International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.913

(01/2014)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Audiovisual quality in multimedia services

Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment

Recommendation ITU-T P.913

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
Methods for objective and subjective assessment of quality of services other than voice services	Series	P.1500

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

Recommendation ITU-T P.913

Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment

Summary

Recommendation ITU-T P.913 describes non-interactive subjective assessment methods for evaluating the one-way overall video quality, audio quality and/or audiovisual quality for applications such as Internet video and distribution quality video. These methods can be used for several different purposes including, but not limited to, comparing the quality of multiple devices, comparing the performance of a device in multiple environments, and subjective assessment where the quality impact of the device and the audiovisual material is confounded.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.913	2014-01-13	9	11.1002/1000/12106

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

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

Introduction

[ITU-T P.910], [ITU-T P.911] and [ITU-R BT.500-13] have been successfully used for many years to perform video quality and audiovisual quality subjective assessments. These Recommendations were initially designed around the paradigm of a fixed video service that transmits video over a reliable link to an immobile cathode ray tube (CRT) television located in a quiet and non-distracting environment, such as a living room or office. These Recommendations have been updated and expanded as technology shifted and they have proven to be valuable and useful for the displays and questions addressed in their original scopes.

However, the initial premise of these Recommendations does not include the new paradigms of Internet video and distribution quality television. One new paradigm of video watching is an on-demand video service transmitted over an unreliable link to a variety of mobile and immobile devices located in a distracting environment, using LCDs and other flat-screen displays. This new paradigm impacts key characteristics of the subjective test, such as the viewing environment, the listening environment, and the questions to be answered.

Users of Internet video and distribution quality television are moving from one device to another and from one environment to another throughout the day, perhaps even observing the same video using multiple devices. For example, someone might start watching a sporting event on their computer using IPTV, move to an over-the-air broadcast in their living room when the IPTV connection displays a re-buffering event, and then switch to a mobile Internet device (MID) or even a smart phone when leaving the house. Thus, subjective quality assessments into Internet video and distribution quality television ask unique questions that are not considered in the existing Recommendations. These questions may require situation-specific modifications to the subjective scale (e.g., presentation of additional information defining what "good" means in this context).

Consider the pristine viewing environment defined by [ITU-R BT.500-13], with its exact lighting conditions and non-distracting walls. The intention is to remove the impact of the viewing environment and listening environment from the experiment. For some subjective audiovisual quality experiments, this is not appropriate. First, consider an experiment that investigates the quality of service observed by video-conferencing users in an office with fluorescent lights and the steady hum of a computer. Second, consider an experiment that analyses a communications device for emergency personnel. A highly distracting background may be a critical element of the experiment design (e.g., to simulate video watched inside a moving fire truck with sirens blaring). The impact of environment is an integral part of these experiments.

These questions and environments cannot be accommodated with the existing subjective assessment Recommendations. Modifying these Recommendations would reduce the value of the intended experiments and paradigms addressed therein. The main differences in this Recommendation when compared to existing ITU subjective assessment Recommendations are:

1) inclusion of multiple testing environment options (e.g., pristine laboratory environment, simulated office within a laboratory, public environment);
2) flexibility for the user to modify the subjective scale (e.g., modified words, added information);
3) applicability for interaction effects that confound the data (e.g., evaluating a device that can only accept compressed material, impact of mobility on quality perception);
4) mandatory reporting requirement (e.g., choices made where this Recommendation allows for flexibility, experimental variables that cannot be separated due to the experiment design); and
5) inclusion of multiple display technologies (e.g., flat screen, 2D, 3D).

Recommendation ITU-T P.913

Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment

1 Scope

The devices and usage scenarios of interest herein are Internet video and distribution quality television. The focus is on the quality perceived by the end user.

This Recommendation describes methods to be used for subjective assessment of the audiovisual quality of Internet video and distribution quality. This may include assessment of visual quality only, audio quality only and/or the overall audiovisual quality. This Recommendation may be used to compare audiovisual device performance in multiple environments, and to compare the quality impact of multiple audiovisual devices. It is appropriate for subjective assessment of devices where the quality impact of the device and the material is confounded. It is appropriate for a wide variety of display technologies, including flat screen, 2D, 3D, multi-view and autostereoscopic.

1.1 Limitations

This Recommendation does not address the specialized needs of broadcasters and contribution quality television. This Recommendation is not intended to be used in the evaluation of audio-only stimuli alone, but rather audiovisual subjective assessments that may or may not include audio-only sessions. Caution should be taken when examining adaptive streaming impairments, due to the slow variations in quality within one stimulus over a long period of time.

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 J.144] Recommendation ITU-T J.144 (2004), *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference.*

[ITU-T J.340] Recommendation ITU-T J.340 (2010), *Reference algorithm for computing peak signal to noise ratio of a processed video sequence with compensation for constant spatial shifts, constant temporal shift, and constant luminance gain and offset*.

[ITU-T P.78] Recommendation ITU-T P.78 (1996), *Subjective testing method for determination of loudness ratings in accordance with Recommendation P.76*.

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

[ITU-T P.800.2] Recommendation ITU-T P.800.2 (2013), *Mean opinion score interpretation and reporting*.

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

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

[ITU-T P.1401] Recommendation ITU-T P.1401 (2012), *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*.

[ITU-T X.1244] Recommendation ITU-T X.1244 (2008), *Overall aspects of countering spam in IP-based multimedia applications.*

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

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

[ITU-R BT.1788] Recommendation ITU-R BT.1788 (2007), *Methodology for the subjective assessment of video quality in multimedia applications*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 modality [ITU-T X.1244]: In general usage, this term refers to the forms, protocols, or conditions that surround formal communications. In the context of this Recommendation, it refers to the information encoding(s) containing information perceptible for a human being. Examples of modality include textual, graphical, audio, video or haptical data used in human-computer interfaces. Multimodal information can originate from, or be targeted to, multimodal-devices. Examples of human-computer interfaces include microphones for voice (sound) input, pens for haptic input, keyboards for textual input, mice for motion input, speakers for synthesized voice output, screens for graphic/text output, vibrating devices for haptic feedback, and Braille-writing devices for people with visual disabilities.

3.1.2 subjective assessment [ITU-T J.144]: The determination of the quality or impairment of programme-like pictures presented to a panel of human assessors in viewing sessions.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 diegetic sounds: Sounds produced by objects appearing in the video, or in the film's world but off-screen.

3.2.2 dominant modality: The modality that carries the main information (i.e., audio or video).

3.2.3 double stimulus: A quality rating method where the subject is presented with two stimuli; the subject then rates both stimuli in the context of the joint presentation (e.g., a rating that compares the quality of one stimulus to the quality of the other stimulus).

3.2.4 hypothetical reference circuit: A hypothetical reference circuit (HRC) is a fixed combination of a video encoder operating at a given bit rate, network condition and video decoder. The term HRC is preferred when vendor names should not be identified.

