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

Twentieth Meeting

NTIA/ITS, U.S. Department of Commerce

325 Broadway

Boulder, CO 80303

May 11, 2001

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

Interim Chairman Alan Godber called the meeting to order at 11:37 AM. Attendees introduced themselves. See Appendix B for a list of attendees.

Item 2 - Approval of Draft Agenda

A report on T1A1.1 activity on objective video quality measurements was added before the VQEG report as Item 5A.

Item 3 - Review and Approval of Minutes of the Previous Meeting #19, January 29, 2001

Review of the minutes from Meeting #19 was delayed.

Item 4 - Matters Arising from the Minutes

There were no matters arising from the Minutes.

Item 5A - Report on activity in T1A11 on Technical Reports Describing Full Reference TV Picture Quality Objective Metrics, David Fibush

David Fibush reported on the Technical Reports from committee T1A1.1 covering objective metrics for TV picture quality. The technical reports provide a framework in which a number of metrics can be evaluated and a common objective scale from 0 to 1, to convert from one metric to another. On the scale, zero represents the least degradation and one the most degradation. The current Technical Reports are restricted to Full Reference (FR) methods.

There are four technical reports. The first describes video quality metrics and the methods to calculate accuracy and convert results to a standard scale. This report also contains the procedures and forms for adding methods to the document. The second technical report covers normalization methods for such parameters as gain, DC level and spatial alignment. There are annexes for different normalization methods, although the Tektronix stripe method is the only one currently listed. The third technical report documents the PSNR method for determining objective picture quality. This is the same method described in the ANSI document published years ago, but it now includes documentation of accuracy and scaling based on the VQEG data. It was noted the range of measurement is limited by the VQEG data. The fourth technical report contains full documentation of the Sarnoff PQR method. It includes an equation that gives a result of zero if there is no degradation and one if it is bad. It was pointed out that some VQEG data sets don't fit the equation, but since these technical reports only apply to the U.S., only 525 line data was used. HRCs 15 and 16, the H263 sets, were not used.

5.1 Further Discussion

In response to a question on how a document describing another metric could be added to the technical reports, David Fibush referred to the procedure in the first Technical Report. Fill out the form in that document and submit it with results using the VQEG data. The goal is that eventually we will find a method that users really like and that will be chosen for use in one or two standards worldwide. The open approach taken in the T1A1.1 Technical Report is designed to give a clear analysis of the models and the data, not to pick a winner.

There was a question whether the methods in Technical Report 1 could be used with subjective measurements other than DSCQS, since it can work with data other than that collected by VQEG. It was though that it could. The three methods used in Technical Report 1 for determining accuracy were described: calculating resolving power against a common scale, simple standard deviation, and classification of errors.

A copy of the technical reports and additional information is available in the T1A1 listing in the T1 file index at http://www.t1.org/html/fileindx.htm.

Item 5 - Update Report of ITU Video Quality Experts Group (VQEG) re ongoing testing plans and results of May meeting in Boulder, CO at NTIA - Arthur Webster, David Fibush, Phil Corriveau, Ann Marie Rohaly and other participants

David Fibush reported the new chairman of ITU-R Working Party 6Q, Vittorio Baroncini, is going to work closely with ITU-T on video quality measurement. The way ITU-T measures video quality through a telecommunications or cable system must be consistent with the way ITU-R measures exactly the same video going through a broadcast television system.

The VQEG meeting concentrated on the reduced reference (RR)/no reference (NR) test plan, with half a day devoted to the full reference (FR) tests. The RR/NR work has priority. The FR tests will use the RR/NR data and some of the HRCs. RR/NR use is focused on network monitoring of the system, so it looks at the effects of MPEG bit rates, concatenation, statistical multiplexing and errors. The FR tests have been restricted to error-free circuits.

An ad-hoc group has been formed to prescreen sources and HRCs to try to generate a reasonable spread of impairment for the subjective testing.

Both the RR/NR and FR test plans are near completion. Some types of test material, however, are still needed. VQEG would like to get a few minutes of sports sequences from the networks.

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

It was noted that there is one viewing distance for the tests - 4 picture heights.

There was a comment that the resolving power of the models is not good and it was questioned if this was due to the subjective data. In these experiments, the mean value of the score is considered important, although there is a wide variation in what 100 viewers see. Individual values will vary significantly and this isn't likely

to change, even with a large variety of tests. It was suggested that we rate objective models by how many viewers it is equivalent to Most are one or two now. An objective measurement will give you the same result for the same test every time it is run, unless it uses random sampling. Subjective tests will vary. In the RR test, proponents must state the repeatability of their method.

