Advanced methods for the evaluation of television picture quality : proceedings of the MOSAIC workshop, Eindhoven, 18-19 September 1995

Institute for Perception Research
PO Box 513, 5600 MB Eindhoven

11.09.1995

Rapport no. 1071

Proceedings of The MOSAIC Workshop Advanced Methods for the Evaluation of Television Picture Quality

Roelof Hamberg
Huib de Ridder

Proceedings

of

The MOSAIC Workshop

Advanced Methods for the Evaluation of Television Picture Quality

Institute for Perception Research,
Eindhoven, The Netherlands
18–19 September 1995

edited by
Roelof Hamberg and Huib de Ridder

Institute for Perception Research
P.O. Box 513
5600 MB Eindhoven
The Netherlands
Phone: +31 40 773 873
Fax: +31 40 773 876
e-mail: iposecr@natlab.research.philips.com

Contents

Chapter 1

MOSAIC – A Bridge between Laboratory Quality Evaluation and Reality

David Wood
EBU

MOSAIC is a two-year collaborative research project in the framework of the European Community's RACE programme. It began in 1994, and brings together a half-dozen European organisations with expertise in picture quality evaluation. Its objective is to provide a series of new tools for evaluating television picture quality. These will be more accurate and more useful than those we have today.

But how should picture quality evaluation be placed in the scheme of things for television broadcasting?

What really influences the public in their decision to buy a new piece of media hardware - HDTV, EDTV, Widescreen, digital or analogue?

The first thing is the "service" and not the "system". What the viewer buys is something that will allow him or her to see the programme he wants. Whether it is digital or analogue, hybrid DCT or MAC, does not fundamentally matter. The public buys something that will allow them to watch MTV (or whatever). The first law of new television systems is the primacy of the service over the system.

While the service is the most important factor in the success of a new media product, it is not the only one. This is the second law of new television systems.

The product has to be seen as a package that includes programme content, product hardware packaging, the technical quality delivered, and the cost. All of these play a role in the success or failure of any new system.

It may come as surprise, and even a disappointment, to learn that 50% of the public claim to determine their choice of television receiver by whether they like the cabinet or not. This is one of the sobering realities we have to live with in the research laboratory.

Another critical factor is the cost. The consumer has to decide how much the quality and services are worth to him in the light of his disposable income. There are no laws of physics to tell us how much people are willing to pay for new services or

new hardware.

1.1 The degree of improvement

If we put aside the question of content, it is clear that different technologies offer different degrees of improvement over what the viewer already has. This is also a major determinant of success or failure.

When colour television started in the late 1960s, a colour receiver cost four or five times as much as a black-and-white receiver. The price of a new, large-screen colour television set was around £350 - around the same price as a new mini-van. The choice was between a cheap car or a colour television. In spite of this, the rate of penetration of colour television was comparatively rapid in Europe. Something that provides a dramatically better service will sell, despite an enormous price-tag. Colour was a first-division innovation, at a time when television held an important place in people's lives.

Not all innovations are so striking. When teletext started in Europe, the growth rate was much slower, in spite of the fact that the cost increment was much less than for colour. Teletext is a success today, but it has taken many years to achieve extensive penetration.

How will digital television rate in its "degree of improvement"? There is no single answer to this, because digital television is capable of providing the complete range of picture qualities.

1.2 The concept of enjoyment

Having made the decision to purchase a new item of media technology, the customer has to use it more or less regularly. The extent to which he uses it will be determined by what we might call his "enjoyment". This could be seen as a range of mental states running from high to zero and below.

A number of factors contribute to the viewer's enjoyment. Two of the most important are programme content and picture and sound quality. Provided the programme itself is of at least watchable quality, the most important element is content. Nevertheless, picture and sound quality are part of the equation.

There is evidence that, in the developed countries, the expectations of the public as regards picture and sound quality are rising. Experiments conducted in the 1950s on picture quality and repeated in the 1980s shows that what appeared to be "excellent" in the 1950s was less than excellent in the 1980s. The human race is adapting to technology and is becoming more discriminating. The same is happening in sound. Today, AM radio sounds less than perfect to a sizeable portion of the population. The same fate will eventually befall FM radio with the introduction of DAB. Incidentally, it is probably true that the public is becoming more discriminating in other areas, too, including production grammar.

Picture quality comes down to something like the perceived degree of reality is possible to depict. This, in turn, can help define the viewer's sense of involvement in the programme, and is linked to "enjoyment".

1.3 Quality factors

Just like "enjoyment", picture quality is the cumulative effect of a number of different factors. Determining these factors brings us into the science of picture quality evaluation.

Television picture quality factors are traditionally considered to include - amongst other things - definition, colorimetry, artifacts, and sharpness.

Much research has shown that, although definition is not without influence on picture quality, it is not the most important factor. In fact, the most important quality factor - after the addition of colour - is "apparent sharpness". This is not the same as definition. Apparent sharpness is influenced by contrast and brightness and can be artificially enhanced by adjusting the frequency response without adding more definition. People like bright, high-contrast pictures which have objects with sharp edges. In their minds, this is what corresponds to high-quality pictures.

Within limits, lack of artifacts is also a primary quality factor. Higher definition comes behind all this in terms of its contribution to perceived quality. Of course, it still does make a difference.

1.4 Viewing distance

Viewing distance has a dramatic effect on picture quality evaluation. The further you are from the screen, the less able you are to see any shortcomings. The eye is saturated with picture detail at a given number of elements per viewing arc. The further the viewer is from the screen, the less detail the eye needs, to become saturated. In other words, from far enough back, everything looks like HDTV.

In the field of subjective evaluation, we currently distinguish two viewing distances: the "design viewing distance" and the "preferred viewing distance".

The "design viewing distance" is the minimum distance for which the system is intended. It is approximately that at which the eye is saturated with detail. Viewed from that distance - or further - from the screen, the picture will be about as good as it can be. Viewed from closer distances, the shortcomings of the system will show up.

The preferred viewing distance is the average distance that viewers prefer for a given system and screen size. This is always larger than the design distance.

As an example, for PAL and SECAM the design distance is six times picture height ($6h$). The preferred viewing distance depends on absolute screen size, but for current normal living room sizes of television set it is eight or nine times picture height ($8/9h$).

If we move to larger screens in future, this preferred viewing distance will shrink. If the room size stays the same and the screen size increases, the relative viewing distance will be reduced.

1.5 Classifying television systems

Viewing distance is such a critical factor that the ITU now uses it to classify television systems. For a system to fall into the HDTV window, it now has to be classified in the "excellent" range for virtually all picture material at viewing distances of three times picture height ($3h$). For a system to be rated EDTV (enhanced-definition television), it needs to do the same at $4h$. For SDTV (standard-definition television) the figure is $6h$, and for LDTV (limited-definition television) it is $9h$.

The design viewing distance is at the foundation of the different television systems. The PALplus system, for instance, meets the EDTV criteria, and PAL and SECAM meet the SDTV criteria.

1.6 The influence of content on perceived picture quality

An important hidden influence on the perception of quality is the content of the scene or programme. Irrespective of the technical parameters of the system, the type of programme being viewed affects the perception of quality.

Within this area come elements such as type of programming, production grammar, and production method. The latter includes whether the programme was made on film or video, and possibly which kind. Each of these recording media has its own "look" which, consciously or unconsciously, we notice.

The generic types of programming include movies, electronic drama, light entertainment, documentaries, news, and sport. The technical elements of picture quality have a different degree of importance for each type. The shooting media and/or production grammar used differ, and the viewer does not have the same expectations.

The look that comes with 35mm film production is more immune to quality limitations than other types of production method. The 35mm product has a gloss and a feel which makes it usually look good, whatever the transport. This may be linked to film grain, the shallow depth of focus usually used, or other reasons.

Probably the type of programming most able to benefit from selected higher quality is sport. This is because as sense of reality is much more important here, and many sporting events are played out against a wide canvas, which needs detail.

Other types of programme material lie between these extremes.

VHS was and is a success despite limited quality. Two major reasons may be the overriding benefit of the service provided and the fact that most videotapes are feature films which look comparatively good even at quite low quality levels.

Having seen that different types of programming may not require the same real quality levels to achieve the same perceived levels, there is another and critically important dimension of complexity to grasp. This is the variation in quality directly associated with activity in the picture.

This is particularly important for today's digital systems. It is necessary to understand these concepts to define into which quality window a given system falls.

1.7 Laboratory subjective assessments

Subjective assessments are controlled psycho-physical experiments designed to find out what the average viewer would make of a particular picture or sequence at a given viewing distance. We imagine that there is a quality continuum in his mind running from the best imaginable picture (a look out of an open window?) to the worst imaginable picture (whatever that might be). We ask him to say where, between these two extremes, he thinks the picture lies. We take a whole series of precautions to ensure his assessments are unbiased by any factors other than the real quality of the picture.

By definition, we are assessing the average of the population as a whole, and since most people in the world are not picture quality experts, we use assessors who are ordinary people (of course, we check they are not colour blind) rather than experts. The sample is big enough to be statistically representative.

There are a range of psycho-physical methods and tools to use in making formal subjective evaluations. The distilled experience of many years work, about which methods are the most practical and reliable, is given in the ITU-R Recommendation 500. The MOSAIC partners were largely responsible for this recommendation in years gone by.

Recommendation 500 explains the gamut of evaluation method possibilities, but focuses in detail on two methods. These are the 'Double Stimulus Continuous Quality Scale Method (DSCQS) and the Double Stimulus Impairment Scale method (DSIS)'. There are different occasions when each is appropriate. Generally, the DSCQS method is used to evaluate how well the system can perform in itself, and the DSIS method is used to evaluate how it travels over transmission media.

Workers in the field readily admit that these two methods can only be called the 'least worst' available, rather than being flawless. They certainly do have shortcomings.

