COMMITTEE T1
CONTRIBUTION

Document Number: T1A1.5/94-127

****

STANDARDS PROJECT: Analog Interface Performance Specifications for Digital Video Teleconferencing/Video Telephony Service

****

TITLE: Accuracy in Predicting Subjective Video Quality

****

ISSUE ADDRESSED: Relations Between Objective and Subjective Measures of Video Quality

****

SOURCE: GTE Laboratories, Incorporated
(Gregory W. Cermak)

****

DATE: 28 March 1994

****

DISTRIBUTION TO: T1A1.5

****

KEYWORDS Video Quality, Objective Quality, Subjective Quality, Statistical Measures, Correlation

****

1. Introduction

The T1A1 committee is sponsoring a program to measure the quality of video signals. The general strategy is to empirically determine a set of *objective* measures of video quality that predict users' *subjective* judgments of video quality. Documents in the series T1A1.5/93-014R describe the plan for collecting subjective measures of video quality. Documents T1A1.5/94-101 and T1A1.5/93-152 are sources of references to many T1A1 documents on objective measures of video quality. Document T1A1.5/94-112 describes the data analysis plan in which sets of objective measures will be used to predict subjective measures which have been collected for the same video systems.

One would like a criterion for judging accuracy of prediction. Several criteria are possible. One is to do at least as well as in the past; that criterion could be taken as a lower bound for prediction accuracy. Another is to approach the limit imposed by noise inherent in the subjective and objective measurement systems; that criterion could be taken as an upper bound.

2. Upper Bound

An upper bound on how well objective measures can predict subjective measures (which I'll call "the fit") is given by the proportion of random error in both the objective measures and the subjective measures. Because the T1A1.5 committee has so far spoken much more of error in the subjective measures, and because we have estimates of the error in the subjective measures, this note will concentrate on error in the subjective measures as providing the upper bound. However, take note: There is surely error in any objective measures as well, and the effect of this error can be difficult to quantify when multiple objective measures are used to predict a subjective measure.

Think of a subjective measure as having two components, one systematic and one random. Only the systematic component can be fit by the objective measures. The larger the random component, the poorer the fit that will be observed, even if the objective measures are perfect in the sense that they unerringly predict the systematic part of the subjective measure. Below, I quote from statistics

textbooks to support this claim. Then I show how observed randomness in the subjective data can be used to set an upper bound on the fit of objective measures.

N. R. Draper and H. Smith's Applied Regression Analysis (1966) is a standard text on regression. Section 1.5 is titled "Lack of Fit and Pure Error." Draper & Smith (pp. 26-29) note that the difference between (a) predicted values of the dependent variable and (b) the observed values (the residual) can be decomposed into a systematic component and a random component. The random component is the error variance or "pure error" in the dependent variable (the subjective measure of video quality in our study). The systematic component is the non-random difference between predicted and observed, for example when the prediction is linear but the data follow more of a log function, or when a crucial predictor variable is left out of the set of predictors.

The systematic part of the residual can be fixed (in principle) by finding the right model -- the right form of the relationship between the objective measures and the subjective measure, and / or the addition of the right explanatory variables. The "pure error" part of the residual cannot be fixed because it is in principle unpredictable, like thermal noise or Brownian motion.

K. A. Bollen's Structural Equations With Latent Variables (1989) presents regression and other techniques in the language of modern "measurement models" in which error in variables is treated more extensively. In this treatment as well, the variables one observes consist of a systematic component and a random component. The size of the random component for a variable is related to the size of the variation in the systematic part of the variable by a ratio which is called the *reliability* of the variable (pp. 208-209):

Reliability = variance(systematic) / [variance(systematic) + variance(random)].

The reliabilities of variables provide a bound on their correlation or univariate regression, (p. 157):

Observed R^2 = ("True" R^2) * (reliability of x) * (reliability of y).

Thus, suppose that the reliability of the objective measure x were perfect, and that this measure perfectly predicted some ideal measure of the subjective data, y. Because the subjective data actually have reliability less than 1.0, the observed R^2

will also be less than 1.0. In fact, the observed R^2 in this case will exactly equal the reliability of the subjective data.

