International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

Recommendation ITU-R BT.2095-1

(06/2017)

Subjective assessment of video quality using expert viewing protocol

BT Series

Broadcasting service (television)

International Telecommunication Union

Foreword

The role of the Radiocommunication Sector is to ensure the rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including satellite services, and carry out studies without limit of frequency range on the basis of which Recommendations are adopted.

The regulatory and policy functions of the Radiocommunication Sector are performed by World and Regional Radiocommunication Conferences and Radiocommunication Assemblies supported by Study Groups.

Policy on Intellectual Property Right (IPR)

ITU-R policy on IPR is described in the Common Patent Policy for ITU-T/ITU-R/ISO/IEC referenced in Annex 1 of Resolution ITU-R 1. Forms to be used for the submission of patent statements and licensing declarations by patent holders are available from http://www.itu.int/ITU-R/go/patents/en where the Guidelines for Implementation of the Common Patent Policy for ITU-T/ITU-R/ISO/IEC and the ITU-R patent information database can also be found.

Series of ITU-R Recommendations

(Also available online at http://www.itu.int/publ/R-REC/en)

Series	Title
BO	Satellite delivery
BR	Recording for production, archival and play-out; film for television
BS	Broadcasting service (sound)
BT	**Broadcasting service (television)**
F	Fixed service
M	Mobile, radiodetermination, amateur and related satellite services
P	Radiowave propagation
RA	Radio astronomy
RS	Remote sensing systems
S	Fixed-satellite service
SA	Space applications and meteorology
SF	Frequency sharing and coordination between fixed-satellite and fixed service systems
SM	Spectrum management
SNG	Satellite news gathering
TF	Time signals and frequency standards emissions
V	Vocabulary and related subjects

Note*: This ITU-R Recommendation was approved in English under the procedure detailed in Resolution ITU-R 1.*

Electronic Publication
Geneva, 2017

RECOMMENDATION ITU-R BT.2095-1

Subjective assessment of video quality using expert viewing protocol

(2016-2017)

Scope

This Recommendation describes the method to subjectively assess video quality of moving images by means of the expert viewing protocol, with the participation of a reduced number of viewers, all selected among experts in the relevant video processing area.

Keywords

Television, video quality, subjective assessment, expert viewing

The ITU Radiocommunication Assembly,

considering

a) that source coding technologies for digital television applications are continuously improving both in efficiency and in visual performance;

b) that the continuous evolution of video coding technologies implies an ever increasing demand for evaluation methods to assess technical and visual performances;

c) that the compression efficiency and visual performances of new video source coding technologies require new and more efficient visual assessment and ranking methods;

d) that the evaluation methods specified in current ITU-R Recommendations is highly demanding in terms of time and human resources, and that they often do not take into account the technical evolution of the displays and of the final user fruition;

e) that new approaches in expert viewing protocols have recently shown better efficiency and performance, in terms of time and overall cost, compared to those provided by methods based on the use of non-expert viewers;

f) that if the results of expert viewing protocol cannot be considered as a replacement of the results provided by a formal subjective assessment protocol, the results of expert viewing protocol can be considered a valuable preliminary indication of the performances of the systems under test;

g) that the growing technology evolution in the area of flat panel displays has drastically modified the viewing condition normally used by experts;

h) that ISO/IEC have already successfully used new protocols based on expert viewing in the evaluation of new video source coding technologies,

recommends

1 that, in the assessment of new digital video coding technologies, consideration should be given to the use of the expert viewing protocol, described in Annex 1;

2 that the expert viewing protocol should be implemented using professional flat panel displays and the laboratory set-up described in Annex 1.

Note 1 – Annex 2 (informative) shows results of subjective experiment using the expert viewing protocol as mentioned in *considering h).*

Annex 1

Expert viewing protocol for the evaluation of the quality of video material

1 Laboratory set-up

1.1 Display selection and set-up

The display used should be a flat panel display featuring performances typical of professional applications (e.g. broadcasting studios or vans); the display diagonal dimension may vary from 22’ (minimum) to 40’ (suggested), but it may extend to 50’ or higher, when image systems with a resolution of HDTV or higher are assessed.

