4th Video Quality Experts Group Meeting
13-17, March, 2000
Communications Research Centre
Ottawa Canada

Meeting Report

Co-Chairs: Philip Corriveau and Arthur Webster

Summary, Conclusions, and Major Decisions

The five days of meetings were extremely productive and three parallel areas of activity have been established (FR-TV, NRRR-TV, NRRR-MM). The next meeting is tentatively scheduled for September 2000 in Rome, Italy.

All of the notes of the ad hoc committee meetings, as well as the presentation on the Final Report, are contained in the following annexes:

Annex I	ILG
Annex II	CPG
Annex III	FR-TV
Annex IV	RRNR-TV
Annex V	RRNR-MM
Annex VI	Normalization

Ad Hoc Committees

All the Ad Hoc committees will have Web pages and e-mail reflectors established to help facilitate communication and the distribution of ideas and decisions.

□	**Independent Lab Group**	**Co-Chair**	**Philip Corriveau**
	ILG	**Co-Chair**	Open
	e-mail reflector: ilsc@m31.dgbt.doc.ca		
□	**Collaboration Phase Group**	**Co-Chair**	**Stephen Wolf**
	CPG	**Co-Chair**	Open
	e-mail reflector: fr-collab@its.bldrdoc.gov		
□	**Full Reference Television**	**Co-Chair**	**Vittorio Baroncini**
	FR-TV	**Co-Chair**	**Alexander Schertz**
	e-mail reflector: fr-tv@fub.it		
□	**Reduced Reference/No Reference Television**	**Co-Chair**	**Jamal Baïna**
	RRNR-TV	**Co-Chair**	Open
	e-mail reflector: rrnr-tv@its.bldrdoc.gov		

- **Reduced Reference/ No Reference Multimedia** **Co-Chair** **Jorge Caviedes**
  RRNR-MM **Co-Chair** Open
  e-mail reflector: rrnr-mm@its.bldrdoc.gov

Vittorio Baroncini announced that he would pursue the establishment of a Task Group under the auspices of the ITU-R JWP 10-11Q in order to facilitate the collaborative effort in the FR-TV area. The Task Group would require any collaborative model to be submitted for VQEG validation.

There is concern that if no improved method can be shown to have significantly higher correlation using the present data (VQEG 1) that a phase II should not be pursued. Pilot studies with expert viewers and short viewing distances and using the VQEG I materials would be helpful. Data from such studies will not be used for evaluation of models in Phase II, but will only be used to do a better design of the Phase II validation tests. It was decided to proceed with design of Phase II at the present meeting.

It was agreed to request that proponents interested in participating in Phase II FR-TV testing should submit results to the reflector of their new objective quality models on the VQEG I dataset. Testing will not proceed unless there are new models that perform better than those submitted for VQEG I.

It was agreed to pursue some pilot studies in order to facilitate the decisions on a Phase II testing procedure. The following pilot studies will be carried out by volunteer labs using the VQEG I video data:

- 3 Picture Heights viewing distance (Arthur Webster, NTIA)
- Expert Viewer test (To be determined)
- JND Subjective Test (Andrew Watson, NASA)

The VQEG I dataset will be made available on the VQEG web site at CRC (http://www.crc.ca/vqeg).

Meeting Minutes:

Monday March 13, 2000

Introduction

The meeting began after brief introductory statements from the President of CRC, Gerry Turcotte, and the Vice-President of Broadcast Research, Metin Akgun.

Agenda Approval

The Agenda was approved.

Discussion and Ratification of Final Report

Ann Marie Rohaly presented the results of the final report and covered some “lessons learned” from the first phase of VQEG testing.

The report concludes that: depending on the metric that is used, there are seven or eight models (out of a total of nine) whose performance is statistically equivalent. The performance of these models is also statistically equivalent to that of peak signal-to-noise ratio (PSNR).

The committee ratified the report without comment or objection. An electronic copy of the final report is available at the CRC ftp site (ftp://ftp.crc.ca/crc/vqeg/Final_Report_March00.doc).

Four metrics were used in the final report:

Metrics relating to Prediction Accuracy of a model:

Metric 1: The Pearson linear correlation coefficient between objective and subjective scores, including a test of significance of the difference.
Metric 2: The Pearson linear correlation coefficient between objective and subjective scores.

Metric relating to Prediction Monotonicity of a model:

Metric 3: Spearman rank order correlation coefficient between objective and subjective scores.

Metric relating to Prediction Consistency of a model:

Metric 4: Outlier Ratio of "outlier-points" to total points.

A valuable result of two years of work is the "lessons learned". The following items were discussed during Ann Marie's Presentation (available on the VQEG website http://www.crc.ca/vqeg, also distributed as document VQEG-2000-10):

Lessons Learned

Test conditions

- ❑ The HRCs may have spanned too many application areas, considering the later objection to subset evaluations. It has been stated by many (but not all) members that the test was not designed for subset evaluation.
- ❑ There were too many sequences with near-threshold impairments; as many as 90% had virtually no observable degradation at the 5H viewing distance.

Test design

- ❑ The subjective experiment was not well suited for data analysis, particularly use of the full data set (i.e. combining all four quadrants of the test).

Data analysis

- ❑ There was no metric to evaluate objective versus lab-to-lab correlation
- ❑ Statistical analysis methods were not adequately specified beforehand
- ❑ A benchmark level of performance was not specified. PSNR (which has become a de facto benchmark for our work) was not considered until after the test was completed
- ❑ No criteria for acceptance was agreed upon prior to the test

Verification procedure

- □ A verification of data analysis results was not established

Ann Marie's Powerpoint presentation will be made available on the VQEG website (www.crc.ca/vqeg).