1.8 The problem of choosing test material

A fundamental difficulty in designing subjective evaluations using these methods is knowing what pictures or sequences to use. To some extent, all television systems, new or old, perform differently with different pictures. Taken over a long period of time, a television channel will contain a wide range of picture types, some occurring

more often, some less often. In fact, to describe the contents of a television channel, you need a distribution of different types of picture. Since the performance of the television system will be affected by the picture contents, we could expect that, in addition to a distribution of picture types, there will be a parallel distribution of picture qualities associated with a given television system.

We might therefore take representative samples from all types of picture, but usually, for lack of time and money, we have to be selective. Normally, we select a series of pictures that are of above-average difficulty for the system being evaluated. There would be no point in selecting easy ones - unless we were car salesmen.

The guideline is that the pictures or sequences chosen should be "critical, but not unduly so". The theory is that is you evaluate, say, a dozen test pictures of sequences that are "critical, but not unduly so", then you have a fair measure of how the system performs.

This general philosophy worked well for many years with analogue television systems and then with low-compression digital systems. However, it reaches the limits of its usefulness with high-compression digital systems. The reason is that the available quality depends on content of the picture even more than with analogue systems.

The way codecs work means that - to exaggerate a little - low-entropy pictures look good, whatever the bit-rate ("entropy" is a measure of uncertainty). For instance, when cinema films at 24 picture per second are passed through high-compression codices, this is downhill for the codec. The problems don't come there: they come with pictures having lots of detail and action, like 50 Hz (or, worse, 60 Hz) sporting events shot with video cameras.

This is really the important point to bear in mind when examining digital codecs. When the salesman shows you a codec, asks to see the pictures it won't pass unimpaired. There will always be some. Then you will know what the codec can do.

A key aspect of MOSAIC's work has been to look for methods to measure the way a system performs which takes account of the range of programme material likely to pass through it. This is not an easy matter. If it was, we would have found solutions many years ago. However, today with data processing capacity readily available, more adventurous thinking is possible.

The MOSAIC project has looked at a range of issues which will help to make formal subjective evaluations more useful indicators of how a system is likely to perform. These have included the viewing conditions under which the evaluations are made, the influence that sound has on picture quality, the influence of the display size and the scanning algorithm. We have also developed computer software for collecting and analysing the results.

1.9 Aide memoir for quality

In respect of quality, it may be helpful to bear a number of points in mind.

Quality matters very definitely and can be a key element in the viability of new systems and services. Of course, quality alone is not enough. It must be coupled to

services that the public is willing to pay for.

Quality evaluation for digital systems is a complex affair and it is a mistake to believe that a few simple quality grades characterise a system. Some notion of the statistical distribution of quality is needed. Furthermore there are a range of factors which bias subjective evaluations, and we need to take them into account. We need to understand them and their effect.

Doing so has been the work of the MOSAIC project.

Chapter 2

An Introduction to Advanced Subjective Assessment Methods and the Work of the MOSAIC Consortium

Nick Lodge
ITC

2.1 Introduction

No designer of an image source, transmission, or display system will need reminding that their ultimate objective is to provide pictures which are subjectively optimal for the final human user. It is therefore obvious that methods must be employed, as part of the system development, which will determine viewers' opinions of the subjective quality achieved and/or the subjective nature of any picture impairment which may occur as a result of an external influence (e.g. channel noise). The designer will then wish to feed back the results of these measurements in a systematic way so that he/she is able to improve the design by matching it more closely to the requirements of the viewer and the application. It is a sad fact however, that subjective testing methods have not kept pace with the advances in the television systems which they are being used to assess!

Over the years there has been much careful experimentation to arrive at methodologies by which the opinions of observers on the quality or impairment of a television picture, can be gathered. These specify the procedural and environmental conditions of the subjective tests as closely as possible so that results obtained on one system, on one occasion, with one set of observers and in one laboratory, will be comparable with results obtained elsewhere and at other times. The most recent state of this work is given in [1]. This Recommendation is updated at yearly intervals, but, with full published volumes appearing only every 4 years, it is hardly surprising that the methodologies have struggled to maintain their usefulness in the rapidly advancing world of television technology. The advent of the D1 digital VTR has probably given the biggest boost of recent years, to the reliability of subjective assessment methods,

since it has ensured consistency of reproduction during test sessions and has allowed the compilation of libraries of internationally accepted test picture sequences.

Recently the broadcasting world has seen the development and standardisation of a large number of television picture processing systems intended for professional and domestic transmission applications, as well as for use within the studio. Examples of these are numerous: HD-MAC, 34 Mbit/s inter-studio contribution codecs, PALplus, MPEG-type digital codecs, 'non-linear' editing facilities, digital videotape machines, standards converters, noise reducers and slow motion interpolators. One feature that all these processes have in common is that they are *adaptive*, that is they alter their behaviour depending upon the content of the picture, or part of the picture, that they are handling at any particular time. For efficient image transmission or storage, processes which rely on the reduction of picture signal redundancy, there are good theoretical reasons why this adaptivity is essential.

Redundancy reduction is not only concerned with the more obvious forms of statistical redundancy, such as the likely similarity between successive picture frames, but also with psychovisual redundancy - the exploitation of the inadequacy of the human visual system to perceive certain types of distortion against a background of certain local picture content. Optimisation of these systems requires extensive use of specialised subjective test procedures.

The presence of adaptivity in television processing means that attempts to relate objective television measurements to subjective quality are no longer useful because the nature of artifacts which can arise is so varied. Even employing existing subjective assessment methods is not totally reliable, because adaptive systems usually exhibit scene-dependent quality and this raises the question of how a few 10 s picture sequences should be chosen to be representative of typical broadcast television content.

More than ever before there are strong commercial reasons for taking subjective measures seriously. Collaborative as well as competitive research programmes frequently organise subjective assessment campaigns to determine which of several proponent systems should be selected for further development or standardisation; the financial implications of losing these competitions can be considerable. Even after implementation, service operators will be faced with decisions about system trade-offs such as the quality vs quantity of television services which should be offered through a fixed capacity channel. Buyers of professional studio equipment too, may also organise subjective comparisons between competing systems to assist in their choice. Failure to judge the requirements of customers or to take account of the subjective performance of competing systems or services, could well spell commercial disaster.

2.2 The design of subjective assessment tests

In an ideal world, psychologists and engineers would have published a single method for the reliable and sensitive, subjective assessment of any television system. Neither the application of the method nor the processing of the data, would require any

specialist advice, and the results would be immediately representative of the opinions of normal viewers watching television. The reality today has been somewhat different.

Despite the 'user friendliness' of Rec. 500-6, which advises how to choose an appropriate method from a limited set of options, questions will still arise in the mind of the non-specialist: How should the 'reference' picture condition be chosen in an impairment test? Will revealing clues in the nature of certain distortions, invalidate the methodology? How should observers be briefed before each session? How many observers are necessary to ensure the statistical significance of the results? When is it appropriate to use an 'anchor'? Remember too, that subjective assessment is expensive, not only in terms of observer time, but also in picture sequence preparation, videotape editing and statistical analysis. It is little wonder that system designers embark far too infrequently upon subjective evaluation to guide their work.

Let us examine the key elements involved in subjective test design:

Basic methodology Here it must be considered whether quality or impairment should be measured, or whether a comparison between two or more systems should be made, perhaps in order to rank their subjective performances. It will be necessary to decide if a single or double stimulus method is more appropriate and how the 'test' and 'reference' (if used) conditions should be presented. For the benefit of the observers' attentiveness, and therefore test validity, it is important to ensure that the test duration is not excessive.

Viewing conditions So that results can be compared with, or contributed to, assessments performed elsewhere, it is essential to adopt standard conditions for room illumination, monitor set-up and the seating of observers with respect to the monitor.

Choice of observers It will be necessary to decide how observers should be screened (e.g. for visual acuity, colour blindness) prior to their participation and whether there are certain classes of 'expert viewers' who should be excluded.

Scaling method A key issue in the registering of observers' votes is to decide which scale they should be recorded against. There are many scales from which to choose, each having particular advantages and disadvantages.

Reference conditions The choice of a 'reference' picture condition for comparative or double- stimulus testing is not always a simple matter. What for example would one use in the subjective evaluation of a film to video transfer system?

Presentation timing The pattern and duration of the presentation of picture sequences, and the period allowed for voting must be carefully determined and is a compromise between: allowing sufficient time for observers to make reliable judgements, but not so long that their memories of earlier conditions have faded, and not so long that the total session time becomes excessive. Also allowing sufficient repetition of test picture sequences that various factors under

study can be explored, but not so much that observers become over-familiar with them.

Test picture scenes This is a particularly important issue, which will be thoroughly considered later. The picture material used must be chosen scientifically, so that on the one hand it is demanding for the system under test, but on the other hand it should be understood just how representative of real television content it is. A set of scenes could easily be chosen, for example, having a predominance of colour transitions and high frequency stripes, to demonstrate that the PAL television system is completely unusable for broadcasting.

Analysis of voting Obtaining the required information from the raw votes of the observers is, of course, essential but it is important to understand what processing may be validly done on the data and what limitations may be imposed by the basic methodology. Is it possible, for example, to identify those observers who have been inattentive or confused, and remove their votes from the analysis? Remember too, that analysis of the results not only yields the mean opinion scores sought, but also variances and significance information which is vital in interpreting the results and ultimately judging how successful the evaluation has been.

Results presentation Choosing the most appropriate way in which to present the results of subjective evaluations is important, both for ease of interpretation and for purposes of comparison between different experiments. Members of the MOSAIC consortium believe that the processing and presentation of the results of subjective evaluations should be a subject of standardisation and are working toward this end.

2.3 A brief look at current methodologies

There are two methods which together account for the vast majority of subjective assessments performed today, they are the double stimulus impairment scale (DSIS), and the double stimulus continuous quality scale (DSCQS) methods. There is no rigid advice on whether one should choose a 'quality' or an 'impairment' test in a particular situation. Generally, if an experimenter is causing different amounts of degradation to a picture, and the relevant issue is the difference between the original and the distorted versions, then an impairment scale should be used. Where one or more systems are tested under normal operating conditions (no external distortions are introduced), so that interest lies in the fidelity of the reproduced picture with respect to the source, then a quality scale should be used. Occasionally a sensitive differentiation between pairs of systems is required. Here a comparison scale may be employed and can also prove useful for rank-ordering small numbers of systems.

