INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.911

(12/98)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Audiovisual quality in multimedia services

Subjective audiovisual quality assessment methods for multimedia applications

ITU-T Recommendation P.911

(Previously CCITT Recommendations)

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series P.10
Subscribers' lines and sets	Series P.30
	P.300
Transmission standards	Series P.40
Objective measuring apparatus	Series P.50
	P.500
Objective electro-acoustical measurements	Series P.60
Measurements related to speech loudness	Series P.70
Methods for objective and subjective assessment of quality	Series P.80
	P.800
Audiovisual quality in multimedia services	**Series P.900**

For further details, please refer to ITU-T List of Recommendations.

ITU-T RECOMMENDATION P.911

SUBJECTIVE AUDIOVISUAL QUALITY ASSESSMENT METHODS FOR MULTIMEDIA APPLICATIONS

Summary

This Recommendation describes non-interactive subjective assessment methods for evaluating the one-way overall audiovisual quality for multimedia applications such as videoconferencing, storage and retrieval applications, telemedical applications, etc. These methods can be used for several different purposes including, but not limited to, selection of algorithms, ranking of audiovisual system performance and evaluation of the quality level during an audiovisual connection. When interactive aspects are to be assessed, conversation test methods described in Recommendation P.920 should be used. This Recommendation also outlines the characteristics of the source sequences to be used, like duration, kind of content, number of sequences, etc. Finally it provides indications about relation between audio, video and audiovisual quality, as they were derived from results of tests carried out independently in different laboratories.

Source

ITU-T Recommendation P.911 was prepared by ITU-T Study Group 12 (1997-2000) and was approved under the WTSC Resolution No. 1 procedure on the 3rd of December 1998.

FOREWORD

ITU (International Telecommunication Union) is the United Nations Specialized Agency in the field of telecommunications. The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of the ITU. The ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Conference (WTSC), which meets every four years, establishes the topics for study by the ITU-T Study Groups which, in their turn, produce Recommendations on these topics.

The approval of Recommendations by the Members of the ITU-T is covered by the procedure laid down in WTSC Resolution No. 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation the term *recognized operating agency (ROA)* includes any individual, company, corporation or governmental organization that operates a public correspondence service. The terms *Administration, ROA* and *public correspondence* are defined in the *Constitution of the ITU (Geneva, 1992)*.

INTELLECTUAL PROPERTY RIGHTS

The ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. The ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, the ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

CONTENTS

Recommendation P.911

SUBJECTIVE AUDIOVISUAL QUALITY ASSESSMENT METHODS FOR MULTIMEDIA APPLICATIONS

(Geneva, 1998)

1 Scope

This Recommendation is intended to define non-interactive subjective assessment methods for evaluating the one-way overall audiovisual quality for bit rates specified in classes TV 3, MM 4, MM 5 and MM 6, as specified in Tables B.2 and B.4, for applications such as videoconferencing, storage and retrieval applications, telemedical applications, etc. The methods can be used for several different purposes including, but not limited to, selection of algorithms, ranking of audiovisual system performance and evaluation of the quality level during an audiovisual connection. When interactive aspects are to be assessed, Recommendation P.920 should be used.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; all users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

[1] ITU-T Recommendation P.930 (1996), *Principles of a reference impairment system for video.*

[2] ITU-T Recommendation P.920 (1996), *Interactive test methods for audiovisual communications.*

[3] ITU-R Recommendation BT.601-5 (1995), *Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios.*

[4] ITU-R Recommendation BT.500-8 (1998), *Methodology for the subjective assessment of the Quality of Television Pictures.*

[5] IEC/TR3 Publication 60268-13 (1998), *Sound system equipment – Part 13: Listening tests on loudspeakers.*

[6] ITU-T Recommendation P.800 (1996), *Methods for subjective determination of transmission quality*.

[7] ITU-R Recommendation BT.814-1 (1993), *Specifications and alignment procedures for setting of brightness and contrast of displays*.

[8] ITU-R Recommendation BT.1128-2 (1997), *Subjective assessment of conventional television systems.*

[9] CCITT Recommendation J.61 (1990), *Transmission performance of television circuits designed for use in international connections.*

[10] ITU-T Recommendation P.810 (1996), *Modulated Noise Reference Unit (MNRU).*

[11] ITU-T Recommendation P.910 (1996), *Subjective video quality assessment methods for multimedia applications*.

[12] CCITT Recommendation G.722 (1988), *7 kHz audio-coding within 64 kbit/s*.

[13] CCITT Recommendation G.711 (1988), *Pulse Code Modulation (PCM) of voice frequencies*.

[14] CCITT Recommendation G.728 (1992), *Coding of speech at 16 kbit/s using low-delay code excited linear prediction.*

[15] ITU-T Recommendation G.114 (1996), *One-way transmission time*.

[16] ITU-R Recommendation BS.775-1 (1993), *Multi-channel stereophonic sound system with and without accompanying picture*.

[17] IEC Publication 60651 (1979), *Sound level meters.*

[18] ITU-T Recommendation P.931 (1998), *Subjective audiovisual quality assessment methods for multimedia applications*.

3 Terms and definitions

This Recommendation defines the following terms:

3.1 explicit reference (source reference): The condition used by the assessors as reference to express their opinion, when the DCR method is used. This reference is displayed first within each pair of sequences. Usually the format of the explicit reference is the format used at the input of the codecs under test (e.g.: ITU-R BT.601, CIF, QCIF, SIF, etc.). In the body of this Recommendation, the words "explicit" and "source" will be omitted whenever the context will make clear the meaning of "reference".

3.2 gamma: The relationship between the screen luminance and the input signal voltage is non-linear, with the voltage raised to an exponent gamma. To compensate for this non-linearity, a correction factor that is an inverse function of gamma is generally applied in the camera. Gamma also has an impact on colour rendition.

