COMMITTEE T1
CONTRIBUTION

Document Number: T1Q1.5/92- 112

****

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

****

TITLE: Objective Measures of Video Impairment: An Update on the ITS Derivation Process

****

ISSUE ADDRESSED: Objective Quality Assessment of VTC/VT

****

SOURCE: National Telecommunications and Information Administration
Institute for Telecommunication Sciences (ITS)
(Stephen Voran)

****

DATE: January 22, 1992

****

DISTRIBUTION TO: T1Q1.5

****

KEYWORDS: Video Teleconferencing, Video Telephony, Subjective Quality, Objective Quality

****

Objective Measures of Video Impairment: An Update on the ITS Derivation Process

1. Introduction

The Institute for Telecommunication Sciences, in support of T1Q1.5, is conducting research to determine what objective measures of video impairment correlate well with human perception. Our work to date is described in a sequence of contributions to T1Q1.5. This contribution provides a set of truly encouraging, very recent results pertaining to subjective and objective measurements of 12 VTC/VT video scenes and 25 transmission service channels. This work continues in a careful and deliberate fashion. Additional results will be provided to T1Q1.5 as they become available.

For a video impairment measurement to be meaningful, it should provide the same type of feedback that the users of the video service would provide. An ideal measurement might simulate a large, unbiased, repeatable user survey and would yield results such as: "15% of the users will find that the impairments created by this transmission service channel are very annoying" or "The mean opinion of the users will be that the service contains impairments that are perceptible but not annoying". An objective measure of this type could provide highly relevant, user survey data without consulting any users. In effect, we are looking for an implementable "black box" that evaluates the transmission service channel and thus replaces the user survey. The following results show considerable progress towards that goal. First, we summarize our procedure for arriving at objective measurements that emulate viewer perception. Next, we describe our process for building a linear model to fit our experimental data. We provide a third-order linear model and evaluate the fit of that model. Finally, we argue that the model has value as a predictor of subjective impairment scores.

2. Summary of the Derivation Procedure

The ITS derivation procedure is fully described in contribution T1Q1.5/91-119. In short, it is a perception-based process, where distributions of viewers' impairment judgments form a subjective data set that is correct by definition. Our job is to define objective video measurements that reproduce this subjective data a nearly as possible. Given enough data for this modeling step, these measurements should provide accurate estimates of subjective responses when the subjective responses are not available. Many of the objective measurements under consideration are described in T1Q1.5/90-123. Each proposed measure must be evaluated on the same video sequences that are evaluated by the panel of viewers. This is accomplished by digitizing the video sequences and performing various digital video processing functions on them. The results of this processing form the objective data set. Further details on this stage are provided in T1Q1.5/91-118. Additional information can be gained by applying separate measures to the motion and still portions of each video frame. The process of segmenting frames into motion and still portions is described in T1Q1.5/91-110.

The subjective viewing tests are conducted in a highly controlled environment (CCIR recommendation 500-3). For details on this step see T1Q1.5/91-119. The results of our most recent set of subjective tests are given in T1Q1.5/91-133. The response forms used and additional description are found in T1Q1.5/91-123. The final stage of the derivation process centers on the simultaneous statistical analysis of the subjective and objective data sets. This

process reveals which of the proposed objective measures have the potential to predict the responses of the viewer panel, and how that prediction might be accomplished. A preliminary run through this stage is described in T1Q1.5/91-124.

3. Derivation Procedure Details

The following analysis pertains to a set of 46 video sequences, each of 9 second duration. These sequences were formed by passing each of 12 different scenes through 3 or 4 transmission service channels. The scenes are all highly relevant to VTC/VT applications. Included are scenes with one through five actors, with and without graphics or other props, a lecture scene, and several close up graphics scenes. The transmission service channels used include 19 configurations of 9 different codecs. These configurations were applied a total of 34 times. The remaining 12 sequences were generated using analog test conditions such as NTSC encoding and decoding, VHS and S-VHS recording and playback, and RF modulation and demodulation under noisy channel conditions (i.e. snow).

