# VQEG Validation and ITU Standardisation of Objective Perceptual Video Quality Metrics

Kjell Brunnström, David Hands, Filippo Speranza and Arthur Webster

For industry, the need to access accurate and reliable objective video metrics has become more pressing with the advent of new video applications and services such as mobile broadcasting, internet video, and internet protocol television (IPTV). Industry-class objective quality measurement models have a wide range of uses, including equipment testing (e.g. codec evaluation), transmission planning and network dimensioning tasks, head-end quality assurance, in-service network monitoring, and client-based quality measurement. The Video Quality Experts Group (VQEG) is the primary forum for validation testing of objective perceptual quality models. The work of VQEG has resulted in International Telecommunication Union (ITU) standardisation of objective quality models designed for standard definition television and for multimedia applications. This article reviews the work of VQEG with particular consideration of the group's approach to validation testing.

## BACKGROUND

### MOTIVATION

The VQEG was founded in 1997 by a small group of experts in subjective and objective video quality from ITU-T and ITU-R Study Groups. The general goal of VQEG is to advance the field of video quality assessment by investigating new and advanced subjective assessment methods and objective quality metrics and measurement techniques.

### OBJECTIVES

VQEG aims at providing a forum where algorithm developers and industry users can meet to plan and execute validation tests of objective perceptual quality metrics. VQEG applies a systematic approach to validation testing that typically includes the collection of several subjective databases whose results are to be predicted by the objective video quality models under examination. An important facet of the VQEG approach is the formulation of test plans that specifically define the procedures for performing objective model validation. These test plans describe the format of source content, the nature of degradations that may be applied to the content, the subjective methods to be used to collect the subjective data, the test laboratories who perform the subjective assessment tests, the type of objective quality models that may be submitted, the submission procedures for objective quality models, and the statistical techniques and model evaluation metrics to be used. Importantly, the test plans are approved by consensus among all VQEG participants — including model proponents, subjective test laboratories, industry representatives, academia, and representatives of several Standards Developing Organisations.

### ISSUING BODIES AND SCHEDULE

Once a validation test has been completed, VQEG submits a final report to the ITU, which is ultimately responsible for preparing new standards for objective perceptual quality measurement.

To date, VQEG has completed three validation tests. The first two tests, called VQEG Full-Reference Television Phase I (FRTV- I) and Phase II (FRTV- II), covered quality measurement of standard definition television services using full-reference models[1]. The first test, FRTV- I [1], was completed in 2000. None of the models tested outperformed Peak Signal-to-Noise Ratio (PSNR), which is the performance benchmark against which the ITU has tended to make decisions on standardising objective models. Accordingly, the initial standard, published by the ITU-T Study Group 9 as Recommendation J.144 [2], included only informative appendices detailing objective models. The second test, FRTV-II [3], was completed in 2003. At the end of this validation effort, the ITU-T published an updated version of Recommendation J.144 in which four objective models were included as standardised objective perceptual quality measurement methods. Scatterplots illustrating the predictive performance of two of these methods are shown in Fig. 1. A full functional description of each model is included in a normative annex to the standard. In addition to their publication in ITU-T Rec. J.144 which applies to cable television services, a mirror standard was published for baseband television services by ITU-R Study Group 6 as Recommendation BT.1683 [4].

