Subjective quality of internet Video codecs

— Phase 2 evaluations using SAMVIQ

Franc Kozamernik *(EBU Technical Department)*, **Paola Sunna** *(RAI CRIT)*, **Emmanuel Wyckens** *(France Telecom R&D)* **and Dag Inge Pettersen** *(NRK)*

In order to evaluate the performance of video codecs for the internet, EBU Project Group B/VIM has developed a new *subjective* evaluation methodology called SAMVIQ (Subjective Assessment Methodology for Video Quality). This new methodology was used recently during B/VIM's Phase 2 subjective evaluations of four codecs designed for internet use: Envivio MPEG-4, QuickTime 6, RealNetworks 9 and Windows Media 9.

This article gives a short description of SAMVIQ and summarizes the main findings of the Phase 2 subjective evaluations.

Assessing the performance of audio and video compression technologies is generally considered a very controversial subject. This is particularly true of commercial audio and video codecs designed for the internet and webcasting. There are those who consider it a science but there are also some who say it is a black art. As the codec market grows, the stakes become high; one may read bold statements from the codec manufacturers that their codec is significantly "better" than their previous version and indeed "better" than any codec marketed by their competitors. Quite often these claims are not supported by the necessary evidence about which feature of the codec has actually been improved and what are the conditions under which these measurements have been made.

In order to avoid relying on the studies of other organizations which may often have some commercial bias, EBU Members decided to perform regular independent codec tests by themselves. To this end, the EBU Broadcast Management Committee (BMC) set up two Project Groups: **B/VIM** (Broadcast / Video In Multimedia) and **B/AIM** (Broadcast / Audio In Multimedia). The evaluation results from these two groups are now among the most frequent downloads from the EBU's website [1]. As our findings are solicited worldwide, B/VIM and B/AIM tend to perform evaluation campaigns at regular intervals, around every two years or so. For example, B/VIM conducted its Phase 1 series of tests in 2002 and published its findings in EBU document BPN 055 [2]. The second series of tests using more recent video codecs was conducted during 2003/2004.

It is interesting that both Groups, quite independently, took a similar course of action. Before they actually started to evaluate the available codecs, they both developed new criteria and methodologies which they considered suitable for assessing "intermediate-quality" multimedia codecs. For multimedia *audio* subjective evaluations, Project Group B/AIM developed the **MUSHRA** methodology [3] which is now used worldwide. Similarly, B/VIM developed the **SAMVIQ** methodology [4][5] to perform *video* subjective evaluations.

Whereas evaluating performance is a useful measure of a codec's figure-of-merit, the quality performance itself is by no means the sole criterion for codec selection. Cost/complexity and processing delay are the other two, slightly contradictory, requirements. It is virtually impossible to satisfy all three criteria at the same time. Different applications may require a different emphasis.

For example, broadcasting applications may require not only real-time decoding but also real-time encoding. If real-time processing is not critical, it may be possible for video to pass through the encoder more than once (i.e. "two-pass encoding") which increases the encoding efficiency significantly. Another example is video conferencing which requires very low delay. This requirement is contradictory to using frame stores (e.g. "B-frames") which may improve coding efficiency but introduce unwanted delay. Other requirements may involve error robustness, scalability and random-access editing.

This article focuses on the *performance* aspects of internet video codecs and neglects evaluation of other codec characteristics.

General considerations

It is evident that a magic formula allowing us to determine the coding performance does not exist. The performance of a codec needs to be evaluated by physically measuring it using a suitable methodology.

Some typical questions relating to codec performance and quality are as follows:

- How does video codec A compare with video codec B (e.g. how does Windows Media compare with RealVideo)?
- What is the quality improvement of generation Y in comparison to generation X for the same codec standard (say, Windows Media)?
- What is the difference in codec performance between implementation C and implementation D for the same codec standard (e.g. MPEG-4)?
- For a given bitrate available in the transmission path, which codec gives the best quality?
- Conversely, given a required level of quality (e.g. "transparent" [1] quality), what is the required bitrate for different codecs?
- Which codec performs best for "my" content (e.g. sport events, films, news bulletins, etc)
- Which codecs gives the most consistent quality across a range of different content?

We will attempt to provide some informed answers to the above questions in the summary section of this article.

Broadly speaking, evaluation methodologies are either *objective* or *subjective*. The former use mathematical models to mimic the behaviour of human visual systems, or can be based on feature extraction from a bitstream. The latter use a group of subjects (“assessors”) who are presented with decoded video pictures and have to judge the perceived quality using a tailored evaluation methodology.

Objective tests can be used for quick and cost-efficient assessment of media quality. They are particularly useful for assessing the progress in codec design for a particular algorithm. However they have several drawbacks. They are strongly dependent on the type of codec used and the parameters chosen: they have very limited correlation with the subjective test results, especially at lower bitrates where distortions are high.

Subjective methodologies are more time consuming, require more effort and are more costly than objective methods. But subjective methods generally give reliable and accurate results if correctly applied. Objective methods are not capable of providing the full truth about codec quality. If a conclusive decision is needed, we can only rely on “real eyeballs”.

The process of evaluating codec performance subjectively can be summarized in the following simple steps:

[1] Transparent quality is the quality of the codec which is indistinguishable from the uncompressed source quality.

1) Select the codecs under test and define their parameters (pre-filtering, buffering, key frame distance, etc);
2) Select the test sequences (uncompressed);
3) Define the coding conditions: bitrates [2], format resolutions;
4) Compress the test sequences to produce coded representations of the test sequences;
5) Select the evaluation methodology and establish a reproducible test environment;
6) Organize test sessions, invite the test subjects (assessors), present them with the decoded test sequences and ask them to determine the quality as they perceive it;
7) Collect the evaluation results, perform some statistical analysis on the voting data and remove inconsistent subjects;
8) Publish a test report.

The rationale for SAMVIQ

Table 1
Main differences between digital TV and multimedia domains

	TV domain	Multimedia domain
Types of codecs	MPEG-2	Open standards (e.g. MPEG-4, AVC) Proprietary codecs (e.g. Windows Media [a], Real Video, QuickTime)
Image format	Fixed (720 x 576 pixels for active area)	CIF, QCIF, SubQCIF, SIF, VGA, SVGA, etc.
Rate of image refresh (frame rate)	Fixed (25 Hz)	May vary from 0 to 30 Hz
Decoder type	Standardized	Various
Display type	TV	PC, PDA, mobile phone

a. Windows Media is in the process of being standardized within the SMPTE and may become an open international standard known as VC-1.

The television and multimedia domains differ significantly in a number of aspects, which may justify using different assessment methodologies [5]. Some of these differences are summarized in *Table 1*.

The table shows that the multimedia domain offers a large choice of parameters and uses a variety of proprietary and standardized decoders and players, as opposed to television where the system parameters do not vary so much. Whereas in television the rate of temporal refresh (update) of the image data is fixed, it can vary significantly in multimedia video sequences. In addition, in the case of television, the assessor's judgment of images is based on one axis of perception only: spatial sharpness of the images (i.e. quantization). In multimedia, however, the rate of updated video data is not constant. Thus, the assessors combine perception of the sharpness axis with perception of the fluidity of images. Because of this two-dimensional perception (i.e. fluidity and sharpness), it is far more difficult to evaluate the quality of multimedia images than those in the TV domain. Experience shows that this combination strongly influences the final subjectively-perceived picture quality.

Furthermore, multimedia image formats may vary with the types and characteristics of the transmission network, whereas in television these are fairly constant.

Another important difference between TV and multimedia is the viewing distance. In multimedia the viewing distance may vary significantly. In TV the viewing distance used for subjective evaluation is

[2] Assuming constant bit rate (CBR) codecs in which the encoder produces a bitstream at its output which has a constant bitrate (using a suitable buffer size in order to average out the bitrate fluctuations).

well defined and can be between 4 H and 6 H (where H is the display height). In multimedia the viewing distance depends on both the image size (a combination of image and display formats) and the *punctum proximum* [3] which varies from one user to another. Consequently, almost any viewing distance can be considered for evaluations.

In multimedia there is also a practical difficulty: multimedia images cannot be recorded directly on tape – they are only available as data files. As direct stream does not allow for repeatable playout of the sequences, it is not possible to use traditional ITU-R Recommendation BT.500 (e.g. DSCQS) for the multimedia tests. Because of all the reasons given above, a new evaluation approach – specifically designed to assess multimedia – was patently necessary.

