Question(s):		**Meeting, date:** VQEG - Rome, 14-18 June 2004
Study Group:	**Working Party:**	**Intended type of document** (R-C-D-TD):
Source:	Psytechnics Limited	
Title:	**Subjective Video Quality Assessment Methods for Multimedia Applications**	

Contact:	Quan Huynh-Thu Psytechnics Limited United Kingdom	Tel: +44-1473-261839 Fax: +44-1473-261880 Email: quan.huynh-thu@psytechnics.com
Contact:	Antony Rix Psytechnics Limited United Kingdom	Tel: +44-1473-261862 Fax: +44-1473-261880 Email: antony.rix@psytechnics.com

Summary

This contribution presents the results of 2 subjective experiments for video quality assessment using the Absolute Category Rating (ACR) and Double Stimulus Continuous Quality Scale (DSCQS) methods [ITU-T P.910, ITU-R BT.500]. Video contents, codecs, bitrates, frame rates and error profiles (packet loss) were selected to be representative of mobile video streaming multimedia applications. The subjective data from the two tests shows that the MOS scores given by subjects in a double-stimulus-based method (DSCQS) are highly correlated with MOS scores given in a single-stimulus-based method (ACR). Furthermore, the DSCQS reference MOS for each HRC was found to be statistically indistinguishable from the average reference MOS over all HRCs. It is proposed that ACR (with Hidden Reference) method should be used for all subjective tests for multimedia applications. While a relatively low number of HRCs may be sufficient for applications with small impairments such as digital broadcast television, a significantly important number of test conditions is required to test and validate objective models for multimedia applications. ACR allows to subjectively assess four times more test conditions, while providing the same essential information than DSCQS for each condition.

Note: This document is an extension of the Delayed Contribution D108 presented at the ITU-T SG9 meeting in Geneva 10-14 May 2004 and is submitted for discussion within VQEG.

Psytechnics Limited

SUBJECTIVE VIDEO QUALITY ASSESSMENT METHODS FOR MULTIMEDIA APPLICATIONS

Table of contents

1 Introduction

Video quality assessment for television applications has received extensive attention from the industry and from the international standardization bodies over the last few years. The Video Quality Expert Group (VQEG) carried out work on video quality for digital television [2], which led to a new ITU-T Recommendation J.144 [3].

However, applications over packet networks such as the Internet and over wireless networks such as video on mobile phones raise different technical issues and challenges. They cover a much wider range of frame sizes, frame rates, bitrates, and errors in the transmission channel; thus they exhibit a wider range of distortions, and in many cases the picture shows significant visible distortion. Network conditions (e.g. congestion or packet loss) are different from the ones occurring in TV transmission. Also, the content is viewed at a short distance on smaller LCD screens with progressive display.

This contribution describes two experiments of subjective video quality assessment designed to simulate typical uses such as Internet video streaming or mobile video streaming on 3G. The two different subjective methods used for comparison are the Double Stimulus Continuous Quality Scale (DSCQS) method as defined by ITU-R BT.500 [1], which has typically been used for testing conditions with small impairments, and the Absolute Category Rating (ACR) method as defined by ITU-T P.910 [4]. The procedures of the two recommendations were slightly altered to use a PC-based testing: voting scales were on-screen graphical sliders for DSCQS, on-screen buttons for ACR, and voting time was not constrained to a certain limit. Video material was processed through a number of Hypothetical Reference Circuits (HRCs) representative of the degradations of the target applications, including compression by an encoder and transmission over a network with packet losses. The same processed video material was used in both ACR and DSCQS experiments.

The document is organized as follows. Section 2 gives a description of the two subjective methods used for the experiments. Section 3 describes the source material and HRCs to produce the test data. Section 4 describes the laboratory set-up. In section 5, we discuss the analysis of the subjective data.

2 Subjective video quality assessment methods

2.1 Double Stimulus Continuous Quality Scale (DSCQS)

DSCQS is included in ITU-R BT.500. This testing methodology was used by VQEG [2] for all the subjective experiments used to test, select and recommend the objective video models included in ITU-T J.144.

In the DSCQS method, a subject is presented with a pair of sequences (A and B) twice; one of the two sequences is the source material while the other is the test material obtained by processing the source material. The subject is asked to evaluate the picture quality of both sequences using a continuous grading scale. The order by which the source and the processed materials are shown is random and is unknown to the subject.