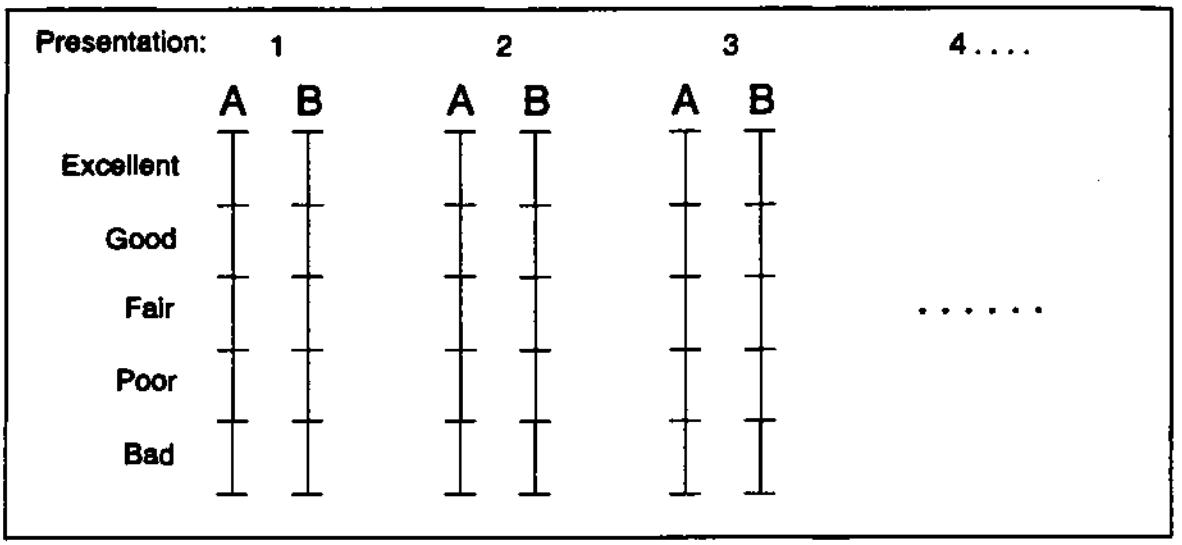


Fig. 2.1. Portion of the double stimulus continuous quality scale voting form

2.3.1 The double stimulus impairment scale

This method uses a cyclic presentation where observers are first shown an unimpaired reference picture sequence (for ~10 s) and then the same scene subjected to the impairment under test (for ~10 s). They are informed which of the pictures is the reference and which the test condition, and then asked to vote on the second, keeping the first in mind. Throughout the session, which should last no longer than 30 minutes, all impairments of interest are shown to the observers in a random order of combinations covering all the test scenes. This time permits 40 presentations to be made. The unimpaired picture is also included as one of the assessed conditions, and all presentations are repeated twice within the session. Care is taken to ensure that, although the presentation order is random, the same scene is never used on two successive occasions.

Voting is on the basis of the observer's 'overall impression' of each test picture, and is recorded using one of five discrete impairment grades: *Imperceptible, Perceptible but not annoying, Slightly annoying, Annoying, Very annoying.* For processing, each grade is assigned the numerical representation 1–5, and results are presented by mean and standard deviation for each test parameter.

2.3.2 The double stimulus continuous quality scale

This is another cyclic method employing pairs of picture sequences. One picture of the pair is directly from the source and the other results from the source signal after it has passed through a system under test (although the condition where both pictures are from the source is also used). These pictures are randomly designated

A and B, and shown (for ~10 s each) to the observers, who are asked to grade both, with no knowledge as to which is the source and which the test picture. Throughout a session, which should be limited to 30 minutes, all combinations of systems under test and scenes are shown, not only in a random order, but are also twice, where the designations to A and B are reversed on the second occasion to eliminate bias.

Voting is performed by the observer making a mark at an appropriate point on a continuous scale for both pictures A and B. Fig. 2.1 illustrates a portion of the voting form where the five adjectives: *Excellent, Good, Fair, Poor, Bad* are shown as a guide for the observer. Statistical processing of the presentations is usually based on the measured difference between the marks on scales A and B expressed as a percentage of the scale length. The differences are often re-converted to equivalent quality grades, and mean scores for the source and test conditions are presented for each combination of variables.

The use of adjectival descriptors has been criticised on two grounds. First, the perceptual intervals between the terms used are known not to be equal [2], with for example, *poor* and *bad* being perceived to be much closer than *excellent* and *good*. Secondly, when translated into other languages the terms give rise to yet different perceptual intervals, so that for example, *mäßig* and *schlecht* are further apart in the mind of a German than *poor* and *bad* are in that of a native English speaker. This has consequences for the validity of subsequent linear statistical processing and international subjective evaluation campaigns.

2.3.3 Single Stimulus Methods

Are those in which no comparison with an unimpaired reference condition is invited during every presentation, i.e. only one picture sequence is shown each time. These are not as popular as double stimulus methods but are most often used whenever an appropriate reference sequence cannot be derived. They do have the advantage that they are quicker to perform, but are known to be sensitive to the range and distribution of the conditions shown. This problem can be partly alleviated by including in the presentation, 'anchor' sequences, which represent the most extreme conditions displayed.

The definitions of standard viewing conditions for all methods are given in [1] for conventional TV, and [3] for HDTV. These references also advise on the choice of observers, as does [4] pp. 45, 46.

2.4 What is wrong with current methodologies?

For evaluating many established systems, the answer to this question is 'nothing', however recently two factors have completely changed the environment in which we wish to apply subjective evaluation methods. The first is the advent of adaptive systems (e.g. digital compression), which give rise to scene-dependent quality, as was noted in the introduction - existing methods cannot meaningfully describe this

behaviour. The second is that technological advances mean that adaptive systems are no longer solely the domain of the studio, but are poised to be used in direct-to-home delivery. The consequence of this is that we may wish to take a *commercial* rather than *professional* view of picture quality: no longer shall we consider laboratory viewing conditions, but domestic ones instead; and no longer shall we aim to deliver excellent quality across all types of picture material, but shall ask such questions as 'how much occasional distortion will the viewer tolerate before he will switch to another service?'. In ascertaining the worth to the viewer of picture quality, against say, number of available channels or waiting time for a near-video-on-demand movie, it may well be that we would wish to get away entirely from existing methods and employ marketing-type comparisons along the lines: 'would you swap two packets of Brand X for your regular soap powder?'

Let us briefly examine the failings of existing methods in this new environment [5]:

1. **Test picture sequences are too short** - Existing methods use sequences of 10–15 s duration, this is not only inadequate to judge scene-dependent quality variation, but is also very different from real domestic viewing, where scenes are watched in the context of a programme lasting tens of minutes.

2. **Choice of test material is unscientific** - Choice of test material has usually been governed by politics, availability, and vague concepts, such as that it should be 'critical but not unduly so' (CCIR Rec. 500–4). No attempts have been made to ensure that it is statistically representative of real TV content.

3. **Comparisons are usually made with a 'reference'** - In the usual double stimulus procedures, test pictures are compared with an unimpaired reference. While this results in sensitive laboratory evaluations, such side-by-side comparisons are not available to the domestic viewer, who would very probably not notice some occasional distortions.

4. **Test picture sequences are repeated** - Existing methods employ repetition of test sequences, so that observers have time to seek out particular details in the scene which reveal distortion most readily. They then base subsequent judgements solely on these. In real TV viewing, scenes are shown only once (or with sufficient time in between to forget details), so intimate familiarity with scene content does not have time to develop. Some observers also report lack of attentiveness due to boredom from the presentation of repeated sequences.

5. **Viewing distances are short** - Studies of domestic viewing conditions consistently show that most viewers sit at a far greater distance from their screens than the existing $4h$ and $6h$ assessment standards. This means, for example, that some occasional resolution variation would probably not be noticed by the domestic viewer.

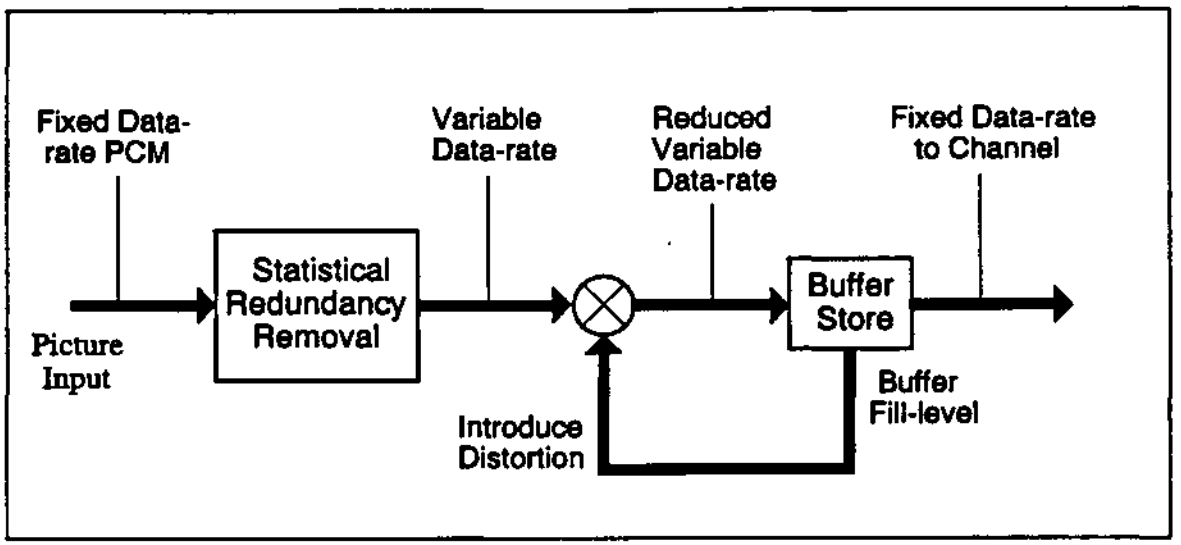


Fig. 2.2. Model of a digital television compression encoder

2.5 Progress towards a new methodology

2.5.1 Scene-dependent quality

The first step in developing a new methodology is to understand how the performance of advanced systems, which exhibit scene-dependent quality, can be characterised. Certainly the most common of these systems, and the one which will be concentrated upon here, is digitally compressed television. The compression process, explained for example in the tutorial paper [6], operates by removing redundant information such as the similarity between adjacent frames and pixels, from the picture signal. A model of this is shown in Fig. 2.2.