Conventional video evaluation approaches

ITU-R Recommendation BT.500 [6] is a reference document for conventional ***television-centred*** subjective evaluations. It proposes several subjective test methodologies. The most important subjective methods used for television assessments are as follows:

- **DSIS**: Double Stimulus Impairment Scale;
- **DSCQS**: Double Stimulus Continuous Quality Scale;
- **SSCQE**: Single Stimulus Continuous Quality Evaluation;
- **SDSCE**: Simultaneous Double Stimulus for Continuous Evaluation.

BT.500 contains the test methodologies used for both quality assessments and impairment assessments [4]. The most commonly used is the DSCQS method in which an assessor is presented with a pair of images or short video sequences A and B, one after the other, and is asked to give A and B a quality score by marking on a continuous line with five intervals ranging from Bad to Excellent. For each pair of sequences, one is an unimpaired reference sequence, and the other is the same sequence, modified by the coding system under test. The order of the two sequences is randomised, so that the assessor does not know which is the original and which is the impaired sequence. The result of the evaluations is a "Mean Opinion Score", which indicates the *relative quality* of the impaired and reference sequences.

Table 2 gives some details of the BT.500 methodologies, in comparison with the SAMVIQ methodology, which is now described in some detail.

The SAMVIQ methodology

SAMVIQ has specifically been designed for ***multimedia*** content. It takes into account a range of codec types, image formats, bitrates, temporal resolutions, zooming effects, packet losses, etc. The SAMVIQ methodology was submitted to ITU-R 6Q in 2003 and has achieved the status of a Draft New Recommendation [7]. This section gives a broad outline of SAMVIQ; a detailed description is given in *Appendix A*.

Compared to BT. 500, a major difference is in the way video sequences are presented to the assessor. In SAMVIQ video sequences are shown in multi-stimulus form, so that the user can choose the order of tests and correct their votes, as appropriate. As the assessors can directly

[3] *Punctum proximum* is defined as the nearest viewing distance, subjectively determined by the viewer's eyes, for optimum accommodation to a given display.

[4] *Quality assessments* are defined as those that establish the performance of systems under optimum conditions. *Impairment assessments* are used to study the systems subjected to non-optimum conditions such as error-prone transmission and emission.

Table 2
ITU-R BT.500 and SAMVIQ

Parameter	DSIS	DSCQS	SSCQE	SDSCE	SAMVIQ
Explicit reference	Yes	No	No	Yes	Yes
Hidden reference	No	Yes	No	No	Yes
High anchor	No	Yes	No	No	Hidden reference
Low anchor	No	Yes	No	No	Yes
Scale	Bad to excellent	Bad to excellent	Bad to excellent	Bad to excellent	Bad to excellent
Sequence length	10s	10s	5 min	10s	10s
Picture format	All	All	All	All	All
Two simultaneous stimuli	No	No	No	Yes	No
Presentation of test material	I: Once II: Twice in succession	Twice in succession	Once	Once	Several concurrent (multi-stimuli)
Voting	Only test sequence	Test sequence and reference	Test sequences	Difference between the test sequence and the reference simultaneously shown	Test sequences and reference
Possibility to change the vote before proceeding	No	No	No	No	Yes
Continuous quality evaluation	No	No	Yes (moving slider in a continuous way)	Yes (moving slider in a continuous way)	No
Minimum accepted votes	15	15	15	15	15
Assessors per display	One or more	One or more	One or more	One or more	One
Display	Mainly TV	Mainly TV	Mainly TV	Mainly TV	Mainly PC [a]

[a] B/VIM is in the process of studying whether SAMVIQ can also be applied to standard television displays, rather than solely PC displays.

compare the impaired sequences among themselves and against the reference, they can grade them accordingly.

SAMVIQ is based on random playout of the test files. The individual assessor can start and stop the evaluation process as he wishes and is allowed to determine his own pace for performing the grading, modifying grades, repeating playout when needed, etc. With the SAMVIQ method, quality evaluation is carried out scene after scene including an explicit reference, a hidden reference and various algorithms (codecs). There is no continuous sequential presentation of the sequences as in the DSCQS method, where the assessor can make errors of judgement due to a lack of concentration. As a result, SAMVIQ offers higher reliability, i.e. smaller standard deviations.

In SAMVIQ there is only one assessor at a time, which alleviates a "group effect".

Both an explicit and a hidden reference are used. The explicit reference is an uncompressed version of the original sequence and allows the assessor to determine a near-absolute measure of video quality [5]. A hidden reference is technically identical to the explicit reference but is not readily available to the subject. It is actually hidden among other stimuli and the subject should be able to identify it. In SAMVIQ the hidden reference is mandatory.

The SAMVIQ method provides an overall quality score for relatively short multimedia sequences. The duration of a sequence is typically in the range of 10 to 15 seconds in order to give the subject sufficient time to formulate a stable grading. The content of a sequence has to be homogeneous.

A large quality range is required to stabilize the assessor's quality scores; otherwise, when the quality range is reduced, assessors try to discriminate among the quality of the sequences even if the differences are not perceptible. Therefore, the reliability of results decreases, as the quality of the codecs tested is similar.

SAMVIQ includes improved rejection criteria (compared with those used in BT.500). The multimedia image quality is to be assessed on a multimedia screen and platforms, and not on conventional TV displays, in order to avoid the artefacts due to interlace and flicker.

Similar to all other subjective methodologies, SAMVIQ requires careful consideration of the test arrangements. If these arrangements are not scrupulously adhered to, the results of the evaluations may not be as expected.

Subjective evaluations of internet video codecs – Phase 2

Phase 2 evaluations were performed by Project Group B/VIM during 2002 and 2003. The following four codecs were included:

- Windows Media 9 – Microsoft;
- RealVideo 9 – RealNetworks;
- MPEG-4 – Envivio implementation;
- QuickTime 6 – Apple.

As in Phase 1, the following bitrates were used for the QCIF and CIF resolution formats:

QCIF format	56 kbit/s	128 kbit/s	256 kbit/s	500 kbit/s
CIF format	256 kbit/s	500 kbit/s	700 kbit/s	1400 kbit/s

The test sequences used were taken from the Phase 1 viewing tests and are listed in *Table 3* and shown in *Fig. 1*. The sequences represent typical broadcast programmes and are fairly critical but not unduly so. Some of the sequences, however, contain difficult scenes (fast movements, details, colours) that may challenge the performance of the codecs under evaluation. The duration of the sequences was typically set to 10s.

Two organizations performed the subjective tests: NRK (the Norwegian public broadcaster) and France Telecom R&D (FTRD). Each site organized a test panel consisting of 15 to 20 subjects. Generally, half of the subjects were experienced while the other half were not regularly involved in this kind of test but showed some interest in video evaluations and were subjected to some initial training.

[5] By comparison, DSCQS and DSIS are capable of establishing a quality level *relative* to a reference sequence.

For the viewing and lighting conditions, the tests used the data given in ITU Recommendation BT-500.11.

Each of these codecs was tested according to the parameters listed in *Table 4*.

Summary of the Phase 2 codec evaluations

Detailed results of the B/VIM video evaluation tests are given in *Appendix B*. In the following, some analysis of these results is performed by responding to the questions raised in the introductory section.

How does video codec A compare to video codec B?

