

Perceptual Quality Measurement—Towards a More Efficient Process for Validating Objective Models

The Quality of Service Metrics (QoSM) Committee of the Alliance for Telecommunication Industry Standards (ATIS) Internet Protocol Television (IPTV) Interoperability Forum (IIF) is tasked with defining how objective quality metrics can provide meaningful IPTV performance measures. This group has reviewed current objective quality models as well as the processes by which such models are validated. This article describes current practices in validating objective quality models and presents a new, streamlined process that can be implemented to achieve more efficient and effective model validation. Of main interest for IPTV are models for predicting video and audiovisual quality; however, the process also applies to the validation of perceptual quality models (PQMs) for other modalities. The proposed process offers vendors a fast route to validating objective PQMs while providing industry with the assurance of independent, unbiased model evaluation.

BACKGROUND

Service providers are rolling out IPTV services to slow down erosion of revenues from circuit-switched voice services and to keep up with the competition to deliver multiplay service offerings. To support IPTV operations, the need for service performance measurements that can provide insights into the customer's perception of the quality of IPTV content is apparent. Vendors, standards groups, and researchers are actively investigating meaningful algorithms and tools for conducting these measurements.

Subjective quality tests are widely used and support the development and testing of objective perceptual quality models (or objective models) that predict customer perception as a benchmark. However, subjective quality tests are not a practical solution for in-service performance monitoring. The purpose of objective models is to replace subjective tests by estimating the perceptual quality of voice, audio, video, and multimedia. Objective models can use different techniques to predict subjective quality. These techniques include full-, reduced- and no-reference methods and may utilize pixel-domain, bit stream, packet

THE MAIN PREMISE IS THAT AN OBJECTIVE MODEL ALGORITHM DOES NOT NEED TO BE STANDARDIZED IN ITSELF, AS ITS PRIMARY REQUIREMENT IS MEASURING QUALITY WITH A CERTAIN LEVEL OF ACCURACY.

data, or some combination of these information sources to extract parameter values that then are used to predict quality [1]–[3]. The industry has an increased need for objective models as competition increases, and as quality becomes both a critical part of the value chain [e.g., high-definition TV (HDTV)] and a potential market differentiator between service providers.

The creation of objective models to compute an estimated customer opinion score is a complex process. Fundamental to the success of objective models is how accurately they can predict subjective quality ratings. A set of statistical methods

has been defined to determine the accuracy of objective models [9].

The accuracy of objective PQMs is currently validated through various routes, including self-validation, contracted external validation, and independent validation (e.g., by the Video Quality Experts Group (VQEG) [13]). Clearly, the industry will find great value in model accuracy data that is obtained through independent validation, as well as data that is based on appropriate subjective testing methods and model performance metrics.

This column considers the limitations of current validation procedures, such as those practiced by VQEG, ITU-T Study Group 9 (SG9), and ITU-T Study Group 12 (SG12), presents work in progress within relevant standards groups (in particular the ATIS IIF QoSM Committee) to address these problems, and outlines a proposal for providing more effective and efficient model validation.

CURRENT VALIDATION PROCESSES: VQEG AND ITU

VQEG [13], [14] has been central to coordinating efforts to perform independent validation of objective perceptual quality models in a competition-style process. VQEG has completed several phases of testing to date, and the model performance data obtained from these tests has been used by the ITU to produce international standards [7], [8], [10]–[12]. The VQEG process is based on voluntary contributions from government organizations, research centers, universities, and industry. For agreed projects, VQEG prepares a test plan, in collaboration with those who participate, that defines the scope of testing, the types of objective models that may be submitted, subjective test methods and test laboratories that

may perform subjective tests, model evaluation criteria, and so on.

The current VQEG process has the advantage of bringing together the premier experts in objective and subjective assessment to perform independent validation of objective models. Unfortunately, the relatively slow progress of VQEG projects means that the validation of models does not keep pace with industry requirements, and standardized models become outdated. The test plans often take several years to define, and once they are agreed upon, the test phase itself (including the accumulation of suitable test content, preparation of test sequences, and completion of subjective tests) is a lengthy process. After project completion, the best performing models may be standardized, and VQEG then moves on to the next project. The approach adopted by VQEG has the consequence that once a particular form of objective model has been validated and subsequently standardized, it may take many years before the group is able to perform a second validation test for that form of model. In fact, to date VQEG has not run a second validation round for any form of model. For example, full-reference TV (FR-TV) models were validated by VQEG in 2003 and standardized by the ITU in 2004. These models remain the standard so far as no further FR-TV validation tests have been performed, yet superior models may well have been developed in the meantime. The FR-TV test in particular did not include H.264 compression artifacts or IP loss impairments; consequently, the current standardized models have not been tested for the conditions that are present in most of today's IPTV systems.

