

# Video Streaming Quality Measurement with VSQI

Technical Paper

© Ascom 2009. All rights reserved.

TEMS is a trademark of Ascom. All other trademarks are the property of their respective holders.

No part of this document may be reproduced in any form without the written permission of the copyright holder.

The contents of this document are subject to revision without notice due to continued progress in methodology, design and manufacturing. Ascom shall have no liability for any error or damage of any kind resulting from the use of this document.

# Contents

<b>1. Introduction .....</b>	<b>1</b>
<b>2. Human Visual Perception .....</b>	<b>1</b>
<b>3. Assessing Video Quality with Objective Measures .....</b>	<b>2</b>
3.1. Reference vs. No-reference Methods .....	2
3.2. Perceptual vs. Non-perceptual Input.....	2
3.3. Technical Properties of Video That Affect Perceived Quality.....	3
3.4. Degradations during Data Transfer .....	3
3.5. Behavior of Streaming Client Application .....	4
<b>4. VSQI .....</b>	<b>5</b>
4.1. What VSQI Is Based On .....	5
4.2. What VSQI Does Not Consider.....	5
4.3. Static and Dynamic VSQI.....	6
4.3.1. Static VSQI .....	6
4.3.2. Dynamic (Realtime) VSQI.....	8
4.3.2.1. VSQI Intermediate Score.....	9
4.3.3. Comparison of Static and Dynamic VSQI.....	9
<b>5. ITU-T Standardization .....</b>	<b>10</b>



## 1. Introduction

The high data rates of 3G networks enable video services such as video telephony, on-demand streaming, and realtime streaming. Like the voice service, video services need to be monitored to ensure that users experience them as being of adequate quality. These quality monitoring procedures must necessarily be automated, since it would be obviously impracticable to have a panel of flesh-and-blood test persons continuously evaluating an entire cellular network (not to mention the prohibitive cost it would entail).

For voice, quality assurance has reached a mature state, and standardized methods and tools exist for objective speech quality monitoring and for troubleshooting of the service. Video services, on the other hand, are not yet mature in this regard. Assessing the quality of video is also more difficult because of the greater complexity of the signal as well as its perception by humans. For multimedia, which combines video and audio, this complexity is compounded.

This paper deals with video streaming and describes an algorithm called VSQI, developed by Ericsson<sup>1</sup> for objectively judging the quality of that service. VSQI is primarily intended for on-demand streaming, but the dynamic version of the algorithm could conceivably also be applied to realtime streaming.

Video telephony is not within the scope of VSQI; for this service, TEMS products provide a separate quality measure, VTQI (Video Telephony Quality Index). See the document Video Telephony Quality Measurement with VTQI.

## 2. Human Visual Perception

Visual perception by humans is a highly complex affair that involves multiple mechanisms and is influenced by expectations and prior knowledge. Quality judgments are inextricably tied to perceptual mechanisms. The perceived quality of a video clip will therefore depend not only on its technical quality but also on other factors, such as its content. When viewing a football game, it is of vital importance to be able to track the ball; in a news clip featuring the anchor, on the other hand, that person's facial expression is what needs to come through clearly.

The viewer's emotional involvement is another major factor. A viewer who enjoys the content might be more tolerant of quality degradations, while an uninterested and bored viewer might tend to be more critically inclined even if the technical quality is the same. However, the opposite could equally well be true: a sufficiently enthusiastic viewer might be *more* sensitive to disruptions and other quality problems, precisely because he is anxious not to miss any of the content. Clearly there is no question of this being an exact science.

Expectations on quality are naturally also dependent on the equipment used. People will accept lower quality when watching a mobile phone screen than they do when watching a DVD movie at home.

---

<sup>1</sup> The TEMS business was owned by Ericsson until 2009, when it was acquired by Ascom.

## 3. Assessing Video Quality with Objective Measures

The term *quality of experience* (QoE) has been coined to differentiate between user-perceived quality and technical quality measures relating to data transport, commonly denoted *quality of service* (QoS). QoE could be defined as “the overall acceptability of an application or service, as perceived subjectively by the end-user”.

