Video Quality Evaluation for Internet Streaming Applications

Stefan Winkler[a] and Ruth Campos[b]

[a]Audiovisual Communications Laboratory and [b]Signal Processing Laboratory
Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland

ABSTRACT

We carried out a number of subjective experiments using typical streaming content, codecs, bitrates and network conditions. In an attempt to review subjective testing procedures for video streaming applications, we used both Single Stimulus Continuous Quality Evaluation (SSCQE) and Double Stimulus Impairment Scale (DSIS) methods on the same test material. We thus compare these testing methods and present an analysis of the experimental results in view of codec performance. Finally, we use the subjective data to corroborate the prediction accuracy of a real-time non-reference quality metric.

Keywords: Quality assessment, multimedia, DSIS, SSCQE

1. INTRODUCTION

Quality assessment for television applications has become quite well established, as evidenced by the number of publications and products available, as well as the work of the Video Quality Experts Group (VQEG).[8,9] Video streaming over packet networks such as the Internet is an entirely different matter. It comprises a wider range of frame sizes, frame rates and bitrates, and thus exhibits a wider range of distortions. Network conditions (e.g. congestion or packet loss) are different from the ones occurring in TV transmission. Also, the content is viewed at a short distance on smaller screens with progressive display.

This paper describes the extensive experiments we conducted with different types of streaming video. It discusses the goals, the procedures, and the results of these tests. The experiments were designed to simulate typical streaming applications and viewing conditions on a PC screen. The source material and test conditions were selected with two distinct sets of video streaming applications in mind:

1. Medium-bitrate, medium-size (CIF) streaming clips (e.g. news, sports, music videos, ads);
2. High-bitrate, high-resolution film content for on-demand movie applications.

The first set is similar to the tests with different multimedia codecs conducted previously by one of the authors,[11] but the choice of codecs and network conditions made in this paper is better adapted to video streaming. A third set of tests focusing on low-bitrate video for mobile streaming applications with WCDMA bit-error patterns is described in an upcoming paper.[13]

The source material was subjected to a number of Hypothetical Reference Circuits (HRC's), comprising not only compression by the encoder, but also transmission of the video over a network with packet losses.

Subjective ratings were obtained for the resulting test sequences using two methodologies defined by ITU-R Recommendation BT.500,[5] namely Single Stimulus Continuous Quality Evaluation (SSCQE), which measures the time-varying quality of the sequences, as well as Double Stimulus Impairment Scale (DSIS), which measures the global amount of degradations perceived.

The paper is organized as follows. Section 2 lists the source material and HRC's used to produce the test videos. In Section 3 we discuss the test methods, the lab setup and the presentation of the sequences. The data obtained in the subjective experiments is analyzed in Section 4. Finally, *Stream PQoS*^{TM}, a non-reference quality metric with MOS prediction, is evaluated with these data in Section 5.

E-mail of corresponding author: stefan.winkler@epfl.ch
This work was done in part while the authors were with Genista Corporation, Tokyo, Japan.

2. TEST MATERIAL

2.1. Source Sequences

The source sequences were carefully chosen from over 20 hours of original content. Scenes were selected to be representative of two distinct sets of streaming applications:

1. Medium-size (CIF) streaming clips such as news, sports, music videos, ads;
2. High-resolution film content in wide-screen (2.35:1) format for on-demand movie applications.

The sequence selection was governed by the following considerations:

- Preferably, a sequence should not have scene cuts more frequently than once every 10 seconds.
- At least one scene must fully stress some of the HRC's in the test.
- The set of test sequences should span the entire range of coding complexity.

More specifically, the ensemble of scenes should contain the following elements:

- Flat areas, complex patterns, masking effect;
- Object and/or camera motion (zoom, pan) at different speeds;
- Objects appearing, crossing the scene, moving in different directions;
- Faces, landscapes.

A detailed description of the selected scenes is given in Tables 1 and 2. Several scenes of streaming content are taken from clips used in previous tests by MPEG-4[1] and by VQEG.[9] If necessary, clips were cropped and rescaled to the same frame size. All clips were displayed at 25 fps, independent of their original frame rate (no frame rate conversions were performed).

From each of these two sets, two source sequences were compiled, each of approximately 1 minute duration, comprising scenes A–F and G–K, respectively. These 1-minute source sequences were then processed with the HRC's described in the next section. In the SSCQE tests, the processed clips were shown to every subject in random order. For the DSIS test, the processed 1-minute sequences were broken down again into the individual scenes listed in the Table 1, and each sequence was paired with its original for evaluation by the viewers.

