Measuring the End-to-End Performance of Digital Video Systems

Stephen Wolf
Institute for Telecommunication Sciences (ITS)
325 Broadway
Boulder, CO 80303

Abstract

Significant research and development efforts by industry and government laboratories were brought to fruition in 1996 with the approval of American National Standard (ANSI) T1.801.03 entitled "American National Standard for Telecommunications - Digital Transport of One-Way Video Signals - Parameters for Objective Performance Assessment." This standard provides a set of objective parameters that have consistently demonstrated high correlation levels with subjective evaluations of digital video impairments. The parameters are technology-independent and may be used to measure the performance of a wide range of digital video compression, storage, and transmission systems. This paper presents an overview of the ANSI T1.801.03 parameters and summarizes other relevant standards activities and contributions.

I. Performance measurement issues

A. Input scene dependencies

The advent of video compression, storage, and transmission systems has exposed fundamental limitations of techniques and methodologies that have traditionally been used to measure video performance. Traditional performance parameters have relied on the "constancy" of a video system's performance for different input scenes. Thus, one could inject a test pattern or test signal (e.g., a static multiburst), measure some resulting system attribute (e.g., frequency response), and be relatively confident that the system would respond similarly for other video material (e.g., video with motion).[1] A great deal of research has been performed to relate the traditional analog video performance parameters (e.g., differential gain, differential phase, short time waveform distortion, etc.) to perceived changes in video quality [1, 2, 3]. While the recent advent of video compression, storage, and transmission systems has not invalidated these traditional parameters, it has certainly made their connection with perceived video quality much more tenuous. Digital video systems adapt and change their behavior depending upon the input scene. Therefore, attempts to use input scenes that are different from what is actually used "in-service" can result in erroneous and misleading results. Variations in subjective performance ratings as large as 3 quality units on a subjective quality scale that runs from 1 to 5 (1=lowest rating, 5=highest rating) have been noted in tests of commercially available systems. While quality dependencies on the input scene tend to become much more prevalent at higher compression ratios, they also are observed at lower compression ratios. For example see [4], where subjective test results of 45-Mb/s contribution quality systems (i.e., systems used to transmit high quality video from studio to studio) revealed one system whose subjective performance varied from 2.16 to 4.64 quality units.

A digital video transmission system that works fine for video teleconferencing might be completely inadequate for entertainment television. Specifying the performance of a digital video system as a function of the video scene coding difficulty yields a much more complete description of system performance. Recognizing the need to select appropriate input scenes for testing, algorithms have been developed for quantifying the expected coding difficulty of an input scene based on the amount of spatial detail and motion [5, Annex A of 6]. Other methods have been proposed for determining the picture-content failure characteristic for the system under consideration [Appendices 1 and 2 to Annex 1 of 7]. National and international standards have been developed that specify standard video scenes for testing digital video systems [8, 9]. Use of these standards assures that users compare apples to apples when evaluating systems from different suppliers.

B. New digital video impairments

Digital video systems produce fundamentally different kinds of impairments than analog video systems. Examples of these include tiling, error blocks, smearing, jerkiness, edge busyness, and object retention [10]. To fully quantify the performance characteristics of a digital video system, it is desirable to have a set of performance parameters, where each parameter is sensitive to some unique dimension of video quality or impairment type. This is similar to what was developed for analog impairments (e.g., a

[1] The subjective, or user-perceived, quality of analog video systems can also depend upon the scene content. For example, a fixed analog noise level may be less objectionable for some scenes than others.

multiburst test would measure the frequency response, and a signal-to-noise ratio test would measure the analog noise level). This discrimination property of performance parameters is useful to designers trying to optimize certain system attributes over others, and to network operators wanting to know not only when a system is failing but where and how it is failing.

Also of interest is how a user weights the different performance attributes of a digital video system (e.g., spatial resolution, temporal resolution, or color reproduction accuracy) when subjectively rating the quality of the experience. The process of estimating these subjective quality ratings from objective performance parameter data is an important new area of work that will be discussed later in this paper.

