HIGH DYNAMIC RANGE SUBJECTIVE TESTING

M. E. Nilsson and B. Allan

British Telecommunications plc, UK

ABSTRACT

This paper describes of a set of subjective tests that the authors have carried out to assess the end user perception of video encoded with high dynamic range technology when viewed in a typical home environment.

Viewers scored individual single clips of content, presented in High Definition (HD) and Ultra High Definition (UHD), in Standard Dynamic Range (SDR), and in High Dynamic Range (HDR) using both the Perceptual Quantiser (PQ) and Hybrid Log Gamma (HLG) transfer characteristics, and presented in SDR as the backwards compatible rendering of the HLG representation.

The quality of HD SDR was improved by approximately equal amounts by either increasing the dynamic range or increasing the resolution to UHD. A further smaller increase in quality was observed in the Mean Opinion Scores of the viewers by increasing both the dynamic range and the resolution, but this was not quite statistically significant.

INTRODUCTION

UHD televisions are now retailing in significant numbers, and UHD services are starting to appear in the market. But while these services offer higher resolution than HD services, further improvement could be made in due course to provide an even better viewing experience.

The next improvement in viewing experience is likely to come from the use of a higher dynamic range for video. Consumer televisions are already shipping with much higher brightness and much higher dynamic range than televisions of only a couple of years ago, and non-consumer displays are capable of much higher brightness still. Standards bodies are debating around the world how high dynamic range should be supported from content capture, through broadcast and distribution channels, to end users on television screens.

In this paper we report the methodology and results of a set of subjective tests to determine how viewers perceive high dynamic range content on a current high-end consumer television, for what we considered to be typical content, mostly shot outdoors in sunny conditions in the UK. We wanted to quantify the benefit of adopting new technological solutions that support higher dynamic range for the delivery of content services to current high-end consumer televisions.

We also wanted to compare two non-linear transfer functions that have been standardised to support high dynamic range video, the Perceptual Quantiser (PQ) as defined in SMPTE ST 2084 (1) and Hybrid Log Gamma (HLG) as defined in ARIB STD-B67 (2). We also

wanted to quantify the effectiveness of the implicit backward compatibility of HLG with the quality of standard dynamic range delivery to current high-end consumer televisions.

TEST CONTENT

1) Bermuda

2) Crowd

3) Cup

4) Lifeguards

5) Mouse

6) Pilots

7) Sailing

8) Sausage

9) Victory

10) Windsurfer

Figure 1 – Single low resolution still images representative of the ten test clips, using BT.709 colour primaries and BT.709/BT.1886 transfer characteristics (4)

BT Sport, with support from BBC and Arri, captured content during an America's Cup World Series event in Portsmouth, UK, 23-26 July 2015, in UHD resolution at 50 frames per second with BT.709 colour primaries (3) using Arri Alexa Mini and Arri Amira cameras. We reviewed the many hours of content captured and selected ten test clips of ten seconds duration for use in subjective testing, as shown in Figure 1. These clips are quite varied, including one indoor scene and one outdoor night-time scene, but are dominated by scenes with bright sunshine and water. We feel these scenes are representative of content that would be broadcast during coverage of an event like the America's Cup.

PROCESSING OF TEST CONTENT

The image processing suite DaVinci Resolve was used to reverse the LogC transfer characteristics applied in the camera during capture, outputting EXR files at UHD

resolution at 50 frames per second with linear light RGB samples in half float format, with the RGB samples being relative to the BT.709 colour primaries.

We developed software to convert these source EXR images to TIFF format, applying first a matrix to map the samples to BT.2020 primaries (5), then applying a single power function, 'gamma', to each sample of each component, then applying a linear scaling factor, and finally applying a non-linear transfer function. The equation below shows the part of this mapping for the red component expressed relative to BT.2020 primaries, from linear sample R, to non-linear sample R', using scaling factor s, exponent γ (hereafter gamma), and an Opto-Electrical Transfer Function (OETF).

$$R' = OETF(s \times R^{\gamma})$$

For SDR, we selected the inverse of the Electro-Optical Transfer Function (EOTF) specified in BT.1886 as the OETF. For PQ we selected the Inverse EOTF specified in SMPTE ST 2084. For HLG, we used the concatenation of an inverse OOTF and the OETF specified in ARIB STD B-67 as the OETF, where the inverse OOTF comprised scaling the colour components by a factor dependent on the luminance, L, as in the equation below, where R_{scaled} would then be subject to the OETF. The value of 1.2 was chosen as the peak brightness of the television is below 1000cd/m^2, and the contrast and gain were determined for black level zero and white level of 800cd/m^2, all as specified in BT.2100-0 (6).

