APPLICATION OF THE NTIA GENERAL VIDEO QUALITY METRIC (VQM) TO HDTV QUALITY MONITORING

Stephen Wolf and Margaret H. Pinson

National Telecommunications and Information Administration (NTIA)

ABSTRACT

This paper summarizes results from an experiment whose goal was to assess whether the NTIA General Video Quality Metric (VQM) is an acceptable objective metric for measuring High Definition TV (HDTV) video quality. The HDTV subjective test that was performed to evaluate the NTIA General VQM contained 60 30-second video clips that were rated using the Single Stimulus Continuous Quality Evaluation (SSCQE) method. The 60 clips included twelve 1080i HDTV originals and 48 processed versions of these originals from 16 different video systems. The video systems included 5 different HDTV codecs running at bit rates from 2 to 19 Mbps and broadcast transmission errors (i.e., RF transmission with poor signal-to-noise-ratio). Excellent objective-to-subjective correlation results for this experiment demonstrate the potential application of the NTIA General VQM to HDTV quality monitoring.

1. INTRODUCTION

The National Telecommunications and Information Administration (NTIA) developed the General Video Quality Metric (VQM) as a means for quantifying perceptual quality degradation in video systems that utilize compression. As a result of good correlations to subjective quality ratings in the Phase II validation tests performed by the Video Quality Experts Group (VQEG) [1], the NTIA General VQM was adopted as both a national standard and an international recommendation [2] [3] [4] [5]. The scope of these standards includes quantifying and comparing the quality of Standard Definition TV (SDTV) systems that utilize error-free digital transport, i.e., video systems that contain an encoder, an error-free transmission channel, and a decoder.

To assess the applicability of the NTIA General VQM for measuring the quality of High Definition TV (HDTV) systems, NTIA designed and conducted an HDTV subjective experiment. HDTV systems are different from SDTV systems in that they normally include the use of large, high resolution screens. While the viewing distance is closer (in terms of picture height), the spatial resolution is also higher so one has approximately the same number of pixels per degree of viewing angle. Thus, from a pure human visual system modeling standpoint, no adjustment to the objective model should be required.

The total horizontal viewing angle, however, is much larger (i.e., approximately 30 degrees for HDTV versus 12 degrees for SDTV), and this creates other potential differences that may influence quality decisions. Since the human visual system only achieves high spatial resolution over several angular degrees, the eye must roam the picture when looking at HDTV in order to track specific objects and their motion. Impairments that are present outside of the immediate attention of the viewer will be less visible than in SDTV systems. Such aspects of viewer attention are not normally included in current objective video quality models. The HDTV subjective experiment described here is the first attempt to quantify the applicability of the NTIA General Model to HDTV.

This paper is organized as follows. Section 2 describes the HDTV subjective test design, including a description of the scenes, video systems, and subjective viewing sessions. Section 3 discusses the subjective and objective data processing that was applied to the raw data, while Section 4 presents the objective-to-subjective correlation results. Finally, Section 5 summarizes the conclusions of the study.

2. SUBJECTIVE TEST DESIGN

The presence of coding artifacts and transmission errors was transitory in many of the HDTV systems that were examined. Single Stimulus Continuous Quality Evaluation (SSCQE) testing [6] was chosen to be able to track these time varying quality changes. In SSCQE testing, viewers move a quality slider (see Figure 1) in real time and the position of the slider is sampled several times per second.

2.1 Description of Scenes

The test scenes were drawn from a pool of uncompressed and mildly compressed material (compression ratios ranging from 4:1 to 10:1) shot in 1080i format (1920 x 1080 pixels). Twelve 30-second scenes were selected that spanned a wide range of coding difficulty (motion and detail), color, contrast, and brightness. While scene cuts were present within the individual 30-second clips, the scene content for most of the twelve scenes was similar throughout the entire 30-second period. Copyright restrictions prevent the inclusion of sample video frames for most of the scenes in this paper. The following is a brief description of the twelve test scenes:

1. People preparing for a scuba-diving mission on a tropical island and in a boat.
2. Real and computer-generated fish and underwater scenes.
3. Aerial views of different cities during the day, including camera pans.
4. Multiple city scenes at sunset/night as viewed from helicopters.
5. Horizontal and vertical pans of red tulip gardens.
6. Mix of nature scenes, including rippling water, a bird, a crab, a honeybee, and flowers.
7. Flyby of waterfalls with fades/scene cuts.
8. A farm tractor plowing a field and a combine harvesting a corn field. This scene contained pans, zooms, and scene cuts.
9. Horse race on green grass arena with pans and scene cuts.
10. A commercial for a wireless mouse.
11. People walking in the city, including a shot of a mime actor in the city square.
12. A collection of scenes shot in Stockholm, including a man pointing at shields, a calendar and toy train, and a man running along a river bank.