Update and discussion on release of VQEG test materials and data

A statement previously circulated on the email reflector was discussed and revised as follows.

> VQEG validation subjective test data is placed in the public domain. Video sequences are available for further experiments with restrictions required by the copyright holder. All video sequences have been approved for use in research experiments. Most may not be displayed in any public manner or for any commercial purpose. Some video sequences (such as Mobile and Calendar) will have less or no restrictions. Stripes will be removed from the video sequences prior to any further distribution, however stripes may be used with permission from Tektronix. VQEG objective validation test data may only be used with the proponent's approval. Results of future experiments conducted using the VQEG video sequences and subjective data may be reported and used for research and commercial purposes, however the VQEG final report should be referenced in any published material.

The video source and normalized video sequences will be placed on a CRC ftp site for a period of 6 months to 1 year. Format of the files is defined in the Objective Test Plan and may not be the same as Abekas or any standard format, as it is a color difference data format. The test plans, final report, normalization process description, and other final result files will be made available at a specific location on the VQEG ftp site. All of the copyright issues have been dealt with and the sequences will be made available.

Proponent Presentations

Proponents were encouraged to make presentations on their measurement method and any improvements that have been made, to facilitate possible collaboration in future VQEG tests.

All proponents who gave presentations are asked to provide Philip Corriveau with an electronic copy, these will then be made available to the VQEG body via the website (http://www.crc.ca/vqeg).

KPN – Andries Hekstra

KPN's commercial interests are now primarily in the area of low bit rate applications over ADSL type circuits. The use of single ended methods seems most appropriate. They no longer have much interest in broadcast quality work.

NTIA – Stephen Wolf

This presentation was a resume of the information contained in their recent SPIE paper, "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system." An electronic copy is available at their website (http://www.its.bldrdoc.gov/n3/video/pdf/spie99.pdf). NTIA/ITS has 17 subjectively scored data sets including 1557 data points, 115 unique scenes and 158 unique HRCs. Correlations are in the range of 0.9 and above. Their conclusion is that they have well documented, well tested data and methods available for collaborative efforts. Mr. Wolf proposed that the next test should use an ANOVA for the objective data analysis, including HRC variance (averaging over the scenes) and Scene variance (averaging over the HRCs).

NHK – Yukihiro Nishida

The presentation was titled, “Real-time Picture Quality Assessment System”. Three characteristics are used in the model: spatial-temporal frequency response of human vision, dependency of spatial-temporal response on picture brightness, and dependency of noise perception threshold on picture brightness including chroma. The model applies the difference pictures to a brightness controlled 3-D filter containing spatial and temporal low pass filters. A simplified spatial adaptive filter block diagram shows a three-channel approach. The hardware contains: real-time measurement, built-in test sequences for out of service measurement, automatic compensation of spatial temporal alignment and results logging. Possible improvements include introduction of object-based and cognitive processing. NHK is very interested in collaborating for the next phase of testing.

CPqD – Ricardo Nishihara

The presentation was titled, “Image Evaluation Based on Segmentation”. The original scene is segmented providing context information for the objective measurements. Segmentation is based on edge detection and texture analysis. The result of the measurement is processed by an impairment level estimation producing one value for each frame. A reduced impairment model is also available. A graph was presented, showing only the “Mobile and Calendar” sequence results for each impairment model without correlation values. If predetermined context information is available (out of service), correlation coefficients of 0.9 and higher are obtained, and for unknown scenes the correlation is in the range of 0.85 to 0.88. Results using the VQEG data are said to be in the range of 0.83 to 0.86. The largest improvement was with HRC 1 (which exhibited color problems, particularly with “Susie”). An extended impairment model was discussed along with potential improvements. CPqD is open to collaborative work with other proponents.

Philips Research – Jorge Caviedes

The presentation was titled, “From Impairment Metrics to Single-ended Models”. A simple processing chain consists of encoder, decoder, and post-processing. This provides four measurement points. The measurement is a combination of different components, such as blockiness, ringing and others. Interaction between the components must be considered for good correlation with subjective results. If some components are missing the model will be blind to some artifacts. The main interests of the presenter are in the results of post processing, where reduction of blockiness would produce blurring. Collaborating proponents could contribute various metrics. Measurements can be used for monitoring or control. In the latter case, measurements could be used to improve the post processing block results, based on picture type and complexity. Specification of a multi- dimensional control signal would be required, including information about the various impairments. A single-ended measurement made at different points in the system with communication of the data is essentially a reduced reference double-ended method.

KDD – Takahiro Hamada

The presentation was titled, “Proposal of an Object Based Model Considering its Complexity”. An electronic copy of the presentation is available on the VQEG website. Mean square error is calculated and applied to several filters: pixel based, frame based, block based with noise masking effect and sequence based filtering with motion vector and object segmentation. The object-based filter (F3) is now improved with respect to their previous VQEG contribution. This filter can be used with any of the previous proponent models. Improved results using PSNR as the base were shown. Evaluation of the subjective to objective variations as a function of picture complexity provides a method to improve the results of each proponent’s objective data. It is proposed to use F3 plus PSNR as a benchmark. KDD is very interested in collaboration.

