COMMITTEE T1

CONTRIBUTION

Document Number: T1A1.5/95-102

****

STANDARDS PROJECT: All Performance Standards Projects

****

TITLE: Assessing the Composite Performance of a Hypothetical Reference Circuit (HRC) for an Ensemble of Source Material

****

ISSUE ADDRESSED: Objective and Subjective Performance Testing

****

SOURCE: NTIA/ITS - Stephen Wolf

****

DATE: January 9, 1995

****

DISTRIBUTION TO: T1A1.5

****

KEYWORDS: Video Performance Testing, Objective Video Quality, Subjective Video Quality

****

Abstract

A matter of considerable importance to Working Group T1A1.5 is the method used to obtain a composite rating of Hypothetical Reference Circuit (HRC) performance for an ensemble of source material. Two possible methods have been discussed in T1A1.5. In the first method, subjects are asked to separately observe and rate each of several HRC - source material combinations, and the separate ratings are then averaged over subjects and source materials to produce a composite rating for each HRC. In the second method, subjects are asked to rate each HRC based on a single observation period in which all the source material is presented; the individual subject ratings are then averaged to produce a composite rating for each HRC. This contribution evaluates the two methods in light of accepted industry practice and statistical considerations. It is shown that the first method is far more widely used and offers substantial statistical advantages over the second.

1. Introduction

When assessing the performance of a Hypothetical Reference Circuit (HRC), decisions are almost never based on the use of a single piece of source material. Instead, the performance of an HRC is assessed using an ensemble of various source materials that are relevant for the intended application and subject population. For subjective testing of non-interactive video services, CCIR Recommendation 500-5 [1] specifies that a minimum of 4 test scenes be used. Similarly for audio services, CCITT Recommendation P.80 [2] specifies 2 to 5 sentences per Group with about 5 to 10 Groups being used. For some decisions, the computation of a composite Hypothetical Reference Circuit (HRC) performance rating for the ensemble of source material is required. At the October 1994 T1A1.5 meeting, participants discussed two possible methods that might be used to obtain this composite HRC performance rating.

Method 1: Subjects are asked to rate the perceived quality of each combination of HRC and source material (for video, a piece of source material is a test scene; for audio, a piece of source material is an audio recording). An HRC score is produced by averaging over viewers and source material (in the most general case, a weighted mean over source material could be computed if some material is deemed more important). This average HRC score is then used to assess the relative performance of the HRCs.

Method 2: Viewers are shown the entire ensemble of source material for an HRC and are then asked to rate the perceived quality of the HRC. An HRC score is produced by averaging over viewers. This average HRC score is then used to assess the relative performance of the HRCs. This method has been occasionally mentioned by members of T1A1.5 as a possible method for determining HRC performance.

This contribution compares both methods and examines their use by the national and international community. As detailed in sections 2 and 3 below, the two methods differ substantially in their industry utilization and statistical effectiveness. The contribution shows that not only is the former method widely used and accepted by the national and international community, but that this method has significant statistical advantages over the latter method.

2. Industry Practice

When a composite HRC performance rating is required, method 1 is the method that has been used almost exclusively by the national and international community. For subjective video testing, the fully randomized testing procedure specified by Recommendation 500-5 excludes the possibility of using method 2 (see section 3.1). The "Presentation of Results" sections (3.11 and 4.11) in Recommendation 500-5 use method 1 averaging. For subjective audio testing, the "Statistical Analysis and Presentation of Results" section (B.4.7) in Recommendation P.80 specifies that the male speech mean and the female speech mean should be calculated and presented for each condition (HRC).

Specific recent examples that have utilized method 1 for audio include the CCITT 16 kbps voice codec tests which resulted in G.728 [3], the CCITT 8 kbps voice coder tests [4], and the very low bit rate visual telephony voice coder tests [5]. Recent examples for video in the national and international community include the 45 Mbps coding/ transmission tests [6], MPEG 1 tests [7, 8, 9], and the high definition television (HDTV) coding/transmission tests [10]. For the video examples cited above, method 1 has been used even when the scene ratings for a given HRC span a large portion of the quality scale. For voice, the span in quality can also be substantial -- particularly for low bit rate codecs when child speakers are included. This has not, however, precluded the use of method 1 type averaging (see for example [5] where the female mean, male mean, child mean, and overall mean are all calculated and presented).

Clearly, method 1 is the accepted industry practice. Although an exhaustive search of the literature has not been performed, the use of method 2 appears to be quite rare.

3. Statistical Effectiveness

3.1 The Effect of HRC and Scene Presentation Order

Method 1: The presentation order is fully randomized over both HRC and scene so that an individual viewer is not exposed to many consecutive showings of either an HRC or a test scene. For instance, if 5 HRCs are being tested with 5 test scenes, there are a total of 25 HRC x scene combinations that are presented to the viewer in a random order. This full randomization is recommended in CCIR Recommendation 500 [9] -- the statistical purpose of which is to greatly reduce the effects of HRC and scene ordering on a test subject.

