INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 1997 - 2000

COM 12-
September 1998
Original: English

Questions: 10/12, 11/12

SOURCE: Co-chair objective testgroup VQEG, KPN Research

TITLE: Evaluation of new methods for objective testing of video quality: objective test plan

Abstract

The ITU is currently in the process of developing one or more recommendations for the objective measurement of video quality. This contribution presents the objective test plan that has been drafted by members of the VQEG (Video Quality Experts Group) ad hoc committee for the objective test plan. This test plan will be used in the bench marking of the different proposals and was offered to the participating ITU Study Groups (ITU-T Study Groups 9 and 12 and ITU-R Study Group 11) for further review and comment in the beginning of 1998. It was further modified during the second VQEG meeting (Gaithersburg, USA May 1998), taking into account their comments. The objective test plan will be used to evaluate video quality in the bit rate range of 768 kbit/s to 50 Mbit/s. In conjunction with the subjective test plan, it will be used to evaluate several proposed methods for objectively assessing video quality by measuring the correlation between subjective and objective assessments. It is expected that parts of this test plan will be included in new Draft Recommendations in the area of video quality, probably as an Annex.

Contact Persons: Mr. John G. Beerends
Tel: +3170 3325644
Fax: +3170 3326477
E-mail: j.g.beerends@research.kpn.com

Mr. Mihir Ravel
Tel: +1 503-627-2768
Fax: +1 503-627-5177
E-mail: mihir.ravel@tek.com

VQEG OBJECTIVE VIDEO QUALITY MODEL TEST PLAN

1 Introduction

The ITU is currently in the process of developing one or more recommendations for the objective measurement of video quality. The Video Quality Expert Group (VQEG[1]) drafted an objective test plan which defines the procedure for evaluating the performance of objective video quality models as submitted to the ITU. It is based on discussions at the 1st Meeting of VQEG, October 14-16, 1997, Turin, Italy. This test plan was offered to the participating ITU Study Groups (ITU-T Study Groups 9 and 12 and ITU-R Study Group 11) for further review and comment in the beginning of 1998. It was further modified during the 2nd VQEG meeting, May 27-29,1998, Gaithersburg, USA and during the period June-September 1998 by e-mail and submitted to the ITU-T SG12 by KPN Research.

The objective models will be tested using a set of test sequences selected by the VQEG Independents Labs and Selection Committee (ILSC). The test sequences will be processed through a number of hypothetical reference conditions (HRC's) as can be found in the subjective test plan.

The quality predictions of the models will be compared with subjective ratings by the viewers of the test sequences as defined by the VQEG Subjective Test Plan. The Subjective Test Plan has two separate but overlapping subjective test experiments to cover the intended bit rate range of 768 kbit/s to 50 Mbit/s, and the model performance will be compared separately with the results from each of the two subjective test experiments. Based on the VQEG evaluation of proposed models, the goal is to recommend method(s) for objective measurement of digital video quality for bit rates ranging from 768 kbit/s to 50 Mbit/s. The preference is one recommended model, but multiple models are possible.

1 Contact: Arthur Webster, +1 303-4973567, E-mail webster@its.bldrdoc.gov

2 Data formats and processing

2.1 Video data format, general

Objective models will take two Rec. 601 digital video sequences as input, referred to as Source and Processed, with the goal of predicting the quality difference between the Source and Processed sequences. The video sequences will be in either 625/50 or 525/60 format. The choice of HRC's and Processing will assure that the following operations do not occur between Source and Processed sequence pairs:

- Visible picture cropping
- Chroma/luma differential timing
- Picture jitter
- Spatial scaling (size change)

2.2 Model input and output data format

The models will be given two ASCII lists of sequences to be processed, one for 525/60 and one for 625/50. These input files are ASCII files, listing pairs of video sequence files to be processed. Each line of this file has the following format:

<source-file> <processed-file>

where <source-file> is the name of a source video sequence file and <processed-file> is the name of a processed video sequence file, whose format is specified in section 2.5 of this document. File names may include a path. Source and processed video sequence files must contain the exact sequence pattern specified in section 2.3 and section 2.5. For example, an input file for the 525/60 case might contain the following:

/video/src1_525.yuv **/video/src1_hrc2_525.yuv**

/video/src1_525.yuv **/video/src1_hrc1_525.yuv**

/video/src2_525.yuv **/video/src2_hrc1_525.yuv**

/video/src2_525.yuv **/video/src2_hrc2_525.yuv**

From these lists the models are allowed to generate their model setting files from which the model can be ran.