NASA – Andrew Watson

A brief description of the outliers for their method was presented. They were primarily due to HRC1 (multi generation Betacam) and transmission errors. The main presentation was to describe a proposed method for determining JND (just noticeable difference) thresholds, such as that being considered for the IEEE work. The idea is to mix the reference and processed sequences with forced comparison viewing to determine a 1 JND threshold. This is done in an interactive manner using SGI hardware and software. Examples of 32 pair comparisons show convergence to a well-defined threshold. For multiple JND values it appears that a 75% increase in artifact from the previous threshold is the value for the next threshold. The presenter proposes that thresholds are objective, unbiased, repeatable and context free as opposed to the variability in the DSCQS (double stimulus continuous quality scale) method. The proposed method is said to be most useful in the high quality range. NASA is open to the possibility of collaboration.

TDF – Jamal Baïna

The goal of TDF’s method is real-time, continuous, in-service monitoring of transmission system performance. It uses reduced reference picture quality evaluation as a method to determine if there are system problems that are not specifically due to the processing of the picture. They do not expect to evaluate encoder or statistical multiplexer performance. The picture quality measurement methods are of the same nature as other feature extraction methods such as blockiness, spatial activity and temporal activity. They have been shown to have good correlation with the SSCQE (single stimulus continuous quality evaluation) subjective method, for real operating system errors such as uncorrected bit errors, or incorrect protocol implementation. It is not clear if picture quality evaluation to locate such system errors will, or will not, prove useful for more general-purpose applications.

Tuesday March 14, 2000

A change was made to the agenda to allow for a presentation by Alexander Schertz about Subset Analysis.

Subset Analysis – Alexander Schertz

An analysis was performed using subset of the most critical sources from the Phase 1 data. As with various HRC subsets shown in the final report, there is little change in the performance of the various proponents. In fact, the correlations appear slightly lower than those obtained by using all the sources. This analysis only used the 60 Hz high-quality quadrant. The conclusion is that the overall results of the tests are reliable.

Recommendations from ITU-T SG12 – Arthur Webster

The prioritization of SG12's most urgent needs is given below.

1. Low bit rate video quality measurements (16 kb/s to 2 Mb/s, systems with variable frame rate, variable temporal alignment, frame repetition, and with transmission impairments). These methods should include alignment and normalization.
2. Optimization/collaboration phase of current VQEG validation test.
3. In-service measurement systems with particular attention to single-ended and reduced reference measurement systems and measurement of transmission impairments.
4. Standard methods for alignment and normalization for double-ended systems at high bit rates.

Study Group 12 looks forward to continued harmonious collaboration with other ITU Study Groups through VQEG. They would also encourage VQEG to investigate new subjective test methods that could be suitably used for the validation of low bit rate video quality measurements (16 kb/s to 2 Mb/s).

Recommendations from ITU-T SG9 – Andries Hekstra

SG9 asks VQEG to consider the following requirements of an objective video quality measurement system for cable television. Some of these requirements may be topics for ongoing study.

1. Although validation of model applicable to all test conditions that usually occur in the audiovisual scenario is a desirable target, this should not preclude the selection of a model(s) for the limited test conditions of cable operations or of contribution and primary distribution in the broadcasting domain. This implies:
   - H.263 should not be considered.
   - The input to a cable television chain may include noisy and/or compressed signals. Therefore, it is required that perceptual objective measurements of a cable television chain should be able to address these conditions.
   - Transmission errors are unlikely in cable operation, but may occur in general broadcasting, which may require separate consideration.
2. The model(s) should apply on a similar basis to 50 Hz and 60 Hz operation.
3. SG 9 is interested in ranges of quality that span contribution, primary distribution and secondary distribution of television programs. High quality and low quality signal conditions may occur within that range. It would be convenient to be able to use the same measurement equipment for all scenarios, and possible differences in the two applications should be noted.
4. An indication of the computation time and the equipment complexity would be very useful for each case.

At the VQEG meeting it was stated that SG9 is expected to proceed with a "soft" recommendation for the full reference methodology at their May meeting. It would contain various methods in an informative annex, but without sufficient details to implement the algorithms. Pressure is being felt in SG9 to complete a recommendation (standard), although a report might be more appropriate at this time.

Recommendations from ITU-R JWP10-11Q Vittorio Baroncini

1. JWP10-11Q recommends that a collaborative effort be co-ordinated within the framework of the next phase of VQEG testing.
2. For the future work of VQEG, JWP 10-11Q supports the idea to evaluate single-ended, double-ended and reduced reference objective video quality measurement models in parallel. All of these approaches are important and it would be very useful to have timely solutions for each of these methodologies.
3. Since monitoring of the quality of video signals (e.g. over satellite, cable and terrestrial distribution systems) requires compliant methods of objective assessment, all standardizing bodies should agree on one common method for each methodology and application-specific area. Therefore, the reference model described below should be used as a framework for the development and evaluation of objective models covering the above methodologies. This recommended reference model provides a solid foundation for coordinating technical approaches and future recommendations to implement the reduced reference and single-ended methodologies.

Presentation of a "Reference Model" – Jamal Baina

The idea behind the contribution to ITU-R JWP 10-11Q, and subsequently to VQEG, is to provide a model for collaboration on development of objective measurements for various applications. The design and the development of a video quality meter may be based on the general structure of the measurement procedure. Several layers compose this structure.

- ❑ Measurement methodology defines the class or the strategy relative to the application requirement
- ❑ Measurement Method is composed of a set of modules, algorithmic and associated ones, implemented to process inputs such as original signals or processed reference data, and provide output results such as processed reference data, level of impairment or final quality notation,
- ❑ Algorithmic module(s) is the basic block of signal processing functions composing the method. It composes the core of the method from which the final objective qualification is delivered,
- ❑ Associated module(s) is an additional function that aids the algorithmic module(s) in its operation by addressing such issues as dating, synchronization, presentation of data, etc.

