The History of Video Quality Model Validation

Margaret H. Pinson [#1], Nicolas Staelens [*2], Arthur Webster [#3]

[#] *Institute for Telecommunication Sciences (ITS), National Telecommunications and Information Administration (NTIA), U.S. Department of Commerce (DOC), 325 Broadway St., Boulder, CO, USA*

[1] mpinson@its.bldrdoc.gov

[3] webster@its.bldrdoc.gov

[*] *Ghent University – iMinds,*
G. Crommenlaan 8 bus 201, 9050 Ghent, Belgium

[2] nicolas.staelens@intec.ugent.be

Abstract—**This paper describes objective video quality validation efforts conducted in the past two decades. Validation efforts to be examined include a validation test performed by the T1A1 committee in the early 1990's; five rounds of validation testing performed by the Video Quality Experts Group; and validation tests performed by ITU-T Study Group 12. Useful products that resulted from those efforts will be identified, including standards, datasets, and model validation techniques.**

I. Introduction

Shortly after the advent of digital video codecs, there arose a need for objective video quality models that could predict the quality of digitally encoded video. Because this has proved to be a difficult problem, a series of validation tests have been conducted. The goal was to identify objective measures of video performance, to compare the objective measures with user opinion of video quality, and to select from the candidate measures those that were well correlated with user opinion. This is necessary to demonstrate the accuracy of objective models to an uncertain consumer market.

This paper summarizes validation tests that have been performed by Standards Developing Organizations (SDOs) or by others in support of SDOs. Tables I and II summarize the design and outcome of each validation test. Table III summarizes the standards, standards documents, and subjectively rated datasets that have been made available as a result of these validation tests. This document presents a summary of each validation test, followed by some lessons learned in the process.

Some validation datasets are available for research and development purposes (see Table III). Objective models that are trained on these datasets must not, however, be compared to the models submitted for independent validation. Such a comparison is misleading, because the experiments contain primarily source scenes and systems that were unknown to the model developers at that time.

II. Validation Tests

A. T1A1

The first large scale test designed to validate objective measures of digital video quality was executed in 1994–95 by an SDO, the American National Standards Association (ANSI) accredited Alliance for Telecommunications Industry Solutions (ATIS).[1] A subcommittee of ATIS, called T1A1.5 at that time,[2] consisted of around 30 telecommunications and television engineers. The T1A1 validation test became a model for validation tests conducted during the next two decades.

The T1A1 validation test focused primarily on videoconferencing scenes and systems. Subjective tests were conducted using paper score sheets and Betacam SP tapes (i.e., component analog professional video tapes). Standard definition video was evaluated using a variety of algorithmic approaches, including full reference (FR) parameters, reduced reference (RR) parameters, and specialized video test patterns.

The resulting ANSI Standard T1.801.03 did not standardize a model for predicting mean opinion scores (MOS). Instead, it standardized 13 RR parameters submitted by NTIA [3] that could be used to build such a model (see T1A1 Contributions [2] and [3]). All other parameters and methods were withdrawn. These parameters were removed from T1.801.03 when it was revised in 2003.

Another standard resulting from this first large scale test is ANSI Standard T1.801.01, which specifies a collection of 22 source video sequences (SRC) for future objective and subjective assessment of videoconferencing systems. These SRC are in the public domain. The entire T1A1 dataset is available for research and development purposes in the Consumer Digital Video Library (CDVL, www.cdvl.org, [4]).

B. The Role of the Video Quality Experts Group

The Video Quality Experts Group (VQEG, www.vqeg.org) was born from a need to bring together international experts in subjective video quality assessment and objective quality measurement. The first VQEG meeting, held in Turin in 1997, was attended by a small group of experts drawn from

1 ATIS is a North American organization whose member companies work together to develop and propose telecommunications standards to ANSI and international SDOs.

2 ATIS Committee T1A1 is now called PRQC, which stands for Network Performance, Reliability and Quality of Service Committee.

