# VQEG Meeting Report

## Briarcliff Manor, NY
## February 25-28, 2002

MYLENE FARIAS

# VQEG MEETING

**Location:** Philips Research USA, Briarcliff Manor, NY
**Date:** February 25-28, 2002
**Attendees:**

| | | |
|---|---|---|
| Arthur Webster | NTIA/ITS | 1 303 497 3567 |
| Philip Corriveau | Intel | 1 503 712 4335 |
| Ann Marie Rohaly | Tektronix | 1 503 627 3048 |
| Stephen Wolf | NTIA/ITS | 1 303 497 3771 |
| Andrew Watson | NASA/AMES | 1 650 604 5419 |
| Jamal Baima | TDF | 333 87 207628 |
| Mylene Farias | UCSB | 1 805 893 8312 |
| Filippo Speranza | CRC | 1 613 998 7822 |
| Ron Renaud | CRC | 1 613 998 7822 |
| Hans Puttenstein | Philips (Netherlands) | 31 40 27 45314 |
| Jorge Caviedes | Philips (USA) | 1 914 945 6430 |
| Vittorio Baroncini | FUB | 39 06 54 80 2134 |
| Alan S. Godber | Consultant | 1 732 846 4476 |
| Alexander Woerner | Rhode & Schwartz | 1 410 910 7836 |
| David Hands | British Telecom | 44 1473 64 8615 |

Cermak
Geoff
Alan Godber

## SUMMARY

This report contains the notes taken by Mylene Farias during the last VQEG meeting on Briarcliff Manor, NY on February 25-28, 2002.

## VQEG BRIARCLIFF NEW YORK MEETING AGENDA

| | |
|---|---|
| **Monday** 9:00-12:00 | Introductions<br>Approval of Agenda<br>Distribution of Documents<br>(Meeting documents must be sent to reflector by Feb 18)<br>Short Overview of Topics<br>RRNR-TV Test<br>Test Plan<br>Scenes<br>HRCs<br>FR-TV Test<br>Test Plan<br>Scenes<br>HRCs<br>Independent Lab Group<br>Status Report<br>Possible Viewing<br>Goals of meeting |
| **Monday** 1:00 - 6:00 | FRTV<br>Test Plan Discussion and approval<br>Expert Viewers Test<br>Naive Viewers Test<br>Test Scene guidelines for ILG<br>HRC guidelines for ILG |
| **Tuesday** 9:00 - 6:00 | RRNR-TV<br>Test Plan Discussion and approval<br>Test Scene guidelines for ILG<br>HRC guidelines for ILG<br>10:00 am - Philips Demo of HDTV |
| **Wednesday** 9:00 - 12:00 | Scene/HRC Viewing (Open to all of VQEG) |
| **Wednesday** 1:00 - 6:00 | Liaisons to ITU-R WP6Q, ITU-T SG9, T1A1, IEEE |
| **Thursday** 9:00 - 12:00 | Multimedia Test Status Report<br>Other Business<br>Combined Model Effort<br>Software Tools Repository |
| **Thursday** 1:00 - 6:00 | Independent Lab Meeting/Viewing/Selection<br>(Only Independent Lab Representatives) |
| **Friday** | (Reserved if needed) |

## MONDAY - 25TH OF FEBRUARY

MORNING

- Introductions
- Approval of Agenda - The original agenda was modified to accommodate everybody's schedule. (See previous page)
- Presentation of the independent labs which will run the subjective test: CRC (Canada) and TDF (Italy)
    - Fees for proponents to cover costs of independent labs: Initial estimate of U$2,500 for test. (70 – 80 working days * U$250.00 = U$35,000 to U$40,000). Cost may change depending of how many proponents actually participate on the experiment.

| Task | Time allocated |
|---|---|
| Selection of sequences, HRCs, etc | 2 days |
| Coding | 20 days |
| Editing | 20 days |
| Test | 10-20 days |
| Expert test | 3 days |
| Data analysis | 15 days |
| Total | 70-80 days |

    - Call for proposal will be released ~~on~~ March.
- The FR-TV and RRNR-TV Test plans have to be finalized in this meeting.
    - Subjective tests

There was a long discussion about the subjective tests that are going to be used. Some people argued that a more discriminating technique was needed. The last VQEG experiment generated results that showed that the performance of all objective measures performed as good (statistically) as PSNR. The ideal goal was to obtain an easy to understand result form the test. This time, it was agreed that there was not going to be a competition for the best model. VQEG will not pick winners.