There are various ways to measure the reliability of a variable such as the subjective video quality data. We have a measure from an analysis of variance of the GTE Labs subjective data. A model of the 9287 data points we collected is

Rating = F(HRC, Scene, Subject, all 2-way interactions).

This model does not take account of any objective measures at all, but it does concede that the rating for an HRC-scene pair depends on which HRC and scene are being rated. This model can be taken as an estimate of the systematic variance in the subjective data. It accounts for 0.85 of the variance. That is, this estimate implies that the reliability of the subjective data is 0.85, and that is the upper bound for the fit of any regression predicting subjective measures using objective measures. (Note that the residual term in this model actually includes more than just "pure error;" it also includes a three-way interaction and any unmeasured effects due to the lab itself. A bound that used only "pure error" would be larger than 0.85).

This bound of 0.85 (or larger) would be appropriate if the individual subjects were "accounted for," as they are in the analysis above. Analysis of Variance (ANOVA) accounts for individual subject differences analytically. Other ways of "accounting for" individual subjects are (a) to standardize their data, and (b) to average across subjects before analyzing any data. Thus, if the data analyst averaged all subjective ratings for a given HRC-scene pair before comparing the results to objective measures, then a bound of 0.85 would be appropriate. (Tukey, in a tutorial dated 11/7/93, argues that averaging before doing the regression leads to less robust estimates than doing the regression with raw data.)

However, if one were analyzing raw, unstandardized rating data, then an estimate of the systematic variance would be given just by variance accounted for by HRC and Scene, and their interaction. In the case of the GTE Labs subjective data, that variance is 0.68 of the total. Thus, an estimate of an upper bound on the fit of a regression model to individual subjects' raw data would be an R^2 of 0.68, quite a bit smaller than when the data are first averaged. (Tukey demonstrates exactly this point in his tutorial of 11/7/93.) Using data from all three labs, in which

various lab effects add to the data variability, the corresponding R^2 would presumably be somewhat less.

3. Lower Bound

In document T1A1.5/92-112, Voran and Wolf report results from a testing program very similar to the current one. The subjective testing methodology was the same, and the range of HRC's tested was quite similar. The unit of analysis reported was the HRC-scene pair; apparently the subjective judgments of 48 subjects were averaged for each pair. For 64 pairs the fit of the best set of objective measures to the subjective measures was $R^2 = 0.84$. The residual variance, 0.16 of the total, is some mixture of effects due to the particular lab in which the study was conducted, effects of differences among individual subjects, "pure error" due to lack of repeatability of individual subjective responses, and genuine lack of fit of the objective measures. Because of the averaging over subjects, the effects due to subjects and to "pure error" must have been quite small.[1]

The Voran and Wolf paper does not report fit of objective measures to the raw, unaveraged subjective judgments. Therefore this paper does not give us a lower bound for the case of raw data.

Comparing the actual fit of objective measures in the Voran and Wolf paper (0.84) to an estimate of the systematic portion of the subjective responses in the GTE Labs study (0.85) suggests that the bounds on the fit of objective measures to subjective measures are quite tight for the class of studies we are considering. Naturally, the bounds themselves are subject to some uncertainty, but given the data in hand we should expect that a good model of subjective data would account for about 0.85 of the variance in data averaged across subjects. While we do not have a lower bound for fit to raw subjective ratings, the upper bound from the GTE Labs data was 0.68.

---

[1] Errors due to subject effects and to "pure error" tend to cancel unless the experimental situation is very peculiar. As the number of subjects increases, averaging responses over subjects produces estimates of the mean opinion score that are increasingly accurate. The effects of individual subjects and "pure error" are effectively eliminated as the sample size becomes very large.

Appendix: What Do We Mean By "Random"?

Randomness in Principle

Randomness in principle can be thought of as the fundamental limit on how well we can measure some phenomenon due to its nature and due to the nature of our measuring instrument. Example 1: In experimental design and the branch of statistics that grew up with it, ANOVA, one performs some operations and observes an outcome. If the experiment were in agriculture (as, in fact, many of the early applications of statistics were), one might apply various treatments of fertilizer to various soil types and observe the weight of plants growing in the experimental plots. Randomness in principle enters in (at least) two ways: (A) The phenomenon itself, plant growth, has some inherent variability; two plants in the same plot will not weigh the same. (B) The scales used to weigh the plants, and the operators of the scales, will also show some variability. These two sources of randomness contribute to a fundamental limit in the ability to measure plant response to fertilizer and soil. This kind of error is usually identified with the residual error in ANOVA, i.e., what is left after all the systematic effects of the treatment variables, and their interactions, are accounted for.