3 These parameters formed the basis of later research that led to the development of the NTIA General Model, "VQM" [1]

MMSP'13, Sept. 30-Oct. 2, 2013, Pula (Sardinia), Italy.
978-1-4799-0125-8/13/$31.00 ©2013 IEEE.

participants in the International Telecommunication Union Telecommunication Standardization Sector (ITU-T) and Radiocommunications Sector (ITU-R) Study Groups. The general motivation of VQEG was and is to advance the field of video quality assessment by investigating subjective assessment methods and objective quality measurement techniques.

For the first decade, VQEG concentrated its efforts on the validation of new objective quality metrics for standardization purposes. Reports for each VQEG validation effort are available at [5]. More recently, VQEG efforts have also included collaborative research efforts and new subjective methods that are outside of the scope of this paper (e.g., [6]–[8]).

C. Full Reference Television Phase I

During 1999 and 2000, VQEG conducted its first validation experiment, the full reference television (FRTV) Phase I test. This test examined FR and no reference (NR) objective video quality models that predicted the quality of standard definition television. All NR models were withdrawn.

The subjective tests were designed and conducted by the Independent Lab Group (ILG) after model submission. Unfortunately, each dataset spanned a narrow range of quality, which made it difficult to detect differences between the submitted models. VQEG concluded that no model was statistically better than Peak Signal to Noise Ratio (PSNR) and, generally, no model was statistically better than the rest.

Despite that, the test was a success. Perhaps the most important achievement was the collection and redistribution of this important dataset. The FRTV Phase I subjective data are included in the final report, and all video sequences are available at [9]. ITU-T Recommendation J.144 (2001) was approved by ITU-T Study Group 9 (SG9) with eight FR models described in non-normative Appendices.

D. Full Reference Television Phase II

VQEG's FRTV Phase II test addressed the design flaws of the FRTV Phase I effort. Conducted from 2002 to 2003, FRTV Phase II examined the performance of FR and NR models on standard definition video. Subjective tests were designed and conducted by the ILG after model submission but the model proponents provided comments on the source scene pool and the levels of impairments to include in the test. The goal was to avoid subjective test design flaws that might cause problems similar to those seen in FRTV Phase I.

FR models from BT, Yonsei University, Centro de Pesquisa e Desenvolvimento (CPqD), and NTIA were standardized for use with standard definition video by ITU-T SG9 in Revised ITU-T Rec. J.144 (2004) and ITU-R Rec. BT.1683. Both versions of J.144 are available for free (as are all ITU Recommendations) on the ITU website. All NR models were withdrawn. Unfortunately, the datasets cannot be distributed due to licensing restrictions on the source video.

E. Multimedia Phase I

VQEG's Multimedia (MM) Phase I examined the performance of FR, RR, and NR video quality models for multimedia resolutions (i.e., VGA, CIF and QCIF video without audio). The MM models were analyzed per resolution (see Table I and II). These tests were conducted in 2007 and 2008 under the direction of the ILG, with proponents doing some of the necessary work. A total of thirteen organizations performed subjective testing on 41 datasets using 5320 processed video sequences (PVSs). This was the largest video quality test ever performed—at least up until that time. None of the datasets can be redistributed due to a multiple party non-disclosure agreement among participants.

Based on the results of the VQEG MM1 test, two standards were approved by the ITU-T SG9 and two by the ITU-R Working Party 6C. FR models from Nippon Telegraph and Telephone Corporation (NTT), OPTICOM, Psytechnics, and Yonsei University were standardized in ITU-T Rec. J.247 and ITU-R BT.1866; and an RR model from Yonsei University was standardized in ITU-T Rec. J.246 and ITU-R BT.1867. The ITU decided not to standardize an NR model because the performance of NR models in this test did not warrant standardization.

F. Reduced Reference/No Reference Television

VQEG's reduced reference/no reference (RRNR-TV) Phase I test examined the performance of RR and NR models on standard definition video. These tests were conducted in 2008 and 2009 under the direction of the ILG, with proponents doing some of the necessary work. RR models from Yonsei University, NEC, and NTIA were standardized by ITU-T SG9 in ITU-T Rec. J.249. All NR models were withdrawn. These two datasets cannot be redistributed due to licensing restrictions on the SRCs.