The possibility of using expert viewing in the test was also discussed. The initial idea was to do a special subjective testing only with experts, where videos would be shown in a different way, allowing for example, zooming and still frames. The goal was to "clean" (decrease variance) the data. The majority of the participants did not agree with the idea. First, because they didn't believe this was going to clean the data. Second, video quality is different from still picture quality and allowing such detailed examination of the data does not seem right. Third, there is no good way of combining the data from the two sets of subjects because they are under different rules. (for more detail see old test plan.)

The need for keeping the name of the participating labs secret was also discussed. The companies were afraid the results could be used in a non-ethical way. The video sequences will be kept secret. 25% of the material will be completely new.

    - HRCs need to be chosen

It was argued that 40 variations of the same originals should be avoided. This makes the experiment very tiring and increases the "noise" in the results because subjects start seeing errors (artifacts) where there are none. Several suggestions were given about how to create the HRCs needed:

1. Change the viewing distances: Bo Watson argued that this is an easy way of differentiating the models from PSNR, which certainly does not account for different viewing distances.
2. Eliminate high quality conditions: Subjects cannot differentiate them from originals, therefore the annoyance response saturates. It was argued then that this would only be a waste of time.
3. Several codecs must be used.

- Liaisons to ITU-R WP6Q, ITU-T SG9, T1A1, IEEE should be written & sent.

AFTERNOON

- Discussion about distribution of HRCs and scenes for FR.

  There were 2 main proposals.

  - Full Matrix – All conditions would be varied in linear form. The big advantage of this approach is that it is simpler to implement. It is the same approach used last time. The disadvantage is that a lot of these conditions are lost because they are not perceptually visible (Watson).
  - Bigger hybrid (Sparse) Matrix – In this approach, instead of covering all conditions, a set of important conditions is chosen. Nevertheless, a smaller full matrix would be kept. The idea is not to waste so many conditions on being able to include different HRCs.

  VQEG has decided to go for a hybrid approach for the rest of the material, which will include a 6×6 full matrix design spanning the range of quality and an additional 19 points chosen by the ILG to fill out the test. (The total of SRC×HRC conditions will be 36 + 19 + 9 = 64.) The 6×6 will be chosen to include HRCs and SRCs covering the range of quality of test filling out two tables (one for 525 and one for 625) like the one in Figure 1.

  The population of the matrix in Figure 1 will be full and will be done applying the rules given below:
  - No more than 3 HRCs will utilize a coder from one given manufacturer,
  - Try to use as many manufacturers as possible,
  - Include compression and two transcoding (or cascading) processes.

  The selection of the additional 19 points will be done applying the rules below:
  - To fill out the quality range spanned by the test.
  - To satisfy secrecy conditions.
  - To include other conditions such as noise, slice errors, etc. (as much as is feasible)

| | $SRC_1$ | $SRC_2$ | $SRC_3$ | $SRC_4$ | $SRC_5$ | $SRC_6$ |
|---|---|---|---|---|---|---|
| $HRC_1$ | | | | | | |
| $HRC_2$ | | | | | | |
| $HRC_3$ | | | | | | |
| $HRC_4$ | | | | | | |
| $HRC_5$ | | | | | | |
| $HRC_6$ | | | | | | |