To measure QoE as such with objective methods is patently impossible, since it is dependent on the factors mentioned in chapter 2, and on many other things besides. Fortunately, it is possible to obtain a fair approximation of QoE by studying technical properties of the transferred video. What we are then measuring is not QoE itself, but aspects of video quality that are related to QoE. Chapter 3 takes a look at some of these aspects, and against that background chapter 4 sketches the workings of VSQI.

### 3.1. Reference vs. No-reference Methods

Some methods of objective quality assessment compare the signal presented to the end-user with the original, undistorted signal. The original then serves as a *reference* against which the end-user’s signal is measured. If a video frame presented to the viewer is identical to the original, the highest possible score is obtained for that frame. The more the original has been distorted, the lower the score. A synchronization algorithm is required to align the two signals correctly before the comparison is made.

A *no-reference* method, in contrast, deals only with the received signal. Consequently, it does not measure degradation but judges the quality of the received signal on its own merits, extracting and assessing some judiciously chosen properties of the signal.

Between these two extremes we find *reduced-reference* methods, where the quality assessment algorithm does not consider the reference as such but does receive some information about it.

Full-reference methods have the advantage of greater precision: the correlation with subjective perceived quality is normally somewhat higher than for a no-reference method. This is to be expected, since the reference method has more information to go on. No-reference methods, on the other hand, are more generally applicable: access to the original may be difficult, or its capture may be impractical. No-reference methods also do not require synchronization.

### 3.2. Perceptual vs. Non-perceptual Input

Input to quality assessment algorithms can be *perceptual* or *non-perceptual*. Perceptual input is related to what humans perceive; in the case at hand, video and audio. Non-perceptual input is data that cannot be perceived by a human, such as throughput or block errors over a radio link.

An algorithm that uses perceptual input normally extracts artifact properties such as blockiness, blurriness and jerkiness from the video images. These properties are then used to estimate perceived quality on a scale analogous to that of MOS (Mean Opinion Score), usually with reference to a model of the human visual system. Algorithms taking perceptual input are optimally suited to detect artifacts in individual video frames.

Algorithms with non-perceptual input estimate perceived quality based on parameters such as the choice of codecs and frame rate, the occurrence of buffering, the packet loss level, and the achieved throughput. Such an algorithm will not be as versatile as an algorithm with perceptual input; rather, it must be tuned for a specific setup, say, for a specific video codec and a limited set of bit rates. Properly trained, however, the algorithm can perform excellently within its application area. It will not detect single-image artifacts, but it will report the same performance *on average* as an algorithm using perceptual input. Furthermore, the average performance is usually the focus of interest, as opposed to detailed information on transient phenomena.

An indisputable advantage of dispensing with perceptual input is that it permits computationally more efficient implementations.

### 3.3. Technical Properties of Video That Affect Perceived Quality

In spite of the difficulties mentioned in chapter 2, there are concrete properties of video footage that correlate closely with perceived video quality.

The following are some key parameters relating to the process of recording and encoding the video signal:

- Codecs
- Frame rate
- Quantization
- Picture resolution

When a video is transmitted over a wireless link, the limited bandwidth imposes constraints on these parameters. Normally the video needs to be encoded with a lossy compression algorithm, which irreversibly degrades the video quality to some extent: a trade-off has to be made between frame rate, quantization, and picture resolution.

### 3.4. Degradations during Data Transfer

Difficult network conditions may cause some packets to be lost or delayed at the transport level. Such errors may in turn result in visible artifacts at the application level, i.e. in the video replay. Examples of such phenomena are blockiness, frame rate jitter, and low frame rate.

If the radio conditions are so bad that the data throughput falls below the required rate for an extended period of time, the streaming application may have to rebuffer and halt the replay. See section 3.5.

### **3.5. Behavior of Streaming Client Application**

Every streaming application buffers some data at the beginning of the transmission before starting the replay. If the data throughput in the network is low, this initial buffering may drag on to such an extent as to annoy the user.