Table 1. Scenes for streaming content (360×288, 25fps); duration in seconds:frames

Scene #	Description	Characteristics	Duration
A	Letters	Letters with different colors flying in all directions over dark background	10:10
B	News	Male and female speaker in newsroom, almost still	11:23
C	F1 car	Object motion, camera following car, 2 angles	8:20
D	Fast food	Texture, people, fast pans, 2 angles	8:20
E	Coastguard	Two boats crossing on river, medium motion, water motion	11:24
F	Balloons	Amusement park, saturated colors, people, motion	8:05
G	Foreman	Talking head, with pan to construction site, geometric shapes	16:00
H	New York	Slow city flyover, skyscrapers at sunset, detailed texture	10:10
I	Football	Fast camera and object motion, colors	10:10
J	Live concert	Dark scene, spotlights, 3 angles	13:15
K	Cartoon	Characters dancing through scene, with pan	12:14

Table 2. Scenes for film content (844×360, 25fps); duration in seconds:frames

Scene #	Description	Characteristics	Duration
A	Movie credits	Text on forest flyover towards city skyline	15:03
B	City street	Man leaving shop, walking around building, detail, camera pan	12:01
C	Action	Helicopter crashing into building, explosion, 3 angles	9:12
D	Country road	Camera on car following road, 2 angles	6:17
E	Casino outdoors	Car driving up to casino at night, camera pan, object motion, detail	9:22
F	Casino indoors	People passing through hall, camera follows them	9:22
G	Bridge	Pan to two people crossing bridge, faces	22:12
H	Dinner	Woman talking at dinner table, faces	7:18
I	Living room	Woman in red dress walking down stairs and across room, camera follows her	10:15
J	Desert race	SF race through canyon/desert landscape, several angles	13:02
K	CG movie	Camera pan over characters, very colorful, fade to other scene	9:14

2.2. Hypothetical Reference Circuits (HRC's)

The HRC's were chosen to be representative of the most common applications of video streaming over the Internet. Two sources of artifacts were taken into account (see Figure 1):

- Video compression (source encoding);
- Transmission over a packet network.

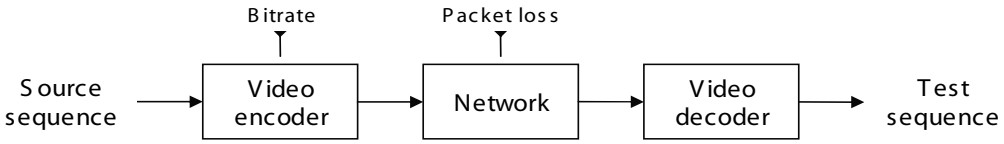


Figure 1. HRC generation chain

Although this chain appears simple, many configurations are possible. At the source encoding stage, the following encoders and video formats were used:

- Windows Media Video 8,*
- Real Video 8,*
- ISO MPEG-4[2] (Microsoft implementation).†

The encoding and transmission parameters were chosen such that they would deliver typical video quality for each of the two applications and at the same time achieve a good distribution of qualities for the different scenes. For the medium-size streaming content, this comprises bitrates of 256 kb/s (except for MPEG-4, whose quality was not satisfactory at this bitrate) and 512 kb/s. For high-resolution film content, bitrates of 512 kb/s and 1 Mb/s were selected.

At the transmission stage, an IP network simulator (SHUNRA\Cloud) was used to simulate different network conditions. Specifically, different packet loss rates (PLR) were selected for each set of applications. Many simulations were necessary to find loss rates and cases that resulted in interesting test videos. It can be noted

* Versions 9 of the Real and Windows Media codecs were not yet available at the time of the tests.

† To facilitate streaming with the tools at hand, the ISO MPEG-4 codec provided with the Windows Media Encoder was used. It encapsulates the MPEG-4 stream inside the WMV file format.

that HRC's with packet losses are only present for Real Media. The reason is that only Real proved robust to packet losses in the sense that there were visible artifacts or brief frame freezes and frame drops, but after a few moments the video continued playing. The Windows Media decoder never recovered from packet losses before the end of the video, which did not result in interesting additions to the test set.

The final HRC lists are shown in Tables 3 and 4.

Table 3. HRC's for medium-size streaming content

HRC #	Codec	Bitrate	PLR
1	WM	256 kb/s	–
2	Real	256 kb/s	–
3	Real	256 kb/s	2%
4	WM	512 kb/s	–
5	Real	512 kb/s	–
6	Real	512 kb/s	3%
7	MPEG-4	512 kb/s	–

Table 4. HRC's for high-resolution film content

HRC #	Codec	Bitrate	PLR
1	WM	512 kb/s	–
2	Real	512 kb/s	–
3	Real	512 kb/s	4%
4	MPEG-4	512 kb/s	–
5	WM	1 Mb/s	–
6	Real	1 Mb/s	–
7	Real	1 Mb/s	12%
8	MPEG-4	1 Mb/s	–

At the receiving end, the respective decoders were used. The video was captured using a proprietary video stream capture tool, which keeps track of the exact timing of frame display as encountered during playback (including picture freeze and playback irregularities). It stores the received video in an uncompressed AVI and the associated time stamps in a separate log file.

3. SUBJECTIVE ASSESSMENT

3.1. Assessment Methods and Tools

The following two assessment methods specified in ITU-R Recommendation BT.500[5] were used in the subjective experiments:

1. In the Single Stimulus Continuous Quality Evaluation (SSCQE), a series of video sequences is presented once to the viewer. The video sequences may or may not contain impairments. Subjects evaluate the *instantaneous* quality in real time using a slider with a continuous scale.

