RECOMMENDATION ITU-R BT.1129-2

SUBJECTIVE ASSESSMENT OF STANDARD DEFINITION DIGITAL TELEVISION (SDTV) SYSTEMS

(Question ITU-R 211/11)

(1994-1995-1998)

The ITU Radiocommunication Assembly,

considering

a) that a number of administrations and organizations throughout the world are currently evaluating digital systems and that, in many parts of the world, digital broadcasting is likely to become the primary medium of the next century;

b) that subjective assessments are a vital element in the design and comparison of digital systems;

c) that Recommendation ITU-R BT.500 outlines a number of preferred methods for subjective assessments, many details of which are appropriate in the context of digital television;

d) that implicitly, and by virtue of the data transport techniques used, digital systems offer the opportunity of introducing multiple programming or scalable or hierarchical coding schemes in the broadcast channel;

e) that in digital television systems, program contents may have a significant influence on the picture quality,

recommends

1 that the general methods for the subjective assessment of digital standard definition systems, including those offering multiple programme streams or scalable or hierarchical coding schemes, be as described in Recommendation ITU-R BT.500;

2 that specific procedures for the subjective assessment of digital systems, including those offering multiple programme streams or scalable or hierarchical coding schemes, be as given in Annex 1.

ANNEX 1

1 Introduction

This Annex, which is intended to be used in conjunction with Recommendation ITU-R BT.500, provides details concerning the application of the general methods given in the Recommendation to subjective assessments of digital systems offering levels of quality at, or near, those of conventional television systems. The procedural details given here, together with relevant background information, pertain to tests of codecs (or systems) used to convey material originated according to Recommendation ITU-R BT.601 in contribution and distribution applications as well as to those used in emission applications.

For distribution applications, quality specifications can be expressed in terms of the subjective judgement of observers. Such codecs can in theory therefore be assessed subjectively against these specifications. The quality of a codec designed for contribution applications however, could not in theory be specified in terms of subjective performance parameters because its output is destined not for immediate viewing, but for studio post-processing, storing and/or coding for further transmission. Because of the difficulty of defining this performance for a variety of post-processing operations, the approach preferred has been to specify the performance of a chain of equipment, including a

post-processing function, which is thought to be representative of a practical contribution application. This chain might typically consist of a codec, followed by a studio post-processing function (or another codec in the case of basic contribution quality assessment), followed by yet another codec before the signal reaches the observer. Adoption of this strategy for the specification of codecs for contribution applications means that the measurement procedures given in this Recommendation can also be used to assess them.

Although progress is being made, there is currently insufficient experience to give details of objective picture quality assessment methods for codecs. In the area of subjective assessment, where much experience exists, test conditions and methodologies can be recommended. It must be remembered, however, when specifying quality or impairment targets, that existing methods cannot give absolute subjective ratings but rather results which are influenced to some extent by the choice of the reference and/or anchor conditions. The same methodologies may be adopted for both fixed and variable word-length codecs, and for intrafield and interframe codecs although the choice of test images sequences may be influenced.

At the present time, the most completely reliable method of evaluating the ranking order of high-quality codecs is to assess all the candidate systems at the same time under identical conditions. Tests made independently, where fine differences of quality are involved, should be used for guidance rather than as indisputable evidence of superiority.

A useful subjective measure may be impairment determined as a function of the bit error ratio which occurs in the transmission link between coder and decoder. At present there is insufficient experimental knowledge of true transmission error statistics to recommend parameters for a model which accounts for error clustering or bursts. Until this information becomes available Poisson-distributed errors may be used.

2 Viewing conditions

The general viewing conditions for subjective assessments are those given in Recommendation ITU-R BT.500, § 2.1. Specific viewing conditions for subjective assessments of digital systems are given in the following paragraphs.

2.1 Laboratory environment

The laboratory environment is intended to provide critical conditions to check systems. Specific viewing conditions for subjective assessments in the laboratory environment are given in the following Table 1.