Notice that the system employs a control loop to ensure that the buffer store, which matches the irregular redundancy-reduced bits entering it, to a constant rate channel, does not overflow. It does this by introducing into the transmitted picture, some degree of distortion which is tuned to the properties of the human visual system in such a way that for most television scenes it remains imperceptible. Occasionally however, very busy scenes will occur in normal television which require a large number of bits for their adequate reproduction, these will stress the buffer store and will consequently appear noticeably distorted at the receiver.

The first question which one might seek to answer then, is: given a particular compression method operating at a particular bit-rate, how often and to what degree will subjectively noticeable distortion occur on typical TV programmes? One way to answer this would be for a panel of observers to vote on the quality of many thousands of test scenes which have passed through a compression system. The resulting distribution of percentage occurrence vs subjective quality could then describe

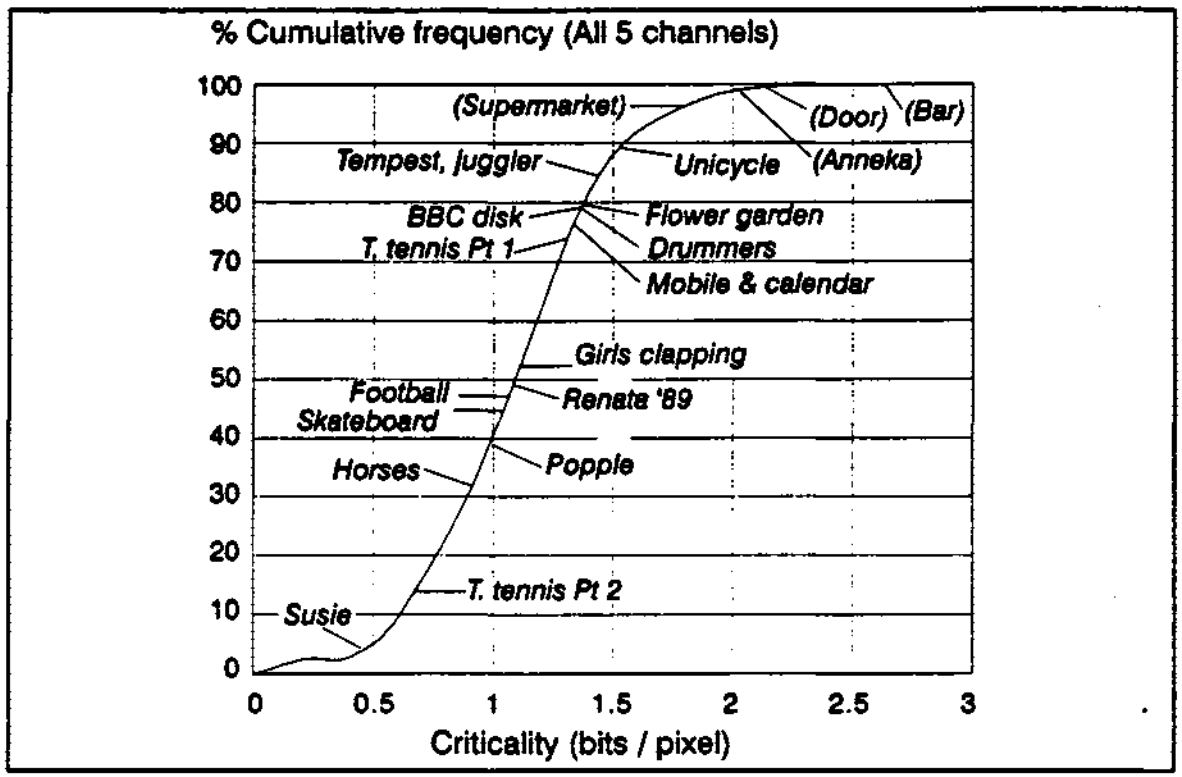


Fig. 2.3. Cumulative criticality distribution for 27369 television scenes

the scene-dependent quality variation. Naturally this is impractical, but in an early experiment [5] a closely similar result was reached by a more realistic means:

Rather than employ the observers to grade all the scenes, a rank order was first established according to the difficulty which a particular compression system would experience when presented with each scene (measured by using the rate-buffer occupancy or a similar analytical approach). The observers then needed only to assess compressed versions of a representative sample of the ranked scenes. This method was attempted in an experiment which employed 27369 short scenes recorded automatically on D1 tape over a week, from 5 channels carrying component television. These were the former BSB channels: 'Power Station' (pop music videos), 'Galaxy' (general entertainment & news), 'Sports Channel', 'Movie Channel', and 'Now' (news & features). Not only did this set of scenes provide a wide range of material, but it also allowed analysis channel by channel, to discover if some carried more demanding material than others.

The resulting rank ordering was produced in days using a large parallel processing computer which simulated the behaviour of a normalised motion-compensated hybrid DCT compression codec employing subjectively optimised quantisers (typical of MPEG-type systems). The cumulative distribution of ranked scenes itself, is a valuable tool since it permitted, for the first time, a calibration to be made of the

CCIR/EBU standard library of test scenes in terms of typical television content. To do this, the standard scenes were subjected to the analysis programme, and their resulting 'criticalities' marked (those not shown in parentheses) onto the cumulative distribution (Fig. 2.3). It proved to be a revelation that, despite the fact that these scenes had been very widely used in all international studies on digital compression, none was representative of the most critical 10% of scenes occurring in typical television - 6 minutes in every hour would look worse than anything that had been seen in evaluations performed using the library! The scenes shown in parentheses were examples identified from the 27369-scene sample.

In the frequency distribution of the same results (Fig. 2.4) some interesting characteristics are revealed, in particular the presence of three distinct peaks. Examination of the source material showed that the small left-hand one was due to the presence of captions, which being computer-originated, exhibited very little noise and achieved low criticality scores. This peak was strongest in the 'Sports Channel' statistics. The second peak consisted of scenes of talking presenters and news readers. These were characterised by a still camera, no background movement, and were often live and so contained low noise levels. This peak was strongly evident in the channels 'Now' and 'Galaxy' which carried news, and was totally indiscernible in the 'Movie Channel' statistics. It is perhaps most interesting because it is representative of the statistics of video conference scenes, which had not previously been compared in the same graph with those of entertainment television. The third major peak is indicative of the spread of criticality in general television scenes. Its character did not vary greatly across the channels, however the 'Sports' and 'Power Station' channels did exhibit a slightly higher proportion of very critical scenes, with almost all of the most critical 1% of scenes being attributable to 'Power Station'.

2.5.2 Subjective calibration

A representative sample of the ranked scenes was compressed to the bit-rates of interest in order to obtain subjective quality grades for them. Here the DSCQS method was used so that the results obtained could be interpreted in terms of the many documented tests which have employed this method. In the assessments, three DCT-based compression approaches were used, the results of one, the MPEG-1+ algorithm at 5 Mbit/s, are shown in Fig. 2.4. This characterisation of variable quality revealed that 84% of scenes were reproduced in the 'excellent' range (0 - 20% of the DSCQ scale), 15% were in the 'good' range (20% - 40%), and 1% of scenes fell into the 'fair' range (40% - 60%).

This measure is invaluable for comparing the performance of different systems since it is determined across a very wide range of different picture material, it should however, be interpreted with care, since statistics can be considerably different in whole atypical programmes. As an example of this, consider watching television through a system which does not reproduce well, thin black lines moving against a white background. While for most programmes this may not be a severe handicap,

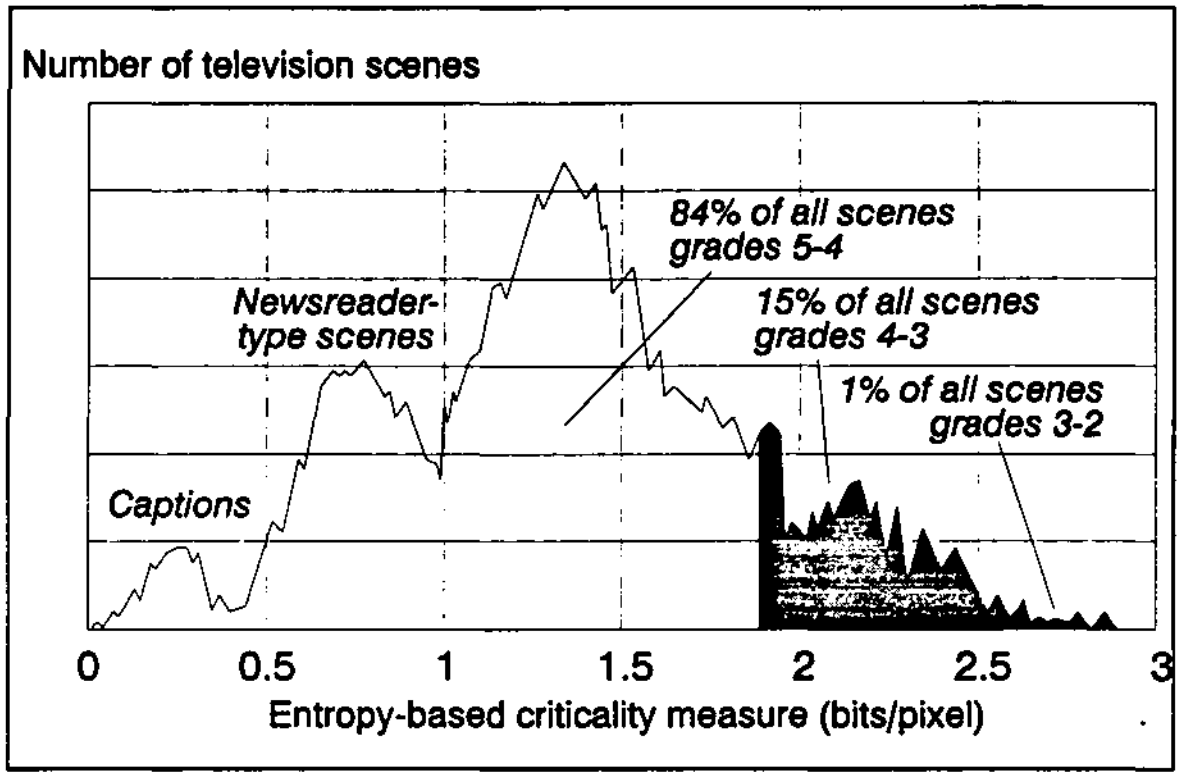


Fig. 2.4. Distribution of scene-dependent subjective quality variation for MPEG-1++ at 5 Mbit/s