It is allowed to use a reduced portion of the active viewing area of a display; in this case the area around the active part of the display should be set to mid-grey. In this condition of use it should not be allowed to set the monitor to a resolution different from its native one.

The display should allow a proper set-up and calibration for luminance and colour, using a professional light-meter instrument. The calibration of the display should comply with the parameters specified in the relevant Recommendation for the test being undertaken.

1.2 Viewing distance

The viewing distance at which the experts are seated should be chosen according to the resolution of the screen, and to the height of the active part of the screen, according to the design viewing distance as described in Recommendation ITU-R BT.2022 or shorter viewing distance, according to the requirements in terms of critical viewing conditions.

1.3 Viewing conditions

An expert viewing protocol (EVP) experiment should not necessarily be run in a test laboratory, but it is important that the testing location is protected from audible and or visible disturbances (e.g. a quiet office or meeting room may be used as well).

Any direct or reflected source of light falling on the screen should be eliminated; other ambient light should be low, maintained to the minimum level that can allow filling scoring sheets (if used).

The number of experts seated in front of the monitor, may vary according to the screen size, in order to guarantee the same image rendering and stimulus presentation for all the viewers.

2 Viewers

The viewers participating to an EVP experiment should be expert in the domain of study.

Viewers should not necessarily be screened for visual acuity or colour blindness, since they should be chosen among qualified persons.

The minimum number of different viewers should be nine.

To reach the minimum number of viewers, the same experiment may be conducted at the same location repeating the test, or in more than one location. The scores from different locations participating to an expert viewing session may be statistically processed together.

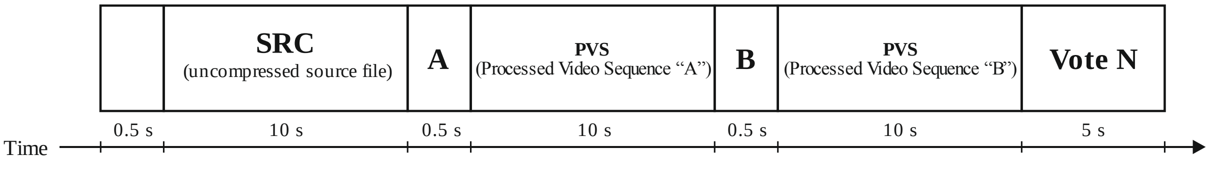
3 The basic test cell

The material to be presented to the experts should be organised creating a basic test cell (BTC) for each couple of coding conditions to be assessed (see Fig. 1).

The source reference sequences (SRC) and the processed video sequences (PVSs) clips to consider in a BTC should always be related to the same video sequence, in order that the experts may be able to identify any improvement in visual quality provided by the compression algorithms under test.

FIGURE 1

Timings of a basic test cell for the expert viewing protocol


SRC
(uncompressed source file)
A
PVS
(Processed Video Sequence "A")
B
PVS
(Processed Video Sequence "B")
Vote N
0.5 s
10 s
0.5 s
10 s
0.5 s
10 s
5 s
Time

BT.2095-01

The BTC should be organised as follows:

- 0.5 seconds with the screen set to a mid-grey (mean value in the luminance scale);
- 10 seconds presentation of the reference uncompressed video clip;
- 0.5 seconds showing the message "A" (first video to assess) on a mid-grey background;
- 10 seconds presentation of an impaired version of the video clip;
- 0.5 seconds showing the message "B" (second video to assess) on a mid-grey background;
- 10 seconds presentation of an impaired version of the video clip;
- 5 seconds showing a message that asks the viewers to express their opinion.

The message "Vote" should be followed by a number that helps to get synchronised on the scoring sheet.