3.2.5 least distance of distinct vision: The closest distance at which someone with normal vision (20/20 vision) can comfortably look at something. This is sometimes known as "reference seeing distance".

3.2.6 non-diegetic sounds: Sounds produced by objects outside of the film's world, such as a narrator's voice-over.

3.2.7 processed: The reference stimuli presented through a system under test.

3.2.8 processed video sequence: The processed video sequence (PVS) is the impaired version of a video sequence.

3.2.9 reference: The original version of each source stimulus. This is the highest quality version available of the audio sample, video clip or audiovisual sequence.

3.2.10 reference seeing distance: The closest distance at which someone with normal vision (20/20 vision) can comfortably look at something. This is sometimes called "least distance of distinct vision".

3.2.11 sequence: A continuous sample of audio, video or audiovisual content.

3.2.12 single stimulus: A quality rating method where the subject is presented with one stimulus and rates that stimulus in isolation (e.g., a viewer watches one video clip and then rates it).

3.2.13 source: The content material associated with one particular audio sample, video clip or audiovisual sequence (e.g., a video sequence depicting a ship floating in a harbour).

3.2.14 stimuli: Audio sequences, video sequences or audiovisual sequences.

3.2.15 subject: A person who evaluates stimuli by giving an opinion.

3.2.16 temporal forgiveness: Impairments in video material which are to some extent forgiven if poor quality video is followed by good quality video.

3.2.17 terminal: Device or group of devices used to play the stimuli during a subjective experiment (e.g., a laptop with earphones, or a Blu-ray player with an LCD monitor and speakers).

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR	Absolute Category Rating
CCR	Comparison Category Rating (also known as DSCS)
DCR	Degradation Category Rating (also known as DSIS)
DMOS	Differential Mean Opinion Score
DSCS	Double Stimulus Comparison Scale (also known as CCR)
DSIS	Double Stimulus Impairment Scale (also known as DCR)
DV	Differential Viewer scores
HRC	Hypothetical Reference Circuit
LDDV	Least Distance of Distinct Vision
LPCC	Linear Pearson Correlation Coefficient
MOS	Mean Opinion Score
MUSHRA	Multi Stimuli with Hidden Reference and Anchor points
PVS	Processed Video Sequence
RSD	Reference Seeing Distance
SAMVIQ	Subjective Assessment of Multimedia Video Quality

5 Conventions

None.

6 Source stimuli

In order to evaluate quality in various circumstances, the content should cover a wide range of stimuli. The stimuli should be selected according to the goal of the test and recorded on a digital storage system. When the experimenter is interested in comparing results from different laboratories, it is necessary to use a common set of source stimuli to eliminate a further source of variation.

The selection of the test material should be motivated by the experimental question addressed in the study. For example, the content of the test stimuli should be representative of the full variety of programmes delivered by the service under study (sport, drama, film, speech, music, etc.).

6.1 Source signals recordings

The source signal provides the reference stimuli and the input for the system under test.

The quality of the reference stimuli should be as high as possible. As a guideline, the video signal should be recorded in uncompressed multimedia files using one of the following two formats: YUV (4:2:2 or 4:4:4 sampling) or RGB (24 or 32 bits). Usually the audio signal is taken from a high quality audio production. The audio CD quality is often the reference (16 bits, 44.1 kHz) such as the sound quality assessment material (SQAM) from the European Broadcasting Union (EBU), but if possible audio masters with a minimum of 16 bits and 48 kHz are preferred.

See clause 11.5.1 for more information on compressing reference video recordings.

6.2 Video considerations

The selection of source video is an important issue. The spatial information (detail) and temporal information (motion) of the video are critical parameters. These parameters play a crucial role in determining the amount of video compression that is possible and, consequently, the level of impairment that is suffered when the scene is transmitted over a fixed-rate digital transmission service channel. Fair and relevant video test scenes must be chosen so that their spatial and temporal information is consistent with the video 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 of interest to users of the devices under test.

Post-production effects and scene cuts can cause different portions of the encoded video sequence to have different quality levels. This can confuse subjects (e.g., make the subject unsure how to rate the video). Depending upon the purpose of the experiment, it may be advisable to avoid such video sequences.

6.3 Audio considerations

When testing the overall quality of audiovisual sequences but *not* speech comprehension, the speech need not be in a language understood by all subjects.

All audio samples should be normalized for a constant volume level (e.g., normalize between clips, leaving volume variations within each clip alone).The audio source should preferably include a variety of audio characteristics (e.g., both male and female talkers, different musical instruments, different dynamic ranges). The dynamic range of an audio signal plays a crucial role in determining the impact of audio compression.

Post-production effects and scene cuts can cause different portions of the encoded audio sequence to have different quality levels. This can confuse subjects (e.g., make the subject unsure how to rate

the video). Depending upon the purpose of the experiment, it may be advisable to avoid such audio sequences.

Items have to be chosen to be realistic types of audio excerpts as much as possible, keeping in mind that they must remain as critical as possible as well (this means that transparency is not often achieved by famous encoders when encoding these audio sequences).

6.4 Audiovisual considerations

Specific care has to be taken when choosing source stimuli for audiovisual quality subjective assessments, since some degradation may have different impacts according to the relationship between audio and video. Aspects that should be considered are:

- Diegetic or non-diegetic sounds: Diegetic sounds are produced by objects appearing in the video (e.g., a person visible on the screen is talking) or in the film's world but off-screen (e.g., traffic noise, crowd noise). Non-diegetic sounds include voice-overs and background music.
- Dominant modality (audio or video). For example, in TV news sequences, the main information is carried by audio modality, whereas a sport sequence would be more characterized by a video dominant modality.

Both aspects have been shown to have an impact on audiovisual quality (see [b-Lassalle]). For example, the perception of de-synchronization between image and sound is influenced by diegetic aspects.

6.5 Duration of stimuli

The methods in this Recommendation are intended for stimuli that range from five seconds to 20 seconds in duration. Eight- to ten-second sequences are highly recommended. For longer durations, it becomes difficult for the viewers to take into account all of the quality variations and score properly in a global evaluation. The temporal forgiveness effects become important when the time duration of a stimulus is high (see [b-Hands]).

Extra source content may be required at the beginning and end of each source stimuli. For example, when creating a ten-second processed stimulus, the source might have plus two seconds of extra content before and after for a total of 14 seconds. The purpose of the extra content is to allow the audio and video coders to stabilize, and prevent the propagation of unrelated content into the processed stimuli (e.g., after the occurrence of digital transmission errors). The extra content should be discarded during editing. This technique is advised when analysing hardware coders or transmission errors.

In order to limit the duration of a test, stimuli durations of 10 seconds to one minute is preferred. This also diminishes subjects' fatigue.

