Video Clarity

Tools for Video Analysis

# Video Quality Testing for the IP and Internet Domains


## Executive Summary

As broadcasting and content distribution move ever farther into the Internet domain, satellite, broadcast, cable, and IPTV network technologists, system architects, and engineers must continually test new codec technologies to ensure that distribution setups are primed to deliver the best possible video quality at multiple bit rates, frame rates, and resolutions. Reference-based video quality testing — whereby the quality of the output is measured against the source content — is still the most accurate way to arrive at an idea of human-perceived quality for deployment of video services. But for IP- and Internet-based delivery methods, the delivered content may need deinterlacing and require the network to apply both full and reduced resolutions that are lower than actual TV screen resolutions and those of the source material. Scaling the source to match the downstream video poses a problem when attempting to compare IP-encoded video to its source material in order to make proper numerical quality measurements. At the same time, no content provider makes a decision about quality based solely on numerical ratings. They also do a visual comparison of the downstream video to the original, and weigh the visual comparison against the numerical results.

In the midst of that testing landscape, adaptive delivery systems that automatically adjust for network conditions and target devices are increasing the required number of profiles that content distributors must be prepared to deliver. Creating those profiles requires testing many permutations, which can be daunting without the right set of test tools.

This paper addresses the requirements and challenges of testing video quality that is processed for multiple resolutions on adaptive-bit-rate networks in the IP domain, and explores a reference-based test method that will satisfy most content originators and media entertainment delivery service providers.


## Introduction

With the proliferation in the amount of programming and the number of devices and screen sizes today, testing the output of various video delivery protocols and equipment is a crucial step in deciding which combination yields the best downstream video quality — and therefore which systems to invest in. It's a never-ending process for satellite, broadcast, cable, and IPTV network technologists, system architects, and engineers, because techniques, equipment, standards, and formats are always evolving.

Full-reference testing is the most accurate method for assessing changes in video quality between the source video and the downstream version. The basic idea is to take a short video clip, send it through a system to be tested, and then compare the system's output to the original (See Figure 1). If the system causes any differences in any of the video frames, those differences can be measured using a variety of objective metrics to yield numerical results — results that are tied to accepted databases derived from standardized human-vision-based studies. In other words, the number will tell you, based on the scale you're using, how close the processed signal is to the original, and the results will very closely approximate what human viewers would judge the quality to be when viewing the same video sequence on their own screens.

**Full-reference testing is the most accurate method for assessing changes in video quality between the source video and the downstream version.**

Today most traditional television signals are maintained at specific resolutions and frame rates from the start of production through transmission to the end viewer. This consistency makes reference-based testing in the traditional television domain relatively easy to do because you can use the original, unprocessed or minimally processed source video as the basis for comparison.

However, performing reference-based testing in the IP domain is not so simple. The advent of IPTV and internet-enabled devices means the video coming out of the system will often be delivered at a lower resolution and frame rate than the source material in order to accommodate equipment other than televisions.

To complicate the matter further, it is becoming more and more economical to deploy delivery systems that adapt to the conditions of the network and the requirements of the end

Frame 6– Encoder B Reference 480x270 @ 15Mbps

Frame 6 – Encoder B Test 480x270 @ 350kbps

**Figure 1: Encoder B Reference @ 15Mbps vs. Encoder B Test @ 350kbps**

devices. These adaptive systems make the testing process even more complex because it means that content providers must be prepared to deliver multiple profiles (resolutions, bit rates, frame rates) for every asset, with understood levels of quality for every instance of the delivery chain and end-device type.

In this scenario, full-reference testing is still the best way to assess video quality. The method applies broadly, not just to functional testing of an adaptive-bit-rate (ABR) service, but to lab testing that helps determine the optimal combination of bit rates, frame rates, resolutions, and equipment for the various profiles. (For example, full-reference testing can be especially useful when trying to decide the quality of the new HEVC encoders.)