In the Phase 2 evaluations, we compared four codecs: Real Networks 9, Windows Media 9, Envivio MPEG-4 and QuickTime 6. For the CIF image format, RealNetworks 9 is the only codec that reaches transparency level at 1.4 Mbit/s. Windows Media 9 is next best

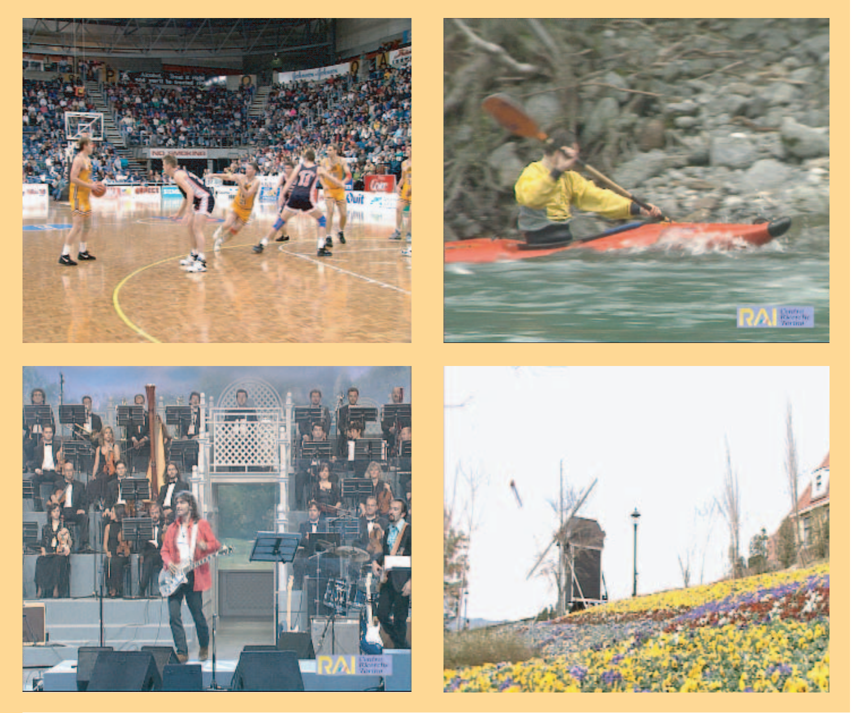


Figure 1
Test sequences used

Table 3
Video test sequences used in Phase 2

1	**Basket**	MPEG	Sport footage with vigorous movements and extensive details
2	**Kayak**	RAI	Sport footage with background panning motion
3	**Entertainment**	RAI	Concert footage with camera motion and details
4	**Flower Garden**	MPEG	Detail and colour rendition

Table 4
Parameters of the codecs under test

#	Channel type	Nom. bitrate (kbit/s)	Net bitrate (kbit/s)	Audio (kbit/s)	Video (kbit/s)	Frame rates and Formats	
						QCIF (176 x 144)	CIF (352 x 288)
1	Modem/PSTN	56	40±10%	8 mono	32±10%	6.25	
2	Dual ISDN	128	100±10%	20 mono	80±10%	12.5	6.25
3	DSL/Cable 1	256	200±10%	32 st music	168±10%	25	12.5
4	DSL/Cable 2	500	400±10%	48 st music	352±10%	25	25
5	DSL/Cable 3	700	560±10%	64 st music	500±10%		25
6	Cable 1	1400	1160±10%	128 stereo	1032±10%		25
8	Reference					AVI RGB 24 @ 18 Mbit/s	AVI RGB 24 @ 68 Mbit/s

but falls 10 points short. It is interesting to see that the quality range of the four codecs at 250 kbit/s is about 10 points but this range increases steadily to some 35 points at 1.4 Mbit/s (which is equivalent to the difference between RealNetworks 9 and QuickTime 6 at that bitrate). This conclusion applies to both CIF and QCIF.

What is the quality improvement between one generation and the next of the same codec standard?

The new generation of codecs (e.g. WM9) is not always better than the previous generation ones (WM8). Sometimes, the reverse is true: WM8 performs better at 500 and 700 kbit/s than WM9. Similarly, RealNetworks 9 is slightly worse than RealNetworks 8 at 250 kbit/s but similar at higher bitrates. The QuickTime codec has the same difficulty: the grades of the new version of QT6 are consistently lower at all bitrates than those of the older version (at 1.4 Mbit/s this difference amounts to 15 points).

Example 1
Windows Media 9 vs. Windows Media 8, CIF, the same laboratory (FTRD)

	250 kbit/s	500 kbit/s	700 kbit/s	1400 kbit/s
WM 8	52	55	62	73
WM 9	50	43	50	70

What is the difference in codec performance between different implementations of the same codec standard?

Some evidence is available from the Phase 1 evaluations: both QuickTime 6 and Dicas Mpegable implemented the MPEG-4 Part 2 video standard. Dicas is better for both CIF and QCIF for all bitrates but the difference is relatively small (about 5 points). In addition, Dicas seems to render “Flower Garden” (which is critical for colour rendition) significantly better than QT6.

In Phase 2 the Envivio codec and the QT6 codec, both based on the MPEG-4 algorithm, can be compared. Our results show that Envivio performs better for both CIF and QCIF at all bitrates. The difference increases with bitrate and reaches 10 to 15 points at 1.4 Mbit/s.

For a given transmission bitrate, which codec gives the best quality?

RealNetworks 9 is undoubtedly the winner. It gives the best quality at all bitrates considered for both CIF and QCIF. It also performs very well for all content types, in particular “Flower Garden”. Windows Media 9 comes second with scores of some 10 points lower at the higher bitrates.

What is the bitrate at which the required level of quality is achieved?

Example 2
Bitrates neede to achieve a score of 50 points (“Fair”)

	Real 9	WM 9	Envivio	QuickTime
CIF	550	650	1100	>1400
QCIF	180	300	370	>1400

Assuming that 50 points ("fair") is an acceptable quality for an application, the bitrates required for different codecs to achieve the 50-point quality mark is shown in the table to the left.

Transparent quality (80 points) was achieved only by the RealNetworks 9 codec at 1.4 Mbit/s for CIF and 500 kbit/s for QCIF.

Which codec performs best for a given content?

- **Basketball:** RealNetworks and Windows Media are very close – the former is slightly better at higher bitrates, the latter is slightly better at lower bitrates.
- **Entertainment:** For CIF, RealNetworks is the winner but Windows Media follows closely behind.
- **Flower Garden:** Undeniably, Real Networks is the best candidate. It achieves transparency at 1.4 Mbit/s for CIF. At 700 kbit/s, RealNetworks exceeds QuickTime by more than 40 points (two categories).
- **Kayak:** RealNetworks and Windows Media are equivalent at lower bitrates but, at higher bitrates, the RealNetworks codec is better by about 10 points.

Which codec gives the most consistent quality across a range of different content?

It is interesting to observe how the quality varies with scene content. The table on the right shows the France Telecom R&D results for the range of variation or "spread" across different scenes for the codecs under test.

Both the NRK and France Telecom tests confirm that Windows Media codec has the least dependency (most consistent quality) across the different scene contents.

Quality variation (in points) across different scenes for the four codecs evaluated: CIF format

	Quality variation (points)	
Codec	**500 kbit/s**	**1400 kbit/s**
Envivio	10	10
QuickTime	14	9
RealNetworks	23	6
Windows Media	4	7

Summary of the inter-laboratory tests

Concerning the CIF format, the NRK and FTRD results are well correlated in terms of the mean scores. The confidence intervals are also of the same size in both cases. It should be pointed out however that the NRK results are generally slightly higher than those of France Telecom. The difference may amount to 10 points (in the case of Windows Media and CIF).

For the QCIF format, the results between labs are again very similar for both RealNetworks and Windows Media. Also, the confidence intervals are almost the same.

A general conclusion can be drawn from these tests (Phase 2) – the evaluation results from both the laboratories are quantitatively the same, regardless of the codec tested, the image format and the bitrate.

This conclusion confirms that SAMVIQ gives consistent results from one laboratory to another.

Conclusions

The first conclusion from the Phase 2 subjective evaluations of four internet video codecs is that the SAMVIQ methodology provides satisfactory reproducibility and repeatability of results. This means that the results will be coherent and indeed consistent from one laboratory to another. This new method is able to discriminate efficiently between the different quality levels in low, intermediate or high quality ranges. The method can combine quality evaluation capabilities with the ability to discriminate between similar levels of quality, using an implicit comparison process.

SAMVIQ is simpler, faster and more user-friendly than traditional subjective evaluation methods.

Concerning the quality performance of the video codecs tested, we believe that our findings may provide useful guidance to EBU Members trying to make a commercial decision about which codec to purchase or use. It should be pointed out, however, that the choice of the most suitable codec is often not a simple matter: it will not only depend solely on the codec quality performance. Often, non-technical considerations prevail, such as cost, rights management and security.

Project Group B/VIM has now embarked on the third phase of its studies and plans to evaluate H.264/AVC and other more recent developments. In addition, the Group plans to study the subjective quality of video streams subjected to some IP packet loss and jitter. Another interesting area of activity is to expand the use of SAMVIQ to standard television. To this end, the Group will carry out some studies aimed at assessing whether SAMVIQ can be used also in the traditional television environment.

