A New Subjective Audiovisual & Video Quality Testing Recommendation

Margaret H. Pinson and Lucjan Janowski

Introduction

ITU-T Rec. P.913 is a new subjective video quality testing standard that was approved in January 2014. This Recommendation focuses on the evaluation of flat screens, laptops, and mobile devices. P.913 emphasizes flexibility of environment, rating scale, display technology, and stimulus modality (video, audio, or audiovisual). To balance this flexibility, P.913 includes mandatory reporting requirements.

This paper introduces ITU-T Rec. P.913. The reader is assumed to have some knowledge of subjective video quality testing. Pinson et al. [1] provides a suitable tutorial on this topic. ITU-T Rec. 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." is freely available on-line at http://www.itu.int/rec/T-REC-P.913/en.

Environment & Reporting

When testing consumer grade devices, most aspects of the viewing environment have a minimal impact on mean opinion score (MOS) [2]. Consequently, P.913 does not rigidly constrain the environment and does not include monitor calibration procedures. Instead, the experimenter chooses an environment that is suited to the experiment. This alternate paradigm encompasses distracting environments, monitors

that cannot be calibrated, mobile devices that only play highly compressed signals, questions that can only be answered using modified rating scales, and mixed evaluation of video and audio.

P.913 includes two environment choices: controlled and public. A controlled environment is non-distracting: a comfortable and quiet room that is devoted to conducting the experiment. Examples include a sound isolation chamber, a laboratory, a simulated living room, a conference room, or an office. The P.913 controlled environment allows experimenters to choose an environment where the subject could imagine using the device under test. Lighting is chosen by the experimenter to suit their situation.

A public environment intentionally includes distractions. A public environment can change over time or include people not involved in (or unaware of) the experiment. Examples include a cafeteria, a bus, a busy office, the subject's home, and an otherwise controlled environment with intentionally distracting background noise (e.g., crowd noise, traffic noise, sirens). A public environment should represent a distracting environment where a person would reasonably use the device under test.

The importance of the public environment can be seen in Harrison et al. [3]. This literary overview summarizes a large variety of studies that evaluate the usability of mobile applications. Of the 163 studies discussed in [3] and conducted from 2008 to 2010, 50% were performed in controlled environments, and 27% were field studies.

Because the experimenter has full control of the environment choice, P.913 mandates that subjective test results carefully document the environment. The report should include:

- a picture of the environment
- type of environment (controlled or public)
- noise level (e.g., quiet, bystanders talking)

- lighting level measured in lux
- viewing distance in picture heights
- type and size of video monitor
- type of audio system
- placement of speakers

Also, a full description may not be possible; for example, if each subject takes a mobile device to their home. Depending upon the type of stimuli, some of these values may be inapplicable.

Types of Stimuli

Quality evaluations of mobile devices and modern video systems can include multiple types of stimuli.The subjective quality test methods used for video are very similar to those used for speech and audio (see for example ITU-T Rec. P.800, ITU-R Rec. BS. BS.1534). One option is to design a series of experiments, as suggested in ITU-T Rec. P.1301, "Subjective quality evaluation of audio and audiovisual multiparty telemeetings."

Another option is to design a single experiment that includes multiple stimuli, and P.913 encompasses this solution. P.913 can be applied to video-only stimuli, audio-only stimuli, audiovisual stimuli, and 3D video stimuli. These can be evaluated in separate sessions or mingled into a single session. Naturally, other ITU Recommendations are better suited to experiments that only evaluate speech or audio quality. Special consideration for 3DTV subjective tests is the focus of several Recommendations that are nearing completion.

Vision Testing

BT.500 and P.910 require that all subjects have normal visual acuity (e.g., on a Snellen chart) and normal color vision (e.g.,

using Ishihara plates). By contrast, the visual screening of subjects is optional within P.913.

We are not aware of a definitive study that analyzes the impact of abnormal visual acuity and/or abnormal color vision on subjective quality ratings. Cermak and Fay [4] analyzed the T1A1 dataset's 625 processed video sequences (PVS) and 114 subjects. They concluded that visual acuity and color vision should not be used to screen subjects, because those subjects' data was not significantly different from the rest of the population's data. This hypothesis is supported by [2] and private communication from Cermak describing later experiments. Bovik [6] questions the validity of vision screening, because the general population includes people with normal vision and people with impaired vision. The usual goal of behavior research is to choose a pool that is representative of the general population.

P.913 leaves the choice of visual screening to the researcher, based upon the purpose of the experiment. Visual screening may be desirable when fine tuning compression algorithm improvements yet undesirable when performing a cost / benefit analysis on a product.

Rating Scales

P.913 includes four rating scales that answer different questions (see Fig. 1):

- Absolute category rating (ACR): the subject views one video sequence, then rates the quality on a 5 level scale (excellent, good, fair, poor, bad).