**Challenges arise when de-interlacing with tools that do not match the deinterlacing and scaling algorithms in the encoder under test.**

### The Challenge of Full Reference Testing of IP Content

The basis for reference quality testing is that the source and after-processed sequences are required to be at the same resolution and frame rate to create and maintain an accurate measurement of the video quality as tests are created at multiple downstream profiles. But this is harder to do for IP content than for traditional television content because you have to reduce the resolution/frame rate of the source material, which is most often HD or UHD, in order to create an acceptable reference. Only then does it match for comparison to the encoded/transcoded material that the end user will see.

Many operators have opted to offer a multi-profile adaptive streaming service in which they perfect a fixed set of delivery profiles that will satisfy most of the network delivery characteristics and target devices at any given time — as devices call for adaptation to variances in network conditions. For example, some services might have at the ready three low-bit-rate delivery profiles for cell phones, three in the mid range for tablets, and two high-end profiles for televisions. The idea is to provide multiple profiles so that any end device can adapt its playback for any change in network conditions. All of the six lower-resolution profiles may be created from an original UHD, 1080i, or 720p source. In order to test those profiles, you must compare the source video to the possibly deinterlaced and/or lower-resolution video that will be sent by the encoder/transcoder and also what will ultimately come out of the device on the viewer's end.

To create the content streams for IP delivery, you must scale (and, in the case of a 1080i source, deinterlace) the original video for each profile. And in order to test those streams, you must create similarly processed streams to use as references during the test.

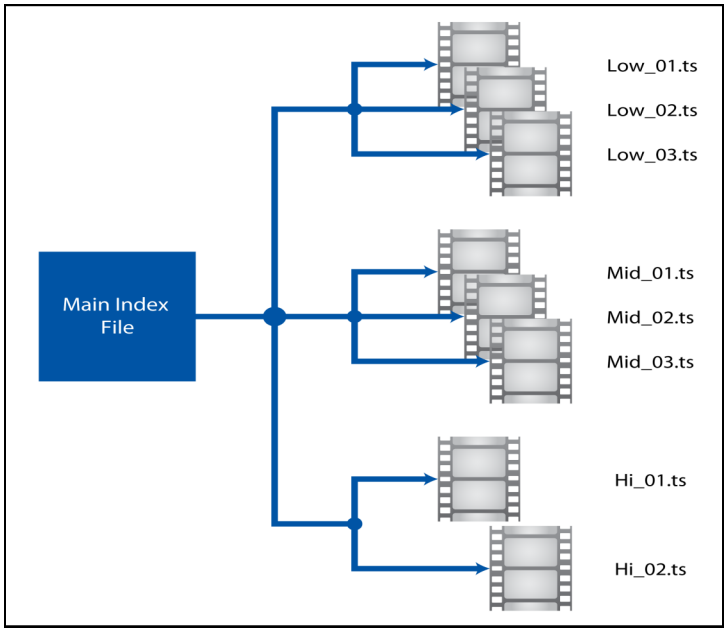



**Figure 2: An Example Number of Streams Per Program**

### Using Multiple Tools Skews Measurements

Challenges arise when deinterlacing and/or scaling the video with separate tools that do not match the deinterlacing and scaling algorithms in the encoder under test. Because the target encoder — and in the case of ABR services, the transcoder — has its own set of processes, the practice of using separate deinterlacing and scaling tools on the front end can create additional artifacts that exacerbate the differences between the reference material and the downstream output, which could result in a lower score on the chosen video quality index.

### An Alternative Method for Reference Sequence Preparation

In ABR testing the reference sequence must be altered to match downstream image characteristics and resolutions. This may be done with technologies external to the processing device that may be determined to be of equal or higher quality process. An alternative method is based on the same concept of deinterlacing and/or reducing resolutions to match profiles, but more importantly, using the same transcoder to create the reference video sequence so as to minimize the differences in artifacts and/or degradations that may result. This method is a proven, mathematically accurate means of objectively comparing different profiles and equipment.

