

RECOMMENDATION ITU-R BT.1676

**Methodological framework for specifying accuracy
and cross-calibration of video quality metrics**

(Question ITU-R 44/6)

(2004)

The ITU Radiocommunication Assembly,

considering

- a) that digital TV and HDTV utilizing bit-rate reduction technologies such as MPEG-2, DV and others have achieved widespread use;
- b) that the Radiocommunication Sector is responsible for setting the overall quality performance of broadcasting chains;
- c) that impairments to television pictures can be shown to correlate with measurable features of the signals;
- d) that overall picture quality is related to the combination of all impairments;
- e) that in the case of digital TV it is necessary, in particular, to assess the performance of bit-rate reduction methods both in terms of subjective and objective parameters;
- f) that for television systems a number of objective picture quality parameters as well as associated performance measurement and monitoring methods have been developed for the studio environment and in broadcasting;
- g) that full reference objective picture quality measurement methods are useful in evaluating studio and broadcasting systems;
- h) that data sets of test materials, subjective scores and objective values are used in the validation testing of objective picture quality measurement methods;
- j) that there are a number of proposed full reference video quality metrics (VQMs) that can be used to provide objective picture quality ratings;
- k) that there are a number of well-known statistical evaluation methods documented in the literature that can be used to validate and compare VQMs based on data sets of test materials, subjective scores and objective values;
- l) that when one or more VQMs are accepted as normative in ITU Recommendations there will still be a need to estimate the mathematical accuracy (resolving power) of the VQM being used;
- m) that cross-calibration of full reference objective picture quality measurement methods based on available data sets is important for international exchange of measurement and monitoring results,

recommends

- 1 that the calculations specified in Annex 1 be used to estimate the accuracy and cross-calibration of objective picture quality measurements utilizing the full reference method;
- 2 that the calculations specified in Annex 1 may be used as one of several methods to determine the accuracy in evaluation and validation of various objective picture quality measurements utilizing the full reference method.

Annex 1

Method for specifying accuracy and cross-calibration of VQMs

1 Scope

VQMs are intended to provide calculated values that are strongly correlated with viewer subjective assessments. This Recommendation provides:

- Methods for curve fitting VQM objective values to subjective data in order to facilitate the accuracy calculation and to produce a normalized objective value scale that can be used for cross correlation between different VQMs.
- An algorithm (based on statistical analysis relative to subjective data) to quantify the accuracy of a given VQM.
- A simplified root mean square error calculation to quantify the accuracy of a VQM when the subjective data has roughly equal variance across the VQM scale.
- A method to plot classification errors to determine the relative frequencies of “false tie”, “false differentiation”, “false ranking”, and “correct decision” for a given VQM.

The methods specified in this Recommendation are based on objective and subjective evaluation of component video such as defined by Recommendation ITU-R BT.601 using methods such as described in Recommendation ITU-R BT.500 – Methodology for the subjective assessment of the quality of television pictures. A data set for a VQM will consist of objective values and mean subjective scores for a variety of motion video sources (SRCs) processed by a variety of hypothetical reference circuits (HRCs). An example of such a data set is given in ITU-T Document COM 9-80-E, Final report from the video quality experts group on the validation of objective models of video quality assessment.

The methods specified in this Recommendation are directly applicable to a defined data set as described above. For measurements not specifically part of the data set the methods specified in this Recommendation provide a reasonable estimate of accuracy and cross-calibration for applications that can be considered to be similar to and within the scope of the defined data set.

The methods specified in this Recommendation are appropriate for use in combination with other statistical calculations in order to evaluate the usefulness of a VQM. Informative information regarding the use of the methods is presented in Appendix 1. A complete verification process by suitable independent laboratories is required for a VQM to be considered for inclusion as a normative part of an ITU-R Recommendation.

2 Accuracy of a VQM

In order to use an objective VQM, one must know whether the score difference between two processed videos is statistically significant. Hence, a quantification is needed of the accuracy (or resolving power) of the VQM. To visualize this resolving power, it helps to begin with a scatter plot in which the abscissa of each point is a VQM score from a particular video SRC and distortion HRC, and the ordinate is a subjective score from a particular viewing of the SRC/HRC. Each SRC/HRC combination (associated with a particular VQM score) contains a distribution of mean subjective scores, S , based on a number of viewers, which represents (approximately) the relative probabilities of S for the particular SRC/HRC combination. The resolving power of a VQM can be defined as the difference between two VQM values for which the corresponding subjective-score distributions have means that are statistically different from each other (typically at the 0.95% significance level).

