

DRAFT MEETING RECORD  
Video Compression Measurements Subcommittee G-2.1.6  
Audio Video Techniques Committee G-2.1

Broadcast Technology Society  
Institute of Electrical and Electronics Engineers  
Fifteenth Meeting

Ramada Inn Silicon Valley/Sunnyvale  
1217 Wildwood Avenue  
Sunnyvale, CA 94089-2701  
January 23, 2000

Item 1 - Welcome and Introduction by Interim Chairman, of IEEE G-2.1.6.

The meeting was called to order at 11:36 AM by Interim Chairman Alan Godber.

Item 2 - Approval of Draft Agenda

After merging items 6 and 7, the agenda was approved.

Item 3 - Review and Approval of Minutes of the Previous Meeting #13, July 27th, 1999

The Minutes from Meeting 14 were accepted as prepared.

Item 4 - Matters Arising from the Minutes

There were no matters arising from the minutes.

Item 5 - Update Report of ITU Video Quality Experts Group (VQEG) re results of the tests conducted - Arthur Webster, David Fibush, John Libert, Al Morton and other participants.

*The Final Report From The Video Quality Experts Group On The Validation of Objective Models of Video Quality Assessment, Draft 4c, VQEG, December 1, (IEEE Doc. G-2.1.6/107) was discussed.*

Except for a correction in metric one, there were no substantial changes since the preliminary report was released. The conclusions didn't change. Nine of the models were statistically equivalent. A map showing Pearson correlation had all models except P6 clustered at the top. Spearman rank order showed the same clustering. The draft final report is available on the VQEG web site.

The next VQEG meeting will be in March at CRC in Ottawa, Canada. One issue to resolve at the meeting will be determining what conclusions people are looking for. The analysis would not support comparisons from model to model. John Libert reported that his experiments showed that even if the models had performed better, we would still have trouble separating the top six models.

### *5.1 Further Discussion and Recommendations from the Subcommittee to the next meeting of VQEG*

It was noted that lab to lab correlation put a limit on the measurements. It was suggested that rather than try to improve performance of the metrics used on the VQEG data, we should try to increase their generality in a way that the PSNR measurement isn't general. One way would be to include a variable such as viewing distance. Another observation was that future tests should use less facilities and be more focused. There were comments that too many conditions were too close to discriminate. Viewing distance was an issue, with one comment that we can't avoid having multiple viewing distances and resolutions. If this isn't considered, the value of vision models becomes less important. There seemed to be agreement that we had to find factors that will differentiate one method from another. In defense of VQEG, some compromises were necessary to get the work done and those included using one resolution and one viewing distance.

The discussion moved to the question of whether the VQEG tests were a success. This, of course, depends on how success is defined. If the definition is correlation, the problem is the PSNR method looked as good as it did. The set of tests did not pick out the particular elements in the other models. If subsets of the results were removed, in general the results always got worse. There were some cases where correlation improved when some HRCs were removed. In one case, where one model had better correlation than the others, it turned out to be a case where a 60 Hz model was used on 50 Hz just to get a number. It performed better than other models designed for 50 Hz. There was a comment that a full HRC is a crude scope. In the H.263 HRC, if sequences that have high motion are taken out, correlation improves. However, if the H.263 HRC is entirely removed, it hurts results.

Even if everyone improves their models for the next round of testing, they may still be equal. For people without a specific model in the tests, they may consider it a success.

Future VQEG work and participation by G-2.1.6 was discussed. There was a comment that we need to consider the range of applications the methods will be used for, ranging from satellite distribution for production to distribution to viewers at home. Viewing distance is a factor. Faults can be seen at a viewing distance of 1.5H that aren't noticed at 5H. The JND work would be useful in testing this data. Could the VQEG tests be repeated with more variables? While this could be done as a way to validate the models. It was agreed that rerunning the same study wouldn't work. A study the magnitude of the VQEG tests won't happen again. The idea of a test with fewer labs and more subjects, or more subjects under different conditions was suggested. We would have to devise a testing plan the subjective labs will buy into and will take a vested interest in.

