Subjective quality evaluation – The SSCQE and DSCQE methodologies

Th. Alpert (CCETT)
J.-P. Evain (EBU)

The EC project, RACE MOSAIC, was set up to find ways of overcoming specific digital picture quality issues (e.g. content-dependent encoding performance, codec cascading and dynamic statistical multiplexing). In the framework of this project, a new methodology has been designed to allow subjective assessment of both picture and service quality, in conditions that are closer to the actual home environment.

This Article describes the new method – known as Single-Stimulus Continuous Quality Evaluation and, more particularly, "SSCQE Stage 1" which was recently introduced in ITU-R Recommendation BT.500-7.

The double-stimulus DSCQE methodology – recently studied in the EC project, ACTS TAPESTRIES – is an adaptation of SSCQE. DSCQE has been proposed to the MPEG-4 group to address the specific issue of error-robustness evaluation, and is briefly described here.

Original language: English
Manuscript received 6/2/97.

This article is adapted from a paper presented by the Authors at the EBU/IAB Seminar "Made to Measure", November 1996.

1. Introduction

The number of digital services is now growing rapidly, giving rise to a number of questions such as:

– at which bit-rate should I deliver my services?

– how can I best use the transmission resources available?

– does my system functionality fulfil the application requirements?

These questions cover two complementary aspects: the picture quality and the overall service quality. The notion of a *service* is particularly important in a digital environment, as the programme "bouquets" (i.e. multiplexes) can be organized in a flexible manner (number of digital programmes per bouquet, their individual quality, functionality, etc.). The overall service quality can thus be seen as the optimal balance between the operational constraints of the service provider and the expectations of the targeted audience.

In the process of continuously improving the subjective assessment methodology, and adapting it to the most recent technological developments (i.e. to the digital multimedia world), a Consor-

tium of European specialists – under the European Commission (EC) umbrella – launched the RACE MOSAIC project. From the work of this project, the **S**ingle-**S**timulus **C**ontinuous **Q**uality **E**valuation (SSCQE) method was developed [1][2]. Even if designed to answer emerging needs in a digital environment, SSCQE can perfectly be used to evaluate analogue systems alone, or for comparison with digital systems.

MOSAIC did not directly address the objective measurement issue. Nevertheless, the SSCQE methodology was rapidly seen to work in conditions close to those of objective evaluation. The SSCQE continuous acquisition of votes (two votes-per-second) looks very similar to the acquisition of specific data from a real-time system, both being closely related to time and picture content. Studies are currently being undertaken to derive global ratings from SSCQE assessments, and the definition of a link between subjective and objective evealuation of picture quality may be proposed in the not-too-distant future.

A new file format for data interchange was defined by MOSAIC and has been introduced in ITU-R Recommendation BT.500-7 [3]. This format already offers the possibility of storing objective measurement data and subjective assessment data in a compatible way for parallel processing.

With the exception of Swiss Telecom PTT being replaced by the Italian private research laboratory, CSELT, the same Consortium launched the ACTS project, TAPESTRIES. Arising from the work of this group, an adapted version of the SSCQE methodology has been proposed, using simultaneous double visual stimuli. This new method is called **D**ouble-**S**timulus **C**ontinuous **Q**uality **E**valuation (DSCQE) and has been proposed for adoption by MPEG in the framework of the MPEG-4 tests on transmission error-resilience [4][5][6].

Abbreviations

ACTS	Advanced Communications Technologies and Services
CCETT	Centre Commun d'Etudes de Télédiffusion et de Télécommunications (France)
DSCQE	Double-stimulus continuous quality evaluation
DSCQS	Double-stimulus continuous quality scale
DSIS	Double-stimulus impairment scale
ISO	International Standards Organisation
ITU	International Telecommunication Union
MOSAIC	Methods for Optimization and Subjective Assessment in Image Communications
MPEG	(ISO) Moving Picture Experts Group
PS	(SSCQE) programme segment
QP	(SSCQE) quality parameter
QUOVADIS	QUality Of Video and Audio for DIgital television Services
RACE	R&D in Advanced Communications technologies in Europe
SSCQE	Single-stimulus continuous quality evaluation
TAPESTRIES	The Application of Psychological Evaluation to Systems and Technologies in Remote Imaging and Entertainment Services
TC	(SSCQE) test condition
TP	(SSCQE) test presentation
TS	(SSCQE) test session
VS	(SSCQE) vote segment

2. The SSCQE concept

SSCQE was originally designed to perform time-efficient subjective quality evaluations of digital services, in conditions near to the home environment. It also overcomes most of the difficulties encountered when using conventional double-stimulus methodologies to assess the picture quality of digital systems (see the article starting on *page 21* [7].

