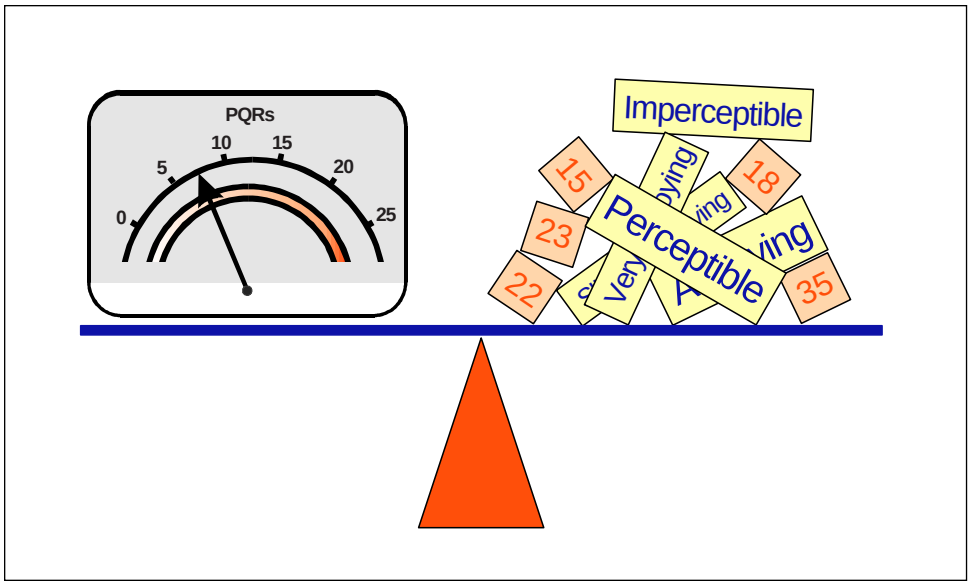
Tektronix

Technical Brief

Comparing Objective and Subjective Picture Quality Measurements



TOPICS DISCUSSED:

- An Overview of Subjective Measurements
- Objective Measurements with the PQA200
- A Comparison Between Subjective and Objective Measurements

Relating Picture Quality Rating (PQR) values to Subjective Scores

Introduction

The advent of the PQA200 Picture Quality Analysis system has enabled rapid objective picture quality measurements, allowing codec developers and those evaluating codecs to perform a broader range of tests than would otherwise have been possible.

Objective picture quality measurements using the PQA200 provide results that are in Picture Quality Rating (PQR) units. Many users and potential users have raised the question, "How does the PQR value relate to the units I have been used to from tests performed under the ITU-R BT.500 standard for subjective measurements?".

In order to answer this, it's necessary to look at both subjective and objective measurements to see how they are performed. It's also valuable to assess other factors which may be apparent, such as how repeatable the measurement may be.

Finally, a comparison may be made between the subjective and objective results. This can provide a figure for correlation, and can also provide a direct mathematical relationship between the two. Factors that will alter this relationship are discussed.

Subjective Measurements

For many years, ITU-R BT.500 has provided standards for subjective picture quality measurement. Over the last few years, this has been extended to deal with the changing measurement requirements for compressed systems. Several measurement techniques are described within this standard, two of which are described below in greater detail.

The first, referred to as "The double-stimulus impairment scale method (the EBU method)" provides a measurement that can be represented within a five-grade impairment scale.

5 = Imperceptible
4 = Perceptible
3 = Slightly annoying
2 = Annoying
1 = Very annoying

Observers see the reference image sequence followed by three seconds of mid-grey followed by the test image sequence. Voting on the impairment is performed based on the adjectival scale described above.

