# Estimating the impact of single and multiple freezes on video quality

S. van Kester[1], T. Xiao[3], R.E. Kooij[1,2], K. Brunnström[3], O.K. Ahmed[2]

[1]University of Technology Delft, Fac. of Electrical Engineering, Mathematics and Computer Science
[2]TNO, Delft, the Netherlands,
[3]NetLab: IPTV, Video and Display Quality, Acreo, Kista, Sweden

## ABSTRACT

This paper studies the impact of freezing of video on quality as experienced by users. Two types of freezes are investigated. First a freeze where the image pauses, so no frames were lost (frame halt). In the second type of freeze, the image freezes and skips that part of the video (frame drop). Measuring Mean Opinion Score (MOS) was done by subjective tests. Video sequences of 20 seconds were displayed for four types of content, to a total of 23 test subjects. We conclude there is no difference in the perceived quality between frame drops and frame halts. Therefore one model for single freezes was constructed. According to this model the acceptable freezing time (MOS>3.5) is 0.36 seconds.

Pastrana – Vidal et al. (2004) suggested a relationship between the probability of detection and the duration of the dropped frames. They also found that it is important to consider not only the duration of the freeze but also the number of freeze occurrences. Using their relationship between the total duration of the freeze and the number of occurrences, we propose a model for multiple freezes, based upon our model for single freeze occurrences. A subjective test was designed to evaluate the performance of the model for multiple freezes. Good performance was found on this data i.e a correlation higher than 0.9.

**Keywords:** channel video freezing, video, QoE, MOS, subjective testing

## 1. INTRODUCTION

The fixed voice telephony market continues to decline as mobile and IP-based fixed services replace traditional fixed PSTN services. Incumbent operators are looking to multiple play strategies, including selling media content through IPTV services, for new streams of revenue. Providing multiple play bundles of services is also expected to reduce customer churn towards competitor operators [1]. Service providers are also rolling out video services for mobile devices. Service providers and network equipment manufacturers must first verify that video services will in fact meet user quality expectations, because video quality is the primary reason for customer churn[2].

Quality of Experience (QoE) refers to how well the video service satisfies users' expectations. The quality experienced by subscribers must be equal to or better than today's cable and satellite TV services or service providers run the risk of significant subscriber churn and the resulting loss in revenue. Furthermore, the cost of customer support is very high, so proactive measures can reduce network management costs significantly. Hence service providers are taking QoE of video very seriously.

“Measuring” QoE of video refers to testing the technical aspects that influence the subscriber's service experience. There are two fundamental areas of QoE testing:

- Channel zapping measurements,
- Media (audio and video) quality metrics.

In this paper we were only focusing on the second area. In particular we are focusing on the impact of freezing (i.e. when the image freezes for some time) on the Quality of Experience. Almost 90 percent of the telecom operators with IPTV offerings regularly experience video freezes[2]. The relation between freezing time and QoE was investigated. The QoE was expressed in terms of Mean Opinion Score (MOS) values[3].

Previous research at TNO on zapping time of video[4,5] and web browsing[6] suggests that there could be a generic way that people experience waiting time. Waiting time in that research was caused by a user action. In our experiment the waiting

Human Vision and Electronic Imaging XVI, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7865, 78650O · © 2011 SPIE-IS&T
CCC code: 0277-786X/11/$18 · doi: 10.1117/12.873390

time is not a consequence of a user action, but can occur due to inherent degradations in the transmission speed, decoding of video data, etc. However, it is possible that there are similarities in both models.

Two types of freezes are investigated. First a freeze where the image pauses, so no frames were lost (frame halt or "freezing without skipping" as defined by the Video Quality Experts Group (VQEG)[7]). This can occur in progressive download such as YouTube. In the second type of freezes, the image freezes and skips that particular part of the video (frame drop or "freezing with skipping[7]"). This can occur in broadcast streaming video where no retransmission is performed. Only single freezes were considered in this part of the work. Furthermore, the freeze was inserted at a scene change.

Most encoded videos streams use adaptive GOP structures for grouping consecutive frames. Such a GOP starts with an I-frame, a full reference frame and is followed by B- and P-frames. Loss of such a full reference frame generally leads to a freeze of the complete GOP. Encoders that use adaptive GOP structures insert a reference frame at a scene change. Loss of a reference frame can lead to a frame drop. Therefore research on freezes on scene changes is relevant. The reference frames are relatively large compared to B- and P-frames. It is therefore likely that a rebuffering takes place when downloading a reference frame.

