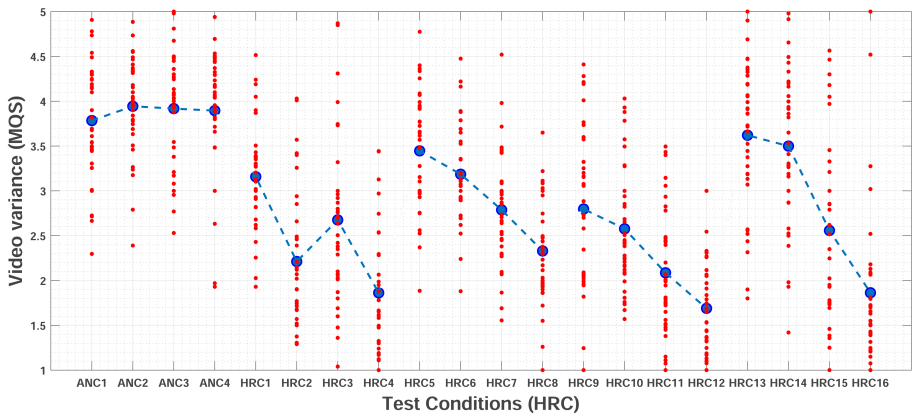
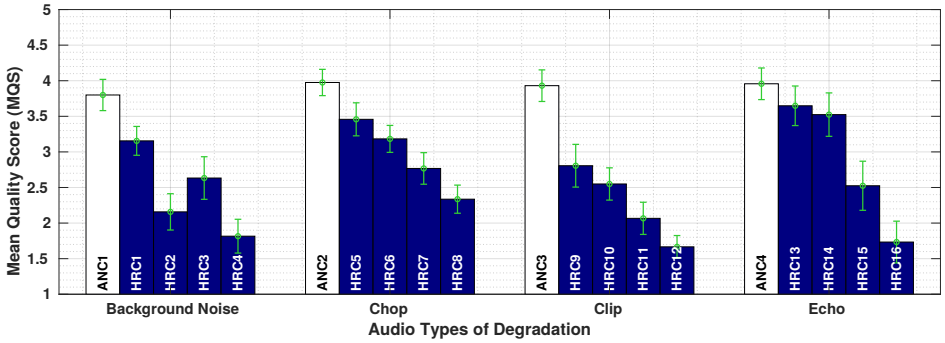
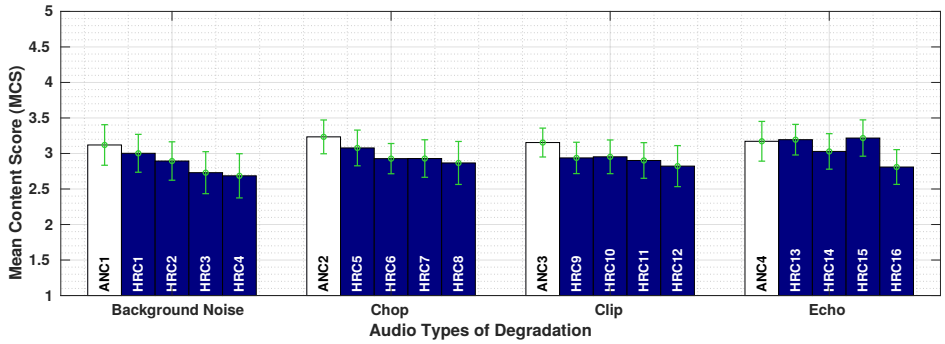


Figure 7. Mean Quality Score (MQS), and its respective spread of scores, for the different Hypothetical Reference Circuit (HRC) in Experiment 2.

For Background Noise, combining the type of noise and the SNR value associated with it resulted in four HRCs (Table 8). By observing the results for HRC2 (noise = bable, SNR = 10 dB) and HRC3 (noise = office, SNR = 10 dB), we notice that the babble noise was perceived by participants as more annoying than the office noise. Similar results were observed in a previous study using only the audio component [30]. For Chop, each HRC corresponds to the combination of three parameters: rate, period, and mode (Table 8). The reported MQS_{HRC} values vary from 3.5 to 2.5, decreasing from HRC5 to HRC8. Notice that the MQS_{HRC} values decrease as the chop rate increases, independent of the period or the chop mode. For the particular case of HRC6 (rate = 2 chops/s, mode = zeros) and HRC7 (rate = 2 chops/s, mode = previous), repeating previous portions of samples (previous mode) was slightly more annoying than inserting silence portions (zeros mode).



(**a**) Mean Quality Score



(**b**) Mean Content Score

Figure 8. Mean Quality Score (MQS_{HRC}) and Mean Content Score (MCS_{HRC}) for the different Hypothetical Reference Circuit (HRC) in Experiment 2.

For Clip, a multiplier factor was the only parameter assigned to each HRC. The reported MQS_{HRC} values varied from 3 to 1.5, decreasing from HRC9 to HRC12. Clip results presented the lower MQS_{HRC} values among all 4 types of audio distortions, which indicates that clipped distortions were perceived as the most annoying ones. Finally, for Echo, each HRC was associated with the combination of three parameters: alpha, delay, and feedback (Table 8). The reported MQS_{HRC} values vary from 3.7 and 1.7, decreasing from HRC13 to HRC16 with an abrupt drop in the MQS_{HRC} between HRC14 and HRC15. This drop might be related to the inclusion of a feedback percentage, which affected the perceived quality in a similar way as observed in previous studies [30].

Figure 6b presents the MCS_{HRC} values for each HRC, corresponding to the four audio distortions groups. As observed in the previous MQS_{HRC} results, the range where these values vary is very small, fluctuating around '3' ("Neutral" Content). Almost no difference was observed between scores for distorted conditions (HRCs) and no-distorted conditions (ANCs). This indicates that, for this experiment, participants were not able to trace a correspondence between the perceived quality and the content of the sequence.

8. Perceptual Experiment 3 (Audio-Visual)

In the last experiment, the goal was to estimate the overall quality of audio-visual video clips containing combinations of audio and video distortions. The video distortions were Bitrate compression, Packet-Loss, and Frame-Freezing, replicating the distortions in Experiment 1. The audio distortions were Background noise, Chop, Clip, and Echo, replicated conditions of Experiment 2.

Forty (40) source stimuli were considered for this experiment. Both audio and video degradations were organized into sixteen (16) different HRCs. As with previous experiments, 4 anchors (ANCs) were included in the HRC set. Parameter details of both HRCs and ANCs are presented in Table 9. Altogether, 40 source stimuli were processed at 20 different test conditions (16 HRCs plus 4 ANCs). This resulted in 800 PVSs, containing different audio and video distortions. In each experimental session, the participant watched (only) 40 test stimuli out of the 800 test sequences, as recommended by the immersive methodology. These 800 sequences compose the third part of the UnB-AV dataset [47]. The sub-set is labelled as UnB-AV-Experiment-3 and is available for download along with the entire UnB-AV dataset.

Table 9. HRC and ANC parameters used in Experiment 3.

