Standardization Activities in the ITU for a QoE Assessment of IPTV

Akira Takahashi, NTT Service Integration Laboratories
David Hands, Orion
Vincent Barriac, France Telecom

ABSTRACT

This article gives an overview of the state of the art of objective quality assessment of audio and visual media and its standardization activities in the ITU. IPTV services are becoming one of the most promising applications over next generation networks. To provide end users with comfortable, stable, and economical services, QoE assessment methodologies for quality design and management are indispensable.

INTRODUCTION

To ensure that IPTV (Internet Protocol television) services meet the high expectations of end users, factors affecting the quality of service (QoS) or a user's quality of experience (QoE) must be properly considered. A variety of factors can affect the quality of IPTV audio and video, such as the content preparation process, network reliability, and terminal performance. Industry must have access to tools designed to assess the QoE of IPTV services. For this purpose, it is essential to develop a quantitative method to evaluate audio and visual quality efficiently and accurately. The QoE of audio and visual media should be discussed in subjective terms. Subjective quality assessment is the most fundamental method of evaluating subjective quality. However, subjective testing is time-consuming, expensive, and requires special assessment facilities to produce reliable and reproducible test results. Therefore, objective means to predict subjective quality solely from physical characteristics are necessary. These are called *objective quality assessment* methods. Objective quality assessment methods are extremely useful for in-service quality monitoring and management, as well as in codec optimization, codec selection, and quality design of networks or terminals. This article focuses on the state of the art of technologies and standardization activities related to objective quality assessment for IPTV.

STANDARDIZATION BODIES

In the International Telecommunication Union — Telecommunication Standardization Sector (ITU-T), Study Group 12 (SG12) has been investigating QoE requirements and assessment methods for multimedia services including IPTV. Study Group 9 (SG9) studies cable television and also covers video and multimedia quality assessment methods. To harmonize the work of these two SGs, a joint group, the Joint Rapporteur's Group on Multimedia Quality Assessment (JRG-MMQA), has been organized. In addition, the Video Quality Experts Group (VQEG), whose activities are thoroughly introduced in this article, has been acting as a technical advisory team mainly for the objective quality assessment of video. This article focuses on the activities of ITU and VQEG, but important work on IPTV quality measurement also is in progress in other standardization bodies. In particular, both the Alliance for Telecommunications Industry Solutions — IPTV Interoperability Forum — QoS Metrics (ATIS IIF QoSM) group and the European Telecommunications Standards Institute — Technical Committee for Speech, Transmission Planning, and Quality of Service (ETSI STQ) are working on defining appropriate methods for evaluating the quality of IPTV services.

To coordinate and promote the development of global IPTV standards, taking into account the existing work of the ITU study groups, as well as standards-developing organizations, fora, and consortia; ITU formed the Focus Group on IPTV (FG-IPTV) in July 2006. The mission of FG-IPTV includes the study of the QoE/QoS aspects of IPTV. This has been assigned to Working Group 2 (WG2) of FG-IPTV, which is currently developing the following four documents for standardization in ITU-T:

- Quality of experience requirements for IPTV
- Traffic management mechanism for the support of IPTV services
- Application layer reliability solutions for IPTV
- Performance monitoring for IPTV

From the viewpoint of quality assessment methodologies, the last document is relevant because it addresses the QoE monitoring framework for IPTV. This is one of the most important application scenarios for objective quality assessment methods that are introduced in the following sections.

