International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.800.2

(05/2013)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Methods for objective and subjective assessment of speech quality

Mean opinion score interpretation and reporting

Recommendation ITU-T P.800.2

ITU-T

International Telecommunication Union

ITU-T P-SERIES RECOMMENDATIONS

TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	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 speech quality	**Series**	**P.80**
		P.800
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400

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

Recommendation ITU-T P.800.2

Mean opinion score interpretation and reporting

Summary

Recommendation ITU-T P.800.2 introduces some of the more common types of mean opinion score (MOS) and describes the minimum information that should accompany MOS values to enable them to be correctly interpreted.

History

Edition	Recommendation	Approval	Study Group
1.0	ITU-T P.800.2	2013-05-14	12

Keywords

Absolute category rating (ACR), mean opinion score (MOS), objective model, reporting, subjective experiment.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. 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 Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 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 expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. 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, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T P.800.2

Mean opinion score interpretation and reporting

1 Scope

This Recommendation introduces some of the more common types of mean opinion score (MOS) and describes the minimum information that should accompany MOS values to enable them to be correctly interpreted.

It should be noted that this text does not aim to provide a definitive guide to subjective or objective testing. The bibliography at the end of this Recommendation provides information on more detailed material.

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; 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. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T P.800.1] Recommendation ITU-T P.800.1 (2006), *Mean Opinion Score (MOS) terminology*.

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 condition: One of a set of use cases being evaluated in a subjective experiment; often referred to as a hypothetical reference circuit (HRC) in video experiments.

3.2.2 sub-condition: A subset of a condition defined by a specific characteristic of the use case, e.g., speech material from a particular talker.

3.2.3 subject: A participant in a subject experiment.

3.2.4 vote: A subject's response to a question in a rating scale for an individual test sample or interaction.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR Absolute Category Rating

DCR Degradation Category Rating

DMOS Degradation Mean Opinion Score

HRC Hypothetical Reference Circuit

MOS	Mean Opinion Score
MUSHRA	Multi-stimulus test with Hidden Reference and Anchor
QCIF	Quarter Common Intermediate Format
SSCQE	Single Stimulus Continuous Quality Evaluation
VGA	Video Graphics Array

5 Introductory information

Audio and video quality are inherently subjective quantities. This means that the baseline for audio and video quality is the opinion of the user. However, one person's opinion of what is 'good' may be quite different to another person's opinion – neither person is correct, neither person is incorrect.

Before a new audio or video transmission technology is deployed, it is good practice to assess the transmission quality using one or more subjective experiments. The purpose of a subjective experiment is to collect the opinions of multiple people ("subjects") about the performance of the system for a number of well-defined use cases ("conditions")[1]. The mean opinion score (MOS) for a given condition is simply the average of the opinions ("votes") collected for that use case.

Objective quality measurement algorithms aim to predict the MOS value that a given input signal would produce in a subjective experiment. Hence, when interpreting an objectively derived MOS value, it is essential to understand the basic design of the experiment being predicted.

There are several different types of MOS value and many different test methodologies for producing them. The purpose of this Recommendation is to give the reader an appreciation of the main points to consider when interpreting MOS values and the minimum information that should accompany MOS values when they are reported.

6 Subjective MOS values

Types of MOS

There is a common misconception that MOS values only pertain to voice services, but the process of asking subjects to provide their assessment of quality can be just as easily be applied to video and general audio services as it can to voice services. It is also possible to ask subjects to rate the overall audiovisual quality of a service. The ITU has produced various standards describing different aspects of subjective testing for video and general audio applications in addition to voice applications, and these are listed in the bibliography.

Subjective experiments may be broadly divided into two types: passive and interactive. In a passive subjective experiment, subjects are presented with pre-recorded test samples representing the conditions of interest. The subjects are asked to passively listen to and/or watch the test material and provide their opinion using the rating scale provided. In an interactive experiment, two or more subjects actively engage in conversation using equipment designed to emulate the use cases of interest. The subjects are often given tasks in order to stimulate conversation and interaction. Most experiments tend to be passive in nature. However, there are some aspects of user experience, for example, the effects of delay and echo, that only become apparent in conversational scenarios.

Test methodology and rating scale

In a subjective experiment, subjects are asked to provide their opinions using a "rating scale". The purpose of the scale is to translate a subject's quality assessment into a numerical value that can be averaged across subjects and other experimental factors.

