Subjective Video Quality Test: Methodology, Database and Experience

Rafael Sotelo, *Senior Member, IEEE*, Jose Joskowicz, *Senior Member, IEEE*, Juan Pablo Garella, *Member, IEEE*, Diego Durán, *Member, IEEE*, Marcos Juayek, *Member, IEEE*

Abstract— **This paper presents a database with video clips and results of subjective video quality tests oriented to Digital Terrestrial Television (DTT). It makes them publicly available for the research community in the Quality of Experience field, in order to train, verify and validate video quality assessment objective models. The database consists of the results of two sets of tests, each one with one hundred high definition (HD) and one hundred standard definition (SD) video clips. The first one contains video clips with encoding impairments and simulated packet losses due to transmission degradation. The second one contains video clips obtained from real recordings from two broadcast TV channels using ISDB-T standard. The paper includes a discussion on the criteria to select the clips and the types of impairments, including packet losses patterns, as well as some experience collected during the subjective evaluation sessions. Finally, it presents a software application developed to automate the subjective tests, which also becomes available to the research community.**

Index Terms— **Quality of Experience, ISDB-Tb, Quality Assessment, video subjective tests database.**

I. Introduction

Subjective tests are the most accurate mean to measure video quality. Since subjects are involved, and their ratings are averaged, a Mean Opinion Score (MOS) can be computed. This MOS reflects the rating that an average viewer would assess a video clip.

There are methodologies that have been normalized and recommended by the International Telecommunication Union (ITU) to perform subjective tests and obtain results that can be considered valid. In particular, ITU-R BT.500-13 [1] describes some ways to show video sequences to viewers and collect their scores.

On the other hand objective metrics are automatic procedures that estimate video quality using models and algorithms based on content features and/or networks parameters. Since it is not feasible to permanently perform subjective tests for real time services, these metrics are necessary to ensure the quality of multimedia broadband or broadcasting operations. There are numerous objective methods to assess video quality [2].

Objective metrics try to mimic as closely as possible the evaluation of an average human being. The better the predictions of an objective method matches the average assessment performed by human beings who saw the same content, the better its performance. Therefore, the objective methods are trained and calibrated based on controlled subjective test results; and moreover, they must be validated by the results of other subjective tests, different from those used for training purposes.

To verify and validate new objective methods it is necessary to have available data sets with video clips and results of subjective tests performed according to appropriate recommendations. However, although there were numerous subjective tests conducted over the years, few databases are openly available to the scientific community. References to most publicly available video databases can be found in [3]. There are also ongoing efforts to build and use large scale databases [4, 5].

In previous works we have presented the development of an objective metric for measuring video quality in Digital Terrestrial Television (DTT) [6,7]. The ISDB-T Standard uses H.264/AVC for video compression and MPEG-2 Transport Stream (TS) syntax for packaging and multiplexing video, audio and data signals in the digital broadcasting system. To calibrate, verify and later validate our metric we have generated our own video database according to the ITU-R BT.500 methodology and also considering the requirements in source and channel coding imposed by the DTT standard.

A first series of subjective tests was performed using one hundred High Definition (HD) video clips presented to twenty-five subjects and one hundred in Standard Definition (SD) resolution presented to nineteen subjects. For this first series five uncompressed source video clips were used in order to simulate different encoding impairments and transmission errors according to the encoding and transmission processes provided by the ISDB-T standard.

Then, with the aim of validate the objective model a second series of subjective tests was performed also using one hundred clips in HD format presented to twenty-seven subjects and one hundred clips in SD format presented to eighteen

This work was supported by ANII-MIEM/Dinatel FST_1_2012_1_8147.

Rafael Sotelo is with Facultad de Ingeniería, Universidad de Montevideo, Montevideo, Uruguay (Phone: +598 27067630; email: rsotelo@um.edu.uy)
Jose Joskowicz is with Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay (Phone: +598 27110974; email: josej@fing.edu.uy)
Juan Pablo Garella is with Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay (Phone: +598 27110974; email: jpgarella@fing.edu.uy)
Diego Durán is with Facultad de Ingeniería, Universidad de Montevideo, Montevideo, Uruguay (Phone: +598 27067630; email: dduran@correo.um.edu.uy)
Marcos Juayek is with Facultad de Ingeniería, Universidad de Montevideo, Montevideo, Uruguay (Phone: +598 27067630; email: mjuayek@correo.um.edu.uy)

subjects, each one with a duration of 9 to 12 seconds. For this purpose the video clips were originated with real recordings from two free-to-air TV broadcast channels from Montevideo, Uruguay, with different conditions of reception. This second series of subjective tests and their corresponding video clips are part of our database as the independent data set for validation purposes.