Until recently, the VQEG process used Independent Test Laboratories (ITL) to perform subjective testing and model validation. More recently, VQEG has allowed model developers to act as test laboratories. This has led to a move away from cleanly separating the model development from the model validation. VQEG has begun working on an alternative process to validating "competing" models, having initiated a Joint Effort

Group (JEG) that will test dedicated model components with the goal of building a model that combines the best performing modules from different organizations. Similarly, ITU-T SG12 has started a series of collaborative

VQEG HAS BEEN CENTRAL TO THE INDEPENDENT VALIDATION OF PQMs. UNFORTUNATELY, ITS RELATIVELY SLOW PROGRESS DOES NOT KEEP PACE WITH INDUSTRY REQUIREMENTS, AND STANDARDIZED MODELS BECOME OUTDATED.

projects directed towards producing "best-of-breed" objective models. The approach taken by SG12 is to develop alternative objective models collaboratively that are then validated by the group. It should be noted that in the SG12 projects, many organizations that contribute objective models also perform the subjective tests used to validate the models and/or model components.

Reviewing the approaches of VQEG and SG12, several limitations in the current validation processes can be identified:

- Validating PQM models requires the acquisition of suitable multimedia content. Once this test material has been made available to model developers, it cannot be reused in future validation tests, requiring the selection and preparation of new content for subsequent tests.
- The current approaches (competition, collaboration, etc.) have strict cutoff dates for model submission, because all models are evaluated in the same exercise.
- Model developers are sometimes involved in the preparation of processed video sequences or in conducting subjective experiments due to ITL budget and time constraints, which is not ideal for an independent evaluation.
- At this time, the entire process for validating PQMs is very lengthy

and can take several years, because a new test plan is written and a new test library is created for every round of testing.

- Once a standard has been defined and approved, it is very difficult to change, which means that standardized models can quickly become outdated, and there is no process for the models or the standards to be updated in a prompt fashion.

A NEW VALIDATION PROCESS

The ATIS IIF QoS Committee has been working on a series of documents that form the backbone to validating objective models.

A general test plan for performing validation tests [4] was standardized to encourage industry developments where multiple organizations could develop PQMs all using the same basic test plan. With such a test plan in place, additional specialized documents, specific for each type of model, would then need to be developed that go into more detail for particular types of PQMs and applications. It is recommended that for each type of model, a single test plan is produced so that multiple organizations that want to test such a model all use the same procedures.

A technical report proposing a new process for validating objective models [5] has recently been completed. To date, standards groups combine the test process and test plan activities with the eventual goal of a standardized PQM solution. ATIS IIF separates these two processes. This column describes the concepts specified in the ATIS technical report.

Completing the series, a third document is planned that specifies the various types of perceptual quality measurements for use in IPTV environments [6]. The purpose of that document is to recommend a variety of IPTV quality of experience (QoE) measurements that predict customer experience, to describe the various types of measurements (e.g., parametric and bit-stream approaches), their inputs and outputs, and also the points in an IPTV system where such measurements could be most useful.

Summary Report

Testing lab: XYZ
 Model developer: ABC Corp.
 Model: DEFG Version 1.0 (Software model)
 Scenario: Standard Definition (SD)
 Application: Linear fixed-line IPTV
 Testing round: 4
 Number of PVs: 110

Prediction performance:
 Correlation: 85% (0.85)
 RMSE: 1.7
 Outlier ratio: 0.02
 Accuracy class: B
 Transformation function: $MOS = f(MOSp, a, b, c, d)$;
 $a = 15.7, b = 846, c = 0.669, d = 5.21$

Computational complexity: The minimum, average, and maximum run times for the model were 2s, 2.6s, 2.8s, respectively. This was performed on an XXX Workstation with a YYY processor rated at 2 GHz. The platform had 2 GB of core memory and used a Linux operating system.

