IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 5, 2020, accepted March 6, 2020, date of publication March 19, 2020, date of current version March 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981861

UnB-AV: An Audio-Visual Database for Multimedia Quality Research

HELARD B. MARTINEZ[1], **ANDREW HINES**[1], **(Senior Member, IEEE), AND MYLÈNE C. Q. FARIAS**[2], **(Senior Member, IEEE)**
[1]School of Computer Science, University College Dublin, Dublin 4, D04 V1W8 Ireland
[2]Department of Electrical Engineering, University of Brasília (UnB), Brasília 70910-900, Brazil

Corresponding author: Helard B. Martinez (hlrdbm03@gmail.com)

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

ABSTRACT In this paper we present the UnB-AV database, which is a database of audio-visual sequences and quality scores aimed at multimedia quality research. The database contains a total of 140 source content, with a diverse semantic content, both in terms of the video and audio components. It also contains 2,320 test sequences with audio and video degradations, along with the corresponding quality and content subjective scores. The subjective scores were collected by performing 3 different psycho-physical experiments using the Immersive Methodology. The three experiments have been presented individually in previous studies. In the first experiment, only the video component of the audio-visual sequences were degraded with compression (H.264 and H.265) and transmission (packet-loss and frame freezing) distortions. In the second experiment, only the audio component of the audio-visual sequences were degraded with common audio distortions (clip, echo, chop, and background noise). Finally, in the third experiment the audio and video degradations were combined to degrade both audio and video components. The UnB-AV database is available for download from the site of the Laboratory of Digital Signal Processing of the University of Brasilia and The Consumer Digital Video Library (CDVL).

INDEX TERMS Audio-visual sequences, quality assessment, multimedia, databases, compression, transmission.

I. INTRODUCTION

The great progress achieved by communications in the last twenty years is reflected by the amount of multimedia services available nowadays. Among these services, internet-based streaming applications are probably the most popular ones. It is understood that the success of these services relies heavily on the quality of the content delivered. Yet, guaranteeing an optimum quality of experience (QoE) can be a challenging task considering the number of distortions that the media is subject to during the delivery process. Therefore, the development of ways to quantify the quality of multimedia content can bring real benefits to internet service providers and broadcast companies. There are currently two ways of estimating quality: subjectively (psychophysical experiments) and objectively (quality metrics).

Subjective methods are known as the most accurate ways of measuring quality. These methods consist of conducting psychophysical experiments, in which a number of human participants rate the perceived quality of a content. Lately, there has been an interest in creating more realistic testing environments. The Immersive Methodology is an example of a methodology that addresses this concern [1], by engaging the participants in the audio-visual experience by attending to specific aspects of the experiment (e.g. experiment stimuli).

Objective quality methods (metrics) are computational algorithms that automatically estimate the quality of a content, as perceived by the end-user. Although several objective quality metrics have been developed, so far, most of the achievements have been in the development of individual image [2], audio [3], and video quality metrics [4]–[6]. In fact,

The associate editor coordinating the review of this manuscript and approving it for publication was Victor Sanchez.

only a few objective metrics have addressed the issue of measuring the quality of audio-visual contents, taking into consideration both the audio and video qualities [7]–[12]. One of the many challenges faced by researchers trying to develop audio-visual quality metrics is the lack of diverse audio-visual quality databases, which represent real multimedia scenarios.

Currently, there are few publicaly available audio-visual databases labelled with subjective quality scores that can be used for testing and developing quality objective metrics. Most of the available databases are limited in terms of content diversity. Considering the current variety of content (Sports, Movies, Series, Videogames, etc.), it is important that these metrics are properly tested with relevant material. In addition, studies that attempt a deeper exploration of the user's level of satisfaction could benefit if the proper material is made available for research [13]–[15].

In summary, the area of multimedia quality depends heavily on the availability of quality databases, which are datasets that contain: (1) source contents (SRCs) in pristine condition, (2) processed versions sequences (PVSs), which are generated by treating the SRCs with various Hypothetical Reference Circuits (HRCs), and (3) the subjective scores of each PVS, collected by performing a psychophysical experiment [16]. Most commonly, subjective scores available in quality databases contain subjective data about some aspect of the content, like the overall quality, the perceived degradations, the level of comfort experienced by the user, etc. Some quality databases also include other types of data, like eye tracking information, qualitative responses, biological signs, etc. An important use of quality databases is to help design objective quality metrics. But, frequently these databases are used in (visual and audio) perception research. For example, they are used to study the impact of system parameters on quality or to analyze the perceptual characteristics of common artifacts [17]–[20].

To address these shortcomings, in this paper, we present a public audio-visual quality database, which contains diverse SRC content and PVSs with audio and video degradations. Besides the common video compression and transmission degradations, the database contains a set of audio degradations that include background noise, chop effect, amplitude clipping, and echo effect. Quality and content scores were collected in 3 psychophysical experiments using the Immersive methodology [21]–[23]. The experiments were conducted using a large number of PVSs while still providing reliable scores.

The paper is organised as follows. Section II discusses the currently available audio-visual quality databases. Section III lists the experimental methodology used in the 3 experiments. Sections IV, V, and VI describe the source content and the audio and video degradations in the database. Sections VII, VIII, and IX describe the HRCs used in each experiment. Section X discusses the reliability of the gathered scores.

II. AVAILABLE AUDIO-VISUAL QUALITY DATABASES

Audio-visual databases, along with their accompanying subjective scores, represent an essential ground truth for quality assessment research. Over time, a number of audio-visual databases have been published and made available for researchers. Table 1 presents a list of the most important audio-visual quality databases currently available. For each database, this table includes its content characteristics, types of distortions, and characteristics of the psychophysical experiments used to collect the subjective data. The first database in this list is the PLYM [24]. It contains data from 3 experiments, providing audio, video, and audio-visual quality scores. Since this is the oldest database of the list, it has some issues that limit its use in modern multimedia scenarios. First, the content has reduced spatial (176 × 144) and temporal (8 and 15 frames per second - fps) resolutions. Second, the 6 SRCs have a low diversity of semantic content, with all scenes having a speaker facing the camera. Finally, the HRCs (H.263 and G.711 codecs) used to generate the PVSc are now outdated.