Reasons given for this approach are:

- ❑ It provides a conceptual model for method description
- ❑ A large set of solutions is covered
- ❑ Internal functions are specified
- ❑ Basic functions can be compared
- ❑ The model is open to future improvements
- ❑ It can be extended to audio

A block diagram of the reference model measurement method is shown in Annex 1. The reference model is described more completely in the distributed document VQEG-2000-11, “ JWP 10-11Q Report on Objective Quality Assessment in a Digital Environment”. Vittorio Baroncini made a strong request for contributions to the September meeting of JWP 10-11Q.

Discussion on the “reference model”:

Although this approach could be considered architecture, the word “model” is considered appropriate for standards committee discussions. Certainly the system approach for Reduced Reference (RR) and single-ended is more complex than that of the Full Reference (FR) methodology. Much of these system aspects were eliminated in the previous tests by utilizing the normalization process. There is some question regarding how the model will be used for developing a standard. Even if not used for collaboration, the reference model may provide a structure for cooperation between the various ITU study groups.

Presentation of Application Matrix – Philip Corriveau

Three matrix documents were presented. One did not provide for use of the three methodologies in all applications. The other two (distributed documents VQEG-2000-8, VQEG-2000-9) divided the measurement space into six cells, three methodologies by two quality ranges as shown in the table below (insert this?). It was agreed to call single-ended “no reference” (NR). The use of FR for in-service was emphasized, based on work said to be “in progress” by one proponent. An important aspect of document VQEG-2000-9 is that the compressed bit stream may provide useful information for NR systems.

Alternate Approach for VQEG work – Andrew Watson

A “Performance-based Standard” rather than development of a specific model(s) was suggested. This suggestion would overcome the difficulty of co-ordinating collaboration and of selecting one single model for a recommendation. Instead, a standard could specify a level of performance, which could then be correlated with a (new) standard data set. This data set could be private (requiring certification of metrics) or public (to allow anyone to check model operation).

VQEG would provide sequences and subjective scores. Recommendations would define the performance level based on certain statistical analysis methods. Again, the concern about the training of the objective models on the data set was raised. Some members believe methods can be trained to data sets, but will not perform well in general situations. Others believe if the data set covers a large enough universe, training will help ensure good performance in a general situation.

If more than one method meets the criteria for acceptance for standardization there is concern that two objective methods will provide two different answers. This makes facility-to-facility measurement comparisons difficult. The problem is the accuracy that can be expected for conformance to the standard data set.

Report on the Subjective Lab Situation – Philip Corriveau

1. Laura Contin (CSELT) is stepping down as co-chair of the Independent Laboratory Group (ILG). A new co-chair from the ITU-T area would be appropriate.
2. The Australian lab has now closed.

3. There is no commitment from RAI for future subjective testing. However, some other labs are expected to become available. With proper controls, some proponents may also be able to help with subjective assessments.

- CSELT - independent
- CRC - independent
- CCETT - independent
- FUB - independent
- NHK - independent
- IRT - independent
- Tek - proponent
- CPqD - proponent
- Philips - proponent
- ITC - proponent
- NTIA - proponent
- TDF - proponent
- ACREO: - proponent ?
- NASA: - proponent ?
- Sarnoff: - proponent ?

Discussion of Future Work

Six proponents indicated a desire to collaborate on future FR validation tests as a continuation of the previous work (TV quality). These proponents will be known as the Collaboration Phase Group (CPG).

Areas of interest for future VQEG work were indicated, categorized by No reference (NR), Reduced reference (RR) and Full Reference (FR) methodologies, and divided into two sections, TV and Multimedia (MM, i.e. low bit rate applications). Proponent interest (by count, one vote per organization) in the six areas, independent of collaboration, is shown below.

FR-TV, 12 FR-MM, 5

RR-TV, 10 RR-MM, 8

NR-TV, 14 NR-MM, 14

Various combinations of organizing tests for these areas were discussed. Some suggestions were in conflict, for example grouping FR with RR because both are double-ended versus grouping RR with NR because both should use continuous evaluation (and thus, testing against subjective methods such as SSCQE). Some compromise on application definition may allow a combination of subjective tests for more than one area. The idea of selecting subjective testing approaches (an example being DSCQS versus SSCQE) and quality range rather than methodology and quality range became the motivating factor for methodological choice. More than one methodology could be evaluated with respect to the same subjective scores, as they were for the previous phase. For discussion purposes, three ad-hoc groups were formed: FR-TV, RR/NR-TV/MM, and CPG.

Creation of Discussion Groups

FR-TV – Full Reference Television Group

- Chaired by Andrew Watson and Michael Brill

RRNR – Reduced Reference and No Reference Group (TV and Multimedia (MM))

- Chaired by Jamal Baïna and Jorge Caviedes

CPG – Collaborative Phase Group

- Chaired by Stephen Wolf

Wednesday-Friday March 15-17, 2000

For the remainder of the meeting the group was broken into these three discussion groups, and work was done on each area. Provided here is a summary of the work completed and the current groups that have been established under the VQEG umbrella.