$$R_{scaled} = \frac{L^{\left(\frac{1}{1.2}-1\right)} \times (s \times R^{\gamma}) - Contrast}{Gain}$$

We adopted this methodology as we had imagined that we could choose suitable values of gamma by viewing still images displayed on a Sim2 monitor with appropriate peak brightness. However, this proved not to be possible, as the selected values of gamma did not result in good quality video clips when played on the television. Hence we adopted the approach described below.

The TIFF images produced by the process above were encoded at HD resolution (1920x1080 at 50fps) as ten second video sequences using an Ateme Titan File Encoder to generate HEVC (7) compressed video streams at a bit rate of 30MBit/s within an MP4 file. MP4Box was used to extract the raw HEVC streams, which we processed with our own software to modify the signalling. SDR streams were signalled in the VUI as having transfer_characteristics equal to 1. PQ streams were signalled in the VUI as having transfer_characteristics equal to 16. Two versions of each HLG stream were produced, both having VUI signalling transfer_characteristics equal to 1, but one of them also having periodic repetitions of the alternative_transfer_characteristics SEI message indicating preferred_transfer_characteristics equal to 18. Thus we could generate one encoded stream for HLG, and signal it in one case as HDR and in another case as backwards compatible SDR.

The resulting modified HEVC streams were multiplexed with an arbitrary audio clip into an MP4 file using MP4 box. The audio was never presented to the viewers, and was included solely to prevent the display of the message "audio format not supported".

The values of gamma and scaling factor were chosen for each of the ten test clips, and the training clips, for each non-linear transfer characteristic, using a time-consuming iterative process, where we tried different values until we were satisfied with the quality of the clip

on the television screen. To get good quality HDR, we found that for many clips we needed to use different values of gamma and scale factor for HLG compared to PQ.

We also found that the quality of the backwards compatible HLG SDR representation could be improved by choosing values of gamma and scaling factor different to those which produced an optimal HLG HDR representation. We ultimately decided to produce three encoded representations of each test clip for HLG: one, for which we use the term HDR-focussed, which provided a good HDR representation, as close to the PQ representation as possible; a second, for which we use the term SDR-focussed, which provided a good backwards compatible SDR representation, as close to the SDR representation as possible, and a third, for which we use the term Balanced, which provided, in our opinion, a reasonable balance between the HDR and backwards compatible SDR representations.

We found that when the HLG HDR quality was high, the backwards compatible SDR representation was often very bright and had low contrast. To get a better backwards compatible SDR representation, we frequently had to increase the value of gamma and reduce the scaling factor, but this had the effect of often making the darker parts of the HDR representation too dark.

Although the selection of the content preparation parameters, scaling factor s, and exponent γ, was made using content encoded at High Definition for speed, these eight representations of each clip were generated and encoded at Ultra High Definition (3840x2160 at 50fps) using an Ateme Titan File Encoder to generate HEVC compressed video streams at a bit rate of 30MBit/s for use in the subjective tests.

We also produced two representations in High Definition in addition to the above eight in Ultra High Definition, one being a down-sampled version of the PQ representation, and the other being a down-sampled version of the SDR representation, but converted from BT.2020 primaries to BT.709 primaries. The PQ representation was encoded with HEVC using the same encoder, but at a video bit rate of 8MBit/s, and the SDR representation, which was also progressively scanned at 50fps, was encoded with H.264 using the same encoder at the same video bit rate of 8MBit/s, to be representative of current HD services.

CHARACTERISTICS OF THE TEST CONTENT

Statistics were gathered during the process of preparing the test content. The histogram of pixel luminance values was collected for a single image from each test clip for each representation. These histograms are in general different for SDR, PQ, and the three HLG variants, as different values of gamma and scale factor were selected.

Table 1 shows, for the PQ representation, these values of gamma, and pixel luminance statistics indicating the mean pixel luminance, the lowest and highest pixel luminance values, the 2.5% and 97.5% percentile pixel luminance values, as well as the corresponding two dynamic ranges. We report these two measures of dynamic range, as one corresponds to the absolute maximum range but is subject to extreme individual pixel values, whereas the other gives a range containing 95% of the pixel luminance values. The table intentionally does not indicate the units in which the luminance is measured as the values are simply numerical values at the input to the PQ Inverse EOTF function. They are nominally in cd/m^2, but it would be misleading to suggest that these were precise values that are displayed on the television.

