A Proposed Validation Process for IPTV Perceptual Quality Measurements

NOTE: This document represents work in progress. The core ideas are presented in which a process for model validation is described. The proposal is centred on the notion of an independent testing body and includes guidelines for generating a library of annotated content used to validate quality measurement algorithms, creation of processed sequences, performance reporting and certification classes. The document is presented for comment and we very much welcome feedback and contribution from VQEG.

Introduction

More and more service providers are rolling out IPTV services. In order to support IPTV operations, the need for measurements that can provide insights into the customer's perception of the quality of the IPTV content is apparent.

A closer look at the manner in which the standards and vendor community offers measurement solutions today reveals that the offered solutions are not in par with today's needs of IPTV service providers. Solutions do not focus on all aspects of the IPTV services, and solutions are either offered through standards or not. Standard based solutions are not offered in an expedient manner to meet today's IPTV service providers' needs, and vendor solutions don't follow any particular standardized process for testing. We can see a need for a formalized test process as well as a formalized test plan for proposed measurement solutions to be used as part of the test process.

One of these issues, the formalization of a test process, is discussed in this document at a high-level. The ATIS IPTV Interoperability Forum (IIF) QoSM Committee proposes a certification/test process for validating the quality of the measurement solutions offered. We hope that this proposal leads to a discussion that introduces changes in this area of the IPTV industry.

With this contribution, we seek comments on this initial high-level proposed process in effort to construct a coordinated industry consensus position for introducing a standardized test process for validating measurements and all the aspects needed to support such an effort. Further development of the concepts proposed in this document is expected and will be guided by the comments received and the support received.

This document includes the following discussion points:

- Certification versus standardization
- Content library
- On-demand testing versus staged testing
- Results publication
- Operational process models
- The role of ATIS IIF

Definitions

The following definitions are maintained in this document:

- Perceptual Quality Measurements (PQM): As specified by ATIS IIF, PQM involves any objective quality algorithm, hereafter referred to as model, capable of predicting subjective measurements to be used in (ATIS IIF defined) IPTV-based applications. The scope of PQM includes video, audio, combinations of audio and video, as well as additional content, such as textual and graphical elements (possibly as overlay) as part of the customer experience. ATIS IIF's position is that all PQM must be validated.
- Model Developer: Creators/developers of algorithms that predict video, audio-video or other media quality.
- Independent Test Laboratory: Laboratories have subjective test responsibilities as well as the responsibility to compare a model's performance with the appropriate subjective test content.
- Model User: Measurement vendors, Service providers, ITU-T, ATIS IIF, etc.

Formalizing Models - Certification

While it is recognized that the design of models can be very complex, guidance in this area does not necessarily mean a path to validation results and standardization alone. At any time, certain models of a particular kind may be recommended, but this should not be perceived to mean no new development should take place. ATIS IIF is interested in a highly stimulated and energetic environment that actively promotes constant model improvements.

Another option that can reach perhaps faster results is that of certification. Certification may still lead to standardization. There may be different types of certification, one example being able to classify models by the results of their comparison with subjective tests.

Model Certification
Since different model categories can have different uses and applications, certification can be specific to certain model categories, depending on what's included in the testing database. In order to make the meaning of model certification clear to the users, it is suggested to make the testing category or categories an integral part of the "certification stamp." Useful categories could include video format (SD/HD), video vs. video-audio, etc. As an example, a model would be certified for "HD video-only quality measurement." However, the number of different categories should be kept to a minimum in order to avoid proliferation of models with too narrow application scope.

It is also proposed to define thresholds for one or more of the model evaluation criteria described in the IPTV test plan [3] that can be used to decide whether a given model

predicts subjective MOS data sufficiently well to be certified. Using the correlation of a model's MOSp with subjective MOS as an example, the criterion could be, "the Pearson linear correlation coefficient has to be at least 80% for the model to be certified." Combinations of different evaluation criteria together with measures of statistical significance may be needed to make the threshold comparison meaningful.

Certification Classes

To refine the simple pass/fail criteria for certification, it is further proposed that different threshold levels for the above-mentioned criteria be used to distinguish different levels or classes of model certification. Care will have to be taken that the threshold levels and ranges defined are relevant to the category or scenario in question. Furthermore, as above, combinations of different evaluation criteria together with measures of statistical significance may be needed to make the class distinctions meaningful.

Using again only correlation as a simplified example, the following classes could be defined (CC = correlation coefficient) for a specific model category:

CC > 90%	Class A
CC > 85%	Class B
CC > 80%	Class C
CC > 75%	Class D

It is believed that the definition of different certification classes as above will encourage competition between vendors and improvement of models over time. It will also allow vendors to market and price models based on prediction performance.

Content Library

In order to validate models, a library of processed video and audio-video sequences (PVS) will be created and held by the test laboratories. This content library will be annotated with subjective scores (PVSa). The subjective test process is depicted in the following figure.

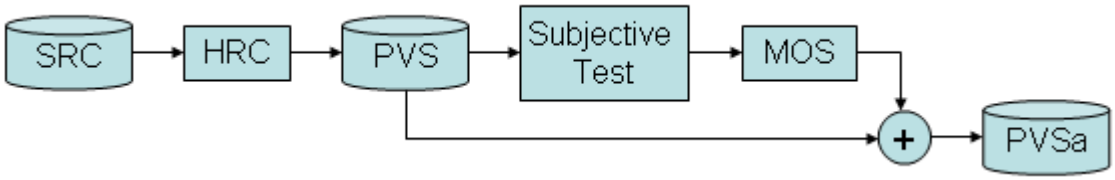

Figure 1: Subjective Test Process

Subjective scores will be obtained in line with the appropriate standardised subjective test procedures. All source (SRC) and processed (PVS) sequences will remain secret, known only to test laboratories.

A small subset of sequences, known as test vectors, and not to be used for validation purposes, will be created by the test laboratory. These test vectors will be publicly available and will be used by model developers and test laboratories to verify that models run correctly at the test laboratory's premises.

Genres and Applications
The library of test content should be representative of different content genres. For example, movies, sports, animation/cartoon, music videos, documentaries, videoconferencing. Within each genre, content must be representative. For example, sports content may include football, tennis, gymnastics, golf and pool.

Content should be correctly formatted for different application scenarios. Examples of target applications include, but are not limited to:

- Linear TV delivered on a fixed-line broadband network (full frame-rate, various resolutions (HD – ¾ SD), progressive and interlace scan)
- TV delivered on a cellular wireless network (variable frame-rates, reduced resolutions (VGA - QCIF), progressive scan)
- Others to be defined.

Source Sequences
The test laboratory will have access to a large selection of high quality source sequences (SRC). A small set of lower quality source sequences should be created (e.g. containing analogue artefacts) and used to produce a special case set of processed test sequences (PVS). All source sequences will be available only to the test laboratory and will remain secret to model developers.

Processed Sequences
Content should be processed using a range of representative coding and transmission methods appropriate to the application being targeted (HRC). For example, processing of linear TV scenario content may use:

- H.264/AVC, VC-1, MPEG-2 video codecs; AC3, AAC+ audio codecs
- UDP packet loss
- Decoders using different error concealment methods
- FEC, ARQ, etc.

The test laboratory should capture both the bitstreams and the uncompressed versions of processed sequences. Care must be taken when storing the files (e.g. the same bitstream may be decoded using different decoders resulting in different subjective scores, the test laboratory must ensure the bitstreams can be matched to the correct uncompressed file; it may be desireable for bitstreams to be captured both with and without encryption).

Service Scenarios
Processed content may belong to different service scenarios. Content should be labelled correctly in order for service scenarios to be easily identified. Example service scenarios include:

- Linear SD IPTV
- Linear HD IPTV
- HD VoD
- Linear IPTV, maximum throughput = 6 Mbit/s, SD, General Entertainment
- Linear IPTV, maximum throughput = 20 Mbit/s, HD, General Entertainment
- Cellular TV, maximum throughput = 384 kbit/s, QCIF, Sports

This will enable model developers to submit models for validation against content datasets that are associated with one or more specific service scenarios.

For each service scenario, the test laboratory should process a minimum of 50 different source sequences (SRC) using a representative set of codecs and transmission errors conditions. The processed sequences along with the associated subjective scores (PVSa), will form the basis of a test library against which models will be validated. Prior to model validation, selection of test scenarios should be agreed between the test laboratory and the model developer. Where a model developer wishes to disclose a model's performance, performance must be reported against each service scenario in full, unaltered form. Model developers may choose to withhold performance details for one or more service scenarios.

Publicly Available Documentation
A publicly available document providing a written description of the test content will be produced by the test laboratories. This written record of test content should provide a detailed description of the video and, where appropriate, audio component of each test sequence (motion, detail, objects present in the video; detail on any audio to include direct speech, commentary, soundtrack details, background noise). In addition to the written documentation detailing scene content, thumbnails (single frame taken from a video sequence) from a representative sample of test sequences should be publicly available.

In addition to a written description of source content, the method used to process sequences should be documented. This publicly available document should provide details of how sequences were processed (e.g. pre-processing applied to source, coding scheme and implementation used, method and process for introducing any transmissions errors, decoder, video capture method and so on).

Finally, a document should be available that lists each source sequence with the processing applied (HRC).

On Demand Testing

Instead of staging a test of various candidate models of a particular kind, a model process that supports on demand testing at any time with a short test period in the order of weeks, rather then years, is highly preferred.

The rational is that while the standards industry currently experiences some very exciting results, ATIS IIF hopes that a continued improvement in the accuracy of the proposed models continues to be a goal for all model developers. In addition, the type of model proposed most likely will evolve over time. Both these developments may impact what has already been standardized. Naturally, on demand testing can only work if there is a complete and secret content library available.

The on-demand test process would then be executed as depicted in the following figure. A model would be validated against a pre-selected set of PVS, and the results of the objective test would be compared with PVSa from the subjective tests.

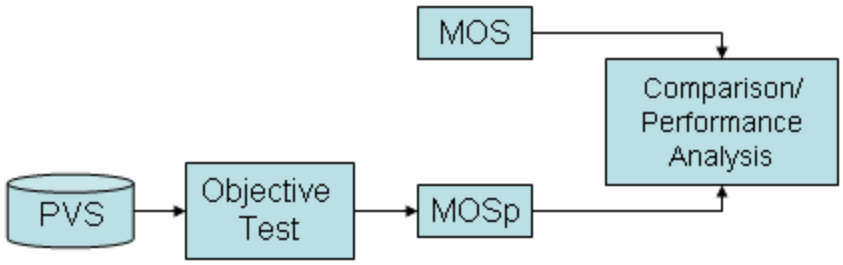

Figure 2: Objective Test Process

Results Publication

On demand testing is initiated by the model developer. The test laboratory conducts the validation and generates the appropriate reports.

There are two types of reports the laboratory generates:

a. Detailed. This is generally shared only with the model developer. The model developer may share it with others, or authorize the ITL to do so. For the agreed upon category (service scenario) the report will include the MOS and MOSp for each PVSa, in the format specified in the test plan.
b. Summary. This is what would normally be shared with model users either by the model developer or the ILT once the certification results are released by the model developer.

The summary report content includes:

1. Reference to test plan, category/service scenario/application tested, PVSa database, and the number of PVS that were used in the validation test

2. model developerPrediction performance of the model for the set of PVSs in terms of evaluation criteria and corresponding certification/class.
3. Identification of the following:
    a. Test laboratory name
    b. Model developer organization, model identifier and version number model developer
    c. Testing round. This identifies whether the test laboratory has tested the model before, including instances where the results were kept confidential.
    d. Outline of model inputs.
    e. Computational complexity (if a uniformly applicable way of describing computational complexity is unavailable, each test laboratory should provide a description of how it determines computational complexity).

The format of a "typical" summary report may be as follows:

Summary Report

Testing lab: XYZ
Model developer: ABC Corp. Model: DEFG Version 1.0
Scenario: SD
Application: Linear fixed-line IPTV
Testing round: 4
Number of PVSs: 110

Prediction performance:
Correlation: 85% (0.85)
RMSE: 1.7
Outlier ratio: 0.02
Certification class: B

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 100 Mb of core memory and used a Linux operating system.

With reference to ATIS-0800008 [1].

Operational Process Models

Models may be submitted as software or as hardware implementations. It is the responsibility of model developers to ensure that the version of the model delivered to test laboratories works correctly. Model developers should provide full installation instructions and a user guide or, where necessary, send a representative to correctly install the model at the test laboratories premises.
For models to be used operationally, the following requirements must be met:

- Model must be a no reference measurement method
- The model must be capable of producing real-time measurements
- Outputs to align with standardised conventions (e.g. for perceptual quality measurements, models should output a MOSp)

Model developers should provide processing requirements to cover:

  - OS
  - CPU load
  - RAM peak usage

ATIS IIF QoSM Committee Role

As indicated earlier, ATIS IIF is currently in the process of creating a specification for PQM tests [3]. While its current scope is focused on video and audio-video tests, ultimately it plans to expand the scope to include all of PQM. Therefore, that document will likely see various revisions over time. Nevertheless, this document is expected to be the standard test plan for PQM and could thus be input to a PQM test process as discussed in this document.

Additionally, ATIS IIF is in the process of defining PQM requirements, more specifically called IPTV QoE Requirements [2]. This document is expected to provide guidance on where and what types of PQM tools should be used in IPTV deployments. This document is expected to be followed by specific solutions. These PQM solutions may be developed internally within ATIS IIF, or adopted from external activities so long as they are within the scope of ATIS IIF. Any PQM tool requirements could be used as part of the validation process to see if the proposed model is in compliance with the standard.

References

[1] ATIS-0800008, "*QoS metrics for Linear/Broadcast IPTV*", ATIS IIF, December 2007

[2] WT-048, "*IPTV QoE Requirements*", ATIS IIF, work in progress

[3] WT-052, "*Test plan for IPTV quality models*", ATIS IIF, work in progress