Measuring the Mean Opinion Scores (MOS) was done by subjective testing. A total of 23 persons were asked to watch video sequences and assess them. Video sequences of 20 seconds were displayed for different types of content. The four content types were computer rendered animation, an action movie, "talking heads" and sports. The single freeze occurred during a scene change.

Pastrana – Vidal et al. (2004)[8] suggested a relationship between the probability of detection and the duration of the dropped frames. They also found that it is important to consider not only the duration of the freeze but also the number of freeze occurrences. Using their relationship between the total duration of the freeze and the number of occurrences, we propose a model for multiple freezes, based upon our model for single freeze occurrences.

For evaluating the multiple freeze model, we designed a subjective test to evaluate the metric performance by comparing the metric predictions against the video quality scores given by viewers.

## 2. SINGLE FREEZE

### 2.1. Single freeze experiment

For the single freeze experiment 23 test subjects were selected. 11 test subjects participated in the Netherlands, 12 test subjects participated in Sweden. The group of test subjects consisted of 14 males and 9 females, with ages ranging from 22 to 60 years and forming a mix of expert and non-experts viewers. The test subjects were not paid for their services. The ITU-T 5 point Absolute Category Rating (ACR) scale[3] was selected, see Table 1 below.

Table 1: The ITU-T 5 point ACR scale

| MOS | Quality |
|---|---|
| 5 | Excellent |
| 4 | Good |
| 3 | Fair |
| 2 | Poor |
| 1 | Bad |

The videos were obtained in the following fashion. First official DVD's in PAL format were ripped. Audio was not used in the experiment so the audio stream was left out. Then the VOB files were converted to uncompressed AVI files. From these files, 20 second snippets were created. In other subjective tests 10 seconds videos have been used, to prevent the forgiveness effect[9]. This effect causes users to give higher ratings when there is a larger time between distortion and rating period. In this experiment the maximum length of a freeze was 3 seconds. This would take a large portion of the 10 second snippet. Therefore longer snippets of 20 seconds were constructed. In those 20 seconds snippets, two types of distortion were created. The frame rate of the videos was 25 fps. In one set, a frame at approximately 10 seconds was copied multiple times. In this way, videos with 0.120s, 0.200 s, 0.520s, 1 s, 2s and 3s frame halts were created. The videos were also cut to get a length of 20 seconds. In the second set, a frame was copied at approximately 10 seconds.

This frame was pasted over multiple consecutive frames. In this way, videos with 0.120s, 0.200 s, 0.520s, 1 s, 2s and 3s frame drops were created. All videos with distortions and the original snippets were compressed with a MS-WMV9 codec. Note that compression was required because uncompressed movies would have too large data rates to be compatible with most DVD drives.

For the freezing video experiment, an interface in the Internet Explorer 7 browser was used. The videos were shown inside a browser, together with buttons for assessing the videos. First a training session was displayed. After that the real test was shown. At the end of the test, all data could be downloaded to an Excel-file. The interface was generated using a locally run apache web server with MySQL database. This meant that no network errors were introduced, for more details see van Kester (2009)[10].

The hardware used was a laptop with the following specs: Dell Latitude D505, Pentium M 1.6 GHz, 512MB RAM, windows XP SP3, 1400x1050 pixels screen resolution (native resolution of the screen).

The laptop was placed in a living room at TNO and Acreo under normal light conditions. The laptop was viewed in a lean forward position, at about an arm length distance, but this was not particularly controlled

The experiment consisted of two parts, a training experiment and the actual experiment.

During the training session, a test subject was shown five videos:

- Action movie with 3 s frame halt
- Sport movie without distortion
- Animation movie with 1 s frame drop
- Talking heads movie with 0.2 s frame halt
- Sport movie with 2 s frame drop

In this way test subjects could get used to the kind of distortion they could expect and familiarize themselves with applying the ITU MOS scale. After the training session the actual experiment started.

During the actual experiment test subjects assessed a total of 52 videos. The design consisted of four different video content types, two different distortion types (frame halt and a frame drop), crossed with 6 freezing lengths (0.12 s, 0.20 s, 0.52 s, 1.0 s, 2.0 s, 3.0 s). Together with the four undistorted videos this adds up to 52 videos.