Selection of test sequences was discussed. It was argued that sequences like sports have too many elements and add noise to the results. The counter argument was that simple scenes don't equate to real-world video sources. You can improve the accuracy of measurements by narrowing the type of impairments. However, once you have a model that works well on these little pieces you have the problem of trying to put the individual results together to handle complex scenes.

An experiment planned by Vittorio Baroncini may help determine if the variability in subjective testing can be reduced by using expert observers and repeated observations. The experiment will use the same materials used in the FR test. The standard tests will be run, using five or six expert observers. For some sequences, the observers will be able to keep comparing the HRCs and sources will make a group decision as to the HRC rankings for each source. The results from these subjective tests will be compared with the results from the metrics.

When the use of data embedded in the source material in one of the RR/NR methods was mentioned, there was concern that the technique has to be defined so that is doesn't interfere with encoding efficiency. It was noted that people often ruin 100 kb of bandwidth to get 1 kb of useful data inserted. There was another concern that from an artistic content standpoint, people don't want their video data modified.

VQEG tests plans will be available soon on the VQEG reflector. Setting the HRCs and sources is the main problem remaining.

Item 6 - Selecting Test Method, Test Material and Test Labs for a Unit of Measure (JND) & a Means of Calibration for Video Impairment, Chair, Leon Stanger

Andrew Watson presented the proposed draft standard, *JND Visual Impairment Scale Measurement Method (JNDVIS).* (The title was changed to *Draft Standard for the Measurement of Visual Impairments in Digital Video Using a JND Scale* during the meeting. See IEEE Doc. G-2.16/124.) The history of G-2.1.6 and the development of the draft standard were discussed. The measurement technique, discussed in previous meetings, was briefly outlined. Changes to the document included putting in into the IEEE standard form and placing details about the method in an appendix

6.1 Review of first Draft of Standard, prepared by Dr. Andrew B. Watson

Dr. Watson described the apparatus used in the tests. A baseline configuration used for the test is included in the standard. The baseline configuration uses non-expert viewers.

The procedure uses a linear combination (arithmetic) of degraded and non-degraded signals. This allows continuous variation of degradation. Twenty-one different levels of degradation are used in the baseline configuration. The degraded video sequences are created before the start of a session. Two pairs of sequences are presented and the user has an unlimited amount of time to respond.

The trial includes an easy pair. This needs to be specified in the procedure.

Discussion

The use of expert viewers was discussed when reviewing Section 8.3. Phil Corriveau found expert observers gave results with a tighter spread. It was suggested we might want to include expert viewers in the trial run of our tests. ATSC used about sixty experts in their testing. Another comment was that because an expert knows where to look, they are less likely to miss an artifact. This would help eliminate extraneous data.

Dr. Watson tracked results of observations from both expert and non-expert viewers and became convinced there was no big difference between the two groups. It was felt “trained observers” were good enough, but they

had to be told what to look for. It was agreed that standardized instructions are important. (Discussion about Section 8.4.8)

There was a question about the resolving power of the method. Is there a confidence level attached to each JND? In the NASA tests, Dr. Watson had only three observers per condition versus up to eighty for the VQEG subjective tests. For three viewers per situation, the variation was 30 percent. He did not claim the metric's variability is less than the VQEG data, however, at the end of the block of sequences it does give a more precise estimation of JND, particularly in cases with a smaller number of JNDs. (Discussion about Section 8.4.9 and 8.4.10)

In Section 8.5, the parameter "Replications/condition" was changed to "observers per condition".

While reviewing Section 11 in the Appendix, it was noted that a simple computer could handle the calculations. In fitting the function to the data, maximum likelihood estimation, not LMS is used. LMS doesn't apply here because the data collected only describes whether the observer was correct on a particular pair, not the observer's estimate of the difference between the scenes. Simulations showed accuracy continues to improve by a reasonable amount for up to 20 trials per JND.

With regards to telling observers whether they were right or wrong after reach trial, it was reported that a psychologist had said people felt bad if they got the wrong answer. Dr. Watson said the procedure self-adapts so that observers are forced to make some mistakes. The feedback keeps subjects involved in the tests. It was suggested that observers be told in the instructions that the procedure forces them to make some mistakes.

In Section 8.2.3, a requirement for a video display with a diagonal size of at least 20 inches was added.