Digital processing is characterized by the extensive use of statistical methods to manipulate image contents and to exploit the human psycho-visual characteristics. The use of high levels of compression, to varying limits, results in artefacts which are neither regular nor consistent. The MOSAIC Consortium therefore proposed to use test sequences longer than the 10-second sequences of, for example, the **D**ouble-**S**timulus **C**ontinuous **Q**uality **S**cale (DSCQS) and the **D**ouble **S**timulus **I**mpairment **S**cale (DSIS) methods of ITU-R Recommendation BT.500-7.

The use of longer test sequences raised new issues such as how long each sequence should be, and what the voting procedure should be in relation to the behaviour of the observer. Different studies were undertaken to evaluate the *recency* and *forgiveness effects* of the observer, by inserting artefacts at different positions within sequences of varying lengths, and collecting one quality grad-

ing at the end of each presentation. The results showed that the reporting time and the human memory processes (beyond 10- to 15-second time-slots) play an extremely important role. Different tests were performed to confirm that the observers could assess the picture and service quality accurately over sequences of 30 to 60 minutes.

A continuous quality evaluation mechanism was carefully considered. It was thought that this approach would solve the problem of quasi-random appearances of content-dependent artefacts, bearing in mind the recency and forgiveness effects. The maximum frequency of vote acquisition was determined (two votes-per-second) using the results of preliminary studies on the recovery time. Continuous quality evaluation was also found to be closer to the real home environment where programme *zapping* allows an immediate sanction over quality. The continuous evaluation is performed using a sliding device where the observer moves the knob in one direction to show appreciation of the picture quality and in the other direction to indicate concern about it.

Continuous subjective evaluation looks very similar to the objective measurement approach. Even if data acquisition occurs at different frequencies, parallel processing can be envisaged at precisely-defined and common points in time. For example, the subjective quality appreciation may be correlated with the picture content and other physical parameters (e.g. during real-time codec operation) at each voting instant. Additionally, if SSCQE could soon deliver *average* quality ratings (see *Section 3*), a link could also be established with objective measurement results.

The selection of test material was finally addressed by MOSAIC. The use of longer test sequences is causing the old rule "critical but not unduly so" to become less meaningful. Nevertheless, in the case of picture and service-quality evaluation in conditions near to the home environment, the most appropriate criteria was defined as "sequences representative of the programme targeted" (e.g. Sport and/or News and/or Drama and/or Movies for television services). It was also recommended that the test material should have accompanying sound.

All types of test conditions (different bit-rates, transmission parameters, etc.) can be assessed using the SSCQE method. It is also possible to add references (anchors) as part of these test conditions. This is a way of overcoming the inherent difficulty of obtaining acceptability thresholds from image-quality evaluations.

3. The three stages of SSCQE

SSCQE is foreseen as a three-stage method but only "stage 1" has so far been introduced in ITU-R Recommendation BT.500-7.

Stage 1 consists of performing the single-stimulus continuous quality evaluation, and collecting data on the instantaneous grading from the slider device used by each observer. Self-consistent processing is already possible at this stage, resulting in a cumulative distribution of quality variations with time. Stage 1 is particularly suited to the requirements of comparison tests.

The *Stage 2* option is available to extract 10-second sub-sequences from the original test material to perform complementary DSCQS or DSIS tests. An example might be those sub-sequences which correspond to the different percentiles of the cumulative distribution obtained at stage 1. Stage 2 can also be used to calibrate the stage 1 results, using the existing adjectival scales given in ITU-R Recommendation BT.500-7.