The output file is an ASCII file created by the model program, listing the name of processed sequence and the resulting Video Quality Rating (VQR) of the model. The contents of the output file should be flushed after each sequence is processed, to allow the testing labs the option of halting a processing run at any time. Alternately the models may create an individual output file for each setting file and collect all data into a single output file using a separate collect program. Each line of the ASCII output file has the following format:

<processed-file> VQR

Where <processed-file> is the name of the processed sequence run through this model, without any path information; and VQR is the Video Quality Rating produced by the objective model. For the input file example, this file contains the following:

src1_hrc2_525.yuv 0.150

src1_hrc1_525.yuv 1.304

src2_hrc1_525.yuv 0.102

src2_hrc2_525.yuv 2.989

Each proponent is also allowed to output a file containing Model Output Values (MOV's) which the proponents consider to be important. The format of this file will be

src1_hrc2_525.yuv 0.150 MOV_1 MOV_2,... MOV_N

src1_hrc1_525.yuv 1.304 MOV_1 MOV_2,... MOV_N

src2_hrc1_525.yuv 0.102 MOV_1 MOV_2,... MOV_N

src2_hrc2_525.yuv 2.989 MOV_1 MOV_2,... MOV_N

All video sequences will be displayed in overscan and the non-active video region is defined as:

the top 14 frame lines

the bottom 14 frame lines

the left 14 pixels

the right 14 pixels.

Possible small differences between individual monitors are averaged out in the analysis of the subjective data. A sanity check for large deviations from the above non-active region will be carried out by the subjective test labs. If in the normalization a different active region is found and the cropping size is such that it will be visible within the active video region this sequence will not be used.

Models will only get one input parameter, the 525/60 versus 625/50 input format, in the form of two separate lists. All other parameters like screen distance (5H), maximum luminance level (70 cd/m^2), background luminance, video format, gamma of the monitor, etc are fixed for the test and thus are not required for the setting files.

2.3 Test sequence normalization

As a Source video sequence passes through an HRC, it is possible that the resulting Processed sequence has a number of scaling and alignment differences from the Source sequence. To facilitate a common analysis of various objective quality measurement methods (referred to as models), Tektronix will normalize the Processed sequences to remove the following deterministic differences that may have been introduced by a typical HRC:

- Global temporal frame shift (aligned to ±0 field error)
- Global horizontal/vertical spatial image shift (aligned to ±0.1 pixel)
- Global chroma/luma gain and offset (accuracy to be defined)

The normalized sequences will be used for both subjective and objective ratings. The normalized sequences will be sent on D-1 digital video tape to the Subjective Testing Labs for the DSCQS (Double Stimulus Continuous Quality Scale) rating. The normalized sequences will also be used for analysis by the objective models. The sequences will be available on computer tape for the objective ratings in the following two formats:

- 8 mm Exabyte format (archived in UNIX tar format with a block factor of 1)
- 4 mm DDS3 format (details to be defined)

The first and last second of the sequences will contain an alignment pattern to facilitate the normalization operation. The pattern is a coded set of alternating light/dark blocks in the upper half of the image (provided by Tektronix) and will not be included in the portion of the sequence shown to subjective assessors. The required normalization will be estimated with a non-confidential set of algorithms (provided by Tektronix) over the first second alignment pattern portion of the sequence. The normalization from the first second estimate will then be applied uniformly over the length of the sequence on the assumption that the differences needing normalization are invariant over the sequence length. The last second of alignment pattern may be used to determine if the values have remained constant through the length of the sequence. Finally ten frames before the 8 seconds video sequence and ten frames after the 8 seconds video sequence will not be used in both the objective and subjective evaluation. A complete sequence on D-1 tape and Exabyte/DAT will be:

AlignmentPattern(1sec)+ VideoNotUsed(10frames)+ **Video(8sec)**+VideoNotUsed(10 frames)+AlignmentPattern(1 sec)

The normalization will be done by Tektronix and will be completed approximately four weeks after receiving the test sequences (after August 7th when all the proponents have submitted and tested their models in their assigned objective testlabs).