Given this qualitative picture, two metrics for resolving power will be described in this section, each one useful in a different context. The metrics are described in § 2.3 and 2.4. Also, in § 2.5, a method is described for evaluating the frequencies of different kinds of errors made by the VQM. As an example of implementation of all the methods, a computer source code in MATLAB (The Mathworks, Inc., Natick, MA) is provided in Appendix 2.

2.1 Nomenclature and coordinate scales

Let each SRC/HRC combination in a data set be called a “situation”, and let N be the number of situations in this data set. A subjective score for situation i and viewer l will be denoted as S_{il} , and an objective score for situation i will be denoted as O_i . Averaging over a variable such as viewer will be denoted with a dot in that variable location. For instance, the mean opinion score of a situation will be denoted as $S_{i\bullet}$. The subjective-score statistics from each pair (i, j) of these situations are to be assessed for significance of VQM difference, and then used to arrive at a resolving power for the VQM difference, as a function of the VQM value.

Prior to any statistical analysis, the original subjective mean opinion scores $S_{i\bullet}$ are linearly transformed to the interval $[0, 1]$, defined as the Common Scale, where 0 represents no impairment and 1 represents most impairment. If best represents the no-impairment value of the original subjective score and worst represents the maximum impairment of the original subjective scale, then the scaled scores $\hat{S}_{i\bullet}$ are given by:

$$\hat{S}_{i\bullet} = \frac{S_{i\bullet} - \text{best}}{\text{word} - \text{best}}$$

Next, the VQM scores are transformed to this common scale as a by-product of the process of fitting the VQM scores to the subjective data, which will be discussed in the following section.

2.2 Fitting VQM values to subjective data

Fitting removes systematic differences between the VQM and the subjective data (e.g. d.c. shift) that do not provide any useful quality discrimination information. In addition, fitting all VQMs to one common scale will provide a method for cross-calibration of those VQMs.

The simplest method of data fitting is linear correlation and regression. For subjective video quality scores, this may not be the best method. Experience with other video quality data sets indicates chronically poor fits of VQM to subjective scores at the extremes of the ranges. This problem can be ameliorated by allowing the fitting algorithm to use non-linear, but still monotonic (order-preserving), methods. If a good non-linear model is used, the objective-to-subjective errors will be smaller and have a central tendency closer to zero.

Non-linear methods can be constrained to effectively transform the VQM scale to the $[0, 1]$ common scale. Besides improving the fit of data with a VQM, a fitting curve also offers an additional advantage over the straight-line fit implied by the Native Scale (i.e. the original scale of the VQM): the distribution of objective-to-subjective errors around the fitted model curve is less dependent on the VQM score. Of course, the non-linear transformation may not remove all the score dependency of objective-to-subjective errors. To capture the residual dependence, it would ideally have been useful to record objective-to-subjective error as a function of VQM value. However, typical data sets are too small to divide among VQM bins in a statistically robust way. Therefore, as will be clear in § 2.3, a sort of average measure over the VQM range is computed.