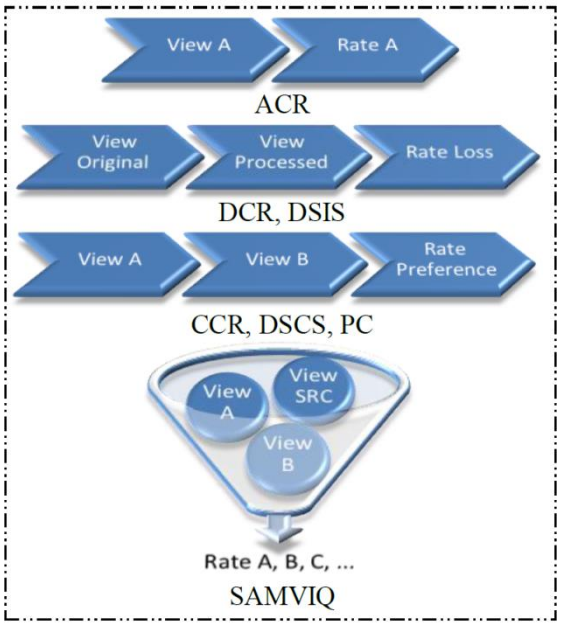


Figure 2. Rating sequence is shown for the four subjective scales in ITU-T Rec. P.913.

- Degradation category rating (DCR) method, also known as the double stimulus impairment scale (DSIS) method: the subject views the original video, views the processed video, and then rates the amount of impairment perceived on a 5 level scale (imperceptible, perceptible but not annoying, slightly annoying, annoying, very annoying).
- Comparison category rating (CCR) method, also known as the double stimulus comparison scale (DSCS) or as pair comparison (PC): two versions of the same source video sequence are viewed in a random order, then the subject rates the second sequence relative to the first on a 7 level scale (much worse, worse, slightly worse, same, slightly better, better, much better).
- ITU-R Rec. BT.1788 (SAMVIQ) and ITU-R Rec. BS.1534 (MUSHRA): a computer interface presents multiple versions of the same source stimuli. The subject may play each stimulus multiple times and chooses the order in which stimuli are rated. SAMVIQ and MUSHRA use a continuous scale with ACR labels.

Each method has a unique design goal. ACR focuses the subject on the task of rating one stimulus in isolation. DCR is an explicit comparison between the reference and impairment. PC allows a direct comparison between two impaired stimuli. SAMVIQ and MUSHRA allow multiple stimulus ratings to be adjusted relative to each other.

P.913 acknowledges that some experiments require modifications to these methods. Some modifications are explicitly identified as acceptable, because prior studies have proven their reliability.

Alternate wording of level labels is the first accepted modification. ITU-T Rec. P.800 has long specified two alternate wordings of the 5 level ACR scale for speech quality tests: listening effort and loudness preference. The MPEG video compression testing [7] used DCR with ACR labels excellent, good, fair, poor, and bad. Other examples are translating into another language, using an unlabeled scale (e.g., endpoints are marked with "+" and "-"), and using a scale with numbers but no words.

Zielinski et al. [8] examines multiple sources of subjective test bias, including prior studies into the impact of the words associated with rating levels. The translation of level descriptors into multiple languages raises a concern that the translated level descriptors will have different distributions in terms of linguistic quality meanings, and that this could bias the MOS ratings. Contrary to this expectation, [8] found that the differences between labeled and unlabeled scales were "negligibly small," indicating that this fear is unfounded. Zielinski theorizes that subjects ignore the verbal level descriptors and either interpret the levels linearly or only take the end points into account. Pinson et al. [2] was also unable to find language or culture based biases. The apparent biases indicated by speech quality experiments, such as Cai et al. [9], can be explained by the use of different speech samples by each lab.

A second accepted modification is ACR with hidden reference (ACR-HR). The source stimuli are rated, and a differential mean opinion score is calculated between the original and processed ACR values. The Video Quality Experts Group (www.vqeg.org) successfully used ACR-HR to validate video quality models. These efforts resulted in ITU-T Rec. J.247, J.246, J.340, and J.341, as well as ITU-R Rec. BT.1866 and BT.1867. This ACR variant has proven value when the choice of method must be a compromise between competing priorities. Examples include measuring difference MOS (DMOS) yet minimizing session duration, and evaluating no-

reference and full-reference objective video quality models on the same subjective dataset. See [1] for more information about the advantages and disadvantages of ACR-HR method.

Increasing the number of levels is discouraged but allowed. An example is implementing ACR as a 9 level, 11 level, or continuous scale. Huynh-Thu et al. [10] and Tominaga et al. [11] compared discrete scales with different numbers of levels (e.g., 5 level, 9 level, 11 level) with continuous scales (e.g., 100 point scales). These studies concluded that continuous scales contain more levels than can be differentiated by people. Increasing the number of discrete levels did not improve the accuracy of the MOS or the corresponding confidence interval. An increase in the number of levels was detrimental, in that the rating task is slower and more cognitively difficult [11].

New Best Practices

Testing of mobile video devices usually requires lossy video playback. That is, the mobile device's video playback introduces quality impairments on the stimuli. P.913 allows for the use of lossy video playback when no alternative exists. Such lossy playback impairments will confound the data being measured, which must be considered during the data analysis.