imagine trying to enjoy a half-hour slalom skiing contest, where the poles defining the gates not only become worst-case features, but also demand the visual attention of the viewer. Do not forget too television advertisements, which employ dynamic presentations and rapid scene changes in order to attract and interest the viewer. These may also be broadcast frequently enough for viewers to become familiar with them. They are likely to be more critical than typical television, however since they represent considerable investment on the part of the advertiser and income on the part of the broadcaster, both are likely to be intolerant of any artifact whatever.

It is also worth pointing out that some of the scenes discovered in the large sample (e.g. Door and Supermarket) are extremely critical because they contain an unusually high level of source noise, which to a compression system appears as a high source information content. Such scenes are however, not suitable as reference sequences in double stimulus tests because after compression they do not look significantly different from before, even though they will actually have undergone considerably distorted. The A–B difference measure in the DSCQS assessment for such scenes, will therefore be very small.

2.5.3 Viewer tolerance to quality variation

So this early work characterised the proportion of television scenes which are likely to exhibit distortion, but did not suggest how tolerant the viewer will be of this. Clearly the durations of presentations in existing subjective assessment methods will be too short for meaningful judgements to be made. The best approach to determining this in a commercial environment is to employ a method which involves exposing viewers to entire programmes which have passed through a system under test. This is now becoming a practical proposition because flexible prototype hardware is more readily constructible, and even where it is not, simulated impairment of a small proportion of scenes can be performed by computer.

Such a method will introduce a range of distractions, which although typical of domestic viewing, may make it difficult to measure the influence of picture quality alone. Studies of some of the influences are in progress and are covered by other contributions to this workshop. They aim to answer such questions as: to what extent will the viewer excuse occasional poor scenes when the vast majority have excellent quality?; will the viewer's final and overall impression of programme quality be based largely upon the quality of scenes shown towards the end of the programme?; and to what extent will interest in parts of the programme content influence the annoyance of distortion? The method used for recording viewers' opinions will form a vital element of a new methodology. One approach is to give each a continuously variable indicator knob, with which they can register the level of their dissatisfaction whenever the picture quality declines. This could perhaps be scaled with respect to some overall rating given by each viewer at the end of the programme. Methods such as this have been used for many years in continuously assessing the enjoyment or dissatisfaction of viewers with the content of pilot programmes and commercials, the MOSAIC project has been carrying out extensive studies on the application of this approach to picture quality and impairment assessment.

Fig. 2.5 shows a flexible arrangement for conducting subjective assessment and optimisation processes. Play-out of source material, which could be in the form of a complete television programme or a large number of short sequences, is controlled by a small personal computer. As the tape plays, parameters of the system, such as the occupancy of the rate-buffer in a digital coder, can be logged against time-code to enable the compilation of the system's statistical response to the input scenes. The system also has the capability to record the continuous or discrete opinions of a number of viewers. The statistical processing and relating of these data can then be performed rapidly to produce the required characterisations and performance measurements of the system under test. After the test, the response of one or all observers can be displayed on-screen as the tape is re-played, in order to discuss or review the session.

The material being played from the tape can, of course, be chosen from a particular class of material (e.g. sporting events or pop music 'videos') or can contain pictures which have already been passed through other potential sources of distortion such as 34 Mbit/s contribution systems or non-linear editing suites. This allows assessments

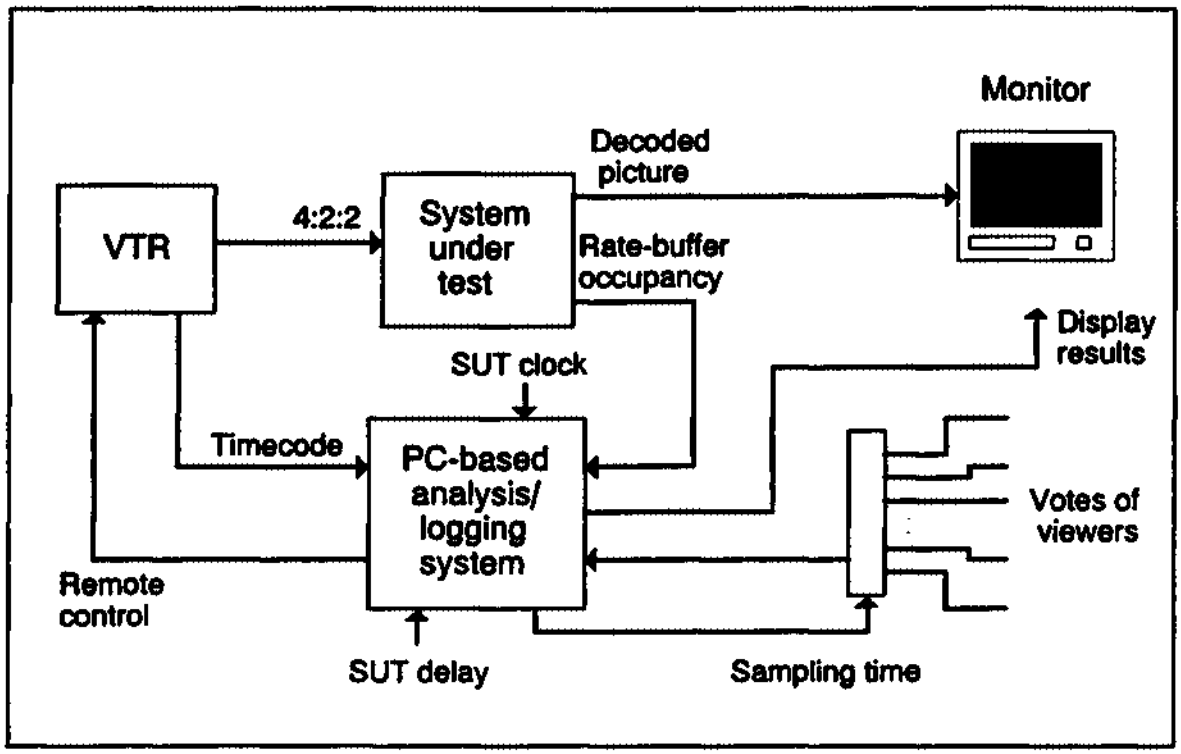


Fig. 2.5. System for the subjective and objective characterisation of picture quality in compression systems

to be carried out in the context of a practical studio environment where cascading of systems will certainly occur.

The use of a tool such as this is not only important for investigating and optimising system performance in the laboratory, but it also provides a convenient means of permitting the specification of the adequacy of systems for use in the studio. Already, digital editing systems, originally designed for off-line working, are being claimed by their manufacturers to be suitable for on-line application and to-date the quality compromise involved is not understood.

2.6 Subjective optimisation

We have looked at the direction which subjective assessment methods are taking to cope with commercial demands and advancing technology. So far we have not explored how a system designer can make use of the results from the assessments to optimise a design. It is difficult to generalise about this, since it is often the case that specialised approaches have to be designed to find optimal values of certain parameters - for many of these approaches the flexible arrangement of Fig. 2.5 will be both applicable and very efficient.

A requirement of most designers is to find test scenes which are demanding for their systems, so that they have convenient material on which to concentrate. In the case of digital compression, it is here that occasional violation of an assumption made in the approach of the previous section proves to be useful.

In Section 2.5, a rank-ordering of a large number of scenes was described according to the occupancy of the encoder rate-buffer, and this was assumed to be the same as the rank-ordering of the subjective quality of compressed scenes which would be experienced by a viewer. This ordered sequence was then 'calibrated' by some subjective tests. Referring to the model of Fig. 2.2, where the buffer fill-level directly controls distortion, this assumption must be true by definition provided that the encoder is perfectly subjectively optimised. Some scenes, however, will appear subjectively worse than they should do, when compared with the majority of other scenes having the same level of criticality (rate-buffer occupancy), and it is the worst of these scenes which provide a powerful pointer to the inadequacy of the subjective optimisation of a system. Poorly subjectively optimised scenes can easily appear worse than scenes which have a much higher statistical criticality measure. Remember too that there is much evidence that viewers in subjective assessment sessions judge the quality of a picture by its worse noticeable part, so local image characteristics which reveal deficiencies in subjective optimisation will have a larger impact than might be expected.

Another important point here is that poorly subjectively optimised scenes can appear more frequently and/or look worse than scenes which have a much higher statistical criticality measure. Simply searching for statistically busy scenes on which to optimise compression algorithms is not therefore the best strategy for improving performance.

Fig. 2.6 presents an illustrative plot of DSCQS subjective quality vs criticality, which show scenes which reveal poor subjective optimisation can be identified as those which excurse to the greatest extent from the average. Points obtained from the study described in Section 2.5 have been marked and reveal the scenes Renata and Mobile & Calendar to exhibit poor subjective optimisation. Interestingly, these scenes are well established in the folklore of compression as being 'difficult', but it has previously assumed that they are statistically demanding or critical, which is not especially the case. The method of plotting DSCQS subjective quality against criticality also provides a means of differentiating those source scenes which are noisy.

