Subject: Objective Methods

Title: Proposals on Methods for Determining Algorithms for Recommendation (Agendum #3)

Date: October 6, 1997

Source: Al Morton, AT&T

Data Analysis Methods

The JRG needs to agree on one or more methods to evaluate the error between Predicted Mean Opinion Score (MP) and the observed MOS of each scene-HRC combination.

John Beerends of KPN has proposed two methods in his May 22 e-mail, one used by ITU-R TG 10/4 (that normalizes MP-MOS to the subjective confidence interval) and simple correlation coefficient.

Dave Fibush of Tektronix suggested an examination of outliers and pair-wise comparison in his May 23 e-mail. The first can be accomplished by examining the MP-MOS difference distribution. The pair-wise analysis uses MOS for two HRCs (same scene). Ideally, MP1-MP2 should have the same relative difference as MOS1-MOS2 when the differences are statistically significant, and should at least have the same sign.

Adding to this discussion, we **propose** the root-mean-square-error (RMSE) between predicted and observed MOS. We prefer RMSE to correlation when examining the predictions of different models.

In this discussion, it may be helpful to introduce the concept of a Main Analysis that would include the primary methods of error evaluation. Other methods could be included in an Auxiliary Analysis. We suggest RMSE for the Main Analysis.

Acceptance Criteria

The criteria for accepting any method depends on the target Recommendation's scope and application. There are proponents of various methods that tend toward one of two classes, In-Service Testing and Out-of-Service Testing. The list below gives some possible applications for Out-of-Service Testing.

1. to assess the continued operational readiness of a video transmission system.
2. to compare the relative levels of impairment between two different systems at the same or different bit rates (both may exhibit supra-threshold impairment).
3. to assess the level of sub-detection threshold impairment of a system, in such a way that the assessments of individual systems can be combined in calculations to design a perceptually loss-less system.
4. to assess the level of supra-threshold impairment of a system, in such a way that the assessments of individual systems can be combined in calculations to design a system to a specific perceived impairment level.

This contribution **proposes** that acceptance criteria be agreed for each application listed above, at least those that the JRG effort will address.

It appears that when a reliable Out-of-Service method is available, selection of an In-Service method that trades some accuracy for wider applicability in network operations (application 1.) would be appropriate. Therefore, additional criteria for In-Service methods should also be agreed and then examined following the acceptance of some Out-of-Service method(s).