Clip Name	Gamma	Mean Luminance	Minimum Luminance	2.5% Luminance	97.5% Luminance	Maximum Luminance	Full Dynamic Range	95% Dynamic Range
Bermuda	2.0	152	0.518	15	312	597	1152:1	21:1
Crowd	2.4	65	0.107	2	192	561	5246:1	96:1
Cup	2.2	45	0.342	2	256	608	1777:1	128:1
Lifeguards	2.0	176	0.922	9	354	622	675:1	39:1
Mouse	3.4	143	0.020	2	338	417	20829:1	169:1
Pilots	2.2	152	0.097	2	348	424	4378:1	174:1
Sailing	1.9	142	0.435	16	307	322	740:1	19:1
Sausage	1.2	14	0.379	2	51	58	153:1	26:1
Victory	2.0	88	0.537	4	186	255	475:1	47:1
Windsurfer	6.0	137	0.013	1	259	2721	202757:1	259:1

Table 1 – Statistics of a representative image from each test clip prior to PQ Inverse EOTF.

SUBJECTIVE TEST METHODOLOGY

The subjective quality evaluation was performed at Adastral Park, in a room with controlled lighting, but not otherwise specifically designed for subjective testing. The aim was to replicate a home viewing environment as closely as we could in a workplace room. The background room illumination was set to be about 20 lux.

Figure 2 – Room configuration for subjective quality evaluation

Test content was presented on a Samsung 65" JS9500 Curved LCD TV, with test clips played back continuously and automatically from USB storage.

Two viewers, separated by a partition, viewed and scored the test content simultaneously. They were seated about 2.6m from the television, a distance that was found by the BBC to be the median absolute TV viewing distance in the UK in a survey carried out in 2014 (8).

We used the Absolute Category Rating method as specified in ITU-T Recommendation P.910 (9). This is a single stimulus category judgment method, intended for multimedia applications, where the test sequences are presented one at a time and are rated independently on a category scale. After each clip is presented, the viewers are asked to evaluate the quality of the clip shown.

The time pattern for the stimulus presentation is shown in Figure 3. Each ten second test clip is preceded by a seven second period during the middle three seconds of which the clip number is presented. This duration preceding each clip was chosen because the software running on the Samsung One Connect box causes a banner consisting of the filename and a progress bar to be displayed at the start of each clip, and we considered it essential that this disappeared at least one second before the start of the actual video clip, and not at the same time that the clip number disappeared. Following presentation of each

clip, after one second of black screen, text asking the viewer to vote on the clip was presented for two seconds. The test unit was therefore 20 seconds.

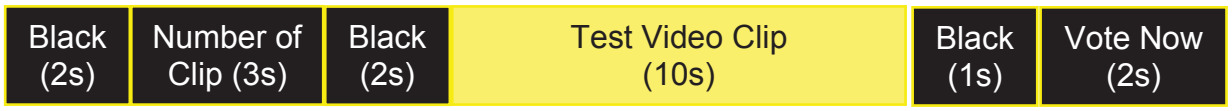

Figure 3 – Stimulus presentation time pattern

Viewers scored each clip independently on the nine-level scale shown in Figure 4.

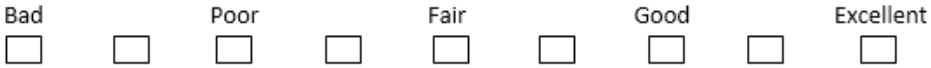

Figure 4 - Nine-grade numerical quality scale

Viewers were shown four training clips, being four additional clips from the content captured at the America's Cup World Series event, represented with gamma, scaling and OETF combinations that achieved video qualities representative of the range of qualities that would be seen during the test.

For each test clip, there were ten representations. Eight of these were at UHD resolution (3840x2160 at 50fps): SDR, PQ, and for HDR-focussed, SDR-focussed, and Balanced HLG, the HDR representation and the backwards compatible SDR representation. Two were at HD resolution (1920x1080 at 50fps): SDR using BT.709 colour primaries and PQ.