The second database is the TUM database [25]. This database contains eight 1080p SRCs and only *video* compression degradations, obtained encoding the SRCs with H.264/AVC and Dirac codecs at several bitrates. One of the interesting aspects of this database is the use of different display technologies (reference/consumer LCDs and home-cinema projectors). The third database is the VQEG audio-visual database [26], which is certainly unique because it contains data from 10 different experiments performed in 6 different laboratories. Experimental settings included different types of displays and environments. The database has ten 480p SRCs and five HRCs consisting of several bitrates of H.264 and Advanced Audio Coding (AAC) enconding.

The fourth database is the VTT database [27], which contains 12 audio-visual SRCs. The HRCs encompass several conditions, including three spatial resolutions (480p, 720p, and 1080p), video (H.264) and audio (AAC) compressions, and transmission degradations. The audio and video compression bitrates varied according to the resolution. Only the video component was affected by the transmission degradations, which consisted of packet loss rates (PLR, given by the percentage of lost packages) and bursts of 3 different sizes. The fifth database is the UnB-AV 2013 database[1] [10], which contains six 720p SRCs. This dataset is comprised of HRCs corresponding to video (H.264) and audio (MPEG-1 Layer 3) compressions at different bitrates. Similarly to the PLYM database, both VTT and UnB-AV 2013 include data from three experiments, providing audio, video, and audio-visual quality scores for each PVS.

The sixth database is the INRS [28], which contains a single 720p SRC of 45 seconds. The database con-

[1] This database is available for download from the site of the University of Brasília (www.ene.unb.br/mylene/databases.html) and at The Consumer Digital Video Library (www.cdvl.org - create an account and search for UNB).

TABLE 1. Publicly available audio-visual quality datasets.

	PLYM [24]	TUM [25]	VQEG [26]	VTT [27]	UNB-AV-2013 [28]	INRS [29]	LIVE-NFLX-II [30]	UNB-AV-2018
Content Characteristics								
Originals (SRC)	6	8	10	12	6	1	15	60, 40, 40
Duration	7s - 14s	10s	10s	10s	8s	42s	25s	19s-68s
Resolution	144p	1080p	480p	480p, 720p, 1080p	720p	720p	1080p	720p
Amount of Motion	low	low-high	low-high	low-high	low-high	low	low-high	low-high
Frame Rate (fps)	8, 15	50, 25	30	25-30	30	10, 15, 20, 25	30	30
HRCs	10	4	5		4, 3, 12	160	28	12, 20, 20
Test Seqs	180	20	60	125	30, 24, 78	160	420	720, 800, 800
Video Distortions								
Compression	H.263	H.264, Dirac: 2-40 Mbps	H.264: 100, 192, 250, 448, 500, 1000 kbps	H.264: 1Mbps (480), 3Mbps (720p), 1Mbps (1080)	H.264: 0.8, 2, 1, 30 Mbps	H.264, QP: 27, 31, 35	H.264: 150Kbps to 6 Mbps	H.265: 0.2, 0.4, 1, 8Mbps; H.264: 0.5, 0.8, 2, 16Mbps
Audio Distortions								
Compression	G.711, 16-bit PCM, mono, 8kHz	PCM, 7.1 surround audio	AAC: 8, 32, 64 kbps	AAC: 96kbps (1080, 720), 64kbps (480)	MPEG-1 Layer 3: 128, 96, 48 kbps	AMR-WB, Mono, 16KHz, 24kbps	MPEG-1 Layer 3	AAC, 16-bits, 48kHz
Chop	-	-	-	-	-	-	-	0.04s; 1, 2, 5 chops/s; Period of 0.02s and ,
Clipping	-	-	-	-	-	-	-	Multipliers: 11, 15, 25, 55
Echo	-	-	-	-	-	-	-	$\alpha = 0.175, 0.3, 0.5\%$ Delay: 25, 100, 140, 180ms; Feedback: 0, 0.8 %
Noise	-	-	-	-	-	-	-	SNR: 15, 10, 5dB Type: car, babble, office, road
Transmission Distortions								
PLR (%)	0.01, 0.05, 0.1, 0.15, 0.20	-	-	(only video) 0.3, 0.6, 1.2, 2.4, 4.8; Burst sizes = 1, 2, 3	- -	0, 0.1, 0.5, 1, 5	-	(only video) 1, 3, 5, 8, and 10
Freezing Frame	-	-	-	-	-	-	-	1, 2, or 3 events of 1, 2, or 3s, at the beginning, middle or end of the video.
ABR transmission	-	-	-	-	-	-	Buffer-based, Rate-Based, Quality-Based, Oracle Quality-Based	-
Experiment Characteristics								
Methodology	ACR	DSUR[28], SSMM [29]	ACR	Modified DCR	ACR	ACR	ACR	ACR, Immersive
Experiments	3	2	10	1	3	1	2	3
Subjects per Exp.	16, 16, 16	21	9 - 35	24	16, 16, 17	30	65	60, 40, 40
Audio MOS	●			●	●		No	
Video MOS	●			●	●		●	
Audio-Visual MOS	●	●	●	●	●	●		●
Rating scale	1-9	0-10	1-5	1-5	0-100	1-5	1-100	1-5

tains 160 HRCs, consisting of video (H.264) compression bitrates and network settings (frame rate, packet loss rate, quantization, and noise reduction parameters). The seventh database is the LIVE-NFLX-II [29], which is not exactly an audio-visual quality database, given that the experimental data provided are video quality scores and not audio-visual quality scores. Nevertheless, we added this database to the list because, to our knowledge, this is the only database that has dynamic adaptive streaming over HTTP (DASH) HRCs. This database also contains a very diverse set of SRCs, consisting of 15 1080p sequences. The HRCs include several content-adapted video compression bitrates and 3 adaptive bitrate (ABR) streaming approaches. This database has no audio degradations.