Under *Stage 3*, further developments are currently being considered in TAPESTRIES to apply an

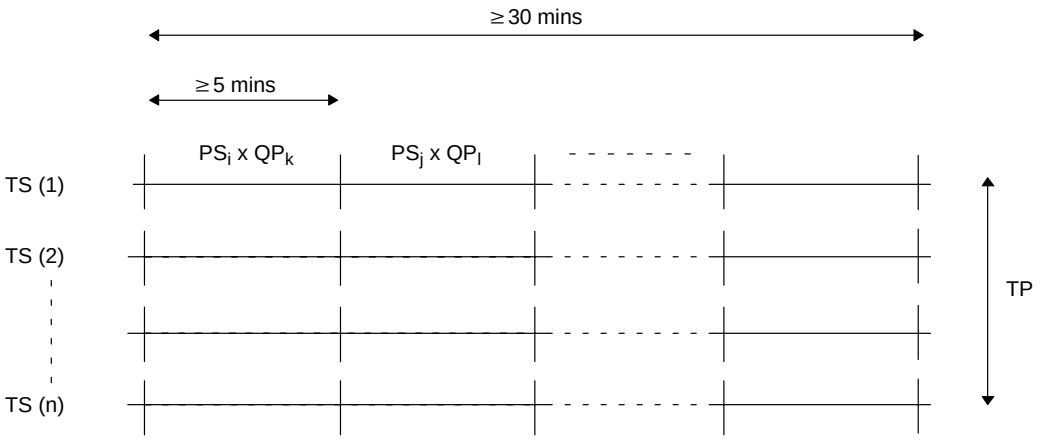


Figure 1
SSCQE – Stage 1 protocol.

overall weighting function (modelling the human memory processes, i.e. the recency and forgiveness effects) in order to arrive at a global *average* for the perceived quality of the sequence being tested.

3.1. Stage 1 test protocol

SSCQE tests can be organized in one or more Test Sessions according to the conditions prescribed in the test protocol. The following definitions and parameters apply *(Fig. 1)*.

A *Programme Segment* (PS) corresponds to one type of video content (e.g. class A, B . . .), processed according to one of the *Quality Parameters* (QPs) – also called *Test Conditions* (TCs) – under evaluation (e.g. the specific picture process or transmission conditions). Each PS must be at least 5 minutes long.

A *Test Session* (TS) is a series of different pairs of PS/QP, without separation, and arranged in a pseudo-random order. Each session contains, at least once, all the Programme Segments and Quality Parameters but not necessarily all the PS/QP combinations. Each session should last at least 30 minutes.

A *Test Presentation* (TP) is a series of TSs that encompasses all the PS/QP pairs.

A *Vote Segment* (VS) is a cluster of votes (e.g. 20 votes for a 10-second VS that is independent of recency and forgiveness effects) on which preprocessing can be made, if required, to smooth out the raw data. Each Programme Segment is therefore made up of a series of Vote Segments.

Each observer is asked to vote continuously during a session using a sliding device with a 10 cm linear range of travel. Vote acquisition is performed automatically, at a rate of two votes per second, with values which lie within the range of the corresponding continuous quality scale.

All the combinations of PS/QP must be assessed by the same number of observers (but not necessarily the same observers).

If audio is introduced, the selection of audio material must be considered as having the same importance as the selection of video material.

In the case of parallel objective evaluation, the approach remains the same but the votes are replaced by data acquired at an appropriate sampling rate by specific devices. Nevertheless, it is still necessary to ensure that a time reference is maintained for correlation between the subjective quality appreciation, the picture content and the corresponding digital processing instants.

At this point, a difference between subjective assessment and objective quality evaluation can be highlighted. Subjective quality assessment allows the impact of audio on video, and video on audio, to be experienced and, hence, assessed. Subjective assessment can also integrate the influence of various environmental and operational factors.

3.2. Data processing for subjective evaluation

When a test has been carried out according to the stage 1 test protocol, one or more data files become available containing all the Test Session votes of the corresponding Test Presentation. A basic test check consists of ensuring that each PS/QP pair has been addressed and that an equivalent number of votes has been allocated to each of them.

A multi-stage analysis is required in each case. Although this is not described in ITU-R Recommendation BT.500-7 as part of the usual test preparation phase, the MOSAIC Consortium has defined this process as follows.

a) In order to achieve maximum flexibility in data processing, it is assumed that each parameter can be selected independently, e.g. suppression of Programme Segments, Quality Parameters or Test Condition, and observers. As data files are closely related to time (two values per second per observer), windowing within the Programme Segments is possible. It is also possible to merge different cases, e.g. in order to calculate global results for all the Programme Segments over one Quality Parameter. Other parameters can be identified and addressed separately through software filters (e.g. laboratories, Test Sessions, viewing distances and vote types).

b) The arithmetic mean and standard deviation is then calculated at each voting instant (every 500 ms) from the votes provided by the individual observers (see *Fig. 2*). Experience has proved it unnecessary to normalize the observer votes before further calculations.

c) Each PS is then considered as a series of 10-second Vote Segments. An arithmetic mean is derived from the 20 preliminary mean values calculated for each Vote Segment. A new standard deviation and/or a confidence interval is also calculated for each Vote Segment (see *Fig. 2*).