---

1 In video experiments, conditions are often referred to as hypothetical reference circuits (HRCs).

There are several rating scales in common use, and the relative benefits of different scales are outside the scope of this Recommendation. The most commonly used scale is the 5-point absolute category rating (ACR) scale:

Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

The ACR scale is a discrete scale, meaning that the subject's response is limited to one of the five values listed above. However, the averaging process used to combine results from different subjects means that MOS values are not confined to integer values. Some rating scales have more than five discrete labels, while others allow the subject to provide intermediate responses at points between the labels.

The "absolute" part of ACR relates to the fact that subjects are asked to independently rate each sample. Some rating scales, such as the degradation category rating (DCR) scale, ask for a subject's opinion about the difference between a sample processed through the condition of interest and an unprocessed version of the same sample. The MOS value produced in such an experiment is often called a degradation MOS or DMOS.

In most experimental designs, subjects are asked to rate the quality of short audio or video samples. The duration of such samples is usually in the range of 6 to 10 seconds, as this provides enough time for the subject to form an opinion without introducing any bias towards the end of the sample. It is difficult for a single sample of this duration to represent a whole condition, and hence subjects are typically asked to rate multiple test samples derived from the same use case. For example, in a voice experiment, each network condition under test might be represented with speech samples from three male and three female talkers. This means that MOS values can be produced for the entire condition, by averaging across both subjects and talkers, or for a sub-condition such as a particular talker or gender of talker.

Test methods such as single stimulus continuous quality evaluation (SSCQE) use much longer test samples, and require the subject to continuously update their opinion of quality as the test sample is being played. This results in a time sequence of quality ratings from each subject, rather than a single opinion value.

Some test methodologies require the subject to answer multiple questions. Not only does this yield more information about the conditions under test, it can be a necessary part of the test design. For example, the ITU-T P.835 test method requires the subject to provide separate opinions about the speech quality and the noise quality of a sample before providing an overall quality score. This process has been found to yield more stable results with noise suppression systems than the single question ACR test method.

It should be noted that some questions may not relate directly to quality, but may address a different aspect of communications, for example, [b-ITU-T P.800] defines a listening *effort* scale for voice experiments. Similarly, some conversational experiments ask the subject about their experience when talking, rather than when listening.

7 Interpreting MOS values

The following discussion initially focuses on voice MOS values; however, many of the points made in the subsections apply equally to video, audio and audio-video MOS values. The main differences for video are described in the following section.

The idea that a particular voice codec has a particular MOS score is another common misconception. One source of this misconception is the widespread use of objective quality assessment models, which produce very repeatable results. Such models are designed to predict or estimate the output of subjective experiments; however, for any given codec at a given bit rate, the MOS value obtained in a subjective experiment can vary substantially from experiment to experiment. There are a number of reasons for this.

First, the exact MOS values obtained for a particular condition in a subjective experiment can be influenced by a large number of factors, including but not limited to:

– the equipment used to present the material (handset, headset, speakers)
– monaural, diotic binaural or stereo presentation
– presentation level
– acoustic environment
– preparation of subjects
– subject profile, e.g., age and technology exposure
– differences in interpretation and use of rating scales across cultures
– speech material (phonetic content and talker characteristics)
– language (presence/absence, prevalence and importance of particular sounds and transitions).

Second, the exact MOS value that is obtained for any given condition in a subjective experiment depends on the quality of the other conditions in the experiment. For example, an ITU-T G.729 voice codec condition may score more than 3.9 in an ACR experiment if most of the other conditions are of worse quality than ITU-T G.729; conversely, the ITU-T G.729 condition may score significantly less than 3.9 if most of the other conditions exhibit better quality.

Third, if an experiment is run with codecs operating at different audio bandwidths, then the presence of higher bandwidth conditions will reduce the MOS produced for conditions with a lower audio bandwidth. The highest audio bandwidth present in a voice experiment is often called the "context" of the experiment. For example, an ITU-T G.711 voice codec condition will often yield a score above 4.0 in a narrow-band (300-3700 Hz) ACR experiment; whereas it is more likely to yield a score in the range of 3.5-3.7 in a wideband (50-7000 Hz) ACR experiment, due to the presence of the higher quality wideband samples.