Method 2: The HRCs are presented in a random order to the subject. The scene ordering can be further randomized for each HRC showing. Although this randomization will reduce the HRC and scene ordering effects for a group of subjects, a particular subject may suffer significant HRC and scene ordering effects. To more fully illustrate this point for scene ordering, suppose that the most difficult test scene to code is shown last for a particular HRC and subject. That subject is likely to be more strongly influenced by the last test scene since it was shown immediately before the subject was asked to rate the HRC quality. Thus, the subject may inappropriately downgrade that particular HRC. To more fully illustrate this point for HRC ordering, consider how two subjects might rate the quality of a "medium" quality HRC in the following example where three HRCs are being rated (a “high” quality, a “medium” quality, and a “low” quality HRC). The first subject is shown the high quality HRC first and then the medium quality HRC. The second subject is shown the low quality HRC first and then the medium quality HRC. To the first subject, the medium quality HRC may look quite poor compared to the previous HRC that was shown, whereas, to the second subject, the medium quality HRC might look quite good compared to the previous HRC that was shown. The HRC and scene ordering effects described here for method 2 can be averaged out if enough viewers are used. However, for a fixed number of viewers, method 2 will have larger order effects than method 1. This will tend to increase the uncertainty of the test (i.e., the average standard deviation of a viewer will be higher for method 2 than for method 1).

3.2 The Ability to Discriminate

Method 1: The standard error of an HRC score using this method tends to decrease as 1 over the square root of {the number of viewers times the number of scenes}.

Method 2: The standard error of an HRC score using this method tends to decrease as 1 over the square root of {the number of viewers}.

Thus, method 1 can provide superior discrimination capability by a factor of 1 over the square root of {the number of test scenes}.

3.3 The Effects of Limited Sampling

When assessing the performance of an HRC, one really desires to obtain an estimate of user satisfaction. User satisfaction could be based on “long term” or “short term” assessments of HRC quality (e.g., my service is normally pretty good but yesterday it was really bad). Months or perhaps years may be required for the user to form a “long term” opinion of the HRC quality. In this respect, neither method can be expected to simulate long-term exposure since neither method can handle the number of minutes of learning a viewer would have over a 6 month to 1 year period.

The subjective rating of an HRC x scene combination is in some sense a “short term” estimate of quality. The two methods above attempt to produce estimates of the “long term” quality, but in different ways. As such, each method has limitations.

Method 1: This method simulates sampling the “short term” user opinion of an HRC over many months or years, since the perceived quality is only dependent upon the scene being shown to the viewer at the time of the sampling. The assumption implicit in this method is that the average of “short term” quality estimates gathered over many months approximates a user’s “long term” estimate of quality. Evidently, this is an assumption that many standards bodies and organizations consider reasonable. Note: This method has the flexibility to accommodate other assumptions. For instance, if the “worst” events are thought to more strongly influence long term quality opinions, one could average the worst 10% of the observed short term opinions.

Method 2: This method artificially condenses months or years of test scene material into a short viewing session. This inappropriate approximation requires the user to hold too much material in memory when forming an opinion of an HRC.

4. Conclusion

For some decisions, the computation of a composite Hypothetical Reference Circuit (HRC) performance rating for the ensemble of source material is desired. This contribution has presented and discussed two methods that might be used to obtain this composite HRC performance rating. Method 1 produces an estimate of the HRC performance by averaging over scenes. Method 2 produces an estimate of the HRC performance by showing viewers the entire ensemble of test scenes and then asking them to rate the performance of the HRC. This contribution has shown that

1. Method 1 is accepted industry practice and is the preferred method of choice by the national and international community for non-interactive audio and video performance testing. In the more general discussion, method 1 is the method of choice in virtually all fields of human endeavor when an overall measure of performance is required (e.g., beauty contests, sporting contests such as figure skating and gymnastics, scholastic testing, computer performance testing).

2. Method 1 has a number of statistical advantages over Method 2. These advantages include decreased HRC and scene ordering effects, and increased HRC discrimination capability.

Considering the above discussion, it is not surprising that method 2 may produce a different result than method 1. This being the case, it may be desirable for T1A1.5 to adopt method 1 as the preferred method when a composite HRC performance rating for an ensemble of source materials is required.

5. References

1. CCIR Recommendation 500-5, “Method for the Subjective Assessment of the Quality of Television Pictures.”

2. CCITT Recommendation P.80, “Methods for Subjective Determination of Transmission Quality.” (T1A1.6/92-098)

3. CCITT Speech Quality Experts Group (SQEG) SQ-10.89R, “Subjective Test Methodology for a CCITT 16 kbit/s Speech Coder.”

4. CCITT Speech Quality Experts Group (SQEG) SQ-2.93R, “Subjective Test Methodology for a 8 kbit/s Speech Coder.”

5. T1A1.5./94-417, “Rapporteur’s Report for SG 15 LBC Experts meeting, July 25-27, 1994, Grimstaad, Norway.”

6. T1Y1.1/90-502R4, “Test Procedure for Evaluating Proposed Algorithms for Broadcast Quality NTSC Television Transmission at DS3.”

7. T1A1.5/94-303, “Consumer Judgements of MPEG 1 Video.”

8. Robert S. Fish and Thomas H. Judd, “A Subjective Visual Quality Comparison of NTSC, VHS, and Compressed DS-1 Compatible Video,” Proceedings of the SID, Vol. 32/2, 1991.

9. C.A. Adesanya, A.M. Lessman, and J.R. Rosenberger, “Video and Multimedia Performance Evaluation - A Beginning Perspective,” Globecom, 1992.

10. Federal Communications Commission, Advisory Committee on Advanced Television Service, “ATV System Recommendation”, Advanced Television Test Center, 1330 Braddock Place, Suite 200, Alexandria, VA 22314.