The original intention of the RRNR-TV test was to conduct the subjective evaluation using single stimulus continuous quality evaluation (SSCQE), so that the models would continuously predict quality when applied in-service. This subjective method was dropped due to the complexities required for model evaluation using the SSCQE data.

G. High Definition Television

VQEG's high definition television (HDTV) project examined the performance of FR, RR, and NR models for HDTV during 2009 and 2010. These tests were conducted entirely by the ILG, with most of the work being performed before model submission. This resulted in model submission deadline delays; however the time between model submission and final report was greatly improved.

An FR model by SwissQual was standardized by ITU-T SG9 in J.341 and an RR model by Yonsei University was standardized in J.342. Two NR models appear in the VQEG final report of the HDTV test; however neither was standardized by the ITU. The HDTV Phase I subjective and objective data are available in the final report. The video sequences for five of the six experiments are available for research and development purposes on the CDVL.

TABLE I
VALIDATION TEST DESIGN (TOP) AND COLUMN DESCRIPTIONS (BOTTOM)

	Org.	Dates	Resolution	PVSs	Common	SLabs	Tests	SRC_i	HRC_i	$PVSs_i$	$Viewers_i$
T1A1	ATIS	1994-1995	NTSC	625	Yes	3	1	25	25	625	27
FRTV Phase I	VQEG	1999-2000	NTSC & PAL	360	No	8	4	10	9	90	71-80
FRTV Phase II	VQEG	2002-2003	NTSC & PAL	110	No	3	2	13	10 or 14	64	27, 66
Multimedia	VQEG	2007-2008	VGA CIF QCIF	1688 1816 1816	Yes Yes Yes	8 9 8	13 14 14	8 8 8	16 16 16	128 128 128	24 24 24
RRNR-TV Phase I	VQEG	2008-2009	NTSC & PAL	312	No	4	2	12	34	156	32
HDTV	VQEG	2009-2010	1080i 59.94 & 50fps 1080p 29.97 & 25fps	830	Yes	6	6	9	15	135	24
P.NAMS/NBAMS LR	SG12	2011-2012	HVGA, QVGA, QCIF	792	Yes	3	4	8	29 or 17	232 or 156	24
P.NAMS/NBAMS HR	SG12	2011-2012	NTSC, PAL, 720p, 1080p, 1080i	1872	Yes	6	8	8	29	232	24

Org	Dates	Resolution	PVSs	Common	Slabs	Tests	SRC_i	HRC_i	$PVSs_i$	$Viewers_i$
Organization responsible	excludes planning	Video resolution	Total # PVSs in all tests	If some clips appeared in all tests	Total # subjective testing labs	Total # of video quality subjective tests	SRC in each test	Impairments in each test	PVSs in each test	Subjects rated each clip

Note 1: Column "PVSs" excludes the SRC and counts each common PVS once (e.g., if a common set of clips appeared in all tests).
Note 2: Column "HRC_i" excludes the ACR-HR hidden reference.
Note 3: Column "PVS_i" excludes the SRC and excludes common set sequences.
Note 4: The P.NAMS/NBAMS "PVSs" cells ignore the subdivision of clips by model scope.
Note 5: The P.NAMS audio and audiovisual quality test statistics are not reported here.
Note 6: All of the validation efforts included transmission errors.

TABLE II
VALIDATION TEST SCOPE AND RESULTS SUMMARY (TOP) AND COLUMN DESCRIPTIONS (BOTTOM)