[FIG1] Example summary report [5].

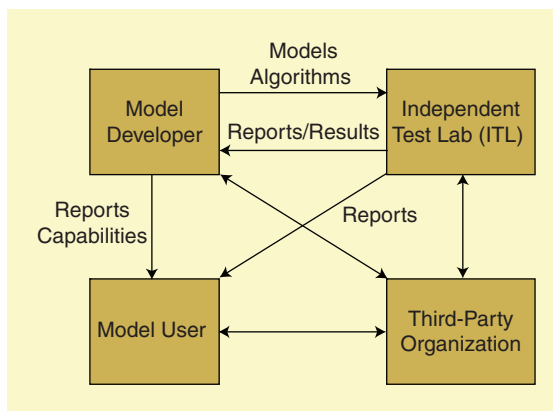
The main premise is that an objective model does not need to be standardized in itself, as its primary requirement is measuring quality with a certain level of accuracy. Also, one could specify various types of perceptual quality models by e.g., their type, expected behavior, inputs, and outputs, thus allowing a black-box approach where the internal details of algorithms do not need to be revealed. Instead, with the test process, test plan, and specification of various types of perceptual quality measurement standardized, a repeatable process for model validation and comparison is created.

Given this, and considering the strengths and weaknesses of VQEG and SG12 approaches to model validation, the ATIS IIF QoSM Committee has produced an alternative process that is similar to the current processes in several ways but is believed to strengthen their weaker aspects. This process has the following unique characteristics:

- An independent validation process using a secret content library of video sequences annotated with subjective rat-

ings, allowing content to be used more than once. The library is prepared and maintained by the ITLs; model developers only have access to the information that is publicly available to everybody and do not become involved in video creation or subjective testing in various ways. Because the library is designed to be reusable, it can be bigger and more varied than for a single test.

- On-demand algorithm validation that allows model developers to have a model evaluated at any time, e.g., at the request of a customer, or when a new model version is released.



[FIG2] Participants of the test process [5].

- Quick turn-around times for model validation rather than multi-year testing events. This is possible because the test procedures and annotated content libraries are prepared in advance, and checking model performance is a simple matter of running a model on the video sequences in the library and compiling results, something that can be done within a few weeks.

- Spur ongoing development and rapid improvement of models, thus increasing model quality and accelerating availability of the best models for model users.

- Clear, well-defined reporting templates, which are designed to provide an overview of the performance of a given model, as well as to facilitate easy comparison of multiple models. Model reports can be requested by model users from model developers or the ITLs. An example report is shown in Figure 1.

- Supporting these process improvements, a validation process is required that consists of clearly defined entities and entity roles, focused on single algorithm submission rather than processes based on competition or collaboration specifically. Collaboratively created models would be validated in the same way as a single algorithm.

- Only the model performance with respect to a standard test plan and library are published. There is no need for algorithm standardization as such. Model developers can keep the details of their algorithms secret, if they so choose, and license their models on their own terms. For example, a model can be developed for a single customer, who can still benefit from independent evaluation.

Other aspects may be quite similar to the current processes. It is envisioned that the process is open and could be as “democratic” in nature as the current processes. To initiate the process, ITLs, model developers, model users, and standards bodies should work together

to define the scope and categories under consideration for model validation.

The validation process is composed of four building blocks (see Figure 2). The blocks represent the different parties needed to provide a rigorous and systematic approach to independently validating objective models.

Fundamental to the process is the existence of an ITL. The ITL may comprise one or more test laboratories and cannot develop objective models. The ITL's operations would be coordinated by a third-party organization that would be the overall sponsor of process activities as well as the business aspects of the process (e.g., relation with content providers, facilitate democratic participation of all parties in this process, fee schedules, and media communications). This third-party organization could be a nonstandards (e.g., commercial) entity or an international standards body such as ATIS or ITU.