In summary, although there are a number of audio-visual databases in the literature, several of them have limitations, either in terms of content diversity, spatial and temporal resolutions, or types of HRCs. In this paper, we present a public audio-visual quality database (UnB-AV 2018). A summary of its specifications is presented in the last column of Table 1. The database contains a large number of SRCs (140) and HRCs (52). Besides the typical compression and transmission degradations, our database has the differential of containing several types of audio degradations. To obtain the audio-visual quality scores, we performed 3 psychophysical experiments using the Immmersive Methodology. This allowed us to have a large number of PVSs and still obtain scores with good reliability. In the following sections we describe in details the UnB-AV 2018 database, which is currently available for download in the site of the Laboratory of Digital Signal Processing (GPDS) of the University of Brasília (UnB)[2] and at The Consumer Digital Video Library (CDVL).[3]

III. APPARATUS AND EXPERIMENTAL METHODOLOGY

We performed 3 psychophysical experiments using the Immersive Methodology [21]–[23]. The first experiment contains degradations only in the video component, the second experiment contains degradations only in the audio component, and the third experiment contains degradations in both components. All three experiments were conducted in a sound-isolated recording studio at the Department of Electrical Engineering of the University of Brasília (UnB). The room had the lights dimmed to avoid light to be reflected on the monitor. Each experimental session was performed with only one participant, seated straight ahead of the monitor, centered at or slightly below eye height. The distance from the subject's eyes to the monitor was 3 screen heights.

Table 2 gives the specifications of the hardware equipment, which consisted of a desktop computer, an LCD monitor, a set of earphones, and a dedicated sound card. The dynamic

[2] `www.ene.unb.br/mylene/databases.html`
[3] `www.cdvl.org` - Create an account and search for UNB.

TABLE 2. Equipment specifications.

Equipment	Technical Details
Monitor	Samsung SyncMaster P2370 Resolution: 1,920x1,080; Pixel-response rate: 2ms; Contrast ratio: 1,000:1; Brightness: 250cd/m2

Earphones	Sennheiser Hd 518 Headfone
	Impedance: 50 Ohm; Sound Mode: Stereo;
	Frequency response: 14–26,000Hz;

Sound Card	Asus Xonar DGX 5.1

TABLE 3. Details about participants in Experiments 1, 2, and 3.

Experiment	Participants	Female	Male	Age Range
Experiment 1	60	18	42	19-36
Experiment 2	40	15	25	21-36
Experiment 3	42	16	26	20-34

contrast of the monitor was turned off, the contrast was set at 100, and the brightness at 50. The experiments were run using a quality assessment browser-based software application developed in GPDS, which was also used to record the subject's data. The experimental interface used a client-server model based on the HTML standard (version 5), using PHP, javascript, and a Postgresql database. The client-server model consists of a web server and a postgresql database running on the same station where the content is reproduced. The experimental sessions were controlled by the browser, using an HTML5 interface to communicate with the server.

All three experiments were performed with volunteers, most of them were graduate students at the University of Brasília. They were considered naive of most kinds of digital video defects. No vision or hearing tests were performed, but unimpaired hearing was a pre-requirement. Moreover, participants were asked to wear glasses or contact lenses if they needed them to watch TV. Table 3 presents an overview of the participants' gender and ages for the three experiments.

The experiment was divided into instruction, training, and main sessions. During the instruction session, participants were given instructions and presented with a set of original content and their corresponding degraded versions. The goal was to familiarize the participants with the quality range. In the training session, subjects performed the same tasks performed in the main session. After each test stimuli displayed, they were asked to rate the quality and the content of the video, using two five-points (1-5) Absolute Category Rating (ACR) scales, in accordance to the Immersive Methodology. The points in the quality scale were labeled (in Portuguese) as "Excellent", "Good", "Fair", "Poor", and "Bad", while the points in the content scale were labeled as "Intriguing", "Interesting", "Neutral", "Uninteresting", and "Boring".

In the main session, the actual experimental task was performed. Participants were presented with a subset of the entire stimuli pool, as detailed in the Immersive Methodology description [1]. None of the participants watched videos with the same content, i.e. all PVSs rated by each subject originated from different SRCs. For each session, subjects rated five stimuli for each HRC. Approximately five subjects rated each single stimuli, from the entire pool of test videos. The experimental session was limited to 50 minutes, with a break introduced in the middle of the session to avoid fatigue.

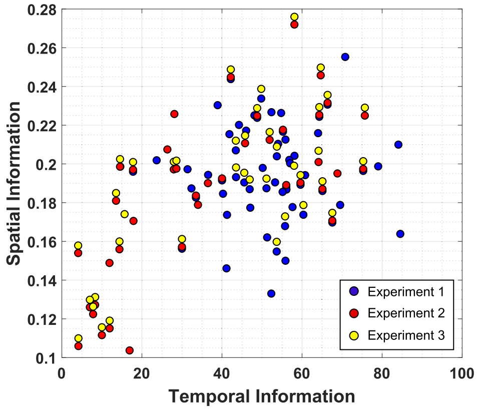

FIGURE 1. Source videos spatial and temporal information measures.

IV. SOURCE STIMULI

To build our database, we used 140 high-definition video sequences (with accompanying audio) as SRCs. Some sequences were generated by parsing larger videos. These SRCs were distributed among all experiments in the following manner: 60 video sequences for experiment 1, 40 for experiment 2, and 40 for experiment 3. The videos have a spatial resolution of 1280 × 720, a temporal resolution of 30 fps, and a 4:2:0 color space. For the audio component, the bit-depth and sample frequency were set to 16 bits and 48 kHz, respectively. The stimuli were 19 to 68 seconds long, with an average of 36 seconds. Figure 1 presents the spatial and temporal information measurements [30], computed for all videos in experiments 1, 2, and 3. Figure 2 shows representative frames of the SRCs.