TABLE 1

Specific viewing conditions for subjective assessments of digital systems in laboratory environment

Condition	Item	Values
a	Ratio of viewing distance to picture height	4 *H* and 6 *H* (1)
b	Peak luminance	70 cd/m^2
c	Viewing angle subtended by that portion of the background that meets specifications	≥43° H × 57° W
d	Display	High quality screen. Size ≥ 20" (50 cm) (2)

(1) 6 *H* is the design viewing distance (DVD) for the assessment of digital standard definition systems, but using assessors at 4 *H* is also acceptable, provided that the results are given separately.

(2) Because there is some evidence that display size may influence the results of subjective assessments experimenters are requested to explicitly report the screen size and make and model of displays used in any experiments.

2.2 Home environment

This environment is intended to provide a mean to evaluate quality at the consumer side of the digital TV chain. Specific viewing conditions for digital standard definition digital television (SDTV) subjective assessments in the home environment are given in Table 2.

TABLE 2

Specific viewing conditions for subjective assessments of digital systems in home environment

Condition	Item	Values
a	Ratio of viewing distance to picture height	6 *H*
b	Screen size for a 4/3 format ratio	From 25" to 29"(1)
c	Screen size for a 16/9 format ratio	From 32" to 36"(1)
d	Monitor standard	SDTV
e	Peak luminance	200 cd/m^2
f	Environmental Illuminance on the screen (Incident light from the environment falling on the screen should be measured perpendicularly on the screen)	200 Lux

(1) This screen size satisfies rules of the preferred viewing distance (PVD) for a PVD = 6 *H*.

3 Assessment methods

3.1 Evaluations of basic picture quality

Where a codec is being assessed for distribution applications, this quality refers to pictures decoded after a single pass through a codec pair. For contribution codecs, basic quality may be assessed after several codecs in series, in order to simulate a typical contribution application.

Where the range of quality to be assessed is small, as will normally be the case for television codecs, the testing methodology to be used is variant II of the double-stimulus continuous quality-scale described in Recommendation ITU-R BT.500. The original source sequence will be used as the reference condition. Further consideration is being given to the duration of presentation sequences. In the recent tests on codecs for 4:2:2 component video, it was considered advantageous to modify the presentation from that given in Recommendation ITU-R BT.500. Composite pictures were used as an additional reference to provide a lower quality level against which to judge the codec performance.

It is recommended that at least six picture sequences be used in the assessment, plus an additional one to be used for training purposes prior to the start of the trial. The sequences should range between moderately critical and critical in the context of the bit-rate reduction application being considered.

In subjective assessments, still pictures and moving sequences may be selected from those listed in Recommendation ITU-R BT.1210, Annex 1. In this respect, it should be noted that digitally stored pictures and sequences, being the most reproducible source signals, are the preferred sources for assessments. Material such as that described in Recommendation ITU-R BT.1210 can be exchanged among laboratories in order to make system comparisons more meaningful. The D-1 4:2:2 tape format described in Recommendation ITU-R BT.657 should provide a basis for such exchange when such machines are widely and economically available. Exchange via computer tape formats also is possible.

Throughout this Annex, the importance is stressed of testing digital codecs with picture sequences which are critical in the context of television bit-rate reduction. It is therefore reasonable to ask how critical a particular image sequence is for a particular bit-rate reduction task, or whether one sequence is more critical than another. A simple but not especially

helpful answer is that “criticality” means very different things to different codecs. For example, to an intrafield codec a still picture containing much detail could well be critical, while to an interframe codec which is capable of exploiting frame-to-frame similarities, this same scene would present no difficulty at all. Some sequences employing moving texture and complex motion will be critical to all classes of codec so these types of sequences are most useful to generate or identify. Complex motion may take the form of movements which are predictable to an observer but not to coding algorithms, such as tortuous periodic motion.

One examination of possible statistical measures of image criticality, such as by correlative methods, spectral methods, conditional entropy methods etc. has revealed a simple but useful measure based on an intrafield/interframe adaptive entropy measurement. This method was used to “calibrate” picture sequences proposed for use in the ITU-R trials of codecs for 34, 45 and 140 Mbit/s and proved useful for the selection of the sequences used. The making of such measurements on picture sequences is most easily accomplished by transferring them to image processing computers and subjecting them to analysis by software.

