Measuring Video Quality in Videoconferencing Systems

By Roger Finger

A paradox in the videoconferencing and multimedia communications industry is that while there are defined international standards such as H.320 and H.323 which set foundations for network and vendor interoperability, there are no standards at all for performance evaluation. It should therefore come as no surprise that every vendor of videoconferencing claims "great quality," a wonderfully subjective measure, while also claiming 15 fps at 128 kbps, a wonderfully quantitative measure with no real substance behind it (Frame rates vary with motion content, among other things). Even worse perhaps is that there is no recognized industry consensus of what really determines conferencing quality, especially video quality. Caveat emptor.

One of the issues at play is that perceived multimedia quality can be captured only partially by quantitative measurements. Other aspects are equally important, but largely subjective. Generally these are too complex to capture with a single quantitative figure. For example, some vendors have engineered their products to maintain constant frame rate by sacrificing clarity when there is a high motion component. This represents a problem since there are no established units of measurement for how clear an image appears, or whether the video is smooth or choppy.

In mid-1997, Intel Corporation began a program to develop a suite of quantitative and qualitative measurements to help users evaluate audio and video quality in videoconferencing systems. We created a laboratory test stand and then evaluated commercially available, standards-compliant products from both Intel and other leading vendors. This article will discuss the test development program and the technical factors, both quantitative and qualitative, which affect perceived conferencing quality.

Criteria for Testing

Besides looking for a set of variables that truly impacted perceived video quality, we wanted a set of tests that would:

- Be reproducible across products and time. We used a common, realistic video source for all tests.
- Avoid loading down the target machine with computational requirements.
- Provide calibration against a known source such as a laser disk counter or SMPTE video track.
- Be platform independent and work with both PC-based and non-PC based platforms.

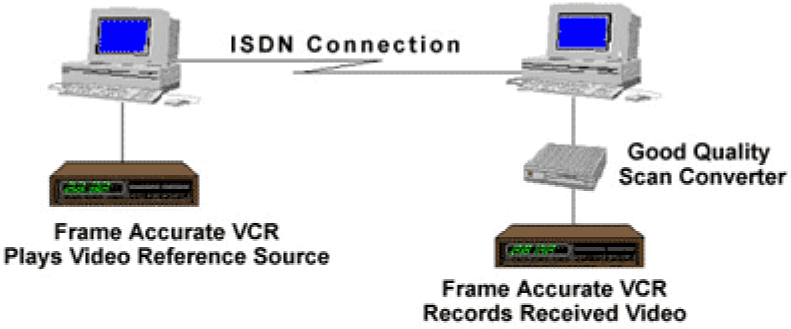

ISDN Connection
Good Quality Scan Converter
Frame Accurate VCR Plays Video Reference Source
Frame Accurate VCR Records Received Video

Figure 1: Typical test configuration

A typical test configuration is shown in **Figure 1**. We used the same videoconferencing product at both the sending and receiving ends of any test. A high quality, frame-accurate VCR player provided a video source to the camera input of the videoconferencing equipment under evaluation. Each frame on the Reference Source Video was pre-marked with a unique frame count. A scan converter on the second unit's VGA output recorded the entire conference (audio/video/data) on video tape for offline processing.

Audio and Video Tests

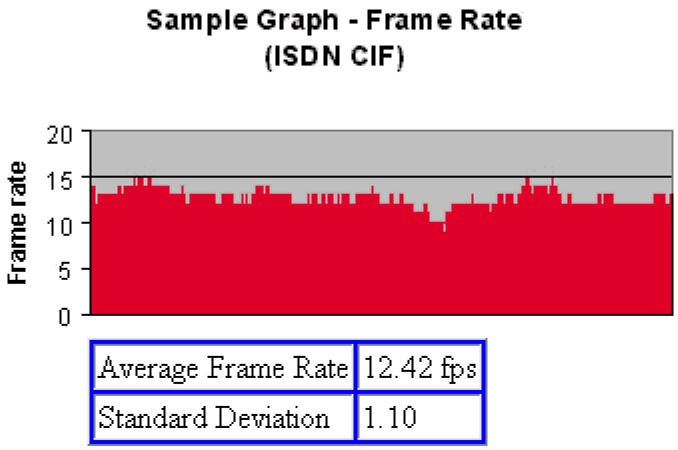
The Reference Video Source was a VHS or Beta SP tape containing a human subject, talking and moving normally as she would during a conference. The video contains embedded frame counters while a special test program was created to read the counters and make frame rate calculations (further details on the test program, reference video source, video calibration procedures, and data handling methods are available on request). Using a reference video source has a major advantage—it removes the data content variation from the product-to-product and test-to-test comparisons and is a crucial step when evaluating how a video codec will handle motion, color changes, and other variables. However, using a tape also removes from consideration the affect of any camera quality variations. Since some of the videoconferencing products in the market provide much higher quality cameras than do others, this important variable is lost in this test suite.