C. The need for technology independence

The constancy of analog video systems over the past 4 decades provided the necessary long term development cycle to produce today's accurate analog video test equipment. In contrast, the rapid evolution of digital video compression, storage, and transmission technology presents a much more difficult performance measurement task. To avoid immediate obsolescence, new performance measurement technology developed for digital video systems must be technology independent, or not dependent upon specific coding algorithms or transport architectures. One way to achieve technology independence is to have the test instrument perceive and measure video impairments like a human being. Fortunately, the computational resources needed to achieve these measurement operations are now available.

II. A new measurement methodology - ANSI T1.801.03-1996

The above issues have necessitated the development of a new measurement methodology for testing the performance of digital video systems. Rather than being limited to artificial test signals, this methodology is one that can use natural video scenes. The methodology can also be extended to include nonintrusive, in-service performance monitoring, making it useful for applications such as fault detection, automatic quality monitoring, and dynamic optimization of limited network resources. Figure 1 presents the reference model for measuring end-to-end video performance parameters and summarizes the principles of the new measurement methodology detailed in ANSI T1.801.03, "American National Standard for Telecommunications - Digital Transport of One-Way Video Telephony Signals - Parameters for Objective Performance Assessment" [11]. This standard specifies a framework for measuring end-to-end performance parameters that are sensitive to distortions introduced by the coder, the digital channel, or the decoder shown in Figure 1.

Features, or specific characteristics associated with individual video frames, are extracted in quantity from both the input and output video streams by performance measurement systems. These performance measurement systems digitize the video signals in accordance with ITU-R Recommendation BT.601 [12] and extract features from these digitized frames of video. The extracted features quantify fundamental perceptual attributes of the video signal such as spatial and temporal detail. Parameters are calculated using comparison functions that operate on two parallel sequences of these feature samples (one sequence from the output video frames and a corresponding sequence from the input video frames). The standard contains parameters derived from three types of features that have proven useful: (1) scalar features, where the information associated with a specified video frame is represented by a scalar; (2) vector features, where the information associated with a specified video frame is represented by a vector of related numbers; and (3) matrix features, where the information associated with a specified video frame is represented by a matrix of related numbers.

In general, the transmission and storage requirements for measuring an objective parameter based on scalar features is less than that required for an objective parameter based on vector features. This, in turn, is less than that required for an objective parameter based on matrix features. Significantly, scalar-based parameters have produced some of the highest correlations to subjective quality! This demonstrates that the amount of reference information required to perform meaningful quality measurements is much less than the entire video frame. The whole output video frame need not be compared to the corresponding input video frame. One only needs to compare features extracted from the output video frame to those same features extracted from the input video frame.

This important new idea for performing video quality measurements has significant advantages, particularly for such applications as long-term maintenance and monitoring of network performance. Since a historical record of the output scalar features requires very little storage, they may be efficiently archived for future reference. Then, changes in the digital video system over time can be detected by simply comparing these past historical records with current output feature values. Another significant advantage of this approach is that performance measurements can be made in-service since an ancillary low bandwidth data channel (e.g., vertical interval, modem connection) can be used to transmit the extracted feature information from the input and output video streams.