	HRC Types	Birates	Subjective Method	Lab-to-Lab Correlation	Max Model ρ
T1A1	H.261, vector quantization, VHS, Proprietary codecs	70.4 kb/s to 45 Mb/s	BT.500-5 DSIS	0.926-0.958	0.805
FRTV Phase I	MPEG-2 (main profile at main level), H.263, multi-generation Betacam SP	768 kb/s to 19 Mb/s	BT.500-8 DSCQS	0.727-0.950	0.827
FRTV Phase II	MPEG-2, H.263	768 kb/s to 5 Mb/s	BT.500-10 DSCQS	0.97	0.912
Multimedia	H.264/AVC, MPEG-2, VC1, RV10, MPEG-4, SVC, WMV, JPEG2000	128 to 4000 kb/s 64 to 704 kb/s 16 to 320 kb/s	ACR-HR P.910 (modified)	0.953-0.996 0.939-0.990 0.943-0.982	0.822 VGA 0.836 CIF 0.841 QCIF
RRNR-TV Phase I	MPEG-2, H.264/AVC	1-5.5 Mb/s	ACR-HR P.910 (modified)	0.925, 0.954	0.901
HDTV	MPEG-2, H.264/AVC	1-30 Mb/s	ACR-HR P.910	0.924-0.990	0.87
P.NAMS LR	MPEG4 very simple profile (VSP), H.264/AVC baseline profile	32 kb/s to 6 Mb/s	ACR P.910		0.830
P.NAMS HR	H.264/AVC main and high profiles	500 kb/s to 30 Mb/s	ACR P.910		0.902
P.NBAMS LR	H.264/AVC baseline profile: x264	50 kb/s to 6 Mb/s	ACR P.910		0.918

HRC Types	Bitrates	Subjective Method	Lab-to-Lab Correlation	Max Model ρ
Types of impairments	Range of video encoding bitrates	Subjective testing method with ITU Recommendation	Pearson correlation between labs, if subjects at multiple labs rated identical video sequences	Maximum model Pearson correlation, using all tests

H. Hybrid Perceptual/Bitstream

VQEG's hybrid perceptual/bitstream (aka hybrid) validation test is currently validating objective video quality models that use both the PVS and bit-stream information. This test will examine WVGA/VGA video and also HDTV video. Hybrid NR, Hybrid RR, Hybrid FR, and NR models could be submitted. Models have been submitted and testing is underway.

I. P.NAMS and P.NBAMS

During 2011 and 2012, ITU-T Study Group 12 (SG12) validated two types of models:

- The non-intrusive parametric model for assessment of performance of multimedia streaming (P.NAMS) examined non-intrusive models for the evaluation of audio quality, video quality, and audiovisual quality based on IP protocol information embedded at the client (e.g., information from the local transport layer, information about the decoder).
- The parametric non-intrusive bitstream assessment of video streaming quality (P.NBAMS) examined non-intrusive models for the evaluation of video quality based on IP protocol and bitstream information (i.e., payload information) embedded at the client.

The term "non-intrusive" generally refers to models that do not require access to the source video as a reference point. An overlapping set of subjective tests were used to validate both types of models. Tables I and II include only information about the video quality subjective tests used for validation.

Independent labs were not available, so proponents conducted separate subjective tests and provided oversight to ensure a fair process. The models were analyzed separately on low resolution (LR) video for mobile applications and high resolution (HR) video for IPTV applications. Testing consisted of a competitive phase (to identify the group of top performing models) followed by an optimization phase (to merge the top performing models into a single model with equal or improved performance).

Based on the P.NAMS results, an LR model by NTT and Huawei was standardized in P.1201.1 and an HR model by Deutsche Telekom and Ericsson was standardized in P.1201.2. ITU-T P.1201 provides an overview of this type of non-intrusive monitoring. Based on the P.NBAMS results, an LR model by Technicolor, Ericsson, and Deutsche Telekom was standardized in P.1202.1. HR applications require further study. ITU-T P.1202 provides an overview for this type of non-intrusive monitoring.

P.NAMS tested LR and HR models against the x264 encoder [10], and LR models against the ffmpeg encoder [11]. P.NBAMS tested models against the x264 encoder [10]. ITU-T Rec. 1201 and 1202 say, "It is assumed that the model can be used for estimating quality when other encoder implementations for the given codec have been used. However, if the encoder performance is significantly worse or better than for the encoder used, the model prediction accuracy will be reduced."