	Audio Component				Video Component			
HRC	**Noise, SNR (dB)**	**Chop-Period (s), Rate (Chop/s), Mode**	**Clip Multiplier**	**Echo-Alpha (%), Delay (ms), Feedback (%)**	**Video Codec**	**Bitrate (kbps)**	**PacketLoss PLR**	**Freezing Pauses, Length (s)**
$HRC1_{E3}$	car, 15	-	-	-	H.264	16,000	-	1, 2
$HRC2_{E3}$	-	-	11	-	H.264	16,000	-	1, 2
$HRC3_{E3}$	-	-	11	-	H.265	8000	0.01	-
$HRC4_{E3}$	-	0.02, 2, zeros	-	-	H.265	8000	0.01	-
$HRC5_{E3}$	-	-	-	0.3, 100, 0	H.264	16,000	-	1, 2
$HRC6_{E3}$	office, 10	-	-	-	H.264	16,000	-	1, 2
$HRC7_{E3}$	-	-	-	0.3, 100, 0	H.265	8000	0.01	-
$HRC8_{E3}$	-	-	-	0.3, 100, 0	H.264	2000	0.05	-
$HRC9_{E3}$	office, 10	-	-	-	H.264	2000	0.05	-
$HRC10_{E3}$	office, 10	-	-	-	H.264	800	-	3, 7
$HRC11_{E3}$	-	-	25	-	H.264	2000	0.05	-
$HRC12_{E3}$	-	-	25	-	H.264	800	-	3, 7
$HRC13_{E3}$	-	-	25	-	H.265	400	0.08	-
$HRC14_{E3}$	-	0.02, 5, zeros	-	-	H.265	400	0.08	-
$HRC15_{E3}$	-	-	-	0.3, 180, 0.8	H.264	800	-	3, 7
$HRC16_{E3}$	-	-	-	0.3, 182, 0.8	H.265	400	0.08	-
$ANC1_{E3}$	-	-	-	-	H.264	64,000	-	-
$ANC2_{E3}$	-	-	-	-	H.265	32,000	-	-
$ANC3_{E3}$	-	-	-	-	H.264	64,000	-	-
$ANC4_{E3}$	-	-	-	-	H.265	32,000	-	-

Figure 9 presents the average quality scores obtained for each HRC (i.e., the MQS_{HRC}), along with the individual scores given by each participant of Experiment 3. For each HRC, the MQS_{HRC} value is represented by the larger dot in the middle of the cloud of values. Notice that there is a certain consistency among results for different test conditions. This characteristic is more evident for extreme test conditions, i.e., test conditions with the higher (ANC1, ANC2, ANC3, and ANC4) and lower (HRC13, HRC14, HRC15, HRC16) quality levels.

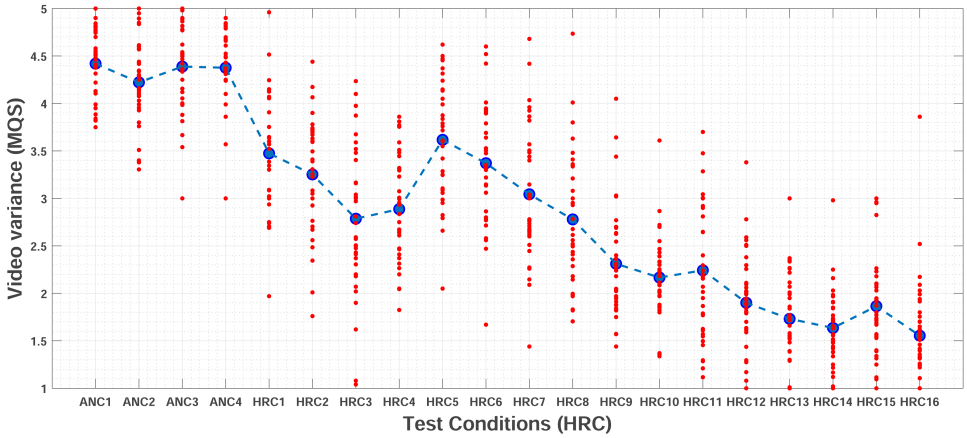
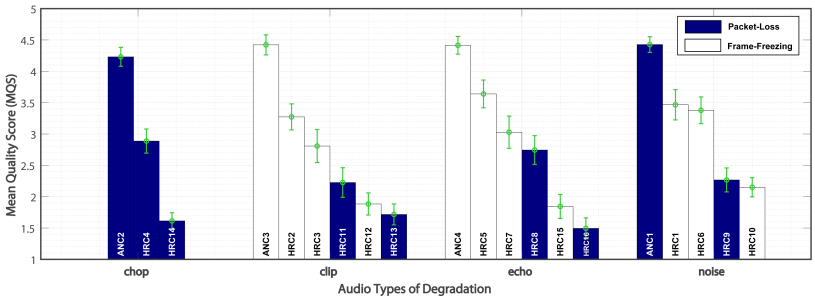


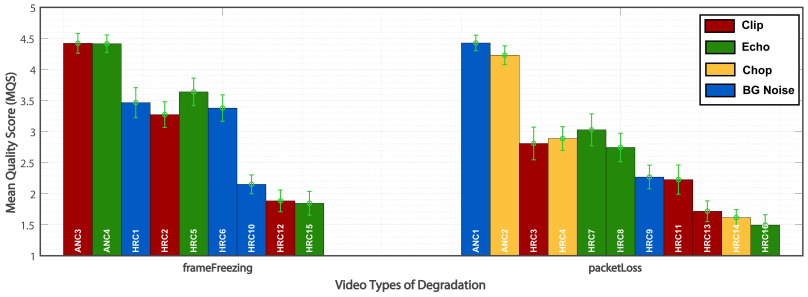
Figure 9. Mean Quality Score (MQS), and its respective spread of scores, for the different Hypothetical Reference Circuit (HRC) in Experiment 3.

In Figure 10a,b, the MQS_{HRC} values gathered for each HRC in Experiment 3 are presented. For better visualization, MQS_{HRC} results were presented in two figures (a) and (b) organized by audio and video types of distortions, respectively. For each HRC, a confidence interval of 95% is included. From a general view, it can be observed that most MQS_{HRC} values fell in the range of 4.4 and 1.5 in the MQS scale. By observing the MQS_{HRC} values for the different audio distortions in Figure 10a, we notice that values vary along all different HRCs, which indicates that participants were able to perceive the variations among the different levels of quality when both audio and video degradations are present.

