ATIS IIF – QoSM Task Force

IIF-WT-nnnn

CONTRIBUTION

TITLE: **An Approach to In-Service Measurement of Video QoE**

SOURCE: **Alan Clark, Telchemy**

PROJECT: **ATIS IIF, QoSM Task Force**

ABSTRACT

This contribution introduces an approach to video performance measurement based on estimation of PSNR.

---

1. Introduction

The approach proposed in this document is based on the estimation of PSNR (Peak Signal to Noise Ratio). PSNR is defined as

$$PSNR = 10 \log_{10}(m^2 / MSE)$$

where m is the pixel range and MSE is the mean squared error across both spatial and temporal dimensions

The concept is fairly straightforward and logical, and has been followed previously within the published literature. PSNR relates closely to perceptual quality, can be estimated fairly readily with knowledge of the codec parameters and packet loss rate and is easily measured using full reference analysis.

- CONTACT: Alan Clark; Telchemy, alan.d.clark@telchemy.com; Tel: 770 614-6944; Fax: 770 614-3951

There are other objective measurement approaches, such as those described in ITU Recommendation J.144, however those that correlate more closely with subjective quality are full reference models that are impractical in an in-service application. Many of these models use PSNR as a basis or as one of their input parameters.

This approach is preliminary, experimental in nature and subject to change.

2. MPEG stream structure

A typical MPEG-2/4 Group of Pictures has the general structure:

[I B B P B B P B B P B B P B B] [I….]

transmitted in the order

[I P B B P B B P B B …]

Each I frame is encoded independently, B and P frames are differentially encoded based on the previous I or P frame. For the above example, with a GOP size of 15 frames, each GOP independently represents approximately 500 milliseconds of video.

I frames typically take 40 percent of the bandwidth with the remaining 60 percent being divided amongst the P and B frames. This means that an I frame takes approximately ten times the number of Transport Units (RTP or MPEG) or IP packets than a B or P frame.

3. Impact of lost packets

MPEG encoders are based on an 8 x 8 (or 16 x 16) pixel block structure. With typical compression ratios a 1500 byte IP packet can carry approximately 90 blocks, and hence if an IP packet is lost then a rectangular strip approximately 90 x 8 pixels wide and 8 pixels high will be impacted. A "slice" structure is also commonly used, which may extend the effects of a lost packet to the edge of the frame.

The proportion of the image impacted by a single lost packet will be the ratio of the number of pixels carried within a packet (more if a slice is impacted) and the number of pixels in a frame.

Assuming that spatial or temporal interpolation is not performed then the difference in value between each pixel in the impaired region and the original pixel value will be a random value with a maximum range equivalent to +/- the range of pixel values. With the further assumption that the range of pixel values tends to be centrally biased, the typical range of errors will be +/- half the range of pixel values.

The Mean Squared Error (MSE) for an image is the sum of the squared errors between individual pixel impaired values and their original values. With the assumption above, the MSE would be:

$$\text{Approximate MSE} = (N_u * 0 + N_i (0.5*R)^2) / (N_u + N_i)$$

For a normalized pixel range of 1 this would give an estimated MSE of $N_i \ 0.25 / (N_u + N_i)$

In the case of video sequences, the MSE is averaged over the frames within the sequence.

The PSNR for an image is given by:

$$\text{PSNR} = 10 \log_{10}(m^2 / \text{MSE}) \text{ where m is the pixel range}$$

For a normalized pixel range, the PSNR is therefore $10 \log_{10}(1 / \text{MSE})$

For an image of typical broadcast resolution, the proportion of the image represented by a single IP packet is small and hence it is reasonable to assume that the proportion of pixels impacted is proportional to packet loss rate p.

$$\text{Average } N_i = N p \quad \text{or} \quad p = N_i / (N_u + N_i)$$

$$\text{Approximate } MSE_{PL} = 0.25 p$$

$$\text{Approximate PSNR} = 10 \log_{10}(1 / \text{MSE}) = 10 \log_{10}(4 / p)$$

In practice, it is necessary to incorporate the error extension effects due to the use of interpolated frames.

Consider the following frame sequence I B_1 B_1 P_1 B_2 B_2 P_2…

$$\text{Proportion of I frame impacted} = Q_0 = N_i / (N_u + N_i) = p$$

Proportion of B_1 or P_1 frame impacted = $Q_1 = Q_0 + (1 - Q_0) N_i / (N_u + N_i)$

B and P frames only contain a proportion (X/N) of the macroblocks, essentially those that represent changes from the earlier I or P frame. Hence, more precisely

Proportion of B_1 or P_1 frame impacted = $Q_1 = Q_0 + (1 - Q_0) p X_1 / N$

Subsequent B and P frames may be derived from an impaired P frame and hence:

Proportion of B_2 or P_2 frame impacted = $Q_2 = Q_1 + (1 - Q_1) p X_2 / N$

Hence the overall expression for the MSE within a GOP is:

MSE = Average($0.25 Q_0 + 0.25 F_1 Q_1 + 0.25 F_2 Q_2$…..)

where F_i indicates the number of frames at a given interpolation level.

4. Impact of bit rate and frame size

The bit rate is affected by image size, frame rate and quantization level.