Ad-hoc FR-TV

The main goal of this Ad Hoc Committee is to co-ordinate a phase II of the previous VQEG validation tests. A shorter time scale and a less complex test based on the lessons learned from phase I is expected. All of the usual aspects were discussed with agreement on two viewing distances (this was then debated later), one being in the range of 2.5 to 4H, the second being 5 or 6H. (Note: this agreement was later questioned and the issue of viewing distance is still open.) Possible HRCs to be eliminated are multi-generation Betacam, H.263 (video conferencing), transmission errors and analog. Composite would be included, but in digital form to reduce the normalization requirements. There would be no division of quality ranges, however 50 and 60 Hz tests still might need to be separated. Normalization may be considered as part of the proponent method rather than a committee activity. As in the past, there was a proposal to use expert viewers, with much discussion as to exactly what that means. It would appear that more training along with the shorter viewing distance might provide a satisfactory solution.

The biggest concern for some members is that the results of a second phase will achieve similar results as the first phase, with poor correlations and no advantage over PSNR. One suggestion was to take a parallel approach, with validation tests as before along with the development of a “model acceptance criteria” as per the method described previously (See section entitled “Alternate Approach for VQEG Work” presented by Andrew Watson). This would not be a significant committee burden, as NASA would do the alternate testing using sequences from the main validation test.

A method to avoid doing the full phase II tests and achieving results similar to phase I, as well as to make better use of the time at this meeting was proposed. Proponents could provide improved models and generate new results using the present data set. If it is clear that good results will be obtained (i.e. much better correlations) phase II would be developed and implemented. However, there is concern that if such a preliminary test showed no improved method with significantly higher correlation, that phase II with smaller viewing distances will not be implemented. Pilot studies with expert viewers and short viewing distances and using the phase I materials would be helpful. Data from such studies will not be used for evaluation of models in phase II, but only for developing a better design of the phase II validation tests. A decision was reached to proceed with design of phase II at the present meeting.

Discussion: Outline of the FR-TV Subjective Test Plan:

- The purpose of phase II is to produce a more discriminating test than was accomplished in phase I. Pilot studies may be executed in parallel based on available time and resources.
- All new scenes will be used if available.
- The method chosen was DSCQS, with sequences of 5-second duration.
- There will be only one quality range, however there are separate tests for 50 and 60 Hz, 2 labs each, with 20 valid observers per lab. Further screening of observers beyond Rec. 500 will be considered to delete erratic or inappropriate results (e.g. observers who only score on modulo 10 boundaries).
- Viewing distances will be 3H and 6H with the same observers for both distances so the analysis can use combined data.
- Viewers will be non-experts, but with good training to help them understand the expected defects.
- Sequences are to be selected by the ILG to minimize low-defect combinations; use of a sparse matrix is to be considered. They will be supplied with tools to help in this selection.
- There will be 10 sequences each for 50 and 60 Hz with no still-scenes. One or two sequences should have a scene cut. Other possibilities are hand-held camera motion, water, grass, and previously compressed.
- There will be 10 HRCs, with a possible scheme of: 4 MPEG-2 mp@ml bit rates 2-10 Mb/s, 1 422 profile, DV-cam, digital composite, cascading M-JPEG with MPEG, and compression by a non-DCT system.
- To the extent possible, different HRC should be implemented with equipment from different manufacturers.
- There is concern the total SRC/HRC combinations may be too much work for subjective labs and the numbers should be reduced.

Discussion: Normalization of FR-TV sequences

- A tentative plan for normalization is that each proponent would include those adjustments appropriate for their method in their objective score calculation. Maintenance of appropriate levels for subjective viewing the HRC processing would require careful monitoring and adjustment of gain for all three channels. Other systematic maladjustments such as chroma/luminance delay should be observed, and the HRC implementation should be rejected if the problem cannot be corrected.
- Alternately, all normalization could be done by one organization (as in phase I) and the resulting processed sequences used for both objective and subjective tests. A public, agreed normalization algorithm would be available for PSNR processing in the first case or uniform processing in the second case.
- There was no resolution of this issueHowever, the diagram in Annex VI shows the current proposal for Normalization.

For the continuing work of the FR-TV (VQEG Phase II) group, Vittorio Baroncini (FUB) and Alexander Schertz (IRT) are ad-hoc co-chairs. Phil Corriveau is the editor of the subjective test plan and David Fibush is the editor of the objective test plan.

Collaboration Phase Group

Five proponents met to discuss ways to collaborate on a phase II, FR-TV model. Included were NTIA, KDD/Pixelmetrix, CPqD, NHK, and ACREO. Due to concern about legal matters relating to collaboration, this work may have to be done under the auspices of an ITU study group .

The purpose of the collaboration is to develop a model for submission to VQEG, not direct development of a recommendation. One unresolved issue concerns voting by the collaborators on test plan design. The following topics were discussed at the meeting of this ad-hoc group:

Project management
Pick up most important modules
Architecture
IP rights
Possible ways to combine modules
Problems, technical combination and logistics

RR/NR Ad-hoc Committee

The plan for this are is much more difficult to develop, since there are a wide range of applications and system approaches possible. The following topics were reported to the main meeting.