These last two points reflect the fact that subjects in experiments tend to adapt their use of the rating scale to the content of the experiment. Indeed, well-designed experiments include a practice period at the start of the experiment when subjects hear examples of a range of conditions, including the best and the worst.

One of the most important consequences of the considerations described above is that it is not meaningful to directly compare MOS values produced from separate experiments, unless those experiments were explicitly designed to be compared, and even then the data should be statistically analysed to ensure that such a comparison is valid.

8 Video considerations

Many of the considerations described above in relation to voice subjective experiments also apply to video experiments. The experimental conditions, often called hypothetical reference circuits (HRCs), typically define various combinations of video codec, bit-rate, frame-rate and transmission conditions, The factors influencing the exact MOS values obtained for a particular condition include, but are not limited to:

– the equipment used to present the material (display technology, refresh rate, contrast, etc.);

– viewing environment (colour, temperature and lighting level);

– viewing distance (usually expressed as the ratio of the viewing distance to the display height);

– video content.

This last point is particularly important for video experiments. The choice of test material is a much stronger factor in video experiments than it is for voice experiments. This is because content of a video sequence can have a highly significant effect on how efficiently it can be encoded. For example, the information content in a fast moving sports sequence is much higher than in a head and shoulders video conferencing sequence.

For video experiments, the primary context is determined by the resolution of the video image. In general, subjective experiments do not mix different resolutions, and therefore video MOS values pertain to a particular resolution, e.g., quarter common intermediate format (QCIF) or VGA. In cases where resolutions are mixed, the context of the experiment will be defined by the resolution with the largest number of lines. In this case, it is important to note whether the smaller resolutions are displayed natively or are resized to the largest resolution in the experiment.

9 Statistical analysis of MOS

The statistical analysis of subjective MOS values is outside the scope of this Recommendation. However, MOS values should be accompanied by sufficient information to allow a basic statistical analysis to be performed, for example, the calculation of a confidence interval for each condition. For any given condition or sub-condition, this information comprises the number of votes, the mean of the votes and the standard deviation of the votes.

10 Objective MOS values

The purpose of an objective quality model is to predict the MOS value that an audio or video signal would obtain in a subjective experiment. As discussed above, the exact MOS value produced in any given experiment for a particular codec or transmission chain depends on many different aspects of the experiment's design and execution. Objective model designers therefore have to predict an idealized experiment. This is typically an experiment that is conducted according to a specific test methodology, usually ACR, and includes a balanced sample of the distortions that will be encountered in the application area of interest.

For example, the mapping defined in Recommendation ITU-T P.862.1 (*Mapping function for transforming P.862 raw result scores to MOS-LQO*) takes the raw output of the ITU-T P.862 objective model and maps it to a range that was determined by averaging the output of a large number of subjective experiments conducted according to the ACR method as described in [b-ITU-T P.800]. A similar mapping is built into the output stage of the ITU-T P.863 model.

One of the advantages of an objective model is that the results are repeatable and hence measurements made at different times and locations can be directly compared. However, care should still be taken as factors such as the choice of test material and any pre or post-processing can still introduce a bias into the results.

For reasons that should now be apparent, different objective models may produce different predicted MOS values for the same conditions. For example, the ITU-T P.862.1 and ITU-T P.863 models do not produce exactly the same predicted MOS values for ITU-T G.729 encoded speech, even though this codec is within the scope of both models. This is partly because the two models have been trained and optimized using different subjective experiments. For this reason, when comparing objective MOS predictions with thresholds, for example to monitor a service level agreement or to raise an alarm, such thresholds should be chosen in the context of the model producing the prediction.

11 Reporting subjective MOS values

Table 1 describes what information must be provided when reporting subjective MOS values, and what additional information is recommended to be provided.

If an experiment has been conducted according to an ITU Recommendation, the information about the methodology can usually be represented with a simple reference to the relevant standard and the particular method used, although variations from the standard procedures should be noted.

It is important to always provide information about the test samples used for passive experiments. In the case of video samples, it can be useful to provide more detailed information, for example whether particular sequences contain panning or scene changes.