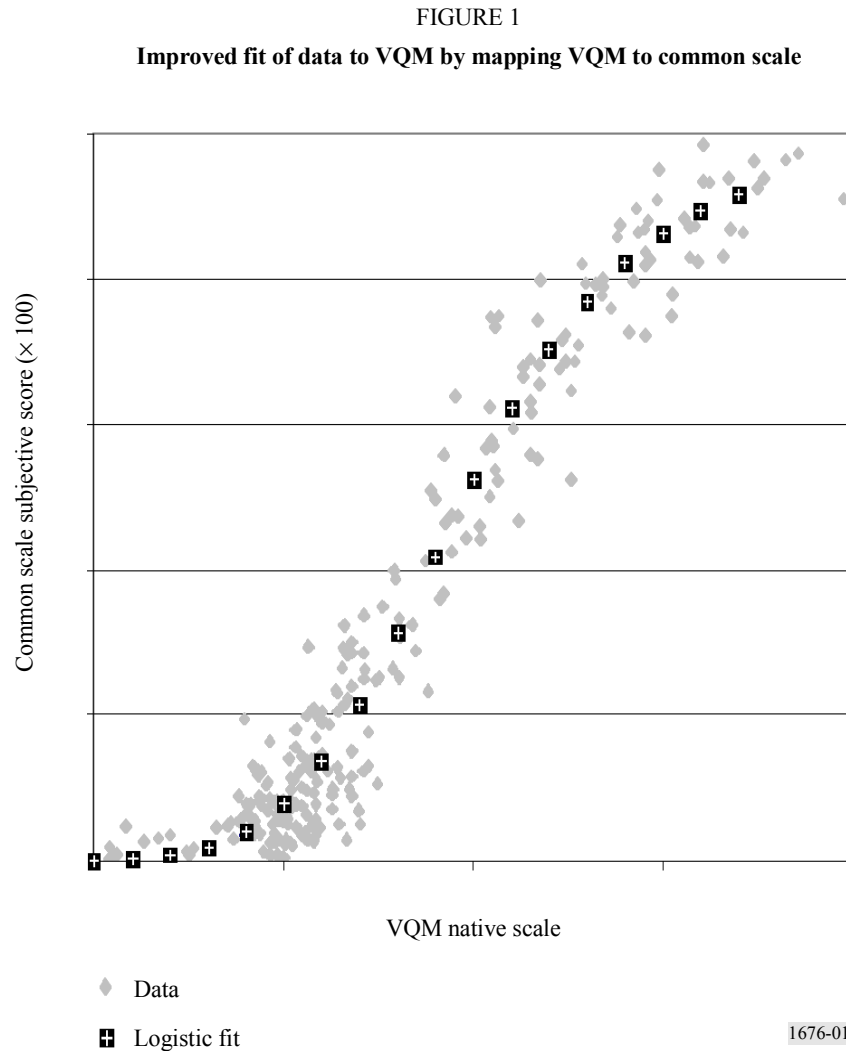
Figure 1 shows the improved fit of model to data incurred by transforming the objective scores using a fitting function. It can be seen that, besides improving the fit of data with VQM, the curve also offers an additional advantage over the straight-line fit implied by the native scale: the distribution of model-to-data errors around the fitted model curve is less dependent on the VQM score.

We denote the original (native scale) objective scores O_i , and the common scale objective scores as $\hat{O}_i$. A fitting function F (depending on some fitting parameters), connects the two. The function used to fit the objective VQM data (O_i) to the scaled subjective data ($\hat{S}_{i\bullet}$) must have the following three attributes:

- a specified domain of validity, which should include the range of VQM data for all the situations used to define the accuracy metric;
- a specified range of validity, defined as the range of common scale scores (a sub-range of $[0, 1]$) to which the function maps;
- monotonicity (the property of being either strictly increasing or strictly decreasing) over the specified domain of validity.

Of course, the fitting function would be most useful as a cross-calibration tool if it were monotonic over the entire theoretical domain of VQM scores, covered the entire subjective common scale from 0 to 1, and mapped to zero the VQM score that corresponds to a perfect video sequence.

(no degradations, hence a null distortion). However, this ideal may not be attainable for certain VQMs and function families used to perform the fit.



One possible family of fitting functions is the set of polynomials of order M . Another is a logistic function with the form:

$$\hat{O}_i = a + b / \{1 + c(O_i + d)^e\}$$

where a , b , c , d and e are fitting parameters. A third possibility is a logistic function with the form:

$$\hat{O}_i = a + (b - a) / \{1 + \exp[-c(O_i + d)]\}$$

where a , b , c , d are fitting parameters and $c > 0$. For convenience, we call these logistic forms Logistic I and Logistic II, respectively. The MATLAB code in Appendix 2 instantiates only a polynomial fit. Appendix 3 discusses possible methods of data fitting using the logistic functions.

The selection of a fitting-function family (including *a priori* setting of some of the parameters) depends on the asymptotic (best and worst) scores of the particular VQM.

The number of degrees of freedom used up by the fitting process is denoted by D . For example, if a linear fit is used, $D = 2$ since two free parameters are estimated in the fitting procedure. The fitting function that transforms objective VQM to the common scale is reported to facilitate industry comparison of two VQMs.

Once transformed to the common scale, any VQM can be cross-calibrated to any other VQM through the common scale. Representing the accuracy of a VQM in common scale facilitates comparisons between VQMs. Also, assuming the resolving power in the common scale does not vary much with the VQM score at which the resolving power is evaluated, the resolving power can be mapped through the inverse of the logistic function to the native scale. In the native scale, the ΔVQM from the common scale generates a VQM-score-dependent resolving power. A table or equation that provides such resolving powers (one at each VQM score in native scale) will have immediate meaning for users of the native scale.