### 2.2. Results

#### 2.2.1. Single freeze experiment

The procedure for screening the observers mentioned in ITU-R BT.500[11] was performed to see if test subjects needed to be eliminated from the data set. Based on this analysis test subject 17 should be excluded from the dataset. However the test should only be performed for relatively small groups (e.g. 20 observers) whom are all non-expert. In our experiment a combination of expert and non-expert viewers was used and the group size is slightly over the maximum (23 test persons instead of a maximum of 20 persons).

This observation should be taken into account when comparing the relatively small groups of Stockholm and Delft, in which the group consists of 11 and 12 test persons, respectively.

Figure 1 shows that for all content types there is a negative relation between the duration of the distortion and the perceived quality. This means that the longer the distortion, the lower the rating by the subject, which is expected.

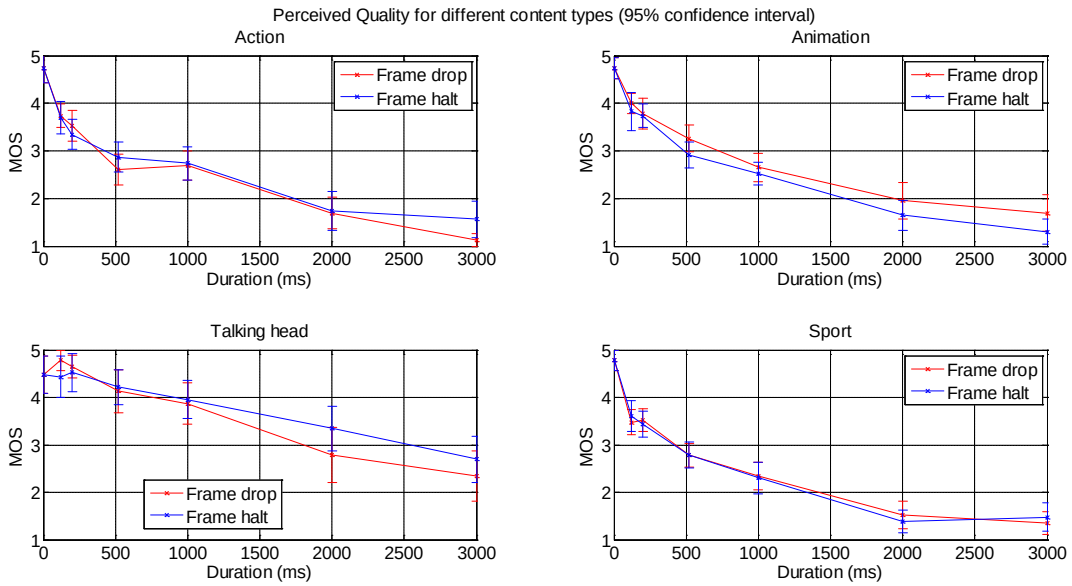


Figure 1: Perceived Quality for different content types

In Figure 1 the results for each content type is shown. For each content type, frame halt and frame drop follow similar shapes. This indicates a large correlation between frame halt and frame drop. This implies that test subjects do not perceive a large difference in quality between the two types of degradation.

We can also deduce from Figure 1 that the talking heads movie has the highest overall rating for frame drop and frame halt. This is probably due to the low movement in the video.

In Pastrana-Vidal et al (2004)[8] a 3 way repeated analysis of variance was performed to check if variables such as content and duration have main or interaction effects. The Mauchly's Sphericity Test was performed to check if the assumptions for performing a repeated ANOVA are valid. The assumption of equal variances proved to be invalid; therefore a regular repeated ANOVA could not be performed.

In the next section we will propose one objective model for the perceived quality. As in Pastrana-Vidal et al (2004)[8]. we fit a Logistic model to the subjective data.

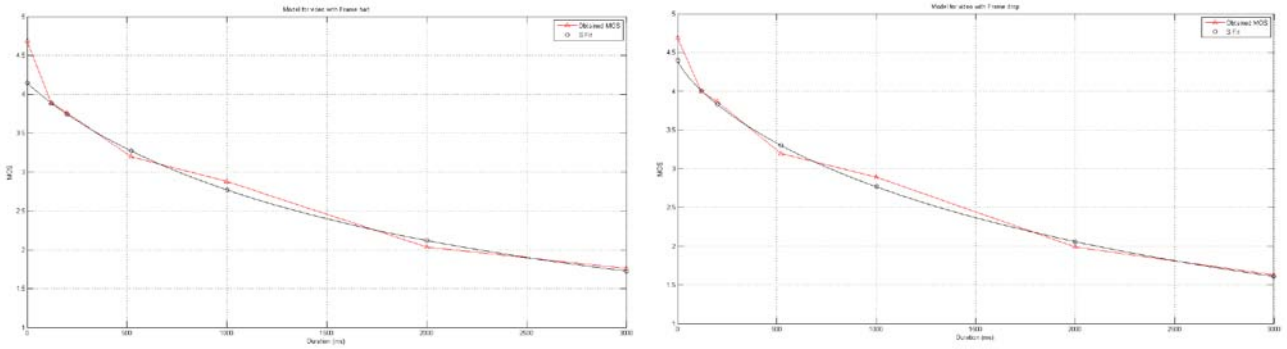

Figure 2: MOS from frame halt and the model predictions (left). MOS from frame drops and the model predictions (right)