0163-6804/08/$25.00 © 2008 IEEE

	Media-layer model	Parametric packet-layer model	Parametric planning model	Bitstream layer model	Hybrid model
Input information	Media signal	Packet header information	Quality design parameters	Packet header and payload information	Combination of any
Primary application	Quality benchmarking	In-service nonintrusive monitoring (e.g., network probe)	Network planning, terminal/application designing	In-service nonintrusive monitoring (e.g., terminal-embedded operation)	In-service nonintrusive monitoring
Existing standards and ongoing projects in ITU					
Speech	ITU-T P.862	ITU-T P.564	ITU-T G.107	—	ITU-T P.CQO
Audio	ITU-R BS1387	ITU-T P.NAMS [IPTV]	ITU-T G.1070 [videophone] ITU-T G.OMVS [IPTV]	ITU-T P.NBAMS [IPTV]	—
Video	ITU-T J.144[SD] ITU-T J.vqhdtv[HD] ITU-T J.mm**[PC]				ITU-T J.bitvqm [IPTV]
Multimedia	(ITU-T J.148)				—

Table 1. *Objective quality assessment models.*

Overview of Objective Quality Assessment Models

In general, objective quality assessment methodologies can be categorized into five types. These are media-layer models, parametric packet-layer models, parametric planning models, bitstream-layer models, and hybrid models (Table 1).

A media-layer model utilizes speech or video signals to predict QoE. Because it does not require a priori knowledge about the system under testing, such as the codec type or packet-loss rate, it can be applied to the evaluation of unknown systems (e.g., codec comparison/optimization). However, by definition, it cannot be used in scenarios in which media signals are not available. For example, it is difficult to obtain media signals at the network mid-point although one can decode the payload of packets. Examples of such a model are ITU-T Recommendations J.144 for video and P.862.1 for speech. This topic is thoroughly discussed in a later section.

A parametric packet-layer model predicts QoE solely from packet-header information, enabling very lightweight measurement without handling the media signal itself. However, it has difficulty evaluating the content dependence of QoE, for example, because it does not look at the payload information. For speech, ITU-T Recommendation P.564 provides the framework and performance requirements for such models, and some commercial products are available.

A parametric planning model is one that takes quality planning parameters for networks and terminals as its input. It requires a priori information about the system under testing. A good example is ITU-T Recommendation G.107, the *E-model*, which has been widely used as a network planning tool for the public switched telephone network (PSTN) and for voice-over-IP (VoIP) services [1]. Recently, ITU-T adopted a new model for videophone services as Recommendation G.1070 [2, 3]. Such a model also is needed for IPTV services. This topic, together with parametric packet-layer models, is discussed later.

A bitstream-layer model is a new concept. It occupies a position between media-layer models and parametric packet-layer models. It utilizes encoded bitstream information, in addition to the packet-layer information that is used in parametric packet-layer models, so that it can take into account the content-dependent quality evaluation characteristics with a relatively light computational load. This is also discussed later.

The last type is a so-called hybrid model, which is a combination of the previously mentioned technologies. It is effective in terms of exploiting as much information as possible to predict QoE. For example, [4] takes a media signal and bitstream information to exploit the information available to both media-layer and bitstream-layer models.

QoE Measurement Framework

The ability to measure the quality of speech, audio, and multimedia services is important to service providers and network operators in ensuring that content is prepared, transmitted, and received according to an appropriate standard of quality. For services such as VoIP, IPTV, video streaming, and mobile TV, it is necessary to assure and monitor quality to identify and resolve problems before they have a negative impact on the customer experience. The preceding sections introduced different measurement methods for voice, audio, and video. For each media type, various measurement methods are available. These methods have quite different computational and operational requirements. Now, we introduce a quality measurement framework that defines both the measurement points in an IPTV system and the objective method most suitable for each measurement point. The framework is confined to quality

The framework can be delineated into three specific measurement points. First, measurements are obtained prior to transmission. The second measurement point is in the network itself. Finally, measurements are obtained at the receiving device.

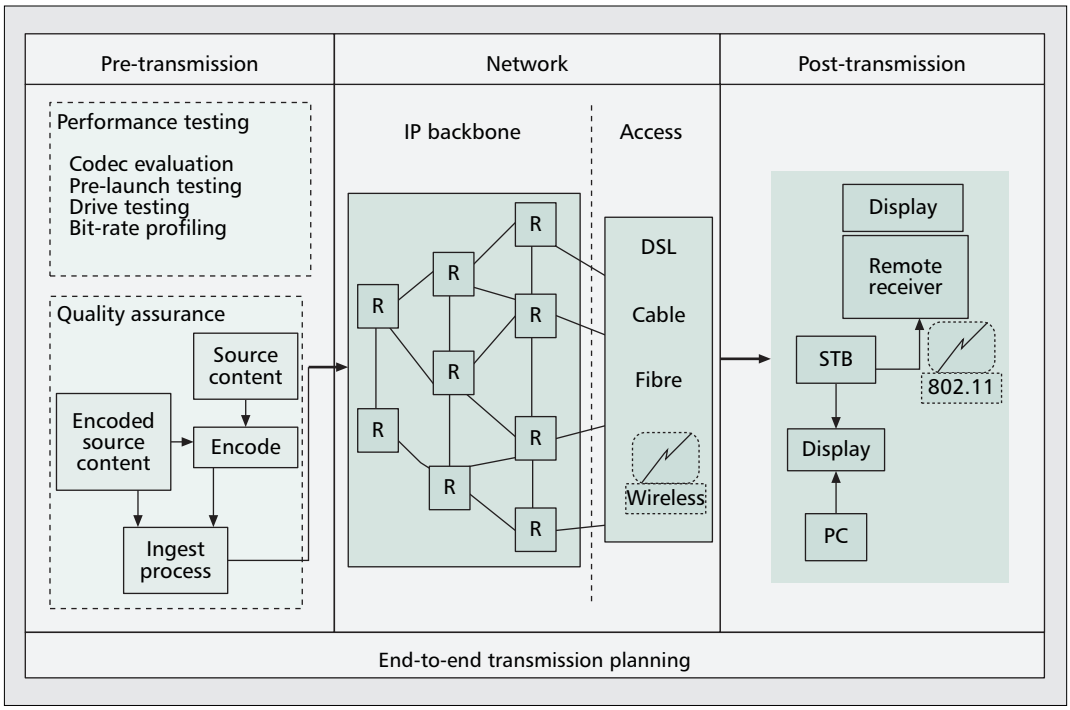

Figure 1. *Quality measurement framework. The framework shows three distinct measurement points. Pre-transmission measurement is especially suited to media layer and hybrid methods. Parametric and bitstream methods are most appropriate for network measurements. No-reference media layer, bitstream layer, and hybrid methods are appropriate for post-transmission measurement. Transmission planning tasks are suited to parametric methods.*

measurement methods that can be defined for transmission planning, network performance measurement (e.g., IP performance metrics as set out in IETF RFC 2330), quality assurance, and monitoring performed before, during, and after transmission. QoE metrics (e.g., channel-changing times, electric-program guide (EPG) response functionality and ease of use, and device-dependent customer expectations) are not considered here, but FG-IPTV and ATIS IIF QoSM are considering QoE in its totality. A general overview of the IPTV quality measurement framework is shown in Fig. 1.

The framework can be delineated into three specific measurement points. First, measurements are obtained prior to transmission. The second measurement point is in the network itself. Finally, measurements are obtained at the receiving device. Each measurement point is discussed in more detail in the following.

Pre-Transmission Measurement

Measurements derived before transmission can be used for a number of purposes, including performance testing and quality assurance. For performance testing, application developers can test the quality of media applications (e.g., codec development and error concealment methods). For industry, reliable measurement methods are required to evaluate the performance of competing vendor products (e.g., codec evaluation).

After a service provider decides on its operational environment, quality measurements become important to ensure that content is received, encoded, and provided at appropriate quality prior to distribution to customers. The left-hand side of Fig. 1 illustrates the pre-transmission content processing chain. Content often is provided by third parties to service providers. The service provider can receive content at distribution quality (e.g., digibeta tapes) or already encoded for transmission. If service providers receive distribution quality content, they must encode it themselves prior to transmission. After content has been encoded according to the desired profile, a provision procedure is performed to place the media in the appropriate transmission package.

For both performance testing and quality assurance, perceptual quality measurements are appropriate. Perceptual quality measurements, such as pixel-based methods using full reference procedures (ITU-T Rec. J.144) are particularly useful for performance testing where different coding schemes (e.g., G.722, G.729, AAC+, MPEG-2 audio, MPEG-2 video, H.264, and VC-1) are under evaluation. Full-reference (FR) methods tend to be computationally intensive, require access to both the original and processed versions of content, and require spatio-temporal registration to be performed. As a result of these limitations, FR methods are not ideal for quality assurance (QA) tasks. Instead, real-time measurement methods are required. Furthermore, it is desirable that QA methods apply no-reference (NR) procedures because the operating environment may not always allow access to the source content.

Network Measurement

There are a number of methods for obtaining network performance measurements: the simplest ones reside on network routers that can

be probed to determine traffic management efficiency and identify and quantify problems. More specific network measurement probes have been defined (e.g., IETF RFC 4445 and IETF RFC 2330) but have not become commonplace because of the increased overhead and installation effort associated with their application. Perceptual quality metrics, particularly, reduced-reference (RR) and NR methods and parametric packet-layer methods have been identified as potential tools for in-service network-based monitoring. Although all forms of measurement probe may be applied to network monitoring, simplicity is especially important for network measurement, and minimal network measurement is expected. As a result, parametric or bitstream measures may prove suitable to this task; although for these methods to be widely accepted, they must be proven to be superior to basic channel error statistics.

Post-Transmission Measurement

The point of reception is the most important measurement point for monitoring customer experience. It may be argued that knowledge of the quality of the encoded and transmitted signal, together with information related to network performance (e.g., packet-loss rate and latency), is sufficient to understand the user experience. Where this measurement scenario fails is that it does not consider problems with the receiving device (or home network environment); nor is it able to adequately accommodate error concealment. For accurate post-transmission measurement, NR quality methods, which can be media-, packet-, or bitstream-layer or hybrid models, are most appropriate. RR methods may be applied where a sidechannel is made available, but the requirement for a sidechannel limits the usefulness of RR methods. The right-hand side of Fig. 1 outlines post-transmission measurement options.

In addition to the three measurement stages identified previously, a fourth form of measurement is commonly performed, namely measurement during transmission planning. Parametric planning models have been defined to examine the impact of different transmission performance characteristics on the quality of speech (ITU-T Rec. G.107) and videophone (ITU-T Rec. G.1070) services. Work is in progress to define a general parametric model for IPTV service.

Media-Layer Models

Objective perceptual video quality measurement was pioneered in the early 1990s [5]. Media-layer quality models utilize knowledge of the human visual system to predict the subjective quality of video. Human visual processing can be represented in two steps: psychophysical models account for fundamental, low-level visual information processing (e.g., spatial and temporal frequency response) and cognitive models account for high-level functions (e.g., memory biases and judgment operations). Objective media-layer methods can model human visual processes either directly or indirectly. For example, the methods of Daly and Lubin include a direct model of human contrast sensitivity, and spatial frequency response, as well as luminance and contrast masking. Watson's DCTune method uses discrete cosine transform (DCT) frequency components to indirectly model human responses to different frequencies within visual stimuli and the subsequent impact on error visibility.

The early objective media-layer models, along with the majority of those that have been proposed since, have used a full-reference method. Full-reference (FR) methods extract information from the source video (usually high quality or undegraded) and its processed counterpart. The reference signal acts as a consistent baseline for comparison. A special instantiation of the full-reference method, termed reduced reference, may be applied for in-service quality monitoring [6]. In RR methods, information is extracted from the reference signal and packaged and transmitted alongside the processed video. RR models assume that a sidechannel will be available to send the reference signal parameter data. The richness of information describing the properties of the reference signal is dependent on the capacity of the sidechannel. NR methods operate solely on information extracted from the processed signal. NR methods are normally defined for specific coding schemes (e.g., MPEG-2) or error types (e.g., blur).

For both FR and RR methods to operate effectively, the reference and processed video sequences must be closely aligned. This spatio-temporal alignment (or registration) requirement represents a major obstacle to the operational application of these models. NR models do not require registration and as such, represent the most efficient means of measuring quality in an operational environment. The problem for NR methods is their predictive ability, as it is especially difficult to obtain high accuracy without any content or quality benchmark. Although published descriptions of both RR and NR models indicate promising quality prediction performance, independent validation is required. Several standards bodies have performed validation testing of objective models. For video and multimedia model evaluation, VQEG has become a prominent forum for establishing the predictive performance of objective quality methods.

VQEG [7] was formed in 1997 to provide a forum for objective video model developers to propose and test their algorithms. The output from VQEG is reported to various standards bodies, most notably ITU-T and ITU — Radiocommunication Sector (ITU-R). In 1999, VQEG performed a set of subjective quality tests designed to evaluate the performance of objective media-layer quality models. In the tests, material was selected to be broadly representative of standard definition television content. The double-stimulus continuous quality scale (DSCQS) method was used to collect subjective scores. For each test sequence, a difference mean opinion score (DMOS) was produced. These tests provided a set of test sequences, each annotated with a DMOS value. A total of ten proponents submitted objective models to

Parametric planning models have been defined to examine the impact of different transmission performance characteristics on the quality of speech (ITU-T Rec. G.107) and videophone (ITU-T Rec. G.1070) services. Work is in progress to define a general parametric model for IPTV service.

The standardization body in charge of the development of such models is currently ITU-T SG12. Based on contributions from several organizations, SG12 decided to launch new study items on parametric planning/packet-layer models for IPTV services.

VQEG for evaluation. The annotated test sequences were processed by each model and the performance evaluated. The primary performance statistic was Spearman's correlation coefficient. This indicates the strength of the relationship between the output of the objective model and the subjective rating. The peak signal to noise ratio (PSNR) also was calculated for the set of test sequences. In this test, none of the models was found to be statistically superior to PSNR in predictive performance. The VQEG Phase I testing had limitations in the test design that may have contributed to the failure to differentiate between objective methods. In particular, the quality range of the test sequences was unbalanced, with the majority of sequences rated as good quality or better. This led to a second phase of VQEG testing.

In the VQEG FR-TV Phase II tests, eight new objective quality metrics in total were submitted for evaluation. Subjective tests were performed by three independent laboratories. Video material covering a representative range of television and film content was encoded, on the whole, using the MPEG-2 compression scheme. Model performance was described using both the Spearman and Pearson correlation coefficients along with root mean square error (RMSE). To examine whether there were statistically significant differences in predictive performance between models, F-tests were performed. In these tests, a number of objective models were found to outperform PSNR [8]. The performance of the different models was considered by the ITU. Following discussions at the ITU, two new international standards were drafted and agreed upon. These two standards (ITU-T Rec. J.144 and ITU-R Rec. BT.1683) describe four objective media-layer quality models that are considered suitable for use in the objective measurement of standard definition television pictures.

The VQEG FR-TV tests were completed in 2003. Since then, the group has concentrated its efforts on designing tests to evaluate objective quality metrics for different service environments and applications. Work has focused on three main activities, namely the validation of RRNR methods for standard definition TV, validation of multimedia objective quality methods, and validation of high-definition television (HDTV) objective quality models.