- Bit rates for the reduced reference (may be up to 100% of the available channel)
- Applications primarily related to transmission system monitoring
- Description of output data
- Features to be analyzed, includes but not limited to artifacts
- Possible subjective test methods
- Selection criteria for subjective test method to be used

Discussion: Outline of the RR/NR Subjective Test Plan

- A single stimulus subjective method seems appropriate for the subjective test plan However, there is strong support for a double stimulus method as well. One possibility is Double Stimulus Continuous Quality Evaluation for a short period (e.g. 10 seconds) with 2 scoring samples per period and a summary score at the end.
- The primary objective is to evaluate models by correlating their quality output results with subjective quality evaluation. Models can also provide additional outputs useful for tasks such as control and troubleshooting, for example impairment metrics, classification of defects etc.
- Subjective test data will be collected using SSCQE, producing a continuous score with a resolution minimum of 2 samples per second. Sequence length is at least 1 minute.
- Collected data may be processed and interpreted to create target points and associated variance to be emulated by proposed modules.
- The subjective test for MM needs to be investigated further, e.g. a standard method for subjective testing of MM.

- Applications are video conferencing, streaming video (real-time and non-real time), PC-TV Video conferencing could be point-to-point and multi-point.
- HRCs to consider are: different bit rates, frame rate, packet losses, and format (CIF, etc.).
- Applications where a quality measure is important to the users, such as video-on-demand and content providers, need to be identified. A questionnaire was developed to solicit user requirements for applications for MM due lack of experts in this field still available in the meeting.

Further discussion of RR/NR-TV:

- Maximum bit rate of reference channel needs to be set, for example: zero (NR), 10 kb/s, 56 kb/s, 256 kb/s, etc. Synchronization information is also needed.
- Proponents could contribute subjective assessment services to the ILG if there are two separate tests.
- Sequences would be similar to those for FR-TV, except that longer sequence durations are needed (i.e. greater than 1 minute). It is possible have multiple HRCs within a sequence.
- The number of labs and use of both 50 and 60Hz are still in question.
- Non-expert viewers will be at a viewing distance of 5H.
- HRCs will include transmission errors, and will also include various bit rates, statistical multiplexer, and possibly composite input and non-MPEG encoding. Analog will not be included. One HRC will be the reference material.
- Multiple viewing distances are said to be important for evaluation of perceptual models, however some proponents are more interested system fault finding.

For continuing work on RR/NR-TV, Jamal Baïna (TDF) is the ad-hoc chairman.

For continuing work on RR/NR-MM, Jorge E. Caviedes, (Philips) is the ad-hoc chairman.

Distributed Documents, VQEG2000:

1. Draft Agenda
2. A Perceptual Distortion Metric for Digital Color Video (Swiss Federal Institute of Technology)
3. A Semi-automated Approach for In-line Single-ended Quality Monitoring (ITC)
4. Liaison from SG12
5. Liaison from SG9
6. Liaison from JWP 10-11Q
7. JWP 10-11Q Draft Chairman’s Report
8. Comment on the Proposed Application Matrices (Tektronix)
9. Diagram of Applications of Objective Video Quality Measurements (NHK)
10. Review of final report
11. JWP 10-11Q Report on Objective Quality Assessment in a Digital Environment
12. Real-time Picture Quality Assessment System (NHK)
13. Proposal for an Object Based Model Considering Object Complexity (KDD)

14. Liaison from T1A1
15. T1 Standard on Quality of Service for Business Multimedia Conferencing

Participants

Participants of the 4th VQEG meeting were:

NAME		ORGANIZATION	TELEPHONE	E-MAIL
Ali	Walid	Philips Research, USA	(914) 945-6497	walid.ali@philips.com
Baina	Jamal	TDF	33-3-87-20-75-99	jamal.baina@c2r.tdf.fr
Baroncini	Vittorio	FUB	39-06-54802134	vittorio@fub.it
Bichlmaier	Thomas	Rohde & Schwarz	49-89-4129-3489	Thomas.Bichlmaier@RSD.rsd.de
Blanchfield	Philip	CRC	(613) 998-2761	phil.blanchfield@crc.ca
Blin	Jean Louis	CCETT/CNET	33-2-99-12-41-67	jeanlouis.blin@cnet.francetelecom.fr
Brill	Michael	Sarnoff	(609) 734-3037	mbrill@sarnoff.com
Brunnström	Kjell	ACREO	46-8-6327732	kjell.brunnstrom@acreo.se
Caviedes	Jorge E.	Philips Research, France	33-1-45-10-68-73	jorge.caviedes@philips.com
Corriveau	Philip	CRC	(613) 998-7822	phil.corriveau@crc.ca
Fibush	David	Tektronix	(503) 628-3040	davefi@jps.net
Hamada	Takahiro	KDD Media Will Corporation	81-3-3794-8174	ta-hamada@kdd.co.jp
Hekstra	Andries P.	KPN	31-70-33-25787	a.p.hekstra@research.kpn.com
Libert	John M.	US Department of Commerce/NIST	(301) 975-3828	john.libert@nist.gov
Lubin	Jeffrey	Sarnoff	(609) 734-2678	jlubin@sarnoff.com
Mainguy	André	CRC	(613) 990-4495	andre.mainguy@crc.ca
Myler	Harley R.	University of Central Florida/TeraNex	(407) 823-5098	hrm@engr.ucf.edu
Nishida	Yukihiro	NHK	81-3-5494-2227	ynishida@strl.nhk.or.jp
Nishihara	Ricardo	CPqD	55-19-7056751	nishihar@cpqd.com.br
Rohaly	Ann Marie	Tektronix	(503) 627-3048	ann.m.rohaly@tek.com
Roitman	Peter	US Department of Commerce/NIST	(301) 975-2077	peter.roitman@nist.gov
Schertz	Alexander	IRT	49-89-32399-286	schertz@irt.de
Slater	Norman	Bell Canada	(613) 781-6300	n.slater@bell.ca
Van Dyke-Lewis	Michele	TeraNex	(407) 517-1453	michele.vandyke-lewis@teranex.com
Vincent	André	CRC	(613) 998-2299	andre.vincent@crc.ca
Watson	Andrew B.	NASA	(650) 604-5419	abwatson@mail.arc.nasa.gov
Webster	Arthur A.	US Department of Commerce, NTIA/ITS	(303) 497-3567	webster@its.bldrdoc.gov
Wilson	Danny	Pixelmetrix	65-547-4935	danny@pixelmetrix.com
Wolf	Stephen	US Department of Commerce, NTIA/ITS	(303) 497-3771	steve@its.bldrdoc.gov

Annex I
Independent Lab Group ILG

Tasks

Find a new Co-Chair

Since the responsibility of the ILG is large it is imperative that another Co-Chair be found to help in the logistics and distribution of work for the Independent facilities.