### 2.2.2. Single freeze model

Because Frame drop and Frame halt follow very similar curves, a combined model for single freeze can be constructed.

$$MOS_{s-fit} = 4.3971 - \frac{6.3484}{1+\left(\frac{4400}{t}\right)^{0.72134}}, 0 \le t \le 3000ms$$

In Figure 3 the single freeze model fit is shown, to the left model predictions are plotted against the freeze duration time and to the right a scatterplot between model prediction for each video clip in test and the its corresponding MOS is plottedThe correlation for the model compared to MOS values for the different content types were (frame drop value first and then frame halt value) : General (0.99, .0.99), Action 0.98, .0.97), Animation (0.99, 0.99), Talking head (0.95, 0.95), Sport (0.97, 0.97). The overall correlation as depicted by the scatterplot in Figure 3 (right) was 0.88.

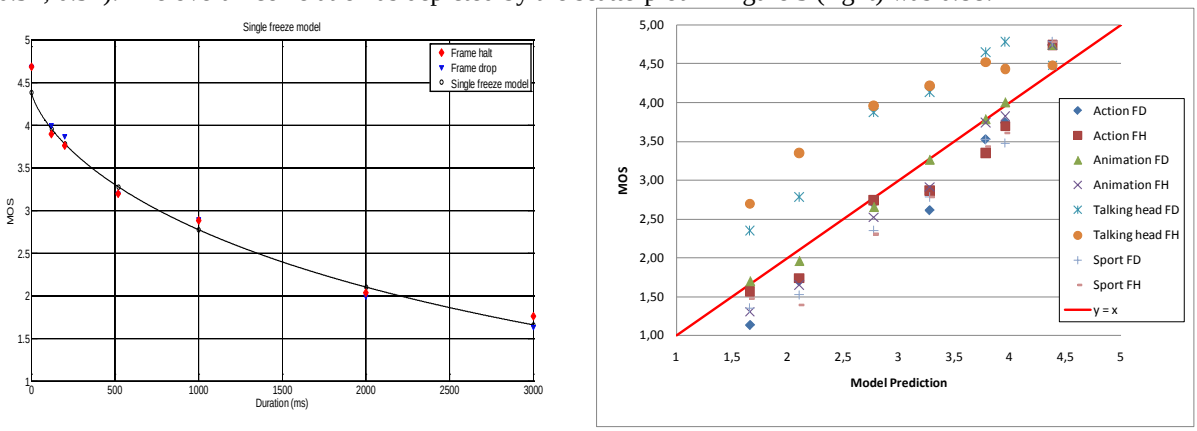


Figure 3: (Left) MOS values for Frame drop, Frame halt and the general freezing model. (Right) Scatterplot of the model predictions for each video clip and its corresponding MOS.

## 3. MULTIPLE FREEZE

An experiment by Pastrana – Vidal et al. (2004)[8] was conducted to quantify the effect of several dropped frames in video clips. Video clips containing one or more freezes with each one having a length of either 160ms or 280ms were used in the test. The number of freezes were specifically set to 1, 3, 5, 8 freezes respectively. The result can be seen in Figure 4.

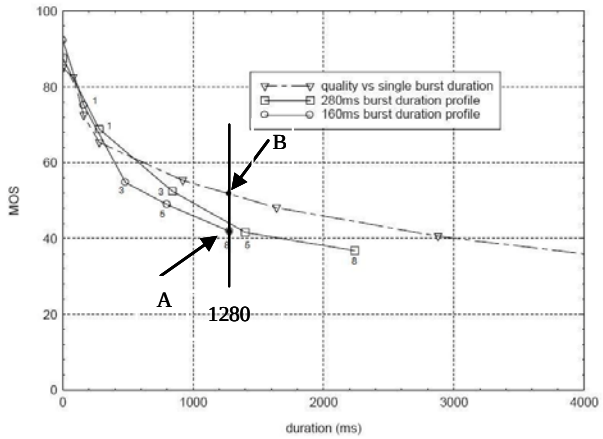


Figure 4: MOS vs. duration of different number of freezes(from Pastrana – Vidal et al. (2004)[8])