As for the audio component, stimuli containing a variety of music, speech, smooth and rough sounds were included in the database. To describe the audio content, we used the algorithm proposed by Giannakopoulos *et al.* [31]. This algorithm divides the audio streams into several non-overlapping segments and classifies each segment into one of the following classes: music, speech, others1 (low environmental sounds: wind, rain, etc.), others2 (sounds with abrupt changes, like a door closing), others3 (louder sounds, mainly machines, and cars), gunshots, fights, and screams. Figure 3 depicts the audio classification for all three experiments, which shows a good distribution of the audio content for all three experiments.

V. VIDEO DEGRADATIONS

The database contains 3 types of video distortions: coding, packet loss, and frame freezing. The video compression HRCs corresponded to Low, Medium, High, and Very High bitrate settings, obtained with the H.264/MPEG-4 Advance

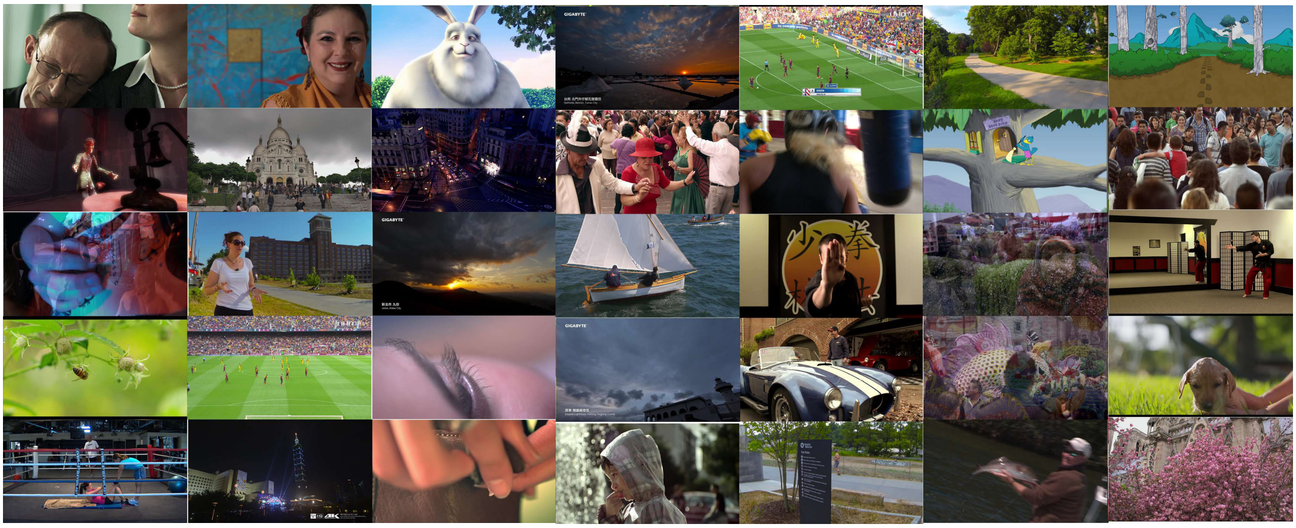

FIGURE 2. Sample frames of original videos (SRCs) of the database. Video genres include: TV Commercials, Sports, Music Videos, Cartoons, Interviews, Documentaries, Movie Trailers, Landscape videos and Computer Graphics.

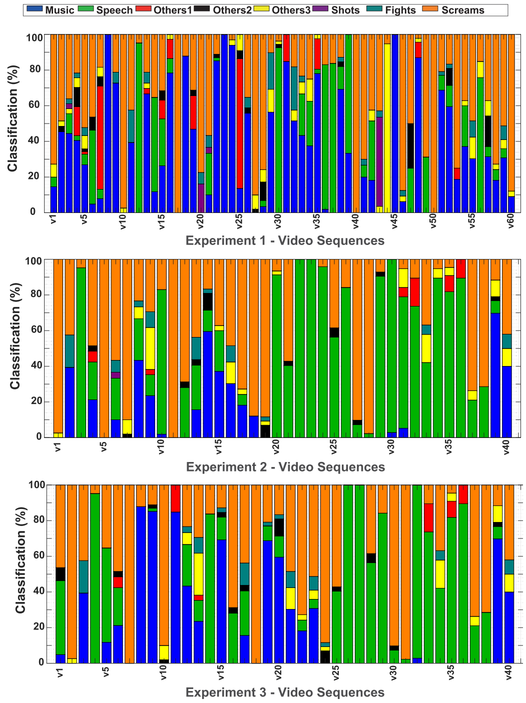

FIGURE 3. Audio classification of video sequences. Eight audio classes: music, speech, others1 (low environmental sounds: wind, rain, etc.), others2 (sounds with abrupt changes, like a door closing), others3 (louder sounds, mainly machines, and cars), gunshots, fights, and screams.

TABLE 4. Bitrate values for each codec.

	Low	Medium	High	Very High
H.264/AVC	0.5 Mbps	0.8 Mbps	2 Mbps	16 Mbps
H.265/HEVC	0.2 Mbps	0.4 Mbps	1 Mbps	8 Mbps

Video Coding (AVC) and the H.265 High Efficiency Video Coding (HEVC) [32], [33] codecs, as shown in Table 4. To select these bitrate values, a number of bitrate values were selected from the literature taking into account previous works [34], [35]. The values ranged from 0.2 – 16 Mbps (H.264/AVC) and 0.1 – 8 Mbps (H.265/HEVC). Two source stimuli (not considered for the main experiment) were processed using these bitrate values. Then, an empirical criteria was used to select four very clear quality levels (for H.264/AVC and H.265/HEVC).

