V Q E G

THE VIDEO QUALITY EXPERTS GROUP

RRNR-TV Group
Test Plan

Version 2.2

Contact: Alex Bourret
Tel: +33 1 55 20 24 28
Fax: +33 1 55 20 24 30
e-mail: alex.bourret @ bt.com

Chulhee Lee
Tel: +82 2 2123 2779
Fax: +82 2 312 4584
e-mail: chulhee @ yonsei.ac.kr

Table of Contents

Editorial History

Version	Date	Nature of the modification
1.0	01/09/2000	Draft version 1, edited by J. Baïna
1.0a	12/14/2000	Initial edit following RR/NR meeting 12-13 December 2000, IRT, Munich.
1.1	03/19/2001	Draft version 1.1, edited by H. R. Myler
1.2	5/10/2001	Draft version 1.2, edited by A.M. Rohaly during VQEG meeting 7-11 May 2001, NTIA, Boulder
1.3	5/25/2001	Draft version 1.3, edited by A.M. Rohaly, incorporating text provided by S. Wolf as agreed upon at Boulder meeting
1.4	26/2/2002	Draft version 1.4, prepared at Briarcliff meeting.
1.4a	6/2/2002	Replaced Sec. 3.3.2 with text written by Jamal and sent to Reflector
1.5	3/12/2004	Edited by Alexander Woerner, incorporating decisions taken at Boulder Meeting January 2004
1.6	5/2/2004	Editorial changes by Alexander Woerner: - Correction of YUV format in 3.2.3 - Included Greg Cermak's description of F-Test in 5.3.6 - CRC suggested modifications (doc. 3/31/04) items #1-6,11 incorporated - Minimum number of HRCs per SRC reduced to six (incl. reference) - Included table of actually available HRC material
1.7	21/6/2004	Edited by Alex Bourret during the Rome meeting in June 2004.
1.8	22/6/2006	Edited by Alex Bourret following the 21/06/2006 audiocall. - HRCs can now be obtained using H264 and VC1 codecs.
1.9	28/9/2006	Edited at Tokyo meeting to update schedule.
2.0	5/29/2007	Edit after Paris meeting, changing details
2.1	10/29/2007	Edit after Ottawa meeting
2.2	Mar 2007	Edit after Kyoto meeting

1. List of Acronyms

ANOVA	ANalysis Of VAriance
ASCII	ANSI Standard Code for Information Interchange
CCIR	Comite Consultatif International des Radiocommunications
CODEC	Coder-Decoder
CRC	Communications Research Center (Canada)
DVB	Digital Video Broadcasting
FR	Full Reference
GOP	Group of Pictures
HRC	Hypothetical Reference Circuit
IRT	Institut für Rundfunktechnik (Germany)
ITU	International Telecommunications Union
MOS	Mean Opinion Score
MOSp	Mean Opinion Score, predicted
MPEG	Motion Pictures Expert Group
NR	No (or Zero) Reference
NTSC	National Television Standard Code (60 Hz TV)
PAL	(50 Hz TV)
PS	Program Segment
PVS	Processed Video Sequence
QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase Shift Keying
RR	Reduced Reference
SMPTE	Society of Motion Picture and Television Engineers
SRC	Source Reference Channel or Circuit
SSCQE	Single Stimulus Continuous Quality Evaluation
VQEG	Video Quality Experts Group
VTR	Video Tape Recorder

2. Introduction

This document defines the procedure for evaluating the performance of objective video quality models submitted to the Video Quality Experts Group (VQEG) RRNR-TV formed from experts of ITU-T Study Groups 9 and ITU-R Study Group 6. It is based on discussions from the following VQEG meetings:

- March 13-17, 2000 in Ottawa, Canada at CRC
- December 11-15, 2000 in Munich, Germany at IRT (ad-hoc RRNR-TV group meeting)
- May 7-11, 2001 in Boulder, CO, USA at NTIA.
- Feb 25-28, 2002 in Briarcliff, NY, USA at Philips Research
- Jan 26-30, 2004 in Boulder, CO, USA at NTIA
- May 7-11, 2007 in Paris at BT
- Sep 10-14, 2007 in Ottawa at CRC
- Mar 3-7, 2008 in Kyoto at NTT

The key goal of this test is to evaluate video quality metrics (VQMs) that emulate ACR and objective amplitude scaling. The evaluation performance tests will be based on the comparison of the ACR-HR MOS and the MOSp predicted by models.

The goal of VQEG RRNR-TV is to evaluate video quality metrics (VQMs). At the end of this test, VQEG will provide the ITU and other standards bodies a final report (as input to the creation of a recommendation) that contains VQM analysis methods and cross-calibration techniques (i.e., a unified framework for interpretation and utilization of the VQMs) and test results for all submitted VQMs. VQEG expects these bodies to use the results together with their application-specific requirements to write recommendations. Where possible, emphasis should be placed on adopting a common VQM for both RR and NR.

The quality range of this test will address secondary distribution television. The objective models will be tested using a set of digital video sequences selected by the VQEG RRNR-TV group. The test sequences will be processed through a number of hypothetical reference circuits (HRCs). The quality predictions of the submitted models will be compared with subjective ratings from human viewers of the test sequences as defined by this Test Plan. The set of sequences will cover both 50 Hz and 60 Hz formats. Several bit rates of reference channel are defined for the model, these being zero (No Reference), 15 Kb/s, 80 Kb/s and 256 Kb/s. Proponents are permitted to submit a model for each of the four bit rate. Model performance will be compared separately with the results from each of the four classes, then compared between them.

3. Division of Labor

This test plan includes certain sub-optimal decisions that reflect the limited resources available to the ILG. The following decisions are pragmatic compromises intended to enable implementation of a sufficient but sub-optimal RRNR-TV test plan; rather than waiting for resources to become available to implement a more ideal RRNR-TV test plan.