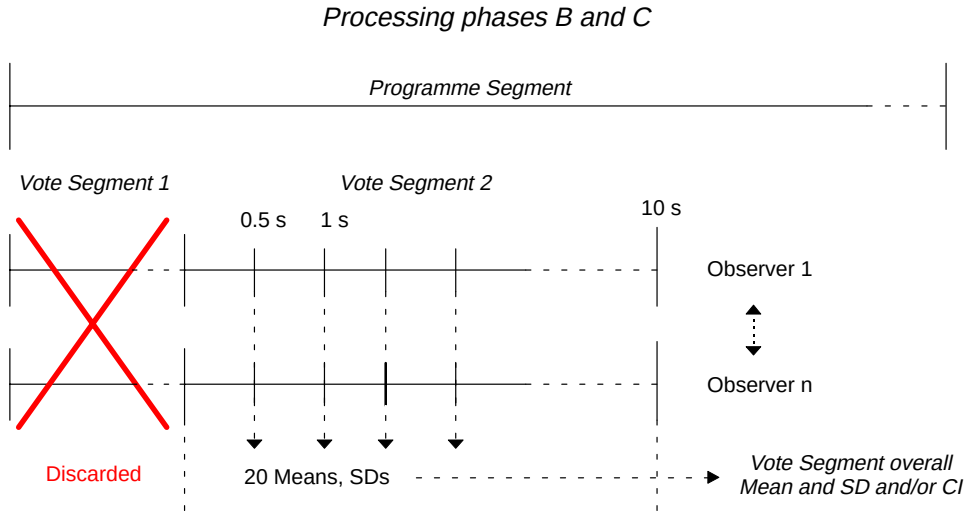


Figure 2
SSCQE – Stage 1 data processing.

d) An analysis of the statistical distribution of the mean values per Vote Segment is performed, in order to find the frequency of occurrence of each average quality grade, and to allocate this value to a particular quality class. (A quality class is defined by splitting the 0 – 100 continuous quality scale equally into, for example, 10 quality classes.) The first Vote Segment is rejected at each PS/QP transition in order to avoid recency effects from the previous PS/QP pair.

e) The global quality distribution can then be calculated (e.g. for each PS or QP, or for an overall estimation which combines all the PSs for a particular QP by accumulation of the frequencies of occurrence). A global quality distribution corresponds to a cumulative statistical distribution function by showing the relationship between the mean values for each Vote Segment (closely related to a Programme Segment) and their cumulative frequency of appearance.

f) The final stage 1 results can then be given in the form of a matrix containing the mean, standard deviation and/or confidence interval of each Vote Segment. It can also be presented more appropriately in the form of a histogram which shows the cumulative distribution of the mean values (the quality distribution). The standard deviation or confidence interval can also be represented on this quality distribution histogram.