References

[1] **http://www.ebu.ch/en/technical/index.php**

[2] EBU BPN 055: **Subjective viewing evaluations of some internet video codecs – Phase 1**
Report by EBU Project Group B/VIM (Video In Multimedia), May 2003.

[3] ITU-R Recommendation BS.1534: **MUSHRA: Method for subjective assessment of intermediate audio quality**
See also EBU BPN 049: **The EBU Subjective Listening Tests on Low Bitrate Audio Codecs**
Report by EBU Project Group B/AIM (Audio In Multimedia), September 2002.

[4] EBU BPN 056: **SAMVIQ – Subjective Assessment Methodology for Video Quality**
Report by EBU Project Group B/VIM (Video In Multimedia), May 2003.

[5] **SAMVIQ – A New EBU Methodology for Video Quality Evaluations in Multimedia**
F. Kozamernik, V. Steinman, P. Sunna, E. Wyckens, IBC 2004, Amsterdam, pp. 191 - 202.

[6] ITU-R Recommendation BT.500–11: **Methodology for the Subjective Assessment of the Quality of Television Pictures**
Question ITU-R 211/11, Geneva, 2004.

[7] ITU-R Document 6Q/57-E: **Draft New Recommendation for Subjective Assessment of Streaming Multimedia Images by Non-expert Viewers**
Source: EBU, Geneva, 27 April 2004.

[8] EBU B/VIM 043: **Rejection criteria of assessors in the context of image quality assessments using a continuous quality scale**
Contribution from France Telecom R&D. Authors: Emmanuel Wyckens, Jean-Louis Blin, Jean-Charles Gicquel, November 2002.

Acknowledgements

This report would not have been possible without the kind support of those who devoted much of their precious time and effort to the underlying work. The bulk of the work was performed by the following EBU member organizations and their staff:

- France Télécom R&D
- Institut für Rundfunktechnik (IRT)

- Norwegian Broadcasting (NRK)
- Radiotelevisione Italiana (RAI) – CRIT

Many thanks to all of them.

Acronyms used

CIF	Common Image Format (352 pixels/line, 288 lines per picture, 30 pictures per second)
QCIF	Quarter CIF (176 x 144)
SubQCIF	Subquarter CIF (128 x 96)
VGA	Video Graphics Array (640 x 480)
SVGA	Super Video Graphics Array (800 x 600)

Appendix A: The SAMVIQ methodology

Test material

The choice of material is crucial to the success of the tests and is far from being a simple matter. It is recommended that we should use a variety of unprocessed, ordinary broadcast programme sequences, addressing different quality aspects (e.g. codec artefacts, motion portrayal, colour rendition, sharpness, etc. The sequences should be chosen not to stress or indeed break the codecs tested. In selecting a range of test sequences it is important to achieve some balance between being not critical enough and being too critical. In the former case, all codecs would appear to be very good. In the latter case, all codecs would appear to be bad. The length of the sequences should typically not exceed 20s to avoid fatiguing the observers and to reduce the total duration of the tests. The test sequences should normally be different from those used by the manufacturers in optimizing the coding algorithms.

Training phase

The training session is an integral part of the SAMVIQ methodology. It is absolutely essential to train the subjects (assessors) in a special training session in advance of the tests proper. The appropriate training helps to obtain more reliable and more consistent results. The subject should be handed a written instruction sheet. As the same instructions should be used in all laboratories involved in the measuring campaign, there should be no statistically inconsistent results from one laboratory to another.

The purpose of the training phase is to allow the subject to achieve the following objectives:

- to become familiar with the kind of artefacts of the compressed sequences
- to learn how to use the test equipment (user-interface) and the grading scale.

The subjects should be told that the hidden reference is included in the tests but should not necessarily score it “100”. They should use the full range of the continuous scale, as they find it appropriate.

Viewing conditions

The tests follow the conditions given in Rec. BT.500-11 for the ambient light, colour of the walls, type of monitor, etc. It is of paramount importance to create viewing conditions that can be reproduced in other laboratories around the world. Influence of the laboratory setup should be minimized.

Test organization

Test sessions are organized such that one scene follows the other *(see Fig. 2)*. Only one image format (e.g. CIF or QCIF) is considered per session. The number of algorithms is limited to ten per scene, e.g. five algorithms for Codec 1 and five algorithms for Codec 2.

For a scene it is possible to play and grade any sequence in any order. Each sequence can be played and assessed as many times as the assessor wants – the last grade remains recorded. Each algorithm must be played out and viewed completely in each scene at least once. Grading of an algorithm can only be made after at least one complete viewing of that algorithm.

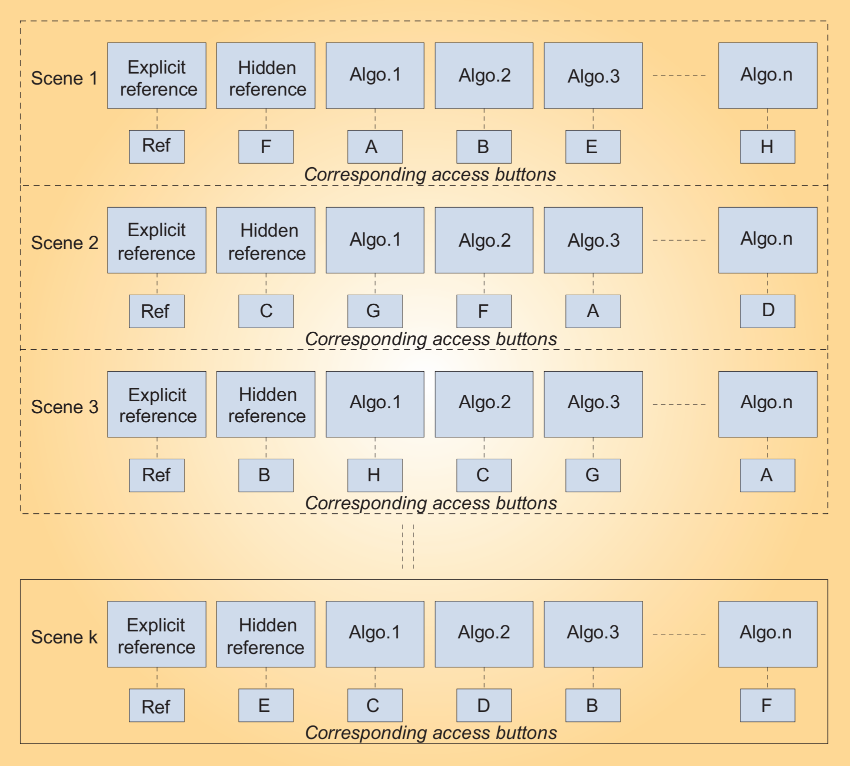


Figure 2
An example of the test organization in SAMVIQ

From one scene to the next, the sequences are randomised. This prevents the assessors from attempting to vote in an identical way according to an established order. Nevertheless, within a test, the algorithm order remains the same to simplify the analysis and presentation of results. Only the corresponding access from an identical button is randomized.

The assessor is allowed to proceed to the next scene only after the evaluation of the previous scene was accomplished successfully. To finish the test, all the sequences of all the scenes must be scored.

Explicit and hidden reference

Many subjective evaluation methods commonly use quality anchors in order to stabilize the results. SAMVIQ uses two high-quality anchors: an explicit reference and a hidden reference. According to our extensive studies, the explicit reference is necessary in order to improve the consistency of the scores, thus minimizing the standard deviation [6]. This explicit reference is coded as an uncompressed, resized, AVI file at 25 frames per second. A hidden reference is coded identically to the explicit reference but it is not readily accessible to the subject. It is actually "hidden" among other stimuli.

The hidden reference is useful as it helps to evaluate the intrinsic quality of the reference, particularly at the lower resolution image formats because, in such a case, the perceived quality of the reference is less than perfect. Our experience shows that about one third of assessors score the explicit and hidden references differently. While they assign the explicit reference the highest possible score (100), they score the hidden reference much lower. We have run some tests without the two references and the standard deviation dramatically increased.

Encoding of multimedia pictures

The encoding process is critical for the proper conducting of SAMVIQ tests. Particular attention should be paid to the length of test sequences. On the one hand, sequences should be sufficiently short, usually between 10s and 15s, so that the quality is relatively uniform throughout the sequence. On the other hand, the scenes should be sufficiently long – typically from 40s to 60s. Such a length is required in order to achieve well stabilized bitrate control towards the end of the sequence. The experience shows that, if very short sequences are used, the results obtained may be too optimistic. Namely, if the ratio between the buffer size and the sequence length is high, the codec can use a higher bitrate than required by the test. This may lead to a higher perceived quality which may give the codec unfair advantage compared to other codecs.

To this end, SAMVIQ requires the construction of longer sequences (e.g. 40 - 60s) which consist of four replications of the basic scene. Encoding/decoding is performed on these longer sequences, but the last portion of the longer sequence is retained for the subjective evaluations.

Assessors

SAMVIQ requires experienced, properly trained assessors (or "subjects" or "evaluators"). They should however not be expert assessors, professionally involved in picture quality assessments. Expert assessors often have preconceived judgements of video artefacts, resulting in somewhat biased scoring. However, assessors must acquire certain experience about the types of impairments and the quality ranges likely to occur. Such experience may be obtained through a prior training session. Assessors need some motivation and patience to observe the test sequences in a critical way. Test sessions should not be too long and the number of test sessions not too large, so that the concentration (focus) of assessors does not suffer. Prior to the test, the assessors should be screened for normal visual acuity on the Snellen or Landolt chart, and for normal colour vision using specially selected charts (Ishihara).

In order to obtain statistically valid results, SAMVIQ requires that the number of subjects involved in the tests should be large enough. As a rule-of-thumb, a panel of at least 15 valid assessors should be available after post-screening.

The SAMVIQ interface

A typical SAMVIQ interface is shown in *Fig. 3* [7]: Seven anonymous algorithms plus an explicit reference for five scenes are to be evaluated and scored. The slider is directly implemented on the

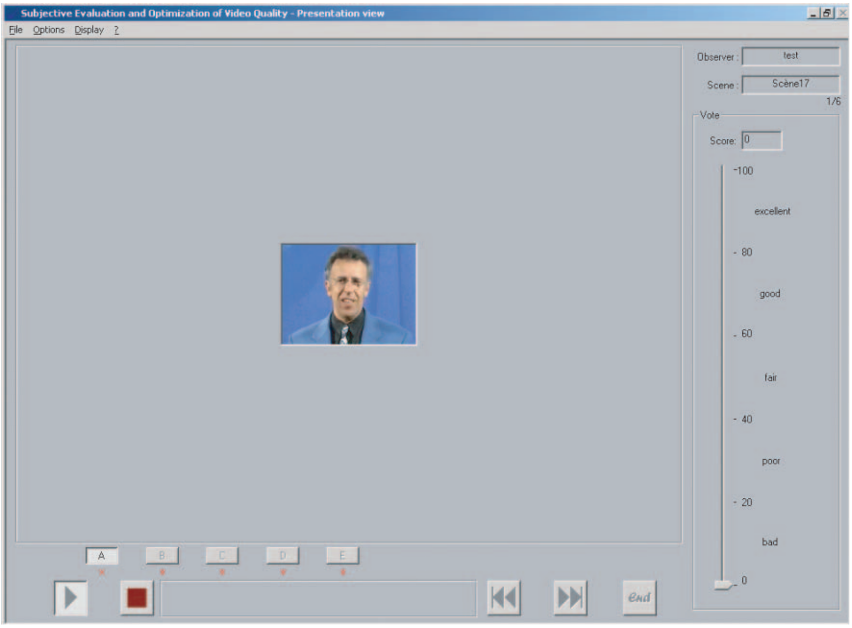
[6] The lower the standard deviation, the higher is the consistency of results.

[7] This example is taken from France Telecom's software package.

screen. In the screenshot, some algorithms have already been evaluated – their scores are written under the corresponding access button – and the explicit reference is currently under evaluation. The following buttons for controlling the playout of the test sequences are shown on this screenshot:

- Selection of algorithm (A, B, C, …, Explicit Reference).
- Play, Stop, Previous scene, Next scene, End.

The "Ref button" (bottom left) represents the Explicit Reference, while the other buttons (A, B, C…) represent the algorithms including the hidden reference and the low anchor. On the right side, a slider is present in order to allow the assessor to grade the quality of the test item according to the continuous quality scale used.



Observer : test

Scene : Scène17

1/6

Score: 0

Figure 3
Interface that implements the SAMVIQ method
(Courtesy: France Telecom R&D)

Grading scale

The assessors are asked to assess the overall picture quality of each presentation by inserting a slider mark on a vertical scale. The scales provide a continuous rating system to avoid quantizing errors, but they are divided into five equal lengths which correspond to the normal ITU-R BT.500 five-point quality scale. The associated terms categorizing the different levels are the same as those normally used; but here they are included for general guidance. The grading scale is continuous and is divided in five equal portions, as follows:

Excellent	(80 to 100 points)
Good	(60 to 80 points)
Fair	(40 to 60 points)
Poor	(20 to 40 points)
Bad	(0 to 20 points)

The lowest quality perceived should be scored "0" (bottom of the scale) and the highest quality should be marked "100" (top of the scale).

Viewing Distance

SAMVIQ does not require any specific viewing distance range. Each assessor adjusts his own optimal viewing distance according to his preference for comfortable viewing. Especially for small images, the viewing distance depends on both the image size (a combination of image and display formats) and the *punctum proximum* (see *Footnote 3.* on page 4) which may vary from one user to another.

Rejection criteria

All the assessors who have taken part in the evaluation process must be screened in order to establish the consistency of their scores. Inconsistent assessors who produced unstable or even contradictory scores are discarded from the final statistics. Compared to BT.500, SAMVIQ developed more accurate and reliable rejection criteria [8]. In SAMVIQ (as in DSCQS), all sequences including the hidden reference, low anchor and encoded sequences are considered.

The rejection criteria use the Pearson decision criterion which is based on a correlation "r" of individual scores and corresponding mean scores from all the assessors. The Pearson algorithm assumes a linear relationship between the quality scale and score range of assessors. If this rela-

Franc Kozamernik graduated from the Faculty of Electrotechnical Engineering, University of Ljubljana, Slovenia, in 1972.

He started his professional career as an R&D engineer at Radio-Television Slovenia. Since 1985, he has been with the EBU Technical Department and has been involved in a variety of engineering activities covering satellite broadcasting, frequency spectrum planning, digital audio broadcasting, audio source coding and the RF aspects of various audio and video broadcasting system developments, such as Digital Video Broadcasting (DVB) and Digital Audio Broadcasting (DAB).

During his years at the EBU, Mr Kozamernik has coordinated the Internet-related technical studies carried out by B/BMW (Broadcast of Multimedia on the Web) and contributed technical studies to the I/OLS (On-Line Services) Group. Currently, he is the coordinator of several EBU R&D project groups including B/AIM (Audio in Multimedia), B/VIM (Video in Multimedia) and B/SYN (Synergies of Broadcast and Telecom Systems and Services). He also coordinates EBU Focus Groups on Broadband Television (B/BTV) and MultiChannel Audio Transmission (B/MCAT). Franc Kozamernik has represented the EBU in several collaborative projects and international bodies, and has contributed a large number of articles to the technical press and presented several papers at international conferences.

Paola Sunna was born in Italy in 1971. In 1997 she received a Degree in Electronic Engineering from the *Politecnico* of Turin. Her thesis was on objective video quality assessments in the MPEG-2 domain. Since then, Ms. Paola has been working at RAI CRIT (the Centre for Research and Technological Innovation of the Italian broadcaster RAI); her activities are mainly focused on full lab testing of video quality assessment schemes for broadcasting, webcasting and 3G applications. Since 2002, she has been the chair of EBU project group B/VIM (Video in Multimedia) that has defined the new subjective methodology for multimedia quality evaluation – SAMVIQ.

Emmanuel Wyckens was born in 1969 and graduated in Electronic Engineering from Valenciennes University (France) in 1994, having specialized in images,. In 1998, he joined France Telecom R&D and became involved in the development of digital processing algorithms for improving the video quality in MPEG-2 codecs.

Mr Wyckens' current activities are in the field of subjective video quality in multimedia (streaming and videoconferencing) applications, and in the standard and high-definition television domains. He also participates in the work of EBU project group B/VIM (Video in Multimedia).

Dag Inge Pettersen received a degree in Electronic Engineering from HiBU in Kongsberg, Norway in 1993. In 1998 he received his MSc in signal processing from the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway.

Mr Pettersen joined NRK, the Norwegian public broadcaster, in 1999 and works mainly as a technical advisor in DAB and other digital broadcast technologies. He also works as a software developer in NRK.

tionship is not supposed to be linear, the Spearman rank correlation may be applied. In practice the Pearson and Spearman correlation results are very close indeed. By taking into account the Spearman rank and Pearson correlation results, an assessor may be discarded if "r" is less than the correlation threshold, which is normally set to 0.85.

Presentation of the results

The results of assessments should be presented in a standardized SAMVIQ form, so that they can be compared among the different laboratories. The EBU plans to establish a standard evaluation protocol which will include the following information:

- Test configuration;
- Test sequences;
- Type of picture source and display computer monitor (screen size, make and model number of displays used);
- Number and type of assessors (age and gender composition of the panel, education or employment category of the panel);
- Reference systems used;
- The grand mean score for the experiment;
- Original and adjusted mean scores and 95% confidence interval if one or more assessors have been eliminated.

Appendix B:
Selected test results

On the following pages are some example results from the Phase 2 subjective evaluations on internet video codecs.

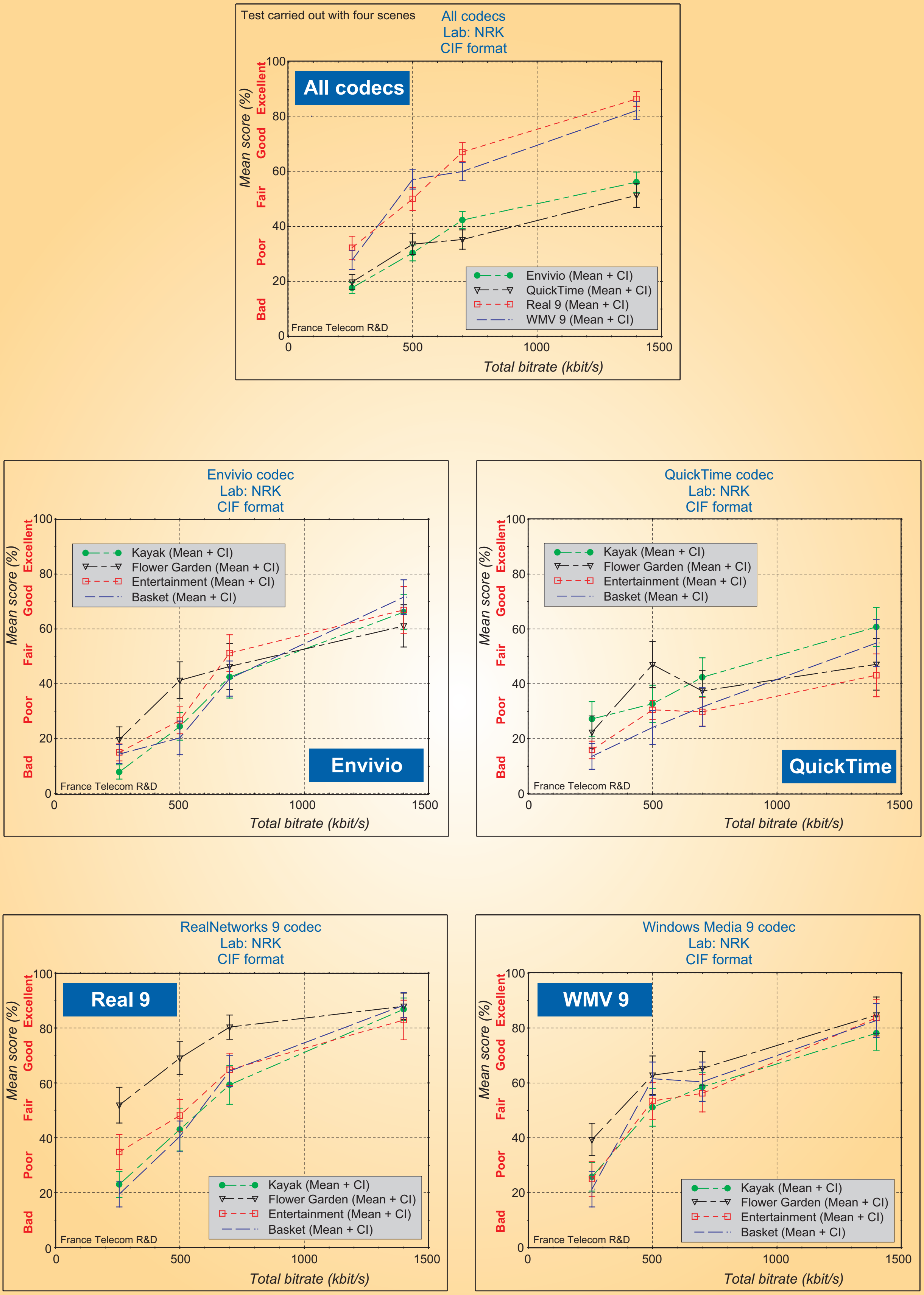


Figure A1
NRK results – CIF format
All codecs, Envivio MPEG-4, QuickTime 6, RealNetwoks 9 and Windows Media 9

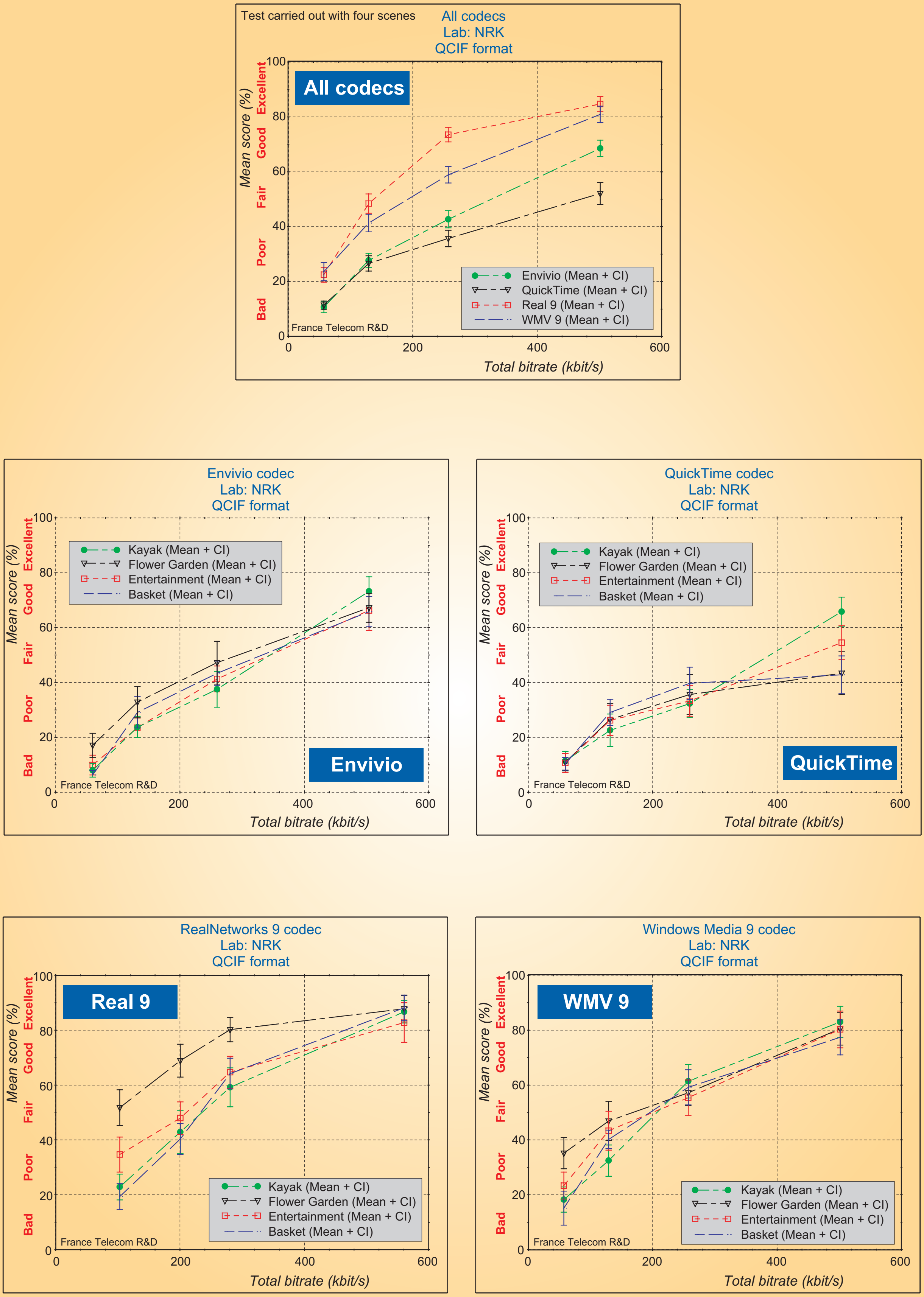


Figure A2
NRK results – QCIF format
All codecs, Envivio MPEG-4, QuickTime 6, RealNetwoks 9 and Windows Media 9