- Change from SSCQE to ACR-HRR
- Task ILG only with those tasks that are necessary to ensure independent validation
- ILG design tests prior to model submission
- Proponents run HRCs after model submission

The ILG will perform only the following tasks:

- Coordinate & accept fee payment.
- Choose (identify only) SRC from those provided by proponents and other organizations.
- Choose (identify only) HRCs for the two tests (one 525-line and one 625-line). ILG designs of 525-line and 625-line tests should be finished two weeks prior to model submission.
- Supply secret SRC for each test. If ILG cannot provide secret SRC, then the ILG will identify SRC material that can be purchased by each proponent for a small fee. Such SRC will be identified to proponents and purchased by them after model submission. Alternatively, ILG may purchase directly such SRC, if the fee is small enough.
- Supply secret HRCs for each test, if possible. If ILG cannot supply secret HRCs, then there will be no secret HRCs.
- Create SRC / HRC listing for each subjective test, matching SRC to HRC and identifying which proponent creates which HRC.
- Accept model submissions & perform minimal model validation
- Run 34% of viewers for each subjective test. Preferably, ILG will run all viewers through all RRNR-TV subjective tests.
- Verify data analysis if resources permit

Proponents will perform the remaining tasks, including:

- Edit SRC.
- Run and edit HRCs (after model submission).
- Re-distribute all SRC and HRC to other proponents and ILG as needed.
- Check calibration limits on each PVS.
- Establish standard calibration values for each PVS, if needed.

- Create test tapes (if required).
- Run up to 66% of viewers in each subjective test.
- Perform data analysis.

4. Subjective Evaluation Procedure

4.1. Subjective Test Methodology

The RRNR-TV subjective tests will use the absolute category scale (ACR) [Rec. P.910] for collecting subjective judgments of video samples. ACR is a single-stimulus method in which a processed video segment is presented alone, without being paired with its unprocessed ("reference") version. The present test procedure includes a reference version of each video segment, not as part of a pair, but as a freestanding stimulus for rating like any other. During the data analysis the ACR scores will be subtracted from the corresponding reference scores to obtain a DMOS. This procedure is known as "hidden reference removal."

4.2. Test Design

The test design is a partial design matrix and balanced design to allow analysis of variance (ANOVA). The following presents a brief overview of the test design for each video format (i.e., 525-line, 625-line):

1. A total of 160 PVSs (processed video sequences) will be used, each eight seconds long.
2. The raw, unprocessed reference video sequences (SRCs) are included within the 160 PVSs
3. These sequences are created by processing source sequences (SRCs) using various HRCs (hypothetical reference circuits)
4. The goal of this collection of PVSs is to obtain uniform distribution across the ACR quality scale.

This will produce a total of 23 minutes of ACR video (plus rating time).

4.3. Randomization and Viewing Sessions

Video clips will be presented in a random order, with care taken not to present the same SRC twice in a row, and not to present the same HRC twice in a row.

Subjective testing may be conducted using viewing tapes or any appropriate technology with studio quality playback.

A minimum of two (2) viewer orderings will be used for each test.

4.4. Viewing Conditions

Viewing conditions should comply with those described in International Telecommunications Union Recommendation ITU-R BT.500-10. An example schematic of a viewing room is shown in Figure 1. Specific viewing conditions for subjective assessments in a laboratory environment are:

- Ratio of luminance of inactive screen to peak luminance: ≤ 0.02

- 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
- Display brightness and contrast: set up via PLUGE (see Recommendations ITU-R BT.814 and ITU-R BT.815)
- Maximum observation angle relative to the normal: 30^0
- Ratio of luminance of background behind picture monitor to peak luminance of picture: ≈ 0.15
- Chromaticity of background: D_{65}
- Other room illumination: low
- The monitor to be used in the subjective assessments is a 19 in. (minimum) professional-grade monitor, for example a Sony BVM-20F1U or equivalent.
- The viewing distance of 4H selected by VQEG falls in the range of 4 to 6 H, i.e. four to six times the height of the picture tube, compliant with Recommendation ITU-R BT.500-10.
- Soundtrack will not be included.

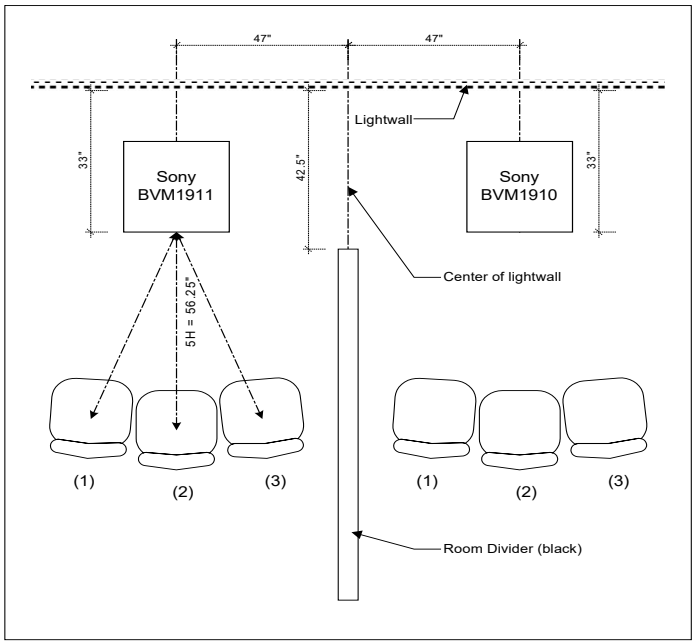

Figure 1. Example of viewing room.

4.5. Instructions to Viewers for Quality Tests

The following text should be the instructions given to subjects. It is noted that the exact text need not to be used.

"In this test, we ask you to evaluate the overall quality of the video material you see. We are interested in your opinion of the video quality of each scene. Please do not base your opinion on the content of the scene or the quality of the acting. Take into account the different aspects of the video quality and form your opinion based upon your total impression of the video quality.