2.3 METRIC 1: VQM accuracy based on statistical significance

We define a new quantitative measure of VQM accuracy, called resolving power, defined as the ΔVQM value above which the conditional subjective-score distributions have means that are statistically different from each other (typically at the 0.95 significance level). Such an “error bar” measure is needed in order for video service operators to judge the significance of VQM fluctuations.

Of several possible approaches to assessing a VQM’s resolving power, the Student’s t -test was chosen. This test was applied to the measurements in all pairs i and j of situations. Emerging from the test are the ΔVQM (i.e. the difference between the greater and lesser VQM score of i and j) and the significance from the t -test. This significance is the probability p that, given i and j , the greater VQM score is associated with the situation that has the greater true underlying mean subjective score. Thus, p is the probability that the observed difference in sample means of the subjective scores from i and j did not come from a single population mean, nor from population means that were ordered oppositely to the associated VQM scores. To capture this ordering requirement, the t -test must be one-tailed. For simplicity, the t -test was approximated by a z -test. This approximation is a close one when the number of viewers is large, as was the case for the Video Quality Experts Group (VQEG) data set (ITU-T Document COM 9-80-E).

An analysis of variance (ANOVA) test might seem better than the t -test method. However, although a single application of ANOVA will determine whether a statistical separation exists among a set of categories, further paired comparisons are needed to determine the magnitudes and conditions of the statistically significant differences. Also, ANOVA assumes equal category-data variances (which may not be true). Finally, although ANOVA resides in many software packages, finding the right software package may not be easy (e.g. not all ANOVA routines will accept different quantities of data in different categories).

The algorithm has the following steps:

Step 1: Start with an input data table with N rows, each row represents a different situation (i.e. a different source video and distortion). Each row i consists of the following: the source number, the distortion number, the VQM score O_i , the number of responses N_i , the mean subjective score $S_{i\bullet}$, and the sample variance of the subjective scores V_i .

Step 2: Transform the subjective scores $S_{i\bullet}$ to common scale $\hat{S}_{i\bullet}$ as described in § 2.1. The variance V_i of the subjective scores must also be scaled accordingly as:

$$\hat{V}_i = \frac{V_i}{(\text{worst} - \text{best})^2}$$

Note that transforming the subjective scores and their variances are optional. It will not change the z statistic defined below, but it may change the VQM fitting process. Next, transform the VQM scores O_i to the common scale using a fitting function as discussed in § 2.2, and amplified in Appendix 3. The result of the fitting process is a set of common scale VQM scores $\hat{O}_i$. Display the coefficient values used in the fit, and also the VQM domain over which the fit was done (domain of validity).

Step 3: For each pair of distinct situations i and j ($i \neq j$), use a one-tailed z -test to assign a probability of significance to the difference between the greater and the lesser VQM ($\hat{O}_i$ and $\hat{O}_j$, respectively). The significance is the probability that the greater VQM score comes from the situation with the greater true underlying mean subjective score. The z score is:

$$z = (\hat{S}_{i\bullet} - \hat{S}_{j\bullet}) / \sqrt{(\hat{V}_i / N_i + \hat{V}_j / N_j)}$$

and the probability of significance of the z score $p(z)$ is just the cumulative distribution function of z :

$$p(z) = \text{cdf}(z) = (2\pi)^{-0.5} \int_{-\infty}^z \exp(-z^2/2) dz$$

Step 4: Create a scatter plot of $p(z)$ (ordinate) versus ΔVQM score (abscissa). Given N situations, record each pair (i, j) with $i > j$, record the VQM difference $\hat{O}_i - \hat{O}_j$ in a vector of length $N(N-1)/2$ called ΔVQM (with index k), and record the corresponding z score in a vector called Z with length $N(N-1)/2$ (with the same index k). It is desired to ensure that $\Delta\text{VQM}(k)$ is always non-negative, which can be ensured by definition of the otherwise arbitrary ordering of the endpoints i and j . To ensure that this is so, if $\Delta\text{VQM}(k)$ is negative, then replace $Z(k)$ by $-Z(k)$ and $\Delta\text{VQM}(k)$ by $-\Delta\text{VQM}(k)$.

Step 5: Consider 19 bins (indexed by m) of ΔVQM , each one of which spans 1/10 the total range of ΔVQM . The bins overlap by 50%. Associate ΔVQM_m with the midpoint of each bin and associate p_m with the mean of $p(z)$ for all z in bin m .

Step 6: Draw a curve through the points $(\Delta\text{VQM}_m, p_m)$, to produce a graph of p versus ΔVQM . Note that p can be interpreted as the average probability of significance.