From: ituvidq-admin@its.bldrdoc.gov on behalf of Stephen Wolf [swolf@its.bldrdoc.gov]
Sent: Friday, May 04, 2001 9:44 AM
To: VQEG
Subject: [VQEG] Comments on NR-RR Test Plan

Dear Colleagues,

We would like the following discussion items added to the NR-RR agenda next week.

Stephen Wolf

Balanced Test Plan Design

Agreement must be reached on a balanced test design to assure proper ANOVA. NTIA has submitted the following balanced test design that needs to be discussed and agreed to by the group:

1. Scenes - 6 carefully selected 1 minute segments.

2. HRCs - 10 total (1 original, and 9 processed versions selected from the total set of 40 possible HRCs).

This will produce a total of 60 minutes of SSCQE video. To assure that all the viewers see all the video, we am proposing that each subject view this 60 minutes of video using two 30-minute sessions, separated by a break (each tape will also have a short stabilization segment at the beginning).

Multiple randomizations are desired so we will need to edit more than 2 viewing tapes. This randomization should be performed at the clip level (i.e., the ordering of each one minute scene x HRC should be randomized). Preferably, 3 sets of tapes should be used (lets call these the red, green, and blue tapes). Subjects should be randomly assigned to one of the 6 possible orderings (R1-R2, R2-R1, G1-G2, G2-G1, B1-B2, B2-B1). Each lab should have an equal number of subjects at each ordering. Perhaps 3 subjects per ordering, for a total of 18 viewers per lab.

The first 8 to 10 seconds of each clip should be discarded to allow for stabilization of the viewer's responses. This leaves 50 seconds from each video clip to be considered for data analysis, or 60 clips of 50 seconds each.

Time Alignment of Subjective and Objective Measurements

All the subjective labs must sample the slider positions at the same time codes. In addition, the objective models need to know when these subjective responses are sampled. This is necessary to assure that the objective model's outputs are time synchronized to the sampling of the slider positions.

The data analysis must address the time alignment of the objective model's output and the average slider responses before the correlation analysis is performed. The latency that results from viewer reaction times and slider "stiffness" is uninteresting and should not be evaluated by VQEG. Thus, the time alignment algorithm for aligning the average subjective slider positions and objective estimates of these slider positions needs to be specified in the test plan beforehand and agreed to by the group.

Amplitude Scaling of Objective Measurements to Subjective Measurements

Section 5.2 (post processing quantization) and Section 9 (optimum mean square quantization method) should be deleted and the rest of the test plan re-written to reflect this. Such coarse quantization before objective and subjective data comparison will discard important information that may allow VQEG to differentiate models and will complicate the analysis, not make it simpler. The effect of data outliers on the behavior of the quantizer is potentially troublesome and the model submitters should not be expected to design their systems around this requirement. The proposal is to discard quantization and replace it with a simple linear fitting procedure (i.e., gain and offset) that will be applied to the time aligned objective and subjective time histories.

Temporal Registration of Processed Video

Since tape editing is involved, there is always the possibility that editing errors will be made. In addition, clips placed on the viewing tapes may be off a frame or two because of tape dynamics. Thus, some mechanism need to be established to assure that the final tapes submitted to the models and subjects (both original and processed) are *field* accurate. Otherwise, the objective models will have to continually resynchronize at each clip boundary. The brief method of test sequence synchronization (Section 3.3) that is currently in the plan is inadequate and needs to be expanded upon to reflect these comments.

Ideally, time alignment should not even be part of the submitted model for the following reasons. First, time alignment should only need to be done once for an MPEG-2 HRC (provided frames are not frozen). Thus, time alignment could be performed during setup and systems should not have to use part of the valuable feature bandwidth for time alignment information. Second, some applications that use the objective measurement will inherently have time alignment information available and thus will not require additional time alignment. Third, some participants with valuable models may have little experience with time alignment, and thus the test may unfairly penalize these systems if they are supplied with a poor time alignment technique. Fourth, "black to scene" time alignment methods will be useless in the real world. Five, if time alignment is part of the model, then there will be no way to distinguish between the performance of the model and the performance of the time alignment algorithm. Sixth, the method of time alignment could be submitted separately as an ITU Recommendation in its own right.

Thus, we would suggest one of the following two options for temporal registration of the processed video. The first option would be to assure that the original and processed video tapes that are submitted to the models are correctly time aligned (in reality, this has to be done anyway in light of the comments regarding tape editing errors). The second option is to have a separate program compute the time alignment and provide this temporal registration information to the models.

Spatial Registration of Processed Video

Spatial registration - Since spatial registration is constant for any particular HRC, this type of calibration only has to be performed once when the objective measurement system is attached to the HRC. Thus, quality-monitoring systems should not have to include spatial registration as part of their bit rate allocation. Two solutions are possible. The spatial shift can be estimated externally by some agreed upon method and this information can be provided to the models. Alternatively, the spatial shift can be removed before the processed video stream is provided to the models. In either case, the test plan needs to specify the method for spatial shift estimation and removal, which can be the subject of a separate ITU Recommendation.