2.4 Test sequence objective analysis

Each proponent receives normalized Source and Processed video sequences after September 25th, 1998. Each proponent analyzes all the video sequences and sends the results to the Independent Labs and Selection Committee (ILSC) before December 11th, 1998.

The independent lab(s) must have running in their lab the software that was provided by the proponents, see section ***3.2***. To reduce the work load on the independent lab(s), the independent lab(s) will verify a random sequence subset (about 20%) of all video sequences to verify that the software produces the same results as the proponents within an acceptable error of 0.1%. The random 30 sequence subset will be selected by the ILSC and kept confidential to the ILSC. If errors greater than 0.1% are found, then the independent lab and proponent lab will work together to analyze intermediate results and attempt to discover sources of errors. If processing and handling errors are ruled out, then the ILSC will review the final and intermediate results and recommend further action.

The model output will be a single Video Quality Rating (VQR) number calculated over the sequence length (or a subset) not containing the alignment patterns. The VQR is expected to correlate with the Difference between the Source and Processed Mean Opinion Scores (MOS) resulting from the VQEG's subjective testing experiment. This Difference in subjective MOS's is referred to as DMOS. It is expected that the VQR's and DMOS's will be positive in typical situations and increasing values will predict increasingly perceptible differences between Source and Processed sequences. Negative values of both may occur in certain situations and will be allowed.

2.5 Data format, specifics

The test video sequences will be in ITU Recommendation 601 4:2:2 component video format using an aspect ratio of 4:3. This may be in either 525/60 or 625/50 line formats. The temporal ordering of fields F1 and F2 will be described below with the field containing line 1 of (stored) video referred to as the Top-Field.

Data storage:

A LINE: of video consists of 1440 8 bit data fields in the multiplexed order: Cb Y Cr [Y] ... . Hence there are 720 Y's and 360 Cb's and 360 Cr's per line of video.

A FRAME: of video consists of 486 active lines for 525/60 Hz material and 576 active lines for 625/50 Hz material. Each frame consists of two interlaced Fields, F1 and F2. The temporal ordering of F1 and F2 can be easily confused due to cropping and so we make it specific as follows:

For 525/60 material: F1--the Top-Field-- (containing line 1 of FILE storage) is temporally LATER (than field F2). F1 and F2 are stored interlaced.

For 625/50 material: F1--the Top-Field-- is temporally EARLIER than F2.

The Frame SIZE:

for 525/60 is: 699840 bytes/frame,

for 625/50 is: 829440 bytes/frame.

A FILE: is a contiguous byte stream composed of a sequences of frames as described in section 2.3 above. These files will thus have a total byte count of

for 525/60: 320 frames = 223948800 bytes/sequence,

for 625/50: 270 frames = 223948800 bytes/sequence

Multiplex structure: Cb Y Cr [Y] ... 1440 bytes/line

720 Y's/line

360 Cb's/line

360 Cr's/line

TABLE 2: Format summary

	-- 525/60 --	-- 625/50 --
active lines	486	576
frame size (bytes)	699840	829440
fields/sec (Hz)	60	50
Top-Field (F1)	LATER	EARLIER
Seq-length (bytes)	223948800	223948800

3 Testing procedures and schedule

3.1 Submission of intent before June 22 1998

The submission procedure is dealt with in separate ITU contributions (e.g., COM 12-30, December 1997). All proponents wishing to propose their objective video quality models for ITU recommendation should submit an intent to participate to the VQEG chair (see footnote 2, page 3) by June 22^{nd,} 1998. The submission should include a written description of the model containing principles and available test results in a fashion that does not violate proponents' intellectual property rights.

3.2 Final Submission of executable model before August 7th 1998

A set of 4 source and processed video sequence pairs will be used as test vectors. They were made available to all proponents, at the beginning of April 1998, in the final file format to be used in the test.