3.3 optimization tests: Subjective tests that are typically carried out during either the development or the standardization of a new algorithm or system. The goal of these tests is to evaluate the performance of new tools in order to optimize the algorithms or the systems that are under study.

3.4 qualification tests: Subjective tests that are typically carried out in order to compare the performance of commercial systems or equipment. These tests must be carried out under test conditions that are as representative as possible of the real conditions of use.

3.5 reference conditions: Dummy conditions added to the test conditions in order to better compare the evaluations coming from different experiments.

3.6 reliability of a subjective test:

a) intra-individual ("within subject") reliability refers to the agreement between a certain subject’s repeated ratings of the same test condition;

b) inter-individual ("between subjects") reliability refers to the agreement between different subjects’ ratings of the same test condition.

3.7 replication: Repetition of the same circuit condition (with the same source material) for the same subject.

3.8 Spatial perceptual Information (SI): A measure that generally indicates the amount of spatial detail of a picture. It is usually higher for more spatially complex scenes. It is not meant to be a measure of entropy nor associated with information as defined in communication theory. The Spatial perceptual Information, SI, is based on the Sobel filter. Each video frame (luminance plane) at time n (F_n) is first filtered with the Sobel filter [Sobel(F_n)]. The standard deviation over the pixels (std_{space}) in each Sobel-filtered frame is then computed. This operation is repeated for each frame in the video sequence and results in a time series of spatial information of the scene. The maximum value in the time series (max_{time}) is chosen to represent the spatial information content of the scene. This process can be represented in equation form as:

$$SI = max_{time} \{std_{space} [Sobel(F_n)]\}$$

3.9 Temporal perceptual Information (TI): A measure that generally indicates the amount of temporal changes of a video sequence. It is usually higher for high motion sequences. It is not meant to be a measure of entropy nor associated with information as defined in communication theory.

The measure of Temporal Information, TI, is computed as the maximum over time (max_{time}) of the standard deviation over space (std_{space}) of $M_n(i,j)$ over all i and j.

$$TI = max_{time} \{std_{space} [M_n(i,j)]\}$$

where $M_n(i,j)$ is the difference between pixels at the same position in the frame, but belonging to two subsequent frames, that is:

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j)$$

where $F_n(i,j)$ is the pixel at the i^{th} row and j^{th} column of n^{th} frame in time.

3.10 transparency (fidelity): A concept describing the performance of a codec or a system in relation to an ideal transmission system without any degradation.

Two types of transparency can be defined:

The first type describes how well the processed signal conforms to the input signal, or ideal signal, using a mathematical distance criterion. If there is no difference (i.e. distance = 0), the system is fully transparent. The second type describes how well the processed signal conforms to the input signal, or ideal signal, for a human observer. If no difference can be perceived under any experimental condition, the system is perceptually transparent. The term transparent without explicit reference to a criterion will be used for systems that are perceptually transparent.

3.11 validity of a subjective test: Agreement between the mean value of ratings obtained in a test and the true value which the test purports to measure.

4 Abbreviations

This Recommendation uses the following abbreviations:

ACR Absolute Category Rating

CCD Charge Coupled Device

CI Confidence Interval

CIF Common Intermediate Format (picture format defined in Recommendation H.261 for videophone: 352 lines × 288 pixels)

CRT Cathode Ray Tube

DCR Degradation Category Rating

%GOB Percent of Good or Better (proportion of Good and Excellent)

LCD	Liquid Crystal Display
MOS	Mean Opinion Score
PC	Pair Comparison
%POW	Percent of Poor or Worse (proportion of Poor and Bad votes)
QCIF	Quart CIF (picture format defined in Recommendation H.261 for videophone: 176 lines × 144 pixels)
SI	Spatial Information
SIF	Standard Intermediate Format [picture formats defined in ISO/IEC-11172 (MPEG-1): 352 lines × 288 pixels × 25 frames/s and 352 lines × 240 pixels × 30 frames/s)
S/N	Signal-to-Noise ratio
SP	Simultaneous Presentation
std	Standard Deviation
TI	Temporal Information
VTR	Video Tape Recorder

5 Source signal

In order to control the characteristics of the source signal, the test sequences should be defined according to the goal of the test and recorded on a digital storage system. Fair and relevant audiovisual test scenes must be chosen such that both their video and audio parts are consistent with the services that the digital transmission service channel was intended to provide. The set of test scenes should span the full range of spatial and temporal information and include any kind of audio signals of interest to users of the devices under test.

The duration of the source sequences should be about 10 s, but not shorter than 8 s. This should be the actual duration of the sequence, that means that the sequences cannot be obtained by repeating a shorter sequence. The termination of the scene should not cause an incomplete sentence or musical phrase. An initial and a final silent period, not longer than 500 ms, can make the sequence to sound more natural.

The quality of the source sequences should be as high as possible: video signal should be recorded in ITU-R BT.601 4:2:2 format and audio should be recorded with a sampling rate of 48 kHz and at least 16 bits per sample. Audio and video should be synchronized when the content requires it.

Annex A lists the audio and video categories that can be used to characterize an audiovisual sequence.

When the experimenter is interested in comparing results from different laboratories, it is necessary to use a common set of source sequences to eliminate a further source of variation.

Characteristics of the recording environment and recording systems should be defined according to 5.1/P.910 and 5.2/P.910 and B.1.1/P.800 and B.1.3/P.800 for the video and the audio part, respectively. Examples of suitable test scenes are given in Annex C.

The number of sequences should be defined according to the experimental design. In order to avoid boring the observers and to achieve a minimum reliability of the results, at least four different types of scenes (i.e. different subject matter) should be chosen for the sequences.