The arrangement of Fig. 2.5 can conveniently be used to identify scenes which are poorly subjectively optimised. In this case the tape would contain test sequences under evaluation, avoiding those with a high noise content. Once found, it is the job of the system designer to identify which characteristic of the image is embarrassing the system, and to devise a scheme to improve its reproduction. In image compression, subjective optimisation is concerned with achieving the most appropriate allocation of bit-capacity throughout the picture so that for a given rate-buffer occupancy, subjective quality is constant. Picture characteristics which are not handled well can usually be improved by deriving some function to detect them, and then by adaptively

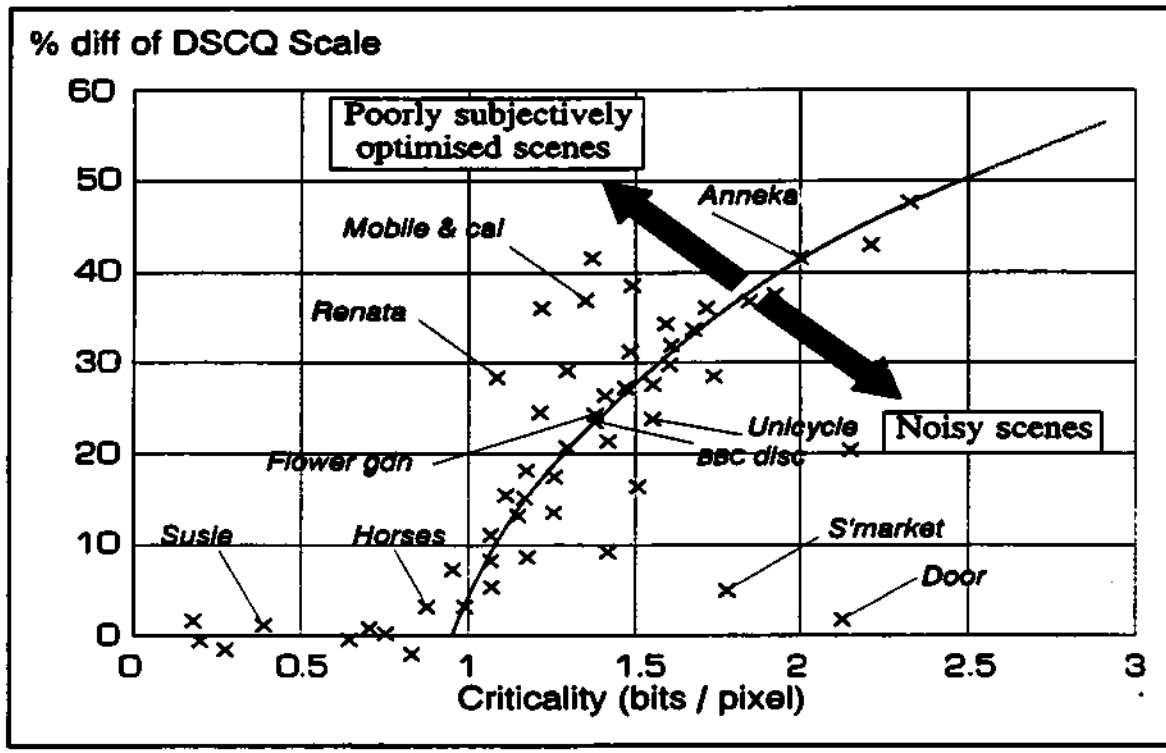


Fig. 2.6. Identification of scenes which reveal deficiencies in the subjective optimisation of compression systems

varying the distortion strategy (e.g. by selecting finer quantiser) in their locality.

2.7 Conclusions

This paper began by reviewing the general principles and the most common current methods for the subjective assessment of television pictures and also discussed their inadequacy to describe meaningfully the performance of advanced television systems. In particular it examined the important case of digitally compressed television, where quality will not be constant, but will to some extent be dependent upon the content of the scene being handled at a particular time. A method for characterising the subjective quality of such systems was presented and demonstrated using an implementation of the MPEG–1+ system. The commercial importance of understanding the viewer's tolerance to occasional distortion was also discussed, and approaches for investigating this using continuous voting procedures, were described. Finally, a general method was described for identifying those scenes which reveal deficiencies in the subjective optimisation of compression schemes. Targeting work on the optimisation of these scenes, through the careful design of bit-allocation processes, is considered to be particularly important for improving compression systems because

poorly optimised scenes are likely to occur more frequently in typical television, than scenes which exhibit high statistical criticality.

This introductory paper has set the scene for the more detailed methodological developments which are described by subsequent workshop papers. MOSAIC has built upon the principles of continuous voting procedures and, through necessary studies of human memory processes, has arrived at reliable methods for picture quality and impairment evaluation. These will be contributed to the World body of knowledge in these areas though the ITU-R.

References

[1] ITU Draft Rec. 500-6. Methodology for the Subjective Assessment of the Quality of Television Pictures. ITU Geneva, 1995.

[2] B.L. Jones and P.R. McManus. Graphic Scaling of Qualitative Terms. *SMPTE Journal*, November 1986, 1166–1171.

[3] CCIR Rec. 710. Subjective Assessment Methods for Image Quality in High Definition Television. ITU Geneva, XI, pt 1, 70–71.

[4] J.W. Allnatt. *Transmitted Picture Assessment.* ISBN 0 471 90113 X, Wiley, 1983.

[5] N.K. Lodge. The Picture Quality Implications of Low Bit-rate Terrestrial Television. *IEE Colloquium on Developments in Terrestrial Broadcasting for the UHF Band*, IEE Digest No. 1992/005, London, January 1992.

[6] N.K. Lodge. Low Bit-rate Video Compression Techniques. *Image Technology - Journal of the BKSTS*, December 1992, 219–223.

Chapter 3

Picture Quality and MPEG Coding Efficiency in HDTV and EDTV

Laurint Boch, Massimo Gunetti, and Mario Stroppiana
RAI

3.1 Introduction

If there were no constraints on transmitted bit-rate, it would be convenient to convey to the user the true high definition source resolution. Consequently, the home user could enjoy the studio quality and could display the pictures on a large screen.

Unfortunately, the bit-rate available on physical channels is limited, depending on the type of channel, i.e. satellite, terrestrial, ADSL, etc. Therefore, the television picture must be compressed and the compression algorithm introduces on the coded picture a coding noise proportional to the compression ratio. The overall quality of the coded picture is then influenced by the picture resolution and by the coding noise.

The picture format, vertical and horizontal resolutions, must match to the channel capacity available to deliver the picture. If the channel capacity is high, an HDTV picture can be delivered. The compression ratio is low and, consequently, the coding noise is low. If the channel capacity is limited, an HDTV picture requires a high compression ratio and then the increase of coding noise can nullify the high picture resolution and impair the overall picture quality. In this case, the overall picture quality could be improved by coding a lower resolution picture with a lower compression ratio [1].

The above discussion is general, and it does not give a precise knowledge of the matter. In fact, the terms "high" and "low" bit-rates and "high" and "poor" picture quality are vague. In order to define better the meaning of these terms it is necessary to have a precise relation between bit-rate and subjective quality of the involved picture formats and resolutions. To achieve this, some simulations and subjective evaluation of HDTV and EDTV pictures have been carried out.

3.2 Subjective assessments

The goal of the scheduled experiments was to compare the overall picture quality of HDTV and up-converted EDTV in function of the coding bit-rate.

The tests highlighted some interesting aspects not considered at the moment of the scheduling, i.e. effect of up-conversion on uncoded and DCT coded pictures, unforeseen effect of viewing distance on picture quality, etc. Consequently, some additional tests have been organised in order to clarify the unexpected results.

3.2.1 Test material

HDTV interlaced pictures with 1440 samples per line and 1152 active lines per frame and EDTV interlaced pictures with 720 samples per line and 576 active lines per frame have been simulated and subjectively assessed. Both HDTV and EDTV pictures have a 16:9 aspect ratio.

Four sequences have been selected: "Drums", a cut from the "Barcelona opening ceremony"; "Baruffa" and "Goal", two different cuts from "Good-bye to New York"; "Banquet", a cut from "Il giro della vita e della morte" which is an HDTV RAI production. The sequences are originally in HDTV format and the EDTV pictures have been obtained by down-conversion using the Snell & Wilcox down-converter.

Therefore two different picture formats with the same contents have been obtained.

3.2.2 Test description

The HDTV pictures have been simulated at 20, 15 and 10 Mbit/s, while the EDTV pictures have been simulated at 15, 10, 6 Mbit/s. The simulations have been performed according to the test model TM5 of MPEG-2. Using the DSCQS - Double Stimulus Continuous Quality Scale method, different assessments have been performed.

3.2.2.1 Comparison between HDTV and up-converted EDTV

The HDTV and up-converted EDTV pictures have been displayed on a 38" HDTV monitor with a viewing distance of $3h$ from the screen.

The picture quality of HDTV and up-converted EDTV have been evaluated by means of non-expert viewing sessions. The reference was the uncoded HDTV for both coded HDTV pictures and up-converted, coded and uncoded, EDTV pictures.

3.2.2.2 Influence of the up-conversion process

The up-converted EDTV coded pictures have been referenced to the up-converted uncoded EDTV pictures and displayed on a 38" HDTV monitor. The test was re-

peated in EDTV without up-conversion process. The EDTV coded pictures have been referenced to the EDTV uncoded pictures and displayed on a 20" 601 monitor.

The assessments have been performed at a distance of $3h$ from the respective screens. In such a way a different quality between the EDTV pictures and the up-converted ones could indicate an impairment provoked by the up-conversion process.

3.2.2.3 Evaluation of EDTV in dependence of viewing distance

The quality of EDTV coded pictures have been evaluated at $3h$, $4h$, $6h$ and $7h$ viewing distance from a 601 monitor. The reference was the EDTV uncoded picture. The assessment at $7h$ viewing distance has been included since that is the preferred viewing distance (PVD) with 20" screens [2].