In order to use a purely PC-based procedure, we slightly altered the standard DSCQS method. Subjects are asked to watch the first presentation of the pair of sequences without voting. Subjects then vote on sequence A after its second presentation and on B after its second presentation. Voting was not time-limited. Figure 1 shows the structure and timing of the basic DSCQS test cell. We do not believe that the modifications brought to the standard methodology will have any significant impact on the ratings given by the subjects.

The order of presentation of basic test cells is randomised over the test session(s) to avoid clustering of the same conditions or sequences.

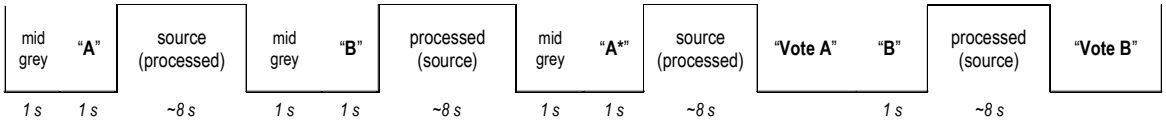


Figure 1 - DSCQS basic test cell

Grading scale

In the test, subjects voted using an on-screen continuous graphical scale divided into five equal intervals with the following adjectives from top to bottom: Excellent, Good, Fair, Poor and Bad. The lowest score is 0 and highest score is 100. The scales are positioned in pairs to facilitate the assessment of the two sequences presented in a basic test cell. The leftmost scale is labelled "A" and the other scale "B". A screenshot from the application is shown in Figure 2.

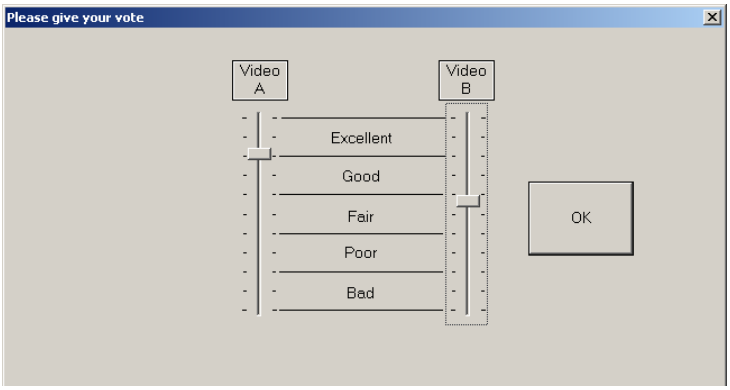


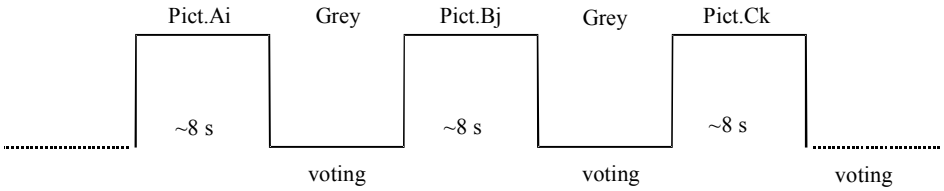
Figure 2 - On-screen voting scales in the DSCQS method

2.2 Absolute Category Rating method (ACR)

The Absolute Category Rating method is a category judgment where the test sequences are presented one at a time and are rated independently on a category scale. This method is also called single-stimulus method. After each presentation, subjects are asked to evaluate the quality of the sequence shown.

The order of the test sequences was randomized such that each subject viewed the video clips in a different order.

The time pattern for the stimulus presentation is illustrated in Figure 3. Voting was not time-limited as a computer-based procedure does not require this.



Ai Sequence A under test condition i
Bj Sequence B under test condition j
Ck Sequence C under test condition k

Figure 3 - Stimulus presentation in the ACR method

Grading scale

In the test, subjects voted using an on-screen five-point scale:

5 Excellent
4 Good
3 Fair
2 Poor
1 Bad

A screenshot from the application is shown in Figure 4.