This article presents our database and makes it publicly available for the research community in the Quality of Experience field. The criteria employed to select the source video clips is described in Section II for both the first and second series of tests. It also presents the controlled-impairments made on the encoding and transmission simulated stages and the impairments obtained in the “real world” recordings.

In Section III the article details the characteristics of the participant subjects and other useful information from the subjective test experience. The incidence of personal information collected from the evaluators on the results is shown (such as age, gender, number of daily hours exposed to television and TV type used). In addition, it is described the number of evaluators per session, the percentage of men and women, the influence of the presence of other evaluators in the same session, among others factors.

Section IV includes a brief description of a platform we developed to automate the subjective tests which made them more efficient, by accelerating the tests and preventing errors. The application is also publicly available to the research community. Finally, Section V presents the main conclusions.

II. Video Clips Selection and Preparation

The database includes information of two sets of video quality subjective tests, each one with one hundred HD and one hundred SD video clips: one with controlled degradation which we will call *Set I HD* and *Set I SD*, and a second one used for validation purposes with clips obtained from real recordings from two broadcast TV channels using ISDB-T standard, called *Set II HD* and *Set II SD*.

The Recommendation ITU-R BT.500-13 describes the methodology to perform subjective tests. In particular, it includes some ways of showing the content to the evaluators. Our database was built using a Single Stimulus (SS) method, including the reference sequence. In particular, the Absolute Category Rating with Hidden Reference (ACR-HR) method, according the Recommendation ITU-T P.910 [8], was used. Each clip, either degraded or not, is presented separated to the evaluator. Typically the clips duration is between 9 and 12 seconds in this kind of tests. The non-degraded (original) clips are shown with no special remarks, mixed with the other clips. The recommendation states that the order of presentation has to be made random.

A. Set I

In order to be useful to train a model to predict video quality, a video clip set must cover a wide range of situations in which the model will be used. In our case, codec H.264/AVC has been used to perform source compression for both HD (1920x1080p@50fps) and SD (720x576p@50fps).

In order to span a wide range of contents, it is necessary to select video clips with different spatial and temporal activity. For this purpose the average amplitude of the motion vectors (MV) and the average sum of absolute differences of residual blocks (SAD) were used as indicators to perform a proper selection.

A total of five uncompressed video clips in HD format of 9 to 12 seconds long were selected as sources. These videos were obtained from the Consumer Digital Video Library (CDVL) [9,10] and from the IRCCyN IVC1080i Video Quality Database [11]. Table I shows a representative frame and a description of the original video clips including content, SAD and MV. Each of these clips was encoded according to the ISDB-T Standard using H.264/AVC with the popular software *ffmpeg* [12]. The encoding settings used are shown in Table II.

To obtain the *Set I HD* five different bitrates were selected aiming to cover a wide range of real DTT broadcasting situation: 3.5, 5, 7.5, 10 and 14 Mbps. This constitutes a set of 5 × 5 = 25 HD video clips degraded only by encoding impairments.

Table I
Original Source Content Selected for *Set I HD & SD*

Preview	Description	
	Name	Fox & Bird
	Description	Cartoon drawings of a fox and a bird.
	Content	Animated Cartoon
	SAD	90
	MV	16
	Name	Golf
	Description	Golf put on the green.
	Content	Sports
	SAD	156
	MV	2
	Name	Concert
	Description	Jean-Michel Jarre is speaking in a concert.
	Content	Music
	SAD	273
	MV	11,6
	Name	Foot
	Description	Goal during a German football match.
	Content	Sports
	SAD	283
	MV	14
	Name	Voile
	Description	A boat on water is moving.
	Content	Movie
	SAD	198
	MV	8,3

With the aim to include transmission errors, a reduced group of the video clips obtained so far underwent a procedure of individual Transport Stream Packet (TSP) extraction to simulate packet losses originated on DTT transmissions. The types of degradations incorporated were based on loss patterns found on actual DTT recordings from free-to-air broadcast transmissions with low reception at the front of the receiver [6]. Throughout the analysis of the recorded TS packets, periods with homogeneous losses (possibly corresponding to signal fading) with a typical length of more than 10 seconds and periods with bursts errors (many losses in a short period of time) were found. In most cases, the bursts length was less than one second, followed by periods of many seconds without losses. Near half of the losses consist of individual TS packets (i.e., the continuity counter of the header of the TSP reports only one missing TS packet). Based on these observations, different pattern losses were simulated and applied to each one of the video clips, according to Table III. In Fig.1 a preview of two representative frames of each one of the contents affected with packet losses is shown.