In full-reference testing, the idea is always to use the most pristine reference possible in order to get a true sense of the quality differences in downstream versions based on the metric and the scoring index that you're using. When testing IP streams, an effective way to arrive at a pristine reference is to use the same transcoder to create the reference as is used to create the downstream deliverables. When applying the target processing device to match the test profile's video resolution at the highest possible bit rate, it minimizes the differences created in picture artifacts when scaling and deinterlacing an image. While any process applied may compromise the video quality of the source slightly, this method insures that the reference and the ABR versions are being deinterlaced and/or scaled according to the same algorithms, and it provides a quality measurement that is as true as possible to the quality index that's being used.

From this point it's recommended to use the same measurement system and scoring index and throughout the ABR stack — such as the Multi-Scale Structural Similarity (MS-SSIM) algorithm and the Differential Mean Opinion Score (DMOS) scale and/or VMAF on its native scale. This will allow comparisons of video quality scores between profiles and allow test thresholds to be set within the same index for different profiles as testing progresses. After testing each profile a video engineer with a trained eye then considers the resulting score while visually comparing the reference and the processed stream side by side.

Operators have already employed this methodology using these generalized steps:

1. Generate a mezzanine-quality reference for each profile using the highest possible bit rate and optimal encoding parameters.
2. Generate each profile test signal using the application's encoding parameters.
3. Calculate quality using the MS-SSIM on DMOS scale , comparing each test profile signal to the appropriate reference profile signal.
4. Analyze results on DMOS scale against visual comparisons of source and downstream profile test segments.

The process is repeated for each stream that needs testing, creating a new version for each profile or device under test, and making incremental adjustments to the variables one at a time in order to test the effect on quality. Given the number of combinations of encoders, bit rates, frame rates, and resolutions in a lab setting — compounded by the fact that there are multiple types of source video — operators need to create a significant number of references and processed test streams.

The first steps in the process — creating the various references and setting up the tests — are manageable, but repeating the test manually for every permutation and then visually comparing the results is a very involved process that can be very time consuming for one person or even a lab full of people to complete in any sort of constructive time frame.

**An alternative method is based on the same concept of deinterlacing and reducing resolutions to match profiles, but more importantly, using the same transcoder to create the reference video sequence so as to minimize the artifacts and/or degradations that may result.**

To manage the testing phase, the methodology calls for an automated tool to measure video quality in the many different combinations of bit rates, frame rates, and lower-than-broadcast resolutions, one that relies on an algorithm that yields numerical results based on human-perception scales. At the same time, the solution automatically creates measurement charts and synchronized, side-by-side picture comparisons that let experienced video engineers view the differences between the reference and the processed video — because even with the most accurate human-vision estimation metrics available today, the decisions still come down to a visual check of the video content.

### Video Clarity's ClearView Video Analysis Solution

Video Clarity's ClearView video quality analyzer records the references created at high bit rates and at the target resolutions from the device under test. It has several automatic frame-based alignment modes to match them with the corresponding profiles created by that same device for testing. This process can be done manually or automatically in a scripted routine to align, measure, and log quality metric results.

Figure 3 shows a simplified view of this process.

A video source feeds an uncompressed or lightly compressed signal to the encoder's input. The video sources could come from captured files or from any number of other sources, such as live IP video streaming from multiple different types of pro-