III. Other Information

A. Dataset Availability

Table III identifies the validation datasets that are freely available. The early availability of the FRTV Phase I dataset made possible research into objective video quality metrics. Arguably, this was partly responsible for the success of later models and validation tests, since some proponents only had this dataset available for training.

The "no redistribution" limitation of the FRTV Phase II, MM and RRNR-TV datasets inspired personnel at NTIA/ITS, Intel and the University of California at Santa Barbara to develop a mechanism for the free redistribution of SRCs and datasets for research and development purposes. This led to the development of the CDVL, which went live in 2009. This resource provided SRCs for all later validation tests. By 2012, 27 subjectively rated image and video datasets were freely available (see Winkler [12]). Validation databases are particularly valuable, due to the careful scrutiny of the test design and the participation of multiple laboratories.

B. Analyzing Multiple Subjective Datasets Simultaneously

Validation tests have conflicting goals of maximizing the total number of PVS (which requires multiple subjective experiments) and calculating a single metric that analyzes model performance on all PVSs.

Some validation tests tried to solve this problem in the reporting of statistics. Examples include averaging Pearson correlation across multiple datasets, and counting the number of experiments for which a model appeared in the top performing group of models. These solutions were not entirely satisfactory from a statistical perspective.

Other validation tests solved this problem using sequences that appear in multiple datasets. The intention was to anchor the subjective ratings from individual experiments onto the same perceptual scale. The T1A1 test divided the viewers and hypothetical reference circuits (HRCs)[4] into three overlapping subsets. All of the PVSs and ratings were combined without modification (i.e., treated as having been drawn from the larger pool of available PVSs and subjects). Multimedia included a common set of video sequences in all experiments, but did not combine datasets. Using data from the Multimedia validation test, Pinson and Wolf [13] proposed an algorithm to map all of the experiments onto a single super-set of subjective data. The HDTV test used this technique to apply the statistical analysis to all PVSs at once. This increased the ability of statistical tests to differentiate between models.

Clause 7.8 of ITU-T Rec. P.1401 describes a statistical technique that combines performance metrics from multiple subjective experiments into an overall measurement. P.1401 weights databases by their importance, computes a statistical significant distance measure for each model on each experiment, and then aggregates the results.

[4] HRC is a fixed combination of a video encoder operating at a given bit-rate, network condition, and video decoder. Vendor names are omitted, because validation tests are not designed to analyze different codec implementations.

TABLE III
VALIDATION TEST OUTPUTS

	Dataset Availability	Standards Documents Produced
T1A1	SRC on CDVL (dataset "ANSI T1.801.01") Full dataset on CDVL (key word "T1A1")	TlA1.5/94-118 Subjective Test Plan T1A1 Test T1A1.5/94-l52 Data Analysis ANSI Standard: T1.801.01-1995 (R2001), Digital Transport of Video Teleconferencing/Video Telephony Signals—Video Test Scenes for Subjective and Objective Performance Assessment (1995) ANSI Standard: T1.801.02-1996 (R2011), Digital Transport of Video Teleconferencing/Video Telephony Signals—Performance Terms, Definitions, and Examples ANSI Standard: T1.801.03-1996 (R2008), Digital Transport of One-Way Video Signals—Parameters for Objective Performance Assessment
FRTV Phase I	Full dataset on VQEG website [9]	ITU-T Rec. J.144 (2001), Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference
FRTV Phase II	None	ITU-T Rec. J.144 (2004), Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference 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
Multimedia	None	ITU-T Rec. J.246 (2008), Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference ITU-T Rec. J.247 (2008), Objective perceptual multimedia video quality measurement in the presence of a full reference ITU-T Rec. J.340 (2010), Reference algorithm for computing peak signal to noise ratio of a processed video sequence with compensation for constant spatial shifts, constant temporal shift, and constant luminance gain and offset ITU-R Rec. BT.1866 (2010), Objective perceptual video quality measurement techniques for broadcasting applications using low definition television in the presence of a full reference signal ITU-R Rec. BT.1867 (2010), Objective perceptual visual quality measurement techniques for broadcasting applications using low definition television in the presence of a reduced bandwidth reference
RRNR-TV Phase I	None	ITU-T Rec. J.249 (2010), Perceptual video quality measurement techniques for digital cable television in the presence of a reduced reference ITU-T Rec. J.340 (2010), Reference algorithm for computing peak signal to noise ratio of a processed video sequence with compensation for constant spatial shifts, constant temporal shift, and constant luminance gain and offset
HDTV	5 tests on CDVL (dataset "VQEG Subjective Tests")	ITU-T Rec. J.341 (2011), Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference ITU-T Rec. J.342 (2011), Objective multimedia video quality measurement of HDTV for digital cable television in the presence of a reduced reference signal
P.NAMS	None	ITU-T Rec. P.1201 (2012), Parametric non-intrusive assessment of audiovisual media streaming quality ITU-T Rec. P.1201.1 (2012), Parametric non-intrusive assessment of audiovisual media streaming quality – lower resolution application area ITU-T Rec. P.1201.2 (2012), Parametric non-intrusive assessment of audiovisual media streaming quality – higher resolution application area
P.NBAMS	None	ITU-T Rec. P.1202 (2012), Parametric non-intrusive bitstream assessment of video media streaming quality ITU-T Rec. P.1202.1 (2012), Parametric non-intrusive bitstream assessment of video media streaming quality – lower resolution application area