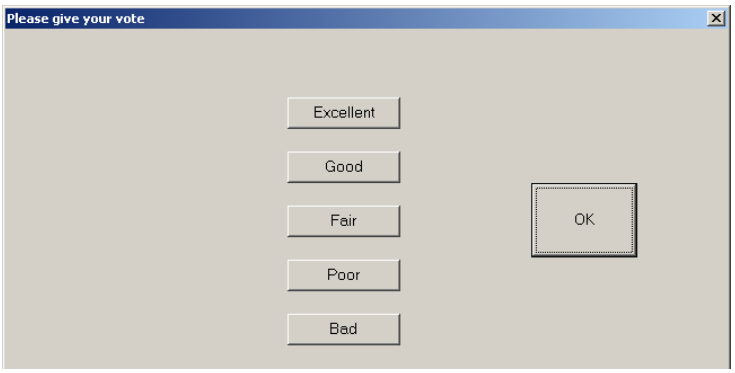


Figure 4 - On-screen voting scale in the ACR method

3 Test material

3.1 Original source material (SRC)

The original source material was selected to include a wide range of genres that are representative of typical mobile video streaming applications such as sports, movie trailers, advertisement, broadcasting news, home video, music video clip, video-conferencing and animation. Contents were further selected to cover a wide range of spatial and temporal complexities based on the spatial (SI) and temporal (TI) perceptual information as recommended by the ITU-T Recommendation P.910. The selected eight pieces of material included the following features:

- Plain areas and complex backgrounds
- Landscapes and human faces

- Slow and simple to fast and complex object motions
- Static camera and moving camera
- Scene changes

Several scenes were obtained from previous VQEG sources while others were in-house sources.

3.2 Test conditions or Hypothetical Reference Circuits (HRC)

Each original source material was processed through a number of Hypothetical Reference Circuits (HRC). Each HRC represents a combination of a codec, a bitrate, a frame rate and a level of packet loss. The HRCs were chosen to be the most representative of target applications. Encoding and transmission parameters were selected to deliver typical video quality for the target applications whilst achieving a wide distribution of qualities over the rating scales. The source material was subjected to both video compression and transmission errors. A total of 13 HRCs were used in each test.

The original source material was processed through the following steps:

1. De-interlacing to progressive display.
2. Cropping of possible black borders if source material was originally prepared for TV broadcast.
3. Frame-resizing.
4. Frame rate conversion.
5. Encoding.
6. Simulation of packet loss.

One HRC in the test was the unprocessed (reference) condition, where the degraded file was identical to the 25fps QCIF reference file. This is termed the hidden reference.

Coding scheme

We considered 2 codecs widely used in multimedia applications such as video streaming and video-conferencing.

Frame size

A frame size of 176 x 144 pixels (QCIF) was used for the two experiments. This frame size is typical for video-capable mobile phones and some video streaming applications on the Internet.

Bit rates

Bit rates ranging from 16 kbits/s to 320 kbits/s were used, as early 3G networks are expected to provide bandwidth up to 384 kbits/s (for both audio and video).

Frame rates

3 different frame rates were considered: 25 fps, 12.5 fps and 5 fps.

Packet loss

A level of 2% of packet loss was introduced in some of the HRCs using in-house software that allows repeatable and controllable simulation of packet loss. Following some previous extensive experiments, this level of packet loss was found suitable to introduce interesting visual distortions, while not completely destroying or completely freezing the contents.

Capture
Video players can introduce additional post-processing (smoothing, filtering) and may apply different error concealment and de-blocking techniques. In order to use video material that are as close as possible to a user's real experience, the encoded video was played using real video playback software and captured using a proprietary capture tool, which keeps track of the exact timing of frame display as encountered during playback (including frame jitter and possible playback irregularities). Captured videos were subsequently trans-coded to uncompressed AVI files. Pixel format in the AVI container was RGB-24 bits uncompressed.

4 Description of the experiment environment

4.1 Testing lab

The subjective testing facility was an acoustically isolated cabinet with controlled lighting following ITU-R BT.500.

4.2 Display

The video sequences were displayed on a standard 15'' LCD screen with a resolution of 1024 x 768 pixels. The use of an LCD screen was motivated by the fact that its usage is becoming more and more popular with PCs, while mobile devices also use LCD screens.

4.3 Presentation

Video sequences were displayed according to the DSCQS and ACR procedures described in section 2, using in-house software that reads AVI files and displays them without introducing any post-processing. 1:1 pixel ratio was used, i.e. QCIF images were not upsampled, and videos were surrounded by a neutral grey background.