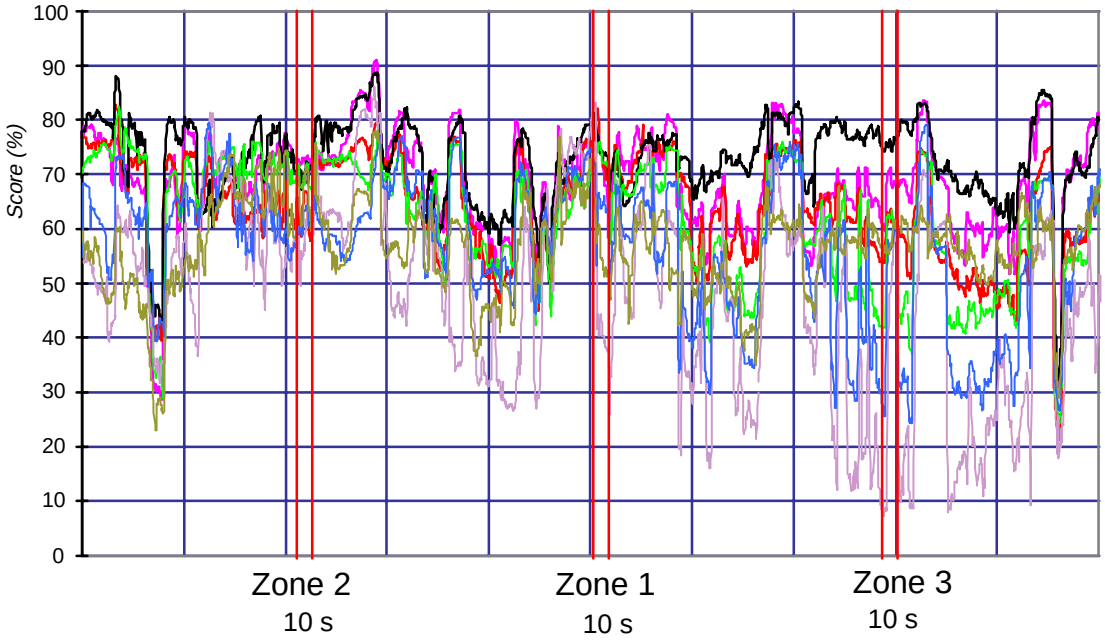


Figure 3
SSCQE – First example of raw data after averaging of the votes (every 500 ms) of all observers. Several Test Conditions (TCs), one Programme Segment (PS) of 30 minutes.

3.3. Examples of SSCQE subjective test results

SSCQE subjective evaluations have already been carried out and have given promising results. *Fig. 3* shows examples of continuous data – averaged over the different assessors of a session – from the same Programme Segment (PS) and for different Test Conditions (TCs).

Fig. 4 shows an example of the comparative quality distributions where different systems (analogue, and digital at different bit-rates) have been ranked in relation to their respective performance with respect to time. It is important to highlight the fact that the various curves – they demonstrate a high degree of stability – were obtained during different sessions, with different observers who were not individually presented with all the Quality Parameters (Test Conditions). This method of assessment prevented the observers from becoming experts on the sequence material under investigation (more realistic assessor behaviour), which is one of the shortcomings of the conventional assessment methodologies.

Further tests have recently been carried out to demonstrate the stability of the method. It confirmed that an observer rejection criteria, like those already used in conventional methodologies (e.g. DSCQS), should be applied to avoid residual variations.

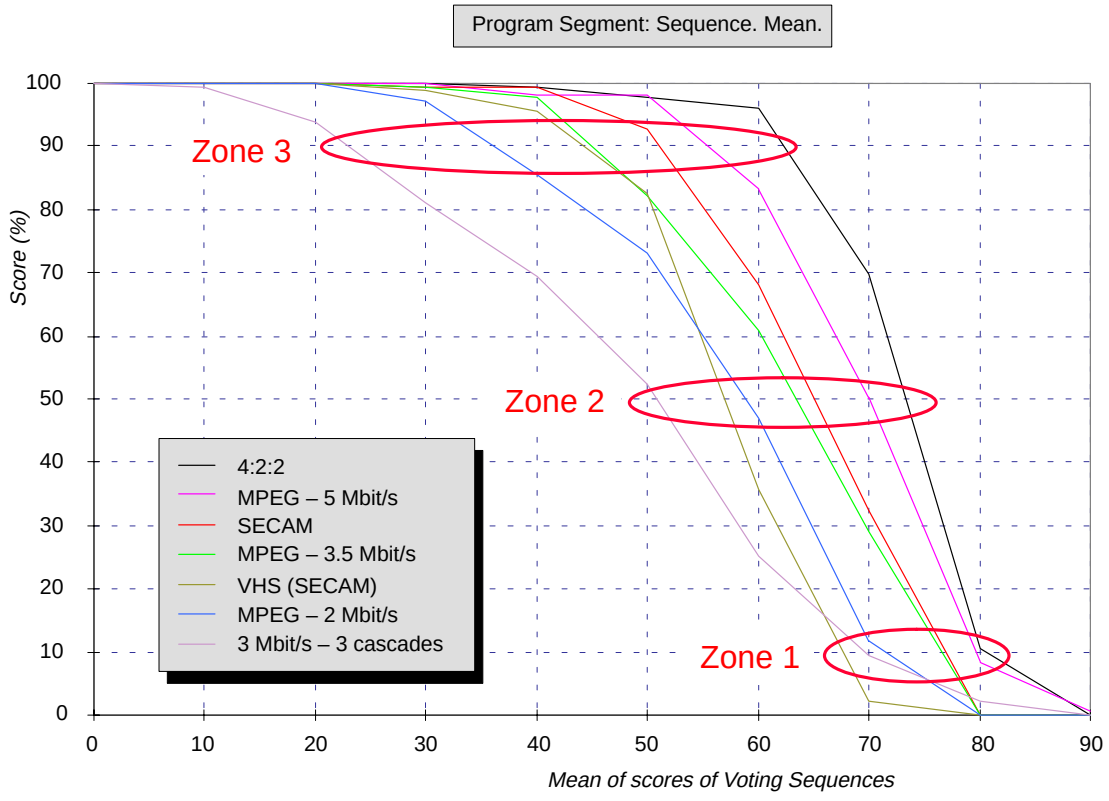
A MOSAIC software has been designed for data-processing and presentation of the results. A powerful advantage of this PC-based solution is that the administration of subjective assessment sessions has been automated as well as the processing of the huge amount of data that is inevitably collected. It also allows the results to become immediately available.

