Video Clarity
Tools for Video Analysis

1566 La Pradera Dr
Campbell, CA 95008
www.videoclarity.com
408-379-6952

White Paper: Understanding MOS, JND, and PSNR

Video Clarity, Inc.

The term “video quality” remains poorly defined even when it seems that it shouldn’t be. We could, for example, try to assess the quality of a video program based on the artistic choices made with the camera work, but that wouldn’t be terribly useful. Rather, our goal is to provide a tool that is useful for troubleshooting a network or testing the ability of a compression algorithm to produce an appealing stream of pictures and sound. To that end, we are going to define “quality” in terms of fidelity; that is, how closely does a processed or delivered signal match the original source (or reference) signal? Our main concern will be to detect and quantify any distortions that have been introduced into the signal as it passes through a network or device.

Quality measurement starts with a simple concept: we must judge video quality in a consistent way regardless of the type of distortion.

Defining Video Quality

Video quality consists of three major components:

- Picture Quality – an index of the eye’s ability to understand a picture
- Audio Quality – an index of the ear’s ability to discern audio
- Lip Sync – a measurement of the audio to video synchronization

This paper will focus on picture quality and ultimately Subjective Testing is the only proven way to evaluate picture quality. Unfortunately, this mode of testing is very expensive, time-consuming, and often impractical. One popular method is Absolute Category Rating (ACR), wherein human subjects are shown two video sequences (original and processed) and are asked to assess the overall quality of the processed sequence with respect to the original (reference) sequence. The test can be divided into multiple sessions and, if so, each session should not last more than 30 minutes. For every session, several dummy sequences are added, which are used to train the human subjects and are not included in the final score. The subjects score the processed video sequence on a scale (usually 5 or 9) corresponding to their mental measure of the quality – this is termed Mean Opinion Score (MOS).

When the MOS score is on a 1 to 5 scale, the scores are

1 Unacceptable
2 Poor
3 Fair
4 Good
5 Excellent

The results can, of course, vary from test to test, but if the pool of human participants is large enough (16 or more), the scores tend to stabilize.

Types of Errors

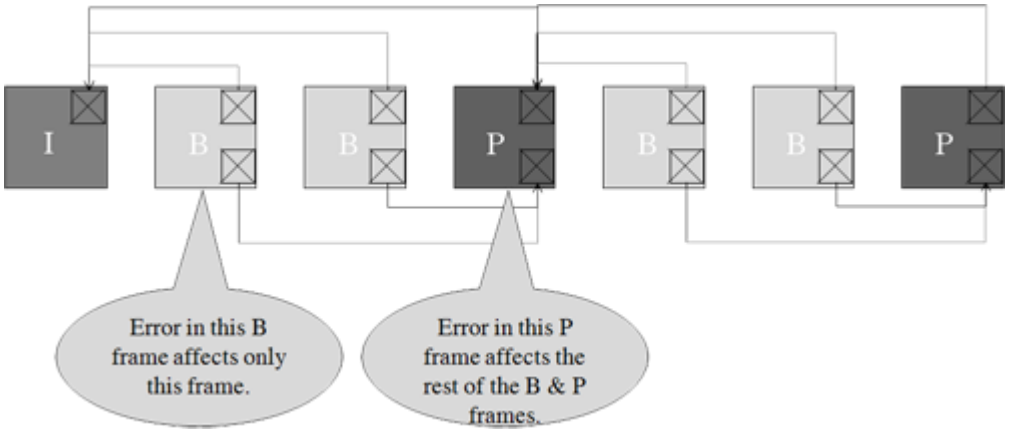
There are two leading sources of problems for digital television:

- The digital transmission path can fall below acceptable levels and cause a complete loss – i.e. no picture and no audio.
- The amount and quality of the compression can lend itself to poor quality.

Checking a digital transmission path for errors is fairly straightforward. It can be accomplished simply by sending a known signal through the path and verifying that the received signal is a bit-for-bit match.

Many video CODECs use a Group of Pictures (GoP) frame structure, which consists of independently coded reference frames (“I” frames), motion changes from the last reference frame

("P" frames) and motion changes from the last reference or next reference frame ("B" frames). If a transmission error occurs, the type of frame lost determines how many other frames are affected. If the compression is too extreme, blocky or blurry images will result.



Most audio CODECs detect high frequency components and encode these with very few bits because the human ear can only hear loud high frequencies. Some algorithms reduce the dynamic range to reduce the amount of data. If a transmission error occurs, the audio will often pop or go silent. If the compression is too extreme, the audio will lack depth – it will sound tinny or hollow.

Perceptual/Objective Quality Testing

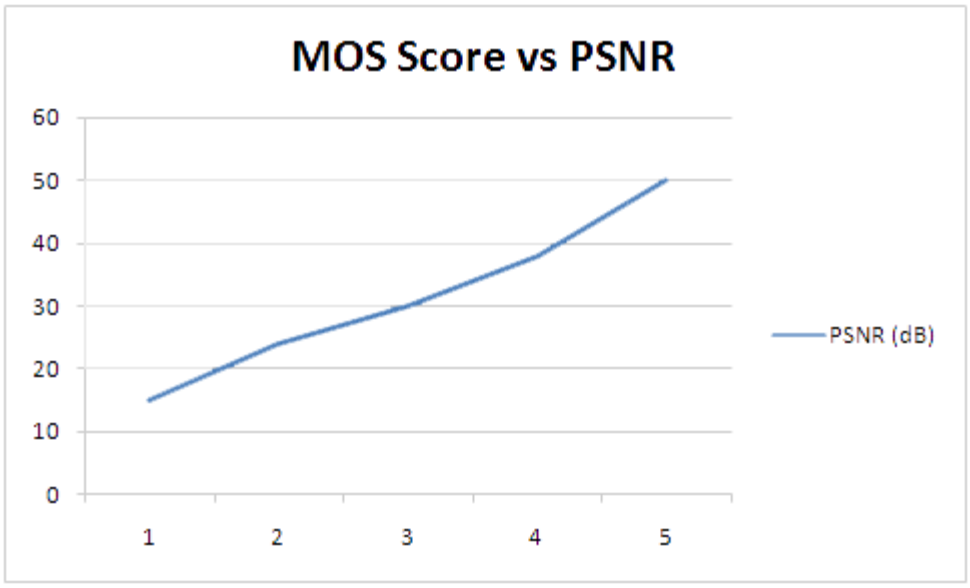
A number of algorithms have been developed to estimate video quality by using mathematical analysis in place of human observers. These algorithms are then fit to the subjective data, which ideally reflects an objective way to measure subjective quality. The algorithms are divided into three general types:

- Full reference algorithms compare the output video stream to the input (or reference) stream.
- No reference algorithms analyze only the output stream.
- Reduced reference algorithms extract specific information from the input stream and use it when analyzing the output stream.

For this paper, we will confine our discussion to full reference algorithms.

To start, the two streams ("reference" and "processed") must be aligned both temporally and spatially. Audio and Video synchronization issues can be detected at this point. Regardless, of whether the audio and video are in-sync or not, both signals can be further analyzed.

The most widely used metrics are PSNR (Peak Signal-to-Noise Ratio) or MSE (Mean Squared Error). Both measure the mean error between input and output. PSNR expresses the result as a ratio of the peak signal expressed in dB. PSNR and MSE are not highly accurate video quality predictors especially with today's video processing techniques, but they do serve an important role. Unlike the indices soon to be discussed, PSNR and MSE are metrics. They measure the absolute difference between two signals, which is completely quantifiable. This is very important in QA and Monitoring where the perceived quality has already been measured in the laboratory environment and what is needed is PASS/FAIL indicator. A PSNR value of 35dB is generally considered good. A general comparison of PSNR to MOS is shown below.



Traditional perceptual video quality index methods are based on a bottom-up approach which attempts to simulate the functionality of the relevant human visual system (HVS) and human audio systems (HAS) components. These methods usually involve:

- Capturing two signals – reference and processed
- Video/Audio alignment
- Calculating the differences that affect the human eye/ear
    - Blockiness
    - Blurriness
    - Lack of Dynamic Range
    - Loss of High Frequencies.
- Classify the types of distortions and adding up the scores
- Scaling the resulting score to correspond to a Subjective MOS

While these bottom-up approaches can conveniently make use of many known psychophysical features of the HVS/HAS, it is important to recognize their limitations. In particular, the HVS and HAS are highly non-linear systems and natural images/sounds are very complex. Most models are based on linear or quasi-linear operators that have been characterized using restricted and simplistic stimuli. Some models that fit into this category are listed below:

- Sarnoff/PQR – First Widely Heralded HVS Metric
- VQM – Video Quality Metric
- PEVQ – Perceptual Evaluation of Video Quality
- PEAQ – Perceptual Evaluation of Audio Quality

The Structural Similarity (SSIM) approach provides an alternative and complementary way to tackle the problem of video quality assessment. It is based on a top-down assumption that the HVS is highly adapted for extracting structural information from the scene, and therefore a measure of structural similarity should be a good approximation of perceived image quality. The eye can recognize a shape even if part of it is missing. It has been shown that a simple implementation of SSIM outperforms state-of-the-art perceptual image quality metrics. However, the SSIM index achieves the best performance when applied at an appropriate scale (i.e. viewer distance/screen height). Calibrating the parameters, such as viewing distance and picture resolution, create the most challenges of this approach. To rectify this, multi-scale, structure similarity (MS-SSIM) has

been defined. In MS-SSIM, the picture is evaluated at various resolutions and the result is an average of these calibrated steps. It has been shown that MS-SSIM out-performs simple SSIM even when the SSIM is correctly calibrated to the environment and dataset. Additional information on the significant advantages of MS-SSIM can be found in a separate Video Clarity published paper http://www.videoclarity.com/WPAdvancingToMulti-ScaleSSIM.html.

In either case, the model produces a score and then needs to be correlated with the subjective MOS. Two methods are available for this when using a ClearView analyzer:

- Differential Mean Opinion Score (DMOS index) with MS-SSIM algorithm
- Just Noticeable Differences (JND index) with Sarnoff/PQR algorithm

DMOS is the difference between "reference" and "processed" Mean Opinion Score in a full reference test system. DMOS can be presented in a 0 – 4 scale as well as a lowest score value of 7 or 10 in a ClearView analyzer. A general guideline for DMOS scoring is as follows.

3.1 - 4	Most Users Dissatisfied
2.1 - 3.0	Many Users Dissatisfied
1.1 - 2.0	Some Users Satisfied
0.7 - 1.0	Most Users Satisfied
0 - .6	Very Satisfied

JND is based on the principle that slight differences between a pair of images are imperceptible to viewers, and that the amount of change required to produce a noticeable difference can be quantified. This change can come in various forms, such as a change in the sharpness of an image, the appearance of blocks in the image, or other similar distortions. To quantify JND, viewers are shown pairs of images, and asked to identify which one is the original, undistorted image and which has been degraded in some way. For image pairs that have imperceptible differences, the viewers will typically split their votes 50/50 between the images, yielding a JND score of 0. For image pairs that have a "just noticeable" difference, the votes will be split 75/25 in favor of correctly identifying the distorted image; this level of distortion will be assigned a JND sore of 1.

To create more steps in the JND scale, comparison results are "stacked" to extend the scale. To do this, images that have a JND score of 1 are now used as a reference, and then distorted once again to produce a 75/25 correct vote. The new distorted image is given a JND score of 2, since it is "just noticeably" different from the JND 1 image. This process is then iterated, with the JND 2 image used as a reference to create an image that has a JND score of 3, and so on. An entire scale with multiple steps can be created using this method, with an image that has a JND score of, say, 7 having four levels of noticeable difference from an image with a JND score of 3.

Comparing the JND Scale to the DMOS scale, shows the following correlation based on our observation.

DMOS	JND	Description
	13+	Probably not aligned check Spatial and Temporal Alignment
3.5000 – 4.0	10 - 12.99	Unwatchable
3.0000 – 3.4999	7 - 9.99	Annoying
0.4000 – 2.9999	2 - 6.99	Broadcast Quality
0.0001 – 0.3999	0.01- 1.99	Production Quality
0	0	Perfect Quality

Subjective Data

The most important item to remember is that lossy, compressed signals have distortions. To understand quality, we must correlate metrics or indices to subjective MOS data. To this end, we must have an open, searchable database of subjective data.

The VQEG (Video Quality Experts Group) created a large database of video. They compressed these using H.263, H.264, and MPEG-2 and conducted subjective tests. These databases are open to member companies, but are not royalty free.

The University of Texas started with a collection of royalty free videos from the Technical University of Munich and distorted these in several ways using MPEG-2 and H.264 and transmitted them over IP networks. They then conducted subjective tests and compiled the results in a royalty-free database called LIVE.

The Sarnoff/PQR and MS-SSIM algorithms are further discussed on our website at www.videoclarity.com/WhitePapers.html.

Video Clarity Solutions

Video Clarity currently manufactures two principal product lines:

- ClearView AV Analyzers
- RTM (Real Time Monitor)
- ClearView Player/Recorders

ClearView Video AV analyzers generate test signals, capture live inputs, and input compressed or uncompressed files. They then align the audio and video automatically and report video and audio perceptual scores with DMOS, JND and PEAQ indices. ClearView also calculates performance metrics such as PSNR for video and aFREQ for audio performance and lip-sync. For the perceptual video measurements the systems use the Sarnoff/PQR algorithm ported to JND (using the VQEG database) and the MS-SSIM algorithm ported to DMOS (using the University of Texas database). It also lets you view the “reference” and “processed” signals side-by-side or their difference maps for your own subjective evaluation.

RTM captures two live inputs, aligns the audio and video signals, reports lip-sync issues, calculates the absolute difference between the two inputs (metric), continually reports the quality score, generates a pass/fail, and saves failures for further offline analysis.

ClearView Player/Recorders provide a reliable solution for preparing uncompressed video recordings or files with audio and VANC for playback in a repeated loop or from a play list for testing purposes.