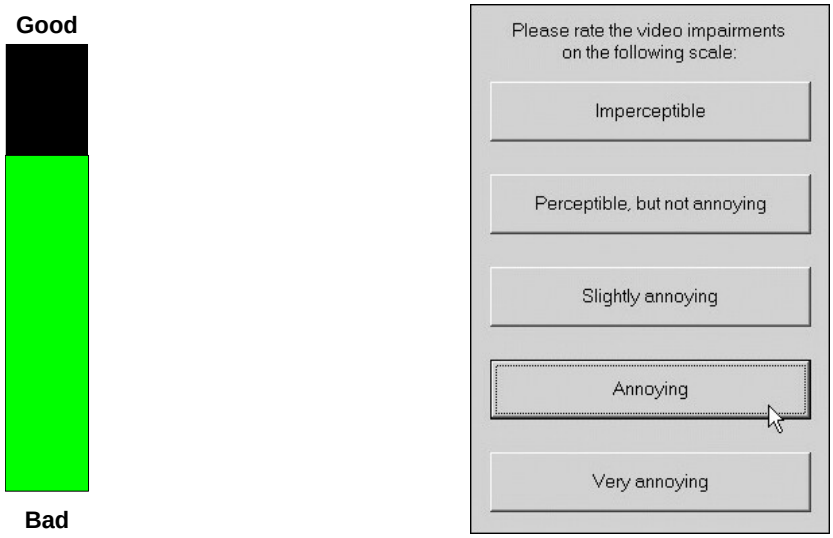
2. In the Double Stimulus Impairment Scale (DSIS) test, viewers are shown multiple sequence pairs consisting of a "reference" and a "test" sequence, which are rather short. Subjects rate the overall amount of impairment in the test sequence on a discrete five-level scale ranging from "imperceptible" to "very annoying".

The SSCQE method yields quality ratings at regular time intervals and can thus capture the perceived time variations in quality. The ratings are absolute in the sense that viewers are not explicitly shown the reference sequences. This corresponds well to an actual home viewing situation, where the reference is not available to the

viewer either. The DSIS method only yields one rating per clip, and viewers evaluate the *relative* degradation of the video with respect to the reference, which is considered an easier task.

ITU-T Recommendation P.910[6] has a clearer focus on multimedia applications than ITU-R Rec. BT.500. It defines a Degradation Category Rating (DCR) method very similar to DSIS, but it does not discuss single stimulus tests. Viewing distance restrictions are also relaxed to anywhere between 1–8 times screen height.

Slight modifications of the procedures described in the ITU recommendations were introduced to adapt them to purely PC-based testing. The slider in the SSCQE test was not a stand-alone hardware device, but a graphical on-screen slider that was steered by moving the mouse up and down, i.e. vertical mouse movements were translated directly into slider shifts. We found this to give viewers a good haptic feeling of where they were on the quality scale (we also tried the scroll wheel, but rejected it for lack of absolute position feedback). People's familiarity with handling a computer mouse is an additional advantage.



(a) SSCQE slider (b) DSIS dialog

Figure 2. Voting devices designed for the subjective experiments.

The on-screen slider we designed for the SSCQE tests is shown in Figure 2(a). We decided not to attach the usual five-level scale of semantic judgment terms ("excellent", "good", "fair", "poor", "bad") on the side of the slider for two reasons: First, none of our videos could be considered "excellent" quality, given that the quality reference for non-experts today is the DVD. Second, studies found that these quality terms may lead to a nonlinear interpretation of the scale by the subjects (i.e. "excellent" and "good" may be considered closer than "poor" and "bad", for example).[10] Therefore, we only put "Good" and "Bad" at the top and bottom end of the slider for general directional guidance.

Furthermore, we decided to make the slider a bright green rectangle ranging from the bottom of the scale to the current slider position (the rest of the slider was black). In initial tests we found this representation easier to follow from the corner of the eye than a plain gray slider, thereby allowing viewers to check the approximate slider position without having to look away from the video. This was especially true for the film content in wide-screen format with an aspect ratio of 2.35:1.

In summary, we found the on-screen visual feedback of the slider position in combination with the haptic mouse feedback to be very user-friendly. Another advantage of a software slider is that it can be automatically reset without having to instruct subjects to do so. We reset the slider to the middle position at the beginning of each SSCQE session.

DSIS ratings were also entered on the PC directly, with the help of a dialog with five buttons showing the DSIS judgment terms (see Figure 2(b)). Subjects voted by clicking on one of the buttons with the mouse.

The videos were displayed with Genista's *QualiView* tool. It reads the frames contained in the uncompressed AVI captured previously and displays them on screen with the precise timing recorded during the simulations.

3.2. Presentation Structure

Instructions were given to the viewers in written form. After they had read the instructions, a training session was run to demonstrate the task that subjects had to perform as well as the range of quality to be expected. The training session was repeated for each set of tests to reset the bounds of the quality range for the subjects.

In order to minimize contextual effects, the order of the test sequences was randomized at the clip level such that every subject viewed the test clips in a different order. This is clearly an advantage of carrying out the tests on a PC — showing video from tapes, it is highly impractical to have more than one or two randomizations.

Every session was limited to 30 minutes, with one or more short breaks, depending on the test method. SSCQE in particular demands constant attention from the subjects, and we felt that frequent breaks would help reduce fatigue. Therefore, subjects were given a short break after at most 8 minutes of SSCQE testing.

The test sessions and their presentation structure are summarized in Table 5.

Table 5. Summary of experiments

Test	Streaming Content	Film Content
SSCQE session	2 sequences (Table 1) 7 HRCs (Table 3) *14 min.* of test material (with break at half-time)	2 sequences (Table 2) 8 HRCs (Table 4) *16 min.* of test material (with break at half-time)
DSIS session	11 sequences (Table 1) 7 HRCs (Table 3) *28 min.* of test material (with break at half-time)	—

3.3. Viewing Conditions

Viewing conditions comply as much as possible with those described in ITU-R Rec. BT.500[5] and ITU-T Rec. P.910,[6] with the necessary modifications of the laboratory set-up according to typical user requirements and conditions for the display of streaming video.

Streaming video on a PC is typically viewed by a single person only. For our test material, we found subjects to be comfortable at a viewing distance of about 3-4 times the height of the video picture.

The monitors used in the subjective assessments are LCD screens. This is motivated by the fact the majority of home PC systems that are bought nowadays have an LCD screen, and that laptops or other mobile devices also use LCD screens. The specific screen used, a 15" Sony SDM-S51, has the following specifications:

Resolution:	1024×768
Dot pitch:	0.297 mm
Peak luminance:	250 cd/m^2
Contrast ratio:	300:1
Viewing angles:	120° horizontal, 90° vertical
Response times:	10 ms (rise time), 20 ms (fall time)

After calibration and black-level adjustment, the screen properties were measured to be as follows:

Gamma:	2.2
Color temperature:	6400 K
White luminance:	77 cd/m^2
Video surround:	20 cd/m^2

The laboratory setup is shown in Figure 3. There was no additional light source in the viewing room.

Figure 3. Laboratory setup

3.4. Viewers

20 non-expert viewers – mostly university students – participated in each of the test sessions; some of them participated in more than one session (on different days). Prior to their first test session, each viewer was screened for the following:

- Normal (20/20) visual acuity or corrective glasses;
- Normal color vision (per Ishihara test);
- Sufficient familiarity with the language to comprehend instructions and to provide valid responses using the semantic judgment terms.

4. SUBJECTIVE DATA ANALYSIS

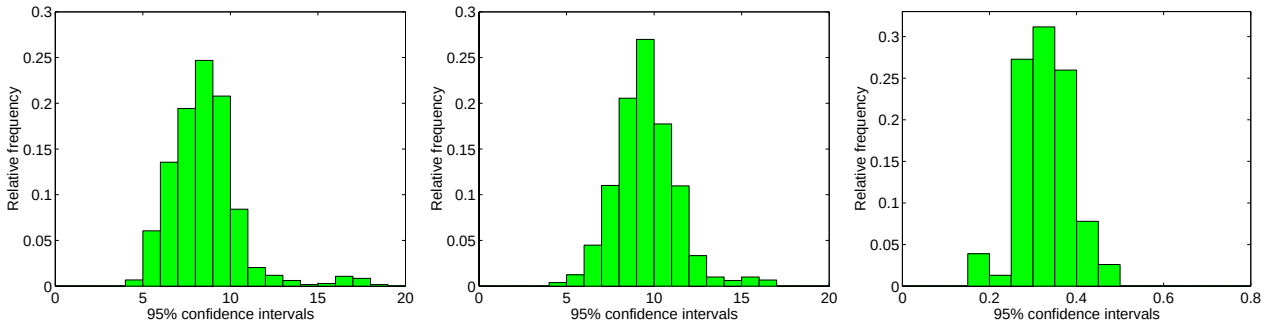
The validity of the subjective test results was verified by screening the observers according to Annex 2 of ITU-R Rec. BT.500. Subsequently, the Mean Opinion Scores (MOS) and the 95% confidence intervals of the subjective ratings were computed. To facilitate numerical analysis and plots, the DSIS ratings are mapped onto a MOS scale from 1 to 5, where 1 indicates the worst quality ("very annoying"), and 5 the best ("imperceptible").

4.1. DSIS-SSCQE Comparison

As a quality indicator of the subjective data, the distributions of the 95% confidence intervals for the different tests are shown in Figure 4. In the SSCQE experiments with the streaming content, the average size of the confidence intervals is ±8.5 on the 0-100 scale, compared to ±9.5 for the film content. This is evidence of the higher variability and faster changes of the quality of the film test set, which oberservers considered more difficult to evaluate. In the DSIS experiment (which was only performed with the streaming content), the average confidence interval is ±0.33 MOS units. Using the mapping from DSIS to SSCQE scores determined below, this corresponds to ±6.2 units on the SSCQE scale. DSIS ratings thus exhibit a lower variability between observers than SSCQE ratings.