They showed in this experiment that with the same total freeze duration, the case were there are more instances of freezes, the quality were rated lower by the observers. For example, in Figure 4, at point A, there are 8 freeze instances each of which was 160ms and the total freeze duration was 1280ms. At point B, the curve indicates the corresponding point for one single freeze with the same total duration as point A. It can be observed that the MOS at point A was lower than B. Thus it can be concluded from their experiment that the number of freeze instances should also be considered.

We would like to find a way to take this effect into consideration. It can be observed from Figure 4 that the shape of the MOS curves, with different number of freeze instances in the video, are the same. The idea we propose is to, when calculating MOS for several dropped frames in one video clip, assign the total duration of single dropped frames to this video, and the effect of this duration for single dropped frame should be equal to the accumulated effect of all the dropped frames in that video clip. Thus, after mapping the duration of several dropped frames into a new duration of single dropped frame, we can use the new duration in the above described equations for single freeze to calculate MOS.

Furthermore, the new dropped frame duration should be based on the number of dropped frames and the total duration of all the dropped frames in a video clip. We assume that we can write it as:

$$duration\_new = duration\_total \times f(n)$$

where *duration_new* is the single freeze duration time that has the same effect as the multiple freeze, *duration_total* is the total freezing time and $f(n)$ is a function of the number of freeze instances *n.* See Tong (2010)[12] for a more detailed treatment. $f(n)$ was estimated to be $(n)^{1/2.16}$ or $\sqrt[2.16]{n}$ . Then if we apply this to our combined single freeze model, it becomes:

$$MOS = 4.4004 - \frac{5.5906}{1+\left(\frac{3011.5}{t \times \sqrt[2.16]{n}}\right)^{0.8021}}, 0 \le t \le 3000ms$$

### 3.1. Multiple freeze experiment

For evaluating the performance of the combined single freeze model that has be adjusted for multiple freeze, we performed a small subjective test.

In the multiple freeze experiment, six 8 seconds long video clips were used. 3 clips had a frame rate of 30 frames per second (fps) and 3 clips had a frame rate of 25 fps. Each video clip had pixel count of 352*288 pixels i.e. CIF format. The content of the video clips were a mixture of different content containing news, concert, duck, car, soccer, and jogging. All the video clips contained 1, 2, 3, 5, and 8 freeze occurrences. There were several durations for each freeze. Four observers having normal or corrected to normal vision, participated in the test. The test software used was AcrVQWin1.0[13]. The subjective test used an absolute category rating (ACR) method which is described in ITU-T Rec. P.910[3]

Table 2: Duration of freeze occurrences and total duration of freeze in 30fps video clips

| Number of freeze occurrences | duration of each freeze occurrence (ms)/total freeze duration(ms) | | | | | |
|---|---|---|---|---|---|---|
| 1 | 67/67 | 133/133 | 400/400 | 800/800 | 1600/1600 | 3200/3200 |
| 2 | 67/133 | 133/267 | 200/400 | 400/800 | | |
| 3 | 67/200 | 133/400 | 267/800 | 533/1600 | | |
| 5 | 67/333 | 133/667 | 333/1667 | 667/3333 | | |
| 8 | 67/533 | 133/1067 | 200/1600 | 400/3200 | | |

Table 3: Duration of freeze occurrences and total duration of freeze in 25fps video clips

| Number of freeze occurrences | duration of each freeze occurrence (ms)/total freeze duration(ms) | | | | | |
|---|---|---|---|---|---|---|
| 1 | 80/80 | 160/160 | 480/480 | 960/960 | 1920/1920 | 3840/3840 |
| 2 | 80/160 | 160/320 | 240/480 | 480/960 | | |
| 3 | 80/240 | 160/480 | 240/720 | 640/1920 | | |
| 5 | 80/400 | 160/800 | 400/2000 | 800/4000 | | |
| 8 | 80/640 | 160/1280 | 240/1920 | 480/3840 | | |

### 3.2. Multiple freeze results

We have computed the MOS from the experiment and the predictions of the model. Displayed in Figure 5 are the comparison of 1 and 3 freeze instances. In Figure 6 the comparison of MOS and model prediction for 5 and 8 freeze instance are shown. In Figure 7 all cases in the test are compared between MOS and model predictions. It can be noted that some data points are out of bounds and the model should be adjusted for this not to happen, but we wanted to show

that the adjustment factor works well enough when applied unoptimized directly into the single freeze model, so we did not change the model for this. The value of the Pearson linear correlation coefficientwas 0.82.