An alternative measurement referred to as "The double-stimulus continuous quality-scale method" (DSCQS), replaces the five-grade scale above with voting on the quality of both the reference image sequence and the test image sequence. The difference between the two represents the impairment. Typically, the two images are displayed in turn, but the viewer is not told which is the reference. The form used for voting con-

tains five adjectives relating to the absolute picture quality: excellent, good, fair, poor, and bad. However voting scores are not restricted to these five adjectives and are instead given values on a continuous scale with values between 0 and 100, where 0 and 100 represent, respectively, bad and excellent picture quality. The resulting difference values will also generally lie within the range of 0 to 100 where 0 and 100 represent respectively low and high levels of impairments. (Negative values can exist when voters consider the picture quality of the test image to be better than that of the reference image.)

Picture quality measurements obtained from subjective tests are subject to variations caused by many factors. These include physical details of the experiment including viewing distance and instructions to voters to weight their judgement more heavily on loss of detail rather than the overall image (usually observers are asked to assess the overall quality) or interpretation of the adjectives used for describing either the impairment or the absolute quality as they are translated into different languages). Finally, different viewing panels will produce differing subjective results. While increasing the number of participants in the panel may reduce this factor, financial constraints normally preclude this option.

Many of these factors may be regarded as virtues of subjective picture measurement in that they enable the selection of test conditions to meet specific test requirements, but they lead to inevitable differences in results obtained from different laboratories. ITU-R BT.500 states that a study of consistency between results obtained from tests performed by different laboratories shows that there can be systematic differences.

Objective Measurements

As with the subjective tests described above, the PQA200 uses both reference and test image sequences to make its measurements. Measurements of picture quality, again actually measurements of impairments, are provided as PQR values. Within the PQA200, a human-vision model is used to assess visible differences for every pixel. These are used to compute PQRs on a frame-by-frame basis so that a plot can be made of PQR values against time. They are also smoothed (Q_{norm} weighting average) to provide a single PQR value for a sequence.

A PQR value of zero indicates that the test sequence is an exact copy of the reference, as we may anticipate with a D1 dub. Increasing PQR values represent greater impairments and can be loosely interpreted as follows:

- A PQR rating of 1 indicates impairments that have a small perceptual impact.
- A PQR rating of 3 indicates impairments that are almost always observable but not strong.
- A PQR rating of 10 indicates impairments that are clearly observable.

Unlike subjective measurements, PQR values are constant and repeatable. For a given video sequence being passed through a given codec at a specified bit rate, PQR values will remain the same. Compression dictates that the complexity of the sequence will impact the PQR rating. Typically, the more complex the sequence, the greater the impairment and the higher the PQR rating, though this will, of course, be dependent on the device-under-test. Encoders from different manufacturers will provide different PQR values for the same sequence, and will potentially exhibit lowest PQR values with different sequences. Because of this, it is recommended that a wide variety of sequences be used for evaluation of codecs.

A Comparison Between Objective and Subjective Results

It's obviously vital that objective picture quality measurement results can be shown to correlate well with subjective results. An investigation performed jointly by IRT and