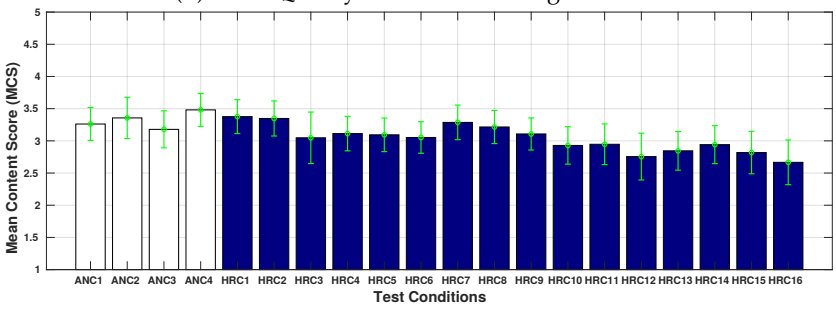
With respect to the Chop type of distortion, the significant difference between HRC4 and HRC14, both with packet-losses, was determined by the combination of a chop rate (HRC4: 2 chops/s, HRC14: 5 chops/s) plus the BR and PLR parameters (HRC4: 8000 kbps, 1%, HRC14: 400 kbps, 8%). This, in accordance with the effects observed by these parameters in Experiments 1 and 2. For the Clip type of distortion, MQS_{HRC} values decrease progressively from HRC2 to HRC13. For the case of HRC2 and HRC3, where the clip multiplier is fixed in 11, an equivalence is expected in terms of bitrate (HRC2: H.264, 16,000 kbps, HRC3: H.265, 8000 kbps). The difference in the quality score was defined by the stronger effect of packet-losses over frame-freezing pauses (HRC2: number = 1, length = 2 s, HRC3: PLR = 1%). This is also the case for HRC12 and HRC13 (see Table 9). Equivalence in terms of bitrate and a certain predominance of packet-loss over frame-freezing is too observed for the Echo type of distortion. More specifically, for HRC5-HRC7 and HRC15-HRC16, both with fixed Echo parameters (see Table 9). For the Background Noise type of distortion, results for HRC1 and HRC6 were almost equivalent. For this case, most of the audio and video parameters were the same (see Table 9), the small difference (no statistical significance) between both test conditions was determined by the car noise type at 15 dB (HRC1) being less annoying than the office noise at 10 dB (HRC6). For the case of HRC9 and HRC10, with fixed parameters of audio (office noise with 10 dB), results were also equivalent, and the small difference was due to the coding parameters (HRC9: 2000 kbps, HRC10: 800 kbps) and the slightly stronger annoyance caused by the frame-freezing pauses over the packet losses (HRC9: PLR = 5%, HRC10: number = 3, length = 7 s). Finally, by observing the MQS_{HRC} values for the different video distortions in Figure 10b, we notice that frame-freezing distortions were less annoying to participants than packet-losses. This is in agreement with the results from Experiment 1.



(**a**) Mean Quality Score—Video Degradations



(**b**) Mean Quality Score—Audio Degradations



(**c**) Mean Content Score

Figure 10. Mean Quality Score (MQS_{HRC}) and Mean Content Score (MCS_{HRC}) for the different Hypothetical Reference Circuit (HRC) in Experiment 3.

Figure 10c presents the MCS_{HRC} values for each HRC gathered in Experiment 3. As in Experiments 1 and 2, MCS_{HRC} values vary in a small range, fluctuating around the value '3' ("Neutral" Content).

9. Discussion and Comparison among Experiments

In this Section, quality and content responses across all three perceptual experiments are compared. We explore the results for equivalent test conditions in Experiments 1, 2 and 3. To this end, only the quality scores of the equivalent test conditions from each experiment are compared. The objective is to study the responses of equivalent test conditions used in different experimental scenarios (video-only distortions, audio-only distortions, and Audio + Video distortions). This analysis helps us understand how and at what level a particular test condition is affected by its accompanying component (impaired or unimpaired).

For this particular analysis, labels assigned to the HRCs of Experiments 1, 2 and 3, were redefined. The goal was to compare equivalent HRCs across all three experiments. For this purpose, the term Video Test Condition (V-TC) was used to denote the test conditions for the video component (Table 10). In the same way, the term Audio Test Condition (A-TC) was used to denote the audio test conditions for the audio component (Table 11). These terms replaced the HRC labels used in previous sections. The analysis is divided into three parts: (1) the visual and (2) auditory component effects for equivalent test conditions, and (3) ranges of quality and content scores for the three experiments. This Section includes an

additional analysis using external data from audio-visual studies available in the literature. This analysis estimates the internal consistency of all three experiments from this study and compares it with results from external audio-visual studies.

Table 10. Video Test Condition (V-TC) corresponding parameters used in Experiment 1, 2, and 3.

Packet-Loss			
Video Test Condition	**Codec**	**Bitrate (kbps)**	**PLR**
V-TC1	H.264	500	10%
V-TC2	H.265	400	8%
V-TC3	H.264	2000	5%
V-TC4	H.265	1000	3%
V-TC5	H.265	8000	1%
V-TC0	H.264	64,000	-
Frame-Freezing			
Video Test Condition	**Codec**	**Bitrate (kbps)**	**Freezing**
V-TC6	H.265	200	S5
V-TC7	H.264	800	S4
V-TC8	H.265	1000	S3
V-TC9	H.264	2000	S2
V-TC10	H.264	16,000	S1
V-TC0	H.265	32,000	-

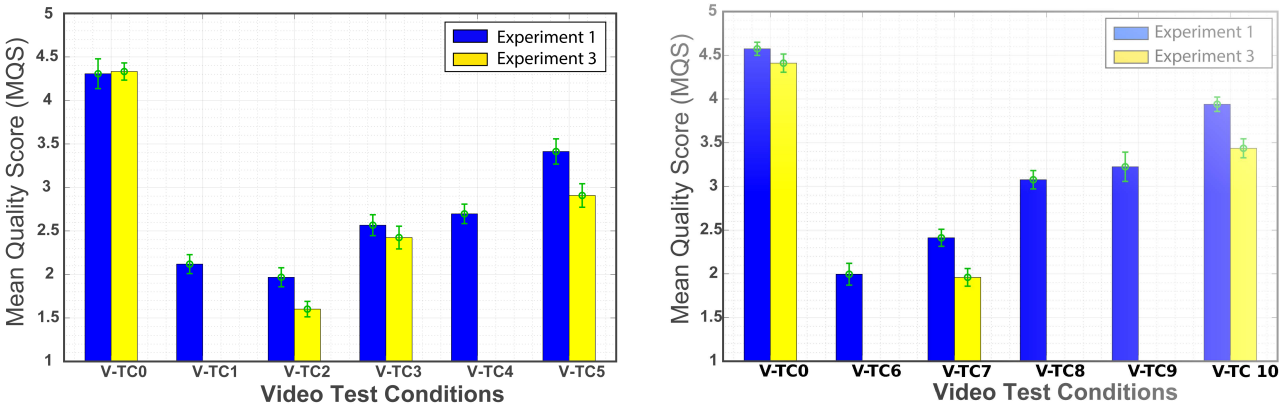
Table 11. Audio Test Condition (A-TC) corresponding parameters used in Experiment 1, 2, and 3.