The RRNR-TV tests will create video content using both MPEG-2 and MPEG-2/H.264 AVC compression schemes and will include conditions containing transmission errors. The scope of the multimedia tests far exceeds other work in the group. The multimedia tests will evaluate FR, RR, and NR methods. Their content initially will be limited to video. Future tests will include audio and audio-video test content. Nevertheless, the multimedia tests have ambitious objectives. For these tests, video content will be created at three different resolutions (VGA, CIF, and QCIF) and a wide range of video codecs, frame rates, and bit rates will be assessed. The multimedia tests will include content exhibiting both compression and transmission errors. The results are expected to be published in late 2007.

Parametric Models

As stated earlier, it seems more and more possible to combine knowledge of the quality of the encoded and transmitted signal and information related to network performance (e.g., packet-loss rate and latency) to understand the user experience. Although there is no standard technology for IPTV services to directly map the objective transmission characteristics to QoE, there are counterpart technologies for speech that were standardized as ITU-T Recommendations G.107 and P.564.

G.107, often called the *E-model*, estimates the QoE of telephone services by taking into account the different aspects of conversational quality. In addition, ITU recently standardized a so-called *non-voice E-model* for videophone applications as Recommendation G.1070. Earlier, we categorized such technologies as *parametric planning models*.

For quality monitoring of VoIP services, where one must monitor QoE solely from a packet stream that one can capture, ITU-T created Recommendation P.564. It defines the framework of models that estimate QoE, based on packet-header information and provides performance criteria in terms of quality estimation accuracy. Previously, this was categorized as a *parametric packet-layer model*. Such models enable real-time and detailed quality monitoring, which is one of the most important operational issues for providing end users with stable and satisfactory QoE.

The standardization body in charge of the development of such models is currently ITU-T SG12. Based on contributions from several organizations, SG12 decided to launch new study items on parametric planning/packet-layer models for IPTV services. A parametric planning model for IPTV services is tentatively called opinion model for video streaming (G.OMVS) applications and is studied under Question 13 of SG12. On the other hand, a parametric packet-layer model is called non-intrusive parametric model for the assessment of performance of multimedia streaming (P.NAMS), and is studied under Question 14 of SG12. Although these two models have different assumptions and input interfaces, the quality estimation function can be common. Therefore, SG12 decided to synchronize these two projects and develop a so-called *core model*. A block diagram of this framework is illustrated in Fig. 2.

In the short term, the first step currently under study is the development of the core model. This will become the basic building block of both P.NAMS and G.OMVS. Each model will run in its specific application domain with its own assumptions. This core model and a global overview of P.NAMS and G.OMVS are given in the figure.

Bitstream Layer Models

For in-service non-intrusive QoE measurement, parametric packet-layer models are very effective in terms of the computational load. Because they do not look at the payload information, however, it is impossible for them to take into