2.2 Description of Video Systems

A goal of the experimental design was to maximize the range of visually different stimuli, in order to best evaluate the NTIA General Model's performance under a variety of operating conditions. Sixteen HDTV video systems were considered in this experiment. Five different software codecs[1] were used to generate constant bit rate encoded bit-streams at rates ranging from 2 Mbits/sec to 19 Mbits/sec. The five encoders included:

1. DivX Pro™ version 5.2.0
2. Windows Media 9™ (WM9)
3. 3MB™ MPEG-2
4. TMPGEnc Plus™ 2.58.44.152 MPEG-2
5. MainConcept™ MPEG-2 that is bundled with Adobe Premiere Pro™ version 1.5.

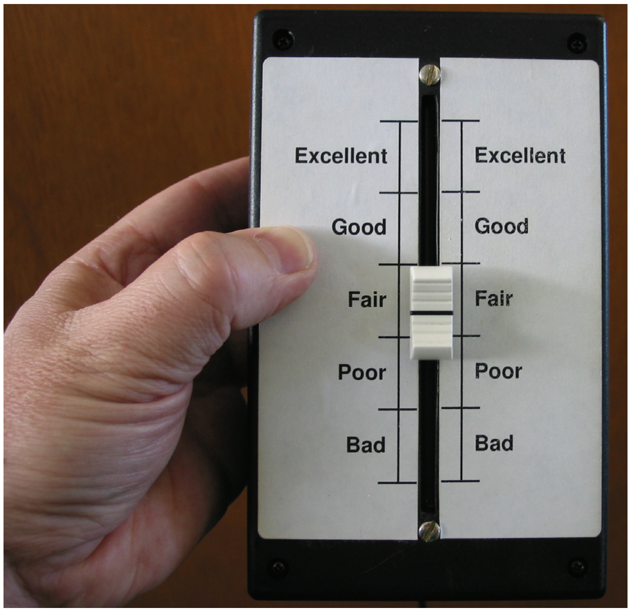

Figure 1. SSCQE Slider.

Lower bit rates (2-8 Mbits/sec) were paired with the newer DivX and WM9 encoders (codecs 1 and 2) while higher bit rates (6-19 Mbits/sec) were paired with the MPEG-2 encoders (codecs 3 through 5). Codecs 1 through 4 were operated at three different bit rates each, for a total of 12 video systems. Codec 5 (which could interface with the 8-VSB RF transmission hardware) generated the remaining 4 video systems, 2 of which had RF transmission errors. These two systems included Advanced Television Systems Committee (ATSC) 8 Vestigial Sideband (8-VSB) Radio Frequency (RF) modulation and transmission over a poor signal-to-noise ratio channel, which caused signal drop-outs in the decoded TV picture.

Scene/system pairs were chosen to establish a diverse range of impairments, instead of using the traditional full matrix design. A roughly periodic sampling of the scene/system matrix was used, where scenes were ordered from difficult to easy (encoding complexity), and systems were ordered from low quality to high quality. Each of the 16 video systems was paired with approximately 3 scenes such that (1) each scene was matched with exactly 4 video systems, and (2) each codec (over all the bit rates)

[1] Certain commercial equipment and material are identified in this paper to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendation or endorsement by NTIA, nor does it imply that the material or equipment identified is the best available for this purpose.

was matched with at least 8 scenes. Each scene appeared a total of 5 times (the original plus 4 processed versions). Altogether, the test contained 48 processed clips, 6 of which contained transmission errors. These 48 processed clips, together with the 12 original clips, resulted in 60 clips, for a total of 30 minutes of viewing material.

2.3 Subjective Viewing Sessions

Two 30-minute HDTV test tapes (in the Panasonic HD-D5 tape format) were generated such that each tape had a unique clip randomization. Clips were randomized within each 30-minute test tape such that the same scene or video system was never consecutively presented. Ten viewers rated each of the two test tapes (where each viewer rated only one test tape), for a total of 20 unique viewers per clip. Video clips were presented to the viewers on a high-end 50-inch HDTV plasma screen with a native resolution of 1366 x 768 pixels and a viewing distance of 3 times picture height.

The viewers used the SSCQE subjective test method to rate each test tape. The viewers' instructions included the following text: "The quality of the video that you will see may change rapidly and span a range of quality from excellent to bad. During the presentation, you are encouraged to move the indicator along the scale as soon as you notice a change in the quality of the video. The indicator should always be at the point on the scale that currently corresponds to your most accurate judgment of the presentation. You are allowed to move the indicator to any point on the scale."