Figure 1 – Full matrix table to be filled out by IL group

## TUESDAY - 26TH OF FEBRUARY

MORNING

- Continuation of the discussion about the FR test plan.
  - Range of bitrates to be used – For broadcasting (TV), compression bitrates typically do not go much lower than 1 Mbps. Is it valid to use values lower than this? Some participants thought that for other applications lower bitrates are used and therefore it would be interesting to include values smaller than 1 Mbps. On the other hand, going too low may cause subjects not to differentiate between very bad quality videos. It was agreed that the quality range should go from **512 Kbps to 6 Mbps**.
  - Some more discussion on the hybrid matrix - The matrix will include a 6×6 full matrix.
  - DSCQS vs. DSIQS – One of the proposals was to change the test plan. The proposal believed using an integer scale decreases the confidence interval. The proposal was rejected. DSCQS will be used
  - Normalization of data for model – Each proponent has to do its own normalization. (as part of the model + object code should be to T1, TG.
  - Evaluation metrics - to be calculated on the subjective and objective data

    For each SRC: Pearson, Spearman, Outlier Ratio, RMS Error, Resolving Power, and Classification Errors. (Lost 2 from T1.TR ---)
- Demo HDTV

  Jorge showed us the HDTV Philips has been working on. The image is really good, however it is ringing artifacts were visible. These artifacts come from enhancement (edges, sharpness, contrast, etc.) in the equipment. They are specially visible near white edges.

AFTERNOON

- RRNR-TV plan
  - What to do when there is a scene change?
  - SSCQE method – in RRNR, contrary to the FR, the SSCQE method will be used. This test uses a slider and does measures real-time measurement.

    A hidden reference will be added to the set of sequences, i.e., the originals are randomized together with the other sequences. The reference is needed because without it the content of the video will have a great influence on the quality (Suzie effect).

    Philip argues that SSCQE works fine. But he thinks the averages of the originals should be subtracted to avoid bias. The bias is created because after the subjects start moving the slider, it is common that it doesn't get the maximum and minimum quality values anymore.

    Watson argues that it might be a better idea to just rescale the data.
  - Shift caused by codecs – Most codecs add a shift to the coded-decoded video, which can be quite large (up to 32 pixels). The question raised is 'Should the model correct it by itself or the amount of shift must be limited?'.

## WEDNESDAY - 27TH OF FEBRUARY

MORNING

- Continuation of the discussion about the RRNR test plan
    - Gain/offset/bias of subjective data

      Data needs to be "registered", i.e., the time delay of the subject needs to be taken into account. Besides that, there is the problem of bias (subjects tend to 'avoid' low and high quality).

      Watson says that the data needs to be normalized for each subject. It was decided that this would only be done if possible.

    - HRCs specification

      It was decided that statistical multiplex was not going to be done. Cascading with noise will not be used neither. But ILG will try to cover as many cases as possible.

- Scenes viewing

  Some of the scenes that could be used in the test were shown. Besides scenes with compression artifacts, it was also shown artifacts form transmission errors (noise). In those cases, the artifacts appeared as missed blocks with some noise inside. The participants picked some of them as representatives of the limit range to be used in the subjective tests.

- Extra discussion on FR-TV test plan
    - So far, only two laboratories will participate in the subjective test
    - Include two viewing distances in the test?

      Watson thinks this could be interesting way of differentiating PSNR from other metrics. Some people argue that this would increase the costs. It was agreed that if possible, the ILG will run a few subjects with another distance (8H).

- Choice of co-chairs
- Discussion of "free" policy of VQEG – some people fear free data from the website can be used in a disloyal way.

AFTERNOON

- RRNR test plan
    - Format for reference and data
    - Synchronization of source and processed sequences

      Codec causes a shift in a field (fields get inverted). This is not so important to FR, but may be crucial to RRNR. Most people think correction should be done by ILGs. This is not a typical situation because only 1 min. of the video is being edited. Wolf and Ron have and idea of how to correct it.

    - Normalization issues – Normalize data from labs to make results more comparable.

## THURSDAY- 27TH OF FEBRUARY

MORNING

- ❑ Discussion of schedule of RRNR-TV (see report)
- ❑ Multimedia
    - o Problems combining video and audio quality
    - o Cross-modal audio/video effects
    - o Lower bitrate (quality) for video
    - o Are the video quality/audio quality models going to work on multimedia?
    - o Request audio expertise
    - o Issues on subjective testing – There has been very little work done on quality assessment of multimedia. There are many open questions such as, how to show videos in a subjective test and how to combine audio and video quality.