Quantitative Tests

The quantitative tests conducted for each product included:

Frame rate, the number of frames per second received during a video connection, measured with the same video source. We chose to calculate the frame rate as an moving average for 30-frame windows, and rolled the average one frame at a time. This has the effect of dampening out any instantaneous blips in frame rate. We also calculated the variation in frame rate across the video sample, which is charted as the standard deviation, though we doubt the statistics actually follow a Gaussian distribution.

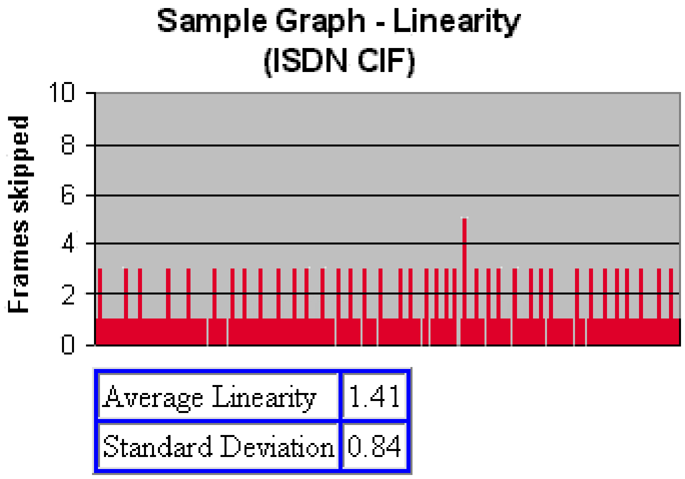
Linearity, which we defined as a function of the number of frames skipped between frames received. Frame rates in conferencing are different from those of broadcast television or film. Videoconferencing frame rates are constantly changing. Based on the amount of motion in the subject, the level of detail, and the percentage of the image that changes from one frame to the next. Often the peak frame rate published by a manufacturer can be achieved with only a non-moving subject, a non-meaningful metric. Unlike the frame rate calculation, our linearity figure is an instantaneous number. We calculated each value, the average for all values, and the standard deviation as well. Our suspicion here is that a codec that provides 26 fps during the first 1/2 second and 4 fps during the second 1/2 second will not produce as pleasing a conference quality as a codec that provides 15 frames per second consistently.



Average Frame Rate	12.42 fps
Standard Deviation	1.10

Figure 2: Typical plot of frame rate test results

A typical plot of one of the test results is shown in **Figure 2**. The frame rate plotted as a function of time is the 30-frame rolling average. **Figure 3** plots the instantaneous value of the number of frames skipped. This "linearity" measure is a strong proxy for a CODEC's sensitivity to motion. Large standard deviation values here indicate that frame rate was highly variable. This may indicate unusual motion sensitivity and loss of lip sync, certainly less desirable characteristics of a CODEC design.

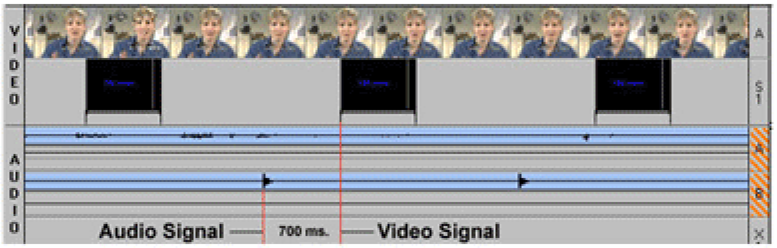


Average Linearity	1.41
Standard Deviation	0.84

Figure 3: Typical plot of linearity test results

The plot in **Figure 3** shows real-time events happening in the system. The large (5) number of frames skipped at one point in the conference shows the effect of a Windows 95 swap buffer, a problem with a direct draw graphics driver. While this is totally outside the domain of the desktop videoconferencing system vendor, the OS glitch had the effect of dramatically reducing video quality when the buffer needed CPU attention.

