



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 Recommendation)

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.

© ITU 1999

All rights reserved. No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from the ITU.

CONTENTS

	Page
1 Scope	1
2 References	1
3 Terms and definitions	2
4 Abbreviations	3
5 Source signal	4
6 Test methods and experimental design.....	5
6.1 Absolute Category Rating (ACR)	5
6.2 Degradation Category Rating (DCR)	6
6.3 Pair Comparison method (PC).....	7
6.4 Single Stimulus Continuous Quality Evaluation (SSCQE).....	8
6.5 Comparison of the methods.....	8
6.6 Reference conditions	9
6.7 Experimental design	9
7 Evaluation procedures	10
7.1 Viewing and listening conditions	10
7.2 Processing and playback system	11
7.3 Subjects.....	11
7.4 Instructions to subjects and training session	12
8 Statistical analysis and reporting of results	12
Annex A – Details related to the characterization of the test sequences	13
Annex B – Video and audio classes and their attributes.....	14
B.1 Video classes and their attributes	14
B.2 Audio classes and their attributes	15
Annex C – Discussion about the Relationship between audio, video and audiovisual quality	17
Appendix I – Test audiovisual sequences.....	18
Appendix II – Instructions for audiovisual tests.....	18
II.1 ACR.....	19
II.2 DCR.....	19
II.3 PC	19
Appendix III – Bibliography	20

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 [$\text{Sobel}(F_n)$]. The standard deviation over the pixels ($\text{std}_{\text{space}}$) 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:

$$\text{SI} = \text{max}_{\text{time}} \{ \text{std}_{\text{space}} [\text{Sobel}(F_n)] \}$$

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 ($\text{std}_{\text{space}}$) of $M_n(i,j)$ over all i and j .

$$\text{TI} = \text{max}_{\text{time}} \{ \text{std}_{\text{space}} [M_n(i,j)] \}$$

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 $\times$ 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.