COMMITTEE T1
CONTRIBUTION

DOCUMENT NUMBER: T1A1.5/93- 163

****

STANDARDS PROJECT: ANALOG INTERFACE PERFORMANCE SPECIFICATIONS FOR DIGITAL VIDEO TELECONFERENCING/VIDEO TELEPHONY SERVICE

****

TITLE: DATA ANALYSIS PLAN

****

SOURCES: DATA ANALYSIS AD HOC COMMITTEE

****

DATE: NOVEMBER 8, 1993

****

DISTRIBUTION TO: T1A1.5 SUBWORKING GROUP ON VIDEO TELECONFERENCING/VIDEO TELEPHONY

****

COMMITTEE MEMBERS:

G. Cermak
W. Coufal
E. Hauch
C. Jones
N. Randall
J. Roth
R. Schaphorst
F. Taylor
P. Tukey
D. Wirth
S. Wolff

1. Introduction

The T1A1 committee is developing a standard for objective rating of video transmission quality, especially for applications in the areas of video telephony and video conferencing. Document T1A1.5/93-107R1 describes the overall project. The general strategy is to consider a number of proposed objective measures of video quality which are based on physical measurements of the video signals before and after passing through various hypothetical reference circuits (HRCs), and to determine how well they predict people's actual subjective impressions of video quality.

The plan involves choosing a collection of video samples, passing them through a representative collection of HRCs, making the proposed objective measurements of output video quality, and separately having a panel of viewers compare the same video scenes both before and after processing through the HRCs and rate their subjective impressions of the visual degradation. (In a few cases, the actual video scenes will differ between the subjective and objective phases.) Once both sets of data are available, they will be combined into a single analysis to determine the relative effectiveness of the various objective measures for tracking the subjects' responses. The final output of the study will be not only an evaluation of the objective measures that have been proposed (possibly modified and combined in certain ways if doing so produces significantly better performance), but also the video test material and fitted statistical models which will provide a test-bed for evaluating new objective video quality measures proposed in the future.

Document T1A1.5/93-014R4 presents a detailed test plan for collecting the subjective evaluation data.

The current document outlines some planned approaches to analyzing and combining the subjective and objective data.

2. Form of Data

Before specific details can be laid out for a data analysis that combines the subjective and objective data and yields the final results, there must be a clear picture of the nature, quantity, and form of the data that will become available for analysis.

The form of the subjective data is completely spelled out in Document T1A1.5/93-014R4. If the test proceeds according to plan, there will be 25 video scenes combined with 25 HRCs, for a total of 625 scene-by-HRC pairings, each embodied in a video clip. A total of 30 test subjects will view these 625 video clips at 3 test laboratories (10 subjects at each). Each viewer will view 1/3 of all the test material, in a statistically partially balanced incomplete block design with randomized presentation sequences. Each video presentation will include both the original video scene and the post-HRC version in juxtaposition, and the viewers will give their ratings of video degradation on a 5 point scale.

The nature of the objective data is less clearly established at the present time. We seem to have emerging three distinct categories of candidate objective tests:

The first category includes three tests, each of which is tied to a particular video test pattern:

A. Rotating Wheel -- which emulates a periodic pattern in the temporal domain to measure motion distortion, and yields two independent measurements.

B. Scene Cut -- which measures the response time to a scene cut, as a frame count until peak output is again achieved.

Since these tests depend on special test patterns, they cannot be applied to the individual 625 scene-by-HRC video clips, but instead will be applied simply to the 25 HRCs alone, so the output will be a set of 25 numbers. The analysis of this data along with the subjective data will involve either (i) regarding groups of 750 (= (25 scenes) x (30 subjects)) observations as replications for 25 different test conditions (the HRCs), or (ii) first collapsing the subjective data across test subjects by averaging, and regarding the scenes as replications, to produce a reduced set of 25 replications for each of the 25 test conditions.

The Scene Cut data is a frame count -- hence likely to be an integer on the scale of 0 to 90 or so, which we can probably regard as a continuous scale. We do not yet know precisely what the scale of measurement for the Rotating Wheel will be, but believe it will be also essentially on a continuous, positive scale.

This category of candidate objective tests also includes several "traditional analog tests" which will be performed using other special test waveforms. Specifically, frequency response will be measured using the static zone plate waveform and signal/noise ratio will be measured using one of the constant IRE pedestals.

The second category of candidate objective tests contains tests that do not depend on special test patterns and can be applied to the actual scenes used in the subjective part of this study. It includes several fundamentally different tests for measuring spatial impairments (e.g. blurring) and temporal impairments (e.g. jerkiness). The tests are non-intrusive, based on in-service measurements that are extracted from natural scenes, as opposed to test patterns. Objective quality parameters are formed by comparing measurements from source scene (i.e. the video scene input to the HRC) to corresponding measurements from the destination scene (the video scene output from the HRC). About 10-15 different comparison functions have been developed that show promise, hence there will about 10-15 different objective parameters (quality measures) per scene/HRC pair submitted for the current study. For example, a blurring parameter is computed as the total edge energy in the source scene minus the total edge energy in the destination scene, normalized by the former.