TABLE II
H.264 ENCODING SETTINGS

Settings	SD	HD
Profile	Main	High
Level	3.1	4.1
Group of Pictures Length	33	33
B Consecutive frames	2	2
Bitrate	CBR	CBR
Min Bit rate	0.7 Mbps	3.5 Mbps
Max Bit rate	4 Mbps	14 Mbps
Scan type	Progressive	Progressive
Frame rate	50 fps	50 fps
Stream syntax	MPEG-TS	MPEG-TS

TABLE III
PACKET LOSS PATTERNS TESTED

Packet Loss Pattern	Percentage of packet loss
No packet loss	0% along all the video clip
Uniform	0.3% along all the video clip
One Burst	0.1% inside the Burst; 0% outside the Burst
One Burst	10% inside the Burst; 0% outside the Burst
Two Bursts	0.1% inside the Burst; 0% outside the Burst
Two Bursts	10% inside the Burst; 0% outside the Burst
Three Bursts	0.1% inside the Burst; 0% outside the Burst
Three Bursts	10% inside the Burst; 0% outside the Burst

To obtain the *Set I SD* each one of the five uncompressed HD clips were down-converted to SD resolution using *ffmpeg*. Analogously, the obtained uncompressed SD clips were encoded to H.264/AVC using *ffmpeg* with four different bitrates commonly used in SD resolution: 0.7, 1.5, 2.8, 4.0 Mbps. Therefore, a set of 5 × 4 = 20 SD video clips degraded only by encoding impairments was obtained.

As it was done with the *Set I HD* a reduced group of the encoded video clips underwent a procedure of individual TS packet extraction, with the loss patterns shown in Table III. This procedure led to a total number of one hundred clips with HD resolution and one hundred with SD resolution.

B. *Set II*

Set II was conceived as a validation data set. Both, *Set II HD* and *Set II SD* have one hundred clips each. In this set the video clips used are real recordings of 9 to 12 seconds long of DTT broadcast transmissions in Montevideo, Uruguay.

Each video was encoded and transmitted to the air according to the ISDB-T Standard. Two TV Broadcasters with different H.264 encoders were selected for this purpose while they were on a test phase and transmitting test signals.

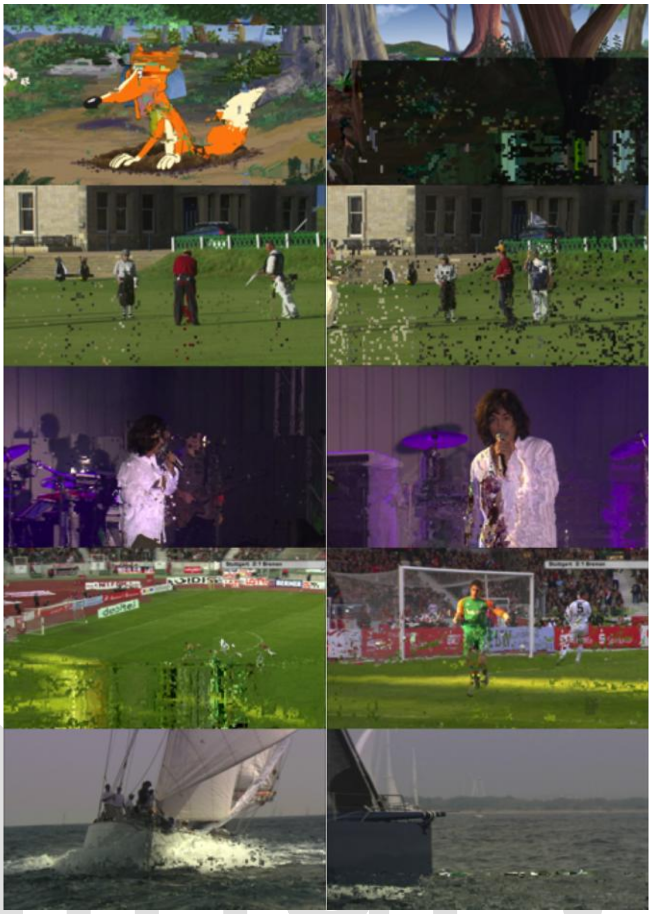

Fig.1 Preview of representative frames showing how transmission errors affect the selected source video clips.