Where access to these techniques is not available, the following presents some general guidelines on how to choose critical material.

a) *Fixed word-length intrafield codecs*

While it is possible and valid to assess these codecs on still images, the use of moving sequences is recommended since coding noise processes are easier to observe and this is more realistic of television applications. If still images are used in computer simulations of codecs, processing should be performed over the entire assessment sequence in order to preserve temporal aspects of any source noise, for example. The scenes chosen should contain as many as possible of the following details: static and moving textured areas (some with coloured texture); static and moving objects with sharp high contrast edges at various orientation (some with colour); static plain mid-grey areas. At least one sequence in the ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic (i.e. computer generated) so that it is free from camera imperfections such as scanning aperture and lag.

b) *Fixed word-length interframe codecs*

The test scenes chosen should all contain movement and as many as possible of the following details: moving textured areas (some coloured); objects with sharp, high contrast edges moving in a direction perpendicular to these edges and at various orientations (some coloured). At least one sequence in the ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic.

c) *Variable word-length intrafield codecs*

It is recommended that these codecs be tested with moving image sequence material for the same reasons as the fixed word-length codecs. It should be noted that by virtue of its variable word-length coding and associated buffer store, these codecs can dynamically distribute coding bit-capacity throughout the image. Thus, for example, if half of a picture consists of a featureless sky which does not require many bits to code, capacity is saved for the other parts of the picture which can therefore be reproduced with high quality even if they are critical. The important conclusion from this is that if a picture sequence is to be critical for such a codec, the content of every part of the screen should be detailed. It should be filled with moving and static texture, as much colour variation as possible and objects with sharp, high contrast edges. At least one sequence in the text ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic.

d) *Variable word-length interframe codecs*

This is the most sophisticated class of codec and the kind which requires the most demanding material to stress it. Not only should every part of the scene be filled with detail as in the intrafield variable word-length case, but this detail should also exhibit motion. Furthermore, since many codecs employ motion compensation methods, the motion throughout the sequence should be complex. Examples of complex motion are: scenes employing simultaneous zooming and panning of a camera; a scene which has as a background a textured or detailed curtain blowing in the wind; a scene containing objects which are rotating in the three dimensional world; scenes containing detailed objects which accelerate across the screen. All scenes should contain substantial motion of objects with different velocities, textures and high contrast edges as well as a varied colour content. At least one sequence in the test ensemble should exhibit just perceptible source noise, at least one sequence should have complex computer generated camera motion from a natural still picture (so that it is free from noise and camera lag), and at least one sequence should be entirely computer generated.

3.2 Evaluations of picture quality after downstream processing

This assessment is intended to permit judgement to be made on the suitability of a codec for contribution applications with respect to a particular post-process e.g. colour matte, slow motion, electronic zoom. The minimum arrangement of equipment for such an assessment is a single pass through the codec under test, followed by the post-process of interest, followed by the viewer. It may, however, be more representative of a contribution application to employ further codecs after the post-process.

The test methodology to be used is variant II of the double-stimulus continuous quality-scale method. Here however the reference condition will be the source subjected to the same post-processing as the decoded pictures. If inclusion of a lower quality reference is considered to be advantageous then it too should be subjected to the same post-process.

Test sequences required for post-processing assessments are subject to exactly the same criticality criteria as sequences for other digital applications. This may be difficult to achieve however in chroma key foreground sequences because they usually have a significant proportion of featureless blue background.

Because of the practical constraints of possibly having to assess a codec with several post-processes, the number of test picture sequences used may be a minimum of three with an additional one available for demonstration purposes. The nature of the sequences will be dependent upon the post-processing task being studied but should range between moderately critical and critical in the context of television bit-rate reduction and for the process under consideration. For slow motion assessment a display rate of $1/10^{th}$ of the source rate may be suitable.

3.3 Evaluations of failure characteristics

In subjective assessments of impairments in codec pictures due to imperfections in the transmission or emission channel, a minimum of five, but preferably more, bit-error ratios or selected transmission/emission conditions should be chosen, approximately logarithmically spaced and adequately sampling the range which gives rise to codec impairments from “imperceptible” to “very annoying”.

It is possible that codec assessments could be required at transmission bit error ratios which result in visible transients so infrequent that they may not be expected to occur during a 10 s test sequence period. The presentation timing suggested here is clearly not suitable for such tests.