As ten test clips were used, there were in total 100 clips to be scored by the viewers. This was too many for each viewer to score each clip. We divided the 100 clips into three groups of 33 or 34, with each group containing six or seven presentations of each source content and six or seven presentations of each encoding format. We created three playlists, termed A, B and C, each comprising two of these three groups. A group that was included in the first half of one playlist, was included in the second half of another playlist. The ordering of clips within a group was different in each playlist in which it occurred.

All viewers scored 66 or 67 test clips, in a test that lasted about 23 minutes, being 67 x 20s, following the period of training.

Approximately the same number of viewers viewed each playlist. Each playlist was structured so that the same content was never shown consecutively. Also the playlists were defined so that each test condition (clip and representation) was never preceded by the same test condition in either of the other two playlists.

A total of 122 viewers (94 male and 28 female) took part in the subjective tests, of which 27 considered themselves to be expert viewers. The viewers had an approximately uniform distribution of ages from 15 to 54, with ten viewers older than 54. Prior to taking part in the testing, viewers were screened for visual acuity using a Snellen chart and for colour blindness using Ishihara charts. One viewer had 20/30 vision, ten had 20/25 vision and the remainder had vision 20/20 or better, a large proportion being significantly better. Four viewers were colour blind, but they had above average visual acuity. We chose not to eliminate any viewers based on their eyesight. We applied the outlier identification process described in Annex 2 of ITU-R Recommendation BT.500 (10), despite having a large number of viewers, and found no outliers.

RESULTS

The viewers' scores for each clip were mapped to values in the range 1.0 (Bad) to 5.0 (Excellent) in increments of 0.5. These were averaged across all viewers to determine Mean Opinion Scores (MOS), and 95% Confidence Intervals, as specified in Annex 2 of ITU-T Recommendation BT.500.

In the following charts showing the MOS and Confidence Intervals, we have used blue for SDR, pink for PQ, and red for HLG, and lighter shades for HD resolution.

Averaged over all ten clips, UHD resolution was statistically significantly better than HD resolution, with the overall MOS increasing by 0.128 from 3.617 to 3.745.

Eight of the ten clips had higher MOS for UHD but, due to the smaller number of samples and hence wider confidence intervals, only Pilots and Victory were statistically significantly better than HD.

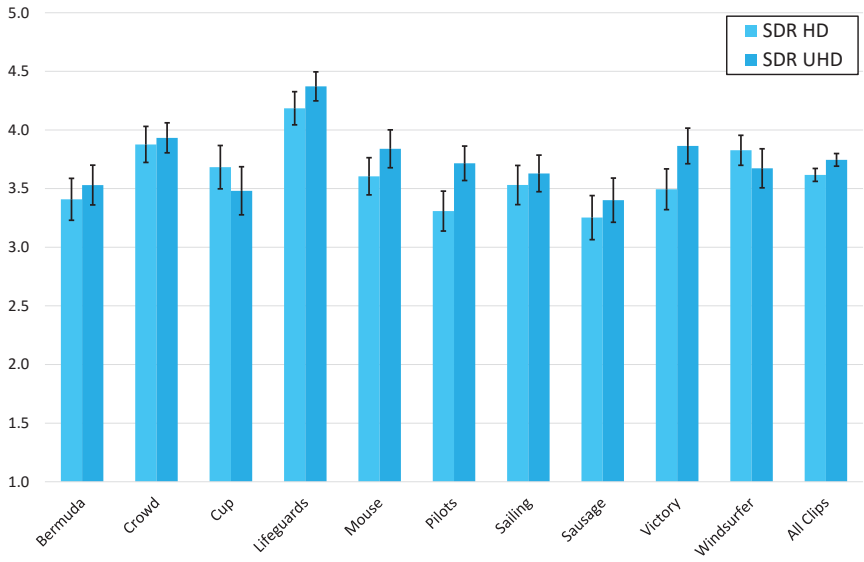

Figure 5 - MOS for each clip in SDR, encoded in HD with H.264 at 8MBit/s, and encoded in UHD with HEVC at 30MBit/s

Averaged over all ten clips, at HD resolution, HDR with PQ was statistically significantly better than SDR, with the overall MOS increasing by 0.126 from 3.617 to 3.743.

Eight of the ten clips had higher MOS for HDR but due to the smaller number of samples, and hence wider confidence intervals, only one of these, Pilots, was statistically significantly better than SDR.