6.6 Number of source stimuli

The number and type of test scenes are critical for the interpretation of the results of the subjective assessment. So, four to six scenes are enough if the variety of content is respected. The audiovisual content must have an interest in audio and video separately and conjointly.

The number of audio excerpts is very important in order to get enough data for the interpretation of the test results. A minimum of five audio items is required with respect to the range of content that can be encountered in "real life" (that is to say when using the systems under test).

The number of five items is also a good compromise in order to limit the duration of the test.

7 Test methods, rating scales and allowed changes

This clause describes the test methods, rating scales and allowable deviations. The method controls the stimuli presentation. The rating scale controls the way that people indicate their opinion of the stimuli. A list of appropriate changes to the method follows.

In-force and superseded versions of [ITU-T P.800], [ITU-T P.910] and [ITU-R BT.500-13] include alternate names for some test methods described in this clause. These alternate names will be identified and may be used.

7.1 List of methods

This clause contains a listing of appropriate subjective test methods and rating scales.

7.1.1 Absolute category rating (ACR) method

The absolute category rating (ACR) method is a category judgment where the test stimuli are presented one at a time and are rated independently on a category scale. ACR is a single stimulus method. The subject observes one stimulus and then has time to rate that stimulus.

The ACR method uses the following five-level rating scale:

5 Excellent
4 Good
3 Fair
2 Poor
1 Bad

The numbers may optionally be displayed on the scale.

Comments

The ACR method produces a high number of ratings in a brief period of time.

ACR ratings confound the impact of the impairment with the influence of the content upon the subject (e.g., whether the subject likes or dislikes the production quality of the stimulus).

7.1.2 Degradation category rating (DCR) method; also known as the double stimulus impairment scale (DSIS) method

The degradation category rating (DCR) method presents stimuli in pairs. The first stimulus presented in each pair is always the reference. The second stimulus is that reference stimulus after being processed by the systems under test. DCR is a double stimulus method. The DCR method is also known as the double stimulus impairment scale (DSIS) 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:

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

The numbers may optionally be displayed on the scale.

Comments

The DCR method produces fewer ratings than ACR in the same period of time (e.g., slightly more than one-half).

DCR ratings are minimally influenced by a subject's opinion of the content (e.g., whether the subject likes or dislikes the production quality). Thus, DCR is able to detect colour impairments and skipping errors that the ACR method may miss.

DCR ratings may contain a slight bias. This occurs because the reference always appears first, and people know that the first stimulus is the reference.

7.1.3 Comparison category rating (CCR) method; also known as the double stimulus comparison scale (DSCS); also known as pair comparison (PC)

The comparison category rating (CCR) method is a method where the test stimuli are presented in pairs. Two versions of the same stimulus are presented in a randomized order (e.g., reference shown first 50% of the time and second 50% of the time). CCR is a double stimulus method. CCR may be used to compare reference video with processed video, or to compare two different impairments. The CCR method is also known as the double stimulus comparison scale (DSCS) method. The CCR method is also known as the pair comparison (PC) method.

The subjects are asked to rate the impairment of the second stimulus in relation to the first stimulus. The following seven-level scale for rating the impairment should be used:

–3	Much worse
–2	Worse
–1	Slightly worse
0	The same
1	Slightly better
2	Better
3	Much better

The numbers may optionally be displayed on the scale.

During data analysis, the randomized order of presentation must be removed.

Comments

The CCR method produces fewer ratings than the ACR method in the same period of time (e.g., slightly more than one-half).

CCR ratings are minimally influenced by a subject's opinion of the content (e.g., whether the subject likes or dislikes the production quality).

Subjects will occasionally mistakenly swap their ratings when using the CCR scale (e.g., mark "Much Better" when intending to mark "Much Worse"). This is unavoidable due to human error. These unintentional score swapping events will introduce a type of error into the subjective data that is not present in ACR and DCR data.

The accuracy of CCR is influenced by the randomized presentation of stimuli one and two. For example, when comparing reference and processed video, if the reference stimulus is presented first 90% of the time, then CCR will contain the same bias seen in the DCR method.

7.1.4 SAMVIQ description for video and audiovisual tests

Subjective assessment of multimedia video quality (SAMVIQ) may be used for video-only or audiovisual test. Where the description below is unclear or ambiguous, see [ITU-R BT.1788]. Where discrepancies exist between the description below and [ITU-R BT.1788], the instructions in this clause are recommended.

SAMVIQ overview

The SAMVIQ method was designed to assess the video that spans a large range of resolutions (i.e., SQCIF to HDTV). The SAMVIQ is a non-interactive subjective assessment method for evaluating the video quality of multimedia applications. This method can be applied for different purposes,

including but not limited to a selection of algorithms, ranking of audiovisual system performance, and evaluation of the video quality level during an audiovisual connection.

SAMVIQ scale

The SAMVIQ methodology uses a continuous quality scale. Each subject moves a slider on a continuous scale graded from zero to 100. This continuous scale is annotated by five quality items linearly arranged (excellent, good, fair, poor, bad).

7.1.5 MUSHRA description for audio tests

The multi-stimuli with hidden reference and anchor points (MUSHRA) method may be used for audio-only tests. Where the description below is unclear or ambiguous, see [ITU-R BS.1534-1]. Where discrepancies exist between the description below and [ITU-R BS.1534-1], the instructions in this clause are recommended.

MUSHRA overview

MUSHRA is a method dedicated to the assessment of intermediate quality. MUSHRA can be used either for monophonic or stereophonic audio excerpts. MUSHRA tends to be used also for 5.1 and binaural audio items. This methodology is run either on headphones or loudspeakers.

MUSHRA can be used to rank the performance of audio systems or evaluate their basic audio quality. MUSHRA can be used for broadcasting applications dedicated to streaming and transmission.

An important feature of this method is the inclusion of the hidden reference and bandwidth limited anchor signals. The chosen anchor points were the band-limited signal with cut-off frequencies of 3.5 kHz (mandatory) and 7 kHz.

The MUSHRA listening panel

The listening panel consists of experts in their subjects, most of whom are experienced users of audio devices but not professionally involved.

The MUSHRA scale

MUSHRA uses a continuous quality scale. Each subject moves a slider along the graded scale from zero to 100 linearly annotated by five quality items (Excellent 100-80, Good 80-60, Fair 60-40, Poor 40-20, Bad 20-0).

Test instructions

The test instructions explain to the subjects how the MUSHRA software works, what they will listen to (briefly), how to use the quality scale and how to score the different excerpts. This is also an opportunity to mention the fact that there is a hidden reference signal to score and consequently, there should be at least one score equal to 100 per excerpt. This will be used later on in the process of rejecting subjects.