To select these bitrate values, we visually examined videos compressed with several bitrate levels and chose four clear quality levels, taking into account previous works [34], [35].

To generate packet loss degradations, all SRCs were first encoded using AVC (H.264) and HEVC (H.265) codecs. Then, we inserted the losses by dropping Network Abstraction Layer (NAL) packets from the video bit-stream [36]. To avoid the generation of unrealistic strong artifacts, the codec's standard error concealment was used, which replaces a lost packet by the co-located packet in the previous frame. To replicate a real video streaming scenario, five packet loss ratios (PLR) were considered [37], [38]: 1%, 3%, 5%, 8%, and 10%.

Services like Video on Demand (VoD) are based on the Transmission Control Protocol (TCP). As a consequence, these services do not experience packet loss distortions. But, when the available throughput is lower than the content bitrate, the reproduction will stall until enough data has been downloaded. This effect is perceived by the end-users as freezing without skipping, commonly known as rebuffering or stalling. The freezing effect is also experienced before the media starts its reproduction, this is known as the 'initial loading'. We considered 3 parameters to create frame freezing: number of freezing events (N), the position of the freezing events in the sequence (P), and length of the freezing event (L).

TABLE 5. Frame freezing settings (N, L, P) for different quality levels.

Level	N	P=1	P=2	P=3	L=1	L=2	L=3
S0	0						
S1	1	2			2		
S2	2	1	3		1	3	
S3	2	2	3		2	2	
S4	3	1	2	3	2	2	3
S5	3	1	2	3	3	3	2

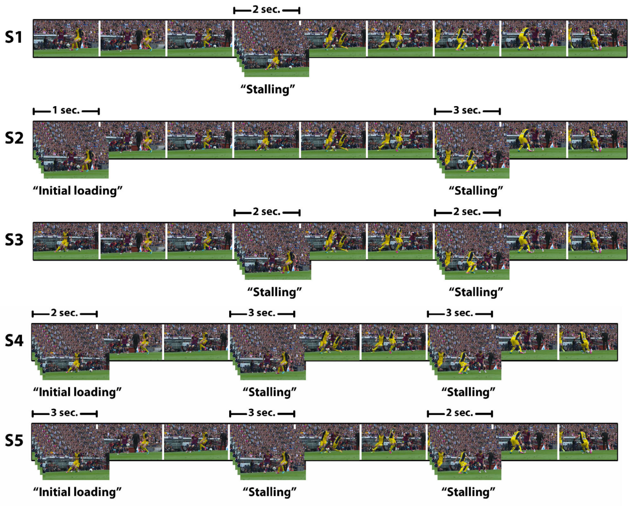

FIGURE 4. Freezing levels of distortion for scenarios S1 – S5.

As for the position of the events (P), 3 possible options were chosen: 1 (beginning), 2 (middle), and 3 (end). These positions were obtained by dividing by three the total length of the sequence and multiplying the result by 0, 1 and 2. A freezing event located at position 1 represents the initial loading experienced before the video starts playing. Finally, the length (L) of the freezing events were fixed at 1, 2, or 4 seconds. Initial loading and stalling were inserted using Avisynth (`www.avisynth.org`). Regarding the audio component, silence was inserted using a faded in and out effect to avoid artifacts at the silence boundaries. All 3 parameters (N, P, and L) were combined to represent the level of discomfort perceived by the user, as depicted in Table 5. These combinations were named as S1, S2, S3, S4, and S5, going from the least annoying combination (S1) to the most annoying combination (S5). Figure 4 illustrates all five levels of freezing distortions.

VI. AUDIO DEGRADATIONS

The TCD-VoIP dataset [39] served as a reference to produce the set of audio distortions used in the UnB-AV database. Four types of audio degradations were selected: background noise, clipping, echo, and chop. Background noise describes any sound that is not the sound under study. Four types of Background Noise (e.g. babble, car, road, and office) were added to the original signal at different SNR levels. Thus, as shown in Table 6, two varying parameters were considered for this type of degradation: the type of background noise and the SNR level associated with the noise.

TABLE 6. Parameters of audio degradations.

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

A clipping distortion appears when a transmitted signal exceeds the maximum amplitude level permitted. This is handled by cutting the signal (clipping) to maintain the maximum amplitude level. As a result, some samples become 'clipped' and the signal quality gets compromised. The clipping effect was generated by amplifying the signal using 4 multiplying factors, as shown in Table 6.

An echo effect normally occurs when a microphone picks up audio signals and sends them back to its origin, creating a feedback loop. We created the echo effect by adding to the original signal its delayed versions. Table 6 shows the three parameters varied to generate different levels of distortion: Alpha, which is the amplitude percentage of the first delayed version; Delay, which is the time length between the first delayed version and the original; and Feedback, which is the percentage reduction of the subsequent delayed versions.

A chop degradation happens when a signal is transmitted with missing samples. When an audio signal is played, missing samples can be discarded, substituted by either silence or previous (repeated) samples, or skipped. Table 6 shows the three parameters varied to produce different levels and types of choppy speech: Period, which sets the length of the discarded samples; Rate, which indicates the frequency of the sample discard; and Mode, which states how the discarded samples are handled.

VII. EXPERIMENT 1

In the first experiment, impairments were only inserted into the video component, while the quality of the audio component remained constant. As previously described, three types of distortions were considered: video coding, packet loss, and frame freezing. The SRCs were compressed at 4 different bitrate levels (low, medium, high, and very high) using 2 codecs (H.264 and H.265). Since frame freezing and packet loss related video distortions do not occur simultaneously in a real transmission scenario [34], two groups of HRCs were set. The first group combines artifacts produced by compression with packet loss distortions (HRC1 to HRC5). More specifically, 5 combinations of bitrates and codecs were selected, representing five levels of quality. For each of these combinations, a packet-loss ratio was assigned (1%, 3%,

TABLE 7. 1st group of HRCs and ANCs of the Experiment 1.