Possible problems in quality include:

- *poor, or inconsistent, reproduction of detail;*
- *poor reproduction of colors, brightness, or depth;*
- *poor reproduction of motion;*
- *imperfections, such as false patterns, blocks, or "snow".*

The test consists of a series of judgment trials. During each trial, a video sequence will be show. In judging the overall quality of the presentation, we ask you to use the judgment scale "excellent", "good", "fair", "poor", and "bad".

Now we will show a short practice session to familiarize you with the test methodology and the kinds of video impairments that may occur. You will be given an opportunity after the practice session to ask any questions that you might have.

[Run practice session, which should include video quality spanning the whole range from worst to best. After the practice session, the test conductor makes sure the subjects understand the instructions and answers any question the subjects might have.]

We will begin the test in a moment.

[Run the session.]

This completes the test. Thank you for participating.

4.6. Viewers

Non-expert viewers should be used. The term non-expert is used in the sense that the occupation of the viewer does not involve television picture quality and they are not experienced assessors. All viewers will be screened prior to participation for the following:

- normal (20/20) visual acuity or corrective glasses (per Snellen test or equivalent)
- normal color vision (per Ishihara test or equivalent)
- sufficient familiarity with language to comprehend instructions and to provide valid responses using semantic judgment terms expressed in that language.

Viable results of at least 24 viewers per test are required, with viewers equally distributed across sequence randomizations. The subjective labs will agree on a common method of screening the data for validity. Consequently, an additional test is necessary if the number of viewers is reduced to less than 24 per lab as a result of the screening.

4.7. Distribution of Viewers Across Labs

Preferably, ILG will run all viewers through both RRNR-TV subjective tests. At least 34% of viewers for each test (525-line and 625-line) must be run by the ILG. The remaining 66% of viewers may be run either by the ILG or by proponent laboratories.

4.8. Subjective Data Format

4.8.1. Results Data Format

Depending on the facility conducting the evaluations, data entries may vary, however the structure of the resulting data should be consistent among laboratories. An ASCII format data file should be produced with certain header information followed by relevant data. Files should conform to ITU-R Recommendation BT 500-10, Annex 3.

In order to preserve the way in which data is captured, one file will be created with the following information:

Test name: tape number:

Vote type: ACR

Lab number:

Number of Viewer:

Number of Votes:

Min vote:

Max vote:

Presentation: Test condition: Program segment:

	Subject Number 1's opinion	Subject Number 2's opinion	Subject Number 3's opinion
	…	…	…
	…	…	…

4.8.2. Viewer Data Format

The purpose of this file is to contain all information pertaining to individual subjects who participate in the evaluation. The structure of the file should be the following:

Lab Number	Subject Number	Month	Day	Year	Age	Gender*
1	1	07	15	2000	32	1
1	2	07	15	2000	25	2

*Gender where 1=Male, 2=Female

4.8.3. Subjective Data Validation

The validity of the subjective test results will be verified by:

1. conducting a repeated measures Analysis of Variance (ANOVA) to examine the main effects of key test variables (source sequence, HRC, etc.),
2. computing means and standard deviations of subjective results from each lab for lab to lab comparisons and
3. computing lab to lab correlation as done for the previous VQEG tests (ref. VQEG Final Report phase 1 and phase 2).

Once verified, overall means and standard deviations of subjective results will be computed to allow comparison with the outputs of objective models (see section 5).

5. Subjective Test Design

This section contains constraints on the design of each subjective test, with regards to SRC, HRC, and PVSs.

5.1. Overview

Prior to model submission, all proponents are encouraged to donate SRC video content, and all proponents will create a list of all HRCs that they can produce for the RRNR-TV test. This document will be submitted to the ILG and other proponents. Proponents will not create example video sequences demonstrating any such HRC.

The ILG will use the lists of proponent HRCs to design two subjective experiments: one containing NTSC/525-line video, and the other containing PAL/625-line video. A total of 160 video sequences will be included in each test, and each video sequence will be 8 seconds long. The raw, unprocessed reference video sequences (SRCs) are included within the 160 PVSs. These test designs will be completed by the ILG prior to model submission.

After model submission, proponents will edit SRC and run HRC as directed by the ILG subjective test plans. If problems occur surrounding an HRC (e.g., requested HRC cannot be created, or a subjective test appears unbalanced), then the problem will be submitted to the ILG for resolution. The ILG will modify the test plan.

5.2. Selection of Source (SRC) Video Sequences

12-second SRC will be used to create HRCs. The first 2s and final 2s of each SRC will then be discarded, such that the viewers and objective models only see the middle 8s of each SRC.

The SRCs (source reference video sequences) shall be selected discretionary by the ILGs taking into account the following considerations:

1. A minimum of twelve 8-seconds SRCs will be used. [Proposed: A minimum of ten 8-second SRC will be used.]
2. A partial matrix will be used (see section 5.3).
3. Video material from the ANSI standard sequences, ITU standard sequences, and the Multimedia test will be used. Proponents and other organizations are encouraged to donate additional source video material.
4. A minimum of 20% new, secret SRCs will preferably be created or added by the ILGs, that no proponent has ever seen before. ILG can use or even shoot in DV25 format, provided the original video quality is acceptable.
5. If necessary, the ILG may include in a test SRC that must be purchased by each proponent from a third party (e.g., film bank) for a small fee.
6. If possible one SRC in each test will contain open source without any copyright protection.

7. Objectionable material such as material with sexual overtones, violence and racial or ethnic stereotypes shall not be included.
8. The scenes taken together should span the entire range of coding complexity (i.e., spatial and temporal) and content typically found in television.
9. At least one scene must fully stress some of the HRCs in the test.
10. No more than 30% SRC shall be from film source or contain film conversions.
11. Downsampled materials from HDTV sources are acceptable. The allowed downsampling procedures will be described in a separate section to be provided.