Hundreds of short video recordings were made. The video clips were finally selected taking into account the temporal activity, scene cuts, content, encoding impairments and packet losses patterns. Figure 2 shows a preview of some of the used clips, without transmission impairments. Figure 3 shows a preview of some of the used clips, with transmission impairments (i.e., with TS packets loss).

III. ABOUT THE SUBJECTS AND THE SUBJECTIVE TESTS

One of the aspects beside subject’s scores that were intended to be evaluated was if there was a relation between previous personal information of the evaluators and their scores. In order to achieve this end subjects were asked to complete a brief survey before each test session started. The survey includes gender (male, female), age, education level (school, high school, university) type of the TV set that usually uses to watch TV (CRT, LCD or LED), size of the TV set that usually uses to watch TV (14”, 21”, 29”, 32” 41” or more) and the number of hours per day that the evaluator spends watching TV.

The subjects’ ages varied from 17 to 68 years old, with an average of 31 years old. 75% of the subjects were males and 25% were females. All the subjects voluntarily agreed to participate in the tests, without any payment.

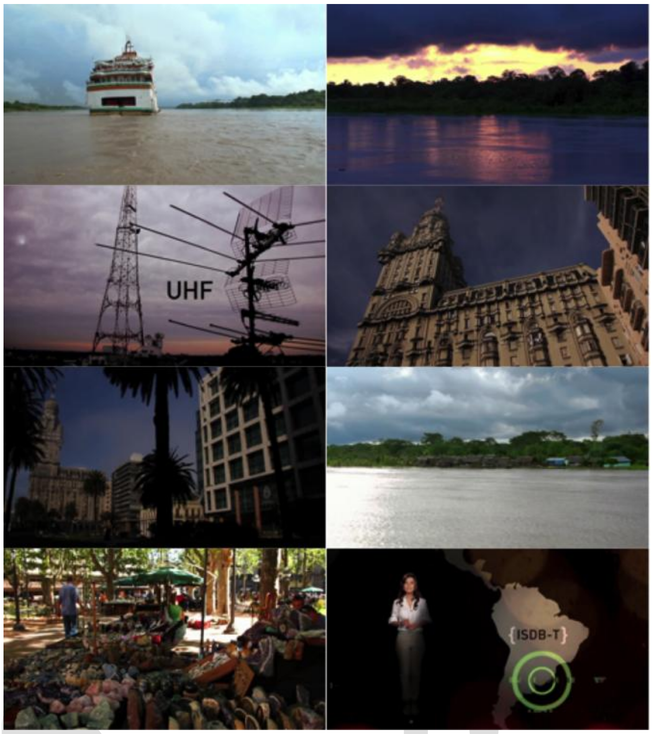

Fig. 2 Preview of representative frames of recorded video clips from the *Set II HD*.

TABLE IV
GENERAL VIEWING CONDITIONS FOR SUBJECTIVE ASSESSMENTS IN LABORATORY ENVIRONMENT

Parameter	ITU-R BT.500-13	Measured at the laboratory
Ratio of luminance of inactive screen to peak luminance	≤ 0.02	0,011
Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white	≈ 0,01	0,013
Display brightness and contrast	Set up via PLUGE	Set up via PLUGE
Maximum observation angle relative to the normal	30°	28°
Ratio of luminance of background behind picture monitor to peak luminance of picture	≈ 0,15	0,14
Chromaticity of background	D_{65}	White
Other room illumination	Low	Low

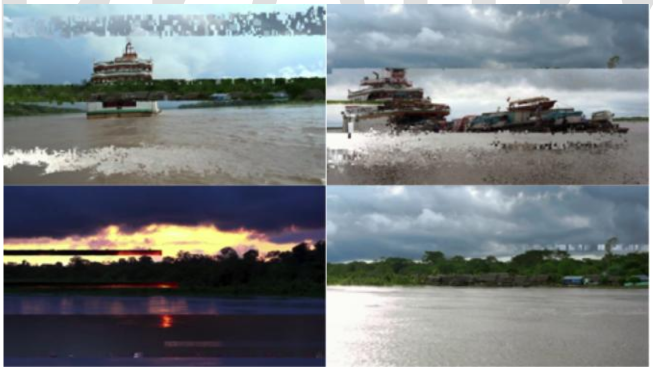

Fig. 4 Preview of representative frames showing how transmission errors affect the frames of recorded video clips from free-to-air DTT.