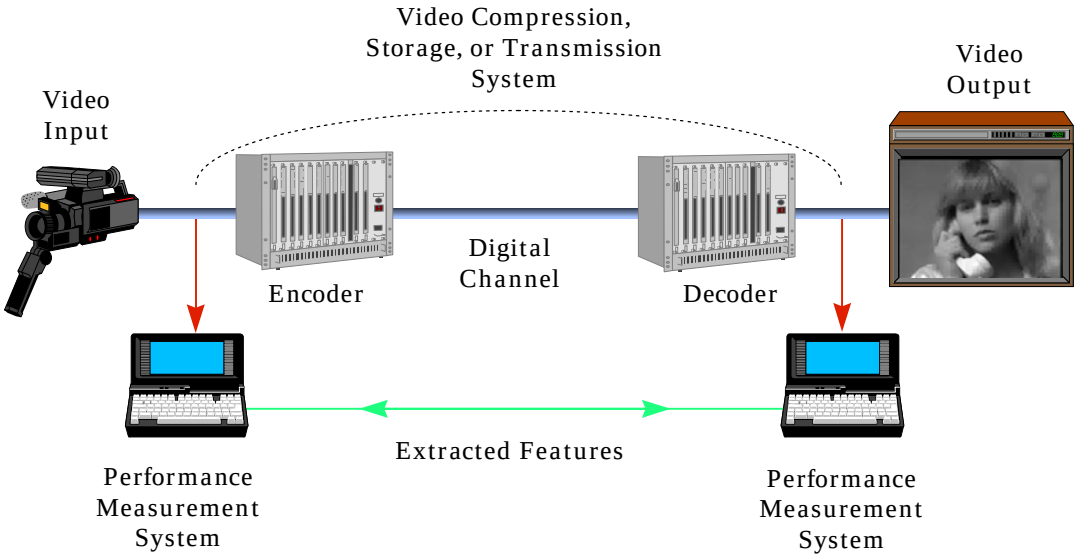


Figure 1. ANSI T1.801.03 reference model for measuring video performance.

III. Example scalar features

A complete description of all the features and parameters in ANSI T1.801.03 is beyond the scope of this paper. However, two examples from the scalar class of features will be presented for illustration purposes. The first is scalar features based on statistics of spatial gradients in the vicinity of image pixels. These spatial statistics are indicators of the amount and type of spatial information, or edges, in the video scene. The second is scalar features based on the statistics of temporal changes to the image pixels. These temporal statistics are indicators of the amount and type of temporal information, or motion, in the video scene from one frame to the next.

A. Spatial information (SI) features

Figure 2 demonstrates the process used to extract spatial information (*SI*) features from a sampled video frame. A gradient or edge enhancement algorithm (i.e., Sobel filtering) is applied to the video frame. At each image pixel, two gradient operators are applied to enhance both vertical differences (i.e., horizontal edges) and horizontal differences (i.e., vertical edges). Thus, at each image pixel, one can obtain estimates of the magnitude and direction of the spatial gradient (the right-hand image in Figure 2 shows magnitude only). A statistic is then calculated on a selected subregion of the spatial gradient image to produce a scalar quantity. Examples of useful scalar features that can be computed from spatial gradient images include total root mean square energy (this spatial information feature is denoted as SI_{rms} in ANSI T1.801.03) , and total energy that is of magnitude greater than r_{min} and within $\Delta\theta$ radians of the horizontal and vertical directions (denoted as $HV(\Delta\theta, r_{min})$ in ANSI T1.801.03). Parameters for detecting and quantifying digital video impairments such as blurring, tiling, and edge busyness are measured using time histories of *SI* features.

B. Temporal information (TI) features

Figure 3 demonstrates the process used to extract temporal information (*TI*) features from a video frame sampled at time *n* (i.e., *frame n* in the figure). First, temporal gradients are calculated for each image pixel by subtracting, pixel by pixel, *frame n-1* (i.e., one frame earlier in time) from *frame n*. The right-hand image in Figure 3 shows the absolute magnitude of the temporal gradient and, in this case, the larger temporal gradients (white areas) are due to subject motion. A statistical process, calculated on a selected subregion of the temporal gradient image, is used to produce a scalar feature. An example of a useful scalar feature that can be computed from temporal gradient images is the total root mean square energy (this temporal information feature is denoted as TI_{rms} in ANSI T1.801.03). Parameters for detecting and quantifying digital video impairments such as jerkiness, quantization noise, and error blocks are measured using time histories of temporal information features.