Figures 5 and 6 show the relationship between mean DSIS scores and mean SSCQE scores for the streaming sequences. For this comparison, the SSCQE scores were time-shifted by 1.5 seconds to compensate for the delay between the display of a video frame and the actual slider response from the subject. Evidence for this time-shift comes from the mapping of MOS predictions discussed below in Section 5 and from previous findings.[3]

In Figure 5, mean SSCQE scores are plotted against mean DSIS scores. Only temporal averaging was applied to the SSCQE scores over the parts of the sequences used in the DSIS test. Despite the lack of a hidden reference in the SSCQE experiments, there is a very good match between the two data sets, characterized by linear and rank-order correlation coefficients of 92%. This corresponds quite well to previous studies comparing SSCQE



(a) SSCQE test (streaming) (b) SSCQE test (film) (c) DSIS test (streaming)

Figure 4. Distribution of 95% confidence intervals

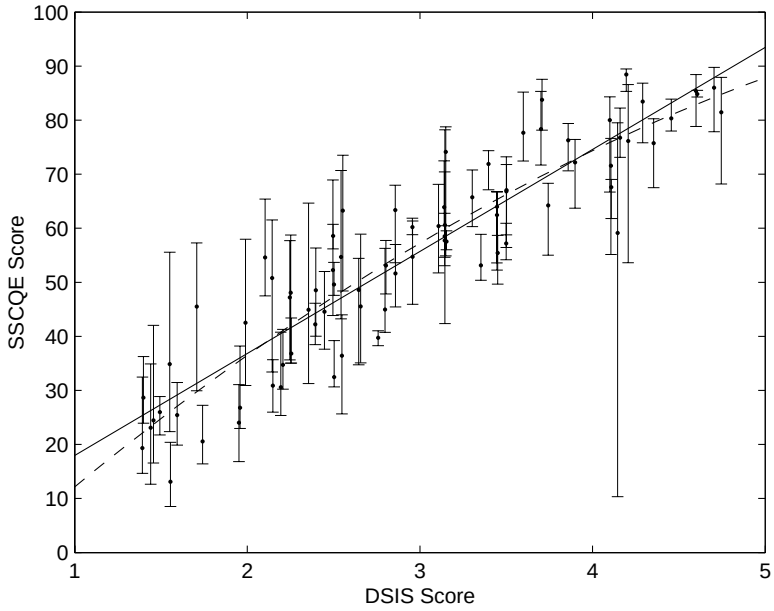


Figure 5. Comparison of DSIS and SSCQE mean scores. The vertical bars indicate the range of SSCQE scores within the corresponding DSIS sequence part; the dots indicate the average over those scores. The solid line represents a linear fit through the data (92% correlation), the dotted line results from a quadratic fit.

and DSCQS data,[4] where more complex models for the temporal pooling of SSCQE scores were used. A linear fit to these data results in the regression line SSCQE $= 18.9 \times$ DSIS $- 0.9$ and an RMSE of 7.7 (units on the SSCQE scale). The quadratic fit (indicated by the dashed line in Figure 5) underlines the slight nonlinearity of the relationship between DSIS and SSCQE ratings, especially the saturation in the high-quality scores. It is also evident that subjets were more hesitant to use the top and bottom ends of scale on the SSCQE slider than in the DSIS test.

Figure 6 shows the mean SSCQE and DSIS scores for the entire streaming test set. The above-mentioned linear transformation was applied to the DSIS scores to fit them onto the same scale as the SSCQE scores. The plots can also be used as the basis for a side-by-side codec comparison.