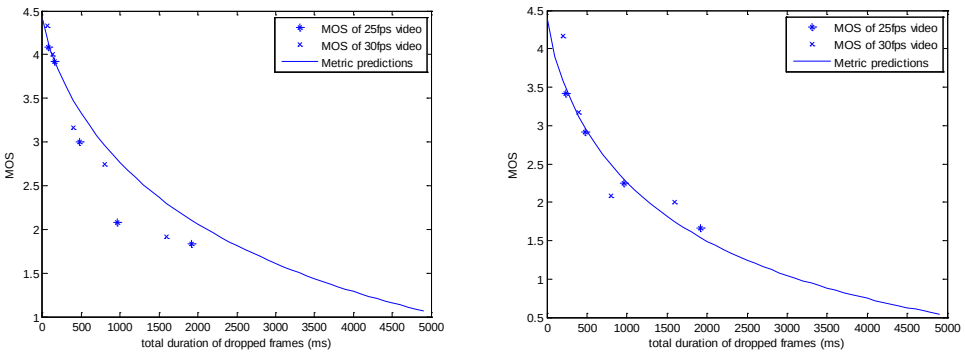


*Figure 5: MOS and multiple freeze model predictions vs. duration for 1 freeze instance (left) and 3 freeze instances (right)*

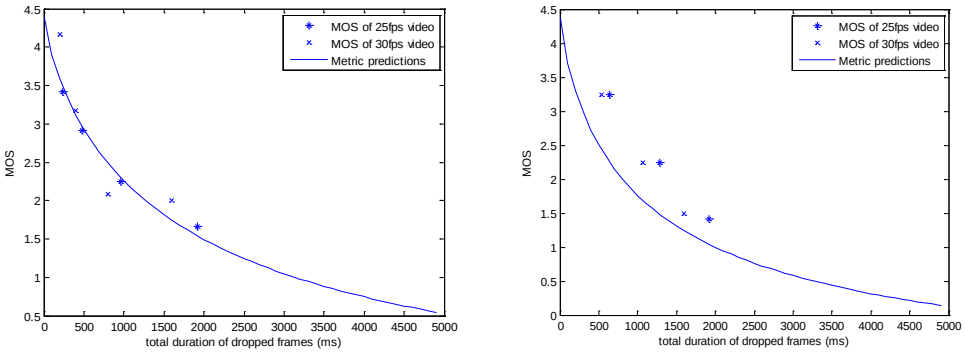


*Figure 6: MOS and multiple freeze model vs. duration for 5 freeze instance (left) and 8 freeze instances (right)*

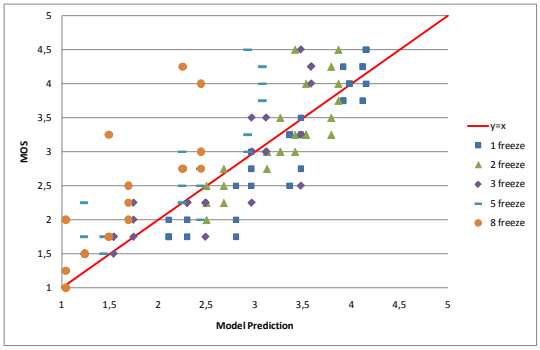


*Figure 7: MOS vs. multiple freeze model predictions*

## 4. DISCUSSION

A comparison was made with three other studies: ITU-T G.1030[6], Pastrana et al.(2004)[8], and Huynh-Thu and Ghanbari (2009)[14]. A one-on-one comparison proved to be difficult because there were differences in the research setup of the different studies. Comparison in Figure 8 should be done with caution and this figure can only be used to compare trends.

### 4.1. . ITU-T G.1030

The standard model of ITU-T G.1030[6] is given below. This is the MOS model for web-browsing applications

$$MOS_{G.1030} = \min(5, \frac{4}{\ln(Min/Max)}(\ln(Sessiontime) - \ln(Min)) + 5)$$

We applied the ITU-T G.1030 model is by using the following parameters.

- Min = 0.120 seconds
- Max = 3 seconds

This results in the following model:

$$MOS_{G.1030} = \min(5, -3.7573 \cdot \ln(Sessiontime) + 7.9665)$$

Where *Sessiontime* denotes the freezing time.