The slider position was encoded using amplitude modulation of an audio test tone. This enabled the slider waveform to be synchronously sampled as a stereo pair together with the Society for Motion Picture and Television Engineers (SMPTE) Time Code (TC) from the viewing tape, which was also available as an audio waveform. These two audio waveforms were synchronously sampled at a rate of 11.025 kHz using a PC audio capture card. This sampling rate was sufficient to decode the amplitude modulation of the SSCQE waveforms and the SMPTE TC. In this manner, each SSCQE sample could be directly related to presentation frames on the viewing tape.

Before each viewing session, a slider calibration waveform was generated by moving the slider in Figure 1 to the bottom (bad) and top (excellent) of the quality scale, and these reference points were used to assign values of 0 and 100, respectively, to the SSCQE waveforms. A program was used to extract and calibrate the SSCQE waveforms between the beginning and ending SMPTE TCs of the viewing tapes. The program returned a sampling rate of 2 samples per second for the final calibrated SSCQE waveforms.

3. DATA PROCESSING

3.1 Subjective Data Alignment

For each of the two viewing tapes, the SSCQE waveforms from the 10 viewers were time aligned to account for the variation in viewer reaction times. The time alignment process allowed a maximum time shift of plus or minus 5 samples (2.5 seconds) between viewers. A cross-correlation process produced a 10 x 10 matrix, where element *ij* provided the optimal time shift of viewer *i* with respect to viewer *j*. The viewer with the smallest total correlation shift (summed over all viewers) was made the reference viewer and the other 9 viewers were time aligned to this reference viewer. Reference viewers selected in this manner resulted in a very low average time shift for each viewing tape, so time alignment between the two tapes was not an issue. SSCQE viewer waveforms with a non-zero time shift were extrapolated by replicating the first or last SSCQE sample.

3.2 Subjective Data Conversion

The NTIA General VQM was designed to measure the perceptual difference in quality between original and processed video clips of 8 to 10 seconds in duration. The subjective testing methodologies that were used to develop the NTIA General VQM included the Double Stimulus Continuous Quality Scale (DSCQS), the Double Stimulus Comparison Scale (DSCS), and the Double Stimulus Impairment Scale (DSIS) [6]. In these double stimulus methods, the viewer is always shown the original and processed video clips and the subjective score is either computed as the difference in quality between the original and processed video clips (where each is rated separately) or the viewer rates the quality difference directly.

In 2003, NTIA performed a series of subjective experiments that related SSCQE with Hidden Reference Removal (SSCQE-HRR) to double stimulus methods [7]. In SSCQE-HRR, the reference video sequences are presented during the test session, but viewers are not aware that they are evaluating the reference video. The viewer's opinion of the reference video sequence is subtracted from the viewer's opinion of the impaired video sequence. It was shown that SSCQE-HRR provides time varying quality assessments that are highly correlated to those obtained by double stimulus testing using short 8-10 second clips provided (1) the SSCQE-HRR sample at the end of the corresponding 8-10 second video clip is used and (2) at least two clip randomizations are used. The current HDTV subjective experiment was designed to meet both of these requirements. Thus, with appropriate data processing, SSCQE-HRR subjective test data is used to evaluate the performance of the NTIA General VQM.

SSCQE-HRR waveforms for each viewer and 30-second scene were obtained by computing

$$U = 100 - (original - processed).$$

Since each SSCQE *original* and *processed* opinion is in the range [0, 100], the difference is in the range [-100, 100]. Adding one hundred to this difference shifts the range to [0, 200]. Here, 0 is the worst quality, 100 is the same quality as the reference, and values greater than 100 indicate quality better than the reference. SSCQE-HRR scores produced in this manner may occasionally be greater than 100 when the original reference is scored by a viewer to be of lower quality than the processed. SSCQE-HRR scores greater than 100 are generally limited to the first several seconds of the video scene (i.e., viewers seem to require about 6 to 8 seconds to move the slider to the proper position after a scene transition from a low quality scene to a reference high quality scene). For our data, this occurred about 6% of the time (when the first 9.5 seconds of each video scene are disregarded, to allow the SSCQE-HRR trace to stabilize). To prevent SSCQE-HRR viewer scores greater than 100 from unduly influencing the mean SSCQE-HRR trace, a crushing function of the following form was applied:

$$C = \frac{120*U}{20+U} \text{ if } U > 100,$$

where U is the uncrushed score, and C is the crushed score.

The SSCQE-HRR traces from all viewers and randomizations were averaged to compute a final SSCQE-HRR Mean Opinion Score (MOS) trace for each of the 48 processed video clips. SSCQE-HRR MOS samples were extracted at times 10, 20, and 30 seconds into each processed video clip, to correspond to the subjective ratings that would have been obtained on 10-second video segments from times 0-10, 10-20, and 20-30, respectively. This resulted in 48*3 = 144 discrete MOS samples.