Objective tests in this category will yield values for all 625 video clips in the subjective study, so we only need to regard test *subjects* as replications, and we hope to learn not only how these objective tests interact with (i.e. predict video quality for) HRCs, but also how they may interact with scenes and indeed how they may interact jointly with scene and HRC.

Although we are told these tests yield values on continuous scales, we anticipate that there may be issues of log (decibel) and other transformations to consider, especially given that some of them are based on measures of "energy". Our need, generally, is to find those transformations of the objective measures that are most nearly linearly related to the natural scales of perceived video quality/degradation, since we prefer to express our final results in terms of linear models (regressions, etc), as described below.

We note that in most cases, the raw versions of these objective measures are computed per frame; hence, there is an issue of how the frame-specific values are combined across a video clip of several seconds duration. Although these decisions will be made before the candidate measures are submitted for this study yielding a single number per

video clip (per objective measure), we recognize the possibility of obtaining and analyzing the more detailed version of the data in relation to our subjective data, if that should prove desirable and profitable.

The third category of objective test is special, consisting of just one physical measurement, the (effective) bit-rate of the HRC. It is not really a candidate objective video quality measure at all, but in a sense a "nuisance parameter" or covariate in this problem. We naturally expect that all reasonable HRCs with very high bit-rates will have very high-quality transmission, and low bit-rate HRCs will give up quality in one way or another. The objective in this study is to learn about dimensions of video transmission quality that go beyond bit-rate. For this purpose, we will need to introduce each HRC's bit-rate as an auxiliary parameter to be "controlled for" or "conditioned on" or "partialed out".

3. Methods of Data Analysis

Earlier contributions describing this work have loosely referred to the data analysis phase in terms of "correlating" the objective and subjective measures of video quality based on the data collected in this study. The word "correlating" in this context should be interpreted only in the general sense to mean quantifying and calibrating the interrelationship between subjective and objective measures. Actual correlation coefficients of the classical Pearson product-moment variety will likely only be an auxiliary output of the analysis.

We expect that the core analysis will take the form of a series of regressions with linear terms and possibly quadratic terms (or terms based on other monotonic transformations of the independent variables, if necessary). The dependent variables will be the subjective responses for each of the 625 HRC/scene pairs in the study, and the independent variables will be the objective measures -- either for the same 625 HRC/scene pairs, or for the 25 HRCs alone, as appropriate.

Some of these regressions will involve using subsets of the objective measures, the actual numbers and choices of objective measures for these groups depending both on prior engineering judgement and on preliminary analyses that will prescreen objective measures to reduce redundancy -- since we realize that many of the objective measures will probably be quite similar to each other.

The prescreening, more specifically, will involve the use of factor analysis and possibly one of several well-known cluster analysis techniques.

The primary analyses will be based on the whole data set in order to establish a baseline of a global relationship between subjective and objective measures of video quality. Similar auxiliary analyses will examine subsets of the data which correspond to particular ranges of bit-rates and intended uses.

Some of the regressions will involve subsets of the scenes, according to preassigned scene categories, as suggested in Document T1A1.5/93-107R1. Also, as discussed in Section 5 below, some of the regressions will involve different weights for different scenes -- at least tentatively.

As an adjunct to the main regression analyses, we will carry out error distribution analyses on the residuals from the regressions. If we use O(i,j) to represent the objective measure for the ith HRC and jth scene, S(i,j,k) to represent the subjective rating for the

ith HRC and jth scene by the kth subject, and s(i,j) to represent the fitted value from a regression (the s(i,j) are "predicted" subjective values, given the corresponding objective values), then the raw residuals (or "errors") entering the error analyses are:

$$E(i,j,k) = S(i,j,k) - s(i,j)$$

where

$$s(i,j) = A + B * O(i,j)$$

and A and B are the fitted regression coefficients. These values will be plotted in various ways and analyzed in relation to other variables not entering the main regressions (e.g. demographics of the test subject, testing lab, presentation sequence number, and higher-level interactions) in order to confirm (or refute) that such other variables have been properly controlled for in the experimental design and primary analysis.

The individual errors and their distribution will be examined over the experiment as a whole, and for subsets of scenes (or Application Categories) -- as discussed below.

If it proves fruitful, the data will be averaged across replications (test subjects) before computing residuals -- yielding smaller and possibly more manageable sets of values to analyze. Also, for those objective test candidates that produce only one measurement per HRC, independent of scene, the fitted values are of the form s(i), and the residuals become E(i,j,k) = S(i,j,k) - s(i).

In those regressions where scene-specific weights are used, the residuals will be standardized (weighted) in a corresponding way before the error analysis is carried out.