Background Noise		
Audio Test Condition	**Noise**	**SNR (dB)**
A-TC1	car	15
A-TC2	babble	10
A-TC3	office	10
A-TC4	road	5
A-TC0	-	-
Chop		
Audio Test Condition	**Period (s)**	**Rate (chops/s)**	**Mode**
A-TC5	0.02	1	previous
A-TC6	0.02	2	zeros
A-TC7	0.04	2	previous
A-TC8	0.02	5	zeros
A-TC0	-	-	-
Clipping		
Audio Test Condition	**Multiplier**	
A-TC9	11	
A-TC10	15	
A-TC11	25	
A-TC12	55	
A-TC0	-	
Echo		
Audio Test Condition	**Alpha (%)**	**Delay (ms)**	**Feedback (%)**
A-TC13	0.5	25	0
A-TC14	0.3	100	0
A-TC15	0.175	140	0.8
A-TC16	0.3	180	0.8
A-TC0	-	-	-

9.1. On the Visual Component Effect

This analysis explores the effect of the visual impairments for equivalent V-TCs in Experiments 1 and 3. Figure 11 presents a comparison between the MQS_{VTC} values (MQS for different V-TCs) collected from these experiments. The MQS_{VTC} values were grouped according to distortion type: packet-loss and frame-freezing. To verify if the differences

between the average MQS values for the equivalent test conditions in Experiment 1 and Experiment 3, a Two-sample *t*-test analysis was made. Table 12 reports this analysis and highlights the equivalent conditions where the difference between results in Experiment 1 and Experiment 3 are significantly different. Figure 11a depicts the results for the packet-loss distortion. Notice that, considering the same V-TCs, the scores were lower when the audio component was impaired (Experiment 3). This effect is more noticeable for V-TC2 and V-TC5 (differences are statistically significant). Figure 11b shows the same comparison for the frame-freezing distortion. More specifically, for V-TC7 and V-TC10 the scores for sequences with audio distortions are lower (differences are statistically significant). Audio distortions affected the audio-visual quality of sequences with packet-loss and frame-freezing distortions similarly. These results show that there is an impact of the audio quality on the perceived audio-visual quality. Moreover, given that the analysis is made per test condition, it looks like this impact did not depend on the content. However, a larger number of test conditions are necessary to analyze the effect of the media content.



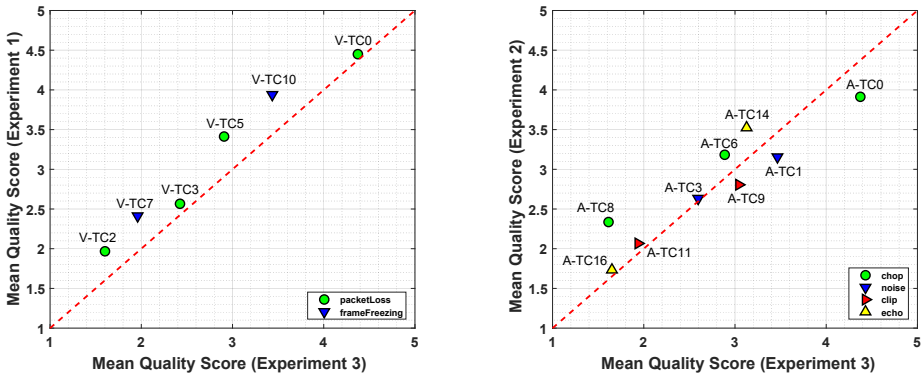
(**a**) MQS—Packet-Loss (**b**) MQS—Frame-Freezing

Figure 11. Average MQS values for different video test conditions in Experiments 1 and 3.

Table 12. Two-sample *t*-test analysis between equivalent conditions from Experiment 1 and Experiment 3 (st. dev.: Standard deviation, *t*: Test statistic, df: Degrees of freedom). Conditions with statistically significant differences are highlighted.

Condition	st. dev.	*t*	df	*ρ*-Value	95% CI
V-TC0	0.543	−0.850	126	0.397	[−0.272, 0.109]
V-TC1	-	-	-	-	-
V-TC2	**0.454**	**4.912**	**168**	**0.000**	**[0.214, 0.502]**
V-TC3	0.622	1.280	168	0.202	[−0.069, 0.325]
V-TC4	-	-	-	-	-
V-TC5	**0.650**	**4.708**	**162**	**0.000**	**[0.288, 0.704]**
V-TC0	**0.376**	**2.533**	**128**	**0.013**	**[0.037, 0.299]**
V-TC6	-	-	-	-	-
V-TC7	**0.477**	**5.919**	**164**	**0.000**	**[0.304, 0.608]**
V-TC8	-	-	-	-	-
V-TC9	-	-	-	-	-
V-TC10	**0.561**	**5.883**	**194**	**0.000**	**[0.340, 0.683]**

Results of equivalent test conditions from Experiments 1, 2, and 3 are compared in Figure 12. The scatter plot from Figure 12a shows a positive correlation between results from Experiment 1 and 3, where video distortions are compared. The plot also shows that scores from Experiment 1, with video distortions only, occupied a larger range in comparison to results from Experiment 3 (audio and video distortions). All markers appear above the red line independent of the type of video distortion (packet-loss or frame-freezing).



(**a**) (**b**)

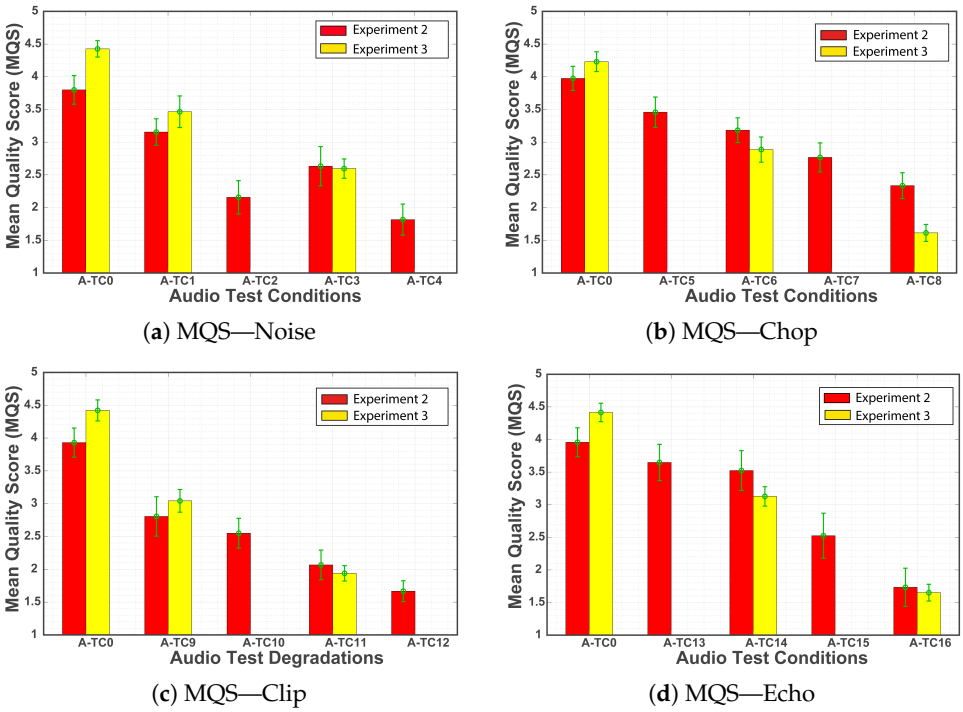
Figure 12. Comparison of MQS values values for: (**a**) Experiment 1 versus Experiment 3 and (**b**) Experiment 2 versus Experiment 3.