5.3. Selection of Hypothetical Reference Circuits (HRC)

The Hypothetical Reference Circuits are chosen to be representative of the most common practices in the field of digital TV broadcast networks, for each of 50 or 60 Hz frame rates. Two stages are taken into account (see Figure 2):

- The encoding of original video, multiplexing and subsequent decoding.
- The modulation stage for transmission purposes.

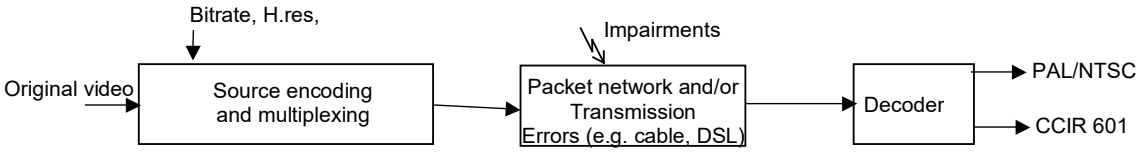

Figure 2. HRC generation chain.

Although this chain appears simple, many configurations are possible. In order to limit the number of HRCs and the overall number of tests to be performed to a practical level, all combinations cannot be tested. Furthermore, the goal of these tests is to discriminate between the proposed models, not to study the impact of specific configurations on the perceived quality. As a consequence, the following directions should be adhered to:

1. Original digital signals are to be used.
2. At the encoding stage, only MPEG2, H.264, or VC1 should be chosen. The proposed range of encoding bit rates is 1 – 6 Mbit/s. Some HRCs must be at 1 Mbit/s (poor quality).
3. At the transmission stage, many configurations are possible
    - Cable network physical layer impairments may be modeled by bit errors of varying lengths. The 64-QAM (e.g. DVB) is a good choice because the noise ranging from an error free output to no output at all at the receiver-decoder is wider than with other modulations (QPSK for example).
    - Video sources may be carried over packet network with different encapsulation schemes (e.g. IP, ATM) and packet loss may occur.

- DSL network physical layer impairments may be modeled by bit errors of varying lengths. If packetized video is carried over DSL network, bit errors rate will translate into packet loss.

4. A minimum of two HRCs, and a maximum of 25% of the processed video sequences shall include transmission and/or packet errors as outlined above. Inclusion of transmission errors for both standards will depend upon the availability of 625-line HRCs with transmission errors. Different types of transmission error HRCs may be selected for the 525-line and 625-line tests
5. A partial matrix design shall be used to create the PVS. This means that not every SRC will be processed using every HRC.
6. A minimum number of eight HRCs plus the original reference sequence shall be used for PVS generation.
7. If possible, a minimum of 25% new, secret HRCs shall be used and selected by the ILGs.
8. No more than 75% of PVSs may be created by any single proponent.

SRC and HRC will be distributed via secure FTP sites or by hard drive. The ILG will not be responsible for distribution costs.

5.4. Video File Format

The test video sequences will be in ITU Recommendation 601-2 4:2:2 component video format as described in SMPTE 125M. This may be in either 525/60 or 625/50 line formats. The temporal ordering of fields F1 and F2 will be described below with the field containing line 1 of (stored) video referred to as the Top-Field.

The following "big YUV" format shall be used for redistribution of video sequences (SRC and PVSs). The final 8s video clips for the subjective tests shall be in "big YUV" format.

Video Data storage:

A LINE: of video consists of 1440 8-bit (Byte) data fields in multiplexed order Cb Y Cr [Y]: Hence there are 720 Y, 360 Cb and 360 Cr Bytes per line of video, 1440 Bytes per line in total:

Multiplex structure: Cb Y Cr Y Cb Y Cr Y Cb Y...

Cb	360 Bytes/line
Cr	360 Bytes/line
Y	720 Bytes/line
Total	1440 bytes/line

A FRAME: of video consists of 486 active lines for 525/60 Hz material and 576 active lines for 625/50 Hz material. Each frame consists of two interlaced Fields, F1 and F2. The temporal ordering of F1 and F2 can be easily confused due to cropping and so it is constrained as follows:

For 525/60 material: F1--the Top-Field-- (containing line 1 of FILE storage) is temporally LATER (than field F2). F1 and F2 are stored interlaced.

For 625/50 material: F1--the Top-Field-- is temporally EARLIER than F2.

The Frame SIZE:

for 525/60 is: 699840 bytes/frame,

for 625/50 is: 829440 bytes/frame.

This video format is also known as YUV Abekas or Quantel.

A SEQUENCE: is a contiguous Byte stream composed of several subsequent frames as described above.

Frame 1, Line 1: Cb Y Cr Y Cb Y Cr... 1440 bytes/line

Frame 1, Line 2: Cb Y Cr Y Cb Y Cr... 1440 bytes/line

Frame 1, Line n: Cb Y Cr Y Cb Y Cr... 1440 bytes/line

Frame 2, Line 1: Cb Y Cr Y Cb Y Cr... 1440 bytes/line

Frame 2, Line 2: Cb Y Cr Y Cb Y Cr... 1440 bytes/line

Frame 2, Line n: Cb Y Cr Y Cb Y Cr... 1440 bytes/line

Frame 3, Line 1: Cb Y Cr Y Cb Y Cr... 1440 bytes/line

Frame 3, Line 2: Cb Y Cr Y Cb Y Cr... 1440 bytes/line

Frame 3, Line n: Cb Y Cr Y Cb Y Cr... 1440 bytes/line

and so on.....

For example, a 10 second length video sequence will have a total Byte count of:

for 525/60 : 300 frames = 209,952,000 Bytes/sequence,

for 625/50 : 250 frames = 207,360,000 Bytes/sequence.

This file format is known also as “concatenated YUV” or “big YUV” format. This file format was used for FR-TV Phase 1 and 2.