The 46 sequences were rated by 48 viewers. The results of this subjective testing process are provided in T1Q1.5/91-133. The results indicate that the 46 sequences span the full range of the impairment rating scale used; from "Imperceptible" to "Very Annoying". Our goal is to predict the distribution of the viewer responses. (e.g. "15% of the users will find that the impairments created by this transmission service channel are very annoying"). We have done some preliminary work on this topic are encouraged by the initial results. In the following however, we have elected to simplify the problem by collapsing each distribution of subjective judgements down to a single central measure according to

$$s = \sum_{i=1}^{5} i p_i \ ,$$

where p_i is the fraction of viewers that responded in i^{th} category. The impairment categories are: 1=Very Annoying, 2=Annoying, 3=Slightly Annoying, 4=Perceptible but not Annoying, 5=Imperceptible. The problem now becomes one of modeling this central measure as accurately as possible for each of the 46 video sequences. In the following, we refer to this central measure as the "true score".

4. The Linear Model and Measurement Selection

In order to solve this problem, we have developed an algorithm that searches through our collection of over 100 proposed objective measurements and selects a much smaller set that forms the basis of a close fitting model. When combined through the appropriate function, the members of this smaller set generate a model for subjective score values that is highly correlated to the true score values. At this stage, we consider only the simplest of all combining functions: the linear combination. Thus, we are looking for p measurements $\{m_i\}$ and p+1 constants $\{c_i\}$, that will allow us to calculate

$$s \approx \hat{s} = \left\{ \sum_{i=1}^{p} c_i m_i \right\} + c_{p+1} \ .$$

For this search procedure, we adopt the least squares error criterion:

$$MIN\left\{\sum_{i=1}^{n}(s_i-\hat{s}_i)^2\right\},$$

where n denotes the number of video sequences involved in the test.

If the p measurements are given, then solution to this least squares problem is well known and can be found in almost every linear algebra textbook. In our case, the measurements are not given. They must be selected from a larger set. Our selection algorithm iterates between a selection step and a least-squares solution step to arrive at a nearly optimal set of measurements and constants. Here we present a third-order solution:

$$\hat{s}=\left\{\sum_{i=1}^{3}c_i m_i\right\}+c_4 .$$

Notice that since we are finding a linear fit for 46 data points using only 4 variables, the problem is well over-determined.

The first of the three measurements that we have selected is:

$$m_1=\max_{time}\left\{\left|\frac{std_{space}(Sobel(O_n))-std_{space}(sobel(D_n))}{std_{space}(Sobel(O_n))}\right|\right\},$$

where O_n denotes the n^{th} frame of the original video sequence (input to the transmission service channel or TSC), D_n is the n^{th} frame of the degraded video sequence (output from the TSC), Sobel indicates the Sobel filtering operation, and std_{space} indicates that a standard deviation of pixel values is computed. Since this measure is computed at each of the 270 frames of the 9 second video sequence, it returns a sequence of 270 scalar values. On the other hand, each viewer returns a single measure for the entire 9 second sequence. Since our goal is the emulation of the human visual and perceptual systems, a reasonable next step is to compress the sequence down to a single value, using an algorithm that might approximate the algorithm viewers use. The selection of such "time collapsing functions" is an integral part of our measurement selection algorithm. For the measurement described above, we find that the maximum value of the time series provides the measurement that agrees best with the subjective data. We indicate this operation with the notation "max_{time}".

The Sobel filtering operation enhances edges and other high frequency content in the video frame. The standard deviation returns the non-dc energy of the filtered frame. Thus, m_1 is a normalized measurement of how the high frequency spatial energy is effected by the TSC. When D_n matches O_n exactly, (perfect TSC) the measurement value is zero. The absolute value function ensures that either a loss (eg: blurring) or a gain (eg: blocking) of high frequency image content will cause a positive swing in m_1. In light of this interpretation of m_1, we categorize it as a "spatial measurement". This helps to differentiate m_1 from m_2 and m_3, which fall into the class of "temporal measurements"; those that measure how the TSC impairs the smooth flow of time and motion.

Since the m_1 tends to increase as impairment becomes more severe, the temporal max function essentially reports the "worst case" frame measurement as the measurement for the entire sequence. It seems quite reasonable that the human perceptual system would also work on a "worst case" basis, at least for video sequences of 9 second duration. The magnitude of the coefficient of correlation between m_1 and the true subjective scores, computed across all 46 video sequences is .912. It is this rather high correlation (1=perfect correlation) that causes m_1 to be picked by our measurement selection algorithm.