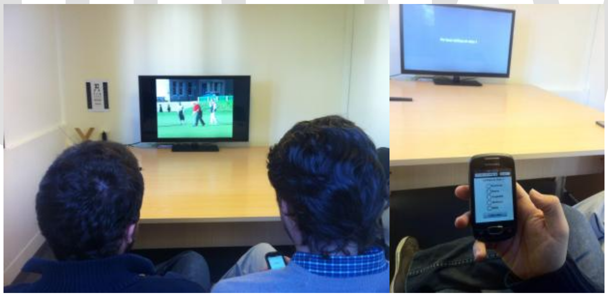

Fig. 3 Controlled environment for subjective tests

A controlled environment was set up, according to the guidelines of Recommendation ITU-R BT.500. We used a 42” LED TV set. Five seats were accommodated in front of the TV, in accordance with the maximum admitted angle and the appropriate distance to the screen. The viewing conditions are summarized in Table IV. Figure 4 shows a picture of the room during an evaluation session.

The *Set I HD* took six different sessions to be completed with a total of 25 participants. Each session was attended with four to five subjects obtaining a total of 2500 ratings. Four subjects failed either the vision test and/or the daltonism (Ishihara) test. In the post-screening stage one subject was identified as an outlier, as the Pearson Correlation between the subject’s scores and the average was below the threshold established in the test plan. Then, MOS values were calculated based on the remaining 20 evaluators.

The *Set I SD* took six different sessions to be completed with a total of 19 participants. Each session was attended with three to five subjects obtaining 1900 ratings. All subjects successfully passed the vision and daltonism tests. In the post-screening stage three subjects were identified as outliers, as the Pearson Correlation between each subject’s scores and the average was below the threshold established in the test plan. Then, MOS values were calculated based on the remaining 16 evaluators.

After analyzing the results of both data sets it has been concluded that there is not meaningful differences between the ratings by men and women and neither between the ratings by people of different ages. Nevertheless, it was observed, as it is shown in Figure 5, that there may be a tendency of how people who regularly watch TV on CRT type monitors tend to rate better the image quality (note that the subjective tests were performed using a 42” LED TV set).

It has also been observed a certain tendency of people that spend more hours watching TV to give worst quality qualifications. In Figure 6 this trend can be observed.

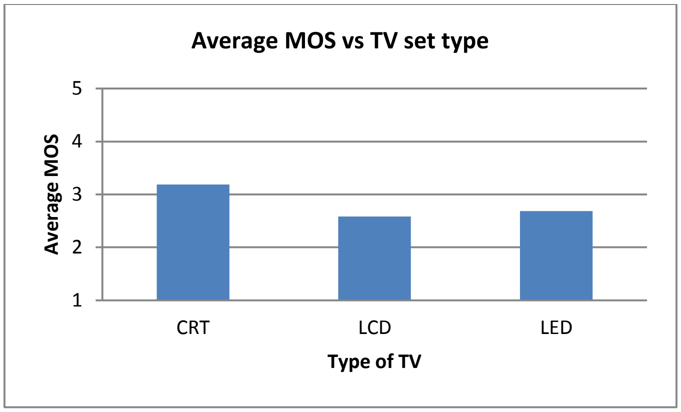

Fig. 5 Average MOS according to the subject's type of TV set: CRT, LCD or LED.

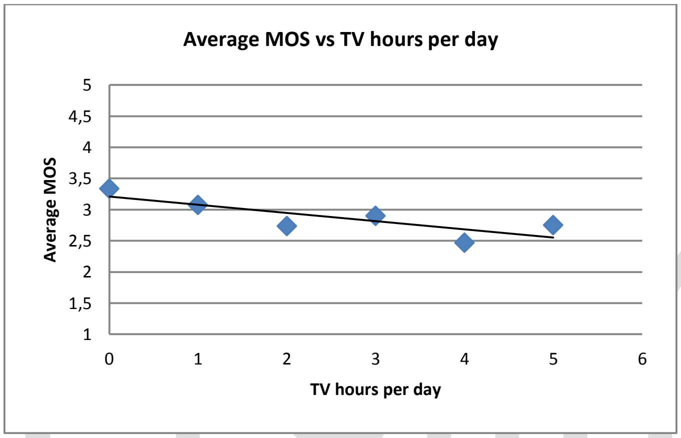

Fig. 6 Average MOS versus hours per day subjects watched TV.