Comments

MUSHRA is sensitive to modifications in methods and environment. The subjective ratings may change significantly depending upon whether the experiment is conducted in a controlled environment (as per clause 8.1) or a public environment (as per clause 8.2). The subjective ratings may also change significantly if the MUSHRA method is modified (see clause 7.2). MUSHRA ratings gathered in accordance with this Recommendation should not be directly compared to MUSHRA ratings that are fully compliant with [ITU-R BS.1534-1].

7.2 Acceptable changes to the methods

This clause of the Recommendation is intended to be a living document. The methods and techniques described in this clause cannot, by their very nature, account for the needs of every subjective experiment. It is expected that the experimenter may need to modify the test method to suit a particular experiment. Such modifications fall within the scope of this Recommendation.

The following acceptable changes have been evaluated systematically. Subjective tests that use these modifications are known to produce repeatable results.

7.2.1 Changes to level labels

Translating labels into different languages does not result in a significant change to the MOS. Although the perceptual magnitude of the labels may change, the resulting MOS are not impacted.

An unlabelled scale may be used. For example, ends of the scale can be labelled with the symbols "+" and "–".

A scale with numbers but no words may be used.

Numbers may be included or excluded at the preference of the experimenter.

Alternate wording of the labels may be used when the rating labels do not meet the needs of the experimenter. One example is using the DCR method with the ACR labels. Another example is using the ACR method with a listening-effort scale as mentioned in [ITU-T P.800].

7.2.2 ACR with hidden reference (ACR-HR)

An acceptable variant of the ACR method is ACR with hidden reference (ACR-HR). With ACR-HR, the experiment includes a reference version of each video segment, not as part of a pair, but as a freestanding stimulus for rating like any other. During the data analysis the ACR scores will be subtracted from the corresponding reference scores to obtain a DMOS. This procedure is known as "hidden reference removal."

Differential viewer scores (DV) are calculated on a per subject per processed video sequence (PVS) basis. The appropriate hidden reference (REF) is used to calculate DV using the following formula:

$$DV(PVS) = V(PVS) – V(REF) + 5$$

where V is the viewer's ACR score. In using this formula, a DV of five indicates "Excellent" quality and a DV of one indicates "Bad" quality. Any DV values greater than five (i.e., where the processed sequence is rated better quality than its associated hidden reference sequence) will generally be considered valid. Alternatively, a two-point crushing function may be applied to prevent these individual ACR-HR viewer scores (DV) from unduly influencing the overall mean opinion score:

$$crushed\_DV = (7*DV)/(2+DV) \text{ when } DV > 5.$$

Comments

ACR-HR will result in larger confidence intervals than ACR, CCR or DCR.

The ACR-HR method removes some of the influence of content from the ACR ratings, but to a lesser extent than CCR or DCR.

ACR-HR should not be used when the reference sequences are fair, poor or bad quality. The problem is that the range of DV diminishes. For example, if the reference video quality is poor on the ACR scale, then DV must be three or greater.

7.3 Discouraged but acceptable changes to the methods

The following acceptable changes have been evaluated systematically. These modifications are discouraged. However, these changes are allowed.

7.3.1 Increasing the number of levels is discouraged

The clause that defines each method identifies the recommended number of levels for that method (e.g., in clause 7.1, a discrete five-level scale is recommended for ACR).

The use of an increased number of levels is allowed yet discouraged. Examples include changing ACR from a discrete five-level scale to a discrete nine-level scale, a discrete eleven-level scale or a continuous scale. This modification is allowed in [ITU-T P.910].

Comments

Tests into the replicability and accuracy of subjective methods indicate that the accuracy of the resulting MOS does not improve. However, the method becomes more difficult for subjects.

Currently published experiments that compare discrete scales (e.g., five-level, nine-level, 11-level) with continuous scales (e.g., 100-level scales) all indicate that continuous scales contain more levels than can be differentiated by people. The continuous scales are treated by the subjects like discrete scales with fewer options (e.g., using five to nine levels). For example, see [b-Huynh-Thu] and [b-Tominaga].

8 Environment

For subjective experiments that fall into the scope of this Recommendation, most aspects of the environment will have minimal impact on MOS. Thus, the environment is not rigorously constrained within this Recommendation. Exceptions include cases where the experiment is designed to investigate the impact of a particular part of the environment on MOS (e.g., the impact of video monitor type on MOS).

This Recommendation allows two options for the environment in which the subjective experiment takes place:

- a controlled environment
- a public environment.

The number of subjects required is impacted by this choice (see clause 9). The environment must be described (see clause 8.4).

Small mobile devices may be either held by the subject or mounted on a stand. The use of a stand will yield a more consistent viewing angle and viewing distance.

8.1 Controlled environment

A controlled environment is a room devoted to conducting the experiment. The room should be comfortable and quiet. People not involved in the experiment should not be present. Examples include a sound isolation chamber, a laboratory, a simulated living room, a conference room, or an office set aside temporarily for the subjective experiment. A controlled environment should represent a non-distracting environment where a person would reasonably use the device under test.

8.2 Public environment

A public environment is any environment where people not involved in the experiment are present. A public environment also includes subjective tests performed in a room where some element of the environment intentionally serves as a distraction from the experiment (e.g., loud background noise). A public environment should represent a distracting environment where a person would reasonably use the device under test.

8.3 Viewing distance

It is important here to differentiate between fixed displays (e.g., TV, monitor, video projector) and mobile displays (e.g., smartphone or tablet). Indeed, for fixed displays, the visualisation distance will not change during the test and is determined by the visual angle perceived, which is described as a minute of an arc (e.g., 3H for HD1080 display). On the other hand, for mobile displays, the subject will adjust the visualisation distance according to the subject's preference, the screen size and the content quality. Thus, for practical purposes in everyday life, the subjects are not constrained while watching content on their mobile device, whereas they are when watching TV or other fixed displays.

The minimum viewing distance should be in accordance with the least distance of distinct vision (LDDV) or the reference seeing distance (RSD).

8.4 Documentation of environment

The subjective test's environment must be reported. The documentation of the experiment must include the following information. Some information only applies for audio and audiovisual subjective tests; while others apply only to video and audiovisual subjective tests.

The luminosity must be measured (e.g., in Lux). The location and direction of the lighting measurement should be identified (e.g., horizontal to the screen and pointing outwards, or at the eye position in the direction of the screen).

If a public environment is changing to a large extent, then a full description may not be possible. For example, if a mobile device is given to each subject to take home with them, or the subject runs the experiment interface on their own mobile phone.

Information	Type of stimuli
A picture of the subjective test environment	All
Lighting level (e.g., dim, bright, light level measured in Lux)	Video, audiovisual
Noise level (e.g., quiet, bystanders talking)	All
Approximate viewing distance in picture heights	Video, audiovisual
Whether a controlled or public environment was used	All
Type of video monitor	Video, audiovisual
Size of video monitor	Video, audiovisual
Type of audio system	Audio, audiovisual
Placement of audio speakers (if used)	Audio, audiovisual

9 Subjects

The number of subjects used in the experiment is extremely important.