9.2. On the Auditory Component Effect

This analysis investigates the effect of audio impairments for equivalent A-TCs in Experiments 2 and 3. Figure 13 shows a comparison between the MQS_{ATC} values (MQS among different A-TCs) gathered from these experiments. Similarly to what was done in the previous analysis, we organized the results according to the audio distortion (background noise, chop, clip and echo). The differences between the average MQS values for the equivalent test conditions in Experiment 2 and Experiment 3 were verified using a Two-sample *t*-test analysis. Table 13 reports this analysis and highlights the equivalent conditions where the difference between results in Experiment 2 and Experiment 3 are significantly different. Figure 13a shows the results obtained for the background noise distortion. For this scenario, the equivalent test conditions A-TC0 and A-TC1 reported statistically significant differences, as for the case of A-TC3, results were very similar for both experiments. Considering that Experiment 3 included visual distortions in addition to the audio distortions in Experiment 2, results in Figure 13a suggest that the background noise (audio-only distortions) affected the overall audio-visual quality at a similar level than the background noise plus the visual distortion (audio + video distortions).

Similarly, Figure 13b presents the quality scores sequences with chop distortions. Results for equivalent test conditions A-TC6 and A-TC8 reported a significant difference in the *t*-test analysis. More particularly, results for A-TC8 shows that there is a difference between results with and without video distortions. This might suggest that chop type distortions by themselves do not have a strong impact on the overall quality. Figure 13c presents the scores for clipping distortions. These results are very similar to the ones observed in the background noise scenario. These results suggest that the combination of clipping and visual distortions (Audio + Video distortions) affected the audio-visual quality at the same level that the clipping distortion on its own (audio-only distortion). None of the equivalent test conditions reported significant differences in the *t*-test analysis.

Figure 13d presents the results for echo distortions. Test condition A-TC14, which reported a significant difference, suggests that video distortions had a higher impact on the perceived quality. As for test condition A-TC16, results seem to suggest that the audio distortion (echo) and the audio plus video distortion from Experiment 3 had an equivalent impact on the perceived quality. However, this test condition did not report a significant difference in the *t*-test analysis.



(**a**) MQS—Noise (**b**) MQS—Chop

(**c**) MQS—Clip (**d**) MQS—Echo

Figure 13. Average MQS values for different audio test conditions in Experiments 2 and 3.

One particular detail that is common to all four audio distortion scenarios is the differences between the equivalent V-TC0 test conditions. Table 13 reports that such differences are statistically significant, plus, in all four cases, a relative higher MQS average was reported for Experiment 3. One possible interpretation of these results is that commonly, speech and audio perceptual tests tend to present a more narrow range of quality when compared to a video perceptual study [7]. Indeed, Experiment 2 contains distortions only in the audio component; meanwhile, Experiment 3 combines both audio and video distortions. This is confirmed by results presented in Figure 11b, where V-TC0 reported a significant difference, only this time it was Experiment 1 (video-only distortions) that had a higher MQS average when compared to Experiment 3.

Table 13. Two-sample *t*-test analysis between equivalent conditions from Experiment 2 and Experiment 3 (st. dev.: Standard deviation, *t*: Test statistic, df: Degrees of freedom). Conditions with statistically significant differences are highlighted.

Condition	st. dev.	*t*	df	*ρ*-Value	95% CI
A-TC0	**0.528**	**−4.988**	**67**	**0.000**	**[−0.889, −0.381]**
A-TC1	**0.629**	**−2.056**	**65**	**0.044**	**[−0.624, −0.009]**
A-TC2	-	-	-	-	-
A-TC3	0.804	0.353	142	0.725	[−0.249, 0.357]
A-TC4	-	-	-	-	-
A-TC0	**0.491**	**−2.373**	**68**	**0.020**	**[−0.513, −0.044]**
A-TC5	-	-	-	-	-
A-TC6	**0.555**	**2.247**	**68**	**0.028**	**[0.033, 0.563]**
A-TC7	-	-	-	-	-
A-TC8	**0.487**	**6.001**	**69**	**0.000**	**[0.463, 0.925]**
A-TC0	**0.556**	**−3.518**	**67**	**0.001**	**[−0.738, −0.204]**
A-TC9	0.786	−1.358	105	0.178	[−0.538, 0.101]
A-TC10	-	-	-	-	-
A-TC11	0.631	1.041	143	0.300	[−0.112, 0.360]
A-TC12	-	-	-	-	-
A-TC0	**0.557**	**−3.664**	**70**	**0.000**	**[−0.7430, −0.219]**
A-TC13	-	-	-	-	-
A-TC14	**0.794**	**2.365**	**136**	**0.019**	**[0.060, 0.6750]**
A-TC15	-	-	-	-	-
A-TC16	0.673	1.199	109	0.233	[−0.107, 0.434]

As for the correlation of responses between Experiment 2 and Experiment 3, Figure 12b shows that there is a small positive correlation between quality scores of Experiments 2 and 3. For most test conditions, markers appear above the red line, which indicated that responses for Experiment 2 were higher. Yet, some test conditions (A-TC 9 and A-TC1) of Experiment 2 presented lower quality scores than their equivalent test conditions in Experiment 3. All in all, results showed that some types of audio distortions have a greater impact on quality. More specifically, background noise and clipping distortions presented a similar impact on the overall audio-visual quality when compared to their equivalent test conditions with visual distortions.

9.3. On the Quality and Content Range of Assessment

This analysis is focused on comparing the quality and content responses of Experiments 1, 2, and 3. To this end, we gather the quality and content range from all three Experiments. Figure 14 presents a comparison of the ranges of the MQSs gathered in Experiments 1, 2, and 3. It can be observed that the MQS ranges for Experiments 1 and 2 (2.58 and 2.4, respectively) are similar. However, there is a negative difference of 0.6 points (maximum range) and 0.42 points (minimum range) in the MQS scale from Experiment 1 to Experiment 2. More interestingly, by comparing the MQS ranges of Experiments 1–3, it can be observed that the range of Experiment 3 overlaps the ranges of Experiments 1 and 2. More specifically, the range of Experiment 3 varies from the minimum range limit of Experiment 2 (1.5) to almost the maximum range limit of Experiment 1 (4.4). This behaviour can also be observed for the MCS range in Figure 14, but at a smaller intensity. As mentioned earlier, there is a correspondence between quality and content scores that needs to be further studied. Overall, this analysis shows that audio-visual quality and content (at a certain level) ranges encompass the audio and video ranges. As stated in the previous analysis, the smaller range of the perceived quality in Experiment 2 might be related to the inclusion of only audio distortions in the sequences.

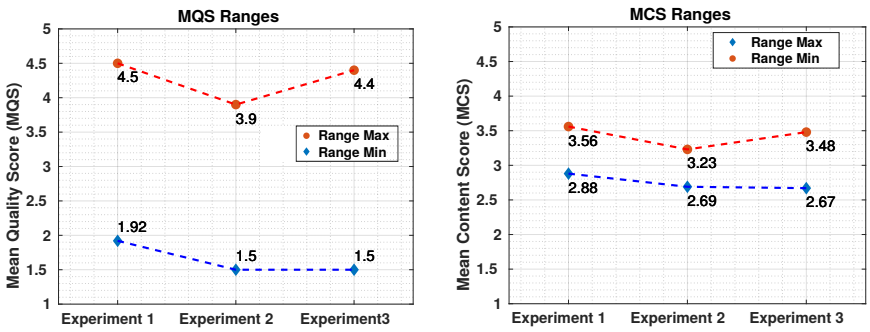


Figure 14. Ranges of quality and content scores in Experiments 1, 2 and 3.