The *Set II HD* took eight different sessions to be completed with a total of 27 participants. Each session was attended with two to five subjects obtaining a total of 2700 ratings. Three subjects failed the vision tests. In the post-screening there were no outliers found.

The *Set II SD* took seven different sessions to be completed with a total of 18 participants. Each session was attended with one to five subjects obtaining a total of 1800 ratings. All subjects successfully passed the vision tests and the daltonism test. In the post-screening stage two subjects were identified as outliers. Then, MOS values were calculated based on the remaining 16 evaluators.

On both *Set II HD* and *Set II SD* the same trends are verified with respect to the worst assessment by subjects who spend more hours watching TV and by those who have CRT's TV set monitors.

Since each session had the same amount of video clips (one hundred) to be displayed with a similar length, all sessions of all data sets lasted approximately the same time. An average session took approximately 50 minutes long and none of them exceeded more than an hour of duration. The time of the session was measured from the moment the subjects were seated and started reading the instructions until they finished rating the last video clip.

A maximum number of five subjects per session were used, seated in the same room and scoring the same video clip at the same time. One may ask if the presence of more than one subject can disturb or distract subjects scoring tasks. At the beginning of the tests campaign it was noted that a presence of an authority inside the room was needed in order to generate an affine environment to work for the subjects. Therefore, in almost all the sessions one of the researchers involved was seated inside the room with the subjects. According to our experience from a total of 27 sessions performed only one was discarded due to disturbing factors between the attending subjects that led to a notorious low correlation with the others test results.

The other precaution that needed to be taken into account due to the amount of subjects per session was to explain that it was an individual task before starting the tests in order to avoid subjects to look at next subject's scores, maintaining the independence of the results. As complementary information, only one session was interrupted during the tests campaign, since one subject needed to leave the room for personal reasons.

Finally it is worthy to mention that the results obtained in all data sets (*Set I HD*, *Set I SD*, *Set II HD* and *Set II SD*) are consider "formal", as the Recommendation ITU-R BT.500 establishes a minimum of 15 assessors to consider the evidence as "formal".

IV. Platform to Automate the Subjective Tests

After selecting the ACR-HR method to perform the tests, available alternatives from traditional paper forms to software tools specifically designed to carry out the tests were analyzed. Traditional paper based forms are tedious to fill by the subjects and error prone during the data collection process and post analysis. Imposing a fixed time to enter the ratings does not help either. Specifically designed hardware devices are difficult to obtain and software applications normally lead to take tests on a PC, where the stimulus and the scale of assessment are presented together on the same screen, limiting the number of subjects per session and the resolution of the stimuli.

In order to overcome these disadvantages, an open source web based software system was designed, developed and deployed. It automates subjective tests for video quality measurement, offering the possibility to create and configure different test sessions and reordering in a random manner between sessions the stimuli for its play back. The evaluators interface runs on a mobile device (smartphones or tablets), provided to each participant at the beginning of the tests.

As an initial step, the system performs a short survey for each evaluator, collecting the required information (age, gender, etc.). The system allows multiple subjects to assess the test stimuli at the same time. After each video clip is played back the evaluators send their ratings to the system using the web-based application that runs on the mobile browser (see Figure 4). Then, after the system received all the ratings it starts playing the next video clip. The system automatically collects all the information in a database and after the session is finished a spreadsheet can be exported with each subject scores and the information collected before the tests began. This facilitates the tasks of collecting and processing the evaluator's ratings. As it has been said above, the application is available under request to any of the authors. An overall illustration of the global system can be seen in Figure 7. For more information regarding the application refer to [13].

With this system, the following advantages can be obtained:

- There is no need for special hardware devices.
- Many viewers can participate in the same session (limited only to respect the viewing angle and distance and room space allowed by recommendations).
- The design can easily be extended to make quality evaluation in real scenarios (i.e. crowdsourcing, people at home), not only at controlled environments.
- Discrete or continuous scales can be implemented.
- There are no cables connected to the devices, neither papers nor pens, making it more user-friendly.