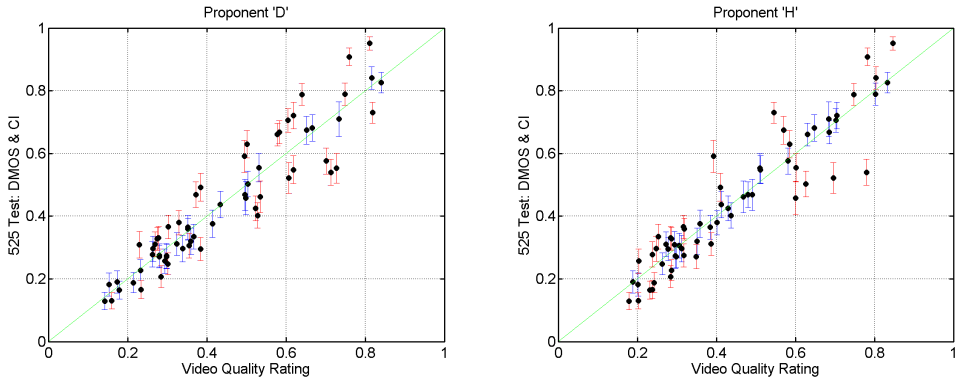



Fig. 1: Scatterplots showing predictive performance of BT's ('Proponent 'D') and NTIA/ITS's (Proponent 'H') objective models. The subjective score is computed using difference mean opinion scores (DMOS), the confidence intervals are also shown. The objective scores are shown on the axis labelled 'Video Quality Rating'. From VQEG FRTV-II final report [3].

[1] Full-reference methods require full access to both the original source sequence and its processed counterpart. They are appropriate for performance testing where time to measure quality is not critical and source video is available. Reduced-reference methods operate by extracting a parameter set from the original reference sequence, and using this set in place of the actual reference video. Some means of transmitting the reference parameters for use with the reduced-reference method is required. No-reference methods operate only on the processed sequence and have no access to source information. Reduced-reference and no-reference methods are appropriate for live monitoring applications.

The third and most recent validation effort was aimed at evaluating objective perceptual quality models suitable for digital video quality measurement in multimedia applications. This project, VQEG Multimedia Phase I (MM-I), was completed in 2008 [5]. Although this validation effort was limited to video only (a second phase concerning both audio and video quality is planned for the near future), it is perhaps the most exhaustive validation test ever performed. A later section below provides more detail on the design and implementation of the MM-I subjective tests. The (MM-I) set of tests were used to validate full-reference, reduced-reference and no-reference objective models.

The VQEG Multimedia Phase I Final Report was completed in March 2008 and ITU-T Study Group 9 has subsequently published two new standards based on that report. ITU-T Rec. J.247 [6] defines four new full-reference objective quality methods for multimedia and ITU-T Rec. J.246 [7] defines one new reduced-reference objective quality measurement method for multimedia.

## TARGET APPLICATIONS

The VQEG reports and associated ITU standards cover both television and multimedia applications. The standard definition standards (ITU-T Rec. J.144, ITU-R Rec. BT.1683) are confined to objective measurement of MPEG-2 encoded 525-line and 625-line television services and are limited to full-reference measurement methods.

The J.247 full-reference and J.246 reduced-reference multimedia standards have been designed for telecommunications services delivered at 4 Mbit/s or less. These standards focus on broadband internet and mobile or PDA video services, which cover applications including videoconferencing, internet and mobile television, and video streaming.

## VQEG MULTIMEDIA VALIDATION TESTING

The VQEG multimedia validation tests, as specified in the test plan [8], examined the performance of objective perceptual quality models for three different video formats: Video Graphics Array (VGA, 640 x 480 pixels picture resolution), Common Intermediate Format (CIF, 352 x 288 pixels) and Quarter Common Intermediate Format (QCIF, 176 x 144 pixels). The validated objective models included full-reference, reduced-reference, and no-reference models submitted by several proponent organisations.

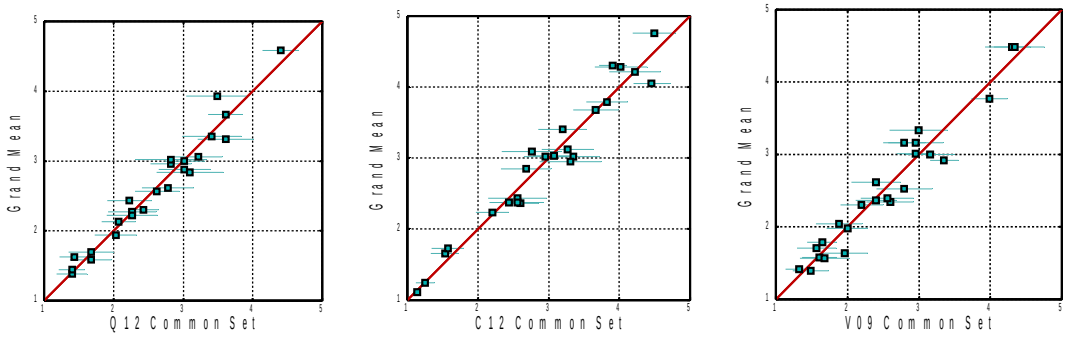
To evaluate the predictive performance of these models, a large number of subjective assessment tests were performed at each of the three video formats: in total 13 VGA, 14 CIF and 14 QCIF tests were completed. The subjective tests were performed by 13 laboratories from 11 different countries in 3 continents. Each test laboratory ran between one and three subjective tests. The tests were conducted in the native language of the test laboratory. Each subjective test included exactly 166 video sequences. Included in the 166 video sequences was a set of 30 sequences which was common to all subjective tests performed with the same video format. The common set allowed researchers to measure the agreement between the subjective data collected by different laboratories. The remaining 136 test sequences differed between

subjective tests. The processed video sequences had been produced in accordance with the guidelines and procedures defined in the test plan which mandated the allowable video codecs, compression levels, frame rates, transmission error levels, and so on.

Subjective video quality was assessed using a single-stimulus presentation method and the Absolute Category Rating (ACR) scale (see ITU-T Rec. P.910) [9]. In this method, the test video sequences are presented one at a time and rated independently on the ITU five-grade quality scale.

The subjective tests included the reference (i.e., unprocessed source) and the processed versions of the reference. The reference sequences were not identified as such to the viewers (hidden reference approach). This ACR method with hidden reference was recently included in a revised version of ITU-T Rec. P.910. The inclusion of the reference video source sequences permitted computation of two types of subjective scores for data analysis: a Mean Opinion Score (MOS) and a Difference Mean Opinion Score (DMOS). The MOS was computed as the average of the absolute ratings obtained for each processed video sequence. The DMOS was computed as the average of the arithmetic difference between the ratings given to the processed video sequence and those given to the corresponding reference video sequence. This latter procedure is known as ACR with hidden reference removal. MOS data were used to evaluate no-reference models, whereas DMOS data were used to evaluate full-reference and reduced-reference models.

The purpose of the subjective tests was to validate objective methods. However, given the scope of the testing it was of interest to investigate the cross-laboratory variation in subjective scores. The insertion of a common set of test sequences was agreed precisely to allow for this comparative analysis. Given the number of tests, picture resolutions, and language differences, it is reassuring to note that the overall correlation in subjective scores for the common set between laboratories was 0.94 for QCIF, 0.94 for CIF and 0.95 for VGA. Fig. 2 shows the scatterplots of the common set of sequences from subjective tests in which the authors were directly involved. The plots show a high degree of consistency in subjective scores between the laboratories. The consistency in cross-laboratory subjective scores is very impressive and provides significant empirical evidence for the reliability of the selected test method.




(a) QCIF (b) CIF (c) VGA

Fig. 2: Scatterplots showing the correlation between the common set of test sequences and the grand mean of these sequences across all tests in a given resolution.

## ASSESSING MODEL PERFORMANCE

For objective quality measurement, there are three aspects to performance: prediction accuracy (i.e., accurate prediction of the subjective MOS of each sequence), computational requirements and run time. The VQEG validation testing does not set requirements with respect to model efficiency. The output from VQEG does not differentiate on the basis of computational requirements or run-time footprint, but only on prediction accuracy. Prediction accuracy is determined by VQEG using three evaluation metrics: Pearson's product-moment correlation co-efficient, root mean square error (rmse) and outlier ratio (see [5] for full details of these evaluation metrics). The F-test statistic [10] is used to differentiate prediction accuracy between models as well as to compare the objective perceptual model performance with that of PSNR.

## FURTHER TECHNICAL DEVELOPMENTS

VQEG has a number of current activities running in parallel. The first project, Reduced-Reference and No-Reference Television (RRNR-TV) project will validate reduced-reference and no-reference objective models for standard definition television. This project complements the original FRTV validation effort which also involved standard definition television. The RRNR-TV project is in its final stages; subjective tests will be completed before the end of 2008 and results are expected to be available in early 2009. The final report will be published in the spring of 2009.

The second project will examine objective models capable of predicting the subjective quality of high definition television. The High Definition Television (HD-TV) validation test will consider full-reference, reduced-reference, and no-reference objective models and focus on objective assessment of secondary distribution video (i.e., video delivered to the home). The test plan for the HD-TV validation test is currently being discussed. Several critical decisions have yet to be made, such as the video formats to be covered in the test, the display technology to use, and the range of error conditions to be included in the test.

Thus far, VQEG has examined only models that consider what is seen by the viewer and thus operate on the decoded video data. Recently a new class of objective models has been proposed, which attempts to measure video quality using information obtained directly from the bitstream. The third VQEG project, termed HYBRID-TV, will evaluate objective models capable of using either one or both of two sources of information: decoded video data and bitstream information. The VQEG Hybrid ad-hoc group is working on defining a test plan that will form the basis for validating objective quality models that can use information obtained from analysis of the packet header, bitstream payload, and decoded picture. This activity is closely associated with ITU-T Study Group 12 projects and is coordinated within the ITU-T Joint Rapporteur Group on Multimedia Quality Assessment (JRG-MMQA).

Finally, VQEG is planning a second phase of the multimedia project, MM-II. As discussed above, VQEG has just completed phase I of the multimedia project, but that project was confined solely to measuring video quality. The second phase of the multimedia project will examine objective quality models that can predict audio-

visual quality; that is, models that can predict both video and audio qualities as well as their interaction.

## SUMMARY

Standardisation of objective quality models has made great progress since VQEG was formed in 1997. Based on the validation tests performed by VQEG, four international standards have been published. In addition, VQEG has also been instrumental in providing a forum for discussions and developments surrounding different facets of quality measurement and assessment. Based on the work of VQEG, new statistical tools for evaluating the performance of objective methods have been proposed, tested, and adopted (e.g., ITU-T Rec. J.149). Subjective test methodologies have been critically examined and modifications to these methods proposed, assessed and implemented.

VQEG has already provided academic, government and industry experts interested in video quality with a suite of tools for advancing their research. Test sequences have been provided along with associated subjective scores that enable researchers to train and test objective models (available from www.vqeg.org). The software that was used to manage all the subjective tests during the MM test is available (www.acreo.se/acrvqwin) [11]. Test plans and final reports provide detailed advice on myriad aspects of video quality assessment and measurement: designing formal subjective tests, selecting source content, introducing compression and transmission errors, evaluating the performance of objective methods, and much more.

Additional video processing tools and data analysis methods as well as improvements to subjective quality test methodologies will continue to be developed and the software and reports will be made freely available from the VQEG webpage (see below).

Recently, VQEG has also begun discussing the possibility of a joint effort to develop objective quality assessment models which combine the best parts of existing models. This opportunity is open to all interested organisations. This joint effort may lead to the establishment of a reference objective metric. VQEG will continue its work to progress knowledge and understanding of issues relating to video and multimedia quality of existing and future technologies such as 3DTV stereoscopic television.

## ACKNOWLEDGEMENTS

The authors would like to thank Margaret Pinson and Stephen Wolf of NTIA/ITS for providing some of the plots in this paper. Kjell Brunnström would like to thank VINNOVA (Swedish Governmental Agency for Innovation Systems) for financial support of his VQEG work.

## OBJECTIVE QUALITY MEASUREMENT RESOURCES

### VQEG RESOURCES

VQEG homepage, www.vqeg.org, has links to project testplans, meeting contributions and test materials. The VQEG web pages provide access to reports from

all completed VQEG validation tests, software tools, as well as details on subscribing to the VQEG reflector.

## ITU RESOURCES

The ITU homepage, www.itu.int, has links to all ITU-T and ITU-R publications. All four objective quality measurement standards are available from the ITU publications website. Standardised methods for performing subjective quality tests can be obtained from the ITU publications web pages. A number of standards documents relevant to the validation and standardisation of objective models have been published and are available from the ITU including recommendations for analysing the predictive performance of objective methods, and calibration methods.

## TUTORIALS

ITU-T (2004). *Tutorial: Objective Perceptual Assessment of Video Quality: Full Reference Television* (includes FRTV-I Test Plan and Final Report and FRTV-II Final Report), http://www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf.

## SUBJECTIVE TEST SOFTWARE RESOURCES

The software used to control and run the VQEG Multimedia tests is available from http://www.acreo.se/acrvqwin.

## REFERENCES

[1] VQEG Final Report of FR-TV Phase I Validation Test (2000). "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase I", Video Quality Experts Group (VQEG), http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI.

[2] ITU-T Rec. J.144 (2004). "Objective perceptual video quality measurement techniques for digital cable television in the presence of full-reference", International Telecommunication Union, Telecommunication Standardization Sector, Geneva, Switzerland.

[3] VQEG Final Report of FR-TV Phase II Validation Test (2003). "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II", Video Quality Experts Group (VQEG), http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII.

[4] ITU-R Rec. BT.1683 (2004). "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full-reference", International Telecommunication Union, Radiocommunication Sector, Geneva, Switzerland.

[5] VQEG Final Report of MM Phase I Validation Test (2008). "Final report from the Video Quality Experts Group on the validation of objective models of multimedia quality assessment, phase I", Video Quality Experts Group (VQEG), http://www.its.bldrdoc.gov/vqeg/projects/multimedia.

[6] ITU-T Rec. J.247 (2008). “Objective perceptual multimedia video quality measurement in the presence of a full-reference”, International Telecommunication Union, Telecommunication Standardization Sector, Geneva, Switzerland.

[7] ITU-T Rec. J.246 (2008). “Perceptual audiovisual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference”, International Telecommunication Union, Telecommunication Standardization Sector, Geneva, Switzerland.

[8] VQEG Multimedia Project (2006). *Multimedia Group Testplan,* Video Quality Experts Group, www.vqeg.org (ftp://vqeg.its.bldrdoc.gov/Documents/Projects/multimedia/MM_new_testplan_v1.21_changes_highlighted.doc ).

[9] ITU-T Rec. P.910 (2008). “Subjective video quality assessment methods for multimedia applications”, International Telecommunication Union, Telecommunication Standardization Sector, Geneva, Switzerland.

[10] Spiegel, M. R. and Stephens, L. J. (1998). *Theory and Problems of Statistics (3rd ed.),* Schaum's Outline Series, New York: McGraw-Hill.

[11] Jonsson, J., Brunnström, K. (2007). *Getting started with ArcVQWin,* (acr022250), Acreo AB, Kista, Sweden.

AUTHORS

Kjell Brunnström (kjell.brunnstrom@acreo.se) is a senior scientist at Acreo AB, Sweden, with a research interest in video and display quality assessment. Kjell Brunnstrom is co-chair of the VQEG Multimedia project and co-chair of VQEG ILG.

David Hands (david.2.hands@bt.com) is a research group leader with BT Innovate, UK. David Hands was formerly co-chair of the VQEG Multimedia project.

Filippo Speranza (Filippo.Speranza@crc.ca) is a research scientist at the Communications Research Centre, Canada, specialising in human visual perception, stereoscopic imaging, and subjective picture quality assessment techniques and methods. Filippo Speranza is co-chair of VQEG.

Arthur Webster (webster@its.bldrdoc.gov) is a lead electronics engineer at NTIA/ITS, USA, and works on the development and standardisation of video and multimedia quality assessment methods. Arthur Webster is founder and co-chair of VQEG.