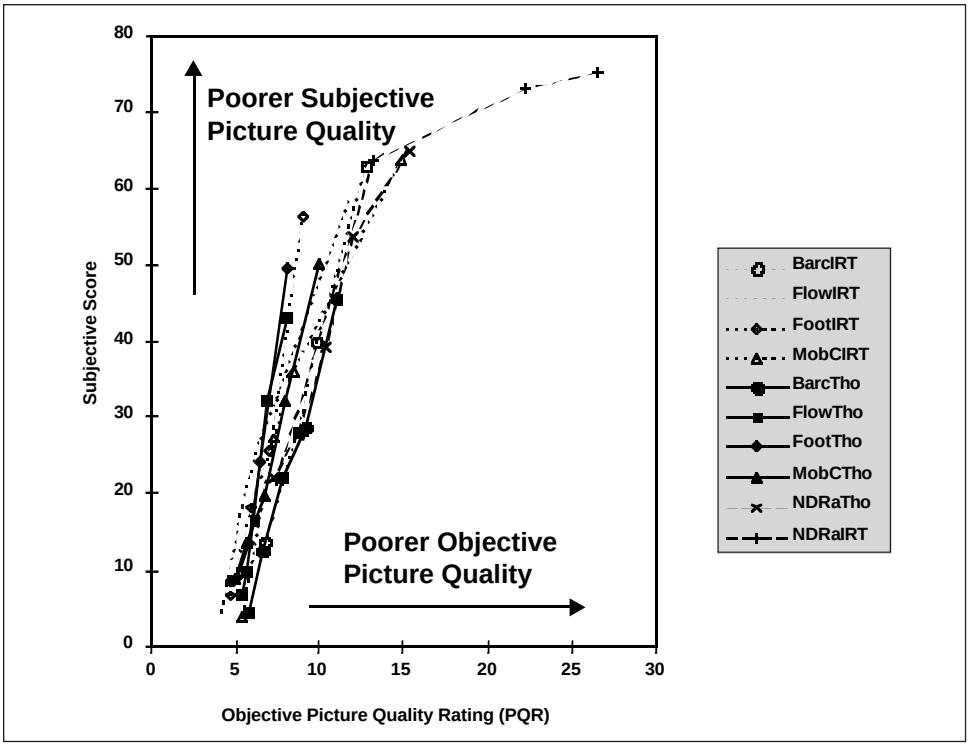


Figure 1. Comparison of Subjective and Objective Picture Quality Rating for 2 to 10 Mbit/s MPEG-2 Video

Tektronix (see Reference 3) has shown very promising results for the PQR ratings provided by the PQA200 and further work is being performed by the Video Quality Experts Group.[1]

The graph in Figure 1 shows plots of PQR ratings plotted against subjective scores obtained from DSCQS subjective tests. Each line plots the values obtained for a specific video sequence being passed through a given codec at a number of different bit rates. Video sequences include Barcelona, Flower Garden, Football, Mobile and Calendar, and a sequence from NDR. Two video coders from IRT and Thomson were employed at bit rates of 2.0, 3.0, 4.5, 7.0, and 10.0 Mbit/s. The high correlation between PQR values and the subjective scores is obvious. (Note that the two points at the upper right represent sequences that contain scenes with the poorest quality, including catastrophic failure of the coder in certain regions. The quality levels for these fall well below what would be expected for broadcast quality.)

As noted above, the PQR ratings for a given sequence and a given codec at a given bit rate will remain constant. However, as noted earlier, the subjective ratings may vary depending on the choice of test conditions and also the group of people selected to perform the test. While this does not change the correlation, it does alter the slope of the graphs. In other words, the variation in subjective scores has the effect of altering the vertical scaling on the graph shown in Figure 1. This makes it difficult to establish a constant relationship between PQR values and different sets of subjective scores.

This can be seen in Figure 2, which shows the three scales and their relationship.

Note also that the PQR scale can be seen to have a zero offset. This is due to the ability of the PQA200 to measure picture impairments that are below the visible threshold. Subjective measurements are obviously limited to measurement of visible impairments.

As noted earlier, PQR values for given sequences and codec set-ups are repeatable and constant. These are represented on the left-hand scale. To the right are the two scales associated with subjective measurements. The first of these is the DSCQS scale with values between 0 and 100, and the second the five-grade impairment scale with values between 1 and 5.