4.4 Viewing distance

Viewing distance complied to the range specified by ITU-R BT-500 and ITU-T P.910, but it was left free to each subject to adjust to the most comfortable viewing distance.

4.5 Subjects

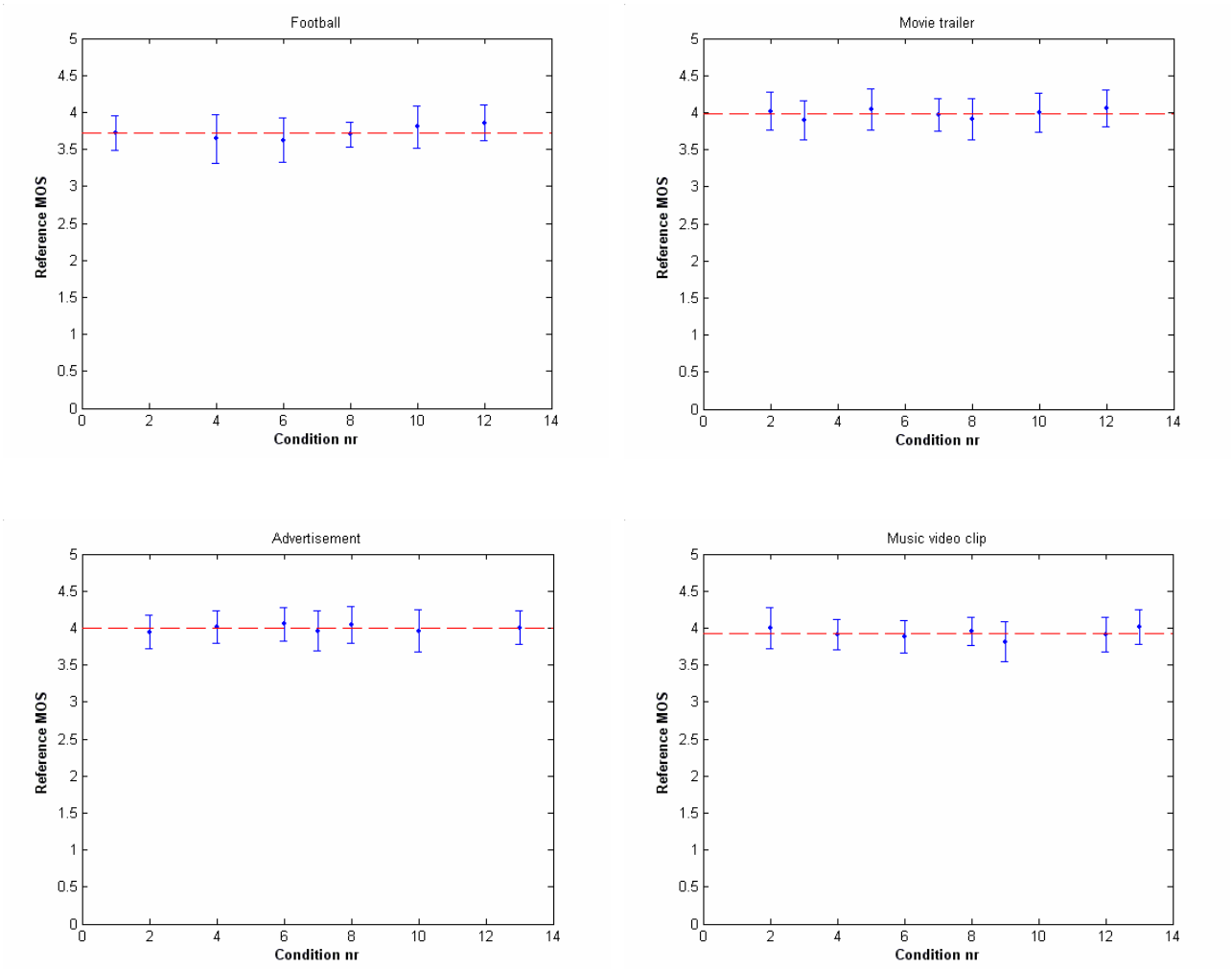
16 non-expert viewers (1 subject viewing at a time) took part to in each of the two subjective tests. No subject took part in both tests. The subjects were given instructions in written form and were shown a short set of preliminary sequences to get familiar with the testing system and the tasks they had to perform.