During the transmission, if the average data throughput stays low for an extended period of time, the streaming client may run out of buffered data and be forced to halt the replay. The client must then build up the buffer again before the replay can continue. (Still, even severe transmission problems do not always result in rebuffering. The streaming application may be able to forge ahead despite bursts of heavy packet losses, if enough fragments of the data come through undamaged in between; but the replay may be completely ruined in the eyes of a human viewer.)

It is important to note here that the buffering behavior is application-specific, so that the properties of the particular application used has a major impact on the perceived quality. This always has to be kept in mind when considering video streaming quality measurements – whether subjective or objective. Compare the much simpler situation with the voice service, where the encoding and decoding procedures are rigidly standardized and no client applications exist that could introduce additional degrees of freedom. In video streaming, too, use is made of many components and procedures that are in themselves standardized (video and audio codecs, UDP, IP and RTP protocols, etc.), but it is unlikely that a comprehensive standard specifying all aspects of video streaming services will ever emerge. The fine points of buffering behavior, for example, will remain features of the individual client applications.

In conclusion, then, quality assessment for video streaming needs to be approached pragmatically. Conditions will differ between phones and applications, but it is better to measure something than to have no data at all. Also, the difficulties in this regard should not be exaggerated; most existing streaming applications are readily comparable.

## 4. VSQI

VSQI is short for Video Streaming Quality Index.

Like SQI, the corresponding TEMS objective quality measure for voice, VSQI is a *no-reference* method (compare section 3.1). The main reason for not using a reference is that the computational complexity would be prohibitive in the dynamic version of VSQI (see section 4.3.2).

The kind of subjective test which VSQI strives to imitate is one where viewers are instructed to assess both video *and audio* and combine their perception of each into an overall “multimedia quality” score.

The output from the VSQI algorithm is expressed as a value between 1 and 5, conforming to the MOS (Mean Opinion Score) scale which is frequently used in subjective quality tests. The unit for VSQI is called “MOS-VSQI”.

VSQI has been tuned for the QCIF video format (176 x 144 pixels).

VSQI works equally well for all kinds of radio networks and bearers.

### 4.1. What VSQI Is Based On

VSQI is based entirely on non-perceptual input (see section 3.2), essentially the following:

1. The quality of the encoded (compressed) signal prior to transmission. This quality is straightforwardly a function of the video and audio codecs used, and their bit rates. Compare section 3.3. The information actually used by the VSQI algorithm is the video codec type and the total (video + audio) bit rate. For the codecs listed in section 4.3.1 and their various bit rates, the “clean” quality has been computed in advance. In practice, what codec and bit rate are used in the streaming session is deduced from the name of the streamed file or from the signaling between server and client.
2. The amount of initial delay and the subsequent interruptions during playback of the video sequence: that is, the time required for initial buffering and the incidence of rebuffering. Compare section 3.5.
3. The amount of packet loss at the application level (i.e. in the video streaming client). Compare section 3.4.

### 4.2. What VSQI Does Not Consider

VSQI does *not* use perceptual input to detect specific visible artifacts as described in section 3.2. The transferred video is not analyzed frame by frame in any way. Thanks to the monitoring of packet loss (item no. 2 in section 4.1 above), however, even slight problems with blockiness, jitter, and so on will still be noticed by the algorithm and affect the VSQI score.

### 4.3. Static and Dynamic VSQI

Two versions of the VSQI algorithm have been devised: one static and one dynamic version.

#### 4.3.1. Static VSQI

The static version of VSQI takes an entire streamed video clip as input and assigns a single quality score to it.