Latency, the amount of time required to transmit and receive an audio/video signal. Given the constraints of a desktop CODEC, and the time to move data around within a desktop system, the minimum H.320 latency is about 200 ms. If one-way latency exceeds 500 milliseconds, then videoconferencing users will experience an annoying delay when they speak. Latency on a circuit-switched network is due mainly to the processing time required to encode and decode the audio and video signals. For the test suite, the Reference Video Source plays an audio and video signal (timing numbers) in perfect sync, then the audio signal is gradually advanced in 100 ms intervals to 1000 ms. The videoconferencing system latency is measured by placing both machines in the same room and comparing the sent audio to the received video. Since audio comes out earlier than video, we can measure latency by listening for the point at which the sender's audio is synchronized with the receiver's video. Typical latency values ranged from 300 to 500 milliseconds, though one product tested much lower.



Lip Sync, the synchronization of the audio and video signals received during a videoconference. In most videoconferencing systems, the audio and video signals are encoded and transmitted separately, and then reassembled on the receiving end. Lip sync can drift in both positive and negative directions during a videoconference, depending on how many frames per second are being displayed and on the linearity of the video stream. Below 8 fps, in our experience, the video quality is too low to make lip sync a meaningful term. Most people barely perceive a 50 ms gap between audio and video (professional musicians tend to be more sensitive), and 100 ms is considered very acceptable for most videoconferencing applications.

T.120 data operations can affect lip sync as well, because the T.120 data stream consumes some of the communications bandwidth. Typically, data packets are sent at higher priority than video packets and cause some reduced frame rate and loss of sync. It is also possible to lose sync due to CPU activity while launching and closing other programs. On a busy LAN, network traffic can influence audio/video synchronization.

To measure lip sync, we used the same Reference Video Source as in the Latency test, but turned down the sound on the television attached to the VCR. We observed both the audio and video signals on the receiving end and listened for the point at which the flashing numbers on the screen corresponded exactly to the audio signal being received. 100 ms measurements were typical.

Video resolution, the ability to discriminate closely spaced lines on the receiving screen. In the real world, the quality of the video camera would be a major consideration, but camera quality has been eliminated as a variable in this test because we used a video tape source. This resolution test therefore measures the ability of the video codec to provide clear and stable backgrounds during a conference.

For this test, the reference video source is a static image of a test pattern. We evaluated the pattern on the receiving screen, by looking closely at the lines, starting from left to right. The viewer should be able to see clearly a white column between the black lines and then identify the most closely spaced set of resolvable lines. The numbers refer to the number of white pixels between lines. Then the viewer proceeds to the plum colored lines (which are thinner, but equally spaced), and verifies the previous observation. Finally, the same test is repeated for the green lines, which are only one pixel thick. If the results differ in each line group, we took the larger (worst) set of numbers.

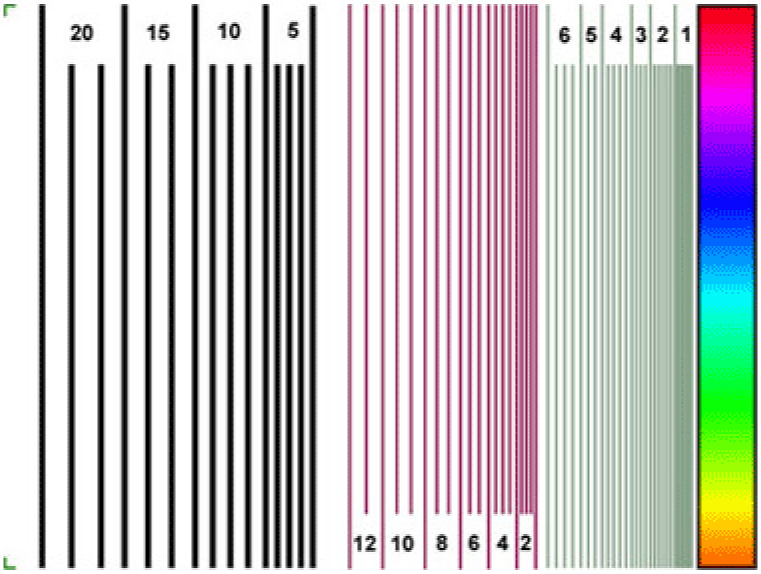
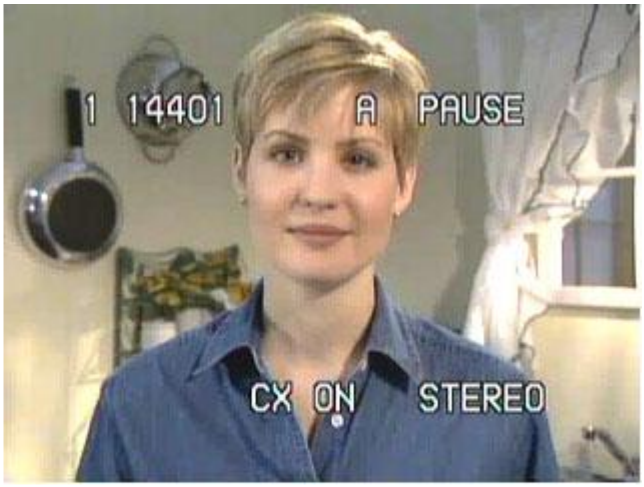


Image color clarity. Overall image color quality in comparison to the original picture. The color spectrum on the right edge of the resolution test image was used to determine if colors were being rendered accurately. Throwing out color values is a key component of most data compression schemes. We looked here primarily for color banding in the spectrum, which leads to less desirable color quality. Ideally the colors should flow smoothly in a gradient.

Subjective Observations



In addition to the quantitative measurements discussed above, we also conducted subjective tests which were particularly useful when comparing two systems side by side. For one test, we used a still shot of the woman on the Reference Video Source as a subject and then adjusting the contrast and brightness to provide the most pleasing image. Then we noted the following characteristics.

1. **Video artifacts:**
   We looked for video artifacts around the subject, especially around the subject's head and shoulders. These might appear as blocks (macrocells), color splotches, image distortions, or areas which are grossly out of focus.

2. **Sharpness:**
   Ideally it should be possible to see individual hairs on the woman's head. The line of her shoulder should be sharp and smooth, not jagged or fuzzy. Her eyes should be crisp and clear.

3. **Contrast, brightness, and color saturation:**
   We looked to see if the compression/decompression had effected the brightness and color saturation. Was the image dull or faded?

4. **Color depth:**
   We looked for color banding in the backgrounds and on the woman's face and compared to the video resolution test for color.

5. **Stability:**
   The image should be perfectly stable, with no motion in the background due to video artifacts, TV interlace jitter, or video noise ('snow'). The picture should not shimmer or deform over time.

6. **Background Clarity:**
   The background on the source image is slightly out of focus, but it is very rich in color and texture. The receiving image should be bright and clear.

A Word on Audio Quality

We did not include audio evaluations in our first round of testing. Like video, there are no metrics for audio quality, but unlike the video industry, the audio industry uses a standard evaluation approach based on "Mean Opinion Scores," which are really the equivalent of focus group ratings under carefully controlled conditions. Audio quality in a videoconference depends on several factors, including the codec that is used for voice compression. The ideal videoconferencing audio codec is the one that offers the widest frequency response while using only a small amount of conferencing bandwidth. Audio quality is also very dependent on the particular speakers and microphones.

For open audio and speakerphone solutions, there should be some form of echo cancellation to avoid picking up the received audio and re-transmitting it back to the source. Today, more and more videoconferencing systems have echo cancellation built-in. There may also be some form of automatic gain control to optimize the volume levels on the audio inputs. In addition, noise reduction can be applied in a variety of ways to improve the signal-to-noise ratio and to mute background noise when no one is speaking.

Summary and Conclusions

Frame Rate, While Quantitative, Isn't Everything!

If a product claims 15 frame per second video rates, does this mean that you are seeing great video? Not necessarily. Much higher frame rates are possible even on standard analog phone lines if you reduce the resolution of the image or transmit fewer bits per frame. Peak quality is achieved by tuning the audio and video compression rates based on the speed of the communication channel and the amount of processing power available.

The frame rate versus video quality trade-off can be seen by comparing 384 Kbps room conferencing systems. Some of these systems operate at 30 frames per second, and the video quality is almost as good as broadcast television. But many 384 Kbps systems operate at only 15 frames per second, and provide the same level or even higher perceived quality by sending more video information per frame. Focus groups are unable to tell the difference—both systems look great!

The differences become more pronounced at lower bit rates, where it becomes increasingly difficult to provide an acceptable level of quality while maintaining sufficient frame rate. If the subject is relatively still, then very high frame rates are easy to achieve. But if there is a lot of motion in the video, then either frame rates will vary or the number of video artifacts is likely to increase.