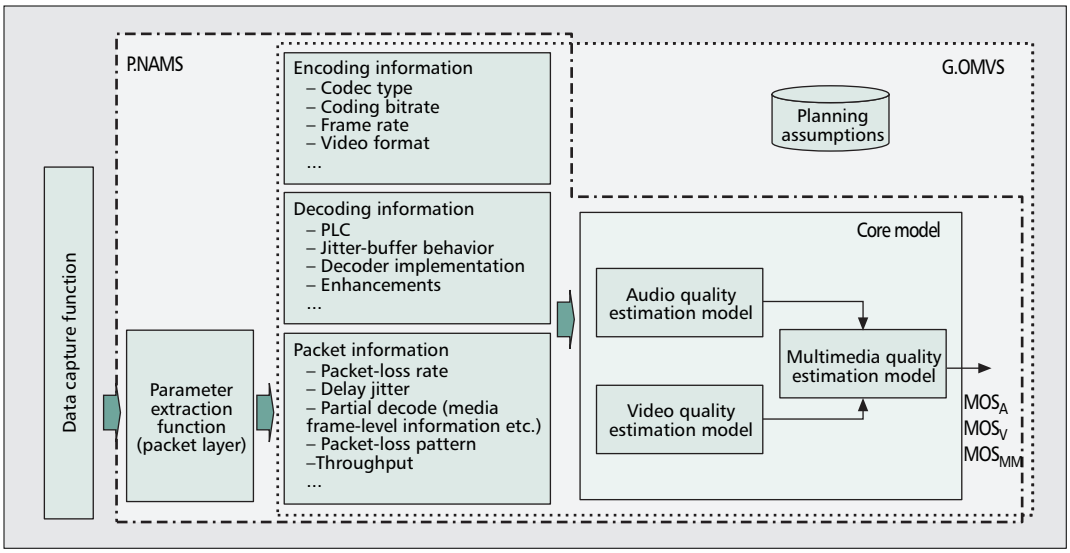

Figure 2. *Framework of parametric planning/packet layer models: G.OMVS and P.NAMS.*

account the dependence of quality on audio and video content. That is, the quality estimated by parametric packet-layer models is the average quality over some of the content used in training the models. This is not appropriate in the scenario in which one must monitor the QoE of individual users. On the other hand, media-layer models can solve this problem by analyzing the characteristics of audio and video content. Nevertheless, obtaining the media signal is not easy in the QoE monitoring scenarios.

One solution to this problem is to utilize the coded bitstream information to analyze the characteristics of source content. For example, DCT coefficients in MPEG-type coding tell us about the spatial complexity of video scenes, which affects the coding performance at a given bit rate and the robustness against packet loss. Figure 3 compares bitstream-layer, media-layer, and parametric packet-layer models on protocol stacks. Recently, VQEG has launched a new study item regarding bitstream models, although the definition and standardization scope are still under discussion.

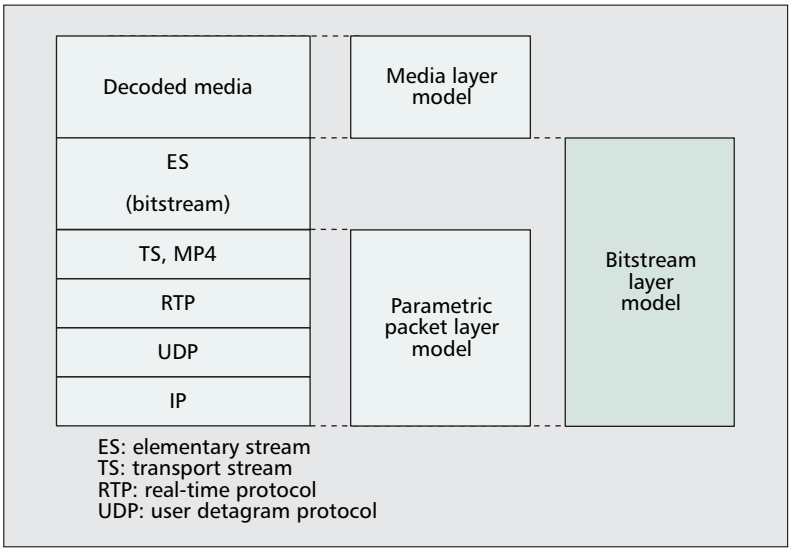

Figure 3. *Bitstream model in comparison with media and packet layer models.*

Future Work

In addition to the existing agenda that VQEG has set itself, the group is confronted with fresh ideas and problems. As novel measurement methods are proposed, these must be defined and evaluated. For example, pixel-based measures, particularly those in which models of the human visual system are present, tend to be computationally intensive. Furthermore, both FR and RR methods require that a registration process is performed in order to operate effectively. The prediction accuracy and generality of FR methods will ensure their continued usage, primarily for performance testing where processing time is not critical (e.g., evaluation of different video codecs). However, many industrial applications require real-time measurement and in many instances, the reference signal may not be available. In such circumstances, a valid NR approach is desirable. Pixel-level NR methods, such as those to be evaluated in VQEG RRNR-TV and multimedia tests may meet the requirements of the industry. An alternative will be required if NR pixel-level methods fail to provide sufficiently accurate quality measurements or are too computationally intensive. A new NR approach [4] has been designed for codec-specific quality measurement where fast and highly accurate perceptual quality measurement is essential. In this method, information is extracted directly from the bitstream by using supplementary parameter values obtained from the decoded picture. A working version of this hybrid bitstream/decoder perceptual quality method has been developed for MPEG-2/H.264 AVC. The performance of this method indicates real promise, providing predictive accuracy superior to FR methods. VQEG, ITU-T SG9, and ITU-T SG12 have plans to evaluate and standardize such hybrid and bitstream models as those introduced in the previous section. In the