If recordings of a codec output under fairly low bit error ratio conditions (resulting in a small number of visible transients within a 10 s period) are to be made for later editing into subjective assessment presentations, care should be taken to ensure that the recording used is typical of the codec output viewed over a longer time-span.

Because of the need to explore codec performance over a range of transmission bit error ratios, practical constraints suggest that three test picture sequences with an additional demonstration sequence will probably be adequate. Sequences should be of the order of 10 s in duration but it should be noted that test viewers may prefer a duration of 15-30 s. It should range between moderately critical and critical in the context of television bit-rate reduction.

As the tests will span the full range of impairment, the double-stimulus impairment scale method is appropriate and should be used.

3.4 Picture-content failure characteristics

The general concept of picture-failure characteristics is given in Appendix 1 to Annex 1 to Recommendation ITU-R BT.500. To apply this concept to digital television systems, the following procedure should be used.

3.4.1 Definition of criticality

A certain measure called “criticality” which represents the characteristics of the digital television system under test and is measured by objective measurement should be defined. As an example of digital television system, MPEG-2 MP@ML is used and the fixed quantizer method of entropy based criticality, which is described in Recommendation ITU-R BT.1210, is applied.

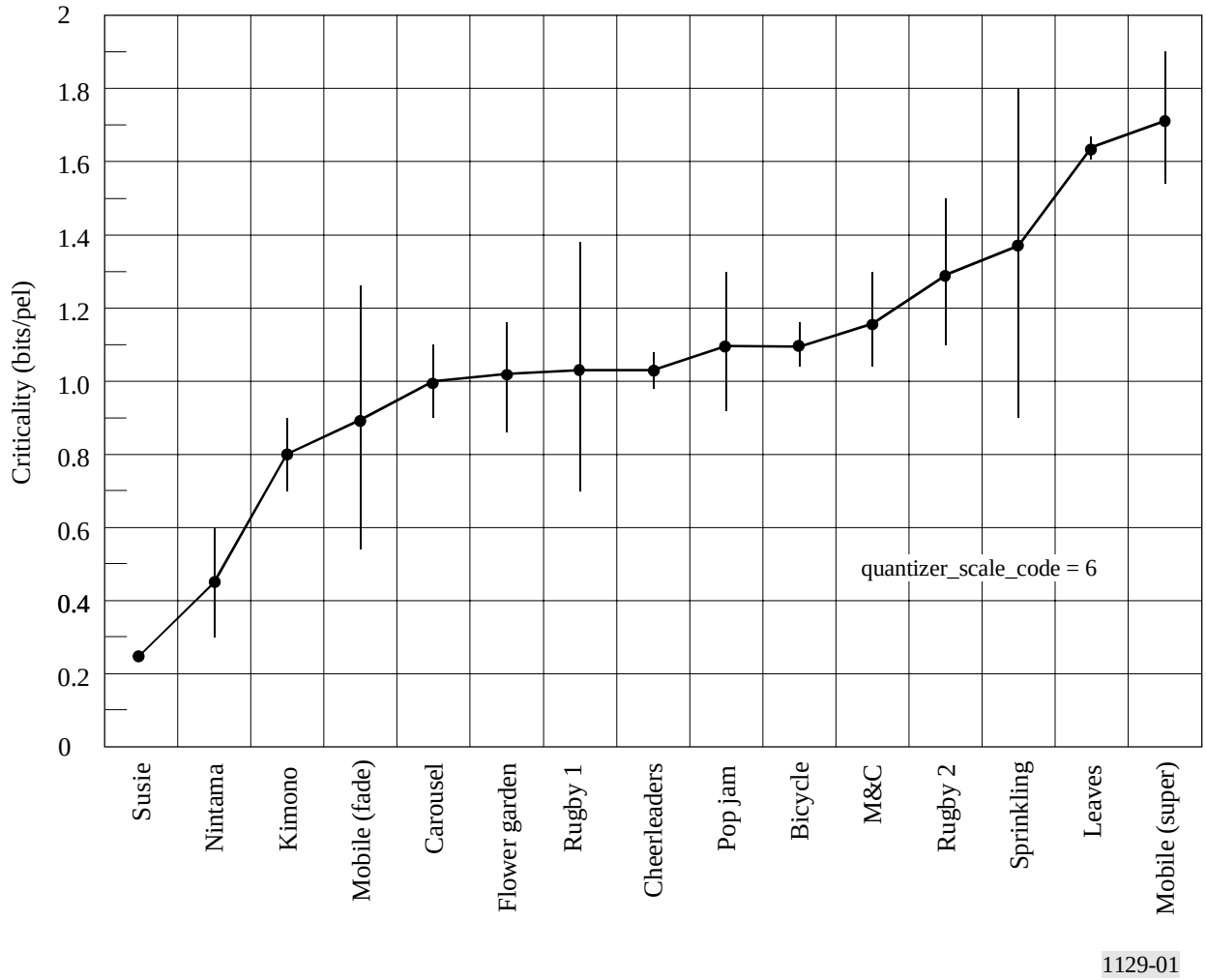
3.4.2 Procedure of derivation of picture-content failure characteristics