Resource Allocation

Depending on the required resources for each of the parallel activities, facilities will have to be assigned to cover all the needs.

Independent Facilities conducting subjective and objective tests

CSELT
CRC
CCETT
FUB
NHK
IRT

Proponent Facilities able to conduct subjective tests

Some system has to be put into place to ensure that there are controls placed on proponents who conduct subjective tests and there is a need for verification of the results.

Tektronix
CPqD
Philips
ITC
NTIA
TDF

Proponent Facilities that might conduct subjective tests

ACREO
NASA
Sarnoff

Annex II
Collaboration Phase Group CPG

Approximately 5 proponents met to discuss ways to collaborate on a phase II, FR-TV model. Included were NTIA, KDD, CPqD, NHK, and ACREO. Due to concern about legal matters relating to collaboration this work will have to be done under the auspices of an ITU study group (probably ITU-T SG12). The purpose of the work is to develop a model for submission to VQEG, not to directly develop a recommendation. One concern is voting by the collaborators on test plan design. Topics reported discussed at the ad-hoc meeting are:

Project management

IP rights

Pick up most important modules

Possible ways to combine modules

Architecture

Problems, technical combination and logistics

Annex III
Full Reference Television FR-TV

new test

- ***purpose***
  - **more discriminating dataset**
    - viewing distance
    - reduced variance
  - **validate improved models**
- ***method***
  - **DSCQS**
  - **no low and high experiments**
  - **2 labs each for 50/60**
  - **20 qualified observers/lab/Hz**
  - **more subjective data screening**
  - **Pilot Experiments**
    - JND
    - expert viewers
    - test sequence length -- 5 sec, 5 sec repeated twice, 10 sec (to be done w/in the next month)
    - 3H viewing distance (to be done by NTIA w/in 60 days of receipt of test tapes)

viewing distances

- **3H and 5H or 6H**
- **same observers on both, randomized order**

viewers

- **non-expert**
- **better training/practice**

sources

- **sparse matrix**
- **range of criticality/quality**
- how to measure?
  - NTIA method
  - mjpeg method
- access to coded materials

re-use? no, or only if necessary

10 sequences

1. no still
2. color
3. movement
4. brightness
5. contrast
6. synthetics
7. text
8. scene cuts between two similar scenes
9. detailed scene, motion
10. water flowing
11. sports
12. grass
13. compressed reference 9Mb

50 & 60 (20 sequences)

- **commercially available source material**

hrcs

- **don’t vary hrc over src**
- **10 HRCs**
- **different vendor for each mpeg bitrate**
- **(4) mpeg 2 mp@ml x 4 bitrates 2-10 MBps**
- **(1) 422 profile**
- **Dvcam (might be transparent)**
- **(1) DV**
- **(1) digital composite**
- **cascading**
- **mjpeg + mpeg**
- **(1) wavelet**
- **(1) sorensen**
- **no transmission errors**
- **no analog**
- **include normalization in models?**

normalization

- **public algorithm, proponent responsibility**
- **alignment**
- **chroma/gain**

Objective test plan

- **VQEG1 Method**
- **secrecy**
- **metrics**
  - chi square
  - correlation

Schedule

- **1year**
- **separate schedules for subjective and objective testing**
- **ILG meeting, summer 2000**

Tasks

- **Identify resources (w/in one month)**
- **source selection**
- **hrc selection**
- **coding**
- **tape editing**
- **normalization**
- **distribution**
  - coded sequences
  - edited tapes
- **subjective testing**
- **objective testing**
- **statistical analysis**

Open issues

- **Number of HRCs and SRCs**
- **Test sequence length**
- **Viewing distance**
- **performance based standard**
  - Alexander’s idea: two step process
  - performance criterion, absolute (Arthur)

Annex IV
Reduced Reference No Reference Television RRNR-TV

Objective methods

Max bitrate of reference channel needs to be set – zero has to be one (NR case), 10 kb/s, 56 kb/s, 256 kb/s (question whether synchronization info must be carried in this channel too)Ground rules –can objective methods take bitstream as input?? Can proponents act as subjective testing facilities??