9.4. On the Internal Consistency and External Comparison

In order to evaluate the reliability of the results of this study, the agreement among subjects on their quality perceptions responses is reported. To do so, the internal consistency of the perceptual responses from all three experiments was calculated using the Cronbach's alpha coefficient [66]. In addition, these results are compared with external perceptual responses gathered from the literature. To this end, perceptual data from VQEG-MM [19], UnB-2013 [25], INRS [44], and LIVE-NFLX-II [24] were considered for comparison.

The Cronbach's alpha coefficients for these datasets are reported in Table 14. The interpretation of the results can be made using these criteria: coefficients between 0.00 and 0.69 are associated with a poor internal consistency, a fair internal consistency is associated with coefficients between 0.70 and 0.79, good internal consistency ranges from 0.80 and 0.89, and an excellent internal consistency is associated with coefficients between 0.90 and 1 [67,68].

Table 14. Comparison of the Cronbach's α of all three experiments and other available audio-visual databases.

Database	Experiment	Measurament	SCR	HRC	Stimuli	Distortion	Participants	Cronbach's α
VQEG-MM	AGH_D5	MOS	10	5	60	Audio + Video	15	0.919
VQEG-MM	AGH_Lab	MOS	10	5	60	Audio + Video	14	0.916
VQEG-MM	Intel	MOS	10	5	60	Audio + Video	34	0.920
VQEG-MM	IRCCyN	MOS	10	5	60	Audio + Video	25	0.918
VQEG-MM	IRCCyN_Tablet	MOS	10	5	60	Audio + Video	25	0.910
VQEG-MM	NTIA_cafeteria	MOS	10	5	60	Audio + Video	9	0.894
VQEG-MM	NTIA_Lab	MOS	10	5	60	Audio + Video	28	0.907
VQEG-MM	Opticom	MOS	10	5	60	Audio + Video	15	0.915
VQEG-MM	Technicolor	MOS	10	5	60	Audio + Video	24	0.902
VQEG-MM	Technicolor_patio	MOS	10	5	60	Audio + Video	24	0.918
UnB-2013	Experiment 3	MOS	8	12	72	Audio + Video	16	0.886
INRS		MOS	1	160	160	Video	30	0.869
LIVE-NFLX-II		MOS	15	28	420	Video	65	0.920
UnB-2018	Experiment 1	MQS_{HRC}	60	12	720	Video	60	0.924
UnB-2018	Experiment 2	MQS_{HRC}	40	20	800	Audio	40	0.893
UnB-2018	Experiment 3	MQS_{HRC}	40	20	800	Audio + Video	40	0.896

From the results presented in Table 14, it can be observed that, in general, all studies presented results with a good level of internal consistency (coefficient α above 0.8). In addition, the experiments conducted using the datasets VQEG-MM, LIVE-NFLX-II, and UnB-2018 (Experiment 1) presented an excellent level of internal consistency (coefficient α above 0.90). This indicates that subjects agreed on the quality score when the quality levels, represented by the HRCs, were shifted. If we consider the number of different distortions included in each perceptual experiment, plus the number of test conditions they used (summary of audio-visual studies in Table 2), the quality scores gathered using the UnB-2018 dataset gain more relevance. These results support the reliability of the scores collected in this study. In addition, they validate the application of an immersive approach and encourages the execution of more experiments using this type of methodology.

10. Conclusions

In this paper, we compile and analyze the results of three psychophysical experiments designed to measure the perceived overall audio-visual quality of sequences. In these experiments, impairments were inserted in the audio and/or the visual component of the sequence. An statistical analysis allowed us to understand how audio and video distortions affected the overall audio-visual quality individually and jointly.

A separate analysis of the distortions confirmed that visual degradations significantly impact the overall perceived audio-visual quality. Audio impairments seemed to have a weaker effect on the overall audio-visual quality, although certain degradations (Background noise and Clipping) showed a stronger effect. We can say that these effect is independent of the type of content since these experiments were conducted using a large amount of different content material. It was also observed that participants agree more in their responses when only visual distortions were present (Experiment 1). Whenever audio distortions were introduced (Experiments 2 and 3), participants had more trouble distinguishing the different levels of quality.

A joint analysis allowed us to compare the effects of equivalent test conditions in the presence (or absence) of audio and visual degradations. Audio distortions had a clear effect when responses from sequences with video-only and Audio + Video distortions (with equivalent visual distortions) were compared. Similarly, when comparing the results from sequences with audio-only and audio + video distortions, we noticed that Background noise and Clipping, by themselves, had an effect that was equivalent to the same audio degradation plus the visual distortion. Based on these results, we can assert that visual distortions are not always fully responsible for the overall perceived audio-visual quality. We observed that for certain audio degradations, the overall perceived quality was deter-

mined by the audio distortions. This aspect requires a deeper analysis that might include an analysis per-content to verify the real impact of such degradations. Finally, an internal consistency analysis of the quality scores showed a high level of reliability, despite the large number of distortions and test conditions considered for the study.

Author Contributions: Conceptualization, M.C.Q.F.; methodology, M.C.Q.F.; formal analysis, H.B.M., A.H. and M.C.Q.F.; writing—original draft preparation, H.B.M.; writing—review and editing, H.B.M., A.H. and M.C.Q.F.; supervision, A.H. and M.C.Q.F. All authors have read and agreed to the published version of the manuscript.