6 Test methods and experimental design

Measurement of the perceived quality of audiovisual sequences requires the use of subjective scaling methods. The condition for such measurements to be meaningful is that there exists a relation between the physical characteristics of the "stimulus", in this case the audiovisual sequence presented to the subjects in a test, and the magnitude and nature of the sensation caused by the stimulus, in this case overall audiovisual quality.

A number of experimental methods have been validated for different purposes. Here, three methods are recommended for range of the bit rates and applications indicated in the scope of this Recommendation.

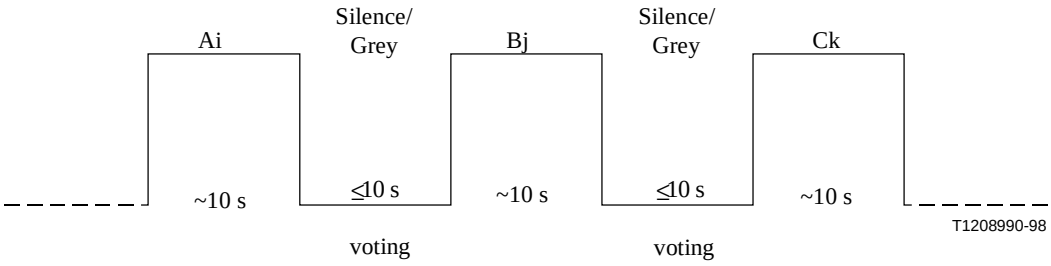
The final choice of one of these methods for a particular application depends on several factors, such as the context, the purpose and where in the development process the test is to be performed.

6.1 Absolute Category Rating (ACR)

The Absolute Category Rating method is a category judgement where the test sequences are presented one at a time and are rated independently on a category scale. (This protocol is similar to protocols described in other ITU-T and ITU-R Recommendations [11], [6] and [4]. In [4], this kind of protocol is referred to as Single Stimulus Method.)

The method specifies that after each presentation, the subjects are asked to evaluate the quality of the sequence presented. The method provides no explicit reference although subjects will always use an implicit reference.

The time pattern for the stimulus presentation can be illustrated by Figure 1. The voting time should be less than or equal to 10 s, depending upon the voting mechanism used. The presentation time may be reduced or increased according to the content of the test material.



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

Figure 1/P.911 – Stimulus presentation in the ACR method

The following five-level scale for rating overall quality should be used (see Table 1):

Table 1/P.911 – Five-level quality scale

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

If higher discriminative power is required, such as in the case of low bit rate coding evaluation, the following nine-level scale should be used (see Table 2):

Table 2/P.911 – Nine-level quality scale

9 8	Excellent
7 6	Good
5 4	Fair
3 2	Poor
1	Bad

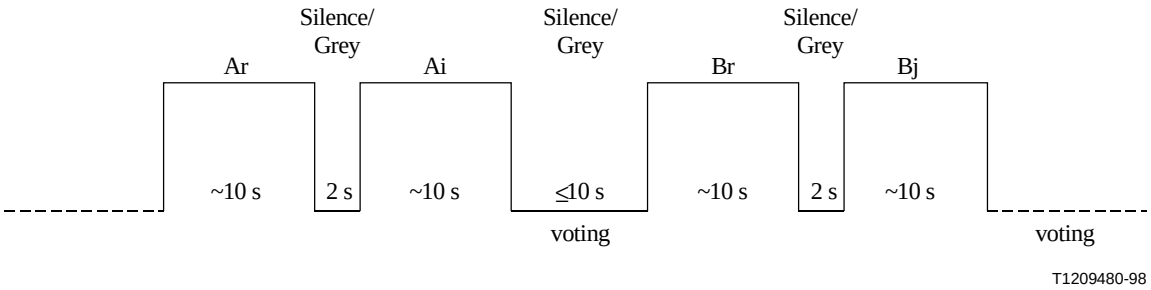
Additional examples of suitable numerical or continuous scales are given in Annex B/P.910, which also gives examples of rating dimensions other than overall quality. Such dimensions may be useful for obtaining more information on different perceptual quality factors when the overall quality rating is nearly equal for certain systems under test, although the systems are clearly perceived as different.

For the ACR method, the necessary number of replications is obtained by repeating the same test conditions at different points of time in the test.

6.2 Degradation Category Rating (DCR)

The Degradation Category Rating implies that the test sequences are presented in pairs: the first stimulus presented in each pair is always the source reference, while the second stimulus is the same source presented through one of the systems under test. (This protocol is similar to protocols described in other ITU-T and ITU-R Recommendations [11], [6] and [4]. In [4], this kind of protocol is referred to as the Double Stimulus Impairment Scale Method.)

The time pattern for the stimulus presentation can be illustrated by Figure 2. The voting time should be less than or equal to 10 s, depending upon the voting mechanism used. The presentation time may be reduced or increased according to the content of the test material.



Ai Sequence A under test condition i
Ar, Br Sequences A and B respectively in the reference source format
Bj Sequence B under test condition j

Figure 2/P.911 – Stimulus presentation in the DCR method

In this case, the subjects are asked to rate the impairment of the second stimulus in relation to the reference.

The following five-level scale for rating the impairment should be used (see Table 3):

Table 3/P.911 – Five-level impairment scale

5	Imperceptible
4	Perceptible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

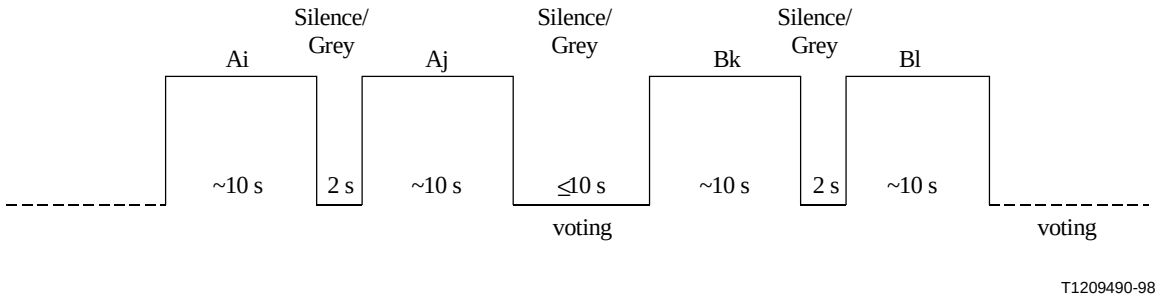
The necessary number of replications is obtained for the DCR method by repeating the same test conditions at different points of time in the test.