At least 24 subjects must be used for experiments conducted in a controlled environment. This means that after subject screening, every stimulus must be rated by at least 24 subjects.

At least 35 subjects must be used for experiments conducted in a public environment.

Fewer subjects may be used for pilot studies, to indicate trending. Such studies must be clearly labelled as being pilot studies.

For SAMVIQ and MUSHRA tests conducted in a controlled environment, the number of subjects that remain after the rejection process should not be less than 15 in order to have significant data for statistical analysis.

10 Experiment design

10.1 Size of the experiment and subject fatigue

The size of an experiment is typically a compromise between the conditions of interest and the amount of time individual subjects can be expected to observe and rate stimuli.

Preferably, an experiment should be designed so that each subject's participation is limited to 1.5 hours, of which no more than one hour is spent rating stimuli. When larger experiments are required (e.g., three hours spent rating stimuli), frequent breaks and adequate compensation should be used to counteract the negative impacts of fatigue and boredom.

The number of times that each source stimulus is repeated also impacts subject fatigue. Among different possible test designs, preferably choose the one that minimizes the number of times a given source stimulus is shown.

10.2 Special considerations for transmission error, rebuffering, and audiovisual synchronization impairments

When stimuli with intermittent impairments are included in an experiment, care must be taken to ensure that the impairment can be perceived within the artificial context of the subjective quality experiment. The first one second and the last one second of each stimulus should not contain freezing, rebuffering events, and other intermittent impairments. When stimuli include audiovisual synchronization errors, some or all of the audiovisual source sequences must contain audiovisual synchronization clues (e.g., lip synch, cymbals, doorbell pressed).

Examples of intermittent impairments include but are not limited to:

- pause then play resumes with no loss of content (e.g., pause for rebuffering);
- pause followed by a skip forward in time (e.g., transmission error causes temporary loss of signal, and system maintains a constant delay);
- skip forward in time (e.g., buffer overflow);
- audiovisual synchronization errors (e.g., may only be perceptible within a small portion of the stimuli)
- packet loss with brief impact.

These impairments might be masked (i.e., not perceived) due to the scene cut when the scene starts or ends. A larger context may be needed to perceive the impairment as objectionable (i.e., audiovisual synchronization errors are increasingly obvious during a longer stimulus). For video-only experiments, the missing audio might mask the impairment, and vice versa. For example, with video-only stimuli, an impairment that produces a skip forward in time might be visually indistinguishable from a scene cut. By contrast, the audio in an audiovisual sequence would probably give the observers clues that an undesirable event has occurred.

10.3 Special considerations for longer stalling events

From prior research, it is known that longer stalling events (e.g., 5 seconds) are perceived differently from shorter stalling events (e.g., 0.5 seconds). In addition to the interruption of the flow, which happens in both cases, longer stalling events may be perceived in terms of waiting time and the need to wait for a service. This may have implications for the instructions given to subjects, which will be addressed in clause 11.

Specific care should be taken in the design of subjective tests that explore the impact of longer stalling events. For example, large confidence intervals may result if some subjects perceive the stalling event as a drop in quality, and other subjects attribute the stalling event to a normal service problem.

11 Experiment implementation

Each subject's participation in an experiment typically consists of the following five stages:

1) Informed consent
2) Pre-screening of subjects
3) Instructions and training
4) Voting session(s)
5) Questionnaire and/or interview (optional)

These steps are described in further detail in this clause.

11.1 Informed consent

Subjects should be informed of their rights and be given basic information about the experiment. It may be appropriate for subjects to sign an informed consent form. In some countries, this is a legal requirement for human testing. Typical information that should be included on the release form is as follows:

- organization conducting the experiment;
- goal of the experiment, summarized briefly;
- task to be performed, summarized generally;
- whether the subject may experience any risks or discomfort from their participation;
- names of all Recommendations that the experiment conforms to;
- duration of the subject's involvement;
- range of dates when this subjective experiment will be conducted;
- number of subjects involved;
- assurance that the identity of subjects will be kept private (e.g., subjects are identified by a number assigned at the beginning of the experiment);
- assurance that their participation is voluntary, and that the subject may refuse or discontinue participation at any time without penalty or explanation;
- name of the person to contact in the event of a research-related injury;
- who to contact for more information about the experiment.

An example informed consent form is presented in Appendix I.

11.2 Optional pre-screening of subjects

Pre-screening procedures include methods such as vision tests, audiometric tests, and selection of subjects based on their previous experience. Prior to a session, the subjects may be screened for normal visual acuity or corrected-to-normal acuity, for normal colour vision, and for good hearing.

Concerning acuity, no errors on the 20/30 line of a standard eye chart (Snellen eye chart) should be made. The chart should be scaled for the test viewing distance and the acuity test performed at the same location where the video images will be viewed (i.e., lean the eye chart up against the monitor) and have the subjects seated. For example, a near vision chart is appropriate for experiments that use laptops and small mobile devices.

A screening test may be performed, as appropriate for the experiment. Examples include:

- Concerning vision test plates (red /green deficiency), no more than two plates (*Pseudo Isochromatic Plates (1940), engraved and printed by The Beck Engraving Co., Inc., Philadelphia and New York, United States*) should be missed out of 12.
- Evaluate with triton colour vision test plates (blue / yellow deficiency).

- Test whether subjects are able to correctly identify colours.
- Contrast test (e.g., Mars Perceptix contrast test, ETDRS Format, Continuous Test).
- Concerning hearing, no subject should exceed a hearing loss of 15 dB at all frequencies up to and including 4 kHz and more than 25 dB at 8 kHz. (Note: hearing specifications were taken from [ITU-T P.78] Annex B.1).
- Stereo-acuity test, with a tentative threshold of 140 seconds.

Subjects who fail such screening should preferably be run through the experiment with no indication given that they failed the test. The data from such subjects should be discarded when a small number of subjects are used in the experiment. Data from such subjects may be retained when a large number of subjects are used (e.g., 30 or more).

11.3 Post-screening of subjects

Post-screening of subjects may or may not be appropriate depending upon the purpose of the experiment. The following subject screening methods are appropriate: [ITU-R BT.500-13] clause 2.3, [ITU-R BT.1788] Annex 2 clause 3, Annex A, and questionnaires and/or interviews after the experiment to determine whether or not the subject understood the task. Subject screening for crowdsourcing may require unique solutions (e.g., clever test preparation).

When subjects are eliminated due to post-screening, it may be appropriate to separately present the data of the screened subjects or to analyse the data both with and without the screened subjects.

The final report should include a detailed description of the screening methodology.

11.4 Instructions and training

Usually, subjects have a period of training in order to get familiar with the test methodology and software and with the kind of quality they have to assess.