For typical MPEG 4 or H.264 encoders with standard resolution of approximately 704x480 and a GOP size of 15, the MSE due to bit rate (quantization level) can be approximated by:

$MSE_{BR} = Z_0 + Z_1 / (B + B^2/Z_2)$

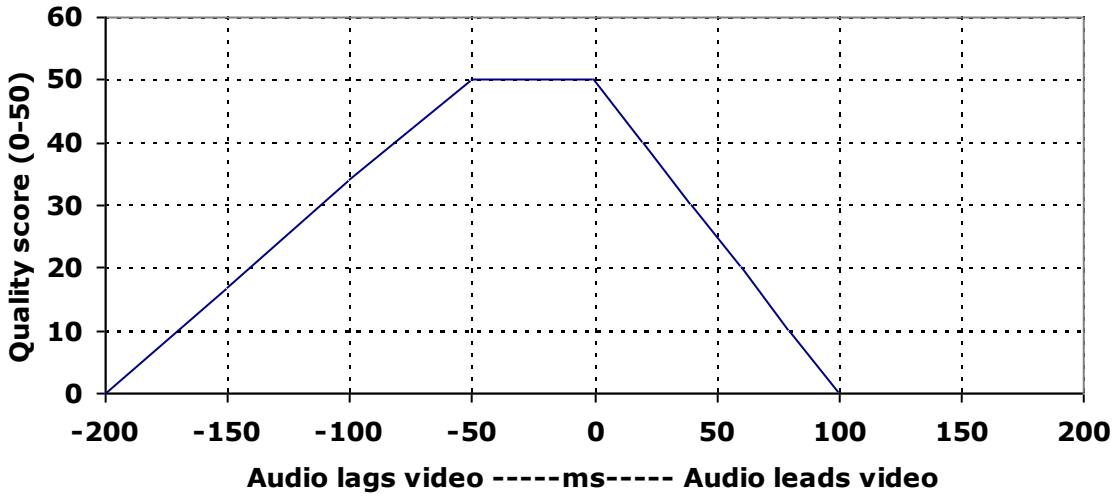
where the bit rate B is given in kilobits per second

The bit rate can be adjusted to an effective bit rate by multiplying by the ratio of the number of pixels in a standard resolution frame N_{SDTV} to the number of pixels in the frame size being used N_{ACT}.

The bit rate will also depend on the proportion of I to P/B frames and frame rate. An I frame is typically ten times the size of a B or P frame. The bandwidth for an MPEG stream consisting of only I frames would therefore be approximately six times as large as an MPEG stream with a typical structure.

5. Audio-Video Sync

The sound channel should not lead the video channel by more than 15 milliseconds or lag by more than 45 milliseconds per ATSC IS191, as people are more sensitive to lagging audio than leading audio. The chart below shows a very rough ("ball park") model for the impact of audio-video sync on perceptual quality - this should be replaced when additional subjective test data is available.



6. Video Quality Model

6.1 Transmission quality VSTQ

The video transmission quality factor - VSTQ - is a codec independent parameter based only on the impact of packet loss on a "nominal" codec

$$MSE_{PL} = \text{Average}(0.25\ Q_0 + 0.25\ F_1\ Q_1 + 0.25\ F_2\ Q_2 \ldots..)$$

$$Q_0 = N_i\ /\ (N_u + N_i)$$

$$Q_i = Q_{i-1} + (1 - Q_{i-1})\ p\ X_i\ /\ N$$

$$PSNR_{PL} = 10 \log_{10}(\ 1\ /\ MSE_{PL}\ )$$

$$VSTQ = \min(\ 0\ ,\ \max(\ (PSNR_{PL} - 18) * 2.2)\ )\ )$$

6.2 Picture quality VSPQ

The picture quality factor - VSPQ - is a codec dependant parameter that incorporates the actual (or estimated) codec performance, frame size, bit rate, frame rate and GOP structure.

$$MSE_{PL} = Average(0.25\ Q_0 + 0.25\ F_1\ Q_1 + 0.25\ F_2\ Q_2 \ldots..)$$

$$Q_0 = N_i / (N_u + N_i)$$

$$Q_i = Q_{i-1} + (1 - Q_{i-1})\ p\ X_i / N$$

$$PSNR = 10 \log_{10}(1 / (MSE_{PL} + MSE_{BR}))$$

$$VSTQ = \min(0 , \max((PSNR - 18) * 2.2)))$$

$$Video\ MOS = 1 + VSTQ * 0.08 + VSTQ * (50 - VSTQ)\ (\ VSTQ - 30) * 0.000056$$

6.3 Audio quality VSAQ

The audio quality factor is calculated using the wideband E Model - R_{WB}

$$VSAQ = R_{WB} / 2.4$$

6.4 Audio Video Sync Quality VSSQ

Temporary method of estimating VSSQ based on Audio-Video Delay (AVD). Note that a positive AVD indicates audio is leading video.

$$VSSQ = \begin{cases} \max(0, 50 - (-50 - AVD) / 5) & AVD < -50 \\ 50 & -50 <= AVD <= +20 \\ \max(0, 50 - (AVD - 20) / 2) & AVD > +20 \end{cases}$$

6.5 Multimedia quality VSMQ

The estimated multimedia quality is determined from the individual components using a Euclidean sum

$VSMQ = sqrt(VSPQ^2 + VSAQ^2 + VSSQ^2)$

Multimedia MOS = 1 + VSMQ * 0.08 + VSMQ * (50 - VSMQ) (VSMQ - 30) * 0.00005

6.6 Control plane quality VSCQ

To be defined.

7. Summary

This contribution proposed a simple computational model for calculating a range of video performance factors. It is proposed that this model be considered as a starting point in determining a practical means of in service estimation of IPTV system performance. The model is at this stage incomplete and in need of refinement however it does provide a logical framework on which to build.

The process of comparing this model to both objective and subjective test data is ongoing, and this may result in suggested updates to the model