If we spell out the G-2.1.6 JND approach well enough, will VQEG buy into it? In a discussion on the merits of absolute measurements, there was a comment that not all the VQEG members buy into JND. In response to that comment, it was noted that the experiments aren't that different. We quantify results in JND, others define them in DSQCS. There was no agreement, however, that both efforts would have the same end result. There was a suggestion that we first need to get a handle on what perceptual phenomenon we are modeling, then see what components are not accounted for.

It was decided to defer making recommendations under after the discussion on JND.

Item 6 - Report of Task Force on "Defining A Unit of Measure & a Means of Calibration for Video Impairment", Chair, Leon Stanger.

This item was merged with Item 7.

Item 7 - Report of Task Force on "Selecting Test Material and Test Labs for a Unit of Measure and a Means of Calibration for Video Impairment", Chair, John Libert.

John Libert reported the task force was trying to lay out how they will implement the JND study to insure a repeatable measure of quality that is sufficient to generate a set of calibrated test materials. The method of generating calibrated test materials has to be proven against a subjective measure of video quality.

#### *7.1 Further discussion and action*

In the discussion, there was general agreement that a pilot test should be done first to see if the procedures selected work. A smaller set of studies would make it easier to find someone willing to do the tests. It was

decided to start with one sequence and one HRC from VQEG. Phil Corriveau said that CRC could run the pilot at relatively low cost by using people from CRC that don't work in the vision science area as subjects.

The question of whether untrained, trained or expert viewers should be used was discussed. It was agreed that we wanted viewers that were more sensitive to image degradations than untrained viewers. It was decided that viewers would be trained by showing them the sequence without degradation and the sequence with the maximum amount of degradation. This should help optimize, within reason, the chance people will see visual degradations that are there, not other distractions. It was also suggested we should train them to look for a single type of impairment. However, it was pointed out that vision models are not able to look at only one type of impairment.

The procedures for the experiment were considered. It was suggested that we create a weighted combination of the end-point and the HRC, to create a set of sequences across a range of degradations. The sequences would be displayed on a screen and the viewer would input a number to indicate the difference between the two sequences. An adaptive procedure was proposed to try to avoid displaying two sequences with large differences between them. The data collected could then be massaged to allow a graph to be drawn that had the weight on one axis and the JND on the other axis. This concept was discussed. One problem is that repeated observations are needed for this method and the VQEG sequences may be too long to allow for a large number of observations. It was pointed out that Thurstone's Law of Comparative Judgements would allow you to get the scale and the distance between, for example, ten steps without pairing everything sequence with every other sequence.

It was agreed to use an adaptive staircase method for the initial experiment to see how it would work and to compare results with the data from the VQEG tests. The Amnon Silverstein method of paired comparisons was suggested. Steps will be generated by taking a VQEG score and dividing the weighted sum into steps containing an equal number of DMOS points. If twenty steps were used, the Silverstein approach would require about 80 pairs or a total number of about 100 trials for one HRC/scene. More information is available in the Silverstein paper at <http://www.best.com/~amnon/Homepage/Research/Papers/TreePaper/TreeMethod.html>.

After discussion, HRC9, SRC 19, "football" was selected for the first tests, because it was one of the scenes with the widest picture quality range.

The first trials are to be done under Rec. 500 conditions, with viewing distances of 5H and 3H.

The white board outline of the experiment is shown in Appendix C.

Under the experiment plan outlined here, preliminary tests were scheduled to start in 6 weeks, with the goal of completing the test by March 12 and presenting the results of the pilot test at the G-2.1.6 subcommittee meeting prior to the VQEG meeting in March. The Task Force was asked to present a progress report on the experiment in four weeks (February 20, 2000).

Item 8 - Further Discussion of Compression Measurement Methodologies.

There was no discussion under this item.

Item 9 - Any Other Business.

There was no other business.

Item 10 - Date(s) of Future Meeting(s).

The next meeting was tentatively planned for the Sunday preceding the March VQEG meeting at the CRC, pending the results of a progress report from Dr. Watson on the plan for the experiment and preliminary tests outlined in Item 7. If, by February 20, it appears the results will not be available to report before the VQEG meeting in March, the next meeting will be scheduled with the T1A1 meetings in Boulder in April. There was a motion to adjourn, which was seconded. The meeting was adjourned at 5:22 PM.

Submitted by:  
H. Douglas Lung  
Secretary

APPENDIX "A"

List of Documents Distributed

23 January 2000

*Draft Agenda - IEEE Compression and Processing Subcommittee G-2.1.6, Fifteenth Meeting, Sunday, January 23, 2000, Alan Godber, Chairman. ([216m15an.html](#))*

*Draft Meeting Record, G-2.1.6, Compression and Processing Subcommittee, Fourteenth Meeting, November 1, 1999, Fort Lauderdale, FL, Doug Lung, Secretary, [IEEE Doc. G-2.1.6/106](#), January 20, 2000.*

*Final Report From The Video Quality Experts Group On The Validation of Objective Models of Video Quality Assessment, Draft 4c, VQEG, December 1, IEEE Doc. G-2.1.6/107, January 20, 2000.*

*Quantifying Perceptual Image Quality, D. Amnon Silverstein, Joyce E. Farrell, Imaging Technology Department, Hewlett Packard Laboratories, March 1, 1998, IEEE Doc. G-2.1.6/108, January 23, 2000.*

APPENDIX "B"  
ATTENDANCE RECORD  
23 January 2000

<b>Name</b>	<b>Affiliation</b>	<b>Telephone</b>	<b>Fax</b>	<b>E-mail</b>
Chairman: Alan Godber	Consultant	(732) 846-4476	(732) 846-4476	<a href="mailto:agodber@idt.net">agodber@idt.net</a>
Secretary: Doug Lung	Telemundo	(305) 884-9664		<a href="mailto:dlung@transmitter.com">dlung@transmitter.com</a>
Philip Corriveau	CRC	(613) 998-7822	(613) 550-6488	<a href="mailto:phil.corriveau@crc.ca">phil.corriveau@crc.ca</a>
David Fibush	Tektronix	(503) 628-3040	(503) 627-4486	<a href="mailto:davef@exgate.tek.com">davef@exgate.tek.com</a>
John Libert	NIST	(301) 975-3828		<a href="mailto:john.libert@nist.gov">john.libert@nist.gov</a>
Jeff Lubin	Sarnoff Corp.	(609) 734-2678	(609) 734-2662	<a href="mailto:jlubin@sarnoff.com">jlubin@sarnoff.com</a>
Rick Redford	Consultant	(917) 441-0105	(212) 664-5222	<a href="mailto:rick.redford@juno.com">rick.redford@juno.com</a>
Ann Marie Rohaly	Tektronix	(503) 617-3048	(503) 627-5177	<a href="mailto:ann.marie.rohaly@tek.com">ann.marie.rohaly@tek.com</a>
Leon Stanger	DirecTV	(310) 726-4676		<a href="mailto:LStanger@compuserve.com">LStanger@compuserve.com</a>
Andrew Watson	NASA	(650) 604-5419	(650) 604-0255	<a href="mailto:abwatson@mail.arc.nasa.gov">abwatson@mail.arc.nasa.gov</a>

APPENDIX "C"

White Board Outline of Proposed Experiment

23 January 2000

Test Material - 1 SRC - HRC from VQEG to start.

MPEG - Low quality HRC 9, SRC 19 (football)

Subjects - Trained (not expert)

Training - test trials show range - anchoring scale

Threshold Proc.

Silverstein -

- DMOS Point to space wt. sum.

Equal # of DMOS points

# Trials prop.  $N * \log_2 * N$   $N = \#wts.$

(guidance from paper)

Rec. 500 Conditions

Viewing Distance - 5H

3H