The training phase is a crucial part of this method, since subjects could misunderstand their task. Written or recorded instructions should be used to be sure that all subjects receive exactly the same information. The instructions should include explanations about what the subjects are going to see or hear, what they have to evaluate (e.g., difference in quality), and how to express their opinion. The instructions should include reassurance that there is no right or wrong answer in this experiment; we are simply interested in the subject's opinion. An example set of instructions is given in Appendix II.

Questions about the procedure and meaning of the instructions should be answered with care to avoid bias. Questions about the experiment and its goals should be answered after the final session.

After the instructions, a training session should be run. The training session is typically identical to the experiment sessions, yet short in duration. Stimuli in the training session should demonstrate the range and type of impairments to be assessed. Training should be performed using stimuli that do not otherwise appear in the experiment.

The purpose of the training session is to (1) familiarize the subjects with the voting procedure and pace, (2) show the subjects the full range of impairments present, thus stabilizing their votes, (3) encourage subjects to ask new questions about their task, in the context of the actual experiment, and (4) adjust the audio playback level, which will then remain constant during the test phase. For a simple assessment of video quality in absolute terms, a small number of stimuli in the training session may suffice (e.g., three to five stimuli). For more complicated tasks, the training session may need to contain a large number of stimuli.

If 3D content is evaluated, the instructions must tell subjects what to do when 3D fatigue is experienced.

The subject should be carefully introduced to the method of assessment, the types of impairment or quality factors likely to occur, the grading scale, timing, etc. Training stimuli should demonstrate the range and the type of impairments to be assessed. The training stimuli should not otherwise appear in the test, but should have comparable sensitivity.

The subject should not be told the type of impairments and impairment locations that will appear in the test.

Subjects should be given instructions regarding any ambiguous issues. That is, an issue that may or may not be perceived as a quality event; and thus it either should or should not impact the subject's quality rating. Without such instruction, different subjects may respond differently to this issue. One example is a long stalling event (see clause 10.3), which can be misinterpreted as a normal service problem or an unintended flaw in the media playback system. A second example is the aesthetic quality of the stimuli. Subjects are typically asked to ignore the stimuli content (e.g., aesthetics, subject matter). See Appendix II for sample training instructions that include the second example.

11.5 Experiment sessions and breaks

Ideally no session should last for more than 20 minutes, and in no case should a session exceed 45 minutes. Every 20 minutes, subjects should be asked to rest a bit by taking in some fresh air.

The stimuli should be presented in a pseudo-random sequence.

The pattern within each session (and the training session) is as follows: play sequence, pause to score, repeat. The subject should typically be shown a grey screen between video sequences. The subject should typically hear silence or instructions between video sequences (e.g., "here is clip one", "please score clip one"). The specific pattern and timing of the experiment sessions depends upon the playback mechanism.

11.5.1 Computer playback and compressed playback

Computerized control of the content playback is only allowed when the playback hardware and software can reliably play the content identically for all subjects. The playback mechanism must not introduce any impairment that is present for some but not all subjects (e.g., dropped frames, pause in playback, pause in the audio).

The ideal computerized playback introduces no further impairments (e.g., audiovisual file is stored uncompressed and is presented identically to all subjects without pauses, hesitation, or dropped frames). See clause 6.1 for information on uncompressed sampling formats.

If the terminal is not capable of playing the uncompressed video as described, then the video can be encoded with a codec that is compatible with the terminal. If no lossless codecs are supported by the terminal, the video must then be encoded using a lossy codec or played as created. Two categories of codecs must be distinguished:

- Lossless. A lossless codec exactly reproduces the uncompressed video. This is preferred whenever it is possible, but the terminal must be able to decode and play back the video in real time. The codec's performance should be tested using the PSNR measurement (see [ITU-T J.340]).
- Lossy. All videos will be identically recompressed using an excellent quality but lossy compression (i.e., for the purposes of computerized playback). The encoded reference video should be considered to be excellent, if expert viewers cannot detect artefacts when the reference video is displayed on the terminal. This expert analysis should be performed before launching the test sessions.
- Not recompressed. In some situations, the compressed stimuli should not be recompressed for experiment playback (e.g., when crowdsourcing, to ensure smooth playback on multiple systems).

The type of computerized playback should be identified in the report.

Any impairment introduced by the playback mechanism that cannot be detected by the subjects may be ignored but must be disclosed in the experiment summary. Preferably, all stimuli should be recompressed identically for playback (e.g., stimuli are lightly compressed to ensure correct playback).

Some computerized playback platforms will introduce impairments that can be detected by the subject, in addition to the impairments intended to be tested (e.g., stimuli are moderately compressed to ensure playback on a mobile device). These impairments will compound the data being measured and must be considered during the data analysis. Such an experiment design should be avoided unless no alternative exists.

If the compressed video quality appears to be different from the uncompressed reference's quality, then a transparency test is recommended. That is, a subjective pre-test that includes uncompressed playback of the reference and compressed playback of the reference as it will be used in the target experiment. This may not always be possible (e.g., some devices do not support uncompressed playback; or uncompressed playback capability is not available). Test stimuli must be created using the uncompressed reference (i.e., not the compressed reference used in such experiments).

11.5.2 Self-paced sessions

Computerized control of content playback usually allows the sessions to be self-paced. With computerized control, it is best to present the subject with silence and a blank screen (typically 50% grey) when transitioning from the scoring mechanism to a scene, and from one scene to the next. The pattern and timing of a single stimulus experiment is typically as follows:

- silence with blank screen for 0.7 to 1.0 s (optional)
- play stimulus
- silence with blank screen for 0.7 to 1.0 s (optional)
- graphical user interface displays scoring option, with a button to be selected after scoring.

The pattern and timing of a double stimulus experiment is typically as follows:

- silence with blank screen for 0.7 to 1.0 s (optional)
- play first stimulus
- silence with blank screen for 1.0 to 1.5 s
- play second stimulus
- silence with blank screen for 0.7 to 1.0 s (optional)
- graphical user interface displays scoring option, with a button to be selected after scoring.

The blank screen with silence serves to separate each stimulus from the visual impact of the computerized user interface.

The experimenter should choose whether or not repeated playback is allowed.

Care should be taken with the background display. If no other considerations are present, a plain grey background is recommended (50% grey), with perhaps a thin border of black surrounding the video. Where possible, icons, operating system menus, and other programs should not be visible. These serve as a distraction and may invite the subject to explore other data on the test computer.

11.5.3 Fixed paced sessions

Some playback mechanisms require a fixed pace of the session. Examples of fixed pace sessions are video tape, DVDs, Blu-ray discs, or a long video file containing one session. When an encoded playback mechanism is to be used, choose the highest possible bit rate that ensures reliable playback (see clause 11.4.1).

The timing of fixed paced sessions must be carefully chosen to allow sufficient time for voting. The pattern and timing of a single stimulus experiment is typically as follows:

- play stimuli
- 10 s for voting
- repeat.