5 Analysis of subjective data

To facilitate comparison between DSCQS and ACR data, the DSCQS absolute MOS scores were linearly mapped from [0 100] to the 1-5 MOS scale of ACR. By absolute MOS scores, we mean the MOS given either to the reference or the degraded, as opposed to the differential MOS (DMOS) representing the difference MOS between the reference and the degraded sequences.

Using the MOS scores of the degraded files only, average 95% confidence interval is ±0.28 for DSCQS and ± 0.34 for ACR.

Figure 5 shows mean DSCQS MOS for the reference sequences for all HRCs (conditions) for which that reference was used. The mean reference MOS is plotted as a dashed horizontal line. The reference sequences all showed the same property that the reference MOS for any HRC is statistically indistinguishable from the mean over all HRCs, at the 95% confidence level. This indicates that capturing multiple votes on the quality of the reference is an unnecessary redundant process. This also indicates that there is not any dependency between reference MOS scores and degraded MOS scores as degraded MOS scores cover a wide range of ratings on the MOS voting scale.



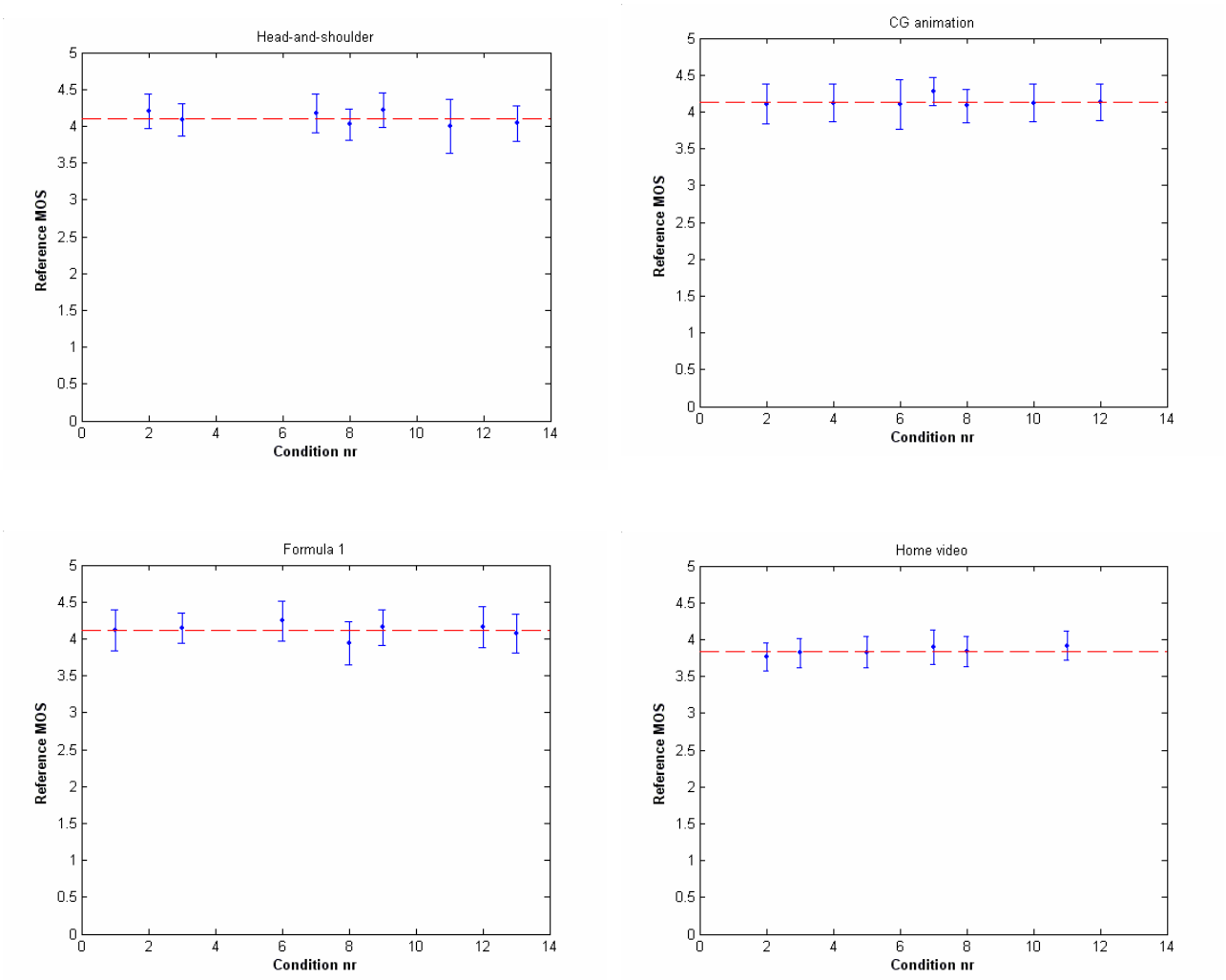


Figure 5 - Mean DSCQS MOS scores for the reference sequences for all HRCs

Figure 6 shows the mean DSCQS MOS scores plotted against mean ACR MOS scores. Figure 6(a) shows the scatter plot for each presented file; (b) shows the results averaged over each HRC. The linear correlation per file in Figure 6(a) is 97.5%, a very high value. The 3rd-order minimum squared error polynomial regression line between the two data sets is also plotted, indicating that there is slight curvature in the relationship. The linear correlation per test condition in Figure 6(b) is 98%.

The approximate symmetric 95% confidence interval is also shown. Note that at the 95% confidence level, it is expected that about 1 in 20 measurements will fall outside the confidence interval and thus it is unsurprising that one test case in Figure 6(a) falls above the interval.