The remaining two measurements are given as:

$$m_2 = f_{time}\left\{\left(\frac{1}{255}\right)\cdot\max\{\,[RMS(D_n - D_{n-1}) - RMS(O_n - O_{n-1})]\,, 0\,\}\right\},$$

$$where \quad f_{time}(\{x_t\}) = \frac{(\max_{time}(\{x_t\}) - \min_{time}(\{x_t\}))}{(std_{time}(\{x_t\}))},$$

$$m_3 = median_{time}\left\{\left|20 \bullet Log_{10}\left(\frac{std_{space}(O_n - O_{n-1})}{std_{space}(D_n - D_{n-1})}\right)\right|\right\}.$$

Here RMS(F) returns the RMS value of the pixels of frame F. Notice that both of these measurements are based on the input and output first-order temporal frame differences: O_n-O_{n-1} and D_n-D_{n-1}. This temporal differencing operation allows m_2 and m_3 to measure how the TSC distorts time and hence motion. Due to the "max" function, m_2 provides an indication of the amount of motion energy present in the degraded video sequence (output of the TSC) that is not present in the original video sequence (input to the TSC). This situation corresponds to what is often described as jerky motion. The time collapsing function for m_2, given by f_{time} is essentially a measure of the spread of the outliers across the time history of the measurement. This is another type of "worst case" collapsing function. The third measurement, m_3, provides a logarithmic indication of motion gained or lost. The time collapsing function for m_3 is simply the median value of the time history.

As indicated above, these two measurements fit into the category of temporal measurements. As such, they complement m_1 by providing additional unique impairment information. In effect, they were selected because they are well correlated with the residual information that m_1 cannot measure. Because it brings new and unique information, m_2 is nearly uncorrelated with m_1, with a correlation coefficient of ρ=.231. The measurement m_3 brings even more new information, and thus is nearly uncorrelated with m_2: ρ=-.051. In the linear algebraic sense of the term, the measurements are nearly *orthogonal*, and this creates a convenient measurement environment.

Having selected the best set of measurements, it remains only to select the four constants that define the best linear combination of them. The least-squares solution yields:

$$c_1 = -4.3246$$
$$c_2 = -0.0937$$
$$c_3 = -0.0917$$
$$c_4 = +5.1945 \quad .$$

Since the three measurements do not share a common scale, the constants given above should not be interpreted as the indications of the relative importance of m_1, m_2, and m_3. Also, we note that this third-order solution is near optimal given our constraints, but it is not particularly unique. That is, there are other models of similar complexity that provide a similar fit.

5. Observations

In a bench test environment, the TSC input and output are co-located. In field tests this is often not the case, and a measurement question arises: How can we measure the impairment when we need access to both the input and the output of the TSC? For the set of three measurements given above, the answer is as follows. In order to implement the test, we must transmit a small amount of side information from one end of the TSC to the other. In particular, for each frame we can send the three scalar values; $std_{space}(Sobel(D_n))$, $RMS(D_n\text{-}D_{n-1})$, and $std_{space}(D_n\text{-}D_{n-1})$, from the output back to the input. This requires an uncoded data rate of roughly 1.5 Kbps. The measurements can then be computed at the input end of the TSC. There is an exactly equivalent scheme for computing measurements at the output end of the TSC, requiring a forward data channel at the 1.5 Kbps rate.

If one feels that this is an unacceptable price to pay for good video impairment measures, one might consider the alternatives. Clearly, impairment measures of output video cannot be made without some knowledge of the input video. If sending side information is unacceptable, then the only alternative is to use video signals that are known a priori. But can a test signal that is known a priori provide a genuine and rigorous test of a video coder? Is there any risk that video coder design will be influenced by these known test signals? Discussion of these open questions could be quite valuable.