The pattern and timing of a double stimulus experiment is typically as follows:

- play first stimuli
- silence and 50% grey for 1.0 to 1.5 s
- play second stimuli
- 10 s for voting
- repeat.

The time for voting should be adjusted to allow sufficient time for voting. Time for voting may be adjusted to help avoid editing mistakes (e.g., placing the beginning of the first stimulus at a predictable minute/second boundary). During voting, spoken or written instructions should appear (e.g., "Here is clip one", "Please score clip one"). This will help the subject keep the proper pace in the experiment (i.e., indicate the proper stimulus number when recording their vote). Preferably, the first and last 0.7 to 1.0 s of the voting time should be 50% grey with silence. This will provide the subjects with a visual and audible separation between the stimuli and the instructions.

11.5.4 Stimuli randomization

Preferably, the stimuli should be randomized differently for each subject. This is typically possible for self-paced sessions. For fixed paced sessions, a randomized sequence for each subject is usually not practical.

A minimum of two tape orderings must be used. Three tape orderings is preferred. This reduces the impact of ordering effects. To create one ordering, the stimuli are randomly divided into sessions and the stimuli within each session are randomly ordered. The sessions themselves must be randomly presented to the subjects.

For example, consider an experiment with three randomized orderings (Red, Green and Blue), each having two session (A and B). 1/6 of subjects would rate Red-A then Red-B; 1/6 of subjects would rate Red-B then Red-A; 1/6 of subjects would rate Green-A then Green-B; etc.

When a small number of randomizations are used, randomization must be constrained so that:

- the same source stimulus does not occur twice in a row
- the same impairment does not occur twice in a row.

These constraints become less important when each subject has a unique ordering.

11.5.5 Types of stimuli in each session

Some experiments that conform to this Recommendation will use only one type of stimuli (e.g., all stimuli contain audiovisual content, all stimuli contain video-only content). Other experiments will use multiple types of stimuli (e.g., audio-only, video-only, and audiovisual stimuli will be rated).

Different types of stimuli may either be split into separate sessions or mixed together into a single session.

11.6 Voting

Each session may ask a single question (e.g., what is the video quality) or multiple questions (e.g., what is the video quality, what is the audio quality).

Voting may be recorded with paper ballots or software.

Paper ballots usually list multiple stimuli on a single sheet of paper. One example ballot for the ACR method is shown in Figure 1. One disadvantage to paper ballots is that the subject can get "off" in time (e.g., observe stimulus 6 and then score stimulus 7).

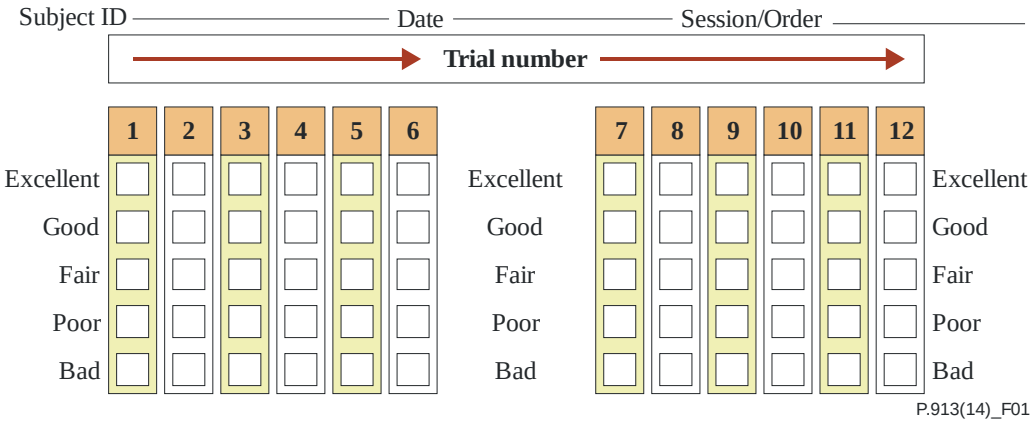

Figure 1 – Example paper ballot for the ACR method showing 12 stimuli

Electronic voting accomplishes the same data entry and has the advantage of automated data entry. An example computer screenshot is shown in Figure 2.

Figure 2 – Example screenshot of electronic voting for the ACR method

11.7 Questionnaires and interviews

For some experiments, questionnaires or interviews may be desirable either before or after the subjective sessions. The goal of the questionnaire or interview is to supplement the information gained by the experiment. Examples include:

- demographics that may or may not influence the votes, such as age, gender and television watching habits;
- feedback from the subject after the sessions; and
- quality experience observations on deployed equipment used by the subject (i.e., service observations).

The disadvantage of the service observation method for many purposes is that little control is possible over the detailed characteristics of the system being tested. However, this method does afford a global appreciation of how the "equipment" performs in the real environment.

12 Data analysis

12.1 Documenting the experiment

Clause 12 of [ITU-T P.800.2] describes the minimum information that should accompany MOS values to enable them to be correctly interpreted.

12.2 Calculate MOS or DMOS

After all subjects are run through an experiment, the ratings for each clip are averaged to compute either a mean opinion score (MOS) or a differential mean opinion score (DMOS).

Use of the term "MOS" indicates that the subject rated a stimulus in isolation. The following methods can produce MOS scores:

- ACR
- ACR-HR (using raw ACR scores)
- SAMVIQ
- MUSHRA

Use of the term "DMOS" indicates that scores measure a change in quality between two versions of the same stimulus (e.g., the source video and a processed version of video). The following methods can produce DMOS scores:

- ACR-HR (average DV, defined in clause 7.2.2)
- DCR / DSIS

  CCR / DSCS

When CCR is used, the order randomization should be removed prior to calculating a DMOS. For example, for subjects who saw the original video second, multiply the opinion score by negative one. This will put the CCR data on a scale from zero ("the same") to three, with negative scores indicating the processed video was higher quality than the original.

[ITU-T P.800.2] provides additional information about mean opinion scores.

12.3 Evaluating objective metrics

When a subjective test is used to evaluate the performance of an objective metric, then [ITU-T P.1401] can be used. [ITU-T P.1401] presents a framework for the statistical evaluation of objective quality algorithms regardless of the assessed media type.

Annex A

Method for post-experiment screening of subjects using Pearson's linear correlation

(This annex forms an integral part of this Recommendation.)

The rejection criterion verifies the level of consistency of the raw scores of one subject according to the corresponding average raw scores over all subjects. A decision is made using a correlation coefficient.

Linear Pearson correlation coefficient (LPCC) for one subject versus all subjects is calculated as:

$$LPCC(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} {x_i}^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} {y_i}^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}} \quad \text{(A1)}$$

Where *x* and *y* are arrays of data and *n* is the number of data points.