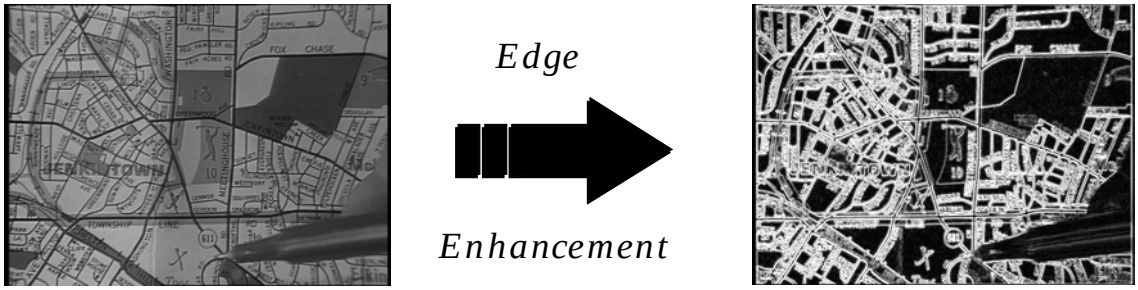


Figure 2. Example spatial information features.

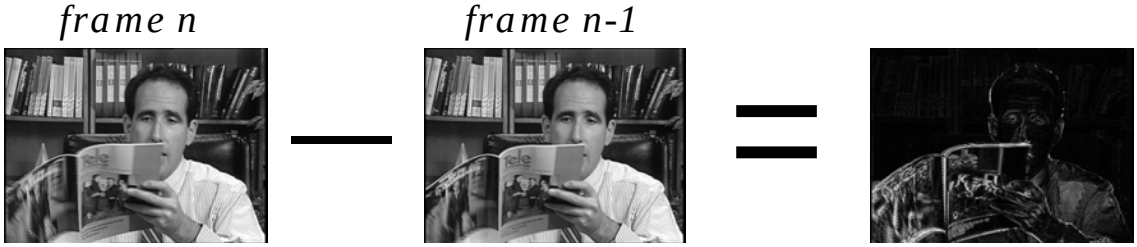


Figure 3. Example temporal information features.

IV. Example scalar parameters

The responses of two ANSI T1.801.03 scalar parameters to common digital video impairments are illustrated with pictorial examples and their accompanying explanations. These pictorial examples are intended to give the reader an intuitive understanding for what is being measured. The reader is directed to ANSI T1.801.03 for a detailed mathematical discourse and additional pictorial examples.

A. Maximum HV to non-HV edge energy difference

The images in Figure 4 illustrate the use of the maximum horizontal and vertical (*HV*) to non-horizontal and vertical (*non-HV*) edge energy difference parameter (section 7.1.9 of ANSI T1.801.03) for detecting tiling.[2] In Figure 4, the output image contains both tiling (i.e., block distortion) and blurring. Tiling creates false horizontal and vertical edges while blurring results in lost edge energy. By examining the spatial information as a function of angle, the tiling effects can be separated from the blurring effects.

To obtain a pictorial representation of this effect, the SI_h and SI_v values were calculated for each image pixel, where SI_h is the horizontal spatial gradient mentioned earlier and SI_v is the vertical spatial gradient. The plots in the third row were generated by counting the number of image pixels at each discrete coordinate (whose abscissa and ordinate values are given by SI_h and SI_v respectively), and then displaying this count as an intensity. Thus, brighter areas indicate more image pixels with those SI_h and SI_v values. The coordinates (SI_h , SI_v) also can be converted into radius (SI_r) and angle (θ) coordinates. As shown in the third row plots, the tiling adds horizontal and vertical spatial information (i.e., the output plot on the right has more spatial information along the horizontal SI_h axis and the vertical SI_v axis than the input plot on the left). The blurring results in a loss of diagonal spatial information (i.e., the output plot on the right has less spatial information along a diagonal direction, such as $\theta = 45$ degrees, than the input plot on the left).

[2] ANSI T1.801.02-1996 defines tiling as "Distortion of the image characterized by the appearance of an underlying block encoding structure."