Here is a final observation on this set of measurements. Frames O_n and D_n appear in pairs, with their matching time indices indicating that some timing information is required in order to make the measurements. This is only the case if one requires that the spatial impairments be measured separately from the temporal impairments. At this stage of our development process, such an orthogonal measurement scheme is helpful. In general, however, this time alignment is not necessary, and single parameters will measure mixtures of spatial and temporal impairments. If the orthogonal scheme is essential, then the required timing information can be extracted from the scalar data sequences that describe the original and degraded video sequences. In the case of m_1, these sequences are $std_{space}(Sobel(O_n))$ and $std_{space}(Sobel(D_n))$.

6. Model Fit

Here we consider the goodness-of-fit of the third-order linear model. Figure 1 is a scatter plot showing the relationship between true scores and those of the model. If the model provided a perfect explanation of why each of the 46 video sequences received their corresponding

subjective scores, then each of the 46 data points would lie on the line y=x. In fact, the data points are spread out in a band about that line and the width of that spread is a measure of the goodness-of-fit of the model. Notice that the model fit is significantly better at the high end of the true score scale than at the low end of the scale. This third-order linear model explains 89.5% of the variance of the true score sequence. Equivalently, the coefficient of correlation between the model scores and the true scores is $[.895]^{1/2}=.946$. It is a consequence of the least squares solution that the errors between the true scores and the modeled scores have zero mean. Their average magnitude is .320 impairment units and their standard deviation is .388 impairment units:

$$e_i = s_i - \hat{s}_i \ , \quad 0 = \frac{1}{46}\sum_{i=1}^{46} e_i \ , \quad .320 = \frac{1}{46}\sum_{i=1}^{46} |e_i| \ , \quad .388 = \sqrt{\frac{1}{46}\sum_{i=1}^{46} e_i^2} \ .$$

As in most areas of science and engineering, the video impairment modeling problem embodies an inherent trade-off between complexity and performance. If an increase in model error is acceptable, the model complexity can be reduced. Figure 2 shows the performance of the first-order linear model:

$$\hat{s} = 4.4275 - 5.1427 \cdot m_1 \ .$$

Here the data points spread farther from the ideal. The errors still have zero mean, but their average magnitude has increased to .428 impairment units and their standard deviation is .492 impairment units. The model explains only 83.1% of the true score variance, with the corresponding correlation coefficient of .912.

7. Model Fit versus Predictor Performance

We have presented two fairly simple mathematical models for our observations of the relationships between objective measurements of video impairments and subjective evaluations of video impairments. We can interpret the third-order linear model as a mapping from three objective measurement variables to a single subjective score variable. The value of this mapping is that it allows us to predict subjective scores when none are available. Any claim that these posterior models possess predictive power would require the critical assumption that the 46 data points form a large and broad enough set to statistically characterize any other data points that we may wish to predict. We refer to this as the generality assumption.

In particular, we would like to claim that average *model fitting* error magnitude over the 46 video sequences is indicative of the average *prediction* error magnitude we would get if using the model to predict subjective impairment values for other video sequences. That is, we would like to claim that by measuring m_1, m_2, m_3, and forming the linear combination defined by c_1, c_2, c_3, c_4, we can predict true subjective scores for arbitrary VTC/VT type video sequences with an average error of zero, an average absolute error of .320 and an error standard deviation of .388. In order to make this claim, we would have to claim that the 46 video sequences statistically characterize all possible VTC/VT video scenes and impairments. This seems unlikely.

We are going to considerable lengths to come as close as possible to satisfying the generality assumption. In general, characterizing the infinite with the finite data is not easy. But by making that finite set as large and diverse as possible, we can come closer to characterizing all video impairments. The diversity requirement led us to use 12 scenes, 19 codec configurations, and some analog channels as well. Clearly a test using a single scene or a single coding technology will not generalize as well as one using multiple scenes and multiple coding technologies. The size requirement led us to approach the modeling problem incrementally; as more data points become available, we will use them to improve our model. We envision that further testing will confirm that we can use a third-order linear model, similar to the one presented here, to predict subjective impairment scores for general VTC/VT imagery and transmission service channels and suffer average prediction error magnitudes on the order of .5 impairment units.