6.3 Pair Comparison method (PC)

The method of Pair Comparisons implies that the test sequences are presented in pairs, consisting of the same sequence being presented first through one system under test and then through another system. The source sequence may be included and would be treated as an additional system under test.

The systems under tests (A, B, C, etc.) are generally combined in all the possible $n(n-1)$ combinations AB, BA, CA, etc. Thus, all the pairs of sequences should be displayed in both the possible orders (e.g. AB, BA). After each pair is presented, a judgement is made on which element in a pair is preferred in the context of the test scenario.

The time pattern for the stimulus presentation can be illustrated by Figure 3. The voting time should be less than or equal to 10 s, depending upon the voting mechanism used. The presentation time should be about 10 s and it may be reduced or increased according to the content of the test material.



Ai, Aj Sequence A under i^{th} and j^{th} test condition, respectively

Bk, Bl Sequence B under k^{th} and l^{th} test condition, respectively

Figure 3/P.911 – Stimulus presentation in the PC method

For the PC method, the number of replications need not generally be considered because the method itself implies repeated presentation of the same conditions, although in different pairs.

A variation of the PC method utilizes a categorical scale to further measure the differences between the pair of sequences. See References [4] and [6].

6.4 Single Stimulus Continuous Quality Evaluation (SSCQE)

The Single Stimulus Continuous Quality Evaluation (SSCQE) method has been included in [4] to take into account temporal variations of quality in digital transmission.

The method considers long-duration sequences (3 to 30 min). When the sequences are displayed, the assessment of subjective quality is continuous. This continuous evaluation is carried out by means of sliders that subjects have to move while looking at and/or listening to programmes or scenarios. It is important to notice that no reference is given for anchoring the subjective assessment.

The results are presented by plotting curves which indicate the percentage of time during which the subjective score was higher than a given score on a 0-100 scale. In that scale, 100 represents perfect quality of the considered service(s).

The advantages of this method is that it is well suited to take into account temporal variations of quality and to make global quality assessments. It allows comparison between services when these services are parametrable. Therefore, the method can be a useful tool for the definition of a service. Additionally, this method can be used in a laboratory and also at a user's home as well. A drawback of this method is that it has no reference. It makes it not well suited for tests which require a high degree of discrimination, e.g. the comparison between coders.

The suitability of the Single Stimulus Continuous Quality Evaluation (SSCQE) method to the range of bit rate and kind of applications addressed in this Recommendation is under investigation.

6.5 Comparison of the methods

An important issue in choosing a test method is the fundamental difference between methods that use explicit references (e.g. DCR) and methods that do not use any explicit reference (e.g. ACR, PC and SSCQE). This second class of method does not test fidelity.

The DCR method should be used when testing the fidelity of transmission with respect to the source signal. This is frequently an important factor in the evaluation of high quality systems. Other methods may also be used to evaluate high quality systems. The specific comments of the DCR scale (imperceptible/perceptible) are valuable when the viewer's detection of impairment is an important factor.

Thus, when it is important to check the fidelity with respect to the source signal, the DCR method should be used.

DCR should also be applied for high quality system evaluation in the context of multimedia communication. Discrimination of imperceptible/perceptible impairment in the DCR scale supports this, as well as comparison with the reference quality.

ACR is easy and fast to implement and the presentation of the stimuli is similar to that of the common use of the systems. Thus, ACR is well suited for qualification tests.

The principal merit of the PC method is its high discriminatory power, which is of particular value when several of the test items are nearly equal in quality.

When a large number of items are to be evaluated in the same test, the procedure based on the PC method tends to be lengthy. In such a case, an ACR or DCR test may be carried out first with a limited number of observers, followed by a PC test solely on those items which have received about the same rating.

6.6 Reference conditions

The results of quality assessments often depend not only on the actual audiovisual quality, but also on other factors such as the total quality range of the test conditions, the experience and expectations of the assessors, etc. In order to control some of these effects, a number of dummy test conditions can be added and used as references.

A description of reference conditions and procedures to produce them is given in Recommendations P.930 [1] and P.810 [10] for video and audio, respectively. The introduction of the source signal as a reference condition in a PC test is specially recommended when the impairments introduced by the test items are small.

The quality level of the reference conditions should cover at least the quality range of the test items.

6.7 Experimental design

Different experimental designs, such as complete randomized design, Latin, Graeco-Latin and Youden square designs, replicated block designs, etc. (see Bibliography [5]) can be used, the selection of which should be driven by the purpose of the experiment.

It is left to the experimenter to select a design method in order to meet specific cost and accuracy objectives. The design may also depend upon which conditions are of particular interest in a given test.

It is recommended to include at least two (in some cases three or four) replications (i.e. repetitions of identical conditions) in the experiment. There are several reasons for using replications, the most important being that "within subject variation" can be measured using the replicated data. For testing the reliability of a subject, the same order of presentation under identical conditions can be used. If a different order of presentation is used, the resulting variation in the experimental data is composed of the order effect and the within subject variation.

Replications make it possible to calculate individual reliability per subject and, if necessary, to discard unreliable results from some subjects. An estimate of both within- and between- subject standard deviation is furthermore a prerequisite for making a correct analysis of variance and to generalize results to a wider population. In addition, learning effects within a test are to some extent balanced out.