– *Step 1*: Measure criticality of the test sequences used in the subjective assessment

Criticality of test sequences used for the subjective assessment described in Step 3 below is measured. Figure 1 shows the mean and standard deviation of each sequence for the example system. Most sequences have criticality measures from 0.8 to 1.4 bits/pixel. Some sequences have a large standard deviation because the picture content varies significantly during the sequence.

FIGURE 1

Means and standard deviations of criticality of test sequences

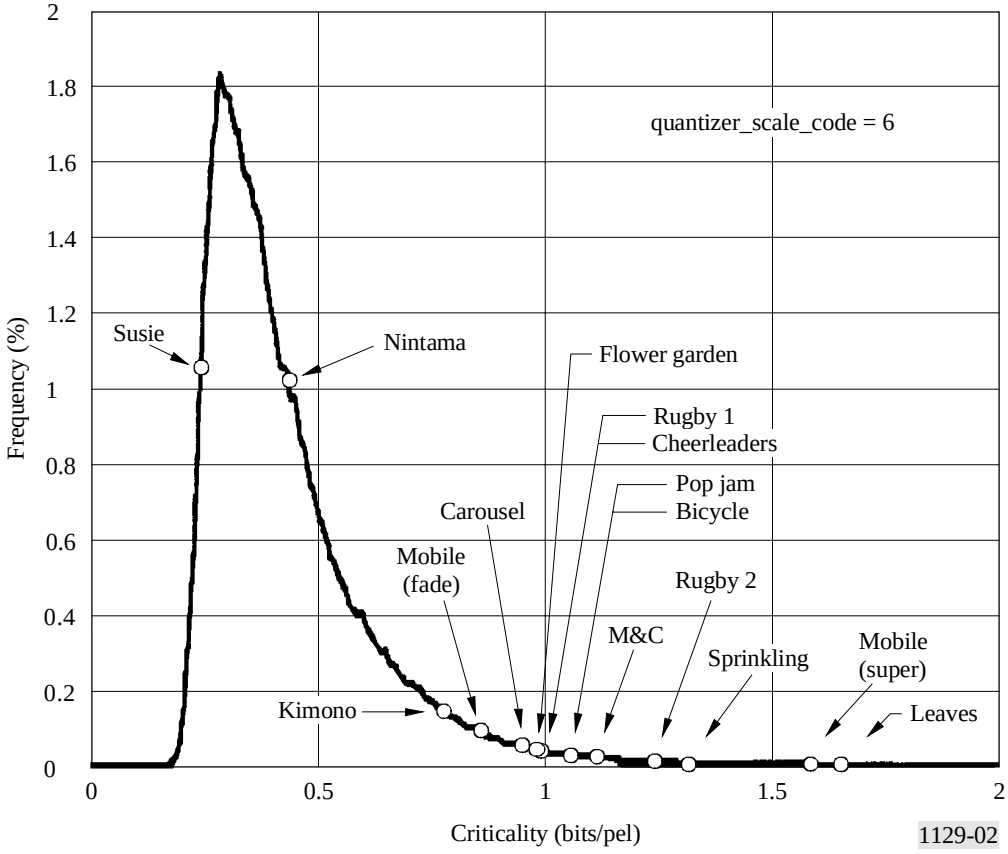


– *Step 2*: Measure criticality distribution of broadcast programs for a long time period

Criticality distribution of broadcast television programs is measured for a sufficiently long time period, e.g. one week. Figure 2 shows an example of the distribution measured for one week, a total of 130 h for NTSC broadcast signals, which were converted into component *Y/C* signals for measurement. The frequency of occurrence of criticality for television programs was calculated every 5×10^{-3} bits/pixel. This figure also shows criticality for the test sequences used for the subjective assessment.