3.4. Comparison with objective evaluation results

Because many different types of parameters and processes are currently envisaged for objective assessment, only preliminary common targets can be proposed.

a) Comparison between quality rating and the corresponding objective data can be achieved at each voting instant. The existing time relationship even allows these values to be correlated with the picture content (type of programme, source, entropy, etc.).

b) Using SSCQE Stage 2, a first attempt can be made to establish a link between quality and the



Zone 1 is representative of what happens 10% of the time, etc.

Figure 4
SSCQE – Example of results presentation after real test data-processing.

adjectival scales given in ITU-R Recommendation BT.500-7 (See *Figs. 3* and *4*).

c) Pending SSCQE Stage 3, the global subjective quality rating could be associated with the equivalent objective measurement result.

It is believed that this procedure could be more appropriate than, say DSCQS only, when evaluating the stability and performance of different objective assessment techniques, particularly if these techniques are aiming to deliver picture and/or programme-service quality estimations.

4. Common interchange datafile format

SSCQE procedures have been defined precisely in order to allow the flexible setting up of tests which require easy test-tape editing and data-processing. A common datafile format was also introduced in ITU-R Recommendation BT.500-7 to help in the interchange of files between laboratories working in the framework of international test campaigns. This format should save a lot of the time currently wasted in re-formatting data from different sources (which often takes more time than the data-processing itself). It should also ease the dis-

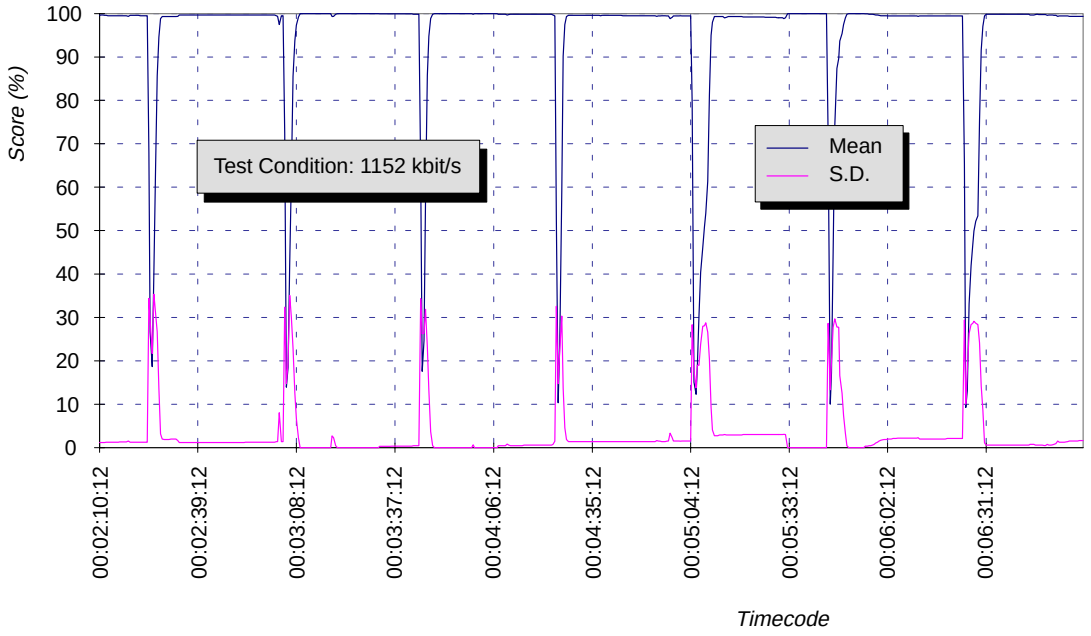


Figure 5
DSCQE – Results after averaging. MPEG-4 error-robustness tests.