Format Summary

	-- 525/60 --	-- 625/50 --
active lines	486	576
frame size (Bytes)	699,840	829,440
fields/sec (Hz)	60	50
Top-Field (F1)	LATER	EARLIER
8s PVS (Bytes)	167,961,600	165,888,000

12s SRC (Bytes) prior to PVS creation)	251,942,400	248,832,000

The total sizes of the sequences in above table are without leading and trailing color bars or gray fields. The total sizes of the sequences in the above table are also without the leading and trailing 2s of extra SRC content, provided for HRC stabilization.

5.5. Calibration Limitations

The video sequences will be Rec. 601 digital video sequences in either 625/50 or 525/60 format. The choice of HRCs and Processing by the ILG will verify that the following limits are not exceeded between Original Source and Processed sequences:

- maximum allowable deviation in Luminance Offset is +/- 10.
- maximum allowable deviation in Luminance Gain is +/- 3%.
- maximum allowable Horizontal Shift is +/- 1 pixels
- maximum allowable Vertical Shift is +/- 1 lines
- maximum allowable Horizontal Cropping is 30 pixels
- maximum allowable Vertical Cropping is 20 lines
- no Vertical or Horizontal Re-scaling is allowed
- Temporal Alignment between SRC and HRC sequences shall be maintained to within +/- 2 video frames
- HRC response to anomalous events (e.g., transmission errors) may result in temporal alignments outside of the above temporal alignment limit. Such PVSs are allowed to temporarily exceed the temporal alignment limits within the following constraints:
    - The first 1s and final 1s of each 8 second video sequence must maintain temporal alignment between SRC and HRC within +/- 2 video frames.
    - At most 25% of any individual PVS's duration may exceed the +/- 2 video frame temporal registration.
    - The SRC and PVS are the same length (8 seconds). Thus, only local temporal variations will be allowed.
- no visible Chroma Differential Timing is allowed
- no visible Picture Jitter is allowed

Proponents will verify adherence of all HRCs to these limits by using at least one, but preferably two software packages, such as those being considered fro ITU standardization in J.cal (TD421 from October 2006 meeting of SG-9). The NTIA, Yonsei, & BT software are suggested. These software checks will be performed in addition to human checking. Proponents will check all PVSs and agree upon calibration correction values.

Proponent software can be used to fix calibration errors in selected video sequences. Preferably, such software should be written in a language that can be easily understood (e.g., Matlab, C++ source code) and posted to the reflector.

The following calibration correction values can (optionally) be passed to models instead of modifying the video sequence:

- horizontal shift (integer value, where positive numbers indicate video has been shifted to the right)
- vertical shift (integer value, where positive numbers indicate video has been shifted down)
- luminance offset (additive value)
- luminance gain (multiplicative factor)

Temporal alignment (i.e., 75% or more of the PVS with +/- 2 frames) should be ensured when editing the 8s PVS from the 12s PVS originally run through the HRC.

VQEG can not guarantee perfect adherence to the calibration limitations, particularly for very degraded HRCs. To prevent inclusion of too many HRC that are nonconforming, proponents analyze video sequences for calibration errors & suggest fixes. The proponents will be given two weeks to perform such verification. If the problem cannot be addressed satisfactorily before the subjective test has been performed, the offending sequence will be replaced. If a sequence is found to not adhere to the calibration limitations after the subjective test has been performed, the offending sequence will not be discarded.

6. Objective Model Submission Guidelines

A reduced reference video quality model consists of two parts. Part one analyzes either the processed video sequence (upstream) or the original reference sequence (downstream) for the purpose of extracting reduced reference data and forwarding it to the second part. The amount of this information determines which class the model belongs to (15, 80, 256 kbit/s).

Part two is typically located at the other end of the transmission line analyzing the "other" video sequence and produces a final video quality estimation by means of using the reference information. With an upstream model the second part analyzes the original video sequence using reference data from the processed video. Part two of a downstream model analyzes the processed video comparing it with reference date from the original sequence. In this scenario a no-referenced (NR) algorithm consists of only part two and doesn't use any reference information (0 kbit/s for the RR channel).

In an effort to limit the amount of variations and in agreement with all proponents attending the VQEG meeting, consensus was achieved to allow only downstream video quality models.

6.1. Input Data Format: SRC Side

The software (model) for the original video side will be given the SRC test sequence in the final file format (i.e., 8s Big-YUV file) and produce a reference data file. The amount of reference information in this data file will be evaluated in order to estimate the bit rate of the reference data and consequently assign the class of the method (NR or RR 15, 80 or 256 kbit/s). NR models will not submit a SRC side program.

The 'SRC side' program will take as input the name of the SRC / PVS pairing file (see section 6.3). The SRC side program will output reduced reference data files, using a naming convention based on the name of the SRC video sequence, the PVS video sequence, the reduced reference bit-rate, or any combination of this information. Example naming conventions include but are not limited to:

<source-file>.dat

<source-file>_<processed_file>.dat

<source_file>_<bit-rate>.dat

6.2. Input & Output Data Format: PVS Side

The software (model) for the processed video side will be given the PVS test sequence in the final file format (i.e., 8s Big-YUV file) and a reference data file that contains the reduced-reference information (see Model Original Video Processing). The processed video side will optionally take as input calibration information. The processed video side will produce an estimated video quality score.

The 'PVS side' program will take as input the SRC / PVS pairing file (see section 6.3) and the results file (see section 6.4). The 'PVS side' is responsible for knowing the naming convention used by the SRC side for reduced reference data files.

6.3. SRC / PVS Pairing File

For each of the two subjective tests (i.e., one 525-line and one 625-line), a SRC / PVS pairings file will be created. This ASCII file will list the pairs of video sequences in the subjective test. Each line of the SRC / PVS pairing file will have the following format:

<source-file> <processed-file>

Where <source-file> is the name of a SRC file and < processed-file> is the name of a PVS created from that SRC. Both video sequences will be in the format specified in section 5.4. File names may include a path.