FIGURE 2

Distribution of criticality for broadcast programs and criticality of test sequences



– *Step 3*: Conduct a subjective assessment of picture quality of the system under test, and derive a relationship between criticality and subjective picture quality

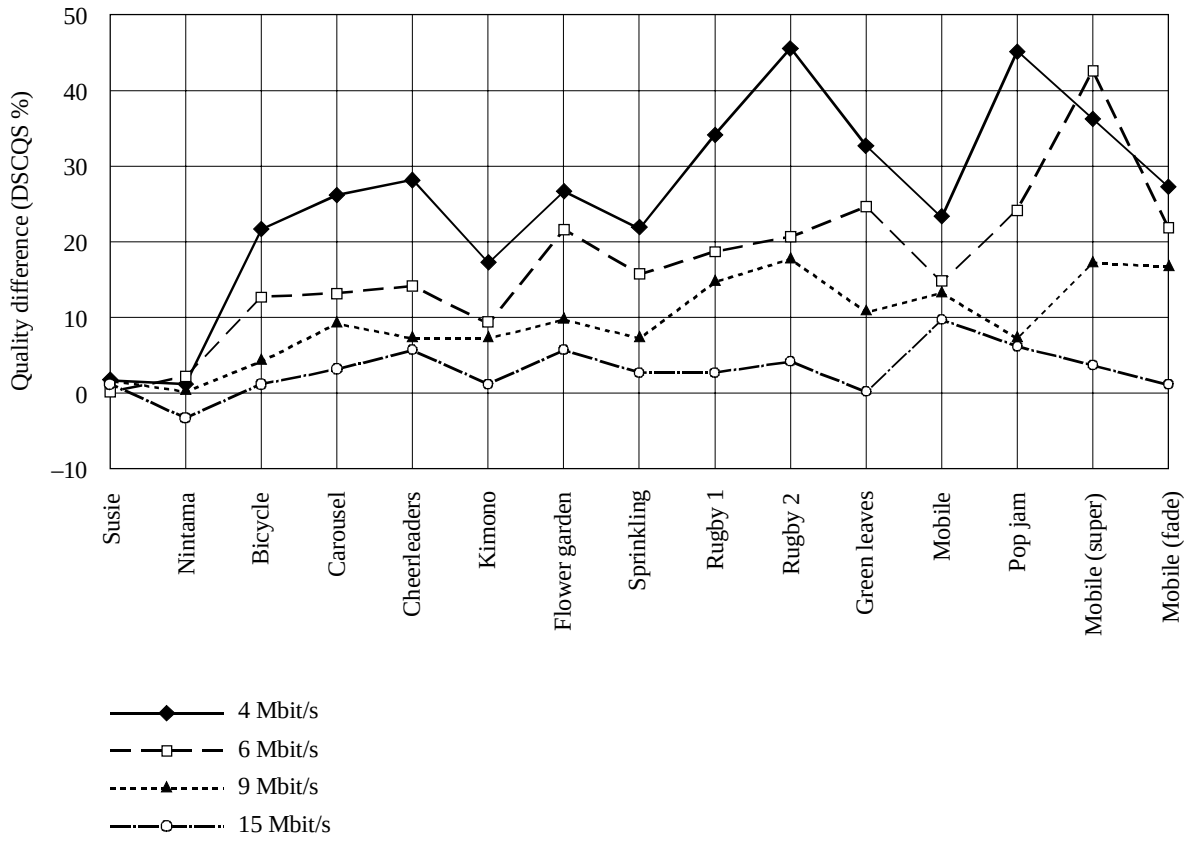
Picture quality of the digital television system is assessed by DSCQS method. Combining the subjective assessment result and the criticality obtained in Step 1, relationship between criticality and the scores of the assessment test is derived. Figure 3 shows the picture quality of the example system at the bit-rates of 4, 6, 9, and 15 Mbit/s. Quality difference (DSCQS %) in the figure represents the degradation from the reference, original 4:2:2 component sequence. Figure 4 shows the relationship between criticality and quality difference. In this example, linear relationship between criticality and picture quality was assumed, and regression lines were derived using the least squares method. The regression line at each bit-rate is illustrated in the figure. In general, nonlinear relationship can be applied depending on the assessment results.