Input parameters to the static version of VSQI are as follows:

- Video codec used (H.263, H.264, or MPEG4)
- Total bit rate (video + audio)
- Duration of initial buffering
- Number of rebuffering periods and their duration
- Amount of packet loss

With some degree of simplification, we may describe the calculation of static VSQI with the following formula:

$$VSQI_{\text{static}} = VSQI_{\text{clean}} - (\text{buffering penalty}) - (\text{packet loss penalty})$$

Here,  $VSQI_{\text{clean}}$  is the “clean value” obtained for the clip prior to transmission. This score is determined by the quality of the encoding, which is in turn dependent on the choice of codecs and bit rate. One typical such relationship is shown in Figure 1:

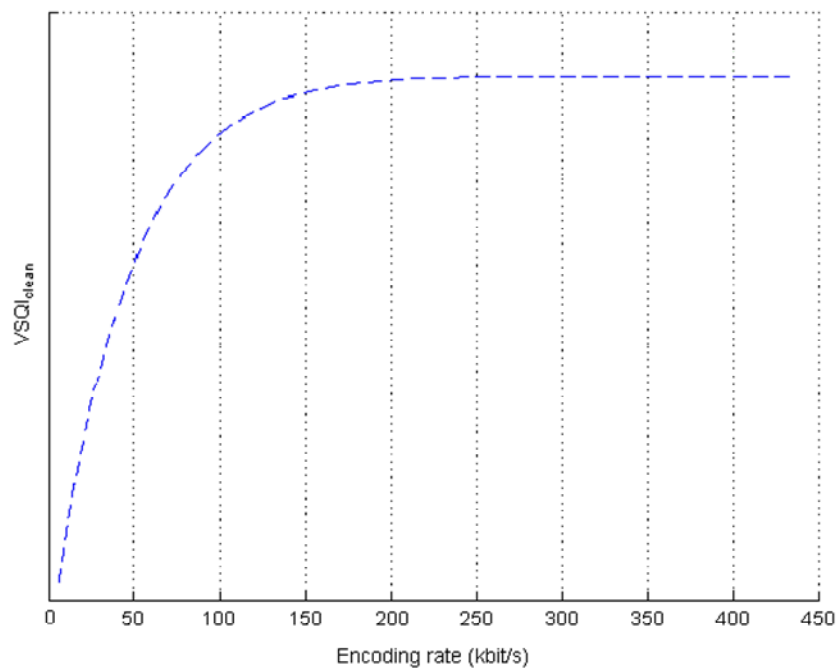


Figure 1 VSQI clean value as function of encoding rate (example).

The size of the buffering penalty depends on the time taken for initial buffering, the time spent rebuffering, and the number of rebuffering events. See Figure 2.

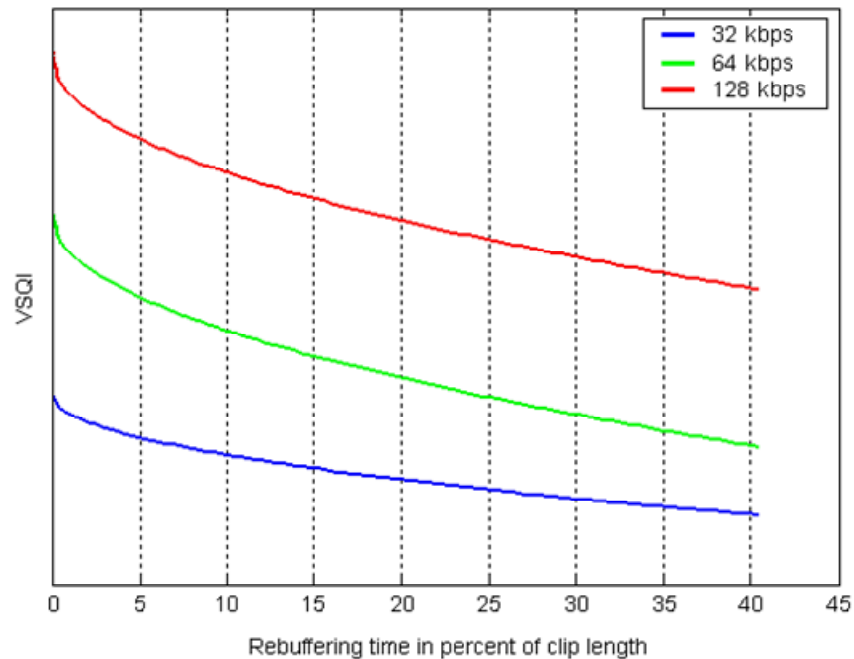


Figure 2 VSQI as function of rebuffering time for a particular codec operating at three different encoding rates.