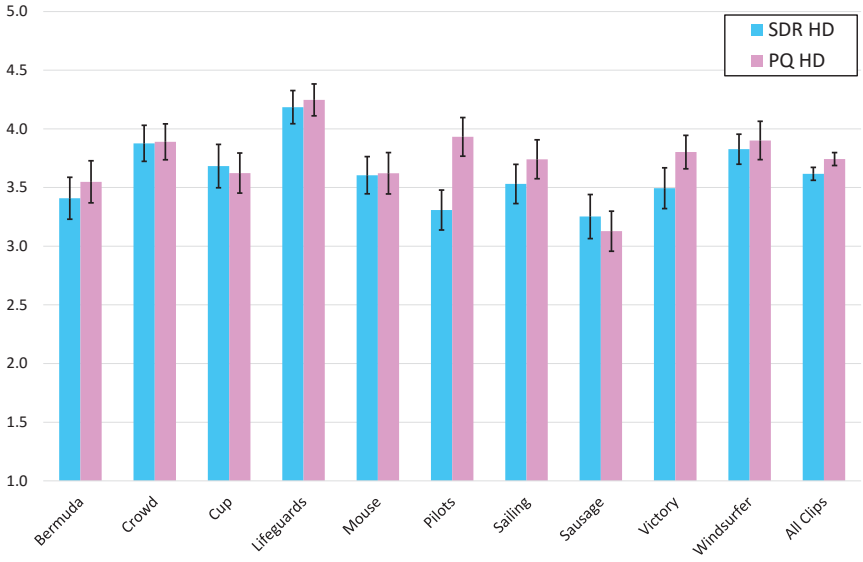

Figure 6 - MOS for each clip in HD, encoded with SDR with H.264 at 8MBit/s, and encoded with HDR using PQ with HEVC at 8MBit/s

Averaged over all ten clips, MOS were higher for UHD HDR, with 3.850 for PQ and 3.817 for HDR-Focussed HLG, compared to 3.745 for UHD SDR.

UHD PQ was nearly statistically significantly better than UHD SDR, with the confidence intervals overlapping by only 0.0038. If the one viewer with near zero correlation with average MOS were eliminated, the confidence intervals would have had a gap of 0.000007.

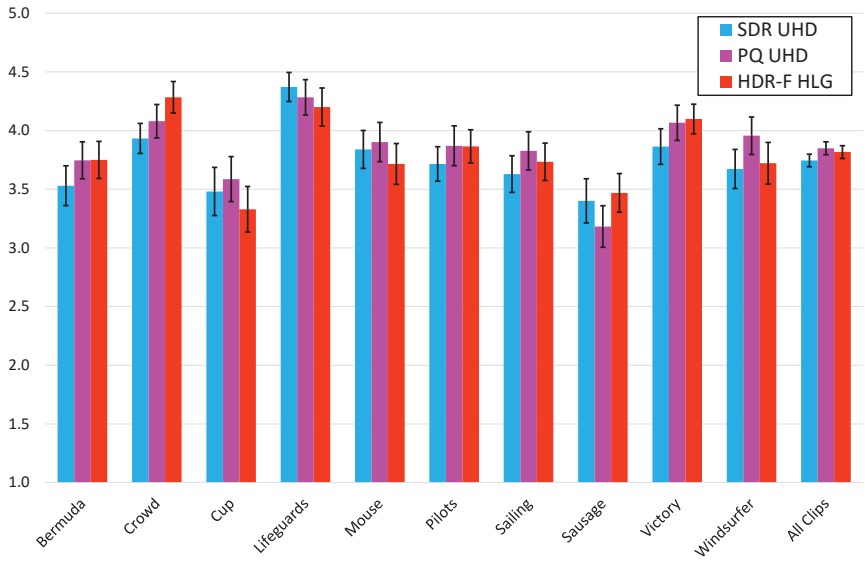

Figure 7 - MOS for each clip in UHD, with SDR, and with HDR using PQ and HDR-Focussed HLG

This result was unexpected as during content preparation we observed improvements when using HDR, notably in the detail in the clouds and in the sparkle on the water. This was presumably not so noticeable to the viewers when the content was presented in a randomised order. We are not surprised by PQ and HDR-Focussed HLG being statistically indistinguishable as we considered them to be very similar during content preparation.

The MOS for each of the encoded formats, averaged over the ten clips, are clustered into four groups.

Within each group the MOS are statistically indistinguishable, but they are statistically distinguishable from all encoded formats in the other three groups.