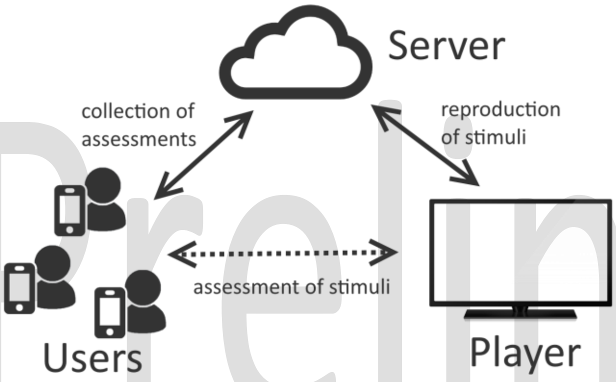

Fig. 7 Illustration of the automatic system for subjective tests.

V. Conclusions

This work provides the community with a database of degraded videos for Digital Television applications. We are including a set of video clips with simulated and controlled impairments, and also a set of videos recorded from two ISDB-T broadcasters in Uruguay (with and without transmission impairments). The vast majority of available databases are designed for IP networks, where the packet loss pattern is different comparing with DTT, so this new database can provide real value for researches in the broadcasting area.

The presented results include some personal characteristics of the evaluators (age, gender, hours spent viewing TV per day, type of TV set used at home, etc.), along with the scores given for each video clip. We have observed that several of these characteristics do not influence the average ratings, but others show clear trends. People who spend more hours viewing TV at home tend to rate worse the video quality, and people using CRT type monitors tend to rate it better. These kind of tendencies should be considered in the future, when performing new subjective video quality tests.

Several recommendations on what to consider when performing subjective tests with multiple simultaneous participants were indicated.

The video database and the automated application used for performing subjective tests are available under request to any of the authors.

References

[1] Recommendation ITU-R. BT.500-13 (2012), Methodology for the Subjective Assessment of the Quality of Television Pictures". *ITU Telecom. Standardization Sector of ITU*.

[2] Chikkerur, S., Sundaram, V., Reisslein, M., & Karam, L. J. (2011). Objective video quality assessment methods: A classification, review, and performance comparison. *Broadcasting, IEEE Transactions on*, *57*(2), 165-182.

[3] Fliegel, K., & Timmerer, C. (2013). WG4 databases white paper v1. 5: QUALINET multimedia database enabling QoE evaluations and benchmarking.*Prague/Klagenfurt, Czech Republic/Austria*. http://dbq-wiki.multimediatech.cz/_media/qi0306.pdf

[4] Leszczuk, M., Janowski, L., & Barkowsky, M. (2013, December). Freely available large-scale video quality assessment database in Full-HD resolution with H. 264 coding. In *Globecom Workshops (GC Wkshps), 2013 IEEE* (pp. 1162-1167). IEEE.

[5] Barkowsky, M., Masala, E., Van Wallendael, G., Brunnström, K., Staelens, N., & Le Callet, P. (2015). Objective Video Quality Assessment—Towards Large Scale Video Database Enhanced Model Development. *IEICE Trans. on Communications*, *98*(1), 2-11.

[6] Joskowicz, J., & Sotelo, R. (2014). A Model for Video Quality Assessment Considering Packet Loss for Broadcast Digital Television Coded in H. 264. *International Journal of Digital Multimedia Broadcasting*, *2014*.

[7] Sotelo, R., & Joskowicz, J. (2013, June). Video Quality Indicators for ISDB-Tb Free to Air Digital Television. In *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on* (pp. 1-6). IEEE.

[8] Recommendation ITU-T P.910 (1999), "Subjective video quality assessment methods for multimedia applications", *ITU Telecom. Standardization Sector of ITU.*

[9] M. Pinson, "The Consumer Digital Video Library [Best of the Web]" IEEE Signal Processing Magazine, vol. 30, no. 4, pp. 172-174, Jul. 2013. doi: 10.1109/MSP.2013.2258265

[10] Consumer Digital Video Library, [online], http://www.cdvl.org

[11] Péchard, S., Pépion, R., & Le Callet, P. (2008). Suitable methodology in subjective video quality assessment: a resolution dependent paradigm. In *International Workshop on Image Media Quality and its Applications*, IMQA2008 (p. 6).

[12] FFmpeg, [online], *http://www.ffmpeg.org*

[13] Joskowicz, J., Sotelo, R., Juayek, M., Durán, D., & Garella, J. P. (2014, September). Automation of Subjective Video Quality Measurements. In *Proceedings of the Latin America Networking Conference on LANC 2014* (p. 7). ACM.