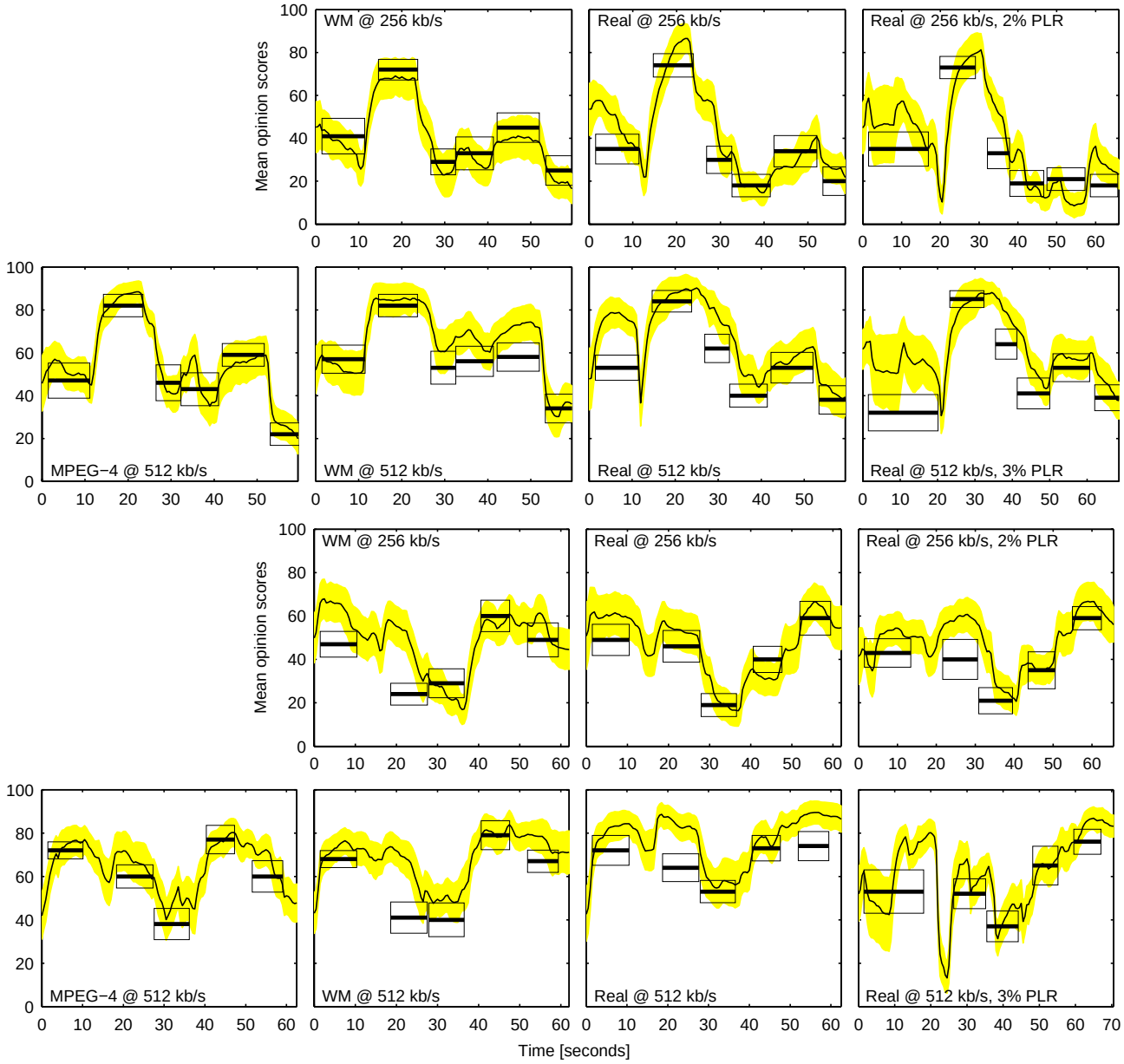
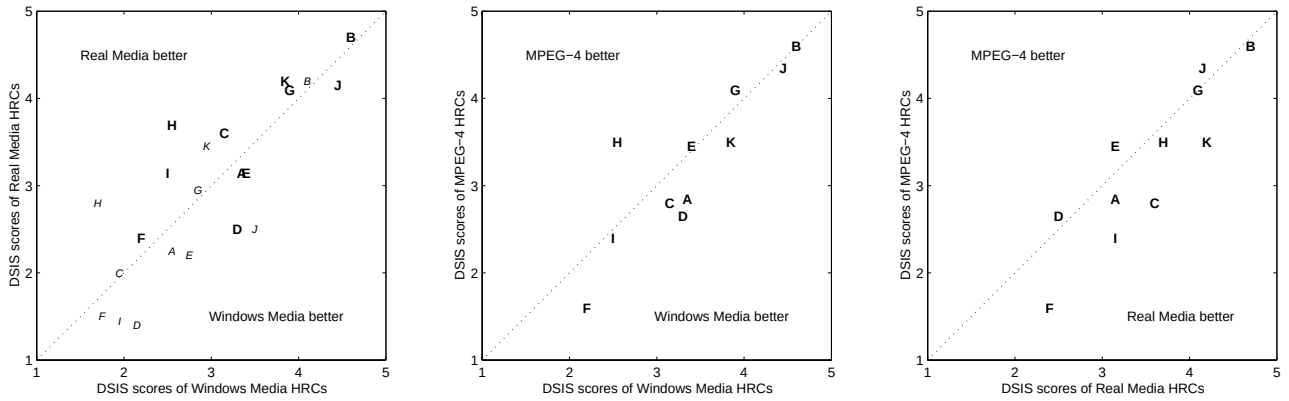


Figure 6. SSCQE ratings (smooth curves) for the 14 streaming test sequences and DSIS ratings (thick line segments) for the corresponding 77 sequence parts. Top two rows: source scenes A–F, bottom two rows: source scenes G–K. The gray bands and the hollow rectangles around the mean values indicate the 95% confidence intervals.