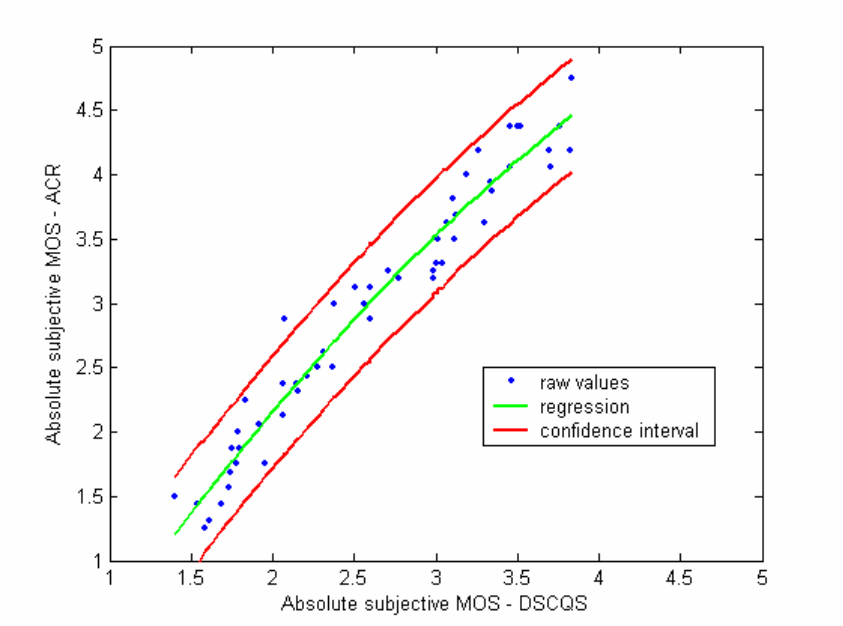


Figure 6 - (a) Degraded DSCQS MOS compared to ACR MOS, per file

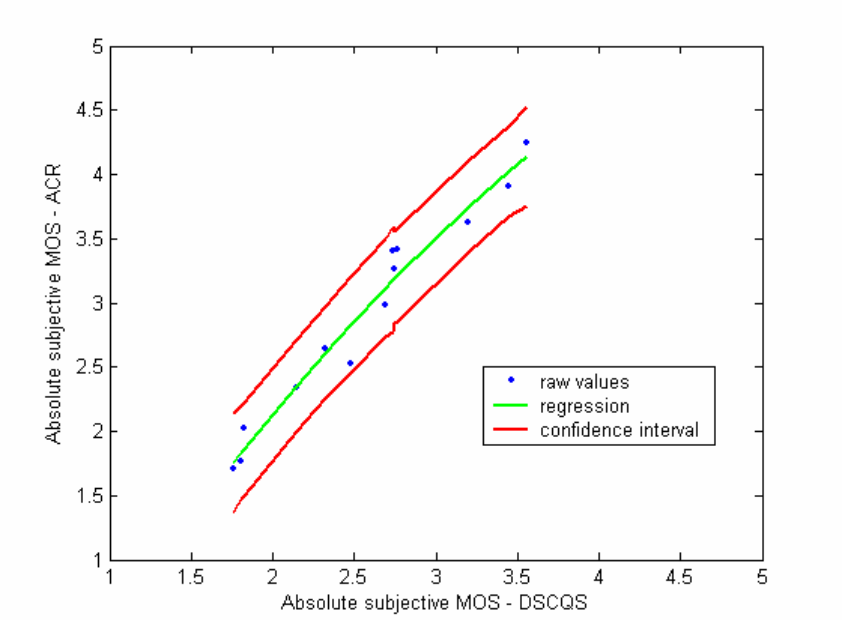


Figure 6 - (b) Degraded DSCQS MOS compared to ACR MOS, averaged over HRC

Additional analysis was carried out using the DMOS scores. For DSCQS, DMOS was computed as the difference between MOS values given to the reference and degraded sequences within each test cell. For ACR, DMOS was computed as the difference between the MOS score given to the hidden reference and the corresponding degraded sequence. Negative DMOS values were left unchanged, i.e. cases where MOS of degraded file was higher than MOS of corresponding reference file. Approximately 8% of the total number of individual DMOS values for all subjects was negative for DSCQS and approximately 3% for ACR. Averaging across subjects, mean DMOS values were

positive for DSCQS while 2 negative DMOS values of -0.06 were found for ACR. These are not statistically significant on the considered scales.

Using the DMOS scores, average 95% confidence interval is ±0.32 for DSCQS and ±0.37 for ACR. Figure 7 shows the mean DMOS values for DSCQS plotted against mean DMOS values for ACR. Figure 7(a) shows the scatter plot for each presented file; (b) shows the results averaged over each HRC. The linear correlation per file in Figure 7(a) is 96.7%, while the linear correlation per condition in Figure 7(b) is 98%. The 3rd-order minimum squared error polynomial regression line between the two data sets is also plotted, together with the confidence interval.