The size of the packet loss penalty is determined as follows. A running packet loss average over the last 4 s is computed approximately every second, and the values thus obtained are weighted and summed to yield an appropriate overall measure of the packet loss. The latter is then translated into a deduction from the VSQI score.

The static VSQI algorithm has been fine-tuned for clips of around 30 s and should therefore in practical use be applied to clips of similar duration. The video sequences must not be too short because of how the buffering works: each instance of rebuffering takes several seconds to complete, and moreover if the clip is short enough it will have been buffered in its entirety before the replay starts, so that no rebuffering will ever occur. For clips considerably longer than 30 s, on the other hand, disturbances towards the end will be more harshly penalized by viewers than those occurring early on, simply because the late ones are remembered more vividly. Therefore, since the current VSQI algorithm does not take into account such memory effects, it would probably perform slightly worse for long clips. (The dynamic version of VSQI is naturally not affected by this limitation.)

### 4.3.2. Dynamic (Realtime) VSQI

The dynamic, or realtime, version of VSQI estimates the quality of a streaming video clip as perceived by viewers *at a moment in time*. It is updated regularly – at intervals of the order of 1 s – while the video clip is playing. Each VSQI output value is dependent on the recent history of the streaming session (i.e. recent packet loss levels and possible recent buffering events).

The design of dynamic VSQI is based on the following:

- Previous research suggesting approximate times taken for the perceived quality to drop to MOS-VSQI 1 (during buffering) and to rise to the highest attainable VSQI (during normal replay)
- Modeling of the impact of packet loss on perceived quality
- Tailoring of mathematical functions for expressing viewer annoyance/satisfaction as a function of time (in each of the states that are possible during replay)
- Codec and bit rate parameters as in the static version

Figure 3 below shows in rough outline the different ways in which dynamic VSQI can evolve during the replay of a streaming video clip. The best achievable quality, i.e. the “ceiling” in the graph, is dependent on the codec/bit rate combination but is also affected by the amount of packet loss. In this example the packet loss is assumed to be constant so that the influence of buffering can be clearly discerned.

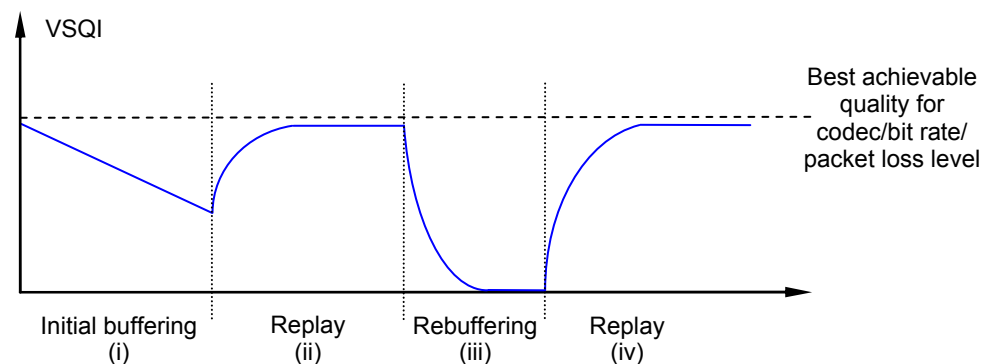


Figure 3 Effect on dynamic VSQI of initial buffering and rebuffering events.

- The user tolerates (and might even expect) a certain amount of initial delay; but the longer the buffering drags on, the more the user loses patience.
- Once the replay gets going, the perceived quality picks up again and soon approaches the highest achievable level.
- If rebuffering occurs, VSQI deteriorates rapidly; compare Figure 2. Rebuffering events are much less tolerated by viewers than initial buffering, especially if repeated; VSQI captures the latter by making the slope of the curve steeper for each new rebuffering event.

- iv. After the replay has recommenced, VSQI recovers reasonably quickly, but not infrequently from a rock bottom level.

An authentic example, which also shows the impact of the packet loss rate on the VSQI score, is given in Figure 4.

