Perceived Quality Evaluation for Multimedia Services

H. Koumaras
University of Athens, Greece

E. Pallis
Technological Educational Institute of Crete, Greece

G. Xilouris
University of Athens, Greece

A. Kourtis
N.C.S.R., Demokritos, Greece

D. Martakos
University of Athens, Greece

INTRODUCTION

The advent of 3G mobile communication networks has caused the fading of the classical boundaries between telecommunications, multimedia, and information technology sectors. The outcome of this convergence is the creation of a single platform that will allow ubiquitous access to the Internet, multimedia services, and interactive audiovisual services, and in addition (and most importantly) offering the required/appropriate perceived quality level at the end user's premises.

In this respect, multimedia services that distribute audiovisual content over 3G/4G mobile communication systems are expected to posses a major part of the bandwidth consumption, making necessary the use of video compression. Therefore, encoding techniques (e.g., MPEG, H-26x) will be applied which achieve high compression ratios by exploiting the redundancy in the spatiotemporal domain of the video content, but as a consequence produce image artifacts, which result in perceived quality degradation.

One of the 3G/4G visions is the provision of audiovisual content at various quality and price levels. There are many approaches to this issue, one being the perceived quality of service (PQoS) concept. The evaluation of the PQoS for audiovisual content will provide a user with a range of potential choices, covering the possibilities of low-, medium-, or high-quality levels. Moreover the PQoS evaluation gives the service provider and network operator the capability to minimize the storage and network resources by allocating only the resources that are sufficient to maintain a specific level of user satisfaction.

The evaluation of the PQoS is a matter of post-encoding procedures. The methods and techniques that have been proposed in the bibliography mainly aim at:

- determining the encoding settings (i.e., resolution, frame rate, bit rate) that are required in order to carry out successfully a communication task of a multimedia application (i.e., videoconference); and
- evaluating the quality level of a media clip based on the detection of artifacts on the signal caused by the encoding process.

The scope of this article is to outline the existing procedures and methods for estimating the PQoS level of a multimedia service.

BACKGROUND

The advent of quality evaluation was based on applying pure mathematical/error-sensitive equations between the encoding and the original/uncompressed video signal. These primitive methods, although they provided a quantitative approach about the quality degradation of the encoded signal, do not provide reliable measurements of the perceived quality, because they miss the characteristics and sensitivities of the human visual system.

The most widely used primitive methods and quality metrics that are based on the error sensitivity framework are the peak signal to noise ratio (PSNR) and the mean square error (MSE):

$$PSNR=10\log_{10}\frac{L^2}{MSE}$$, *where L denotes the dynamic pixel value (i.e., equal to 255 for 8bits/pixel monotonic signal)* (1)

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2$$, *where N denotes the total pixels and* x_i */*y_i *the* i^{th} *pixel value in the original/distorted signal* (2)

Currently, the evaluation of the PQoS is a matter of objective and subjective evaluation procedures, each time taking place after the encoding process (post-encoding evaluation). Subjective picture/audio quality evaluation methods require a large amount of human resources, establishing it as a time-consuming process (e.g., large audiences evaluating video/audio sequences). Objective evaluation methods, on the other hand, can provide PQoS evaluation results faster, but require a large amount of machine resources and sophisticated apparatus configurations. Towards this, objective evaluation methods are based on and make use of multiple metrics, which are related to the content's artifacts (i.e., tilling, blurriness, error blocks, etc.) resulting during an encoding process.

These two categories of PQoS evaluation methods will be analyzed and discussed in the following sections.

SUBJECTIVE QUALITY EVALUATION METHODS

The subjective test methods, which have mainly been proposed by the International Telecommunications Union (ITU) and the Video Quality Experts Group (VQEG), involve an audience of people who watch a video sequence and score its quality, as perceived by them, under specific and controlled watching conditions. Afterwards, the statistical analysis of the collected data is used for the evaluation of the perceived quality. The mean opinion score (MOS) is regarded as the most reliable subjective metric of quality measurement and has been applied on the most known subjective techniques.

Subjective test methods are described in ITU-R Rec. T.500-11 (2002) and ITU-T Rec. P.910 (1999), suggesting specific viewing conditions, criteria for the observer, test material selection, assessment procedure description, and statistical analysis methods. The BT.500-11 describes subjective methods that are specialized for television applications, whereas ITU-T Rec. P.910 is intended for multimedia applications.

The most known and most widely used subjective methods are:

- **Double Stimulus Impairment Scale (DSIS):** This method proposes that observers watch multiple references and degraded scene pairs, with the reference scene always shown first. Scoring is evaluated on an overall impression scale of impairment: imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying. This scale is commonly known as the five-point scale (where 5 corresponds to "imperceptible" and 1 to "very annoying").
- **Single Stimulus (SS) Methods:** Multiple separate scenes are shown. There are two different SS approaches: SS with single view of test scenes and SS where the test scenes are repeated. Three different scoring methods are used:
  - **Adjectival:** The aforementioned five-grade impairment scale, however half-grades are allowed.
  - **Numerical:** An 11-grade numerical scale, useful if a reference is not available.
  - **Non-Categorical:** A continuous scale with no numbers or a large range, for example, 0-100.
- **Stimulus Comparison Method:** This methods exploits two well-matched screens, where the differences between scene pairs are scored in one of the two following scoring methods:
  - **Adjectival:** A seven-grade, +3 to -3 scale labeled: much better, better, slightly better, the same, slightly worse, worse, and much worse.
  - **Non-Categorical:** A continuous scale with no numbers or a relation number either in absolute terms or related to a standard pair.
- **Single Stimulus Continuous Quality Evaluation (SSCQE):** According to this method, the viewers watch a program of typically 20-30 minutes without any reference signal. The viewers, using a slider, continuously rate the instantaneously perceived quality using an adjectival scale from 'bad' to 'excellent', which corresponds to an equivalent numerical scale from 0 to 100.
- **Double Stimulus Continuous Quality Scale (DSCQS):** At DSCQS the viewers watch multiple pairs of quite short (i.e., 10 seconds) reference and test sequences. Each pair appears twice, with random order of the reference and the test sequence. The viewers/subjects are not aware of the reference/test order, and they are asked to rate each of the two separately on a continuous adjectival scale, ranging from 'bad' to 'excellent', which corresponds to an equivalent numerical scale from 0 to 100. This method is usually used for evaluating slight quality differences between the test and the reference sequence.

The aforementioned methods are described in the ITU-R Rec. T.500-11 document and are mainly intended for televi-