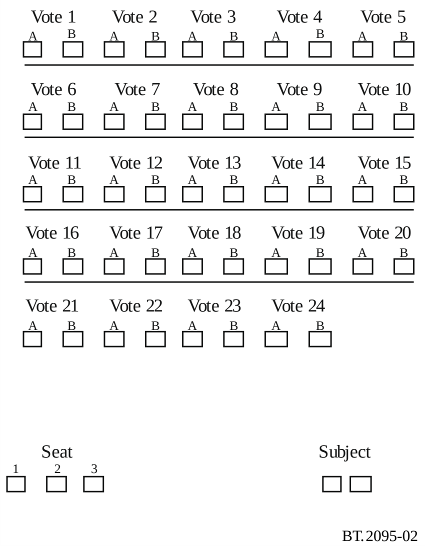
3.1 Scoring sheet and rating scale

As shown in Fig. 1, the presentation of the video clips should be arranged in such a way that the unimpaired reference (SRC) is shown at first, followed by two impaired video sequences (PVS). The order of presentation of the PVS should be randomly changed for each BTC and the viewers should not know the order of presentation.

FIGURE 2

Example of scoring sheet for a 24-BTC expert viewing session

Session 1



An 11 grades numerical scale from 10 (imperceptible impairments) to 0 (very annoying impairments) is used.

Table 1 provides guidance about the meaning of the 11 grades numerical scale.

TABLE 1

Meaning of the 11 grades numerical scale

Score	**Impairment item**	
10	Imperceptible	
9	Slightly perceptible	somewhere
8		everywhere
7	Perceptible	somewhere
6		everywhere
5	Clearly perceptible	somewhere
4		everywhere
3	Annoying	somewhere
2		everywhere
1	Severely annoying	somewhere
0		everywhere

The viewers are asked to fill in a questionnaire made of two boxes (labelled as “A” and “B”) for each BTC, writing in each of the two boxes a score selecting it from the 11 grades numerical scale.

Figure 2 provides an example of scoring sheet for a session consisting of 24 BTC.

For each BTC, viewers fill both the box identified by the letter **A** (to rate the video clip shown as first) and the box identified by the letter **B** (to rate the video clip shown as second).

The presentation of the original unimpaired video clip allows the experts to more easily evaluate any impairment.

The meaning of the 11 grade numerical scale should be carefully explained during "training sessions" as described below.

3.2 Test design and session creation

The order of presentation of the BTC should be set in a random order by the test designer, in such a way that the same video clip is not shown two consecutive times as well as the same impaired clip.

Any viewing session should begin with a "stabilization phase" including the "best", the "worst" and two "mid quality" BTC among those included in each test session. This will allow the viewers to have an immediate impression of the quality range, already at the beginning the test session.

If the viewing session is longer than 20 minutes, the test designer should split it into two (or more) separate viewing sessions, each of them not exceeding 20 minutes. In this case, the "stabilization phase" should be provided before each viewing session.

3.3 Training

Even if this procedure is foreseen for use with the participation of experts, a short (5-6 BTC) training viewing session should preferably be organised prior to each experiment.

The video material used in the training session may be the same that will be used during the actual sessions, but the order of presentation should be different.

The viewers should be trained on the use of the 11-grade scale by asking them to carefully look at the video clips shown immediately after the message "A" and "B" on the screen, and check whether they can see any difference to the video clip shown as first (the SRC).

4 Data collection and processing

The scores should be collected at the end of each session and logged on an electronic spreadsheet to compute the MEAN values.

A "post screening" of the viewers should desirably be performed, using a linear Pearson's correlation.

The "correlation" function should be applied considering all the scores of each subject in relation to the mean opinion scores (MOS); a threshold may be set to define each viewer as "acceptable" or "rejected" (Recommendation ITU-T P.913 suggests the use of a "reject" threshold value equal to 0.75).

5 Terms of use of the expert viewing protocol results

The expert viewing protocol (EVP) may be used when time and resources do not allow running a formal subjective assessment experiment.

EVP requires less time than a formal subjective assessment and may be executed in an "informal" environment, assuming that the ambient in which it is run is protected by any visual and audible external disturbance.

The only mandatory conditions are related to the ambient illumination and to the viewing conditions (display, angle of observation and viewing distance) as described in the above paragraphs.