Processed Valid Region (PVR)

Processed valid region (PVR) (i.e., the portion of the processed video frame that contains valid picture content) can be a function of the input scene. For instance, if letterbox scenes are used, the processed valid region is reduced. Thus, a mechanism must be made in the test plan to allow PVR inputs to the models.

Upstream Monitoring vs. Downstream Monitoring

Figure 3 allows upstream monitoring systems (part 1) to be submitted as well as downstream monitoring systems (part 2). However, section 4.1 (model input and output format) uses the terms “part 1”, “part 2”, “input”, and “output”, in a confusing manner that does not allow an upstream monitoring system to be submitted. We would suggest the following re-wording of section 4.1:

> Model Part 1 Original Video Side:
>
> > The software for the original video side will be given the original test tape in the final file format to be used in the test, and a reference data file that contains the reduced-reference information (see Model Part 1 Processed Video Side).
> >
> > The software will produce an ASCII file, listing the Time Code of the original sequence, and the resulting predictive MOS (MOSp) of the model, with a

resolution of 2 samples per second.

Model Part 1 Processed Video Side:

The software for the processed video side will be given the processed test tape in the final file format to be used in the test, and produce a reference data file. The amount of reference information in this data file will be evaluated in order to estimate the bitrate of the reference data and consequently the class of the method (0, 10, 56 or 256 Kbits/s).

Model Part 2 Original Video Side:

The software for the original video side will be given the original test tape in the final file format to be used in the test, and produce a reference data file. The amount of reference information in this data file will be evaluated in order to estimate the bitrate of the reference data and consequently the class of the method (0, 10, 56 or 256 Kbits/s).

Model Part 2 Processed Video Side:

The software for the processed video side will be given the processed test tape in the final file format to be used in the test, and a reference data file that contains the reduced-reference information (see Model Part 2 Original Video Side).

The software will produce an ASCII file, listing the Time Code of the processed sequence, and the resulting predictive MOS (MOSp) of the model, with a resolution of 2 samples per second.

Note: The file format for the feature reference information is probably not ASCII, but binary, since in general this will contain compressed information.

Overly Restrictive Repeatability Specification (Section 4.3)

In section 4.3, the stated restriction is that the software should produce the same results within an acceptable error of 0.1%. This "acceptable error" is far too tight considering the extremely low bandwidth of the objective measures and the accuracy of the subjective data. When given the same input data, some effective bit rate reduction techniques may produce measurements that are known to vary from one run to the next (e.g., random sub-sampling). A 0.1% repeatability threshold probably means that these types of techniques have to be artificially modified to meet the threshold requirement. The recommendation is to raise the acceptable repeatability threshold to 2%.

Analysis of Subjective Data

The test plan should include a separate section that specifies the analysis of the subjective data. In particular, the subjective data should be analyzed to form a theoretical limit on model performance. This can be used to help answer the question "Are any of the models good enough to recommend?"

No Clear Path to Consensus

Section 5 spells out a lot of nice evaluation metrics but there is no detail as to how a decision will be reached by VQEG. This was a major shortcoming of the first VQEG test plan. We really need to agree on this decision process up front, not after the data is collected. In Section 5.5, the models should be grouped by bit rate for comparison purposes. Each bit rate may have a set of measurement applications and a recommendation could potentially be written for each of the four target bit rates (i.e., VQEG may want to recommend *more* than one model). If a lower bit rate model wins out over all higher bit rate models, then this would be a significant event and may tell us something about the highest bit rate that is required (in this case, all higher bit rate models would be discarded). However, we still have to spell out in the test plan the decision process that will be used and the goal that must be met by video quality models before they are recommended to the ITU study groups.

Proposal to Change VQEG Mission

The probability of reaching consensus on a "winner" as a result of these VQEG tests is extremely low, particularly given the fact that reduced reference systems cannot in principal achieve as high a correlation to subjective score as can full reference systems. In addition, the rapid pace of technology would all but guarantee that even if a winning system were declared, it would be obsolete by the time VQEG makes its recommendation and the ITU adopts that recommendation. There is another option, or path forward, that VQEG can embark on now that is guaranteed to reach convergence. The proposal is to change VQEG's mission to make no attempt to pick a winner. Instead, the purpose of VQEG would be to collect independent test data to determine the accuracy and cross-calibration (i.e., the mapping relationships between video quality models) of every submitted video quality model. This accuracy and cross-calibration data would then be forwarded to the ITU as a technical report. An excellent starting point for this concept is the work of T1A1. The current version of the T1A1 accuracy and cross-calibration document can be found at the following location:

LINK: <http://www.t1.org/FileMgr/GetOneFile.taf?FileName=1A110275&NW=Y>

TITLE: Methodological Framework for Specifying Accuracy and
Cross-Calibration of Video Quality Metrics