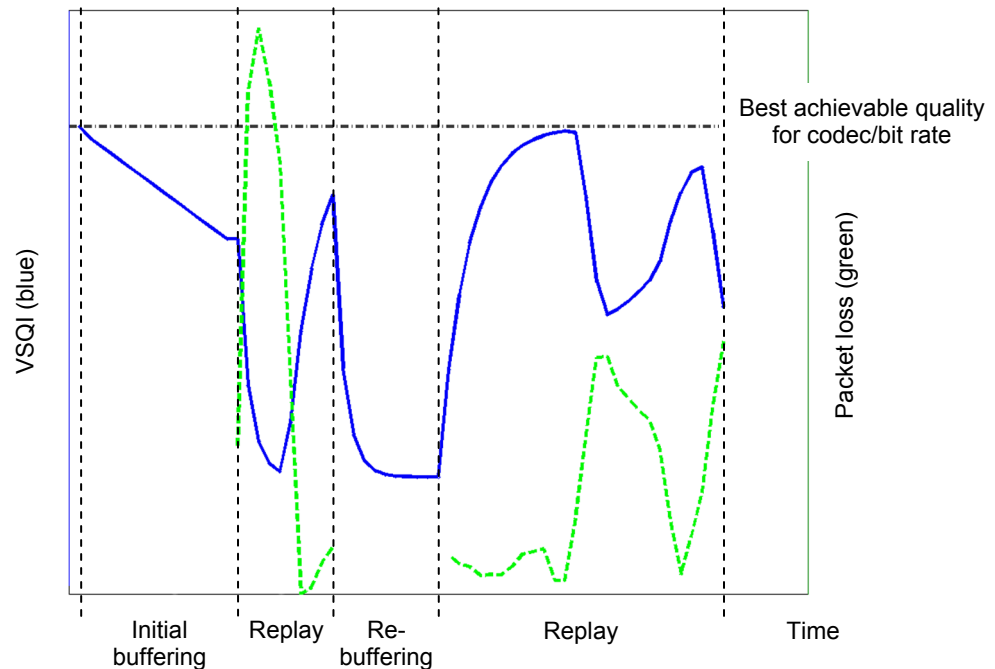


Figure 4 Evolution of dynamic VSQI (blue) in the course of a video streaming session with strongly varying radio conditions (packet loss rate drawn in green).

#### 4.3.2.1. VSQI Intermediate Score

In TEMS products yet another VSQI value is presented, called VSQI Intermediate Score. This is a straightforward average of the latest 30 dynamic VSQI scores.

#### 4.3.3. Comparison of Static and Dynamic VSQI

The average of the dynamic VSQI scores for a clip is not necessarily identical to the static VSQI score. This might be found surprising but is by no means unnatural, since the actions of a human viewer in the two cases are clearly different.

Static VSQI corresponds to the viewer watching the whole clip and then giving it an overall rating. Dynamic VSQI, on the other hand, mimics the process of the viewer judging the quality continuously, from one moment to the next throughout the replay. Since these procedures are markedly different both perceptually (impact of sudden quality changes, memory effects) and practically (assigning a single rating in retrospect vs. producing multiple ratings while continuing to watch and listen), it is only to be expected that they might not yield fully equivalent results.

The two VSQI versions should therefore be regarded each in its own right, and an exact mathematical relationship between two such measures is neither easily derived nor particularly desirable (for the reasons stated above). Needless to say, static and dynamic VSQI will nevertheless be strongly correlated.

## 5. ITU-T Standardization

VSQI is an important part of the P.NAMS standardization project within ITU-T, scheduled to be concluded in 2010. P.NAMS focuses on two main areas: TV (standard resolution as well as HDTV) and low bit rate video (streaming, mobile TV, etc.).

Two other ITU-T standards J.246 and J.247 already exist for evaluation of video quality. J.246 describes a set of reduced-reference models where information about the original signal is transmitted to the receiver. J.247 contains full-reference models where the original and the received signal are compared directly. No assessment of audio quality is performed in either of these standards.