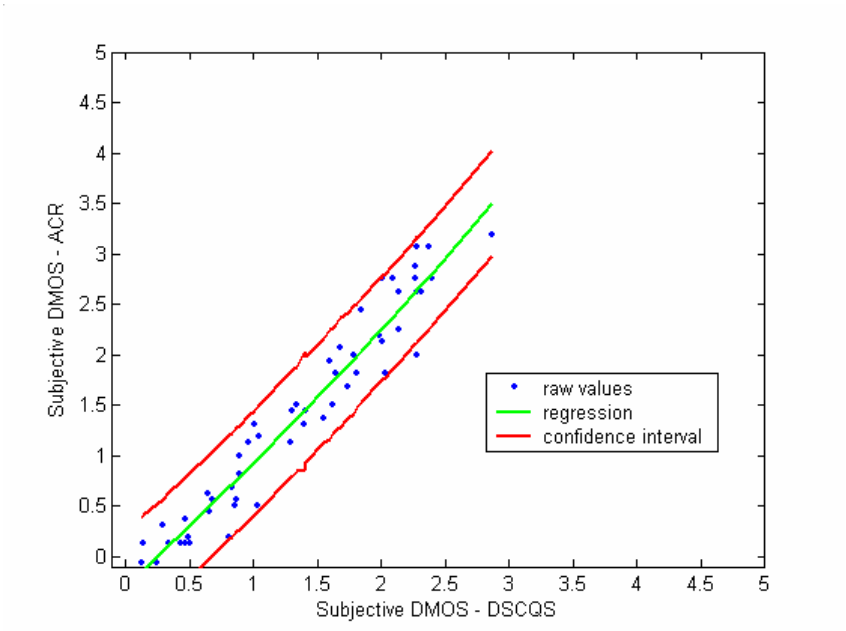


Figure 7 - (a) DSCQS DMOS values compared to ACR DMOS values, per file

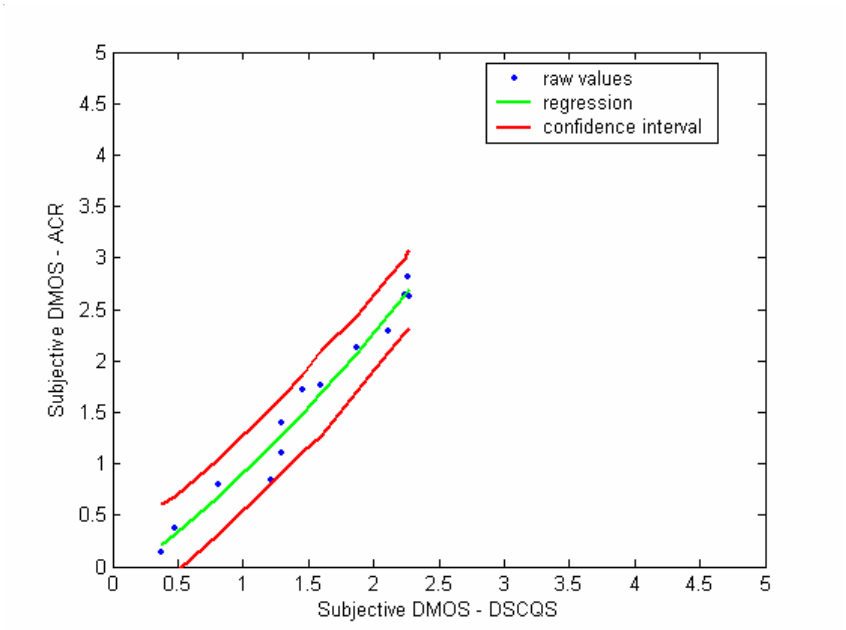


Figure 7 - (b) DSCQS DMOS values compared to ACR DMOS values, averaged over HRC

6 Conclusions

We presented data obtained from 2 subjective experiments for video quality assessment using the Double Stimulus Continuous Quality Scale (DSCQS) and the Absolute Category Rating (ACR) methods. The analysis of the subjective data from the two tests leads to the following conclusions:

- MOS scores given by subjects in a double-stimulus-based method (DSCQS) are highly correlated with MOS scores given in a single-stimulus-based method (ACR with Hidden Reference).
- The DSCQS reference MOS for each HRC was found to be statistically indistinguishable from the average reference MOS over all HRCs. This indicates that it was unnecessary to capture multiple votes on the quality of the reference.
- Thus an ACR method with hidden reference can be used to give equivalent results to DSCQS.

We believe that this close correlation between the ACR and DSCQS methods is found because they both use the same underlying subjective prompt: "excellent"…"bad".

Following discussions with subjects, an issue that was found with DSCQS is an important 'boredom effect', as subjects have to watch each degraded sequence twice before voting and have to watch each reference sequence (SRC) 2 x *m* times over the whole subjective test, where *m* is the number of SRCxHRC combinations considered for that reference sequence. This boredom effect is thought to be more important when viewing videos with small picture size compared to when viewing broadcast-type quality videos on a full screen.

Furthermore, the main disadvantage of DSCQS is that this procedure is time-consuming. For the same length of time, ACR makes it possible to test 4 times as many conditions. ACR (with Hidden Reference) is suitable for subjective tests targeting multimedia applications and allows testing more conditions. As opposed to broadcast quality where a relatively low number of combinations of SRCxHRCs may be sufficient, multimedia applications cover a much wider range of video qualities and a significantly higher number of combinations of SRCxHRCs is required to test and validate objective models.

ACR is already recommended by ITU-T in P.910 as a subjective video quality assessment method for multimedia applications.

7 Proposal

It is proposed that the ITU-T P.910 ACR (with Hidden Reference) subjective video quality assessment method should be used for all subjective tests for multimedia applications in the forthcoming VQEG competition.

8 References

[1] ITU-R Recommendation BT.500-11 (2002), "Methodology for the subjective assessment of the quality of television pictures".

[2] ITU-R Document 6Q/14 (2003), "Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II (FR-TV2)".

[3] ITU-T Recommendation J.144 (2004), "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference".

[4] ITU-T Recommendation P.910 (1999), "Subjective Video Quality Assessment Methods for Multimedia Applications".