Each proponent will send an executable of the model, together with the test vector outputs, by July 22nd , 1998 to an independent lab(s) selected by the ILSC. The executable version of the model must run correctly on one of the three following computing environments:

- SUN SPARC workstation running the Solaris 2.3 UNIX operating system (SUN OS 5.5).
- WINDOWS NT Version 4.0 workstation.
- SGI workstation running IRIX Version no [to be decided].

Alternately, proponents may supply object code working on either the computers of the independent lab(s) or on a computer provided by the proponent. The proponents have until August 7th to get their code running.

The independent lab will verify that the software produces the same results as the proponent with a maximum error of 0.1%. If greater errors are found, the independent lab and proponent lab will work together to discover the sources of errors and correct them. If the errors cannot be corrected, then the ILSC will review the results and recommend further action.

3.3 Results analysis

The results as provided by the proponents and verified by the independent lab(s) will be analyzed to derive the evaluation metrics of section **4**. These metrics are calculated by each proponent and verified by the ILSC, or they may be calculated completely by the ILSC and verified by the proponents. The results will be reported anonymously to the outside world (proponent a,b,c,..) but identified by proponent to VQEG.

4 Objective quality model evaluation criteria

4.1 Introduction to evaluation metrics

A number of attributes characterize the performance of an objective video quality model as an estimator of video picture quality in a variety of applications. These attributes are listed in the following sections as:

- Prediction Accuracy
- Prediction Monotonicity
- Prediction Consistency

This section lists a set of metrics to measure these attributes. The metrics are derived from the objective model outputs and the results from viewer subjective rating of the test sequences. Both objective and subjective tests will provide a single number (figure of merit) for each Source and

Processed sequence pair that correlates with the video quality difference between the Source and Processed sequences. It is presumed that the subjective results include mean ratings and error confidence intervals that take into account differences within the viewer population and differences between multiple subjective testing labs.

4.2 Prediction nonlinearity

The outputs by the objective video quality model (the VQR's) should be correlated with the viewer DMOS's in a predictable and repeatable fashion. The relationship between predicted VQR and DMOS need not be linear as subjective testing can have nonlinear quality rating compression at the extremes of the test range. It is not the linearity of the relationship that is critical, but the stability of the relationship and a data set's error-variance from the relationship that determine predictive usefulness. To remove any nonlinearities due to the subjective rating process (see Figure 1.) and to facilitate comparison of the models in a common analysis space, the relationship between each model's predictions and the subjective ratings will be estimated using a nonlinear regression between the model's set of VQR's and the corresponding DMOS's.

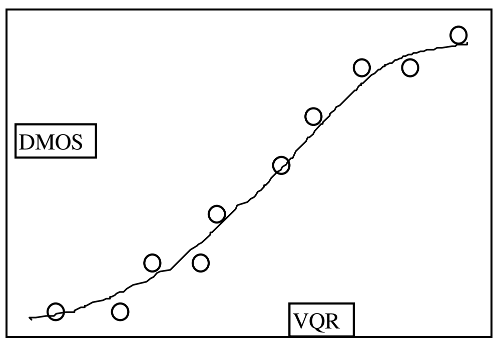


Figure 1. Example Relationship between VQR and DMOS

The nonlinear regression will be fitted to the [VQR,DMOS] data set and be restricted to be monotonic over the range of VQR's. The functional form of the nonlinear regression is not critical except that it be monotonic, reasonably general, and have a minimum number of free parameters to avoid overfitting of the data. As the nature of the nonlinearities are not well known beforehand, several functional forms will be regressed for each model and the one with the best fit (in a least squares sense) will be used for that model.

The functional forms to be regressed are listed below. Each regression will be with the constraint that the function is monotonic on the full interval of quality values:

4.2.1.1 (1) The 4-parameter cubic polynomial

$DMOS_p(VQR) = A0 + A1*(VQR) + A2*(VQR)^2 + A3*(VQR)^3$

fitted to the data [VQR,DMOS].

(2) The same polynomial form as in (1) applied to the "inverse data" [DMOS, VQR].

(3) The 5-parameter logistic curve:

$DMOS_p(VQR) = A0 + (A1-A0)/(1 + ((X+A5)/A3)^{A4})$

fitted to the data [VQR,DMOS].