A further improvement in the handling of learning effects is obtained by including a training session in which at least five conditions are presented at the beginning of each test session. These conditions should be chosen to be representative of the presentations to be shown later during the session. The preliminary presentations are not to be taken into account in the statistical analysis of the test results.

7 Evaluation procedures

7.1 Viewing and listening conditions

Table 4 lists typical viewing and listening conditions as used in audiovisual quality assessment. The actual parameter settings used in the assessment should be specified. For the comparison of test results, all viewing and listening conditions must be fixed and equal over laboratories for the same kind of tests.

Both the size and the type of monitor used should be appropriate for the application under investigation. When sequences are presented through a PC based system, the characteristics of the display and audio transducers must be specified, e.g. dot pitch and of the monitor, type of video display card used, characteristics of either handsets, headphones or loudspeakers, etc.

In particular, in case of loudspeaker presentation, the number and positions of the loudspeakers relative to the image should be reported.

Operational parameters, such as signal level, for the test sequences shall match those of the alignment signal used to verify the viewing and listening conditions. Any operational adjustments performed so that source or processed sequences meet this requirement should be reported.

Concerning the displaying format, it is preferable to use the whole screen for displaying the sequences. Nevertheless when, for some reason, the sequences must be displayed on a window of the screen, the colour of the background in the screen should be 50% grey corresponding to Y = U = V = 128 (U and V unsigned).

Synchronization between audio and video signal should be measured according to P.931 [18] and reported.

Table 4/P.911 – Typical viewing and listening conditions as used in audiovisual quality assessment

Parameter	Setting
Room size (Note 7)	Specify L × W × H
Viewing distance (Note 5)	1-8 H
Peak luminance of the screen	100-200 cd/m^2
Ratio of luminance of inactive screen to peak luminance	≤0.05
Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white	≤0.1
Ratio of luminance of background behind picture monitor to peak luminance of picture (Note 1)	≤0.2
Chromaticity of background (Note 6)	D_{65}
Background room illumination (Note 1)	≤20 lux

Table 4/P.911 – Typical viewing and listening conditions as used in audiovisual quality assessment *(concluded)*

Parameter	Setting
Background noise level (Note 2)	≤30 dBA
Listening level (Note 3)	~80 dBA
Reverberation time (Note 4)	<500 ms, ∀f > 150 Hz

NOTE 1 – This value indicates a setting allowing maximum detectability of distortions, for some applications higher values are allowed or they are determined by the application.

NOTE 2 – If background noise levels used in the application are significantly higher, Hoth noise (see Bibliography [9]) should be used for environments like office. The Hoth noise should be generated in a room with low background noise levels (≤ 30 dBA) and be measured in the acoustical domain. If for some special type of application another type of background noise is used, the spectral power density and dBA level should be specified.

NOTE 3 – This value indicates a setting allowing maximum detectability of distortions, for some applications lower or higher values are allowed. Level setting is measured using the maximum value over the audio sequence using IEC Publication [17] fast averaging. When loudspeakers are used, the sound level might be adjusted according to the viewing distance (see Bibliography [11]).

NOTE 4 – This value is only relevant for loudspeaker presentation, larger reverberation times generally lead to a decrease in detectability of distortions.

NOTE 5 – For a given screen height, it is likely that the viewing distance preferred by the subjects increases when visual quality is degraded. Concerning this point, the preferred viewing distance should be predetermined for qualification tests. Viewing distance in general depends on the applications.

NOTE 6 – For PC monitors, the chromaticity of background may be adapted to the chromaticity of the monitor.

NOTE 7 – Room size is important only for loudspeaker presentation.

7.2 Processing and playback system

There are two methods for obtaining test images from the source recordings:

a) by transmitting or replaying the audiovisual recordings in real time through the systems under test, while subjects are watching, listening and responding;

b) by off-line processing of the source recordings through the device under test and recording the output to give a new set of recordings.

In the second case, the signal should be recorded on digital format to minimize the impairments that can be produced by recording process. The test conditions can be recorded on component ITU-R BT.601 storage device. In any case, taking into account that the impairments introduced by low-bit rate coding schemes are usually more evident than the impairments introduced by professional quality recorders, VTRs such as D2, MII and BetacamSP can be used.

7.3 Subjects

The possible number of subjects in a viewing and listening test (as well as in usability tests on terminals or services) is from 6 to 40. Four is the absolute minimum for statistical reasons, while there is rarely any point in going beyond 40.

The actual number in a specific test should really depend on the required validity and the need to generalize from a sample to a larger population.

In general, at least 15 subjects should participate in the experiment. They should not be directly involved either in picture or audio quality evaluation as part of their work and should not be experienced assessors.

Prior to a session, the observers should usually be screened for normal visual acuity or corrected-to-normal acuity and for normal colour vision.

Nevertheless, in the early phases in the development of audiovisual communication systems and in pilot experiments carried out before a larger test, small groups of experts (4-8) or other critical subjects can provide indicative results.

7.4 Instructions to subjects and training session

Before starting the experiment, a scenario of the intended application of the system under test should be given to the subjects. In addition, a description of the type of assessment, the opinion scale and the presentation of the stimuli is given in written form. The range and type of impairments should be presented in preliminary trials, which may contain audiovisual sequences other than those used in the actual tests.

It must not be implied that the worst quality experienced in the training set necessarily corresponds to the lowest subjective grade on the scale.

Questions about procedure or about the meaning of the instructions should be answered with care to avoid bias and only before the start of the session.

A possible text for instructions to be given to the assessors is suggested in Appendix II.

8 Statistical analysis and reporting of results

The results should be reported along with the details of the experimental set-up. For each combination of the test variables, the mean value and the standard deviation of the statistical distribution of the assessment grades should be given.

From the data subject reliability should be calculated and the method used to assess subject reliability should be reported. Some criteria for subjective reliability are given in [4] and [5].