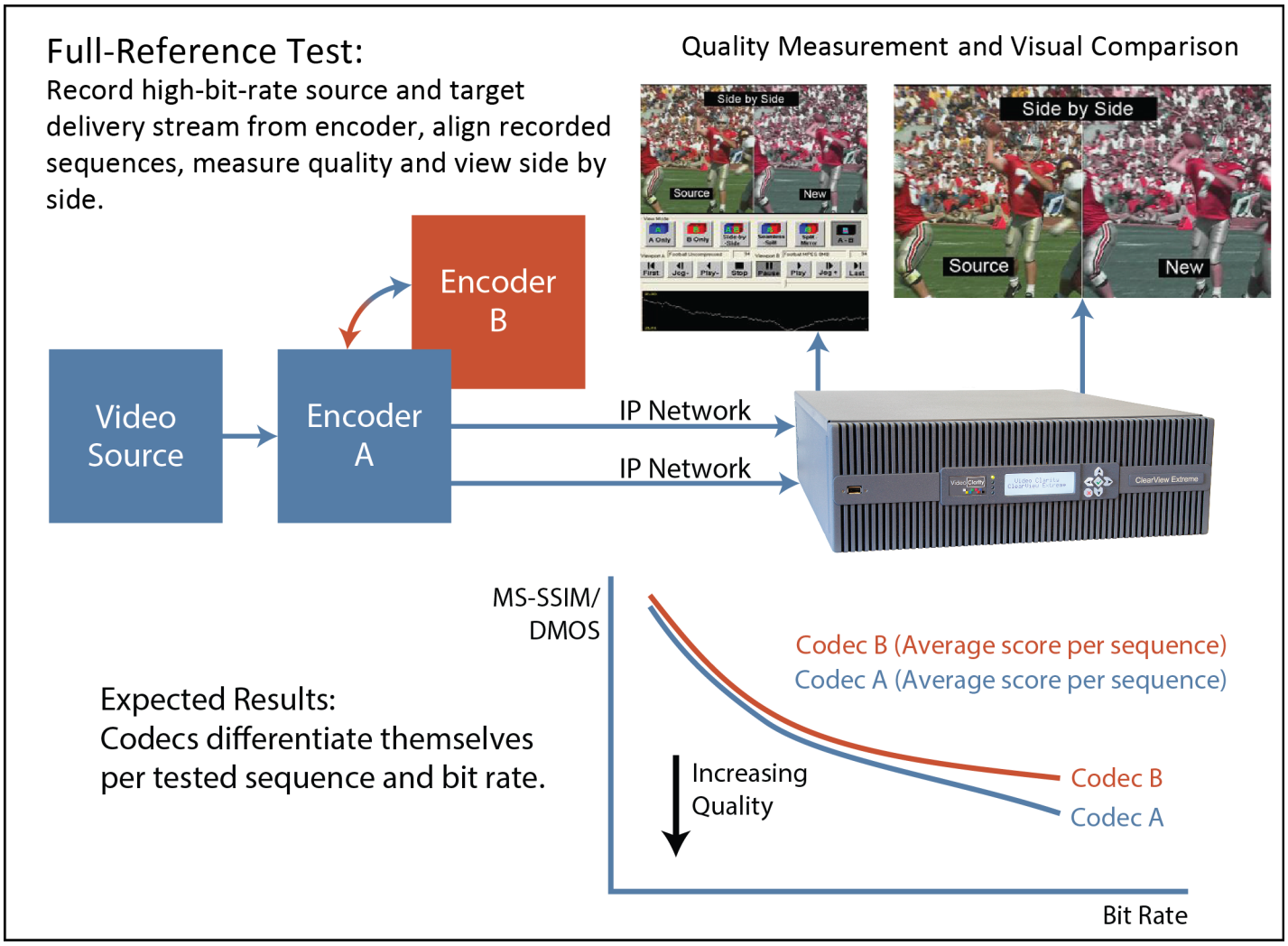



**Figure 3: Full-reference testing with ClearView provides measurement results and a side-by-side view of source and target delivery video.**

gram content.

It is recommended that content sources are compiled with different temporal and spatial complexity characteristics so that the video tests are derived from multiple content types (e.g., sports action, a moving crowd shot, flashing lights, high contrast and color saturation, low motion, and animation). This will provide a comprehensive content set to compare against the additional variables of multiple transcoded bit rates and other settings and/or different transcoding devices. The idea is to envelope the quality differences created by the content itself to fully test the quality performance of the transcoder and possibly exercise the network path all the way through to the endpoint streaming device (e.g., tablet, handheld, or set-top output, *a step not shown in this diagram*).