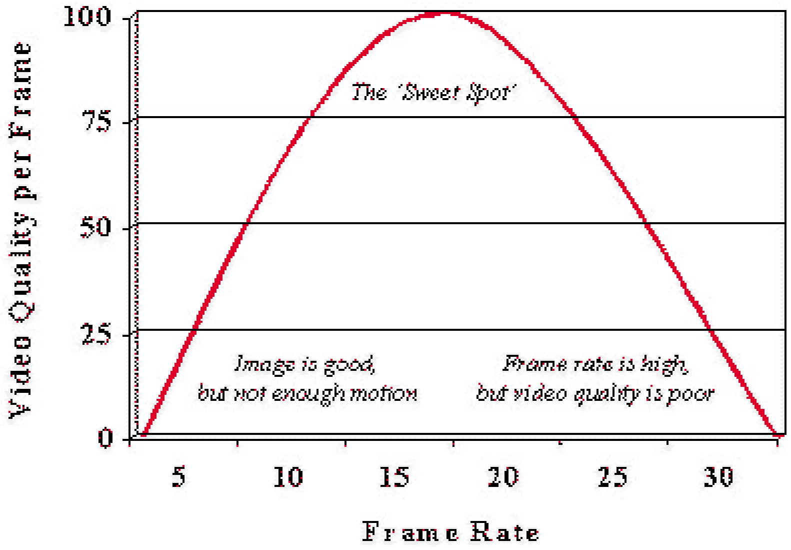
The Importance of Linearity

The abrupt jerkiness that is characteristic of poor quality digital video is caused by sudden variations in the frame rate. Hence, frame rates alone tell us nothing about this video quality parameter. A codec that delivers consistent frame rates over a range of input conditions delivers superior conferencing quality. Instead, we need a new unit of measurement, which we call 'Linearity' for lack of a better term, which measures how consistent the video frame rate is maintained. Our tests show that among codecs with similar frame rates, there can be a wide variation in our linearity measurements. We believe this issue will receive more attention over the coming months.

The Importance of Clarity

Image clarity is just as important as frame rate and linearity in assessing the overall performance of a videoconferencing system. The camera supplied with the system will often be the most important determinant in producing a good quality image, but the tests included in this report do not attempt to measure camera performance.

Assuming you start with a good quality video source, the next most important consideration will be the performance of the video CODEC. The two most commonly used video codecs are H.261 and H.263. Each codec has its own characteristics, and each can be enhanced with vendor-specific proprietary improvements. Video data compression is inherently lossy, therefore some image degradation will invariably occur. Whether implemented in hardware or software, the encoding and decoding of video are computationally intensive operations.



For example, if you were trying to optimize for clarity you could use more CPU power or dedicated hardware to generate the higher order mathematical terms that improve the sharpness of the image. But this takes computational time and tends to decrease frame rates. If you wanted to increase the frame rate to 30 fps, you could easily do so by throwing away the more precise higher order terms. The result would be fuzzy looking images with excellent motion tracking.

The Importance of Audio Quality
Assessing audio quality is still a subjective analysis, but at least the parameters of what is considered to be important can be enumerated. In addition to good frequency response, audio must be consistently delivered at the appropriate volume level with a minimum of background noise and hiss. Without acceptable video, we can talk; without audio, we have no conference.

Level of Integration With Data Tools
Our product evaluation also included a brief evaluation of how the vendors supported T.120 data conferencing and Microsoft's NetMeeting in particular. The details are outside the scope of this article.

Ideally, the conferencing data tools should be tightly integrated with the audio/video features of any videoconferencing system. But in dedicated function devices—such as room system videoconferencing devices, the data tools will probably reside on a separate PC attached via serial cable. This approach requires an additional computer system, but the videoconferencing systems themselves can be dedicated, and user friendly enough to be operated by hand-held remote. A personal computer-based system, on the other hand, with integrated audio/video and data services may be a more functional and economical solution, assuming that the host CPU has enough processing power to handle all of these functions. But simply hosting the data conferencing applications on the same screen as the video is not enough. The data tools must be easily accessible from menus and buttons within the main videoconferencing control panel and the user interface must be intuitive.

Which Product is Best for Me?
The answer to this question will depend on your budget, your application, your users, and your needs as a videoconferencing consumer. Our hope is that the information contained in this report will help create standards for comparison for the many available videoconferencing products and provide a set of metrics that can aid in purchasing decisions. The best consumer is still the educated consumer.