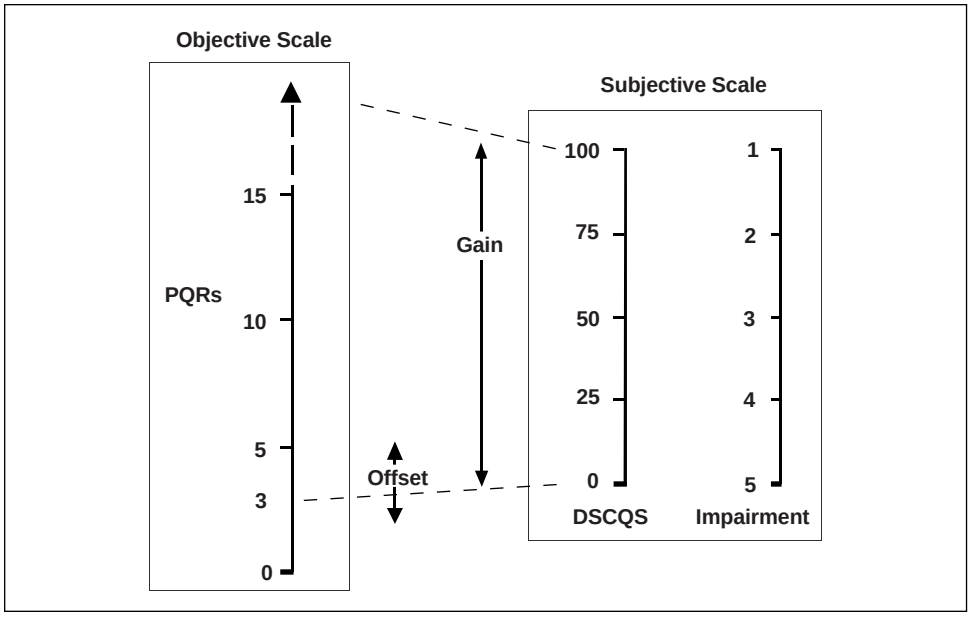


Figure 2. Relationship between Subjective and Objective Picture Quality Rating Scales.

PQR values can be related directly to subjective results for a specific set of subjective tests. By default, this relationship will define a given offset and gain that enables a direct calculation of the DSCQS score from a PQR value. In Figure 2, an offset of 3 PQR is shown. This represents a PQR value that may be equated to imperceptible impairments for a given set of subjective tests. (Note that this offset is also apparent in the results shown in Figure 1, but for these particular subjective tests, the offset is approximately 4.5 PQR.) In addition to this offset, there will be a gain that can be calculated relating the PQR value to the DSCQS score. In the above diagram, this can be seen to be approximately 6.3 (PQR values of 3 and 11 relate approximately to subjective scores of 0 and 50). For the particular subjective tests shown in Figure 1, the gain can be seen to be approximately 9 (PQRs of 5.5 and 11 relate approximately to subjective scores of 10 and 60). Both these values are only valid for the specific set of subjective results. A second set of subjective results is likely to yield a different pair of values. This is a result of the systematic differences between sets of subjective results; i.e., variations in the choice of test conditions and also the group of people selected to perform the test.

It's perhaps easier to define the relationship between the two subjective scales. A DSCQS result of 0 indicates no visible impairments and may therefore be equated to a 5 on the five-grade impairment scale. In the worst case, a value of 100 on the DSCQS scale representing the worst

[1] VQEG is an independent group within ITU with members from Europe, Japan, and the US. Its work involves assessing algorithms that may be used for objective picture quality measurement with the goal of recommending a method, or methods, for objective measurement of digital video quality based on their correlation to subjective results.

impairment may be equated to a 1 on the five-grade impairment scale. It should be noted, however, that this procedure is only an approximation since there is no guarantee of a direct psychophysical correspondence to the formal mathematical conversion from one subjective scale to the other.

While mathematically, knowledge of the offset and gain will enable a calculation of the PQR value that relates to a value of 100 on the DSCQS scale, such calculations are of little value. Correlation between subjective and objective results with such high level of impairments is known to be poor, and one might anticipate that such levels of impairments would be well outside the operational range used within broadcast applications.

Conclusions

It can be seen from the discussion above that objective results represented in PQRs can be related to subjective results in either scale, providing that care is taken.