### 4.2. Sporadic Frame Dropping impact

The goal of the experiment 1 in Pastrana-Vidal et al (2004)[8] was to characterize the effect of sporadically dropped frames on perceived quality under several controlled conditions, which is similar to our goal. There are a few differences in the setup of the experiment:

- An explicit and hidden reference was given. In our experiment only a hidden reference was used.
- A 100 point scale was used, instead of our 5 point scale.
- The freezes occurred away from the scene change. In our experiment the freezes occurred during a scene change.
- The test subjects are allowed to view the videos as many times as they want and are allowed to change the scores. In our experiment all videos are viewed only once and the rating cannot be changed.
- Different content was used.
- The videos are shorter; 10 seconds sequences instead of our 20 seconds.
- The maximum freeze time is 5040 ms. In our experiment freezes up to 3000 ms occur.

### 4.3. Asymmetrical Temporal Masking near Video Scene Change

Huynh-Thu and Ghanbari (2009)[14] assessed the impact of frame freezing impairment on the perceived video quality using a variety of source content and freezing events of different durations placed at different locations in the video. Here we have only compared with the single freeze part of their model.

For our comparison the perceived quality when freezing overlaps a scene change is important. This experiment resembles our research quite a lot. However, there are some differences in the research setup.

- 10 Content types are considered, instead of the 4 content types in our experiment
- Maximum duration of a freeze is 0.8 s, our freezes can take up to 3 s.

It is interesting to note that none of the data points is below MOS = 3.5, which means that in their experiment, the mean opinion is always at least acceptable.

### 4.4. Comparing the results of the different studies

In Figure 8 (left) different model for single freezing are plotted. The curves follow a similar shape, but the differences in MOS scores at 3000 ms are large. The perceptual reasons for these differences are not clearly understood, but we believe that this could partly be explained by the differences in research setup.

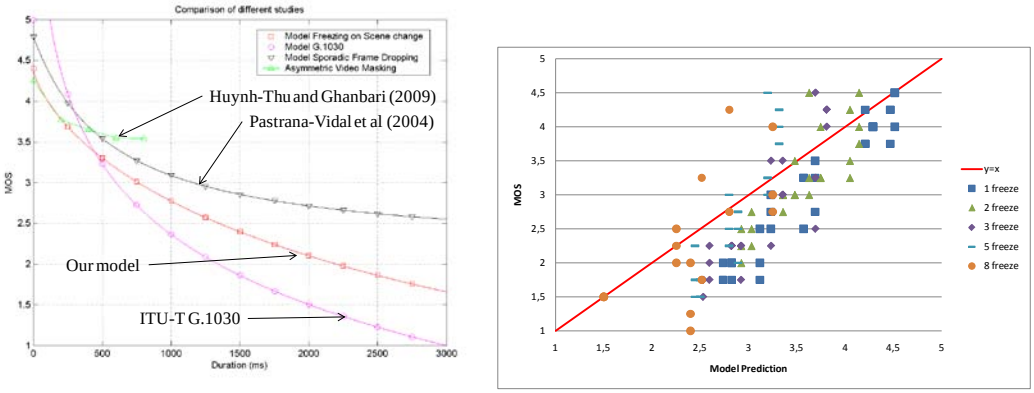


Figure 8: Comparison of different studies on freezing. Left graph show different single freeze models and right graph shows Pastrana-Vidal et al (2004)[8] single freeze model with our multiple freeze extension.

Differences between our freezing model and the ITU G.1030 model for web browsing can be explained by differences in setting. In web browsing the waiting time is caused by a user action, freezes occur without user interaction.

Larger similarities between our freezing model and the sporadic frame dropping model were expected, because both describe similar video freezing phenomena. However, the Root Mean Square Error (RMSE) is similar for web browsing experiment and the sporadic frame dropping experiment (0.5428 and 0.5429 respectively). The difference between the sporadic frame dropping experiment and our experiment can be explained by the large number of differences in research setup.

It is interesting to note that all models cross the MOS = 3.5 at a duration lower than 0.53 sec. Our experiment gives a sharper threshold of 0.36 sec (see Table 4.).