Funding: This publication has emanated from research supported in part by the Conselho Nacional de Desenvolvimento Cientfico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF), the University of Brasília (UnB), the research grant from Science Foundation Ireland (SFI) and the European Regional Development Fund under Grant Number 12/RC/2289_P2 and Grant Number 13/RC/2077_P2.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is available for download from the site of the University of Brasília at www.ene.unb.br/mylene/databases.html (accessed on 22 June 2021). Subjective scores are made available on the same website.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chikkerur, S.; Sundaram, V.; Reisslein, M.; Karam, L. Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison. *IEEE Trans. Broadcast.* **2011**, *57*, 165–182. [CrossRef]
2. Bovik, A.C. Automatic prediction of perceptual image and video quality. *Proc. IEEE* **2013**, *101*, 2008–2024.
3. Rix, A.W.; Beerends, J.G.; Kim, D.S.; Kroon, P.; Ghitza, O. Objective assessment of speech and audio quality—Technology and applications. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1890–1901. [CrossRef]
4. Moorthy, A.K.; Bovik, A.C. Visual quality assessment algorithms: what does the future hold? *Multimed. Tools Appl.* **2011**, *51*, 675–696. [CrossRef]
5. Lin, W.; Kuo, C.C.J. Perceptual visual quality metrics: A survey. *J. Vis. Commun. Image Represent.* **2011**, *22*, 297–312. [CrossRef]
6. Akhtar, Z.; Falk, T.H. Audio-visual multimedia quality assessment: A comprehensive survey. *IEEE Access* **2017**, *5*, 21090–21117. [CrossRef]
7. Pinson, M.; Ingram, W.; Webster, A. Audiovisual quality components. *IEEE Signal Process. Mag.* **2011**, *6*, 60–67. [CrossRef]
8. BT, ITUR. *500-14 (10/2019): Methodologies for the Subjective Assessment of the Quality of Television Images*; ITU: Geneva, Switzerland, 2020.
9. Möller, S.; Raake, A. *Quality of Experience: Advanced Concepts, Applications and Methods*; Springer: Berlin/Heidelberg, Germany, 2014.
10. Yang, J.Y.; Park, K.H.; Chang, J.H.; Kim, Y.; Cho, S. Investigation of DNN based Feature Enhancement Jointly Trained with X-Vectors for Noise-Robust Speaker Verification. In Proceedings of the International Conference on Electronics, Information, and Communication (ICEIC), Barcelona, Spain, 19–22 January 2020; pp. 1–5.
11. Aldeneh, Z.; Kumar, A.P.; Theobald, B.J.; Marchi, E.; Kajarekar, S.; Naik, D.; Abdelaziz, A.H. On The Role of Visual Cues in Audiovisual Speech Enhancement. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 8423–8427.
12. ITU-R. *Recommendation P.920 : Interactive Test Methods for Audiovisual Communications*; Technical Report; ITU: Geneva, Switzerland, 2000.
13. ITU-R. *Recommendation BS.1534 : Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems*; Technical Report; ITU: Geneva, Switzerland, 2003.
14. ITU-T. *Recommendation P.1301 : Subjective Quality Evaluation of Audio and Audiovisual Multiparty Telemeetings*; Technical Report; ITU: Geneva, Switzerland, 2013.
15. ITU-T. *P.913 : Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment*; Technical Report; ITU: Geneva, Switzerland, 2014.
16. Pinson, M.; Sullivan, M.; Catellier, A. A New Method for Immersive Audiovisual Subjective Testing. In Proceedings of the 8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Chandler, AZ, USA, 30–31 January 2014.
17. You, J.; Reiter, U.; Hannuksela, M.M.; Gabbouj, M.; Perkis, A. Perceptual-based quality assessment for audio-visual services: A survey. *Signal Process. Image Commun.* **2010**, *25*, 482–501. [CrossRef]

18. Min, X.; Zhai, G.; Zhou, J.; Farias, M.C.; Bovik, A.C. Study of Subjective and Objective Quality Assessment of Audio-Visual Signals. *IEEE Trans. Image Process.* **2020**, *29*, 6054–6068. [CrossRef]
19. Pinson, M.H.; Janowski, L.; Pépion, R.; Huynh-Thu, Q.; Schmidmer, C.; Corriveau, P.; Younkin, A.; Le Callet, P.; Barkowsky, M.; Ingram, W. The influence of subjects and environment on audiovisual subjective tests: An international study. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 640–651. [CrossRef]
20. Hands, D.S. A Basic Multimedia Quality Model. *Multimedia IEEE Trans.* **2004**, *6*, 806–816. [CrossRef]
21. Garcia, M.N.; Schleicher, R.; Raake, A. Impairment-Factor-Based Audiovisual Quality Model for IPTV: Influence of Video Resolution, Degradation Type, and Content Type. *EURASIP J. Image Video Process.* **2011**, *2011*, 1–14. [CrossRef]
22. Borowiak, A.; Reiter, U.; Svensson, U.P. Quality evaluation of long duration audiovisual content. In Proceedings of the Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, USA, 14–17 January 2012; pp. 337–341.
23. Staelens, N.; Coppens, P.; Van Kets, N.; Van Wallendaef, G.; Van den Broeck, W.; De Cock, J.; De Turek, F. On the impact of video stalling and video quality in the case of camera switching during adaptive streaming of sports content. In Proceedings of the 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), Pilos, Greece, 26–29 May 2015; pp. 1–6.
24. Bampis, C.G.; Li, Z.; Katsavounidis, I.; Huang, T.Y.; Ekanadham, C.; Bovik, A.C. Towards Perceptually Optimized End-to-end Adaptive Video Streaming. *arXiv* **2018**, arXiv:1808.03898.
25. Martinez, H.B.; Farias, M.C. Full-reference audio-visual video quality metric. *J. Electron. Imaging* **2014**, *23*, 061108. [CrossRef]
26. Staelens, N.; Vermeulen, B.; Moens, S.; Macq, J.F.; Lambert, P.; Van de Walle, R.; Demeester, P. Assessing the influence of packet loss and frame freezes on the perceptual quality of full length movies. In Proceedings of the 4th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2009), Scottsdale, AZ, USA, 15–16 January 2009.
27. Moorthy, A.K.; Choi, L.K.; Bovik, A.C.; De Veciana, G. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 652–671. [CrossRef]
28. Vučić, D.; Skorin-Kapov, L. QoE Assessment of Mobile Multiparty Audiovisual Telemeetings. *IEEE Access* **2020**, *8*, 107669–107684. [CrossRef]
29. Wendt, D.; Dau, T.; Hjortkjær, J. Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Front. Psychol.* **2016**, *7*, 345. [CrossRef]
30. Harte, N.; Gillen, E.; Hines, A. TCD-VoIP, a research database of degraded speech for assessing quality in voip applications. In Proceedings of the 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), Pilos, Greece, 26–29 May 2015; pp. 1–6.
31. Schwind, A.; Moldovan, C.; Janiak, T.; Dworschak, N.D.; Hoßfeld, T. Do not Stop the Music: Crowdsourced QoE Assessment of Music Streaming with Stalling. In Proceedings of the Twelfth International Conference on Quality of Multimedia Experience (QoMEX), Athlone, Ireland, 26–28 May 2020; pp. 1–6.
32. Rodrigues, R.; Pocta, P.; Melvin, H.; Bernardo, M.V.; Pereira, M.; Pinheiro, A.M. Audiovisual quality of live music streaming over mobile networks using MPEG-DASH. *Multimed. Tools Appl.* **2020**, *79*, 24595–24619. [CrossRef]
33. Demirbilek, E.; Grégoire, J.C. INRS audiovisual quality dataset. In Proceedings of the 2016 ACM on Multimedia Conference, Klagenfurt am Wörthersee, Austria, 10–13 May 2016; pp. 167–171.
34. Falk, T.H.; Chan, W.Y. Performance study of objective speech quality measurement for modern wireless-VoIP communications. *EURASIP J. Audio Speech Music. Process.* **2009**, *2009*, 12. [CrossRef]
35. Yamada, T.; Kumakura, M.; Kitawaki, N. Subjective and objective quality assessment of noise reduced speech signals. In Proceedings of the NSIP 2005 Abstracts IEEE-Eurasip Nonlinear Signal and Image Processing, Sapporo, Japan, 18–20 May 2005; p. 28.
36. Gogate, M.; Dashtipour, K.; Hussain, A. Visual Speech In Real Noisy Environments (VISION): A Novel Benchmark Dataset and Deep Learning-based Baseline System. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 4521–4525.
37. Michelsanti, D.; Tan, Z.H.; Zhang, S.X.; Xu, Y.; Yu, M.; Yu, D.; Jensen, J. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1368–1396. [CrossRef]
38. Gogate, M.; Dashtipour, K.; Adeel, A.; Hussain, A. CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement. *Inf. Fusion* **2020**, *63*, 273–285. [CrossRef]
39. Brunnström, K.; Beker, S.A.; De Moor, K.; Dooms, A.; Egger, S.; Garcia, M.N.; Hossfeld, T.; Jumisko-Pyykkö, S.; Keimel, C.; Larabi, M.C.; et al. *Qualinet White Paper on Definitions of Quality of Experience*; European Network on Quality of Experience in Multimedia Systems and Services: Lausanne, Switzerland, March 2013.
40. Goudarzi, M.; Sun, L.; Ifeachor, E. Audiovisual quality estimation for video calls in wireless applications. In Proceedings of the IEEE Global Telecommunications Conference GLOBECOM 2010, Miami, FL, USA, 6–10 December 2010; pp. 1–5.
41. Keimel, C.; Redl, A.; Diepold, K. The TUM high definition video datasets. In Proceedings of the Fourth International Workshop on Quality of Multimedia Experience, Melbourne, Australia, 5–7 July 2012; pp. 97–102.
42. Li, Z.; Wang, J.C.; Cai, J.; Duan, Z.; Wang, H.M.; Wang, Y. Non-reference audio quality assessment for online live music recordings. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 63–72.
43. Mäki, T.; Kukolj, D.; Đordević, D.; Varela, M. A reduced-reference parametric model for audiovisual quality of IPTV services. In Proceedings of the Fifth International Workshop on Quality of Multimedia Experience (QoMEX), Klagenfurt am Wörthersee, Austria, 3–5 July 2013; pp. 6–11. [CrossRef]