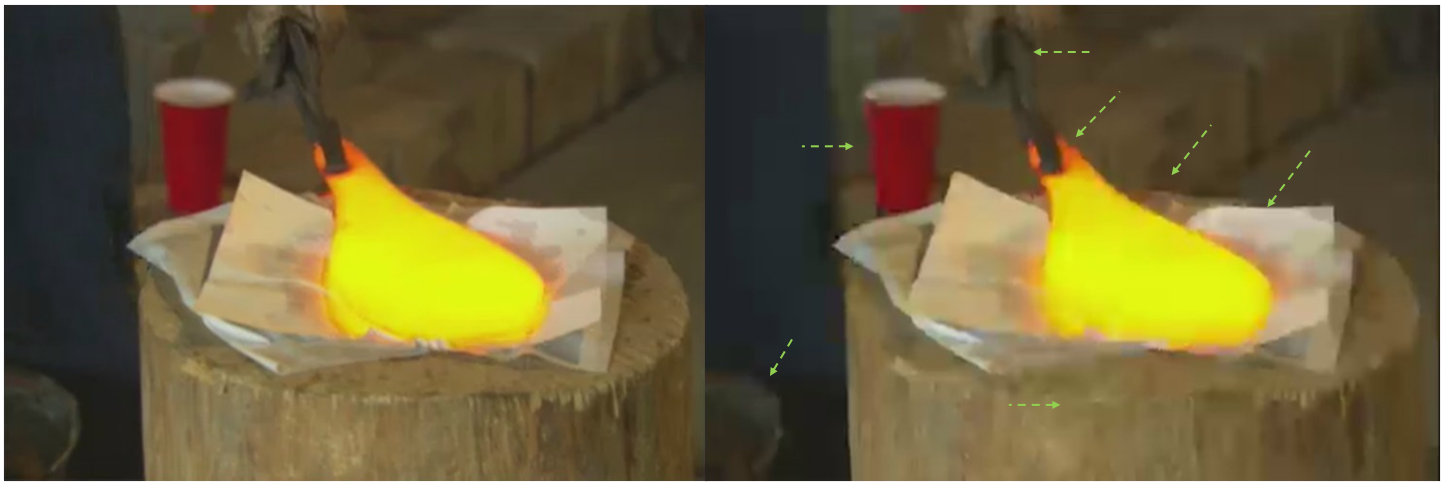
As sets of transcoded test sequences are generated each can be recorded, aligned, and tested against the reference in a scripted alignment plus applied metrics calculation or be aligned and tested after capture or from files in a batch via easily created command line interface script.

With its scripted or graphical user interface control ClearView may perform several types of measurements on a complete set of full-reference test sequences. Using the MS-SSIM algorithm, ClearView systems present results using the native MS-SSIM score with a linear DMOS scale. It can also test using VMAF or the Sarnoff JND perceptual metrics and present results on their native scales. ClearView's Metric Log Grapher also automatically creates comparison charts of all quality scores for each metric, and it also allows instant recall of all logged tests via drag-n-drop of the text log file to recall aligned reference and processed test streams for side-by-side, seamless split, and split mirror view modes for full-resolution viewing on HDTV or UHDTV monitors for visual inspection.

The graph in Figure 3 shows an estimated trend of many sequence score averages that would be expected if quality is measured by using MS-SSIM on the DMOS scale, where a lower score indicates better picture quality. Blue and red lines show a potential differing quality average between two encoders. In general, quality will increase as bit rates increase, signified by downward-sloping curves on the graph.

For any given type of source content, most points in the quality curve of the encode process will perform differently at different bit rates.

