INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2009-2012

COM 12 – C 205 – E

January 2011

English only

Original: English

Question(s): 9, 13, 14/12

STUDY GROUP 12 – CONTRIBUTION 205

Source: Telchemy Inc.

Title: Interpretation of MOS for video services

Introduction

The concept of the Mean Opinion Score (MOS) is fairly straightforward: Looking at the ACR five-point scale, then "1" means *bad quality* and "5" means *excellent quality* (ITU-T P.800). Quantifying service qualities through MOS values (verbally or through integers) as an average user perceives them sufficiently satisfied the needs of the telecom industry in the past. However with the extension of services to provide a multitude of voice bandwidths in the telecom industry or a large number of video resolutions in digital TV broadcasts the previously defined MOS becomes more and more difficult to interpret. This is similar to the issue that occurs with the use of MOS for mixed narrowband and wideband speech services.

This contribution briefly describes the problem and proposes that ITU-T SG12 address the interpretation of MOS for video applications.

Discussion

The MOS as a means to quantify quality as a user perceives it is fairly easy to understand. However with the extension of services to provide a multitude of service options the interpretation of service qualities using a MOS becomes more difficult.

A video subjective test is typically conducted using samples and anchors that are of the same resolution, for example standard definition. Within the industry it is however common for service providers to transcode video content into a range of different formats and to broadcast (or multicast) in HD, SD and mobile formats. Telchemy's experience within the IPTV industry is that service providers often expect that MOS scores would be scaled independently of the format used when the service provider is looking only at a single format (e.g. the expectation that unimpaired SD would have a MOS of 4.5 or more) whereas expect that the MOS for HD would be higher than the MOS for SD; these are obviously inconsistent and cannot be addressed by a single MOS scaling.

Contact: Martin Kastner
Telchemy Inc.
USA
Tel: +1 (678) 387 3000
Fax: +1 (678) 387 3008
Email: martin.kastner@telchemy.com

Contact: Alan Clark
Telchemy Inc.
USA
Tel: +1 (678) 387 3000
Fax: +1 (678) 387 3008
Email: alan.clark@telchemy.com

An example of two use cases for the use of MOS:

Depict a standard definition (SD) video experiment A evaluating the qualities of signals with a frame resolution and rate which are typical for NTSC. A test condition in this test where the encoding bandwidth is high, and apart from minor encoding artefacts with no additional impairments, we could assume that the quality is pretty good thus resulting in a MOS greater than or equal to 4.5.

In addition to the above experiment there is an experiment B, which evaluated the same content but at a high definition (HD) resolution. The average subjective opinion turns out to be a MOS of 4.5 or greater as well.

Both MOS values are rated good to excellent quality even though they have different frame resolutions and degree of detail (up to 6times more pixels in HD). Naturally the HD material could be assumed to have better overall quality than the SD material, but still both their qualities are rated good to excellent.

It is obvious that those two test conditions are not evaluated in an absolute manner but that they have been evaluated relative to their particular context, which is predominantly described by their frame resolution. We could call the resulting MOS values *relative MOS values*.

In order to be able to compare the qualities across the different frame resolutions, an experiment C where both frame resolutions are presented could be designed. Very likely we will find that the HD material will be rated higher than its equivalent in SD. The resulting MOS could be called an *absolute MOS* since we now have an anchor quality condition represented by the excellent HD quality.

Such experiments have already been conducted in the past in the speech measurement domain using so-called “full-scale” experiments. Those experiments were designed such that they contained narrow-band, wideband, and full-band voice signals in the same test. How can such experiments look like in the video domain?

Objective and subjective measurements

What is true for subjective tests is also true for objective measurement tools since latter are usually modelled on the earlier. Unless an objective model has been trained for example on both kinds of subjective databases – “full-scale” and “single-scale” – it will output just one or the other kind of result. And unless it has been explained to the user in what context a MOS needs to be interpreted in, the user does not know if a MOS of “4” is good enough or if it indicates a network problem.

Possible solutions

The following is a list of possible solutions how we could address this issue in the future within the ITU-T. This list is by no means meant to be complete, it just gives a starting point for discussions we would like to initiate within the ITU-T.

1. Subjective tests that cover the whole range of audio bandwidths a/o video resolutions in addition to those that are limited to just a particular context. Especially when those databases are intended for training/modelling of objective measurement tools, both kinds should be included.
2. Development of a mapping function between experiments of different contexts, which could be used in objective models.
3. Introduction of an extended MOS terminology stating the context of the respective experiment.

Summary

It was briefly outlined why the MOS and its interpretation that we have become accustomed to may need to change and how such a change could look like. This contribution is intended to bring this issue to the attention of the delegates and start a discussion within the ITU-T on this subject.