The chosen nonlinear regression function will be used to transform the set of VQR values to a set of predicted MOS values, $DMOS_p(VQR)$, which will then be compared with the actual DMOS values from the subjective tests.

Besides carrying out an analysis on the mean one can do the same analysis on the individual Opinion Scores (OS), leading to individual Differential Opinion Scores (DOS). This has the advantage of taking into account the variations between subjects. For objective models there is no variance and thus $OS_p= MOS_p$ and $DOS_p=DMOS_p$.

4.3 Evaluation metrics

This section lists the evaluation metrics to be calculated on the subjective and objective data. Once the nonlinear transformation of section **4.2** has been applied, the objective model's prediction performance is then evaluated by computing various metrics on the actual sets of subjectively measured DMOS and the predicted $DMOS_p$. The set of differences between measured and predicted DMOS is defined as the quality-error set Qerror[]:

$$Qerror[i] = DMOS[i] – DMOS_p[i]$$

where the index 'I' refers to an individual processed video sequence.

Metrics relating to Prediction Accuracy of a model

Metric1: The Pearson linear correlation coefficient between DOS_p and DOS, including a test of significance of the difference.

Metric2: The Pearson linear correlation coefficient between $DMOS_p$ and DMOS.

Metrics relating to Prediction Monotonicity of a model

Metric3: Spearman rank order correlation coefficient between $DMOS_p$ and DMOS.

A pair-wise comparison of pairs of HRC's on a scene by scene basis has also been proposed for examining the correlation between subjective preferences and objective preferences, and merits further investigation by the VQEG for inclusion in these tests.

Metrics relating to Prediction Consistency of a model

Metric4: Outlier Ratio of "outlier-points" to total points N.

Outlier Ratio = (total number of outliers)/N

where an outlier is a point for which: ABS[Qerror[i]] > 2*DMOSStandardError[i].

Twice the DMOS Standard Error is used as the threshold for defining an outlier point.

4.4 Generalizability

Generalizability is the ability of a model to perform reliably over a very broad set of video content. This is obviously a critical selection factor given the very wide variety of content found in real applications. There is no specific metric that is specific to generalizability so this objective testing procedure requires the selection of as broad a set of representative test sequences as is possible. The test sequences and specific HRC's will be selected by the experts of the VQEG's Independent

Labs and Selection Committee (ILSC) and should ensure broad coverage of typical content (spatial detail, motion complexity, color, etc.) and typical video processing conditions. The breadth of the test set will determine how well the generalizability of the models is tested. At least 20 different scenes are recommended as a minimum set of test sequences. It is suggested that some quantitative measures (e.g., criticality, spatial and temporal energy) are used in the selection of the test sequences to verify the diversity of the test set.

4.5 Complexity

The performance of a model as measured by the above Metrics #1-7 will be used as the primary basis for model recommendation. If several models are similar in performance, then the VQEG may choose to take model complexity into account in formulating their recommendations if the intended application has a requirement for minimum complexity. The VQEG will define the complexity criteria if and when required.

5 Recommendation decision

The VQEG will recommend methods of objective video quality assessment based on the primary evaluation metrics defined in section 4.3 The final decision(s) on ITU Recommendations will be made by the Study Groups involved: ITU-T SG 12, ITU-T SG 9, and ITU-R SG 11.

It is expected that an important measure of model acceptability, and the strength of the recommendation, will be the relative comparison of model rating errors compared to rating errors between different groups of subjective viewers. The selection procedure will require subjective rating cross-correlation data from the DSCQS experiments to estimate individual and population rating variances. This may require both duplication of sequences across different subjective testing labs and duplication of sequences within any one subjective test experiment.

If the metrics of section 4.3 are insufficient for developing a recommendation, then model complexity may be used as a further criterion for evaluation. The preference is one recommended model, but multiple models are possible. If the VQEG judges that a significantly improved recommended model can be developed from some combination of the proposed objective quality models, then this activity falls outside the scope of this plan and the VQEG may charter a follow-on task to address this activity.