The detrimental impact of a distracting environment is a reduction in accuracy. P.913 compensates by increasing the number of subjects. Based on [2], P.913 recommends that 24 or more subjects should be used when ACR, DCR, or PC are conducted a controlled environment. This increases to 35 subjects when using a public environment or a narrow range of audiovisual quality. Based on a study by Péchard et al. [12], a minimum of 15 subjects should be used for SAMVIQ and MUSHRA. For any method, smaller numbers of subjects are suitable for pilot studies, to find trending.

Improved procedures for subjective video quality testing have been developed over the last decade of validation tests

performed by the Video Quality Experts Group (VQEG) and the International Telecommunication Union (ITU). These are included in P.913.

- Intermittent impairments should be avoided during the first 1 sec and last 1 sec of a video sequence. These may not be perceptible as impairments in the artificial environment of a subjective test.
- Subjects may be screened (rejected) by calculating the Pearson linear correlation between each subject and MOS calculated from all subjects. If a subject has a low correlation, their data is discarded. The ITU-R Rec. BT.500 screening method is also allowed.
- Long and short stalling events are perceived differently (e.g., 5 sec versus 0.5 sec). Special care should be taken with the instructions, to avoid differences in subject rating behaviors. For example, one subject could assume rebuffering, while another assumes an unintended problem with the subjective test video playback system.

Basic ethical principles should be considered in any experiment involving human testing. In the U.S., the legal requirement for informed consent resulted from the Belmont Report [13]. Informed consent refers to a document that tells subjects of their rights and gives basic information about the experiment. P.913 lists the information that would typically be included and provides an example.

Conclusions

Researchers are encouraged to try the methods standardized in ITU-T Rec. P913 and send the authors feedback on what they liked and disliked, either informally or formally. Question 12 of ITU-T Study Group 9 welcomes contributions that identify improved methods for conducting subjective testing of modern video devices and systems. See

http://www.itu.int/en/ITU-T/studygroups/2013-2016/09/Pages/rapporteurs.aspx for contact information.

References

[1] M. H Pinson, L. Janowski, and Z. Papir, "Video quality subjective testing of entertainment scenes," IEEE Signal Processing Magazine, January 2015.

[2] M. H. Pinson, L. Janowski, R. Pépion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective tests: an international study," IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6, Oct. 2012, pp. 640–651.

[3] R. Harrison, D. Flood, and D. Duce, "Usability of mobile applications: literature review and rationale for a new usability model," Journal of Interaction science, vol. 1, 2013.

[4] T1A1.5/94-148, "Correlation of objective and subjective measures of video quality," GTE Laboratories Inc. (G.W. Cermak and D. A. Fay), Sept. 20, 1994.

[5] T1A1.5/94-118 R1, "Subjective test plan (tenth and final draft)," AT&T Communications (A. C. Morton), Oct. 3, 1993. Available: ftp://vqeg.its.bldrdoc.gov/Documents/OLD_T1A1/

[6] A. K. Moorthy, L. K. Choi, A.C. Bovik and G. de Veciana, "Video quality assessment on mobile devices: subjective, behavioral and objective studies," IEEE Journal of Selected Topics in Signal Processing, vol.6, no.6, p.652-671, Oct. 2012.

[7] C. Fenimore, V. Baroncini, T. Oelbaum, and T. Tan, "Subjective testing methodology in MPEG video

verification,"' SPIE Conference on Applications of Digital Image Processing XXVII, 2004.

[8] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests—a review," Journal of Audio Engineering Society, vol. 56, no 6, Jun. 2008.

[9] Z. Cai, N. Kitawaki, T. Yamada, and S. Makino, "Comparison of MOS evaluation characteristics for Chinese, Japanese, and English in IP Telephony," 4th International Universal Communication Symposium (IUCS), Oct. 2010.

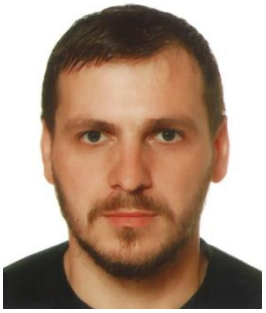
[10] Q. Huynh-Thu, M. Garcia, F. Speranza, P. Corriveau and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," IEEE Transactions on Broadcasting, vol. 57. No. 1, p. 1-14, Mar. 2011.

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

[12] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm." IMQA 2008. Available on-line: http://www.mi.tj.chiba-u.jp/imqa2008/

[13] U.S. Department of Health & Human Services, "Ethical Principles and Guidelines for the Protection of Human Subjects of Research," Apr. 18, 1979. http://www.hhs.gov

Margaret Pinson is an Associate Rapporteur of Questions 2 and 12 in ITU-T Study Group 9. She was the editor for ITU-T Rec. P.913. She investigates improved methods for assessing video quality at NTIA/ITS, in Boulder, Colorado, USA.



Lucjan Janowski is an assistant professor at the Department of Telecommunications, AGH University of Science and Technology, in Krakow, Poland. He is a Co-Chair of the VQEG JEG-Hybrid project (http://www.its.bldrdoc.gov/vqeg/projects/jeg/jeg.aspx).