The Method of Increasing the Accuracy of Mean Opinion Score Estimation in Subjective Quality Evaluation

A. Ostaszewska and S. Żebrowska-Łucyk

Warsaw University of Technology, Faculty of Mechatronics, Sw. Andrzeja Boboli 8 Str., Warsaw, 02-525, Poland

Abstract. The paper presents a method of reliability estimation of scores obtained in subjective quality evaluation, which was inspired by procedures used in between-laboratory tests for determination of repeatability and reproducibility of measurement methods. The proposed use of Madel’s k and h statistics enables for significant decrease of MOS standard deviation, which is exemplified by SSCQE of compressed video results.

1 Introduction

In case of wearable devices each issue starts and ends on a human being and apart from saving life it is also very important to improve its comfort and quality. No matter the definition of comfort or quality is, the only source of knowledge on human sensation and experience is subjective evaluation.

Subjective evaluation is conducted with a group of users of the technology that is being under investigation. Evaluation can be performed with the use of paper questionnaires or with the newer technologies that enable for continuous evaluation for example and need some special equipment. The method of subjective evaluation depends on the character of examined device or technology. Although various methods provide different type of scores, the common problem is the fact, that each of the subjects has his individual requirements, expectations and the level of satisfaction. Hence, the main problem associated with subjective assessments (apart from the fact that they are time- and cost-consuming) is a large dispersion of individual scores, which make estimates of population parameters obtained from experiments - generally speaking – of a little precision. In the language of statistics it means that the confidence intervals of calculated parameters are wide, and in the language of metrology it means a big uncertainty of measurements.

The accurate estimation of the expected value is of great practical importance. The most common estimator of the population mean is Mean Opinion Score (MOS) - mathematically averaged opinionated scores.

MOS is for example used in the area of subjective quality evaluation of compressed video. This area of evaluation came into existence and evolved parallel to...
digital compression. As digital video has a wide audience, a lot had been done to improve the technology. There are recommendations on performing the quality evaluation with the detailed description of the various methods of collecting data. However, there is a problem with wide uncertainty intervals still, despite the fact that there are some ideas proposed of how to deal with that fact.

The paper presents the new method of data filtration, which enables for decreasing confidence intervals for the mean. The method is used for data that regard subjective continuous quality evaluation of compressed video, but the same attitude can be used in any area of measurement with the human audience.

Subjective quality assessment consists in evaluating the quality of the test sequence by the appropriately numerous group of observers and coming to a conclusion on the basis of the set of individual scores.

2 Subjects Selection

In case of subjective quality evaluation the key role is played by the observers. Their ability to focus, perceptiveness, the skill to translate their impressions to a measurement scale and reproducibility (in the meaning of the degree of compliance of scores in the repeated evaluation of the same sequence) influence not only measurement accuracy but also the usefulness of the experiment. Obviously not everyone has the right set of features to be predestined to accurate quality evaluation. Besides even a subject with the best perceptual features may in some adverse circumstances perform poorly due to the attention deficit.

That’s why the procedure of subjects’ selection should include two complementary phases. The first is the preliminary test (pretest), which is the psychometric test for skills and abilities to perform the task of constant observation and evaluation of a test material. The subject would be admitted to the further tests on the basis of the pretest result. The second phase of the selection process should base on critical analysis of individual scores given during the final test session. The scores which don’t fulfill the criteria (for) shouldn’t be taken into consideration whilst MOS computing.

In ITU Recommendations [3, 4, 5], which set the way of data proceeding for R&D centers all over the world, there admittedly is a similar attitude contained, but result of using it is not satisfactory. The pretest segment is limited to screening for normal visual acuity or corrected-to-normal acuity and for normal color vision. According to ITU Recommendations, subjects should not be directly involved either in picture or audio quality evaluation as part of their work and should not be experienced assessors. These undoubtedly are necessary conditions but not sufficient, as it doesn’t take into consideration the ability to express the impression with the use of measurement scale nor the ability to maintain attention for a time of a test session. ITU Recommendations allow to discard the scores from the final test session, although the proposed methods don’t have sufficient mathematical substantiation. Furthermore the experiments conducted in IMIIB PW show the proposed by ITU methods of discarding are not effective.

The necessity of improving methods and criteria of selection of subjects is obvious for computer scientists who work in the area of multimedia data compression.