The ITL would produce an extensive library of test sequences that are annotated with subjective quality ratings. The library of test sequences needs to be sufficiently large and representative of different video oriented services (e.g., HDTV, mobile) for it to be a good test of model performance. Furthermore, the library of test sequences must be secret. By possessing a large, secret library of test sequences, the ITL is able to reuse test materials for validating models. The ITL is expected to maintain and extend the library of test sequences over time, increasing existing data sets and creating new libraries to accommodate technology developments (e.g., new codecs). The library of test content should be representative of different content genres and should be designed with possible different model categories in mind (e.g., linear broadband TV versus wireless TV). A publicly available document providing a written description of the test content will be produced by the ITL. This written record of test content should provide a description of the video and, where appropriate, audio component of each test sequence.

Once the sequence library has been prepared, the ITL conducts subjective

tests on the sequences in the library for annotation with mean opinion scores (MOS). Subjective scores will be obtained in line with the appropriate standardized subjective test procedures. The MOS annotations need to be maintained and extended along with the sequence library.

Once the annotated sequence library is in place, model developers can submit their models to the ITL for validation. The ITL will perform the validation tests by running the model against a large set of secret sequences that meet the defined scope of the tests.

Once completed, the ITL prepares a report that details the scope of the validation test and the performance of the model. The report is sent to the model developer, who can then decide whether or not to publicly release the performance data. The summary report using a well-defined template will

THERE DOES NOT NEED TO BE A STANDARDIZATION COMPONENT FOR OBJECTIVE MODELS AS LONG AS THERE IS A RELIABLE INDEPENDENT VALIDATION PROCESS.

allow model users to compare results from different models and choose the one best suited to their needs. The summary report (see Figure 1) includes reference to the test plan, category/service scenario/application tested, sequence library, and the number of sequences used in the validation test. It also specifies the prediction performance of the model for the set of PVSs in terms of evaluation criteria, such as correlation coefficients, prediction error, or outliers. Finally, the report includes some indications of model complexity and runtime.

To compare PQMs and PQM results from different model developers, especially as multiple different solutions could be used in an operational environment, there is a need to translate (or cross calibrate) the output of one model with that of another. Cross cali-

bration is a transformation of model outputs to a common scale through the annotated PVS database, typically using a linear or nonlinear fitting function that maps the MOS model outputs to the subjective MOS [15]. Computing this fitting function for a model is part of the validation and will be done by the ITL; the function and its coefficients will also be given in the summary report [5].

CONCLUSIONS

We described the shortcomings of current standards-based test processes for evaluating the accuracy of objective models. Based on the work of the ATIS IIF QoS Committee, we introduced an improved process that mitigates the weaker points of the current processes. We also indicated that there does not need to be a standardization component for objective models as long as there is a reliable independent validation process.

The next step is to actually put this process in place. Practical and commercial questions need to be addressed, for example:

- Who are the ITLs?
- Who is the third-party organization?
- What is the fee structure for model validation?
- What is the role of VQEG, ITU, and ATIS in this process, if any?

This is part of an ongoing discussion among various standards groups, including ATIS IIF, VQEG, ITU-T SG9 and SG12.

RESOURCES

ATIS RESOURCES

The ATIS Web site (www.atis.org) provides details of ATIS standards and technical reports. Contributions to the ATIS IIF QoS Committee are available from www.atis.org/IIF/.

VQEG RESOURCES

The VQEG Web site (www.vqeg.org) provides information on its past and present test projects. The test plans and test reports for each project are available for download. Communications between

ATIS IIF QoSM and VQEG can be found under "Meeting Files" for the various VQEG meetings.

ITU RESOURCES

The ITU Web site (www.itu.int) has links to all ITU-T and ITU-R publications. ITU members can access working documents including the test plans for validating parametric models currently under investigation by Study Group 12.

AUTHORS

Robert C. Streijl (robert.streijl@att.com) is a principal member of technical staff in AT&T's architecture and planning organization. He is the cochair of the ATIS IIF QoS Metrics Committee.

Stefan Winkler (swinkler@cheetahtech.com) is chief scientist at Cheetah Technologies. He is an active contributor to VQEG and ATIS IIF and

cochair of the QoE Metrics Activity Group of the Video Services Forum.

David S. Hands (david.2.hands@bt.com) is a research group leader with BT Innovate & Design. He is an active member of ATIS IIF QoS, ITU-T SG9, and VQEG standards groups.