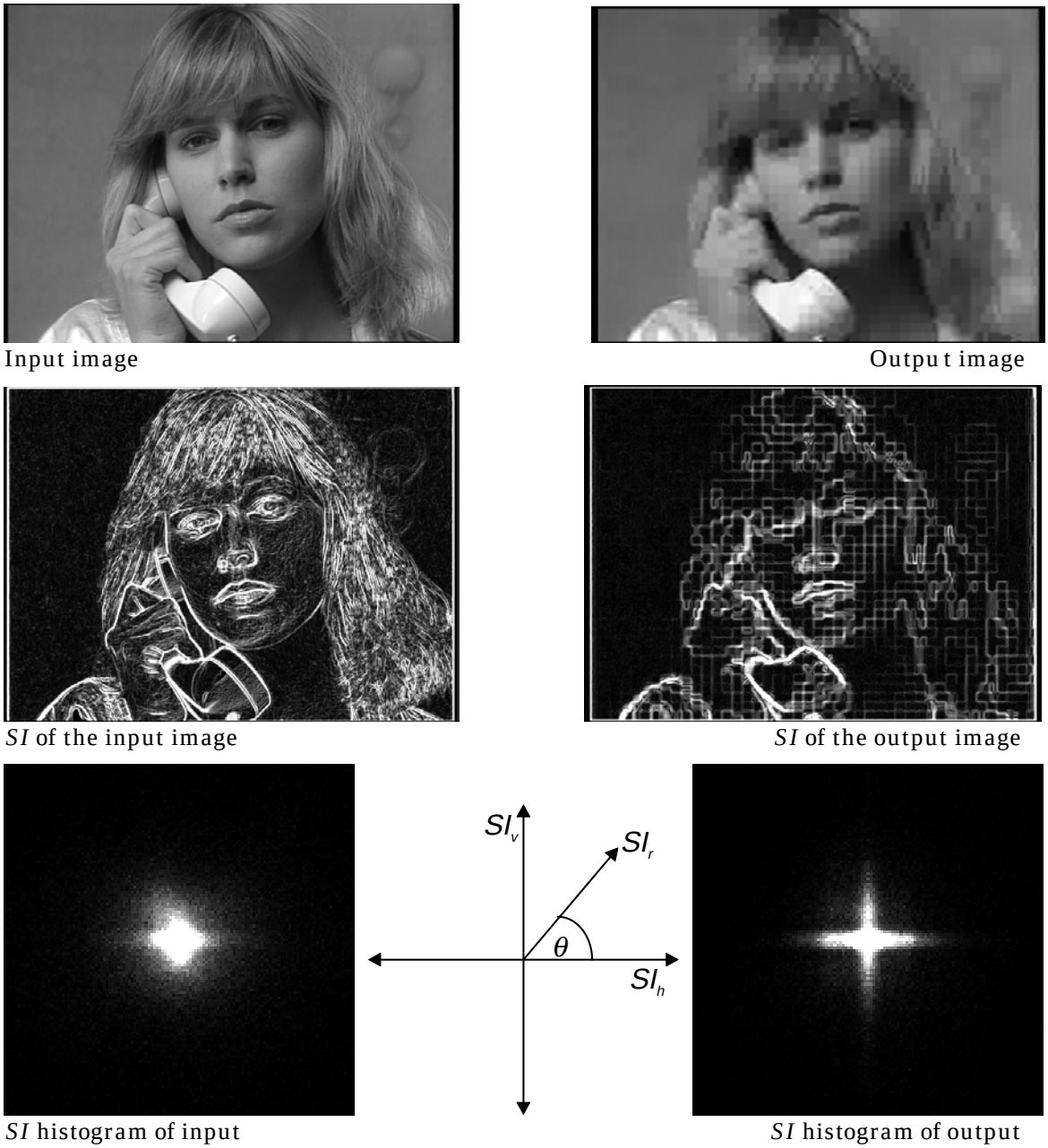


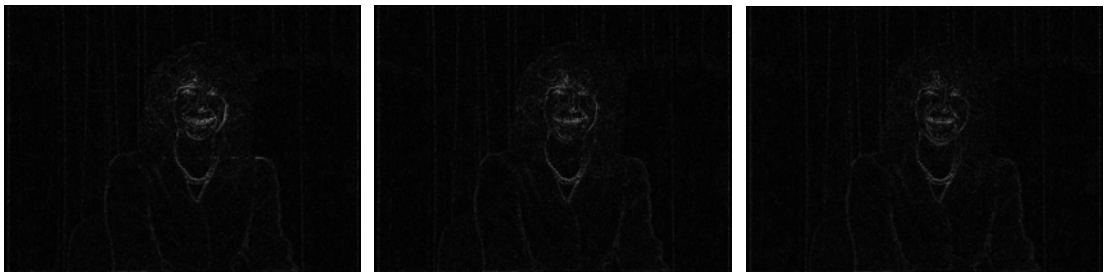
Figure 4. Example of maximum HV to non-HV edge energy difference.

B. Maximum added motion energy

The images in Figure 5 illustrate how the maximum added motion energy parameter (section 7.1.1 of ANSI T1.801.03) can be used to detect error blocks.[3] Three contiguous input and output images are displayed, together with their *TI* images. Increasing values of *TI*, or motion, are shown as whiter areas in the *TI* images. The sudden occurrence of the error blocks in the third output image produces a relatively large amount of added *TI*.

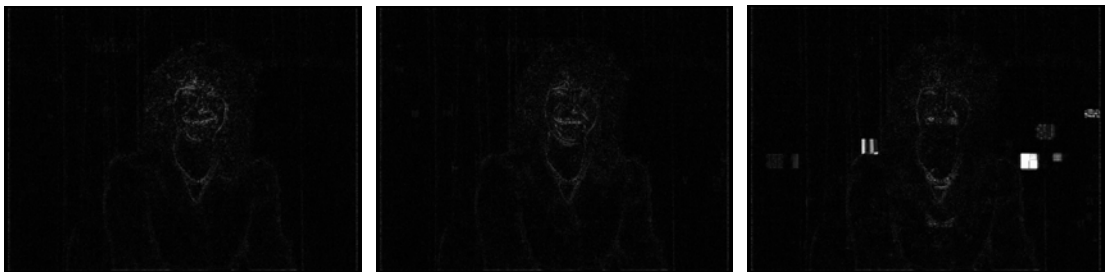
[3] ANSI T1.801.02-1996 defines error blocks as "A form of block distortion where one or more blocks in the image bear no resemblance to the current or previous scene and often contrast greatly with adjacent blocks."

Sequence of input images



TI of the sequence of input images

Sequence of output images



TI of the sequence of output images

Figure 5. Example of maximum added motion energy.

The graph in Figure 6 shows how the appearance and disappearance of an error block causes spikes, or sudden increases, in the *TI* values. The perceptibility of these error blocks is related to the logarithmic ratio of the output *TI* value divided by the input *TI* value. In other words, error blocks become more noticeable in low motion scenes than high motion scenes. Other impairments, such as jerkiness or noise, can also cause sudden increases in the *TI* values.

V. Relationships between ANSI T1.801.03 parameters and subjective quality

For any objective parameter to be useful, there must be some relationship between the objectively measured parameter values and corresponding subjective evaluations of video quality. Statistical analysis techniques such as coefficients of correlation and analysis of variance (ANOVA) provide the fundamental tools for determining how much and what portion of the total subjective variance can be explained by the objective parameters. The parameters in ANSI T1.801.03, when taken as a set, have consistently been able to explain from 75 to 90% of the subjective variance in data sets that cover a wide range of digital video compression and transmission technologies [4, 5, 13-16]. These data sets included coder/decoders (codecs) that were based on technologies such as the discrete cosine transform (DCT) and vector quantization (VQ), and ranged in bit