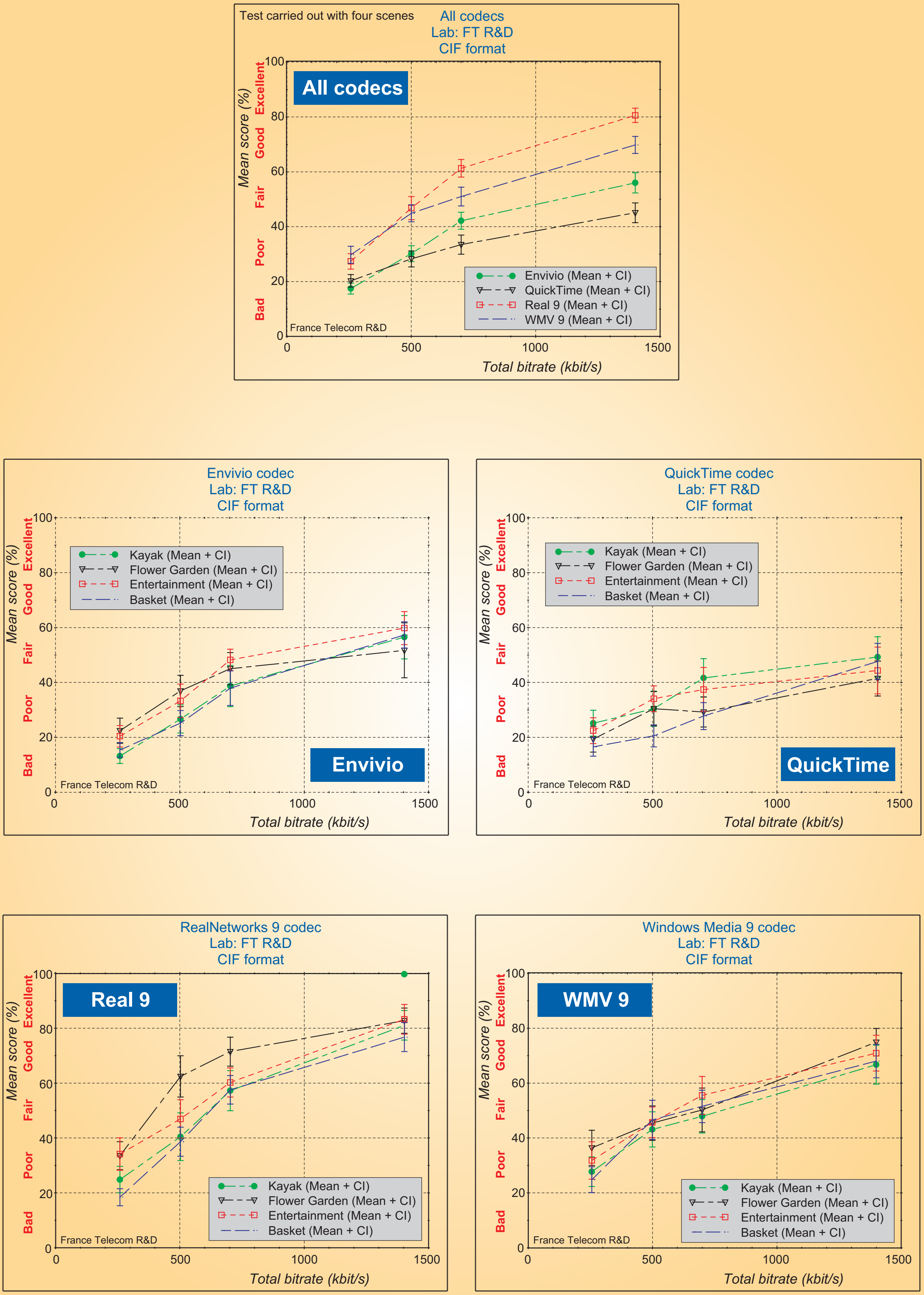


Figure A3
FTRD results – CIF format
All codecs, Envivio MPEG-4, QuickTime 6, RealNetwoks 9 and Windows Media 9

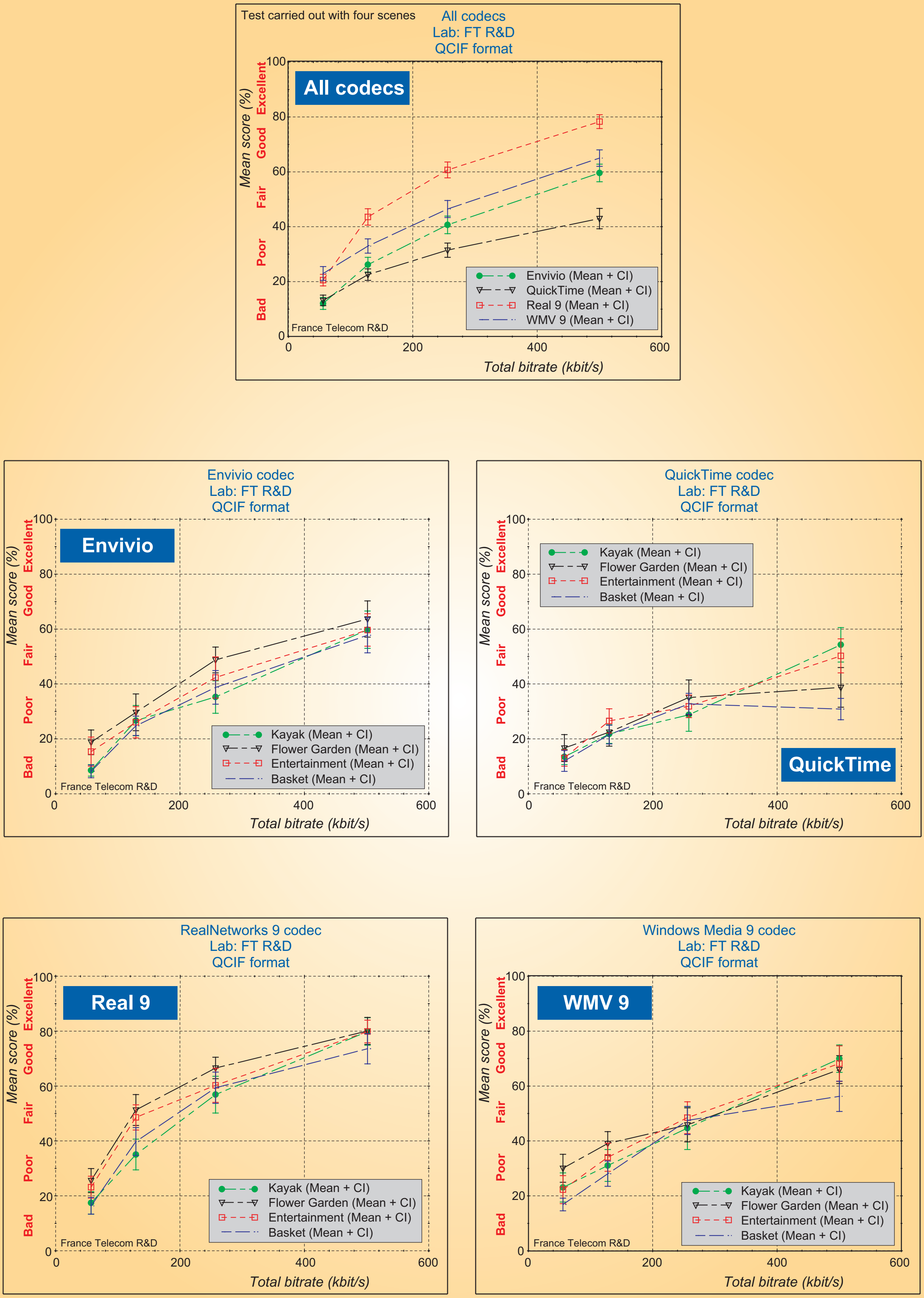


Figure A4
FTRD results – QCIF format
All codecs, Envivio MPEG-4, QuickTime 6, RealNetwoks 9 and Windows Media 9