Perceptual models that accurately predict quality over short periods of time are already available and useful to the industry. Models that account for error aggregation and decay will be invaluable tools in the policy-making and management of video services.

future, such models will be implemented in set-top box and IPTV devices so that service providers can estimate remotely the real QoE that users experience.

Data from subjective tests is essential to develop objective models that produce accurate measurements. The current generation of objective models, on the whole, has been calibrated using subjective data obtained using short video sequences. For the models to exploit their potential fully, subjective data describing natural viewing is required. This data should describe the subjective aggregation of transient errors and weighting errors dependent on their form, duration, and intensity, as well as identifying how the perceptual impact of errors decays. Perceptual models that accurately predict quality over short periods of time are already available and useful to the industry. Models that account for error aggregation and decay will be invaluable tools in the policy-making and management of video services.

References

[1] S. Moeller, *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer Academic, 2000.
[2] K. Yamagishi and T. Hayashi, "Opinion Model for Estimating Video Quality of Videophone Services," *IEEE GLOBECOM '06*, QRP08-1, Nov. 2006.
[3] T. Hayashi et al., "Multimedia Quality Integration Function for Videophone Services," *IEEE Global Telecommun. Conf. 2007*, Nov. 2007.
[4] D. Hands, "Quality Assurance for IPTV," *ITU-T Wksp. End-to-End QoE/QoS*, June 2006.
[5] S. Winkler, *Digital Video Quality: Vision Models and Metrics*, Wiley, 2005.
[6] P. Le Callet *et al.*, "No Reference and Reduced Reference Video Quality Metrics for End to End QoS Monitoring," *IEEE Trans. Commun.*, vol. E89-B, 2006, pp. 289–96.
[7] Video Quality Experts Group, www.vqeg.org
[8] Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II; http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII/index.php

Biographies

AKIRA TAKAHASHI [M'04] (takahashi.akira@lab.ntt.co.jp) received a B.S. degree in mathematics from Hokkaido University, Japan, in 1988, an M.S. degree in electrical engineering from the California Institute of Technology in 1993, and a Ph.D. in engineering from the University of Tsukuba in Japan in 2007. He joined NTT Laboratories in 1988 and has been engaged in the quality assessment of audio and visual communications. Currently, he is the manager of the Service Assessment Group at NTT Laboratories. He has been a co-rapporteur of ITU-T Question 13/12 on multimedia QoE and its assessment.

DAVID HANDS (david.2.hands@bt.com) has been working in the field of objective and subjective video quality assessment for the past 10 years. He was involved in the RACE II MOSAIC project that defined the internationally standardized continuous quality subjective assessment method and is leading British Telecom (BT)'s development of objective video quality tools. He is co-chair of the VQEG multimedia group and is an active member of ITU-T SG9.

VINCENT BARRIAC (vincent.barriac@orange-ftgroup.com) is an electrical engineer who specializes in signal processing. He joined France Telecom R&D in 1993, where he works in the Speech and Sound Technologies and Processes Laboratory. He is a senior expert in voice quality. Initially, he developed algorithms for non-intrusive measurements, in particular for echo. Additionally, he studied and developed different objective methods to model the quality of transmission, particularly VoIP (in mono or multiplay context) and wideband audio. He is currently a rapporteur for ITU-T Question 14/12 on nonintrusive measurement techniques.