Table 4: Acceptable freezing time

| **Model** | **MOS = 3.5** |
|---|---|
| Combined single freeze model | *0.36 seconds* |
| Sporadic frame dropping [10] | *0.53 seconds* |
| G1030 | *0.40 seconds* |

### 4.5. Multiple freeze comparison

We compared our results of multiple freezing with the extension of the model of Pastrana-Vidal et al (2004)[8], see Figure 8 (right). The metric predictions show similiar performance with the MOS from people, as our model. The value of Pearson linear correlation coefficient was 0.81

It should be noted for both these models no extra tuning of the parameters were performed when adding the multiple freeze extension.

## 5. CONCLUSIONS

This work has studied the impact of a single freeze as well as multiple freezes on the perceived quality of the users. A freeze could be of two different types i.e with or without skipping, which means that frames are lost or not. A subjective test was performed to study how these freezes impact the quality perceived by the user. A combined model for these two types of freeze was proposed. We further noted, based on published work[8], that perceived quality depends not only on the duration of the freeze, but also on the number of freeze instances. A correction factor has been derived to take this effect into consideration. We also compared our models with already proposed models and some differences were noted.

It is very interesting to look at the time predicted for when the freeze becomes unacceptable and the combined single freeze model predicts this time to 0.36 sec and the other models are slightly more forgiving. This is in line with the findings in our earlier studies related to the perceived quality of zapping time[4,5]

## 6. ACKNOWLEDGEMENT

In Sweden, the work was financed by VINNOVA (The Swedish Governmental Agency for Innovation Systems). The participation of the observers are gratefully acknowledged.

## 7. REFERENCES

[1] ITU IPTV Focus Group, "Driving the Future of IPTV", http://www.itu.int/osg/spu/stn/digitalcontent/4.9.pdf, International Telecommunication Union (ITU), Place des Nations, 1211 Geneva 20, Switzerland , (2008)

[2] Winkler, S., Measuring Quality of Experience for successful IPTV deployments [on-line], http://images.tmcnet.com/expo/west-08/presentations/iptv03-winkler-symmetricom.ppt, Accessed: 5 Jan. 2011

[3] ITU-T, "Subjective Video Quality Assessment Methods for Multimedia Applications", ITU-T Rec. P.910, International Telecommunication Union, Telecommunication standardization sector, (1999)

[4] Kooij, R., Ahmed, K., and Brunnström, K., "Perceived quality of channel zapping", Vol: *Proc. of 5th IASTED International Conference on Communication Systems and Network, August 28-30, 2006*, (2006)

[5] Kooij, R., Nikolai, F., Ahmed, K., and Brunnström, K., "Model validation of channel zapping quality", *Proc. of SPIE-IS&T Human Vision and Electronic Imaging XII*, Vol: *7240*, B. Rogowitz and T. N. Pappas Eds., paper 31 (2009)

[6] ITU-T, "Estimating End-to-End Performance in IP Networks for Data Applications", ITU-T Rec. G.1030, International Telecommunication Union (ITU), Place des Nations, 1211 Geneva 20, Switzerland , (2005)

[7] VQEG, "Final Report From the Video Quality Experts Group on the Validation of Objective Models of Multimedia Quality Assessment, Phase I", VQEG Final Report of MM Phase I Validation Test, Video Quality Experts Group (VQEG), (2008)

[8] Pastrana-Vidal, R. R., Gicquel, J., Colomes, C., and Hocine, C., "Sporadic Frame Dropping Impact on Quality Perception", *Proc. of SPIE-IS&T Human Vision and Electronic Imaging IX*, Vol: *5292 (paper 15)*, B. Rogowitz and T. N. Pappas Eds., 182-193 (2004)

[9] Seferidis, V., Ghanbari, M., and Pearson, D. E., "Forgiveness Effect in Subjective Assessment of Packet Video", *Electronics Letters* **28,** 2013-14 (1992)

[10] van Kester, S., "Impact of Freezing on Perceived Video Quality", 35144, TNO, Delft, The Netherlands , (2009)

[11] ITU-R, "Methodology for the Subjective Assessment of the Quality of Television Pictures", Rec. ITU-R BT.500-11, International Telecommunication Union, Radiocommunication Sector, (2002)

[12] Tong, X., "Video Quality Measurement for In-Service Monitoring", acr049405, Acreo AB, Electrum 236, 16440 Kista, Sweden , (2010)

[13] Jonsson, J. and Brunnström, K., "Getting Started With ArcVQWin", acr022250, Acreo AB, Kista, Sweden , (2007)

[14] Huynh-Thu, Q. and Ghanbari, M., "No-reference temporal quality metric for video impaired by frame freezing artefacts", Vol: *Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt*, 2221-2224 (2009)