Example 2: In a psychophysical experiment, one might present a human subject with a very long series of pairs of tones. In each pair, the tones are either at the same loudness level (measured, say, in volts delivered to standard headphones), or the tones differ by some small amount. The subject's response is to say that the tones were the same or different. In such experiments one observes error. Subjects do not always say the tones were the same when they were, or that the tones were different when they were; subjects may respond differently to the same pair of tones presented at different points in the sequence of tone pairs. Thus, there is some fundamental amount of randomness in the subject's response. Such randomness can even be observed in the firing of auditory nerve cells. This randomness may obscure the fact that subjects do judge loudness, but randomness does not mean that people cannot judge loudness. By observing a subject's responses to many combinations of tone pairs one finds a very systematic relationship between the probability of a "same" response as a function of the difference in loudness (voltage) between the tones in a pair.

Returning to the present issue, on any single presentation of an HRC-scene pair to a subject, there is a certain amount of fundamental randomness in the subject's perception and in the observed response. I call this the randomness in principle. This randomness is measurable. So far, the committee has spoken as though this were the only sort of randomness that need be considered. An estimate of this randomness is given by differences in a single subject's responses to the same HRC-scene pair.

(Note that in a straightforward ANOVA of the subjective responses in our experiment, we do not have a clean measure of this sort of randomness: If we look at all effects and interactions of the variables HRC, Scene, and Subject, there are not enough "degrees of freedom" to estimate a residual term. Our only residual term is the HRC-Scene-Subject interaction; the randomness in principle is buried in that term. To get a separate estimate of the randomness in principle, we can separately consider the four repeated HRC-scene pairs per subject.)

Randomness in Practice

Randomness in principle applies to the smallest unit of analysis, a single observation. We weigh the single plant; we record a single "same" or "different" response in the loudness experiment. To the extent that the single measurement is not repeatable, there is random error.

However, there are many other sources of variation in a measurement, as cataloged in the paper by Crow, "Plan for analysis of interlaboratory video performance standard subjective test data". Normally one thinks of these sources as systematic, and therefore not "error." But, in practice these sources contribute to lack of fit between objective and subjective measures. We may want to explicitly treat these sources as if they were error in calculating an upper bound for the fit of the subjective and objective measures. Tukey, in his 11/7/93 tutorial, shows the effect of these sources in raw subjective data.

A main source of variation is in the subjects being measured. In experimental design one thinks of many plants in the same plot as identical replications of the same plant. We think of many trials with the same subject and the same pair of

tones as replications. However, in the present case, and in many others, we must recognize that the individual subjects are themselves different. Often, one views such differences as systematic, predictable, and therefore not random. However, in practice we do not have the time and money to collect the huge amounts of data that would be necessary to truly account for individual differences among subjects. The simple kinds of demographic data we can collect cheaply are not good predictors of most human perceptual and cognitive responses.

So, a source of variation is individual subjects, and also their "interaction" with the variables HRC and Scene. These "interactions" are systematic differences between subjects in their preference for various HRC's and Scenes. (Note: the simple effect of Subject alone is that they have different mean ratings, no matter what it is they're rating).

Another source of variation is the Laboratory effect. The mean for each laboratory can be slightly different, independent of the effect due to small samples in each lab. That is, the mean ratings for labs A and B could be different even if the number of tests in each lab was arbitrarily large. And, there could be interaction effects of Lab with HRC and Scene.

In the present study, for each HRC-scene pair we have as many observations as subjects (and a very few replications). These observations are scattered about the mean for the HRC-scene pair. The scatter is attributable to the subject differences as well as to the laboratory effects. The scatter can not be handled by any objective measure of the HRC-scene pair, and so is effectively "error," even though the subject and laboratory effects might, in principle, be systematic. Thus, if we fit objective measures to the raw subjective data, we should not consider a model to be unsuccessful if it accounts for a relatively small amount of the variance.