Table 1 – Minimum information for reporting subjective MOS values

Information	Experiment type	Provision
Methodology Passive or interactive Sample-based or continuous assessment Absolute or relative assessment of samples Question(s) presented to subjects Rating scale labels Whether rating scale is discrete or continuous Sample duration Or ITU Recommendation and method used	All	Mandatory
Test plan Purpose of experiment Date and place test was run Processing information Experimental design, e.g., blocking design Number of sessions and duration Number of subjects Subject profiles, age and gender distributions Type of subjects used, e.g., naive or expert Information about equipment used	All	Recommended

Table 1 – Minimum information for reporting subjective MOS values

Information	Experiment type	Provision
Condition/HRC information Number of conditions List of conditions Average votes per condition (MOS) Standard deviation of votes per condition Number of votes per condition	All	Mandatory
Sub-condition information List of sub-condition factors MOS values for sub-conditions Number of votes and variance per sub-condition	All	Optional
Audio presentation Audio bandwidth(s) Audio channels, e.g., mono, stereo etc. Audio presentation level Audio presentation method, e.g., speakers, headphones (monaural, diotic binaural etc.)	Voice, Audio, AV	Mandatory
Video presentation Video image resolution(s) (Notes 1 and 2) Viewing distance as a function of height, e.g., 3H	Video, AV	Mandatory
Language	Passive Voice, AV	Mandatory
Number and gender of talkers	Passive Voice, AV	Mandatory
Type of video material, e.g., sport, head and torso	Passive Video, AV	Mandatory
Type of audio, e.g., classical music, popular music, movie soundtrack	Passive Audio, AV	Mandatory
NOTE 1 – Use of interlaced images must be noted. NOTE 2 – If the experiment contains multiple image resolutions, information must be provided as to whether smaller image resolutions were presented natively or up-scaled.		

12 Reporting objective MOS values

When reporting an MOS value produced by an ITU-T objective model, it will generally be sufficient to report the model used and any non-default settings. For non-standardized models, the information in the 'Methodology' row in Table 1 must be provided to describe the experimental design being predicted. It is also recommended that information is provided about the type of test material used in the experiments used to test and/or train the objective model.

13 Notation

[ITU-T P.800.1] provides notation that can be used to help identify the source of an MOS value. At present, it only addresses voice quality MOS values.

Bibliography

The ITU has standardized a number of subjective test methods for different applications. Some of the most widely used are listed below.

The ITU-T P.800 series includes numerous Recommendations relating to both the subjective and objective evaluation of voice quality; of particular note are:

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

[b-ITU-T P.805] Recommendation ITU-T P.805 (2007), *Subjective evaluation of conversational quality*.

[b-ITU-T P.835] Recommendation ITU-T P.835 (2003), *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm.*

The ITU-T P.900 series includes multimedia assessment Recommendations:

[b-ITU-T P.910] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications.*

[b-ITU-T P.911] Recommendation ITU-T P.911 (1998), *Subjective audiovisual quality assessment methods for multimedia applications.*

[b-ITU-T P.912] Recommendation ITU-T P.912 (2008), *Subjective video quality assessment methods for recognition tasks.*

ITU-R has also published Recommendations relating to the subjective assessment of audio and video quality:

[b-ITU-R BS.1116-1] Recommendation ITU-R BS.1116-1 (1997), *Methods for the subjective assessment of small impairments in audio systems including multichannel sounds systems.*

[b-ITU-R BS.1534-1] Recommendation ITU-R BS.1534-1 (2003), *Method for the subjective assessment of intermediate quality levels of coding systems.*

[b-ITU-R BT.500-13] Recommendation ITU-R BT.500-13 (2012), *Methodology for the subjective assessment of the quality of television pictures*.

[b-ITU-R BT.710-4] Recommendation ITU-R BT.710-4 (1998), *Subjective assessment methods for image quality in high-definition television.*

NOTE – Standardized methodologies described by the ITU-R do not all typically measure mean opinion scores. The published ITU-R Recommendations provide complete documentation and reference for all corresponding methodologies. For a better description and clarification of the testing methodologies identified above, the reader is directed to individual published Recommendations provided by the ITU-R.

The ITU-T handbook on "Practical procedures for subjective testing" provides an in-depth treatment of subjective test methods and best practices.

[b-ITU-T handbook] *Practical procedures for subjective testing* (2011).

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
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	Cable networks and 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	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Terminals and subjective and objective assessment methods**
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, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects and next-generation networks
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2013