HRC	Codec	Bitrate (kb/s)	PLR
HRC1	H.264	500	10%
HRC2	H.265	400	8%
HRC3	H.264	2,000	5%
HRC4	H.265	1,000	3%
HRC5	H.265	8,000	1%
ANC1	H.264	64,000	0

TABLE 8. 2nd group of HRCs and ANCs of the Experiment 1.

HRC	Codec	Bitrate (kb/s)	Freezing
HRC6	H.265	200	S5
HRC7	H.264	800	S4
HRC8	H.265	1,000	S3
HRC9	H.264	2,000	S2
HRC10	H.264	16,000	S1
ANC2	H.265	32,000	S0

5%, 8%, and 10%), as depicted in Table 7. These 5 HRCs are replicated for all 60 SRCs, resulting in 300 PVSs.

The second group of HRCs combines artifacts produced by compression with frame freezing effects (HRC6 to HRC10). Another 5 combinations of bitrate levels and codecs were used, but no combination used for the first group was used in the second group. Each of these 5 encoding combinations was paired with one of the five levels of the frame freezing discomfort scale (S1, S2, S3, S4, and S5), as depicted in Table 8. These 5 HRCs are replicated for all 60 SRCs, resulting in 300 PVSs. Two video sequences compressed at extremely high bitrate levels, with no packet loss video distortions or frame freezing effects, worked as anchors (ANC1 and ANC2) to help participants recognize the entire range of quality used for the experiment. Pooling all test stimuli, 720 PVSs were generated for this experiment. For each experimental session, the participant was presented with only 60 test stimuli (out of 720), as recommended by the Immersive Methodology. More details about this experiment and an analysis of the subjective scores can be found at [21].

VIII. EXPERIMENT 2

In the second experiment, impairments were only inserted into the audio component, while the quality of the video component remained constant. As previously described, 4 types of degradations were added to the audio component of 40 SRCs: background noise, clipping, echo, and chop. Table 9 depicts the 16 HRCs used in this experiment. We added 4 types of background noise (babble, car, road, and office) to the original signal at different SNR levels (HRC1 to HRC4). We also selected 4 combinations of the choppy degradations (HRC5 to HRC8) and 4 clipping degradations (HRC9 to HRC12). Finally, 4 combinations of echo were selected (HRC13 to HRC16). Additionally, 4 test conditions without degradations were used as anchors (ANC1 to ANC4) to help participants establish the quality range. In total, 800 PVSs with different audio distortions were generated. Again, in each experimental session, the participant was presented with only 40 PVSs (out of 800), as recommended by the Immersive Methodology. Details about the conduction of the experiment and the analysis of the subjective experiments are presented in [22].

TABLE 9. HRCs and Anchor test conditions (ANC) of Experiment 2.

BG Noise	**Noise**	**SNR (dB)**	
HRC1	car	15	
HRC2	babble	10	
HRC3	office	10	
HRC4	road	5	
ANC1	-	-	
Chop	**Period (s)**	**Rate (chops/s)**	**Mode**
HRC5	0.02	1	previous
HRC6	0.02	2	zeros
HRC7	0.04	2	previous
HRC8	0.02	5	zeros
ANC2	-	-	-
Clipping	**Multiplier**		
HRC9	11		
HRC10	15		
HRC11	25		
HRC12	55		
ANC3	-		
Echo	**Alpha (%)**	**Delay (ms)**	**Feedback (%)**
HRC13	0.5	25	0
HRC14	0.3	100	0
HRC15	0.175	140	0.8
HRC16	0.3	180	0.8
ANC4	-	-	-

IX. EXPERIMENT 3

In the third experiment, we introduced audio and video distortions in the audio and video components, respectively, of the original sequences. The HRCs in this experiment were a combination of the HRCs of Experiments 1 and 2. More specifically, the test conditions were organized to produce a set of 16 HRCs and 4 anchors (ANC1 to ANC4), which are depicted in Table 10. Altogether, 40 SRCs were processed at 20 different test conditions (including 4 anchor conditions). This resulted in 800 PVSs with different audio and video distortions. For each experimental session, the participant was presented with only 40 test stimuli, out of the 800 test sequences, as recommended by the Immersive Methodology. An extended description of the experiment plus a discussion of the subjective scores gathered can be found in [23].

X. RELIABILITY OF THE MEAN OBSERVER SCORES

The Mean Quality Score with respect to the j-th HRC, $MQS_{HRC(j)}$, is given by the average of the quality scores, over all subjects, for that particular HRC. In the same way, the Mean Content Score with respect to the j-th HRC, $MCS_{HRC(j)}$, is given by the average of the content scores, over all subjects, for that particular HRC. To measure the internal consistency (reliability) of the quality and content scores, we computed the Cronbach's α coefficient [40]. The α coefficient ranges from 0 to 1, with a greater value being interpreted as a greater internal consistency. Table 11 presents the Cronbach's α coefficients for all scores of the three experiments. For Experiment 1, the coefficient value for

TABLE 10. HRCs and Anchor test conditions (ANC) of experiment 3.