Some lines of the SRC / PVS pairing file may contain calibration values (see section 5.5) in the following format:

<source-file> <processed-file> <lum_gain> <lum_offset> <horiz_shift> <vert_shift>

where <lum_gain> is a double precision number identifying the luminance gain, <lum_offset> is a double precision number identifying luminance offset, <horiz_shift> is an integer indicating horizontal shift, and <vert_shift> is an integer indicating vertical shift. These values are optional, and may be present on some lines but not others.

For example:

\video\susie_original.yuv \video\susie_h264b320k.yuv

\video\susie_original.yuv \video\susie_h264b64k.yuv

\video\susie_original.yuv \video\susie_h264b64k1plr.yuv

boblec_original.yuv boblec_vc1b2M.yuv 0.95 -10.5 10 2

boblec_original.yuv boblec_vc1b1M.yuv 0.95 -10.5 10 2

6.4. Results File

The output of the PVS side is an ASCII text file. For each PVS, the data file will contain one line, listing (1) the PVS file name, (2) the estimated video quality score, and (3) reduced reference data file (RR models only). The results file for a NR model will be formatted:

<processed-file> <score>

While the results file for an RR model will be formatted:

<processed-file> <score> <rr-file>

Below is an example of the contents of a NR model's results file:

susie_h264b320k.yuv 0.2

susie_h264b64k.yuv 0.4

boblec_vc1b2M.yuv 0.3

boblec_vc1b1M.yuv 0.6

Below is an example of the contents of an RR model's results file:

susie_h264b320k.yuv 0.2 susie.dat

susie_h264b64k.yuv 0.4 susie.dat

boblec_vc1b2M.yuv 0.3 boblec.dat

boblec_vc1b1M.yuv 0.6 boblec.dat

6.5. Submission of Executable Model

Proponents may submit up to 4 models, one for each of the reduced reference information bit rates given in the test plan (i.e., 0, 15 kbit/sec, 80 kbit/sec, 256 kbit/sec).

Each proponent will submit an executable of the model(s) to the Independent Labs Group (ILG). Alternatively proponents may supply object code working on any of the computers of the independent lab(s) or on a machine supplied by the proponent. The ILG will verify that the software produces the same results as the proponent. If discrepancies are found, the independent and proponent laboratories will work together to correct them. If the errors cannot be corrected, then the ILG will review the results and recommend further action.

The executable version of the model must run correctly on a Windows 2000, Windows XP, or Windows Vista workstation.

IMPORTANT: test designs will be sent to proponents when the ILG is given ALL proponent's models. No model will be accepted after test design distribution.

7. Objective Quality Model Evaluation Criteria

The performance of an objective quality model is characterized by three prediction attributes: accuracy, monotonicity and consistency.

The statistical metrics root mean square (rms) error, Pearson correlation, and outlier ratio together characterize the accuracy, monotonicity and consistency of a model's performance. The calculation of each statistical metric is performed along with its 95% confidence intervals. To test for statistically significant differences among the performance of various models, the F-test will be used.

The statistical metrics are calculated using the objective model outputs and the results from viewer subjective rating of the test video clips. The objective model provides a single number (figure of merit) for every tested video clip. The same tested video clips get also a single subjective figure of merit. The subjective figure of merit for a video clip represents the average value of the scores provided by all subjects viewing the video clip.

Objective models cannot be expected to account for (potential) differences in the subjective scores for different viewers or labs. Such differences, if any, will be measured, but will not be used to evaluate a model's performance. "Perfect" performance of a model will be defined so as to exclude the residual variance due to within-viewer, between-viewer, and between-lab effects

The evaluation analysis is based on DMOS scores for the FR and RR models, and on MOS scores for the NR model. Discussion below regarding the DMOS scores should be applied identically to MOS scores. For simplicity, only DMOS scores are mentioned for the rest of the chapter.

The objective quality model evaluation will be performed in three steps. The first step is a monotonic rescaling of the objective data to better match the subjective data. The second calculates the performance metrics for the model and their confidence intervals. The third tests for differences between the performances of different models using the F-test.

7.1. PSNR

PSNR will be calculated and reported if someone volunteers to do the calculation.

7.2. Data Processing

Prior to any data analysis, the ILG will perform an inspection of the subjective test data. Any source sequences presented in the test with a MOS rating of <4 will be identified and the file will be examined. If, in the opinion of the ILG the poor MOS vaues for these source sequences are due to inferior quality then they shall be removed and not included in the subsequent data analysis. This data inspection will be completed prior to proponents submitting their objective data to the ILG.

7.2.1. Calculating DMOS Values

The data analysis will be performed using the difference mean opinion score (DMOS). DMOS values are calculated on a per subject per PVS basis. The appropriate hidden

reference (SRC) is used to calculate the DMOS value for each PVS. DMOS values will be calculated using the following formula:

DMOS = MOS (PVS) – MOS (SRC) + 5

In using this formula, a DMOS of 5 indicates 'Excellent' quality and a DMOS of 1 indicates 'Bad' quality. Any DMOS values greater than 5 (i.e. where the processed sequence is rated better quality than its associated hidden reference sequence) will be considered valid and included in the data analysis.

7.2.2. Mapping to the Subjective Scale

Subjective rating data often are compressed at the ends of the rating scales. It is not reasonable for objective models of video quality to mimic this weakness of subjective data. Therefore, in previous video quality projects VQEG has applied a non-linear mapping step before computing any of the performance metrics. A non-linear mapping function that has been found to perform well empirically is the cubic polynomial given in:

$$DMOSp = ax^3 + bx^2 + cx + d$$

where DMOSp is the predicted DMOS, and the x is the model's computed value (VQR) for a clip-HRC combination. The weightings a, b and c and the constant d are obtained by fitting the function to the data [DMOS, VQR]. This function must be constrained to be monotonic within the range of possible values for our purposes.