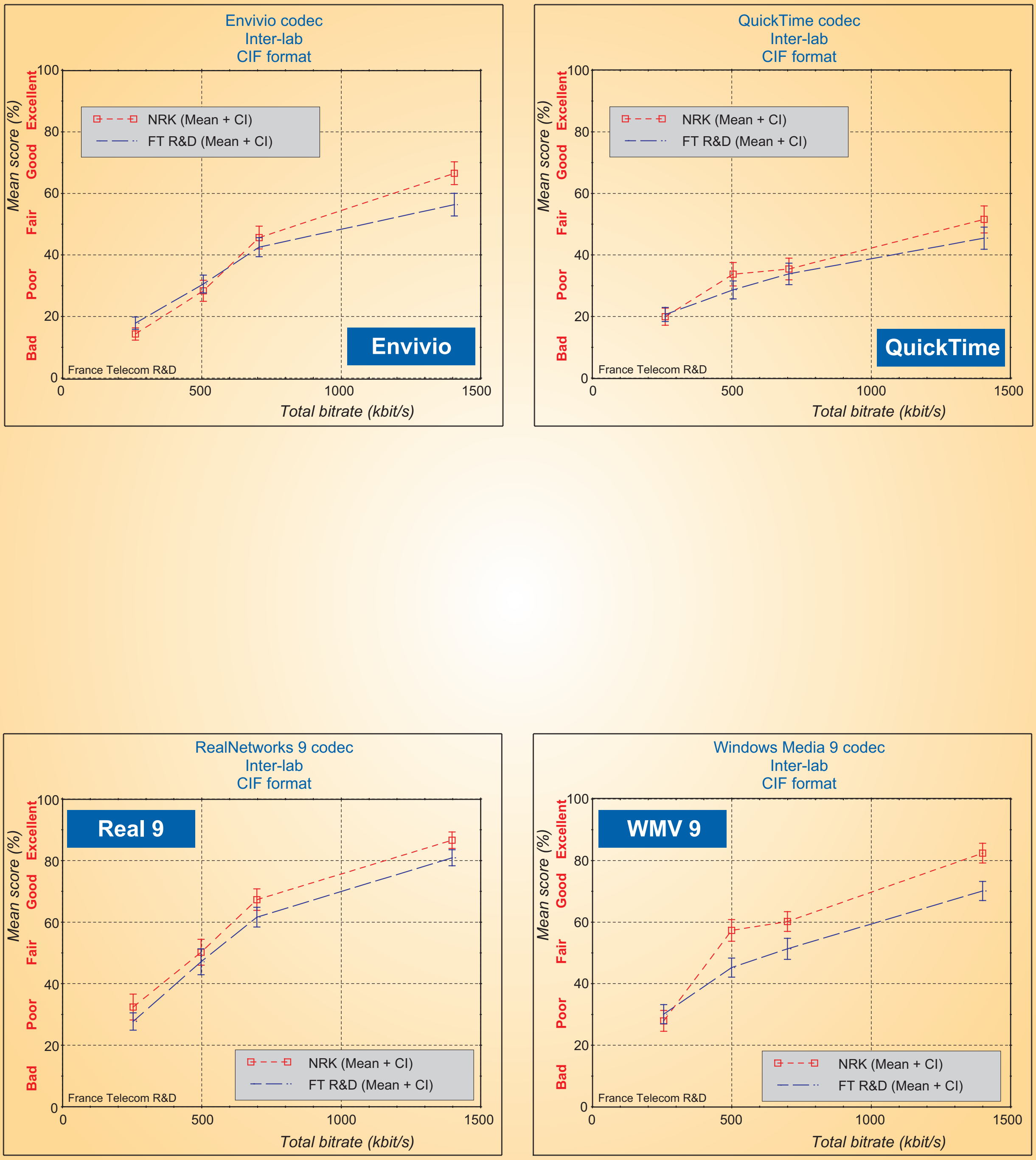


Figure A5
Inter-laboratory evaluations – CIF format
Envivio MPEG-4, QuickTime 6, RealNetwoks 9 and Windows Media 9

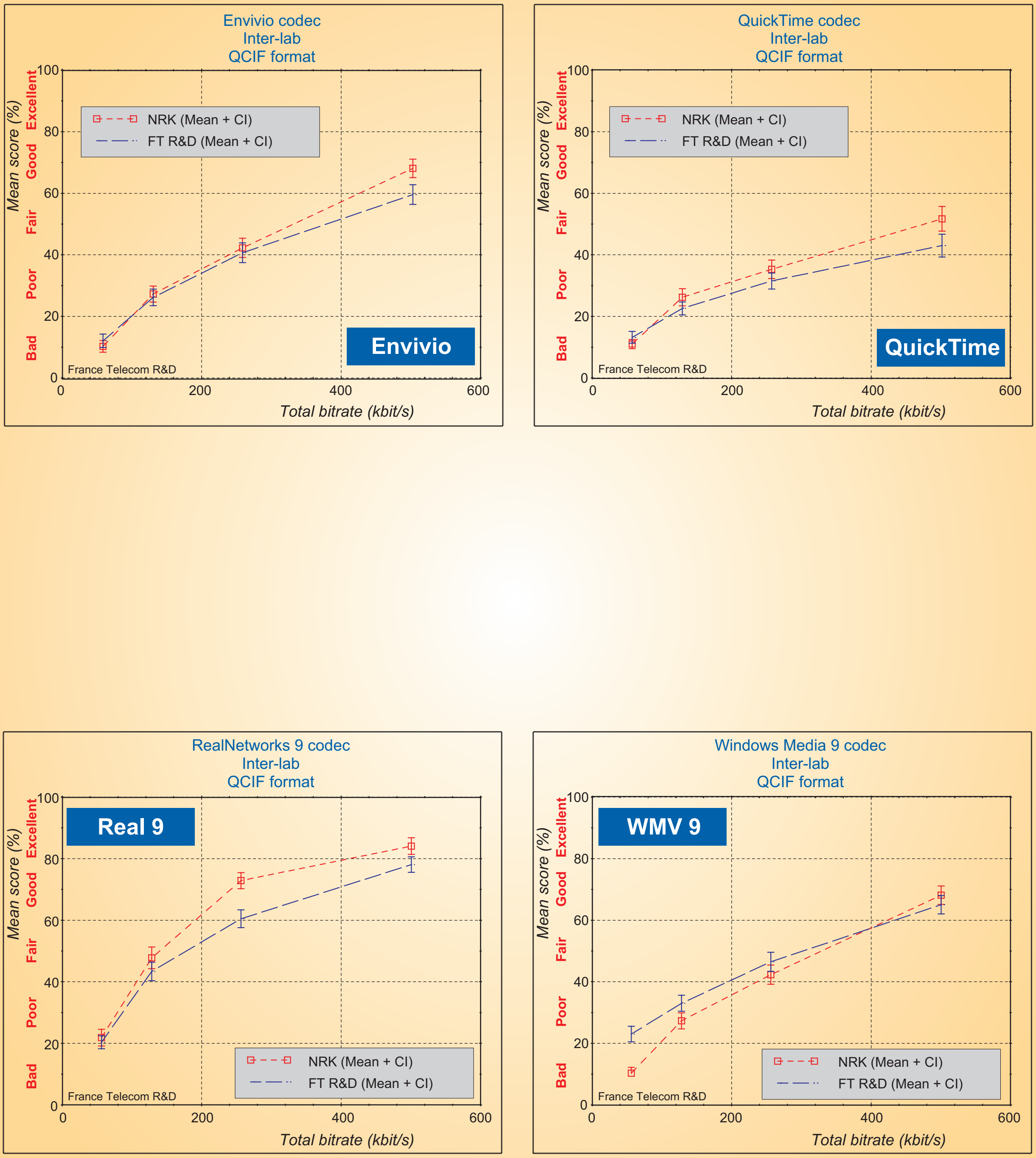


Figure A6
Inter-laboratory evaluations – QCIF format
Envivio MPEG-4, QuickTime 6, RealNetwoks 9 and Windows Media 9