	Audio Component				Video Component			
	Noise	**Chop**	**Clip**	**Echo**	**Video Codec**	**Bitrate**	**PacketLoss**	**Freezing**
HRC	Type, SNR	Period, Rate, Mode	Multiplier	Alpha, Delay, Feedback		(kbps)	PLR	Pauses, Length
HRC1	car, 15 dB	-	-	-	H.264	16,000	-	1, 2s
HRC2	-	-	11	-	H.264	16,000	-	1, 2s
HRC3	-	-	11	-	H.265	8,000	0.01%	-
HRC4	-	0.02s, 2 chop/s, zeros	-	-	H.265	80,00	0.01%	-
HRC5	-	-	-	0.3%, 100ms, 0%	H.264	16,000	-	1, 2s
HRC6	office, 10 dB	-	-	-	H.264	16,000	-	1, 2s
HRC7	-	-	-	0.3%, 100ms, 0%	H.265	8,000	0.01%	-
HRC8	-	-	-	0.3%, 100ms, 0%	H.264	2,000	0.05%	-
HRC9	office, 10 dB	-	-	-	H.264	2,000	0.05%	-
HRC10	office, 10 dB	-	-	-	H.264	800	-	3, 7s
HRC11	-	-	25	-	H.264	2,000	0.05	-
HRC12	-	-	25	-	H.264	800	-	3, 7s
HRC13	-	-	25	-	H.265	400	0.08	-
HRC14	-	0.02s, 5 chop/s, zeros	-	-	H.265	400	0.08%	-
HRC15	-	-	-	0.3%, 180ms, 0.8%	H.264	800	-	3, 7s
HRC16	-	-	-	0.3%, 182ms, 0.8%	H.265	400	0.08%	-
ANC1	-	-	-	-	H.264	64,000	-	-
ANC2	-	-	-	-	H.265	32,000	-	-
ANC3	-	-	-	-	H.264	64,000	-	-
ANC4	-	-	-	-	H.265	32,000	-	-

TABLE 11. Cronbach's α of all experiment scores.

Score	Analysis	Cronbach's α	Experiment
MQS_{HRC}	per-HRC	0.924	Experiment 1
MCS_{HRC}	per-HRC	0.858	Experiment 1
MQS_{HRC}	per-HRC	0.893	Experiment 2
MCS_{HRC}	per-HRC	0.841	Experiment 2
MQS_{HRC}	per-HRC	0.896	Experiment 3
MCS_{HRC}	per-HRC	0.864	Experiment 3

MQS_{HRC} was 0.924, while for MCS_{HRC} it was 0.858. For Experiment 2, the coefficient for MQS_{HRC} was 0.893, meanwhile for MCS_{HRC} it was 0.841. Finally, for Experiment the coefficients for MQS_{HRC} and MCS_{HRC} were 0.896 and 0.864, respectively. In all experiments, the level of consistency is good and it can be concluded that the scores gathered are highly reliable. Prior to this analysis, participants considered as outliers were removed. Two participants were removed from Experiment 1 and one from Experiment 2, meanwhile, no participants were removed from Experiment 3.

XI. CONCLUSION

In this paper, we presented the UnB-AV, which is an audio-visual quality database that can be used in multimedia quality research. The database contains a total of 140 SRCs and 52 HRCs, resulting in a total of 2,320 PVSs. The content is diverse, both in terms of the video and audio components. Besides the typical video and audio compression and transmission degradations, the database has the differential of containing several types of audio degradations. For each PVS, the database has the corresponding (audio-visual) quality and content scores. These scores were collected by running three psychophysical experiments using the Immersive Methodology. We measured the consistency of the collected quality and content scores, obtaining good results.

REFERENCES

[1] M. Pinson, M. Sullivan, and A. Catellier, "A new method for immersive audiovisual subjective testing," in *Proc. 8th Int. Workshop Video Process. Qual. Metrics Consum. Electron. (VPQM)*, 2014, pp. 1–6.

[2] D. M. Chandler, "Seven challenges in image quality assessment: Past, present, and future research," *ISRN Signal Process.*, vol. 2013, pp. 1–53, Feb. 2013.

[3] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOL: The virtual speech quality objective listener," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[4] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[5] U. Engelke and H.-J. Zepernick, "Perceptual-based quality metrics for image and video services: A survey," in *Proc. Next Gener. Internet Netw.*, May 2007, pp. 190–197.

[6] M. A. Usman, M. R. Usman, and S. Y. Shin, "A novel no-reference metric for estimating the impact of frame freezing artifacts on perceptual quality of streamed videos," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2344–2359, Sep. 2018.

[7] M. N. Garcia, R. Schleicher, and A. Raake, "Impairment-factor-based audiovisual quality model for IPTV: Influence of video resolution, degradation type, and content type," *EURASIP J. Image Video Process.*, vol. 2011, pp. 1–14, Dec. 2011.

[8] K. Yamagishi and S. Gao, "Light-weight audiovisual quality assessment of mobile video: ITU-T Rec. P.1201.1," in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2013, p. 1201.

[9] H. B. Martinez and M. C. Q. Farias, "A no-reference audio-visual video quality metric," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2014, pp. 2125–2129.

[10] H. B. Martinez and M. C. Farias, "Full-reference audio-visual video quality metric," *J. Electron. Imag.*, vol. 23, no. 6, 2014, Art. no. 061108.

[11] H. A. B. Martinez and M. C. Q. Farias, "Combining audio and video metrics to assess audio-visual quality," *Multimedia Tools Appl.*, vol. 77, no. 18, pp. 23993–24012, Sep. 2018.

[12] H. Martinez, M. C. Q. Farias, and A. Hines, "NAViDAd: A no-reference audio-visual quality metric based on a deep autoencoder," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.

[13] J. Song, F. Yang, Y. Zhou, S. Wan, and H. R. Wu, "QoE evaluation of multimedia services based on audiovisual quality and user interest," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 444–457, Mar. 2016.

[14] M. J. Scott, S. C. Guntuku, W. Lin, and G. Ghinea, "Do personality and culture influence perceived video quality and enjoyment?" *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1796–1807, Sep. 2016.

[15] J. Guan, S. Yi, X. Zeng, W.-K. Cham, and X. Wang, "Visual importance and distortion guided deep image quality assessment framework," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2505–2520, Nov. 2017.

[16] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[17] G. Cermak, M. Pinson, and S. Wolf, "The relationship among video quality, screen resolution, and bit rate," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 258–262, Jun. 2011.