Annex 1

Objective Video Quality Model Attributes

Section **4** presents several important attributes, and supporting metrics, that relate to an objective quality model's ability to predict a viewer's rating of the difference between two video sequences. This annex provides further background on the nature of these attributes, and serves as a guide to the selection of metrics appropriate for measuring each attribute. The discussion is in terms of the relation between the subjective DMOS data and the model's transformed $DMOS_p$ data. The schematic data and lines are not real, but idealized examples only meant to illustrate the discussion. In the interest of clarity, only a few points are used to illustrate the relationship between objective $DMOS_p$ and subjective DMOS, and error bars on the subjective DMOS data are left out.

Attribute1: Prediction Accuracy

This attribute is simply the ability of the model to predict the viewers' DMOS ratings with a minimum error "on average". The model in Figure 2 is seen to have a lower average error between $DMOS_p$ and DMOS than the model in Figure 3, and has therefore greater prediction accuracy.

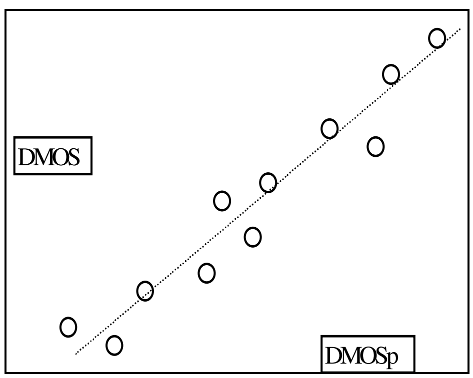


Figure 2. Model with greater accuracy

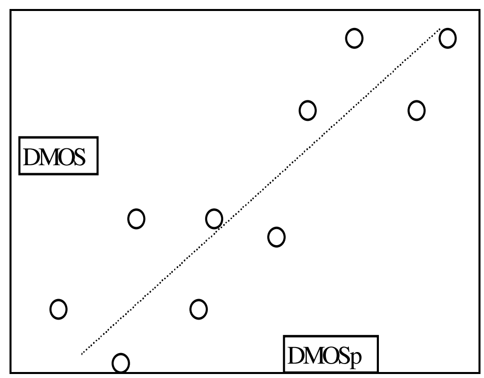
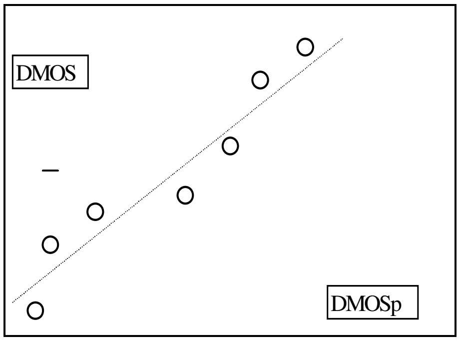


Figure 3. Model with lower accuracy

A number of metrics can be used to measure the average error, with root-mean-square (RMS) error being a common one. In order to incorporate the known variance in subjective DMOS data, the simple RMS error can also be weighted by the confidence intervals for the mean DMOS data points. The Pearson linear correlation coefficient, although not a direct measure of average error magnitude, is another common metric that is related to the average error in that lower average errors lead to higher values of the correlation coefficient.

Attribute2: Prediction Monotonicity

An objective model's $DMOS_p$ values should ideally be completely monotonic in their relationship to the matching DMOS values. The model should predict a change in $DMOS_p$ that has the same sign as the change in DMOS. Figures 4 and 5 below illustrate hypothetical relationships between $DMOS_p$ and DMOS for two models of varying monotonicity. Both relationships have approximately the same prediction accuracy in terms of RMS error, but the model of Figure 4 has predictions that monotonically increase. The model in Figure 5 is less monotonic and falsely predicts a decrease in $DMOS_p$ for a case in which viewers actually see an increase in DMOS.