– *Step 4*: Derive picture-content failure characteristics (quality vs. frequency of occurrence) by combining the results of Step 3 (criticality vs. quality) and Step 2 (criticality vs. frequency of occurrence).

By combining the results obtained in Steps 2 and 3, picture-content failure characteristics, i.e. distribution of picture quality of digitally coded television programs, is derived. The picture degradation in broadcast television programs is converted into cumulative frequency of occurrence. Figure 5 shows the picture-content failure characteristics of the example system.

FIGURE 3

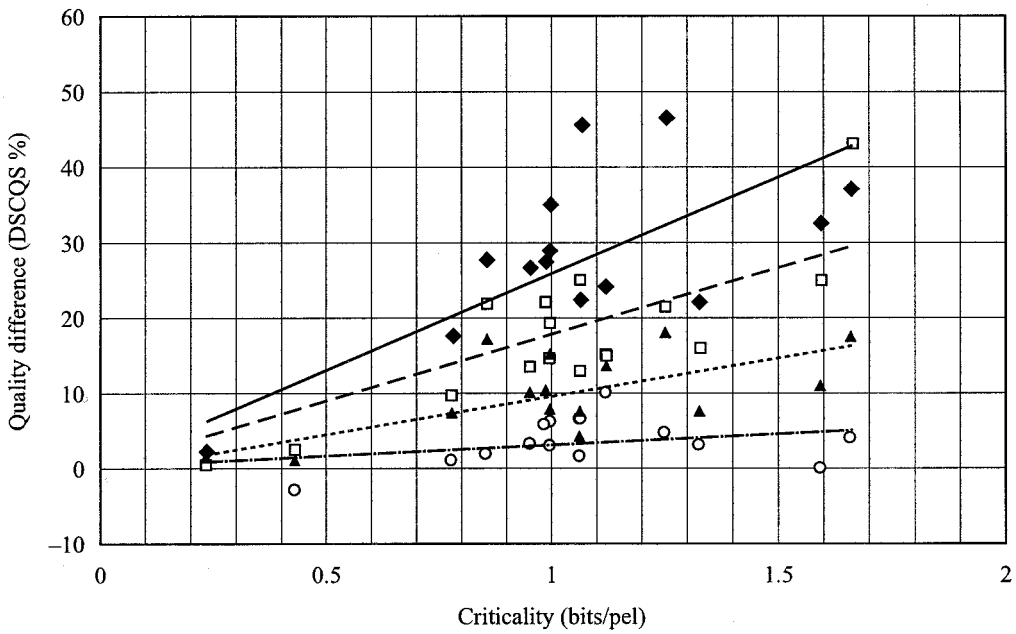
Result of subjective assessment (MP@ML at 6H)