Figure 4 depicts a single example of an actual DMOS test comparing one piece of reference video processed at a single bit rate from two different encoders. The numerical results of this test are plotted on the graph below the images, yielding a



Frame 3608 – Encoder A Test 480x270 @ 350kbps

Frame 3608 – Encoder B Test 480x270 @ 350kbps

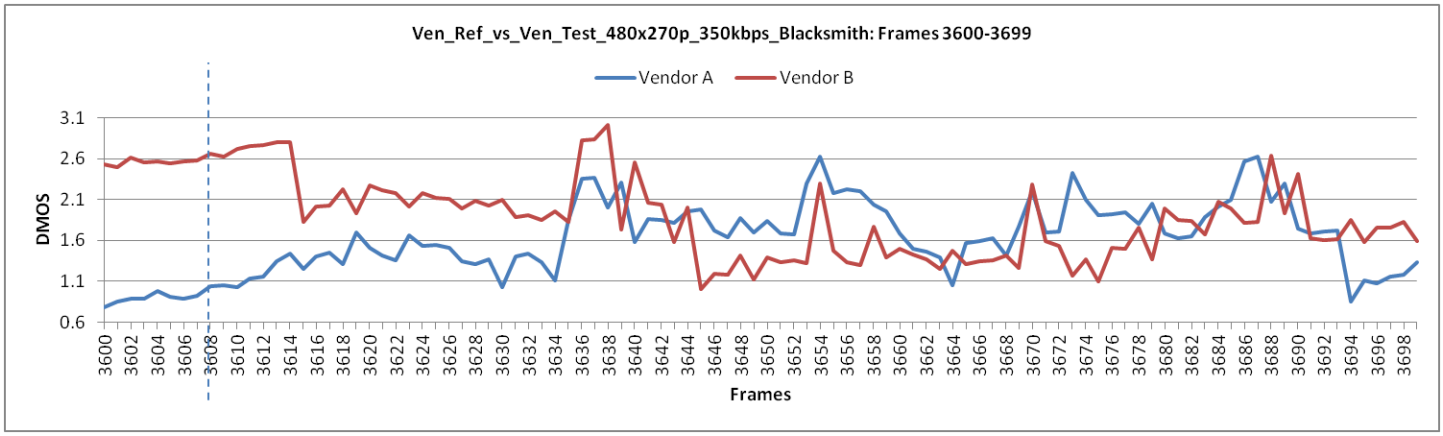



**Figure 4: Encoder A Reference @ 15Mbps vs. Encoder A Test @ 350kbps; Encoder B Reference @ 15Mbps vs. Encoder B Test @ 350kbps**

frame-for-frame comparison of the two codecs under test. For purposes of this paper and an understanding of the visual differences of these two encoded frame sets, a single frame of both encodes is shown as a comparison. By overlaying the results of a given piece of video going through different encoders, the graph demonstrates how one encoder compares to another over a given number of frames from the same source video.

Different types of content will create differing results from each encoder, therefore multiple sources must be encoded and tested at multiple bit rates in order to understand how well each encode process performs.

## Conclusion

Distributing varying resolutions of video via IP is a major part of many entertainment delivery operations today, and operators are using ABR services to do it. The nature of ABR services means there will be downstream deliverables in multiple profiles derived from a single source — all of which require test and measurement to verify their quality. The full-reference test method is generally considered the best test method in this case, and the way to ensure the most accurate test result is to create the reference sequences and the multiple-resolution test streams using the same encoding device so that they are identically formatted. From there, a testing tool can handle aligning the reference and test streams and provide repetitive measurements, generating numerical quality scores as well as visual comparisons. This methodology gives operators the truest possible measure of their IP delivered video quality that is instantly recallable for visual inspection, both of which help them make better decisions about applied processing and equipment investments.

### Works Cited


- "A Proposed VQM Methodology for ABR Networks." Pierre Costa LMTS, and Priyadarshini Anjanappa, AT&T Laboratories. Video Services Forum, April 15, 2014.
- "Achieving Maximum Accuracy in Video Quality Measurement" white paper
- "Analyzing 4K Video Quality" white paper
- ClearView Data Sheet


Video Clarity would like to recognize and thank Pierre Costa and Priyadarshini Anjanappa of AT&T Laboratories for their contributions to this white paper.