It was suggested that there should be a longer list of parameters for the baseline configuration because ITU Rec. 500 includes many options. Non-baseline configurations could still be used. Over time, the different models could be collected in a database.

There was a discussion comparing DMOS and DVQ (NASA's technique) with the proposed JND method. JNDVIS has greater sensitivity for smaller degradations.

6.2 *Selection of Other Test Labs to confirm first lab test results*

Collecting the data for this paper involved 100 observers, more than three per situation in some cases, and approximately 100 hours of testing. How long would it take another lab to begin conducting tests? It was reported that Phil Corriveau said he would give us the names of people at CRC to contact. There is a good chance they could do some testing. There is also interest in testing at Intel. Vittorio Baroncini could do the tests at FUB. Other possibilities are Jeff Lubin at Sarnoff and Steve Wolf at NTIA/ITS. The question is whether they are motivated.

It was pointed out that they wouldn't have to do all 25 points, perhaps only five. Dr. Watson questioned whether it was necessary have the test validated, since it was done scientifically. There was a comment that this method has more validation the many items in ITU Rec. 500. IEEE may consider it sufficient if members of this committee observe the tests and implementation. No one saw a problem with this.

6.3 *Further discussion and action*

The definition of JND was questioned. The draft standard uses the conventional definition. There was a comment that it could also mean the difference determined by a group of expert viewers who could repeatedly view the sequence. This definition is more detailed. The other refers to real-time observations. We must be careful to define the boundaries where we apply the definition. Dr. Watson agreed that experts carefully examining the scenes would see different results. JND is not completely general. It depends on the conditions of the test, which must be clearly defined.

It was noted that if we want to produce calibrated JND tapes, a mandatory set of parameters must be listed in Section 8.5 of the draft standard. There was agreement among the group on this point. Section 8.5 should list

very specific parameters for one experimental design. This matches the requirements in other IEEE standards. Possibly, after the parameters are listed, we could hint that other parameters could be used for other purposes. With regards to ITU Rec. 500, it was noted that background characteristics are very important for the repeatability of the experiment. Should the standard be more specific about this?

The possibility of using four observers was discussed. Dr. Watson said he preferred leaving it as is, but we could see what happens with more observers. We will have to indicate what the margin of error is in any case. It isn't possible to get it down to where we can ignore it.

There was a question about using a viewing distance other than 5H for the standard. While 4H may be better, it was stated that less than 4H is not good. A problem with using 4H is that the data set we have is based on a 5H viewing distance. If we change it to 4H, we may have no data. This method, even at 5H, gives the observer more chances to see the difference between sequences and, as a result, it may be even more sensitive at 5H than DSCQS at 4H. The main issue is resources. If there is data that indicates a 4H distance would be advantageous and we have the resources to do it, there can be a tentative plan to move the standard to 4H. Dr. Watson said he is conducting tests at two more viewing distances. Neither is 4H. The results from these tests should give us a sense about how much variation we can expect at 4H.

Dr. Watson said he had been approached by two organizations that want to help with the production and distribution of sample JND tapes. It is easy for him to do the samples now. The method of distribution was discussed. Each sequence requires about 200MB, uncompressed. Twenty-five sequences would fit on one DVD. Uncompressed Abekas and AVI file formats were discussed.

It was recommended that we standardize particular samples. Vittorio Baroncini said this is required for his ITU 6Q work. The tapes could be distributed now as a de facto standard. Later we can make it a standard. Chairman Alan Godber said he would like to have a letter from Phil Corriveau indicating we had permission to use the VQEG material. No problem was anticipated using sequences from the CCIR tapes. All the sequences except for SRC 13 and 18 are from these tapes.

Vittorio Baroncini, chair of ITU-R Working Party 6Q would like to submit the proposed standard to his committee. He asked IEEE's authorization to work with us on improving the document and invited members of this committee to the 6Q meeting in March 2002. He proposed having 6Q propose this method for application in a formal expert viewing procedure for subjective tests. Dr. Watson indicated there could be some problems using the method for a large number of trials if the experts take more than a second to make a decision.

Vittorio Baroncini stated if this standard comes out of cooperation between the two committees, it would get accelerated approval. There was concern that we could not submit the proposed standard as a contribution until IEEE feels it is finished. It could then go through the U.S. national committee, which would make it more official.

Dr. Watson asked for comments on specific items where clarification is needed. A brief synopsis, a simple explanation, is needed at the front of the document. It could also be a separate document. Addition of a step-by-step procedure was also suggested.

Chairman Alan Godber requested information on liaison for this document. There were suggestions to limit distribution to those involved with video quality metrics – VQEG, ITU-R WP 6Q and ITU-T SG9. This would not include SMPTE. If it is distributed to VQEG members, it should also be posted on the 6Q reflector to reach additional technical experts.

The title of the proposed standard was discussed. The following title was agreed upon by the committee members:

IEEE P1486 – Standard for the Measurement of Visual Impairments in Digital Video Using a Just-Noticeable-Difference (JND) Scale

The following timetable was proposed:

> May 18 - Dr. Watson makes the changes suggested and Alan Godber writes a cover letter for the draft standard.

May 21 – The cover letter and document are sent out for comments.

June 18 – Deadline for comments

June 18 – July 9 – Communicate comments

July 9 – July 10 – IEEE G-2.1.6 meeting in Minneapolis

July 23 – July 24 – Observe tests and edit the document at NASA AMES. (The possibility of July 22/23 was also mentioned.)

July 24 – Document is ready for distribution

[Action Items]

Alan Godber will draft a cover letter for the proposed standard and post it to the email reflector, IEEE members, VQEG members and other interested parties. VQEG members will forward it to the ITU reflectors for Study Group 9 and Working Party 6Q. A copy will also be sent to John Grigg at T1A1

Leon Stanger will receive the comments.

Item 7 - Further Discussion of Compression Measurement Methodologies

Item 7.1 – Discussion of Future Work, Additional Assignments

There was no further discussion.

Item 8 - Any Other Business.

There was no other business.

Item 9 - Date(s) of Future Meeting(s).

The next meeting will be held in conjunction with the T1A1 meeting in Minneapolis, Minnesota July 9, 2001. If T1A1 is willing to schedule audio for Tuesday morning the meeting will continue in July 10. The subcommittee thanked Arthur Webster and NTIA/ITS for arranging the room.

The meeting was adjourned at 6:18 PM.

Submitted by:
H. Douglas Lung
Secretary

APPENDIX "A"

List of Documents Distributed

11 May 2001

Draft Agenda - IEEE Compression and Processing Subcommittee G-2.1.6, Twentieth Meeting, Monday, May 11, 2001, Alan Godber, Chairman, (216m20an.pdf)

Draft Meeting Record, G-2.1.6, Compression and Processing Subcommittee, Nineteenth Meeting, January 29, 2001, Charleston, SC, Doug Lung, Secretary, IEEE Doc. G-2.1.6/123, May 7, 2001.

Draft Standard for the Measurement of Visual Impairments in Digital Video Using a JND Scale, Andrew B. Watson, NASA Ames Research Center, May 21, 2001, IEEE Doc. G-2.1.6/124.

Measurement of Visual Impairment Scales for Digital Video, Andrew B. Watson and Lindsay Kreslake, Proceedings of the SPIE Volume 4299, IEEE Doc. G-2.1.6/125.

APPENDIX "B"

ATTENDANCE RECORD

May 11, 2001

Name	Affiliation	Telephone	Fax	E-mail
Chairman: Alan Godber	Consultant	(732) 846-4476	(732) 846-4476	agodber@earthlink.net
Secretary: Doug Lung	Telemundo	(305) 884-9664	(914) 412-2886	dlung@transmitter.com
Vittorio Baroncini	FUB	+39-06-54802134	+39-06-54804405	Vittorio@fub.it
Alex Bourrel	BT	+44 7710462033	+44 1473644649	Alex.bourrel@bt.com
David Fibush	Tektronix	(503) 628-3040	(503) 627-4486	dave@fibush.net
Michel Poulin	Leitch	(416) 443-2716	(416) 445-4762	Michel.poulin@leitch.com
James Redford	ABC	(212) 456-4450	(212) 456-4451	Rick.redford@juno.com
Leon Stanger	DirecTV	(310) 726-4676		LStanger@compuserve.com
Hideki Takahashi	Pixelmetric	+65-547-4935	+65-547-4945	rstreeter@att.net
Andrew Watson	NASA	(650) 604-5419	(650) 604-0255	abwatson@mail.arc.nasa.gov
Arthur Webster	NTIA / ITS	(303) 497-3567	(303) 497-5969	webster@its.bldrdoc.gov
Alexander K. G. Woerner	Rohde & Schwarz	(410) 910-7836	(410) 910-7801	Alexander.Woerner@rsa.rohde-schwarz.com