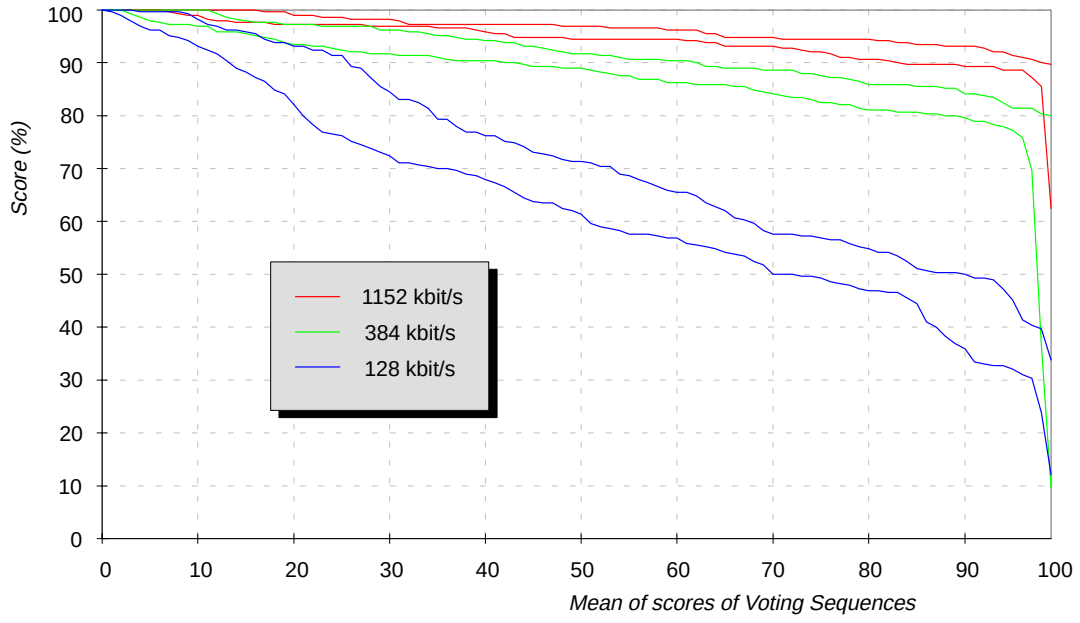


Figure 6
DSCQE – Example of results presentation after real test data-processing. MPEG-4 error-robustness tests.

tribution of evaluation tasks between the collaborating partners (e.g. test preparation, test performing, overall data processing).

This datafile format is compliant with the SSCQE methodology and with the other evaluation methods of Recommendation 500-7. It is made up of text files with a structure that is described in Recommendation 500-7 and its syntax is built around "labels" and "fields" in addition to a limited set of reserved symbols (e.g. [ ] ↵ and =).

The format is also fully compliant with the storage of objective measurement data.

As is is possible to attach a series of related files, there is no intrinsic limitation to the datafile format in terms of capacity, e.g. the number of laboratories, observers, Programme Segments (number and duration), Quality Parameters, voting-scale boundaries or the type of voting peripherals.

5. The DSCQE methodology

The introduction of digital audio-visual services needed a new subjective protocol which is able to measure the quality of service on longer viewing sequences, representative of video contents and statistical error occurrences. The SSCQE method fits this requirement as regards digital TV services. In the case of applications like surveillance, it becomes important to assess not only the basic quality of the images but also the fidelity of the information transmitted. For that reason, it was proposed to adapt the SSCQE method to introduce simultaneous double visual stimuli while still performing continuous quality evaluation.

When performing a DSCQE test, the observers watch two displays. One shows the encoded-decoded video without any transmission errors (i.e. the reference, or source material). The other shows the same video material after alteration by transmission errors. The observers assess the quality by direct comparison, evaluating the fidelity of the video information by moving the slider of a handheld voting device.

An example of data obtained after averaging the votes from the different observers is given in *Fig. 5*. An example of DSCQE results, after data-processing, is given in *Fig. 6*.

6. Conclusions

Most of the MOSAIC specialists, and a few new partners, are now involved in the TAPESTRIES project. In addition to complementary studies in the field of subjective evaluation and psycho-visual perception, TAPESTRIES offers assistance to the other ACTS projects when performing subjective evaluation. Collaboration has been established with other groups dealing with HD-theatres, objective measurements, terrestrial digital services and virtual reality. It is intended to use the SSCQE methodology in each of these areas.