The factors to be considered are summarized below:

- PQR results for a given video sequence and a given codec set-up are constant and repeatable. Subjective results are likely to change with the test conditions and changes in voting panels.
- PQRs enable measurements of impairments below the threshold of visible perceptibility (a PQR of 0 represents a perfect digital copy). Subjective measurements are limited to visibly perceptible impairments.
- Good correlation has been established in the important areas of low and moderate impairments, covering normal broadcast operational usage. At high levels of impairments which fall well outside normal operational practice, correlation is lower.
- Both subjective and objective picture quality measurements of compressed systems are dependent on the video sequence used for the test.

PQR results provide a constant measure of picture quality for a given sequence and codec set up, thus enabling valuable comparison of results taken at different times and in different locations. Evaluation of new firmware in a codec becomes fast and easy. Comparison of codecs can be documented easily, and additional tests performed that may, in the past, have been omitted due to time constraints. In addition, PQR maps indicating the areas of the picture that have the most noticeable differences provide an excellent means of isolating and documenting faults.

While the above refers primarily to codec tests, and implies a single pass, it can be seen that the PQA200 is very valuable for digital video recorders that use compression. It also enables measurements on concatenated codecs or multi-generation recordings. Finally, it provides the ability to measure impairments below the visible threshold, enabling tests that are impossible with traditional subjective testing.

The PQA200 and it's results represent a breakthrough in picture quality measurement. The relationship between PQRs and subjective measurements is clear, but care needs to be taken in the comparison. As with all measurements, a full understanding of the nature of the measurements is required in order to ensure their validity.

References

1) ***ITU-R BT.500-7 Methodology for the Subjective Assessment of the Quality of Television Pictures.***
2) ***A Guide to Picture Quality Measurements for Modern Television Systems***, Tektronix 25W-11419-0.
3) ***IRT/Tektronix Investigation of Subjective and Objective Picture Quality for 2-10Mbit/sec MPEG-2 Video: Phase 1 Results.*** Available at: http://grouper.ieee.org/groups/videocomp/index.html as tek-irt.zip

Acknowledgements

The author wishes to acknowledge discussions with Mr. Alexander Schertz of IRT while writing this paper.

Written by Lionel Durant, European Market Specialist, Video Technologies

For further information, contact Tektronix:

Worldwide Web: For the most up-to-date product information, visit our web site at www.tektronix.com.

ASEAN Countries (65) 356-3900; **Australia & New Zealand** 61 (2) 9888-0100; **Austria, Central Eastern Europe, Greece, Turkey, Malta, & Cyprus** +43 2236 8092 0; **Belgium** +32 (2) 715 89 70; **Brazil and South America** 55 (11) 3741-8360; **Canada** 1 (800) 661-5625; **Denmark** +45 (44) 850 700; **Finland** +358 (9) 4783 400; **France & North Africa** +33 1 69 86 81 81; **Germany** + 49 (221) 94 77 400; **Hong Kong** (852) 2585-6688; **India** (91) 80-2275577; **Italy** +39 (2) 25086 501; **Japan (Sony/Tektronix Corporation)** 81 (3) 3448-3111; **Mexico, Central America, & Caribbean** 52 (5) 666-6333; **The Netherlands** +31 23 56 95555; **Norway** +47 22 07 07 00; **People's Republic of China** 86 (10) 6235 1230; **Republic of Korea** 82 (2) 528-5299; **South Africa** (27 11) 651-5222; **Spain & Portugal** +34 91 372 6000; **Sweden** +46 8 477 65 00; **Switzerland** +41 (41) 729 36 40; **Taiwan** 886 (2) 2722-9622; **United Kingdom & Eire** +44 (0) 1628 403300; **USA** 1 (800) 426-2200.

From other areas, contact: Tektronix, Inc. Export Sales, P.O. Box 500, M/S 50-255, Beaverton, Oregon 97077-0001, USA 1 (503) 627-6877.

ISO 9001

Tektronix

11/98 TD/XBS 25W–12866–0