8. Summary

This contribution has provided an update on ITS research that is applicable to the objective evaluation of VTC/VT transmission service channels. The results are both encouraging and practical. The video acquisition, frame processing, and mathematical computations required can all be performed in real time using off-the-shelf, dedicated digital signal processing cards that reside in an ordinary personal computer. The next phase of this work will add 100 additional data points to the existing 46. This will allow us to improve our model, strengthen and test its predictive power, and enhance its generality. Recommendations to the T1Q1.5 VTC/VT sub-working group for inclusion of these results into the selection process specified in the draft VTC/VT standard will be forthcoming. We encourage interested parties to analyze, critique, reproduce, and refine our results. It will be the collaborative efforts of the video measurement community that will equip the emerging standard with practical, effective measurements.

The research described here is being conducted at The Institute for Telecommunication Sciences in Boulder, Colorado by Stephen Wolf, Arthur Webster, Margaret Pinson, Coleen Jones, and Paul King.

Figure 1: Third-Order Linear Model

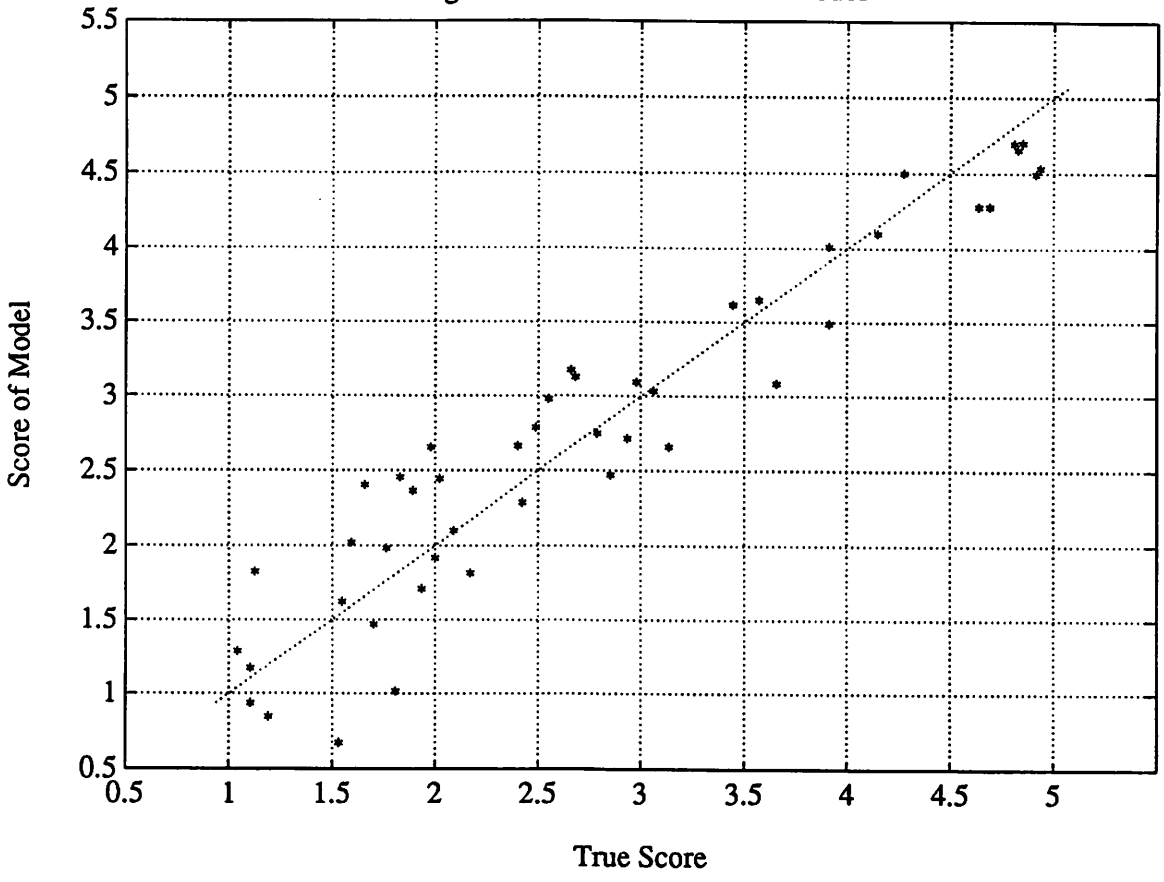

Figure 2: First-Order Linear Model