C. *Minimum Acceptable Performance*

Once models have been analyzed, the question then arises, "how good is good enough?" When can a model be considered good enough for standardization?

While there is no easy answer to this question, PSNR serves as a pragmatic minimum performance benchmark. Although it is imperfect, no superior benchmark has yet been suggested. Generally speaking, an FR model must perform better (statistically) than PSNR. Models with limited access to the original video (e.g. RR models) must be at least as accurate as PSNR. It is an ongoing discussion whether NR models should be expected to perform as well as PSNR.

The form of PSNR used as a benchmark in the Multimedia, RRNR-TV, and Hybrid validation tests can be found in ITU-T Rec. J.340. A free implementation can be downloaded from [14]. This algorithm is optimized for accuracy, not speed, as it is intended to serve as an "idealized" benchmark.

D. *Analysis Metrics*

No single metric can analyze all facets of performance. Some metrics have been tried and discarded as redundant or imprecise. From a reporting standpoint, a key trait is the

ability to calculate statistical significance (e.g., are this model's results statistically better than PSNR?). The following three metrics provide a comprehensive model analysis:

1. *Root mean square error (RMSE)* measures accuracy and has the greatest discrimination capability (i.e., RMSE can better identify differences between models).[5]
2. *Outlier ratio* measures distribution consistency.
3. *Pearson correlation coefficient* measures linearity and yields an easily interpreted range of values (i.e., close to 1.0 is desirable). This is likely the reason for Pearson correlation's continued popularity, despite being closely related to RMSE and arguably redundant due to the trend of removing nonlinearities from objective video quality models prior to analysis.

In July 2012, SG12 approved ITU-T Rec. P.1401, which presents a framework for the statistical evaluation of objective quality algorithms regardless of the assessed media type. This Recommendation standardizes methods, metrics and procedures for statistical evaluation, qualification, and comparison of objective quality prediction models. It can be used to assess any objective model that predicts a subjective judgment of a subjective test procedure.

E. The Role of Proponents in Test Design

In all the validation tests discussed here, the model proponents played a vital role in the design and sometimes in the execution of a given test. While it may seem that it is preferable to have an independent group of labs design and conduct the tests (such as the VQEG's ILG), in fact the model developers are perhaps the most qualified to provide the expertise needed to develop a fair and balanced test. The ILG plays a vital role by ensuring that the tests are conducted according to ITU testing standards and are not biased toward any proponent. The ILG also ensures that the model that was submitted to the test is the same model that is used to provide data after the subjective tests are complete.