1129-03

FIGURE 4

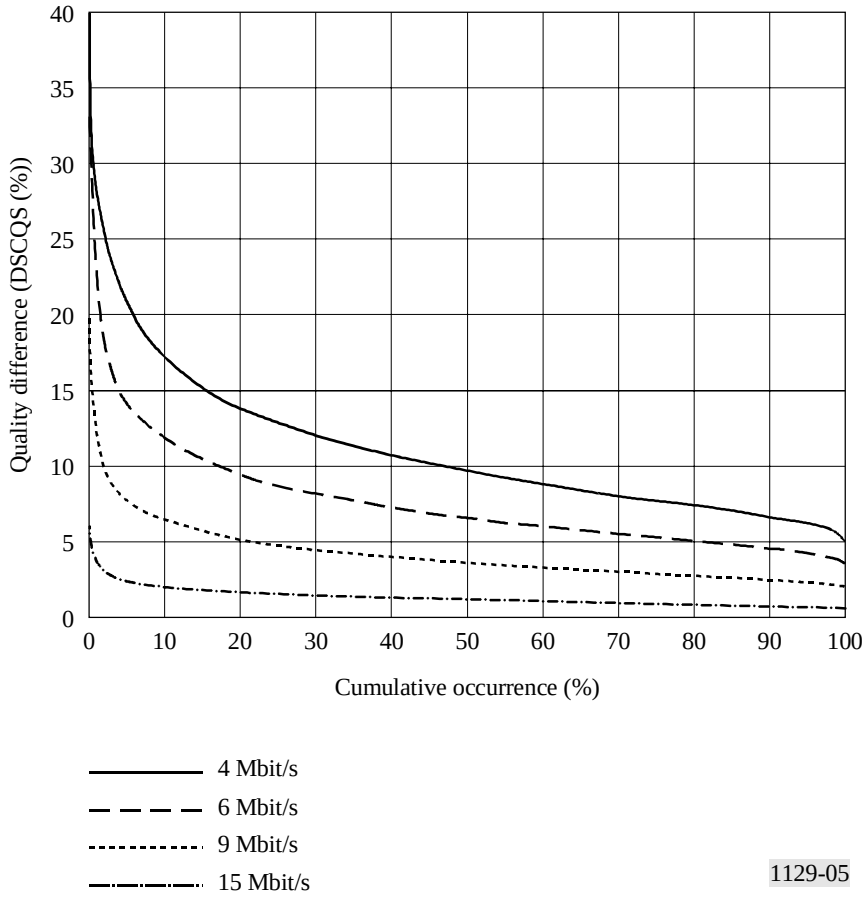
Relationship between criticality and assessment score (MP@ML at 6H)



1129-04sc

FIGURE 5

Cumulative frequency of occurrence of picture degradation (MP@ML at 6 *H*)



4 Application notes

Where a judgement of absolute codec quality or impairment is not required, but only the ranking order, or where confirmation of the ranking order found from double-stimulus results is desired, the method of paired-stimulus comparisons should be used.

As it is described in Recommendation ITU-R BT.500, the method provides a sensitive comparison and a means of determining a measure of the relation between pairs of systems. An extension of this method, to ranking the quality or impairment of more than two systems, is possible. In this approach overall ranking order is derived from the ranking of all possible pairs of picture sequences by the observers.

The analysis is complicated by the fact that an observer can rank, for example, picture A better than picture B, and picture B better than picture C, but also picture C better than picture A. This is termed as “intransitive triad”.

A problem with the method is that the number of presentations required increases as the square of the number of test picture sequences and codecs, and can become impractical.

If the broadcast channel is used to deliver multiple programme streams or scalable or hierarchical coding schemes, it may be necessary to adapt the assessment methodology to take account of the following:

– The criterion for acceptable service may not be transparency in source coding; instead, it may be the ability of the system, at a given bit-rate allocation, to provide a viable alternative to conventional service. Accordingly, as the reference in quality tests, it may be appropriate to use material as delivered by a conventional system under typical reception conditions, rather than material in uncompressed digital form. Further, it may be appropriate to use test

material selected to represent the range of current and future programme content (see Recommendation ITU-R BT.500, Annex 1, Appendix 1). In tests, viewing conditions should be as given in Recommendation ITU-R BT.500 and in § 1 of this Annex, while the general test method should be the double-stimulus continuous quality-scale method (Recommendation ITU-R BT.500, § 5); and

– the ability of the system to maintain the integrity of individual programme streams in conditions of full channel loading and transmission impairment is of issue. Accordingly, in impairment tests, it may be appropriate to ensure full channel loading and to use a range of impairment levels selected to represent the range of likely reception conditions (see Recommendation ITU-R BT.500, Annex 1, Appendix 2). In tests, viewing conditions should be as given in Recommendation ITU-R BT.500 and in § 1 of this Annex, while the general test method should be the double-stimulus impairment-scale method (Recommendation ITU-R BT.500, § 4).

NOTE 1 – When analogue and digital systems are assessed in the same context, it is important to choose a set of test materials that reflects a balanced difficulty for the analogue and digital systems. It may be useful in this case to apply, for supplementary analysis, the multidimensional scaling procedure referred to in Recommendation ITU-R BT.500, Annex 1, Table 2.