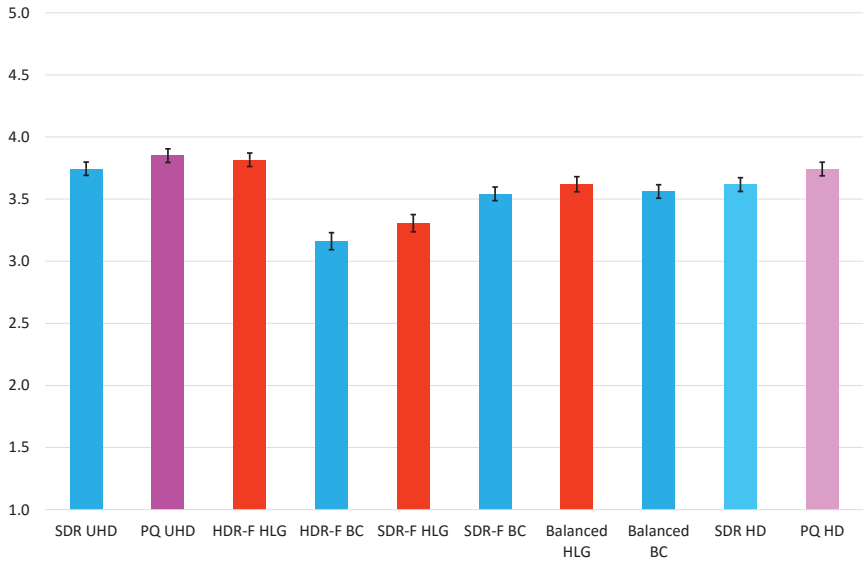

Figure 8 - MOS for each encoded format

The highest quality group contains the four formats, PQ HD, SDR UHD, PQ UHD and HDR-Focussed HLG, which all outperform SDR at High Definition. The lowest quality group contains the Backwards Compatible representation of the HDR-Focussed HLG, which we considered to be bright and low in contrast for many of the clips. The second

lowest quality group contains the HDR representation of the SDR-Focussed HLG, which we considered to be dark for many of the clips.

The poor performance of the two lowest quality groups did not surprise us in one respect, as these results are consistent with our opinions formed during content preparation, but did surprise us in another as before starting the project we expected and hoped for better performance from HLG. We consider it an on-going piece of work to gain a better understanding of these issues with HLG.

The similar MOS for Balanced HLG, 3.620 for the HDR and 3.561 for the Backwards Compatible representation, support the opinion that we generated a good balance between the HDR and Backwards Compatible representations. However, both are statistically indistinguishable from SDR at HD resolution with MOS equal to 3.617.

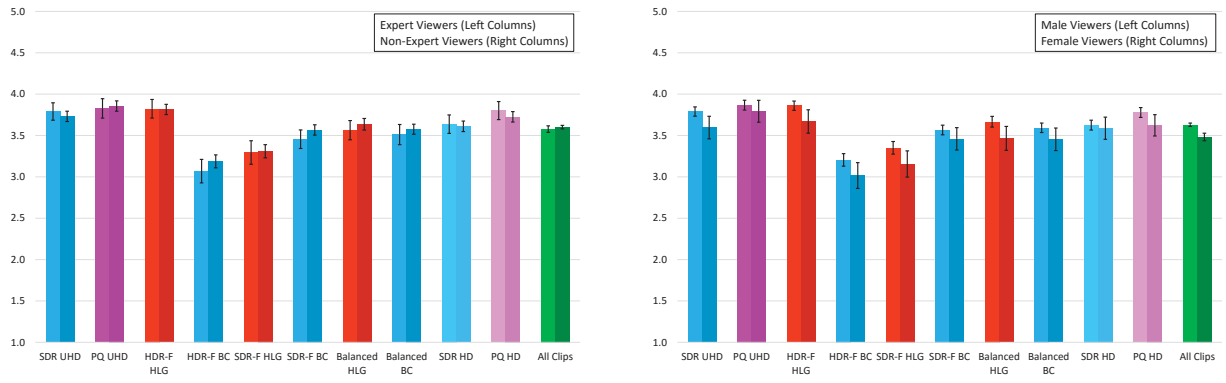

Figure 9 – Overall MOS by viewer expertise (left) and sex (right).

Figure 9 shows that there was no statistical difference between the 27 expert viewers and the 95 non-expert viewers, with the later scoring higher by only 0.024. It also shows that the 28 female viewers' MOS were lower for every format and statistically significantly lower overall, with overall MOS of 3.482 compared to 3.630 for the 94 male viewers.