It is informative to analyse the cumulative distribution of scores. Since the cumulative distributions are not sensitive to linearity, these may be particularly useful for data for which the linearity is doubtful, as those obtained by using the ACR and DCR methods, together with category scales without grading (i.e. category judgement).

The data can be organized for example as shown in Table 5 for ACR:

Table 5/P.911 – Informative table with cumulative distribution of scores for ACR method

Condition	Total votes	Excellent	Good	Fair	Poor	Bad	MOS	CI	Std	%GOB	%POW

Condition	Label indicating a combination of test variables
Total votes	Number of votes collected for that condition
Excellent, Fair ... Bad	Occurrence of each vote

The classical techniques of analysis of variance should be used to evaluate the significance of the test parameters. If the assessment is aimed at evaluating the audiovisual quality as a function of a parameter, curve fitting techniques can be useful for the interpretation of the data.

In the case of pair comparisons, the method to calculate the position of each stimulus on an interval scale, where the difference between the stimuli corresponds to the difference in preference, is described in the *Handbook on Telephonometry*, Section 2.6.2C and in [6].

ANNEX A

Details related to the characterization of the test sequences

The characterization of audiovisual sequences will be based both on audio and video content characteristics. Tables A.1 and A.2 below illustrate the categories of the video and audio part, respectively. The tables are not intended to be exhaustive.

According to this classification, a videophone scene will belong to category A-I.

Table A.1/P.911 – Video content categories

Category	Description
A	One person, mainly head and shoulders, limited detail and motion
B	One person with graphics and/or more detail
C	More than one person
D	Graphics with pointing
E	High object and/or camera motion beyond the range usually found in video teleconferencing

Table A.2/P.911 – Audio content categories

Category	Description
I	Speech/one speaker
II	Speech/Multiple speakers
III	Speech + background music
IV	Music/single instrument
V	Music/multiple instruments

ANNEX B

Video and audio classes and their attributes

B.1 Video classes and their attributes

In this Recommendation, the highest video quality considered is ITU-R Recommendation BT.601, 8 bits/sample linear PCM coded video in 4:2:2, Y, C_R, C_B format.

See Tables B.1 and B.2

Table B.1/P.911 – Definitions of video classes

TV 0	Loss-less: ITU-R Recommendation BT.601, 8 bits per sample, video used for applications without compression.
TV 1	Used for complete post-production, many edits and processing layers, intra-plant transmission. Also used for remote site to plant transmission. Perceptually transparent when compared to TV 0.
TV 2	Used for simple modifications, few edits, character/logo overlays, programme insertion, and inter-facility transmission. A broadcast example would be network-to-affiliate transmission. Other examples are a cable system regional downlink to a local head-end and a high quality video conferencing system. Nearly perceptually transparent when compared to TV 0.
TV 3	Used for delivery to home/consumer (no changes). Other examples are a cable system from the local head-end to a home and medium to high quality video conferencing. Low artifacts are present when compared to TV 2.
MM 4	All frames encoded. Low artifacts relative to TV 3. Medium quality video conferencing. Usually ≥25 fps for 625-lines systems and ≥30 fps for 525-lines systems.
MM 5	Frames may be dropped at encoder. Perceivable artifacts possible, but quality level useful for designed tasks, e.g. low quality video conferencing.
MM 6	Series of stills. Not intended to provide full motion (examples: surveillance, graphics).

Table B.2/P.911 – Attributes of video classes

Video class	Spatial format	Delivered frame rate (Note 1)	Typical latency (Note 2) Delay variation	Nominal video bit rate, Mbit/s
TV 0	ITU-R Rec. BT.601	Max FR	(Note 2)	270
TV 1	ITU-R Rec. BT.601	Max FR	(Note 2)	18 to 50
TV 2	ITU-R Rec. BT.601	Max FR	(Note 2)	10 to 25
TV 3	ITU-R Rec. BT.601	Max FR occasional Frame repeat	(Note 2)	1.5 to 8
MM 4a	ITU-R Rec. BT.601	~30 or ~25 fps	Delay <≈150 ms Variation <≈50 ms	~1.5
MM 4b	CIF	~30 or ~25 fps	Delay <≈150 ms Variation <≈50 ms	~0.7

Table B.2/P.911 – Attributes of video classes *(concluded)*

Video class	Spatial format	Delivered frame rate (Note 1)	Typical latency (Note 2) Delay variation	Nominal video bit rate, Mbit/s
MM 5a	CIF	10-30 fps	Delay <≈1000 ms Variation <≈500 ms	~0.2
MM 5b	≤CIF	1-15 fps	Delay <≈1000 ms Variation <≈500 ms	~0.05
MM 6a	CIF-16CIF	Limit → 0 fps	No restrictions (Note 3)	<0.05, Limit → 0 fps
MM 6b	CIF-16CIF	Limit → 0 fps	No restrictions (Note 3)	<0.05, Limit → 0 fps

NOTE 1 – Normally 30 fps for 525 systems and 25 fps for 625 systems.

NOTE 2 – Broadcast systems all have constant, but not necessarily low, one-way latency and constant delay variation. For most broadcast applications latency will be low, say between 50 and 500 ms. For high quality video conferencing, and conversational types of applications in general, latency should be preferably less then 150 ms [15]. Delay variations are allowed within the given range but should not lead to perceptually disturbing time-warping effects.

NOTE 3 – Differs only in audio bit rate.

B.2 Audio classes and their attributes

In this Recommendation, the highest audio quality considered is 5.1 surround sound, 20 bits/sample linear PCM coded audio [16]. In referring to the audio classes, the channel set-up should be given, e.g. mono, stereo, 3 channel, 3/1 channel, etc.

See Tables B.3 nd B.4.

Table B.3/P.911 – Definitions of audio classes

TV 0	Studio quality, 20 bits/sample, 48 kHz loss-less, linear PCM.
TV 1	Used for complete post-production, many edits and processing layers, intra-plant transmission. Also used for remote site to plant transmission. Perceptually transparent when compared to TV 0.
TV 2	Primary distribution: Used for simple modifications, few edits, programme mixing, and inter-facility transmission. A broadcast example would be network-to-affiliate transmission. Other examples are a cable system regional downlink to a local head-end and a high quality video conferencing system. Nearly perceptual transparent when compared to TV 0.
TV 3	Used for delivery to home/consumer (no changes). Other examples are a cable system from the local head-end to a home and medium to high quality video conferencing. Low audible artifacts are present when compared to TV 2.
MM 4	Low audible artifacts relative to TV 3 using speech and audio. Medium quality video conferencing. Usually full audio bandwidth (20-20 000 Hz), but bandwidths (Note) down to wideband (50-7000 Hz) are acceptable.

Table B.3/P.911 – Definitions of audio classes *(concluded)*

MM 5	Low audible artifacts relative to a narrow-band reference (300-3400 Hz telephony band) using speech and music. Perceptual artifacts possible, but quality level useful for designed tasks, e.g. low quality video conferencing.
MM 6	Severe audible artifacts relative to a narrow-band (300-3400 Hz) telephony application. Speech is however still intelligible.
NOTE – Bandwidth is defined as the perceptual bandwidth (i.e. bandwidth of the system under test without taking into account inaudible components) of the system under test.	

Table B.4/P.911 – Attributes of audio classes

Audio class	Sampling frequency (kHz)	Typical latency (Note 1) Delay variation	Nominal bit rate, kbit/s/channel (Note 2)
TV 0	48.0	(Note 1)	960
TV 1	44.1 or 48.0 or 32.0	(Note 1)	250-500
TV 2	44.1 or 48.0 or 32.0	(Note 1)	120-300
TV 3	44.1 or 48.0 or 32.0	(Note 1)	50-150
MM 4a	44.1 or 48.0 or 32.0	Delay <≈150 ms Variation <≈ ±20 ms	20-100
MM 4b	12-32	Delay <≈150 ms Variation <≈ ±20 ms	10-50
MM 5a	12-20	Delay <≈400 ms (Note 3) Variation <≈ ±100 ms	4-32
MM 5b	8-12	Delay <≈400 ms (Note 3) Variation <≈ ±200 ms	2-16
MM 6a	8	Delay <≈400 ms (Note 4) Variation <≈ ±200 ms	<8
MM 6b	8	Delay <≈4000 ms (Note 4) Variation <≈ ±2000 ms	<8

NOTE 1 – Broadcast systems all have constant, but not necessarily low, latency and constant delay variation. For most broadcast applications latency will be low, say between 50 and 500 ms. For high quality video conferencing, and conversational types of applications in general, one-way latency should be preferably less than 150 ms. Delay variations are allowed within the given range but should not lead to perceptually disturbing time-warping effects.

NOTE 2 – The low frequency enhancement channel (the 0.1 in 5.1 presentation) requires only a marginally higher bit rate. For an N-channel audio signal mutual correlations will be exploited, resulting in a total bit rate that may be significantly lower than N times the rate per channel.

NOTE 3 – For this application, synchronicity between audio and video may not give the highest conversational quality of the communication link.

NOTE 4 – Audiographics conferencing.

ANNEX C

Discussion about the Relationship between audio, video and audiovisual quality

The body of this Recommendation deals with the overall one-way audiovisual quality of audiovisual sequences. The relations between audio, video and audiovisual quality are under study but some stable results are available. The most important one is the mapping from the one-way audio and one-way video quality, as derived from audio only and video only subjective experiments, to the one-way overall audiovisual quality. Four different laboratories found similar mapping results despite the fact that experimental conditions were quite different (Bibliography [13]-[17]). A series of experiments was performed (Bibliography [13] and [14]) using blurring for the video and multiplicative noise for the audio degradations. In Bibliography [15], a similar experiment is described with TV commercials using broadcast television quality as the highest, reference, audio and video quality. Similar distortions were used for the video degradations (VIRIS 601, ANSI [8] extension V3, V6, V8) but band limiting for the audio distortions (ranging from full bandwidth, 20-20 000 Hz, to narrow-band 300-3400 Hz). In Bibliography [16], a one-way video teleconferencing quality assessment was used with several types of coding distortions. Three different video codecs (two proprietary and H.261) and four different audio codecs were used. In Bibliography [17], also a one-way video teleconferencing quality assessment was used with several types of coding distorsions. In this case, also uncoded conditions (Full Band for audio and PAL signal for video) were introduced. The results of the experiments, which used either the nine-point or the five-point ACR quality scale as given in 6.1 show that:

1) Video dominates overall perception. When the variance in audio quality and video quality are about the same, the variance in overall audiovisual quality is dominated by the variance in video quality. The correlation between video and overall audiovisual quality is higher than the correlation between audio and overall audiovisual quality.

2) The one-way overall audiovisual quality can be predicted from the one-way audio and one-way video quality as derived from audio only and video only subjective experiments.

3) The most stable mapping across laboratories from the separate audio and video quality to the overall audiovisual quality was found to be $MOS_{av} = \alpha + \beta * MOS_a * MOS_v$. This mapping is based on four sets of subjective experiments and the correlation between predicted and measured overall audiovisual quality varied from 0.93 to 0.99. The value of α varied between 1.1 and 1.5, the range of β varied between 0.107 and 0.121, when all scales are mapped onto nine-point scale. Provisionally the recommended values are set to 1.3 for α and 1.1 for β.