REFERENCES

- [1] D. Hands. (2007, Mar. 9–10). Video quality measurement: Past, present and future. *Proc. IMQA 2007*, Chiba Univ., Chiba, Japan [Online]. Available: <http://www.mi.tj.chiba-u.jp/IMQA2007/>
- [2] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, Sept. 2008.
- [3] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Process. Image Commun.* (Special Issue on Image and Video Quality Assessment), to be published.
- [4] ATIS, "Test plan for evaluation of quality models for IPTV services," ATIS-0800025, Oct. 27, 2009.
- [5] ATIS, "Validation process for IPTV perceptual quality measurements," ATIS-0800035, Tech. Rep., Dec. 28, 2009.
- [6] *QoE Measurement Recommendations and Framework*, ATIS-0800031, work in progress.

[7] *Objective Perceptual Video Quality Measurement Techniques for Standard Definition Digital Broadcast Television in the Presence of a Full Reference*, ITU-R Recommendation BT.1683, June 2004.

[8] *Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference*, ITU-T Recommendation J.144, Mar. 2004.

[9] *Method for Specifying Accuracy and Cross-Calibration of Video Quality Metrics (VQM)*, ITU-T Recommendation J.149, Mar. 2004.

[10] *Perceptual Visual Quality Measurement Techniques for Multimedia Services Over Digital Cable Television Networks in the Presence of a Reduced Bandwidth Reference*, ITU-T Recommendation J.246, Aug. 2008.

[11] *Objective Perceptual Multimedia Video Quality Measurement in the Presence of a Full Reference*, ITU-T Recommendation J.247, Aug. 2008.

[12] *Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Reduced Reference*, ITU-T Recommendation J.249, Jan. 2010.

[13] Video Quality Experts Group (VQEG) official Web site [Online]. Available: <http://www.vqeg.org/>

[14] K. Brunnstrom, D. Hands, F. Speranza, and A. Webster, "VQEG validation and ITU standardization of objective perceptual video quality metrics," *IEEE Signal Processing Mag.*, vol. 26, no. 3, pp. 96–101, May 2009.

[15] ATIS, "Methodological framework for specifying accuracy and crosscalibration of video quality metrics," ATIS Tech. Rep. T1.TR.72-2001, Oct. 2001. **SP**

The fourth article in this issue, by Pham et al., describes how digital topology is used to compute mathematical representations of the brain's complex and varied structures. Such methods are central to mapping the brain and can help to model global connectivity.

The most sophisticated of today's medical imaging techniques are based on tomographic reconstruction, a general approach in which images of the body's interior are computed from numerous images acquired from outside the body. Tomographic reconstruction is an inverse problem, in which the goal is to invert a sometimes complicated system describing the physical process of data acquisition. Some of the basic concepts of tomography date back to 1917, when Johann Radon described a formalism now known as the Radon transform. Yet, in spite of decades-long interest in the problem of reconstructing medical images, the past few years have seen an explosion of new discoveries about the nature of this inverse problem and its solution.

The fifth article in this issue, by Clackdoyle and Defrise, discusses dramatic recent developments in the solution of the tomographic image reconstruction problems, overturning long-held notions about fundamental issues in this problem domain. In particular, the article reviews advances with respect to reconstruction from incomplete data, and the so-called "interior problem."

Next, Fessler describes so-called model-based approaches to reconstruction in magnetic resonance imaging (MRI), an alternative to classical approaches based on direct Fourier inversion. These approaches recognize the complex nature of real-life MRI data, which include, for example, non-Fourier physical effects and nonlinear magnetic fields. In addition, these approaches can accommodate deliberate undersampling schemes adopted to permit fast scanning; thus, this work relates also to the field of compressive sensing, which was the sub-

ject of a prior issue of *IEEE Signal Processing Magazine*.

Finally, this issue concludes with an article by Ying and Liang, which discusses parallel MRI, an approach in which a phased array of coils is used to perform MRI more rapidly than traditional methods. Parallel MRI is a cutting-edge technology in medical imaging in which signal processing plays a central role. This article focuses on the signal processing issues of multichannel sampling and filter-bank theory.

A WORD OF THANKS

We were pleased and overwhelmed by the large number of outstanding articles submitted for consideration in this issue, and we regret that we could not accommodate all of them. We are very grateful to the authors and reviewers for their exceptional efforts and thoughtful contributions. We also thank Area Editor Dan Schonfeld for his valuable support and assistance throughout the process. **SP**