To calculate LPCC on individual stimuli (i.e., per PVS), compute

$$r1(x,y) = LPCC(x,y) \quad \text{(A2)}$$

where

x_i: MOS of all subjects per PVS

y_i: individual score of one subject for the corresponding PVS

n: number of PVSs

i: PVS index

To calculate LPCC on systems (i.e., per HRC), compute

$$r2(x,y) = LPCC(x,y) \quad \text{(A3)}$$

where

x_i: condition MOS of all subjects per HRC (i.e., condition MOS is the average value across all PVSs from the same HRC)

y_i: individual condition MOS of one subject for the corresponding HRC

n: number of HRCs

i: HRC index

One of the following two rejection criteria may be used.

A.1 Screen by PVS

Screening analysis is performed per PVS only, using Equation (A2). Subjects are rejected if $r1$ falls below a set threshold. A discard threshold of ($r1 < 0.75$) is recommended for ACR and ACR-HR tests of entertainment video. Subjects should be discarded one at a time, beginning with the worst outlier (i.e., lowest $r1$) and then recalculating $r1$ for each subject.

Different thresholds may be needed depending upon the method, technology or application.

A.2 Screen by PVS and HRC

Screening analysis is performed per PVS and per HRC, using Equations (A2) and (A3). Subjects are rejected if $r1$ or $r2$ fall below set thresholds. For ACR and ACR-HR tests of entertainment video, a subject should be discarded if ($r1 < 0.75$ and $r2 < 0.8$). Both $r1$ and $r2$ must fall below separate thresholds before a subject is discarded. Subjects should be discarded one at a time, beginning with the worst outlier (i.e., by averaging the amount that the two thresholds are exceeded) and then recalculating $r1$ and $r2$.

Different thresholds may be needed depending upon the method, technology or application.

The reason for using analysis per HRC using $r2$ is that a subject can have an individual content preference that is different from other subjects. This preference will cause $r1$ to decrease, although this subject may have voted consistently. Analysis per HRC averages out an individual's content preference and checks consistency across error conditions.

Appendix I

Informed consent form: example

(This appendix does not form an integral part of this Recommendation.)

The following is an example of an informed consent form. The **underlined words in bold** are intended to be replaced with the appropriate values (e.g., a person's name, phone number, organization name, ITU Recommendation names).

Users should investigate local regulations and requirements for informed consent notification, and make the necessary changes.

Video quality experiment

Informed Consent Form

Principal Investigator: **Name**, **Phone Number**

Organization is conducting a subjective audio-video quality experiment. The results of this experiment will assist us in evaluating the impact of several different factors on audiovisual quality.

You have been selected to be part of a pool of viewers who are each a potential participant in this subjective audiovisual quality experiment. In this experiment, we ask you to evaluate the audiovisual quality of a set of video scenes. You will sit on a comfortable chair in a quiet, air-conditioned room, watch video sequences on a laptop and listen to audio from earphones. You will specify your opinion of the current quality by selecting buttons on the screen. The participants in this video quality experiment are not expected to experience any risk or discomfort. This experiment conforms to Recommendation ITU-T P.913.

You will be asked to participate in up to **five** viewing sessions. Before the first session, you will listen to instructions for **four** minutes and participate in a **two** minute practice session. During each session, you will rate audiovisual sequences for **twenty** minutes. There will be a break after the practice session to allow you to ask questions, and another break after each session. In all, the time required to participate in this experiment is estimated to be **less than two and a half hours**. Of this time, approximately **two hours** will be spent rating audiovisual quality.

This experiment will take place during **range of dates** and will involve no more than **number** viewers. The identities of the viewers will be kept confidential. Your quality ratings will be identified by a number assigned at the beginning of the experiment.

Participation in this experiment is entirely voluntary. Refusal to participate will involve no penalty, and you may discontinue participation at any time. If you have any questions about research subjects' rights, or in the event of a research-related injury to the subject, please contact **Name** at **Phone Number**.

If you have any questions about this experiment or our audiovisual quality research, please contact **Name** at **Phone Number** or email address **Email Address**.

Please sign below to indicate that you have read the above information and consent to participate in this audiovisual quality experiment.

Signature _____

Appendix II

Example instructions

(This appendix does not form an integral part of this Recommendation.)

The following example instructions cover a two-session experiment rating audiovisual sequences on the ACR scale in a sound isolation booth. However, an experiment could be done in one session or could require more than two sessions. Other modifications may be required.

"Thank you for coming in to participate in our study. The purpose of this study is to gather individual perceptions of the quality of several short multimedia files. This will help us to evaluate various transmission systems for those files.

In this experiment you will be presented with a series of short clips. Each time a clips is played, you will be asked to judge the quality of the clip. A ratings screen will appear on the screen and you should use the mouse to select the rating that best describes your opinion of the clip. After you have clicked on one of the options, click on the "Rate" button to automatically record your response to the hard drive.

Observe and listen carefully to the entire clip before making your judgement. Keep in mind that you are rating the combined quality of the audio and video of the clip rather than the content of the clip. If, for example, the subject of the clip is pretty or boring or annoying, please do not take this into consideration when evaluating the overall quality of the clip. Simply ask yourself what you would think about the quality of the clip if you saw this clip on a television or computer screen.

And don't worry about somehow giving the wrong answer; there is no right or wrong answer. Everyone's opinion will be slightly different. We simply want to record your opinion. We will start with a few practice clips while I am standing here. After that, the experiment will be computer controlled and will be presented in five blocks of about 20 minutes each.

After the first block is finished, the computer will tell you that the section is finished. You should stand up and push open the door and come out of the chamber and take a break. By the way, the door will never be latched or locked. The door is held closed with magnets; much like modern refrigerators [demonstrate the pressure needed to push open the door]. If you have claustrophobia or need to take an unscheduled break, feel free to open the door and step outside for a moment.

During the break between sessions, there will be some light refreshments for you. When you are ready, we will begin the second session. Do you have any questions before we begin?"

Bibliography

[b-Hands] Hands, D.S. (2001), *Temporal characterisation of forgiveness effect*, Electronics Letters, Vol. 37 No. 12, pp. 752-754.

[b-Huynh-Thu] Huynh-Thu, Q., Garcia, M.-N., Speranza, F., Corriveau, P. and Raake, A. (2011), *Study of rating scales for subjective quality assessment of high-definition video*, IEEE Transactions on Broadcasting, Vol. 57, No. 1, pp. 1-14.

[b-Lassalle] Lassalle J., Gros L., Morineau T., and Coppin G. (2012), *Impact of the content on subjective evaluation of audiovisual quality: What dimensions influence our perception?*" IEEE international symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1-6.

[b-Tominaga] Tominaga, T., Hayashi, T., Okamoto, J., and Takahashi, A. (2010), *Performance comparisons of subjective quality assessment methods for mobile video*, Quality of Multimedia Experience (QoMEX).

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, 2014