The validity of the model was proved only for synchronized audio and video signals and only within the bounds of the experiments where impairments in audio and video are approximately equally disturbing. It is possible to design an experiment where the relationship [(item 3) above] would be invalid. For example, if the treatments include range of audio qualities where intelligibility is in question, and a small range of video qualities, then it is expected that the audio assessment would dominate the combined audio and video quality.

APPENDIX I

Test audiovisual sequences

The selection of appropriate test sequences is a key point in the planning of subjective assessment. When results of tests carried out with different groups of observers or in different laboratories have to be correlated, it is important that a common set of test sequences is available.

A first set of such sequences is described in Table I.1. In this table the following information is given for each sequence:

– the category (defined in Tables A.1 and A.2);

– a brief description of the scene;

– the source format (either 625- or 525-lines, either ITU-R BT.601 format or BetacamSP);

– the sampling frequency of the audio signal;

– the values of spatial and temporal information (defined in 5.3.1/P.910 and 5.3.2/P.910 respectively).

All the sequences listed in Table I.1 are in the public domain and may be used freely for evaluations and demonstrations.

Table I.1/P.911 – Test sequences for audiovisual quality assessment in multimedia applications

(To be determined)

Sequence	Category (A + V)	Description	Source format	Language	Sampling frequency	SI	TI
	A-I						
	B-III						
	D-I						
	D-II						
	etc.						

APPENDIX II

Instructions for audiovisual tests

The following may be used as the basis for instructions to assessors involved in experiments adopting either ACR, DCR or PC methods.

In addition, the instructions should give information about the approximate test duration, pauses, preliminary trials and other details helpful to the assessors. This information is not included here because it depends on the specific implementation.

II.1 ACR

Good morning and thank you for coming.

In this experiment, you will be presented with short audiovisual sequences. Each time a sequence is presented, you should judge its quality by using one of the five levels of the following scale.

5 Excellent

4 Good

3 Fair

2 Poor

1 Bad

Your evaluation must reflect your opinion of the overall combined **audio and video** quality.

Observe and listen carefully to the entire sequence before making your judgement.

II.2 DCR

Good morning and thank you for coming.

In this experiment, short audiovisual sequences will be presented. Each sequence will be presented twice in rapid succession: the first sequence is the reference sequence, the second sequence the processed sequence. At the end of each paired presentation, you should evaluate the impairment of the second sequence with respect to the first one. You will express your judgement of the difference between the second and the first sequence by using the following scale:

5 Imperceptible

4 Perceptible but not annoying

3 Slightly annoying

2 Annoying

1 Very annoying

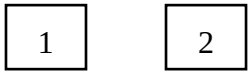
Your evaluation must reflect your opinion of the overall combined **audio and video** quality.

Observe and listen carefully to the entire sequence before making your judgement.

II.3 PC

Good morning and thank you for coming.

In this experiment, short audiovisual sequences will be presented. Each sequence will be presented twice in rapid succession: each time through a different system. The order of the sequences and the combination of codecs in the pairs vary in a random way. At the end of each paired presentation, you should express your preference by ticking one of the boxes shown below. You will tick box #1 if you prefer the first sequence or box #2 if you prefer the second sequence of the pair.



Your evaluation must reflect your opinion of the overall combined **audio and video** quality.

Observe and listen carefully to the entire sequence before making your judgement.

APPENDIX III

Bibliography

[1] GONZALEZ (R.C.), WINTZ (P.): Digital Image Processing, 2nd Edition, *Addison-Wesley Publishing Co.*, Reading, Massachusetts, 1987.

[2] RACE Industrial Consortium Project 1018 HIVITS, WP B5: Picture Quality Measurement, 1988.

[3] Grahm-Field Catalogue Number 13-1240.

[4] Pseudo Isochromatic Plates, engraved and printed by *The Beck Engraving Co., Inc.*, Philadelphia and New York, United States.

[5] KIRK (R.E.): Experimental Design – Procedures for the Behavioural Sciences, 2nd Editions, *Brooks/Cole Publishing Co.*, California, 1982.

[6] VIRTANEN (M.T.), GLEISS (N.), GOLDSTEIN (M.): On the use of Evaluative Category Scales in Telecommunications, HFT 1995, *Human Factors in Telecommunication Conference*, Melbourne, 1995.

[7] GUILFORD (P.): Psychometric methods, MCGRAW-HILL, New York, 1954.

[8] ANSI T1A1.5 contribution 96-109: *VIRIS for ITU-R 601 digital images*, May 1996.

[9] HOTH (D.F.): Room noise spectra at subscribers' telephone locations, *J.A.S.A.*, Volume 12, pp. 99-504, April 1941.

[10] BECH (S.): Calibration of Relative Level Differences of a Domestic Multichannel Sound Reproduction System, *Journal of Audio Engineering Society*, April 1998.

[11] ITU-T Delayed Contribution D.40 (1990), *Some speech quality aspects to be considered in multimedia services*, NTT.

[12] *Handbook on Telephonometry*, ITU, Geneva, 1992.

[13] ITU-T COM 12-20 (December 1993), *Experimental combined audio/video subjective test method*, Bellcore.

[14] ITU-T COM 12-37 (September 1994), *Extension of combined audio/video quality model*, Bellcore.

[15] ITU-T COM 12-19 (February 1998), *Relations Between audio, Video and Audiovisual Quality*, KPN Research, Netherlands.

[16] ITU-T COM 12-64 (November 1998), *Results of an audiovisual desktop video teleconferencing subjective experiment*, USA.

[17] ITU-T COM 12-61 (November 1998), *Study of the influence of experimental context on the relationship between audio, video and audiovisual subjective quality*, France Telecom/CNET.

ITU-T RECOMMENDATIONS SERIES

Series A	Organization of the work of the ITU-T
Series B	Means of expression: definitions, symbols, classification
Series C	General telecommunication statistics
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	TMN and network maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks and open system communications
Series Y	Global information infrastructure
Series Z	Programming languages