44. Demirbilek, E.; Grégoire, J. Towards reduced reference parametric models for estimating audiovisual quality in multimedia services. In Proceedings of the IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, 22–27 May 2016; pp. 1–6. [CrossRef]
45. Perrin, A.F.N.M.; Xu, H.; Kroupi, E.; Řeřábek, M.; Ebrahimi, T. Multimodal dataset for assessment of quality of experience in immersive multimedia. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1007–1010.
46. Ghadiyaram, D.; Pan, J.; Bovik, A.C. A subjective and objective study of stalling events in mobile streaming videos. *IEEE Trans. Circ. Syst. Video Technol.* **2017**, *29*, 183–197. [CrossRef]
47. Martinez, H.B.; Hines, A.; Farias, M.C. UnB-AV: An audio-visual database for multimedia quality research. *IEEE Access* **2020**, *8*, 56641–56649. [CrossRef]
48. Martinez, H.B.; Farias, M.C. Using The Immersive Methodology to Assess The Quality of Videos Transmitted in UDP and TCP-Based Scenarios. *Electron. Imaging* **2018**, *2018*, 233-1–233-7. [CrossRef]
49. Martinez, H.B.; Farias, M.C. Analyzing the influence of cross-modal IP-based degradations on the perceived audio-visual quality. *Electron. Imaging* **2019**, *2019*, 324-1–324-7. [CrossRef]
50. ITU-T. *H.264 : Advanced Video Coding for Generic Audiovisual Services*; Technical Report; ITU: Geneva, Switzerland, 2003.
51. ITU-T. *H.265 : High Efficiency Video Coding*; Technical Report; ITU: Geneva, Switzerland, 2013.
52. Garcia, M.N.; Dytko, D.; Raake, A. Quality impact due to initial loading, stalling, and video bitrate in progressive download video services. In Proceedings of the 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, 18–20 September 2014; pp. 129–134.
53. Horowitz, M.; Kossentini, F.; Mahdi, N.; Xu, S.; Guermazi, H.; Tmar, H.; Li, B.; Sullivan, G.J.; Xu, J. Informal subjective quality comparison of video compression performance of the HEVC and H. 264/MPEG-4 AVC standards for low-delay applications. In *SPIE Optical Engineering + Applications*; International Society for Optics and Photonics: San Diego, CA, USA, 2012; p. 84990W.
54. Redi, J.; Heynderickx, I.; Macchiavello, B.; Farias, M. On the impact of packet-loss impairments on visual attention mechanisms. In Proceedings of the 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, China, 19–23 May 2013; pp. 1107–1110.
55. Boyce, J.M.; Gaglianello, R.D. Packet loss effects on MPEG video sent over the public Internet. In Proceedings of the Sixth ACM International Conference on Multimedia, Bristol, UK, 13–16 September 1998; pp. 181–190.
56. Wenger, S. H. 264/avc over ip. *Circ. Syst. Video Technol. IEEE Trans.* **2003**, *13*, 645–656. [CrossRef]
57. VQEG. Final Report from the Video Quality Experts Group on the Validation of Objective Models of Multimedia Quality Assessment, Phase I. Technical Report. VQEG: Video Quality Experts Group. Available online: https://www.vqeg.org/ (accessed on 22 June 2021).
58. Ostaszewska, A.; Kloda, R. Quantifying the amount of spatial and temporal information in video test sequences. In *Recent Advances in Mechatronics*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 11–15.
59. ITU-R. *Recommendation BT.500-8: Methodology for Subjective Assessment of the Quality of Television Pictures*; Technical Report; ITU: Geneva, Switzerland, 1998.
60. Ohm, J.R.; Sullivan, G.J.; Schwarz, H.; Tan, T.K.; Wiegand, T. Comparison of the coding efficiency of video coding standards—Including high efficiency video coding (HEVC). *Circ. Syst. Video Technol. IEEE Trans.* **2012**, *22*, 1669–1684. [CrossRef]
61. Hoßfeld, T.; Egger, S.; Schatz, R.; Fiedler, M.; Masuch, K.; Lorentzen, C. Initial delay vs. interruptions: Between the devil and the deep blue sea. In Proceedings of the 2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX), Melbourne, Australia, 5–7 July 2012; pp. 1–6.
62. Oztas, B.; Pourazad, M.T.; Nasiopoulos, P.; Leung, V. A study on the HEVC performance over lossy networks. In Proceedings of the 2012 19th IEEE International Conference on Electronics, Circuits and Systems (ICECS), Seville, Spain, 9–12 December 2012; pp. 785–788.
63. Pinol, P.; Torres, A.; Lopez, O.; Martinez, M.; Malumbres, M.P. Evaluating HEVC video delivery in VANET scenarios. In Proceedings of the Wireless Days (WD), 2013 IFIP, Valencia, Spain, 13–15 November 2013; pp. 1–6.
64. Pinson, M.H.; Barkowsky, M.; Le Callet, P. Selecting scenes for 2D and 3D subjective video quality tests. *EURASIP J. Image Video Process.* **2013**, *2013*, 1–12. [CrossRef]
65. Kortum, P.; Sullivan, M. The effect of content desirability on subjective video quality ratings. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2010**, *52*, 105–118. [CrossRef] [PubMed]
66. Bland, J.M.; Altman, D.G. Statistics notes: Cronbach's alpha. *BMJ* **1997**, *314*, 572. [CrossRef]
67. Cortina, J.M. What is coefficient alpha? An examination of theory and applications. *J. Appl. Psychol.* **1993**, *78*, 98. [CrossRef]
68. Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika* **1951**, *16*, 297–334. [CrossRef]