3.3 Objective Data

NTIA General VQM software [8] was used to produce objective scores for the 144 10-second video clips. Video calibration was only necessary for processed clips obtained from the two video systems with 8-VSB RF transmission errors. These two video systems included hardware components (e.g., 8-VSB RF modulator, broadcast MPEG-2 decoder) that introduced horizontal spatial shifts and gain/level offset errors into the processed video. The rest of the video systems were composed of software components that did not introduce any video calibration errors.

4. CORRELATION RESULTS

Figure 2 presents a scatter plot of the subjective SSCQE-HRR MOS results versus the NTIA General VQM for the 144 10-second clips. The NTIA General VQM scores are reported on a nominal range of [0, 1], where zero indicates excellent quality. The Pearson correlation coefficient between the two data sets is 0.84 and the Root Mean Square (RMS) error between the best fit line (shown in red) and the subjective data (on the 0 to 100) scale is 9.7. Processed video clips that included transmission errors are shown with red asterisks.

Figure 3 presents a scatter plot of the subjective SSCQE-HRR MOS results versus the NTIA General VQM for the 16 video systems. For this plot, scores are obtained by averaging (over scenes) the subjective and objective data for each video system. The Pearson correlation coefficient between the two data sets is 0.91 and the RMS error between the best fit line (shown in red) and the subjective data (on the 0 to 100 scale) is 5.0. Video systems that included transmission errors are shown with red asterisks.

5. CONCLUSIONS

The NTIA General VQM has been shown to be highly correlated to subjective ratings of processed video clips from an HDTV experiment that included a fairly wide range of codecs, bit rates, and even some transmission errors. When assessing average video system quality using several different scenes, the correlation results were even more encouraging.

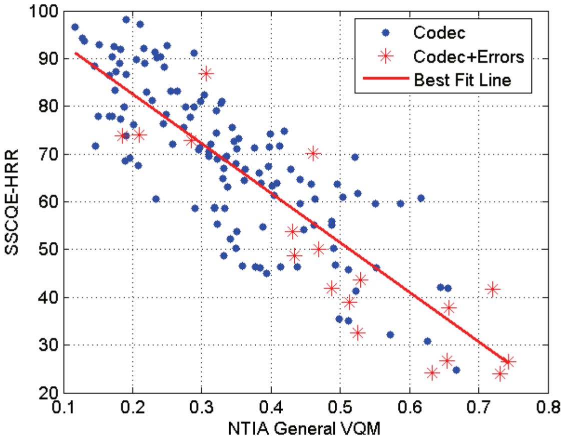

Figure 2. Clip results.

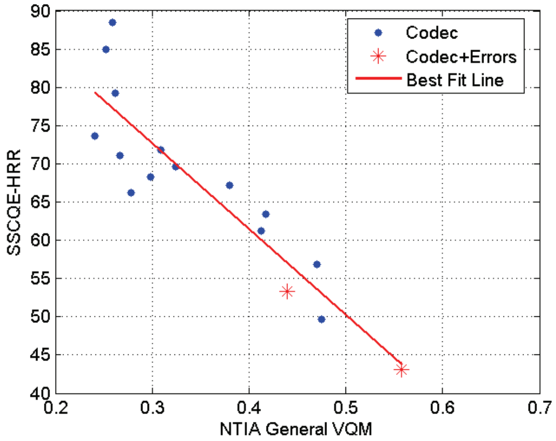

Figure 3. Video system results.

6. REFERENCES

[1] VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, Phase II," August 25, 2003. Available at www.vqeg.org.

[2] ANSI T1.801.03 – 2003, "American National Standard for Telecommunications – Digital transport of one-way video signals – Parameters for objective performance assessment," American National Standards Institute. Available at www.ansi.org.

[3] ITU-T Recommendation J.144 (2004), "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," Recommendations of the ITU, Telecommunication Standardization Sector. Available at www.itu.org.

[4] ITU-R Recommendation BT.1683 (2004), "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference," Recommendations of the ITU, Radiocommunication Sector. Available at www.itu.org.

[5] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, v. 50, n. 3, pp. 312-322, Sep. 2004.

[6] ITU-R Recommendation BT.500-11 (2002), "Methodology for the subjective assessment of the quality of television pictures," Recommendations of the ITU, Radiocommunication Sector. Available at www.itu.org.

[7] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," *Proc. of SPIE Video Communications and Image Processing Conference*, Lugano, Switzerland, Jul. 2003.

[8] NTIA General Video Quality Metric (VQM) Software, available at http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm.