Video Clarity

Tools for Video Analysis

Understanding MOS, JND, and PSNR

Executive Summary

The term "video quality" remains poorly defined even in cases when it seems that it shouldn't be. 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 distortions created by different types of processing applied.

General Video Quality Defined

We are using video quality to define three major components:

- Picture Quality – an index of eyes ability to understand the picture
- Audio Quality – an index of the ears ability to discern the audio
- Lip Sync – a measurement of the audio to video synchronization

We are also going to define 2 terms:

- Metric – an algorithm that quantifies differences
- Index – an algorithm that measures quality using the Human Visual or Audio System (HVS/HAS)

Ultimately, there is only one proven way to evaluate video quality and that is Subjective Testing. However, this is very expensive, time-consuming, and often impractical. The main subjective quality methods are Degradation Category Rating (DCR), Pair Comparison (PC) and Absolute Category Rating (ACR). The human subjects are shown 2 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 is divided into multiple sessions and 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 is large enough (16 or more), the scores tend to stabilize.

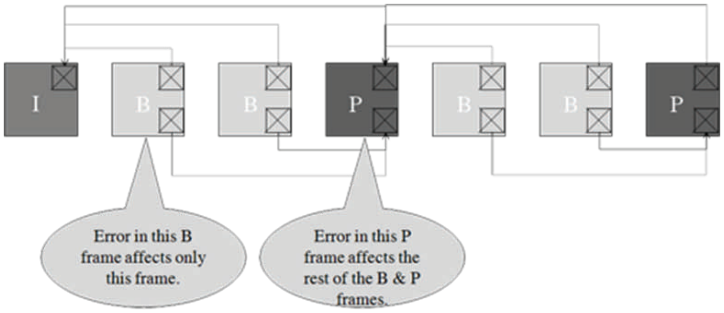
Types of Errors

Two types of problems can arise with 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 digital transmission paths for errors is fairly straight forward. Sending a known signal and verifying that the received path 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 the propagation time of the error. 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 pop or go silent. If the compression is too extreme, the audio will lack depth – i.e. sound tinny or hollow.

Objective Testing

A number of algorithms have been developed to estimate video quality. 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 its 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.

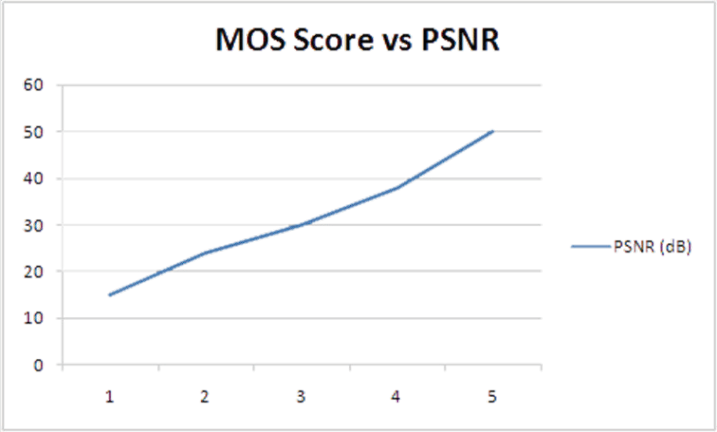
For this paper, we will confine our discussion to full reference algorithms.

At first, 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, 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 early human visual system (HVS) and human audio systems (HAS) components. These methods usually involve

- Video/Audio alignment
- Low pass filtering (to simulate the eye – video only)
- 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



- Applying this Score to the 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 complex and highly non-linear systems and the complexity of natural images/sounds are also very significant, but 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 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 structural similarity (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 outperforms simple SSIM even when the SSIM is correctly calibrated to the environment and dataset.

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

- Differential Mean Option 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 testing.

3.1-4.0	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

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 in the above table.

JND reports how many users need to be put into a room before 1 person thinks that the "reference" video quality is better and 1 person thinks that the "processed" video quality is better. The score is written as NumberOfPeople = $2^{(JND+1)}$. This method is the foundation for T1.TR.75.2001 ("Objective Perceptual Video Quality Measurement Using a JND-Based Full Reference Technique").

JND Score	Experts	Description
0	2	The Experts Disagree on which is better
1	4	3 Experts pick the Reference and 1 picks the Processed
2	8	7 Experts pick the Reference and 1 picks the Processed
3	16	15 Experts pick the Reference and 1 picks the Processed
4	32	31 Experts pick the Reference and 1 picks the Processed
8	512	511 Experts pick the Reference and 1 picks the Processed
12	8192	8191 Experts pick the Reference and 1 picks the Processed

DMOS	JND	Description
	13+	Check the Spatial and Temporal Alignment
3.5-4.0	10-12.99	Most Observers Dissatisfied
3.0-3.49	7-9.99	Many Observers Dissatisfied
0.4-2.99	2-6.99	Broadcast Quality
0.0001-0.39	0.01-1.99	Production Quality
0	0	Perfect Quality

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

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 many ways using MPEG-2 and H.264 and transmitted them over IP networks. They have since conducted additional subjective tests and compiled the results in a royalty -free database called LIVE.

Video Clarity Solutions

Video Clarity currently manufactures three product lines, the ClearView Video Quality Analyzers, RTM (Real Time Monitor-Recorders), and Venue Players with the ClearView Player-Recorder.

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, or VMAF indices. ClearView also calculates performance metrics such as PSNR for video and aFREQ for audio performance and lip-sync.

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.

Venue Players with ClearView Player-Recorder 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 and reliable venue playback purposes.

The Sarnoff PQR/JND and MS-SSIM algorithms are further discussed on the Video Clarity website at:

https://videoclarity.com/videoqualityanalysiscasestudies/