IV. Conclusion

For over 20 years quality experts have been designing objective methods for assessing video and audiovisual quality. The design and execution of fair and honest validation tests have allowed the community of telecommunications and television engineers to standardize methods for objectively assessing video and audiovisual quality for many applications, such as standard definition television, HDTV, and mobile video applications. A major challenge faced by validation efforts is the dependency upon unpaid assistance, which slows execution of the test. For related work, see QUALINET, ISO/IEC JTC1 (JPEG/MPEG) and [15].

To date, objective models for assessing video and audiovisual quality are not as accurate as well-designed subjective tests. However, objective methods are consistent, quicker, and less costly than subjective experiments. The objective methods continue to improve and it is expected that one day they will rival the accuracy of subjective methods. One challenge faced by objective models is how to integrate improved models of visual processes and cognition using artificial intelligence techniques such as object recognition.

Acknowledgment

The authors would like to acknowledge the work of the many participants of ATIS T1A1 (now known as PRQC), VQEG, ITU-T Study Groups 9 and 12 and ITU-R WP6C for contributing to the work on which this paper is based. It took a lot of dedicated work from a multitude of individuals to plan and execute the tests and to take the results forward to produce standardized test methods.

References

[1] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, Sep. 2004, pp. 312–322.

[2] S. Wolf, M. Pinson, C. Jones and A. Webster, "A summary of methods of measurement for objective video quality parameters based on the sobel filtered image and the motion difference image," ANSI T1A1 Contribution T1A1.5/93-152, Nov. 8, 1993. Available: http://www.its.bldrdoc.gov/n3/video/

[3] A. Webster, "Methods of measurement for two objective video quality parameters based on the Fourier transform," ANSI T1A1 Contribution T1A1.5/93-153, Nov. 8, 1993. Available: http://www.its.bldrdoc.gov/n3/video/

[4] M. H. Pinson, S. Wolf, N. Tripathi and C. Koh, "The consumer digital video library," *Proceedings of the Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Jan. 2010, pp. 1–6.

[5] Video Quality Experts Group (VQEG) Reports from Completed Validation Tests: http://www.its.bldrdoc.gov/vqeg/reports.aspx

[6] N. Staelens, I. Sedano, M. Barkowsky, L. Janowski, K. Brunnström and P. Le Callet, "Standardized toolchain and model development for video quality assessment - the mission of the joint effort group in VQEG," *Proceedings of the Third International Workshop on Quality of Multimedia Experience (QoMEX)*, Sept. 2011, pp. 61-66.

[7] M. Barkowsky, N. Staelens, L. Janowski, Y. Koudota, M. Leszczuk, M. Urvoy, P. Hummelbrunner, I. Sedano and K. Brunnström, "Subjective experiment dataset for joint development of hybrid video quality measurement algorithms," *EuroITV - 10th European Conference on Interactive TV*, Jul. 2012.

[8] M. H. Pinson, L. Janowski, R. Pépion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, Oct. 2012, pp. 640–651.

[9] ftp://vqeg.its.bldrdoc.gov/SDTV/VQEG_PhaseI/

[10] http://www.videolan.org/developers/x264.html

[11] http://www.ffmpeg.org

[12] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, Oct. 2012, pp. 616-625.

[13] M. H. Pinson and S. Wolf, *Techniques for Evaluating Objective Video Quality Models Using Overlapping Subjective Data Sets*, NTIA Technical Report TR-09-457, Nov. 2008. Available: http://www.its.bldrdoc.gov/n3/video/

[14] http://www.its.bldrdoc.gov/vqm/

[15] A. Raake *et al.*, "IP-Based mobile and fixed audiovisual media services," *IEEE Signal Processing Magazine*, vol. 28 no. 6, Nov. 2011.

[5] The results reported in "Comparison of Metrics VQEG MM Data," June 2008, by G. W. Cermak to the VQEG MM project, show that (1) correlation, RMSE, and outlier ratio all measure essentially the same thing, (2) RMSE is better at discriminating between models, and (3) the advantage of RMSE over correlation increases as the number of video samples decreases, and vice versa. These conclusions were also true for the VQEG FR-TV Phase 2 data.