Part of the output from the core analyses and error analyses will be estimates of the variance of prediction for the regression equations, computed in a statistically appropriate fashion, taking proper account of weights, categories, and degrees of freedom.

Another important outcome from this work will likely be some insight into how strongly subjective responses are related to the spatial resolution dimension vs. the temporal dimension, etc. We will look for the underlying dimensions of perceived video quality -- to find out what actually *matters* to people -- using some combination of multidimensional scaling, factor analysis and cluster analysis

Part of the follow-up analysis will involve a subjective-data-based cluster analysis of both the scenes and HRC's -- to confirm or refute prior conceptions about which ones should behave similarly. Also, if cluster analysis of test subjects (viewers) reveals recognizable sub-populations defined in terms of their preference patterns, that will be reported as an auxiliary result.

The presentation of the final data analysis will include a discussion of the range of applicability of the results.

4. Honest (Conservative) Standard Errors

In any study of this kind, where a number of candidate measures are being entertained, there is a tendency to focus attention on the one or few that show the best performance (the best prediction of subjective response). But the standard error (attained p-level, etc.) of the best candidate, computed in the normal way, is too optimistic: its performance is likely to appear better, simply because it is the best of the bunch. As an extreme case, if we had 20 candidates that were completely uncorrelated with the response (i.e. no predictive power at all), it is likely that one of them would appear to

have statistically significant predictive power at the 5% level -- if the significance level is computed in the naive way and taken seriously!

This is the so-called multiple comparisons problem in Statistics. The corrective action is to penalize the "best" candidate in accordance to the number of (independent) candidates that entered the comparison, in order to get a more conservative (i.e. statistically unbiased or accurate) picture of its predictive power. We plan to use multiple comparisons corrections in presenting our final results.

A closely related issue is the distortion of prediction accuracy that occurs when a single set of data is used both to fit a model and then to assess its performance in a naive way. Our basic plan calls for outside parties to submit candidate objective measures and for us to analyze them without tinkering with them in any way. Nevertheless, we recognize that in the best interests of the industry, if we discover that some of candidate can be markedly improved through transformations, say, or through combining several of them in linear combinations to synthesize new composite measures, then we will report those results, as well. If this occurs, we will be careful to calculate "honest" standard errors that take proper account of the reuse of the data to both synthesize new composite measures and assess their performance. Methods of cross-validation (including the bootstrap) will be used for this purpose.

5. Levels of Aggregation & Reference Baselines

The committee recognizes that in practice there will be distinct types of applications for which video quality measures will be needed, and not all of the test material in the current study will be equally relevant to all potential types of applications. This section proposes a systematic way of addressing these needs in the context of the current project. Although there is consensus on the overall concept, some of the details are still subject to revision.

Objective measurements are designed to predict subjective test scores. This could be done at several levels. For example, an objective test could be designed to predict the subjective test scores obtained by various HRCs when the input is one of the actual test scenes that we have used (or video material of a very similar kind). In this case, our study provides precisely the required "reference baseline", namely the 25 test scores that we will obtain from the subjective tests for that scene (averaged across viewers). There will, in fact, be 25 different reference baselines -- one for each scene, and 25 separate regressions, each using only 1/25 of the total data.

Other objective tests may be designed to predict subjective test scores when the input to the HRC is material that is generally representative of the specific application for which the HRC is designed. To address this anticipated need, we will aggregate the 25 scenes into several Application Categories, and compute residual errors and a regression analysis separately using the subset of scenes in each category. The Application Categories must be established and approved; they will be the same for each HRC; some scenes can appear in more than one category. Each HRC will be evaluated for all of the Application Categories that are selected and approved. Each Category forms a "reference baseline" for the given type of application. One proposed allocation of scenes into Application Categories is as follows, but it is subject to further revision:

Application Category	Scenes
Video Telephony	vtc1nw, susie, disguy, disgal, smity1 smity2
Distance Learning	vtcmp, vtc2zm, boblec, smity1, smity2, filter, inspec, vowels
Video Teleconferencing	3inrow, 5row1, intros, 3twos, 2wbord, split6
Presentation Graphics	vtcmp, vtc2zm, washdc, boblec, 2wbord, cirk-it, rodmap, filter, ysmite, vowels
Entertainment	flogar, ftball, fredas, vtc1nw, 2wbord, vowels

Following this line of thinking one step further, we anticipate that not only will some scenes be irrelevant to certain applications, but even among the scenes that *are* relevant to a given application, some scenes will be more important than others. To capture this in our analysis, we will try refitting the Application-Category-specific regressions with differential weights assigned to scenes to see whether more incisive results are obtained. If it is perceived that they are, they will be reported.

Still other objective tests may be designed to predict the overall subjective test scores when the HRC is subjected to a wide variety of inputs. In this case, our entire collection of scenes is the appropriate reference baseline, with all the scenes aggregated into a single category to compute one overall regression. Even so, however, we can entertain the use of differential weights for certain scenes, to get a balance that is acceptable to all the participants in the project.