This non-linear mapping procedure will be applied to each model's outputs before the evaluation metrics are computed.

Proponents, in addition to the ILG, may compute the coefficients of the mapping functions for their models and submit the coefficients to ILGs. The proponent who submits the coefficients should also submit his mapping tool (executable) to ILGs so that ILGs can use the mapping tool for other models. It is desirable that the proponent also submit the coefficients of the mapping functions for all the other proponents' models. If a proponent chooses not to exercise this option to compute the coefficients of the mapping functions, the ILG will compute the coefficients of the mapping functions. The ILG will use the coefficients of the fitting function that produce the best correlation coefficient provided that it is a monotonic fit.

Any and all mapping algorithms used for the official data analysis must be referenced.

7.2.3. Averaging Process

Primary analysis of model performance will be calculated per processed video sequence. Secondary analysis of model performance may be calculated and reported on (1) averaged data, by averaging all SRC associated with each HRC (DMOSH), and on (2) averaged data, by averaging all HRC associated with each SRC (DMOSS).

7.3. Evaluation Metrics

Once the mapping has been applied to objective data, the three evaluation metrics: root mean square error, Pearson correlation coefficient and outlier ratio are determined. The

calculation of each evaluation metric is performed along with its 95% confidence interval.

7.3.1. Pearson Correlation Coefficient

The Pearson correlation coefficient R (see Equation 2) measures the linear relationship between a model's performance and the subjective data. Its great virtue is that it is on a standard, comprehensible scale of -1 to 1 and it has been used frequently in similar testing.

$$R = \frac{\sum_{i=1}^{N} (Xi - \overline{X}) * (Yi - \overline{Y})}{\sqrt{\sum (Xi - \overline{X})^2} * \sqrt{\sum (Yi - \overline{Y})^2}} \quad (2)$$

Xi denotes the subjective score DMOS and Yi the objective DMOSp one. N represents the total number of video samples considered in the analysis.

It is known [1] that the statistic z (3) is approximately normally distributed and its standard deviation is defined by (4). Equation (3) is called Fisher-z transformation.

$$z = 0.5 \cdot \ln\left(\frac{1+R}{1-R}\right) \quad (3)$$

$$\sigma_z = \sqrt{\frac{1}{N-3}} \quad (4)$$

The 95% confidence interval (CI) for the correlation coefficient is determined using the Gaussian distribution, which characterizes the variable z and it is given by (5)

$$\text{CI} = z \pm 2 * \sigma_z \quad (5)$$

NOTE. NOTE. If the mean is based on less than thirty samples (ie., N < 30), then the Gaussian distribution must be replaced the appropriate Student's t distribution, depending on the specific number of samples in the mean [1].

7.3.2. Root Mean Square Error

The accuracy of the objective metric is evaluated using the root mean square error (rmse) evaluation metric.

The difference between measured and predicted DMOS is defined as the absolute prediction error Perror (6)

$$Perror(i) = DMOS(i) - DMOS_p(i) \quad (6)$$

where the index i denotes the video sample.

The root-mean-square error of the absolute prediction error Perror is calculated with the formula (7)

$$rmse = \sqrt{\left(\frac{1}{N-d}\sum_{N} Perror[i]^2\right)} \tag{7}$$

Where N denotes the number of samples and d the number of degrees of freedom of the mapping function (1).

The root mean square error is approximately characterized by a □^2 (n) [1 [Ed. Note: a page number or equation should be given here], where n represents the degrees of freedom and it is defined by (8)

$$n = N - d \tag{8}$$

where N represents the total number of samples.

Using the □^2 (n) distribution, the 95% confidence interval for the rmse is given by (9) [1]

$$\frac{rmse*\sqrt{N-d}}{\chi^2_{0.95}(N-d)} < rmse < \frac{rmse*\sqrt{N-d}}{\chi^2_{0.05}(N-d)} \tag{9}$$

7.3.3. Outlier Ratio

The consistency attribute of the objective metric is evaluated by the outlier ratio OR which represents number of "outlier-points" to total points N.

$$OR = \frac{TotaNoOutliers}{N} \tag{10}$$

where an outlier is a point for which

$$\left|Perror(i)\right| > 2.07*\sigma(DMOS(i))/\sqrt{NumberOfObservers} \tag{11}$$

where σ(DMOS(i)) represents the standard deviation of the individual scores associated with the video clip i. The individual scores are approximately normally distributed and therefore 2.07 * σ(DMOS(i)) value represents the 95% confidence interval. Thus, 2.07 * σ(DMOS(i)) value represents a good threshold for defining an outlier point. For tests which differ from 24 viewers, the exact value will depend on the number of viewers.

The outlier ratio represents the proportion of outliers in N number of samples. Thus, the binomial distribution could be used to characterize the outlier ratio. The outlier ratio is represented by a distribution of proportions [1] characterized by the mean (12) and standard deviation (13)

$$p = \frac{TotalNoOutliers}{N} \tag{12}$$

$$\sigma_p = \sqrt{\frac{p*(1-p)}{N}} \tag{13}$$

For N>30, the binomial distribution, which characterizes the proportion p, can be approximated with the Gaussian distribution . Therefore, the 95% confidence interval (CI) of the outlier ratio is given by (14)

$$\text{CI} = \pm 2 * \sigma_p \tag{14}$$

NOTE. If less than N<30 samples are used, then the t-Student distribution with t=1.96 [1] can be used instead..

7.4. Statistical Significance of the Results

7.4.1. Significance of the Difference between the Correlation Coefficients

The test is based on the assumption that the normal distribution is a good fit for the video quality scores' populations. The statistical significance test for the difference between the correlation coefficients uses the H0 hypothesis that assumes that there is no significant difference between correlation coefficients. The H1 hypothesis considers that the difference is significant, although not specifying better or worse.

The test uses the Fisher-z transformation (3) [1]. The normally distributed statistic (15) is determined for each comparison and evaluated against the 95% t-Student value for the two–tail test, which is the tabulated value t(0.05) =1.96.

$$Z_N = \frac{z1 - z2 - \mu_{(z1-z2)}}{\sigma_{(z1-z2)}} \tag{15}$$

$$\text{where } \mu_{(z1-z2)} = 0 \tag{16}$$

$$\text{and } \sigma_{(z1-z2)} = \sqrt{\sigma_{z1}^2 + \sigma_{z2}^2} \tag{17}$$

σz1 and σz2 represent the standard deviation of the Fisher-z statistic for each of the compared correlation coefficients. The mean (16) is set to zero due to the H_0 hypothesis and the standard deviation of the difference metric z1-z2 is defined by (17). The standard deviation of the Fisher-z statistic is given by (18):

$$\sigma_z = \sqrt{1/(N-3)} \tag{18}$$

where N represents the total number of samples used for the calculation of each of the two correlation coefficients.

7.4.2. Significance of the Difference between the Root Mean Square Errors

Considering the same assumption that the two populations are normally distributed, the comparison procedure is similarly to the one used for the correlation coefficients. The H0 hypothesis considers that there is no difference between rmse values. The alternative H1 hypothesis is assuming that the lower prediction error value is statistically significantly lower. The statistic defined by (19) has a F-distribution with n1 and n2 degrees of freedom [1].

$$\zeta = \frac{rmse_{\max}}{rmse_{\min}} \tag{19}$$

rmse,max is the highest rmse and rmse,min is the lowest rmse involved in the comparison. The ζ statistic is evaluated against the tabulated value F(0.05, n1, n2) that ensures 95% significance level. The n1 and n2 degrees of freedom are given by N1-1, respectively and N2-1, with N1 and N2 representing the total number of samples for the compared average prediction errors.

7.4.3. Significance of the Difference between the Outlier Ratios

As mentioned in paragraph 8.3.3, the outlier ratio could be described by a binomial distribution of parameters (p, 1-p), where p is defined by (12). In this case P is equivalent with the probability of success of the binomial distribution.

The distribution of differences of proportions from two binomially distributed populations with parameters (p1, 1-p1) and (p2, 1-p2) (where p1 and p2 correspond to the two compared outlier ratios) is approximated by a normal distribution for N1, N2 >30, with the mean:

$$\mu_{(p1-p2)} = \mu(p1) - \mu(p2) = p1 - p2 = 0 \tag{20}$$

and standard deviation:

$$\sigma_{p1-p2} = \sqrt{\frac{\sigma(p1)^2}{N1} + \frac{\sigma(p2)^2}{N2}} \tag{21}$$

The null hypothesis in this case considers that there is no difference between the population parameters p1 and p2, respectively p1=p2. Therefore, the mean (20) is zero and the standard distribution (21) becomes equation (22)

$$\sigma_{p1-p2} = \sqrt{p*(1-p)*(\frac{1}{N1} + \frac{1}{N2})} \tag{22}$$

where N1 and N2 represent the total number of samples of the compared outlier ratios p1 versus p2. The variable p is defined by 23

$$p = \frac{N1*p1 + N2*p2}{N1 + N2} \tag{23}$$

7.5. References for Evaluation Metrics

[1] M. Spiegel, “Theory and problems of statistics”, McGraw Hill, 1998.

8. Calendar and Actions

Action	Due date	Source	Destination
Call for proponents	July, 2004 Complete	VQEG	Proponents
Test plan final version	TBD	VQEG	Public
Submission of new SRC by proponents	July 31, 2007	Proponents	ILG & Proponents
Proponents inform ILG and other Proponents of what video systems (HRCs) they can produce with as much detail as possible (e.g., brand, bit-rates, ways of creating transmission errors)	July 31, 2007	Proponents	ILG & Proponents
Proponents sign all content NDAs (e.g., KBS)	July 31, 2007	Proponents	VQEG Co-Chair
Distribution of sample sequences for model verification	DONE	TBD	ILG & Proponents
ILG design 525-line and 625-line tests.	Jan 31, 2008	ILG	Proponents
ILG identifies SRC material for purchase, if ILG cannot provide secret SRC.	Nov. 30, 2007 [proposed: Oct. 31, 2007]	ILG	Proponents
Model Submission	Feb. 22, 2008 (Baseline)	Proponents	ILG
	TBD	Proponents	ILG
ILG distributes test design to proponents	Mar. 3	ILG	Proponents
ILG distributes secret SRC to proponents	Mar. 3, 2008	ILG	Proponents
Proponents edit SRC	Feb. 19, 2008	Proponents	Proponents & ILG
Proponents run HRCs	Mar. 4, 2008	Proponents	Proponents & ILG
Proponents distribute HRCs to all Proponents & ILG	March 11, 2008	Proponents	Proponents & ILG
ILG distribute secret HRC (if any)	March 11, 2008	ILG	Proponents & ILG
Proponents check calibration values of HRCs	Mar. 18, 2008	Proponents	Proponents & ILG

Proponents edit viewing tapes (if needed)	Mar. 25, 2008	Proponents	Proponents & ILG
Objective data delivered	TBD	Proponents	ILG
Formal subjective test	Apr. 1, 2008	ILG & ILG	
Results data analysis	Apr 8, 2008	TBD	
Final report.	Apr. 15, 2008	TBD	

9. Conclusions

VQEG will deliver a report containing the results of the objective video quality models based on the primary evaluation metrics defined above. The Study Groups involved (ITU-T SG 9, and ITU-R SG 6) will make the final decision(s) on ITU Recommendations.

10. Bibliography

VQEG Phase I final report.

VQEG Phase I Objective Test Plan.

VQEG Phase I Subjective Test Plan.

VQEG FR-TV Phase II Test Plan.

VQEG MM Test Plan

Recommendation ITU-R BT.500-10.

ITU-R Report BT.2020-1.