3.3 Results

3.3.1 Comparison between HDTV and up-converted EDTV

The compression process and the up-conversion can impair the overall picture quality and the impairment is strongly dependent on the picture contents. Further, the compression algorithm influences more the HDTV picture which requires higher compression factors than the EDTV one, while the up-conversion process is applied only to the EDTV.

The best solution, to code and transmit HDTV pictures or EDTV pictures, is not only related to the bit rate available, but also to the picture characteristics, spatial resolution, entropy, motion, etc.

In addition, spatial resolution has a completely different role if it is connected with regular objects whose spatial distortion is highly perceived or with chaotic objects, as noise or people in sport events.

In the first case, the artifacts introduced by format conversion are very visible and the picture entropy can be limited, therefore the EDTV can be more affected by distortion than HDTV pictures, also at limited bit rates.

In the second case, the format conversion does not impair the pictures, while the compression is a critical process. In this case EDTV pictures present a quality higher than HDTV pictures also at high bit rates.

The above considerations are reflected on the results of the subjective assessment reported in Figure 3.1 at the viewing distance equal to $3h$. The results confirm the difficulty to fix a threshold indicating, for all the pictures, the bit-rate at which the HDTV and the up-converted EDTV present the same picture quality.

Taking into account the precision of the evaluation method, the threshold can be considered at about 10 Mbit/s for "Baruffa", and higher than 15 Mbit/s for "Goal", which is not perceptibly impaired by the up-conversion process.

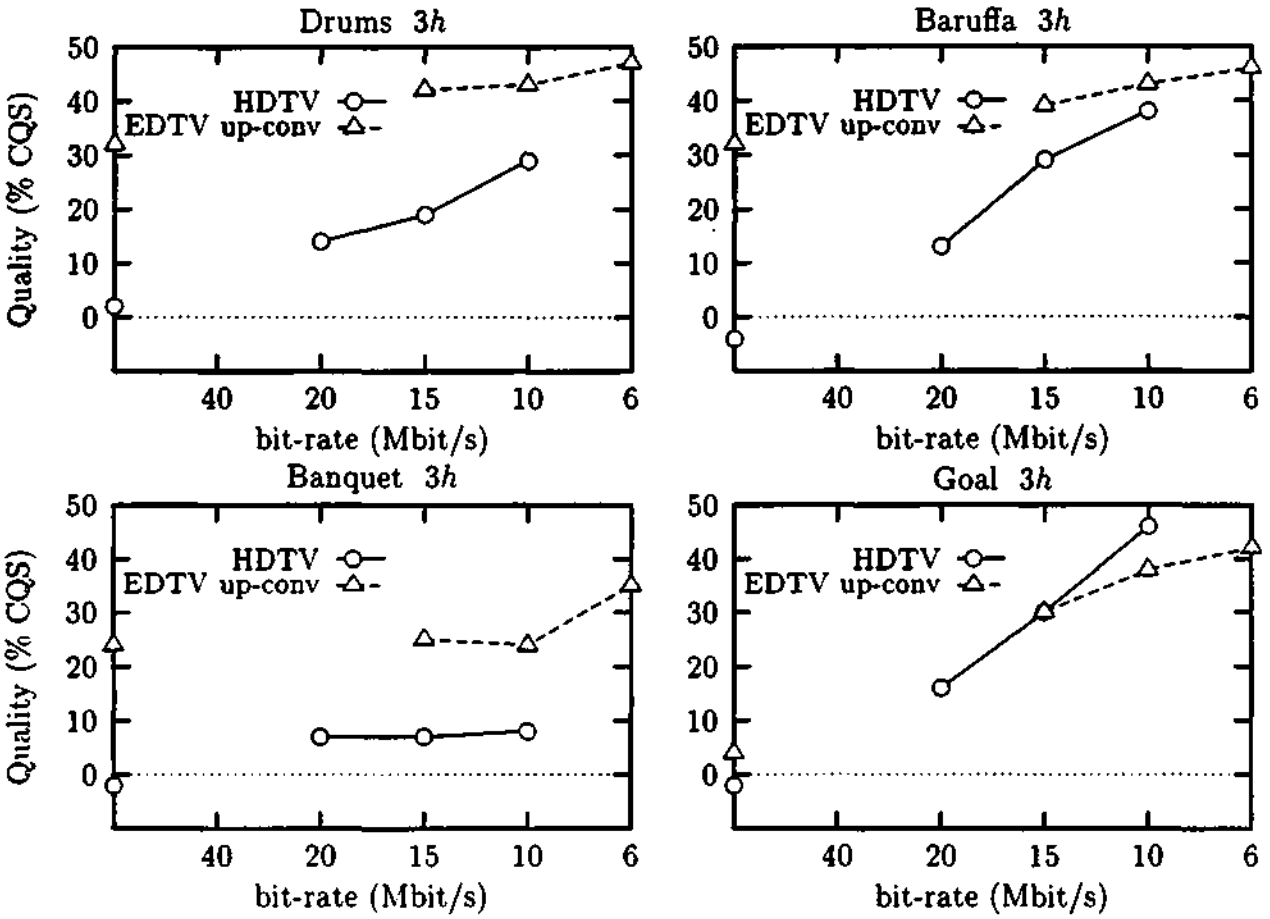


Fig. 3.1. HDTV and Up-converted EDTV. Subjective quality vs. bit-rate. Viewing distance $3h$ from a 38" HDTV monitor.

On the other hand, the sequences "Drums" and "Baruffa" are highly impaired by the conversion processes. The picture quality obtained coding them in the HDTV format is higher, also at low bit-rates, than that achieved by their down-conversion to EDTV format, coding and up-conversion to HDTV.

Considering that at least 4 or 5 Mbit/s are required by scaling factors and motion vectors in the MPEG-2 coding of HDTV pictures, and that the bit-rate for this information is not fixed but varies with the picture contents, it is not convenient to code HDTV sequences at bit rates lower than 8–10 Mbit/s.

3.3.2 Evaluation of EDTV in dependence of viewing distance

The comparison of picture quality of HDTV and up-converted EDTV at different bit-rates, reported in Section 3.3.1, present an unexpected behaviour in function of bit-rates. In fact, Figure 3.1 indicates an impairment of the EDTV picture coded at

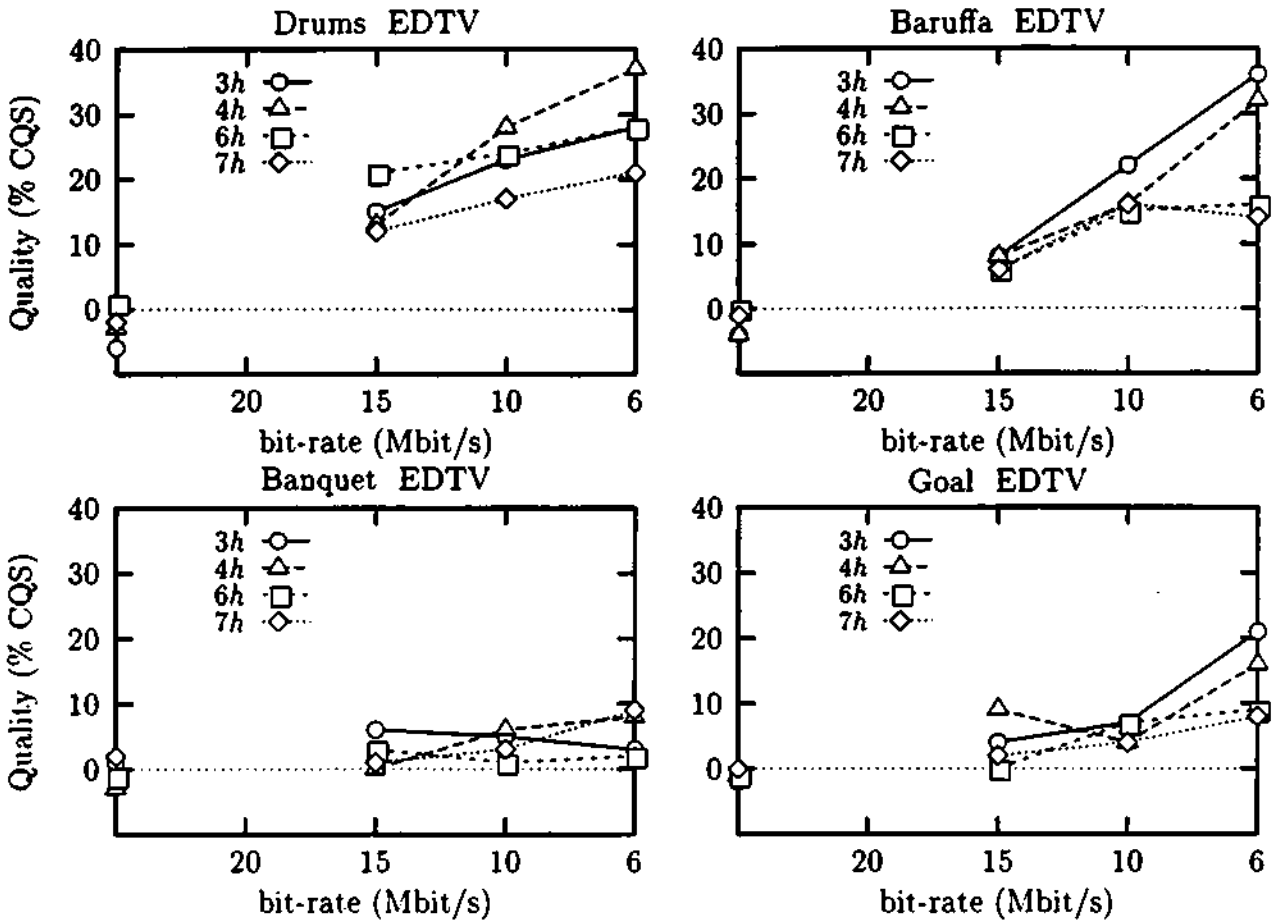


Fig. 3.2. EDTV subjective quality vs. bit-rate. For different viewing distances from a 20" 601 monitor.

10 and 15 Mbit/s and then up-converted with respect to the up-converted uncoded picture.