4.2. Codec Comparison

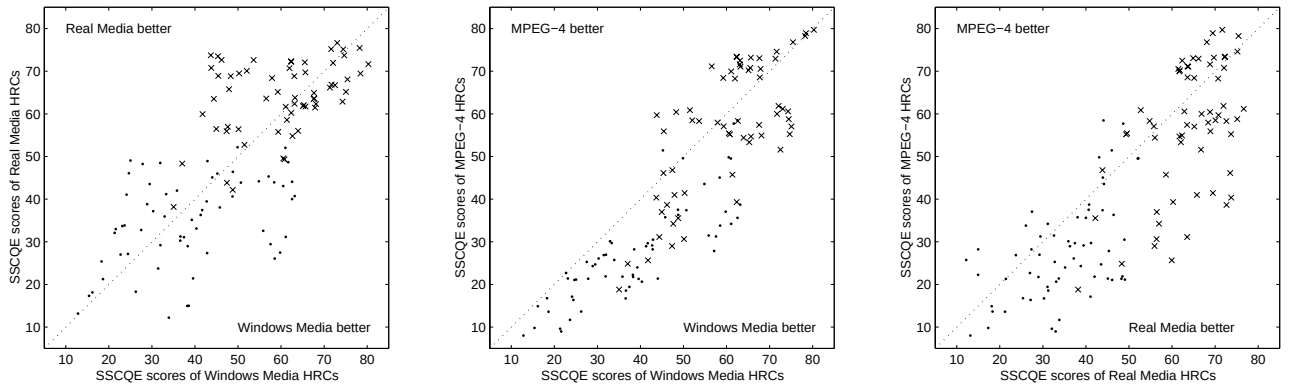
The data also allow us to compare the performance of the three codecs used in the tests. The subjective ratings for the different HRC's (without packet loss) are shown in Figures 7 and 8. It can be noted that Real and Windows Media codecs produced videos with approximately the same coding quality for our range of source material and bitrates. This is true for both streaming and film content and HRC's. The particular scene (cf. Tables 1 and 2) does have an influence on quality, which indicates that the encoder parameters chosen (in particular the trade-off between smooth motion and high detail rendition) may not be entirely comparable.

Both Real and Windows Media codecs outperform the MPEG-4 codec used in our tests, especially at the



(a) Real vs. Windows Media (b) MPEG-4 vs. Windows Media (c) MPEG-4 vs. Real Media

Figure 7. Codec comparisons for streaming content. The letters correspond to the scene numbers from Table 1. Small italic letters denote 256 kb/s HRC's, large bold letters denote 512 kb/s HRC's (only HRC's without packet losses are shown).



(a) Real vs. Windows Media (b) MPEG-4 vs. Windows Media (c) MPEG-4 vs. Real Media

Figure 8. Codec comparisons for film content. 512 kb/s HRC's are represented by dots, 1 Mb/s HRC's are represented by crosses (only HRC's without packet losses are shown).

lower bitrates. This is the reason the 256 kb/s MPEG-4 HRC was excluded from the streaming test set (the quality difference was simply too large). The problem is still present for the 512 kb/s MPEG-4 HRC with the film content, as can be seen from Figure 8. It should be noted, however, that this particular MPEG-4 codec (Microsoft's ISO MPEG-4 implementation) is not as optimized as typical commercial MPEG-4 codecs and does not represent the state of the art.

5. MOS PREDICTION

The data obtained in these experiments were used to tune and evaluate MOS predictions for typical streaming video content and conditions. These MOS predictions are based on existing non-reference metrics for blockiness,[12] blurriness[7] and jerkiness artifacts, as computed by Genista's *Stream PQoS*^{TM} software tool. These artifact metrics are rather simple, which makes it possible to compute them in real-time on a standard PC, in parallel to decoding and displaying the streamed video.

The results presented in this section are based on the SSCQE data for the streaming test set. This emphasis was given to SSCQE because of the intended quality monitoring use of the application, where instantaneous predictions of MOS and especially its time-varying behavior are very important.

The subjective and objective data must be time-aligned for the mapping. The latency that results from viewer reaction times and slider "stiffness" was eliminated from the data by computing and applying one global time shift between objective metrics and MOS data. This time-shift was found to be 1.5–2 seconds (cf. Section 4.1).

The results over all streaming test sequences are shown in Figure 9.

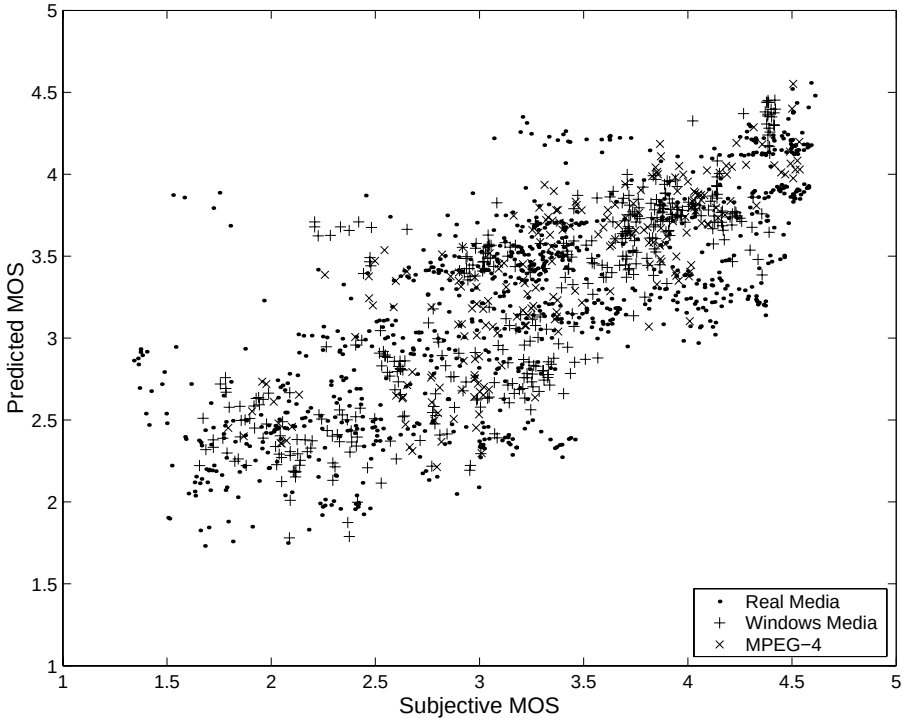


Figure 9. Predicted MOS vs. subjective MOS (both mapped onto the 1-5 scale).