[18] M. C. Q. Farias, J. M. Foley, and S. K. Mitra, "Detectability and annoyance of synthetic blocky, blurry, noisy, and ringing artifacts," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2954–2964, Jun. 2007.

[19] M. C. Q. Farias and S. K. Mitra, "Perceptual contributions of blocky, blurry, noisy, and ringing synthetic artifacts to overall annoyance," *J. Electron. Imag.*, vol. 21, no. 4, Nov. 2012, Art. no. 043013.

[20] K. Yamagishi and T. Hayashi, "Parametric quality-estimation model for adaptive-bitrate-streaming services," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1545–1557, Jul. 2017.

[21] H. B. Martinez and M. C. Q. Farias, "Using the immersive methodology to assess the quality of videos transmitted in UDP and TCP-based scenarios," *Electron. Imag.*, vol. 2018, no. 12, pp. 233-1–233-7, Jan. 2018.

[22] H. Martinez, M. C. Q. Farias, and A. Hines, "Perceived quality of audio-visual stimuli containing streaming audio degradations," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2529–2533.

[23] H. B. Martinez and M. C. Farias, "Analyzing the influence of cross-modal IP-based degradations on the perceived audio-visual quality," in *Proc. 16th Electron. Imag., Image Qual. Syst. Perform.*, Jan. 2019, pp. 324-1–324-7.

[24] M. Goudarzi, L. Sun, and E. Ifeachor, "Audiovisual quality estimation for video calls in wireless applications," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2010, pp. 1–5.

[25] C. Keimel, A. Redl, and K. Diepold, "The TUM high definition video datasets," in *Proc. 4th Int. Workshop Qual. Multimedia Exper.*, Jul. 2012, pp. 97–102.

[26] M. H. Pinson, L. Janowski, R. Pépion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 640–651, Oct. 2012.

[27] T. Mäki, D. Kukolj, D. Dordevic, and M. Varela, "A reduced-reference parametric model for audiovisual quality of IPTV services," in *Proc. 5th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 6–11.

[28] E. Demirbilek and J.-C. Grégoire, "Towards reduced reference parametric models for estimating audiovisual quality in multimedia services," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[29] C. G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, and A. C. Bovik, "Towards perceptually optimized end-to-end adaptive video streaming," 2018, *arXiv:1808.03898*. [Online]. Available: http://arxiv.org/abs/1808.03898

[30] C. G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, and A. C. Bovik, "Towards perceptually optimized end-to-end adaptive video streaming," 2016. *arXiv:1808.03898*. [Online]. Available: https://arxiv.org/abs/1808.03898

[31] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A multi-class audio classification method with respect to violent content in movies using Bayesian networks," in *Proc. IEEE 9th Workshop Multimedia Signal Process.*, Oct. 2007, pp. 90–93.

[32] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[33] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[34] M. N. Garcia, D. Dytko, and A. Raake, "Quality impact due to initial loading, stalling, and video bitrate in progressive download video services," in *Proc. 6th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 129–134.

[35] M. Horowitz, F. Kossentini, N. Mahdi, S. Xu, H. Guermazi, H. Tmar, B. Li, G. J. Sullivan, and J. Xu, "Informal subjective quality comparison of video compression performance of the HEVC and H. 264/MPEG-4 AVC standards for low-delay applications," *Proc. SPIE*, vol. 8499, Oct. 2012, Art. no. 84990W.

[36] J. Redi, I. Heynderickx, B. Macchiavello, and M. Farias, "On the impact of packet-loss impairments on visual attention mechanisms," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2013, pp. 1107–1110.

[37] J. M. Boyce and R. D. Gaglianello, "Packet loss effects on MPEG video sent over the public Internet," in *Proc. 6th ACM Int. Conf. Multimedia*, 1998, pp. 181–190.

[38] S. Wenger, "H.264/AVC over IP," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 645–656, Jul. 2003.

[39] N. Harte, E. Gillen, and A. Hines, "TCD-VoIP, a research database of degraded speech for assessing quality in VoIP applications," in *Proc. 7th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, May 2015, pp. 1–6.

[40] J. M. Bland and D. G. Altman, "Statistics notes: Cronbach's alpha," *BMJ*, vol. 314, no. 7080, p. 572, Feb. 1997.

HELARD B. MARTINEZ received the B.S. degree in computer science from the Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), Peru, in 2010, and the M.Sc. and Ph.D. degrees in computer science from the University of Brasília (UnB), Brazil, in 2013 and 2019, respectively. He was a Visiting Researcher with University College Dublin (UCD), Ireland, in 2017. He was a Researcher with the Samsung R&D Institute Brazil, in 2019. He is currently a Postdoctoral Research Associate with the QxLab, University College Dublin. His current research interests include audio-visual quality of experience, machine learning, and immersive media.

ANDREW HINES (Senior Member, IEEE) is currently an Assistant Professor with the School of Computer Science, University College Dublin, Ireland. He leads the QxLab Research Group, where he is also an Investigator with the SFI CONNECT Centre for Future Networks and SFI Insight Centre for Data Analytics. His primary research interests are in media signal processing and machine learning for data driven quality of experience prediction across a variety of domains.

MYLÈNE C. Q. FARIAS (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the Universidade Federal de Pernambuco, Recife, Brazil, in 1995, the M.Sc. degree in electrical engineering from the Universidade Estadual de Campinas, São Paulo, Brazil, in 1998, and the Ph.D. degree in electrical and computer engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2004. She was previously a Research Engineer with CPqD, Campinas, Brazil, an Intern with Philips Research Laboratories, Eindhoven, The Netherlands, and the Intel Corporation, Phoenix, AZ, USA. She is currently an Associate Professor with the Department of Electrical Engineering, University of Brasília, Brasília, Brazil. Her current interests include video quality metrics, video processing, multimedia, watermarking, and machine learning.