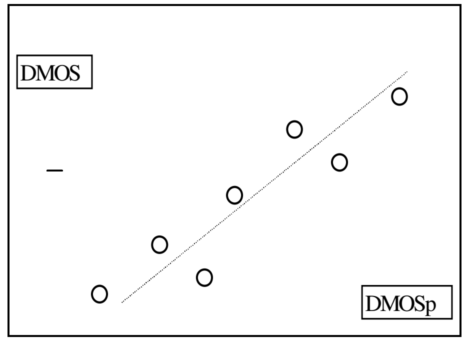
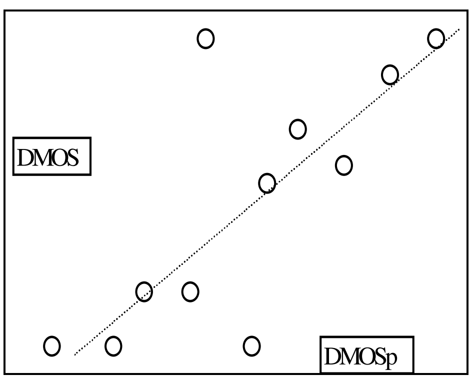


Figure 4. Model with more Monotonicity **Figure 5. Model with less Monotonicity**

The Spearman rank-order correlation between $DMOS_p$ and DMOS is a sensitive measure of Monotonicity. It also has the added benefit that it is a nonparametric test that makes no assumptions about the form of the relationship (linear, polynomial, etc.). Another method to understand model Monotonicity is to perform pair-wise comparisons on HRC's by type of sequence, bitrate, and any other parameters defining an HRC). The change between the pairs in DMOS should correlate with the change in $DMOS_p$.

Attribute3: Prediction Consistency

This attribute relates to the objective quality model's ability to provide consistently accurate predictions for all types of video sequences and not fail excessively for a subset of sequences.



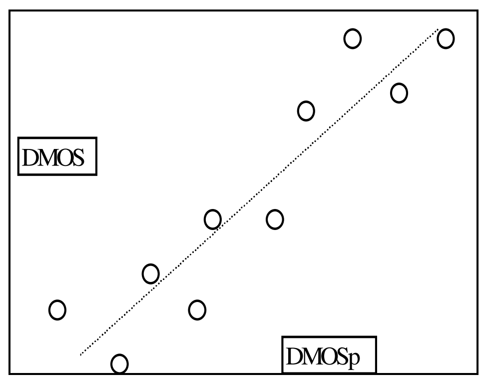


Fig. 6. Model with large outlying errors **Fig. 7. Model with consistent errors**

Figures 6 and 7 show models with approximately equal RMS errors between predicted and measured DMOS. Figure 6 is an example of a model that has quite accurate predictions for the majority of sequences but has large prediction error for the two points in the middle of the figure. Figure 7 is an example of a model that has a balanced set of prediction errors - it is not as accurate as the model of Figure 6 for most of the sequences but it performs "consistently" by providing reasonable predictions for all the sequences. The model's prediction consistency can be measured by the number of outlier points (defined as having an error greater than a given threshold such as one confidence interval) as a fraction of the total number of points. A smaller outlier fraction means the model's predictions are more consistent. Another metric that relates to consistency is Kurtosis, which is a dimensionless quantity that relates only to the shape of the error distribution and not to the distribution's width. Two models may have identical RMS error, but the model with an error distribution having larger "tails" to the distribution will have a greater Kurtosis.

Annex 2

VQEG Objective Video Quality Ad-hoc Group Members

Co-chairs:

Ravel, Mihir <mihir.ravel@tek.com>

Beerends, John <j.g.beerends@research.kpn.com>

Members:

Webster, Arthur <awebster@its.bldrdoc.gov>

Corriveau, Phil <philc@dgbt.doc.ca>

Hamada, Takahiro <ta-hamada@kdd.co.jp>

Brill, Michael <mbrill@sarnoff.com>

Winkler, Stefan <winkler@ltssg3.epfl.ch>

Pefferkorn, Stephane <stephane.pefferkorn@cnet.francetelecom.fr>

Contin, Laura <laura.contin@cselt.stet.it>

Pascal, Dominique <dominique.pascal@cnet.francetelecom.fr>

Zou,William <wzou@gi.com>.

Morton, Al <acmorton@att.com >

Fibush, David <davef@tv.tv.tek.com>

Wolf, Steve <steve@its.bldrdoc.gov >

Schertz, Alexander <schertz@dav.irt.de>

Fenimore, Charles <fenimore@eeel.nist.gov>

Libert, John <libert@eeel.nist.gov>

Watson, Andrew <abwatson@mail.arc.nasa.gov>