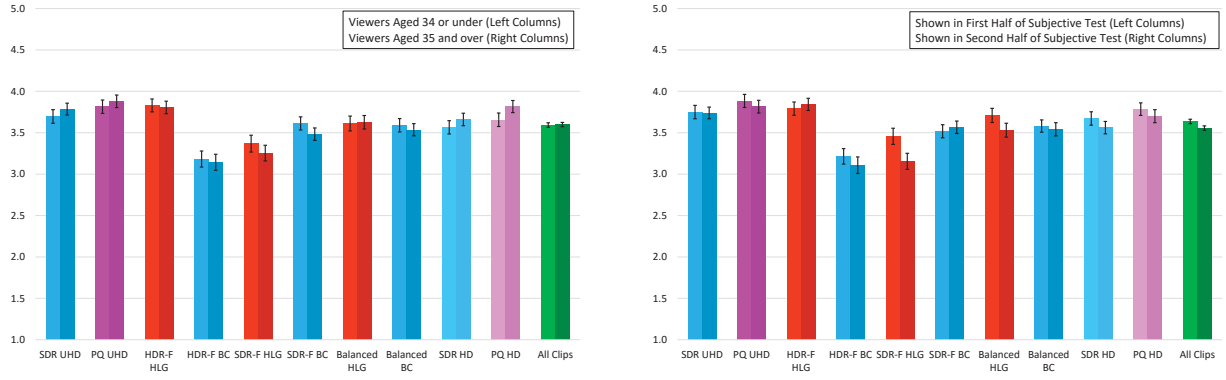

Figure 10 – Overall MOS by viewer age (left) and viewing order (right).

Figure 10 shows that the MOS of the 66 viewers aged 35 or over, 3.599, was statistically indistinguishable from the MOS of the 56 viewers aged 34 or under, 3.593. It also shows that MOS for clips when shown in the first half of the subjective test, 3.636, was statistically significantly higher than the MOS for clips when shown in the second half of the test, 3.556, suggesting that viewers became more critical as the test proceeded.

CONCLUSIONS

We have carried out a set of subjective tests using content that we believe could be typical of a live outside broadcast event, with a large number of viewers in a 'home-like' environment. The test results suggest that the quality of Standard Dynamic Range High Definition services could be improved by approximately equal amounts by either increasing the dynamic range or increasing the resolution to UHD. An additional smaller increase in quality could be achieved by increasing both the dynamic range and the resolution, although this was not quite statistically significant. The tests found that the Hybrid Log Gamma system achieved approximately equal High Dynamic Range video quality as the Perceptual Quantiser scheme, although the performance of its implicit backward compatibility was found to be disappointing. This issue with the Hybrid Log Gamma system requires further study.

REFERENCES

1. SMPTE ST 2084:2014: "High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays".
2. ARIB STD-B67, "Essential Parameter Values for the Extended Image Dynamic Range Television (EIDRTV) System for Programme Production".
3. ITU-R Recommendation BT 709-6, "Parameter values for the HDTV standards for production and international programme exchange".
4. Recommendation ITU-R BT.1886-0, "Reference electro-optical transfer function for flat panel displays used in HDTV studio production".
5. ITU-R Recommendation BT.2020-2, "Parameter values for ultra-high definition television systems for production and international programme exchange".
6. Recommendation ITU-R BT.2100-0: "Image parameter values for high dynamic range television for use in production and international programme exchange".
7. ISO/IEC 23008-2:2015, "Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Part 2: High efficiency video coding".
8. Noland K. and Truong L., 2015. A Survey of UK Television Viewing Conditions. http://www.bbc.co.uk/rd/publications/whitepaper287
9. ITU-T Recommendation P.910 (04/08), "Subjective video quality assessment methods for multimedia applications".
10. ITU-R Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures".

ACKNOWLEDGEMENTS

The authors thank Richard Moreton and his colleagues at Samsung for the loan of a Samsung One Connect box, and for the development of the software for the One Connect box to support the functionality used in the work reported in this paper.

The authors thank Professor Alan Chalmers and his colleagues at the University of Warwick in the UK for the loan of a Sim2 monitor, the provision of a PC player application to use with the Sim2 monitor, and their help and expertise to use this to assist with the preparation of content used in the subjective tests reported in this paper.