6 Limitations of use of the EVP results

Even if the EVP is demonstrating to be able to provide acceptable results with only nine viewers, the MOS provided by an EVP experiment cannot be considered as a replacement of the results obtainable with a formal subjective assessment experiment.

The MOS data obtained using EVP might be used to get a preliminary indication of the level of impairment.

The MOS data obtained using EVP might be used to make a preliminary ranking of the video processing schemes under evaluation.

Where retained convenient or necessary, an EVP experiment can be run in parallel in more locations, assuming the viewing conditions, viewing distance and the test design are identical.

If the number of expert viewers involved in the same EVP experiment, also if running the experiment in different locations, is equal or higher than 15, the raw subjective data might be processed to obtain MOS, standard deviation and confidence interval data, that may help to perform a more accurate ranking of the cases under test. In this last case more accurate inferential statistical analysis may be performed, e.g. T-Student test.

Annex 2
(informative)

Application of the Expert Viewing Protocol and its behaviour in the presence of a large number of expert assessors

This informative Annex provides information on the results of two different subjective assessment EVP sessions on coded HD and UHD video clips, performed during the 117th MPEG meeting applying the provisions of Recommendation ITU-R BT.2095 in order to quickly and reliably rank two different source-coding methods.

Due to the presence of a large number of experts attending the 117th MPEG meeting, the number of assessors participating to the two EVP sessions extended well beyond the number of 9 as recommended in Recommendation ITU-R BT.2095; 30 experts attended the HD EVP test session and 32 experts attended the UHD EVP test session.

The wide participation of expert assessors provided the opportunity to analyse the MOS data, in order to verify the level of reliability inherent in the use of Recommendation ITU-R BT.2095 when ranking coded video clips.

In this assessment four sets of viewers (i.e. 9, 12, 15 and 18) are considered, performing a comparison between the MOS values obtained using 9 experts and the MOS values obtained using 12, 15 and 18 viewers.

The goal was to compare the ranking obtained from 9 experts (and therefore in line with EVP protocol) with the rankings obtained from 12, 15 and 18 experts (and therefore similar to a Formal Subjective Assessment Test).

What appears in Fig. 3 (experiment made on UHD content) and Fig. 4 (experiment made on HD content) is that the results of rankings are very similar for all the four cases considered.

If we consider the results obtained considering 18 viewers like a sort of "ground truth", we can plot the graphs in Fig. 3 and Fig. 4 ranking the test points according to the MOS values obtained considering 18 viewers (continuous red line).

The other lines in the graphs show the results obtained considering 9 viewers (dotted red line), 12 viewers (blue dashed line) and 15 viewers (continuous green line).

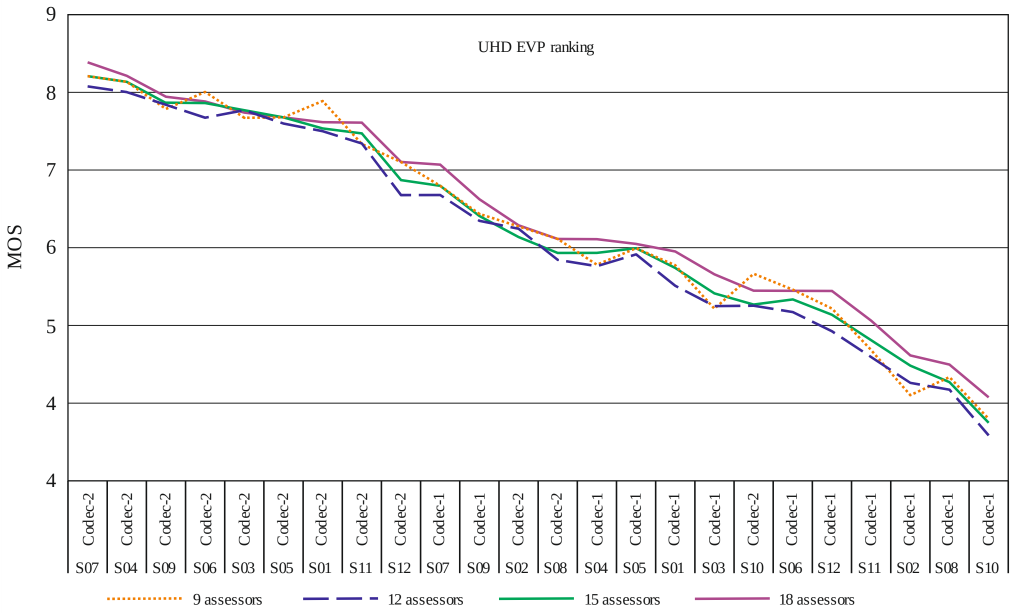
Observing the results plotted in Figs 3 and 4, it can be noted that:

- the 15 and the 18 viewers graphs show an homogeneous slope from high quality to low quality MOS values;
- the 9 and the 12 viewers graphs show some "inversions" of ranking compared to the 18 viewers graph, even if the variations of scores are rather limited in their extension.

In conclusion, the EVP experiments here described show a very good performance of EVP protocol, confirming what stated in the text of the Recommendation ITU-R BT.2095, i.e. the use of the EVP protocol, even if it cannot be considered a full replacement of a formal subjective experiment, might be considered an evaluation procedure stable and providing results very close to those obtained when many more viewers are available and a formal subjective assessment is done.

FIGURE 3

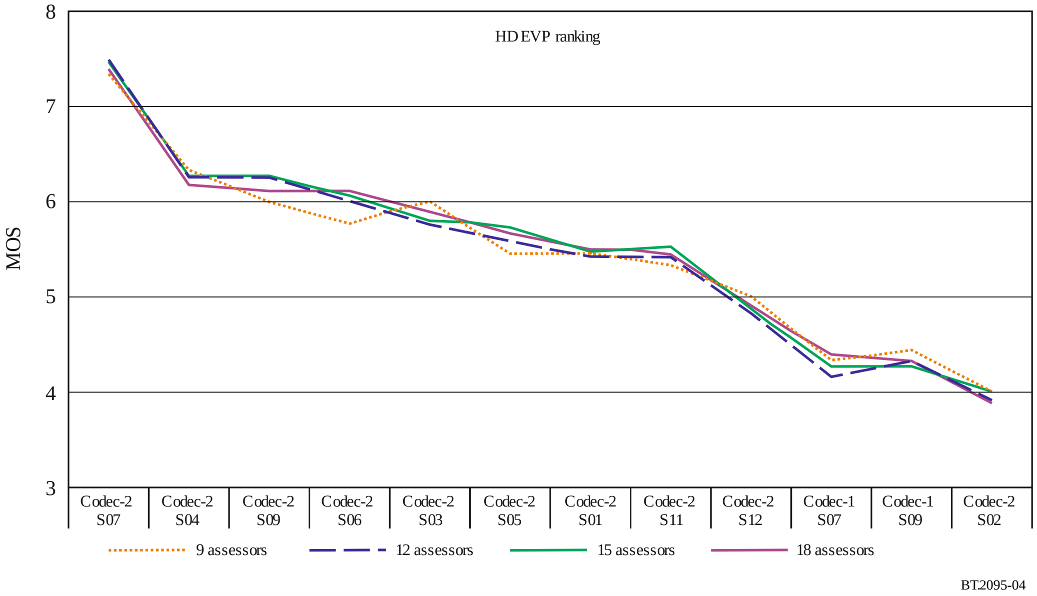
Ranking for the UHD experiment as a function of the number of assessors



BT.2095-03

FIGURE 4

Ranking for the HD experiment as a function of the number of assessors


HD EVP ranking
MOS
Codec-2 S07
Codec-2 S04
Codec-2 S09
Codec-2 S06
Codec-2 S03
Codec-2 S05
Codec-2 S01
Codec-2 S11
Codec-2 S12
Codec-1 S07
Codec-1 S09
Codec-2 S02
9 assessors
12 assessors
15 assessors
18 assessors
BT.2095-04