Subjective test

Sources – use same content types as FR TV, but longer sequences
no stills
color stressing
movement
pan/random camera shake
brightness– range across sequence set
contrast – range across sequence set
synthetic imagery
text
film noise
detailed scene motion – sports
scene cuts
compressed reference
grass
rocks
leaves
method

SSCQE

Sequence length ≥ 1 min
Issue whether or not to apply each HRC to entire sequence length or change HRCs w/in each sequence
# labs ??
50 Hz and 60 Hz tests ??
20 observers/lab
viewing distance 5H
non-expert viewers

hrcs

transmission errors
w/ & w/o encoder included in chain
reference material (4:2:2 profile) should be one HRC
no analog
mpeg 2 mp@ml x 4 bitrates 2-10 Mb/s

statmux
wavelet, other non-MPEG encoders
composite video

Schedule

Annex V
Reduced Reference No Reference Multimedia RRNR-MM

Tasks

Application requirements analysis and design of subjective test
Select/produce material (SRCs, HRCs)
Execution of subjective test
Objective test requirements
Execution of objective test
Analysis and report of results

Milestones

Requirements Analysis and request ofr participation 6-7 months
draft questionnaire ready for May SG12 meeting
Subjective test completed 10-12 months
Objective test completed 13-15 months
Final report completed 15-18 months

Tasks and Time Estimates

1. **Prepare and carry out a survey** to identify, among others: Applications, tasks(e.g. monitoring, control), image formats, bitrates, bandwidth available for reference channel, outputs of interest, expected impact of an objective quality measure (per application). **3 months**
2. **Select SRCs and HRCs and plan s subjective test.** Include verification of appropriateness of subjective test methodology (SSCQE) ad anticipate scenarios to deal with high variance, overall model performance, playback for subjective and objective testing (format constraints, pseudo-reference sequences, repeatability, other experimental design/set-up issues). **3-4 months.**
3. **Conduct subjective testing 3-4 months.**
4. **Conduct objective testing 3 months.**
5. **Evaluation and recommendation 3-4 months**

Total time estimate is 15-18 months.

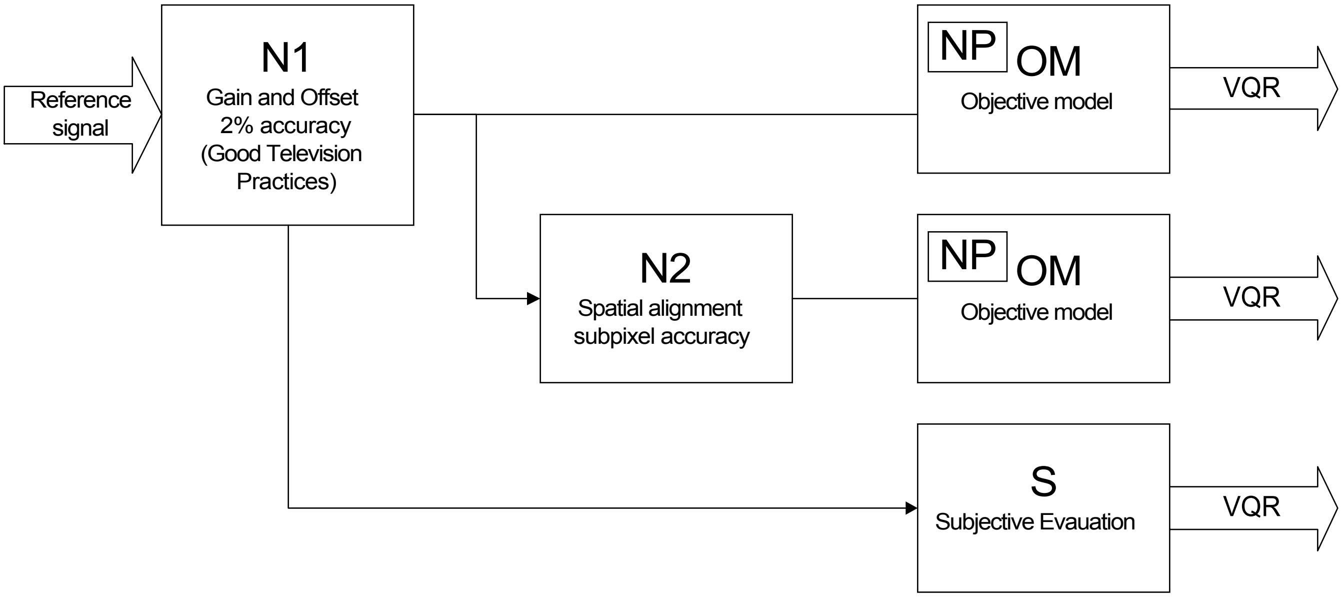
Annex VI
Normalization

NORMALIZATION for FR-TV

DEFINITION

Part 1: Gain and offset (black level) adjustment to match source with 2% accuracy according to good TV operational practice and temporal alignment.

Part 2: Spatial alignment to sub-pixel accuracy.



N1 – Normalization Part 1: Adjustment of Gain and Offset to within 2% accuracy (Good Television Practices)

N2 – Normalization Part 2: Spatial alignment to sub-pixel accuracy

NP – Proponent Normalization: any normalization procedures contained within the proponent algorithm.

SUBJECTIVE ASSESSMENT

Normalization Part 1 will be performed to provide observers with a picture free of systematic adjustment, fault or design problems.

OBJECTIVE MEASUREMENTS

Sequences normalized according to Part 1 only and to both Parts 1 and 2 will be made available to all proponents. (Note: If resources are not available to perform Part 2, only Part 1 will be performed.)

PSNR requires accurate normalization of both parts. Normalization required for PSNR will be implemented by one organization using a defined method to a level of accuracy specified by VQEG.

REFERENCE SIGNALS

Color bars will be added to the start and end of the source sequence tapes primarily as an aid to Part 1 normalization however they may provide some help for Part 2. They are not part of the actual sequences as the stripes were for Phase I VQEG evaluation testing. The color bars are HRC processed and recorded in the same manner as the video. Other test signals may be included (TBD).