TAPESTRIES has established a plan for co-operation with the ACTS QUOVADIS project, working specifically in the field of objective measurement. TAPESTRIES is also working on mat-

Mr Thierry Alpert *was born in Paris in October 1959. He graduated in Biological and Medical Engineering and in Signal Processing from the University of Paris. In 1988, he joined CCETT where he is currently in charge of the Image Quality Laboratory.*

Mr Alpert actively participates in several European projects and on various Standardization Committees, including ACTS, ITU-R, ISO/MPEG and the EBU. His work is focused on objective and subjective visual quality aspects, mainly relating to services based on digital image communication.

Mr Jean-Pierre Evain *graduated from ENSEA, Cergy-Pontoise (near Paris), in 1983. He joined the EBU Technical Department, Geneva, in 1992 as a Senior Engineer and is currently concerned with the co-ordination of Research and Development projects in broadcasting.*

Mr Evain is particularly involved in new television systems. Over the years, he has been a member of many EBU Working Groups, Ad-hoc Groups and Project Groups, and has also taken part in various system evaluation groups, including HD-MAC, PALplus and MPEG.

Jean-Pierre Evain currently represents the EBU in ETSI, ITU-R Study Group 11, ITU-T Study Group 9 and in the European projects, TAPESTRIES, VALIDATE and UNITEL.

ters relating to MPEG-4. An adaptation of the SSCQE method, i.e. DSCQE, has been proposed to address the specific issue of error-robustness evaluation. The protocol remains identical to SSCQE but a reference picture is constantly displayed in parallel to the impaired picture. This test is more fidelity-oriented than quality-oriented.

The EBU intends to help its Members in the adaptation of their test laboratories to obtain wider use of the new SSCQE methodology. If approved by MPEG-4, information on DSCQE will also be offered.

Acknowledgements

The Authors – as responsible for the development of the SSCQE software tools in the MOSAIC project, and as leader/partner of the TAPESTRIES work-package offering evaluation assistance to other ACTS projects – gratefully acknowledge contributions from all the project partners of these two projects (see the text panel below).

Bibliography

[1] Lodge, N. K. and Wood, D.: **New Tools for Evaluating the Quality of Digital Television – Results of the MOSAIC Projec**t
Proceedings of IBC-96, Amsterdam, 1996.

[2] **The MOSAIC Handbook**
Proceedings of the MOSAIC Workshop, Eindhoven, 18 – 19 September, 1995.

[3] ITU-R Recommendation BT.500-7:
Methodology for the subjective assessment of the quality of television pictures.

[4] Alpert, Th., Contin, L., Koenen, R. and Pereira, F.: **Evaluation Protocol for the MPEG-4 Error Robustness Subjective Test**
ISO/IEC – JTC1/SC29/WG11, MPEG 96/0996, July 1996.

[5] Alpert, Th. and Contin, L. (on behalf of Project ACTS 055): **The Application of Psychological Evaluation to Systems and Technologies in Remote Imaging and Entertainment Services (TAPESTRIES).**

[6] Alpert, Th. and Contin, L.: **DSCQE (Double Stimulus using a Continuous Quality Evaluation) experiment for the evaluation of the MPEG-4 VM on error robustness functionality**
ISO/IEC – JTC1/SC29/WG11, MPEG 97/M1604, February 1997.

[7] Abraham, D., Ardito, M., Boch, L., Messina, A., Stroppiana, M. and Visca, M.: **Attempts at correlation between DSCQS and objective measurement**s
EBU Technical Review No. 271 (Spring 1997).

MOSAIC Consortium

Independent Television Commission (ITC), UK
European Broadcasting Union (EBU), CH
Institute of Perception Research (IPO), NL
Radiotelevisione Italiana (RAI), I
Swiss Telecom PTT, CH
Centre Commun d'Etudes de Télédiffusion et Télécommunications (CCETT), F
University of Essex, UK
Additional services were provided by Médiamétrie, F, and Softel, UK

TAPESTRIES Consortium

Independent Television Commission (ITC), UK
European Broadcasting Union (EBU), CH
Institute of Perception Research (IPO), NL
Radiotelevisione Italiana (RAI), I
Corporate research centre of the IRI/STET telecommunications group (CSELT), I
Centre Commun d'Etudes de Télédiffusion et Télécommunications (CCETT), F
University of Essex, UK
AEA Technology, UK.