Due to the different types of artifacts that are produced by the three codecs used in the tests, individual mappings were determined for each codec separately. For example, the MOS prediction for the MPEG-4 videos relies mainly on the blockiness metric.

The overall and the individual prediction qualities are summarized in Table 6. The overall quality of the MOS predictions is characterized by a correlation of 78% and an average prediction error of 0.5 MOS units, which is roughly of the same order as the confidence intervals of the subjective experiments.

Table 6. MOS prediction performance

	Linear correlation	Rank-order correlation	Prediction error
Real Media	76%	76%	0.54
Real Media (no PL)	84%	83%	0.48
Windows Media	84%	85%	0.41
MPEG-4	83%	84%	0.36
Overall	**78%**	**79%**	**0.48**

While the three rather simple artifact metrics for blockiness, blurriness and jerkiness can be successfully combined to achieve MOS predictions with relatively high accuracy, the main problem is the significant deterioration of the prediction accuracy with the inclusion of packet loss HRC's. The reason for this probably lies in the fact that people respond rather slowly to the sudden effects packet losses have on the video. This gradual viewer response obviously cannot be taken into account with memoryless metrics.

6. CONCLUSIONS

We presented a number of subjective experiments for video streaming applications. Test sequences were selected from a wide range of streaming content and created with three different multimedia codecs, typical bitrates and network conditions. Using SSCQE and DSIS methods on the same test material, we found that the data obtained with both methods is highly correlated and of comparable quality, even though there is evidence that DSIS ratings exhibit less inter-subject variability. We also introduced *Stream PQoS*, a real-time non-reference quality metric, whose MOS predictions were shown to correspond well to the subjective quality ratings. Future will focus on the mapping between DSIS and SSCQE scores as well as improvements of the quality metrics and MOS predictions.

ACKNOWLEDGMENTS

We would like to thank Genista Corporation and all the people involved in the design of the experiments, the metrics and the software described in this paper. We also thank the viewers who participated in our tests. Finally, we acknowledge the support of Prof. Sabine Süsstrunk at the EPFL's Audiovisual Communications Lab, whose testing facilities we used for the conducting the subjective experiments.

REFERENCES

1. T. Alpert et al.: "Subjective evaluation of MPEG-4 video codec proposals: Methodological approach and test procedures." *Signal Processing: Image Communication* **9**(4):305–325, 1997.
2. T. Ebrahimi, F. Pereira: *The MPEG-4 Book*. Prentice Hall, 2002.
3. R. Hamberg, H. de Ridder: "Continuous assessment of perceptual image quality." *Journal of the Optical Society of America A* **12**(12):2573–2577, 1995.
4. R. Hamberg, H. de Ridder: "Time-varying image quality: Modeling the relation between instantaneous and overall quality." *SMPTE Journal* **108**(11):802–811, 1999.
5. ITU-R Recommendation BT.500-11: "Methodology for the subjective assessment of the quality of television pictures." International Telecommuncation Union, Geneva, Switzerland, 2002.
6. ITU-T Recommendation P.910: "Subjective video quality assessment methods for multimedia applications." International Telecommuncation Union, Geneva, Switzerland, 1996.
7. P. Marziliano et al.: "A no-reference perceptual blur metric." in *Proceedings of the International Conference on Image Processing*, vol. 3, pp. 57–60, Rochester, NY, 2002.
8. A. M. Rohaly et al.: "Video Quality Experts Group: Current results and future directions." in *Proceedings of SPIE Visual Communications and Image Processing*, vol. 4067, pp. 742–753, Perth, Australia, 2000.
9. VQEG: "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment." 2000, available at http://www.vqeg.org/.
10. A. Watson, M. A. Sasse: "Measuring perceived quality of speech and video in multimedia conferencing applications." in *Proceedings of the ACM Multimedia Conference*, pp. 55–60, Bristol, UK, 1998.
11. S. Winkler: "Visual fidelity and perceived quality: Towards comprehensive metrics." in *Proceedings of SPIE Human Vision and Electronic Imaging*, vol. 4299, pp. 114–125, San Jose, CA, 2001.
12. S. Winkler, A. Sharma, D. McNally: "Perceptual video quality and blockiness metrics for multimedia streaming applications." in *Proceedings of the International Symposium on Wireless Personal Multimedia Communications*, pp. 547–552, Aalborg, Denmark, 2001.
13. S. Winkler et al.: "Video quality evaluation for mobile streaming applications." in *Proceedings of SPIE Visual Communications and Image Processing*, Lugano, Switzerland, 2003.