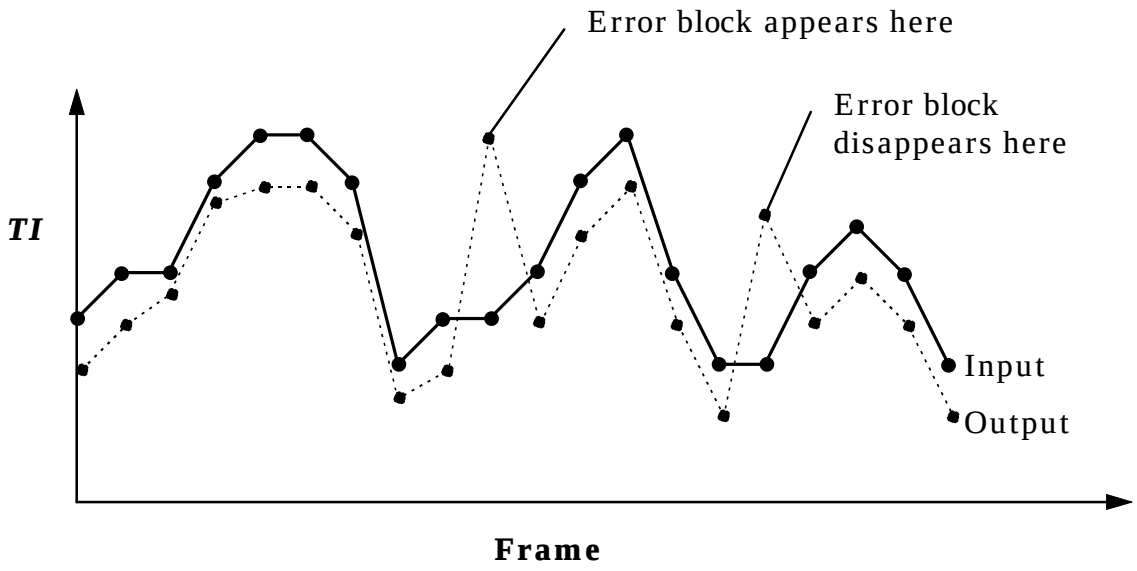


Figure 6. Example time history of TI_{rms} features.

rates from 56 Kb/sec to 45 Mb/sec. Also included were test conditions such as digital transmission errors (random and burst) and tandem connections. The video scenes were also chosen to span a wide range of spatial detail and temporal movement so that the systems would be subjected to various degrees of coding difficulty.

An obvious question is "How good should objective measures be to produce meaningful results?" The ultimate benchmark would be for objective measures to replace subjective experiments altogether. Although this has not yet been achieved by the current set of objective measures, substantial progress toward this goal has been made. Analysis of the repeatability of subjective tests (i.e., conducting the same subjective experiment in multiple laboratories or multiple times) has determined that between 8 and 15% of the variance in the subjective scores from well designed and executed experiments is due to random or unexplained sources [13, 16]. Objective measures cannot be expected to account for this portion of the subjective variance. Thus, the ANSI T1.801.03 parameters can measure system attributes that account for a substantial portion of the useable subjective video quality information.

VI. Future directions

A. Building objective video quality models

Building objective video quality models involves conducting simultaneous subjective and objective tests and determining how objective parameter values can be used to predict the subjective viewer responses. This model building process is necessary for determining the overall accuracy of the objective parameters and for identifying the portion of the subjective responses explained by the objective parameters. However, developing useful video quality models for operational systems is a much more complicated process. Two simple examples illustrate the complexity involved. For the first example, consider two different applications: transmission of high-resolution graphics imagery with pointer capability, and transmission of sign language. Here, these two fundamentally different applications require different performance characteristics for the various dimensions of video quality (e.g., spatial resolution, temporal resolution, or color reproduction accuracy). The graphics application requires very high spatial resolution with low frame rates while the sign language application requires high frame rates at a lower spatial resolution. An objective model that produces overall quality estimates from a set of fundamental objective parameters would have to account for these application-specific effects. For the second example, consider two different viewer populations: the naïve or non-expert viewer, and the critical expert viewer. In this case, the expert viewer will tend to downgrade the quality more than the naïve viewer for the same amount of video impairment. For an actual example that illustrates this viewer population effect, see [4]. Another influence on modeling accuracy is the changing expectations of people over time. This is particularly true for digital video systems where the technology is improving rapidly and the cost is decreasing rapidly.

For these reasons, objective video quality modeling is valid only if the application and viewer population are well defined. Given sufficient time and effort, the objective parameters presented in the ANSI T1.801.03

standard can be used to develop effective video quality models for a large number of video applications and viewer populations.

B. Setting guaranteed levels of service

Many users want assurances that they will receive some guaranteed level of service. This will involve the determination of appropriate thresholds for individual parameter values or video quality model outputs. Different thresholds could be used to define levels or grades of service (e.g., low, medium, high). If the thresholds are violated, the user will have a mechanism to resolve the problem. It is expected that the establishment of standard threshold levels will require a multiyear effort and cooperation by manufacturers, carriers, and users.

VII. Acknowledgments

The author would like to thank William Hughes, Coleen Jones, Dwight Melcher, Margaret Pinson, Stephen Voran, and Arthur Webster at the Institute for Telecommunication Sciences (ITS) for their contributions to this work.

VIII. References

[1] CCIR Recommendation 654, "Subjective quality of television pictures in relation to the main impairments of the analogue composite television signal," Recommendations and Reports of the CCIR, 1986.

[2] CCIR Report 405-5, "Subjective assessment of the quality of television pictures," Recommendations and Reports of the CCIR, 1986.

[3] CCIR Report 959-1, "Experimental results relating picture quality to objective magnitude of impairment," Recommendations and Reports of the CCIR, 1986.

[4] S. Wolf and A. Webster, "Objective and subjective video performance testing of DS3 rate transmission channels," ANSI T1A1 contribution number T1A1.5/93-060, Apr 1993.[4]

[5] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, S. Wolf, "An objective video quality assessment system based on human perception," SPIE Human Vision, Visual Processing, and Digital Display IV, vol. 1913, Feb 1993.

[6] ITU-T Draft Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Recommendations of the ITU (Telecommunication Standardization Sector).

[7] CCIR Recommendation 500-5, "Method for the subjective assessment of the quality of television pictures," Recommendations and Reports of the CCIR, 1992.

[8] ITU-R Recommendation BT.802-1, "Test pictures and sequences for subjective assessments of digital codecs conveying signals produced according to Recommendation ITU-R BT.601," Recommendations of the ITU (Radiocommunication Sector).

[9] ANSI T1.801.01-1995, "American National Standard for Telecommunications - Digital Transport of Video Teleconferencing/Video Telephony Signals - Video Test Scenes for Subjective and Objective Performance Assessment," Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington DC 20005.

[10] ANSI T1.801.02-1996, "American National Standard for Telecommunications - Digital Transport of Video Teleconferencing/Video Telephony Signals - Performance Terms, Definitions, and Examples," Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington DC 20005.

[11] ANSI T1.801.03-1996, "American National Standard for Telecommunications - Digital Transport of One-Way Video Telephony Signals - Parameters for Objective Performance Assessment," Alliance for Telecommunications Industry Solutions, 1200 G Street, N. W., Suite 500, Washington DC 20005.

[12] ITU-R Recommendation BT.601, "Encoding Parameters of Digital Television for Studios," Recommendations of the ITU (Radiocommunication Sector).

[13] G. W. Cermak and D. A. Fay, "Correlation of objective and subjective measures of video quality," ANSI T1A1 contribution number T1A1.5/94-148, Sep 1994.

[14] S. Wolf, "An in-depth analysis of the P6 lost motion energy parameter," ANSI T1A1 contribution number T1A1.5/95-103, Jan 1995.

[15] B. Cotton, "An objective model for video quality performance," ANSI T1A1 contribution number T1A1.5/96-105, Mar 1996.

[16] G. Cermak, P. Tweedy, S. Wolf, A. Webster, M. Pinson, "Objective and Subjective Measures of MPEG Video Quality," ANSI T1A1 contribution number T1A1.5/96-121, Oct 1996.

[4] Copies of ANSI T1A1 contributions can be obtained from the T1 Secretariat, Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington, DC 20005.