This situation has been considered abnormal because the available tests performed on MPEG-2 algorithms gave a high quality, subjective transparency, of sequences coded at 9 Mbit/s.

We suppose the discordance with MPEG-2 tests to be caused by the following reasons: different test material, different viewing distance, $3h$ instead of $6h$, or inclusion of the up-conversion process.

In order to verify the first two conjectures, a subjective evaluation of the EDTV pictures at different viewing distance, $3h$, $4h$, $6h$ and $7h$, from a 20" 601 monitor has been performed. The results, reported as picture quality versus bit-rate, are illustrated in Figure 3.2.

The sequences "Banquet" and "Goal" have been evaluated as high quality pictures at all the tested bit-rates, 15, 10 and 6 Mbit/s. In fact they have been scored less

than 12% of the continuous scale, except "Goal" at $3h$ and $4h$ and 6 Mbit/s that has been scored 21% and 16%, respectively.

Sequences, "Drums" and "Baruffa" present different characteristics.

The sequence "Drums" is very critical at all the viewing distances. In fact, it has been scored from 12% to 21% already at 15 Mbit/s and from 17% to 28% at 10 Mbit/s. At 6 Mbit/s, the worst quality is perceived at $4h$, 37% against 28% at $3h$ and $6h$. At 15 and 10 Mbit/s the picture quality is only slightly dependent on the viewing distance.

The above figures can be explained considering that coding noise causes a flicker on the picture area corresponding to the people inside the stadium. The flicker is already present on the sequence coded at 15 Mbit/s and its perception changes only perceptibly in dependence on the viewing distance and it is generally more annoying at the intermediate distances, $4h$ and $6h$, than at the short and great viewing distances, $\leq 3h$ and $\geq 7h$. The amount of the flicker increases by decreasing the bit-rate, but its perception increases only moderately.

On the contrary, at 6 Mbit/s, the sequence "Baruffa" has been quoted as good quality at $6h$ and $7h$, but only as fair quality at $3h$ and $4h$, 14% and 36%, respectively. The viewing distance has a strong effect on the picture quality evaluation, at least at low bit-rates. On the other hand, considering that the preferred viewing distance (PVD) is about $6h$ for the used 20" monitor [2], this sequence would be evaluated of good quality by the home user.

3.3.3 Influence of up-conversion

Up-converted EDTV pictures can be affected by some different types of impairment with respect to the pictures shot in HDTV:

- reduced spatial resolution which cannot be recovered by the up-conversion process,
- increased visibility of picture noise, e.g. coding noise, because its spectral density distribution is more concentrated at the low frequencies,
- malfunction of adaptive up-conversion algorithms caused by the noise present on the pictures.

In the assumption that the original pictures present a high S/N ratio, the loss of resolution affects both coded and uncoded pictures, and it is picture dependent. In the case of the adopted test material, the impairment caused on the uncoded pictures by the down and up-conversion processes ranges from 4%, "Goal", to 32%, "Drums" and "Baruffa", see Figure 3.1.

The last two items, in the previous list, are related only to the coded pictures. The impairment depends on the sensitivity of up-conversion algorithm to the noise and probably on the picture contents that can be more or less impaired by incorrect

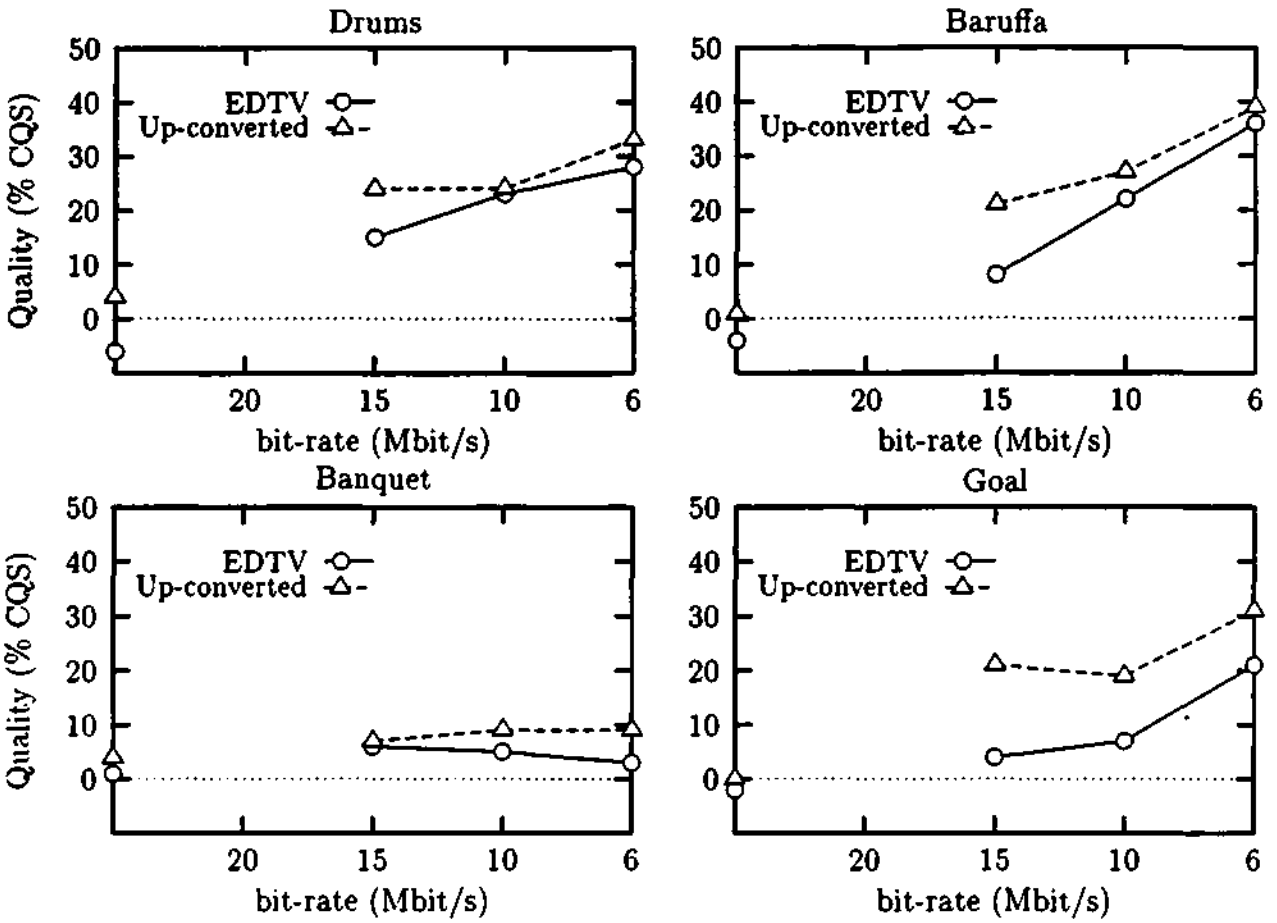


Fig. 3.3. EDTV and Up-converted EDTV. Subjective quality vs. bit-rate. The reference for Up-converted coded EDTV sequences is the Up-converted uncoded EDTV. EDTV: $3h$ viewing distance from a 20" 601 monitor. Up-c. EDTV: $3h$ viewing distance from a 38" HDTV monitor.

choices between spatial or temporal interpolation and by the variation of spectral distribution of the coding noise.

The impact of these effects on the picture quality versus bit-rate is shown in Figure 3.3.

Figure 3.3 graphs compare the EDTV picture quality, displayed on a 20" 601 monitor, with the up-converted EDTV picture quality, displayed on a HDTV monitor.

The up-converted EDTV coded pictures have been evaluated with respect to the up-converted EDTV uncoded pictures.

Up-conversion on coded pictures is not critical with "Banquet" and "Drums" sequences, whose impairment increases only of about 6-8% after the up-conversion.

A different sensibility to up-conversion is shown by "Baruffa" and "Goal", whose impairment varies from 8% to 21% and from 4% to 21%, respectively when coded at

15 Mbit/s.

This fact indicates that also a small amount of noise on the EDTV pictures can perceptibly decrease the quality after the up-conversion process of particular types of pictures.

3.4 Further evaluations

The comparison between HDTV and up-converted EDTV obtained by subjective assessments performed at a viewing distance of $3h$ from the screen would indicate that, when the picture is displayed on HDTV monitors, it is usually better to code HDTV sequences at bit-rates higher than 15-20 Mbit/s, and EDTV sequences at bit-rates lower than 7–8 Mbit/s.

At intermediate bit-rates the best format to adopt can be HDTV or EDTV depending on the sequence.

On the other hand, tests performed by the RAI - Research Centre inside the MOSAIC project indicate that the preferred viewing distance (PVD) varies with the screen size and it is about $6h$ for displays in the range from 28" to 38". These screens can be considered, with the present technology, the larger screens possible in the user homes. Therefore, the comparison between HDTV and up-converted EDTV has been carried out also at the preferred viewing distance of $6h$ and the results are reported in Figure 3.4.

The curves of picture quality versus bit-rate confirm the behaviour indicated by the curves of Figure 3.1, referring to a $3h$ viewing distance, but the absolute differences between HDTV and up-converted EDTV are very limited, less or equal to 10%, for all the sequences and bit-rates. The 10% value corresponds to less than half grade of the five grade quality scale of the ITU-R. Furthermore, in the same viewing conditions, the uncoded up-converted EDTV is scored less than one grade of the ITU-R scale with respect to the uncoded HDTV.

Consequently, a question could arise: is it convenient to deliver the home user with HDTV signal by means of limited capacity networks, since larger screens are unlikely to be available?

3.5 Conclusions

The advantage of HDTV studio quality against up-converted EDTV is generally high: one grade or more on the ITU-R quality scale, when the viewing distance is close to $3h$ (this value corresponds to the preferred viewing distance in the case of very large screens).

On the opposite, the advantage of HDTV against up-converted EDTV is reduced to less than half grade when the viewing distance is equal or higher than $6h$ which represents the preferred viewing distance with 38" or smaller screens.

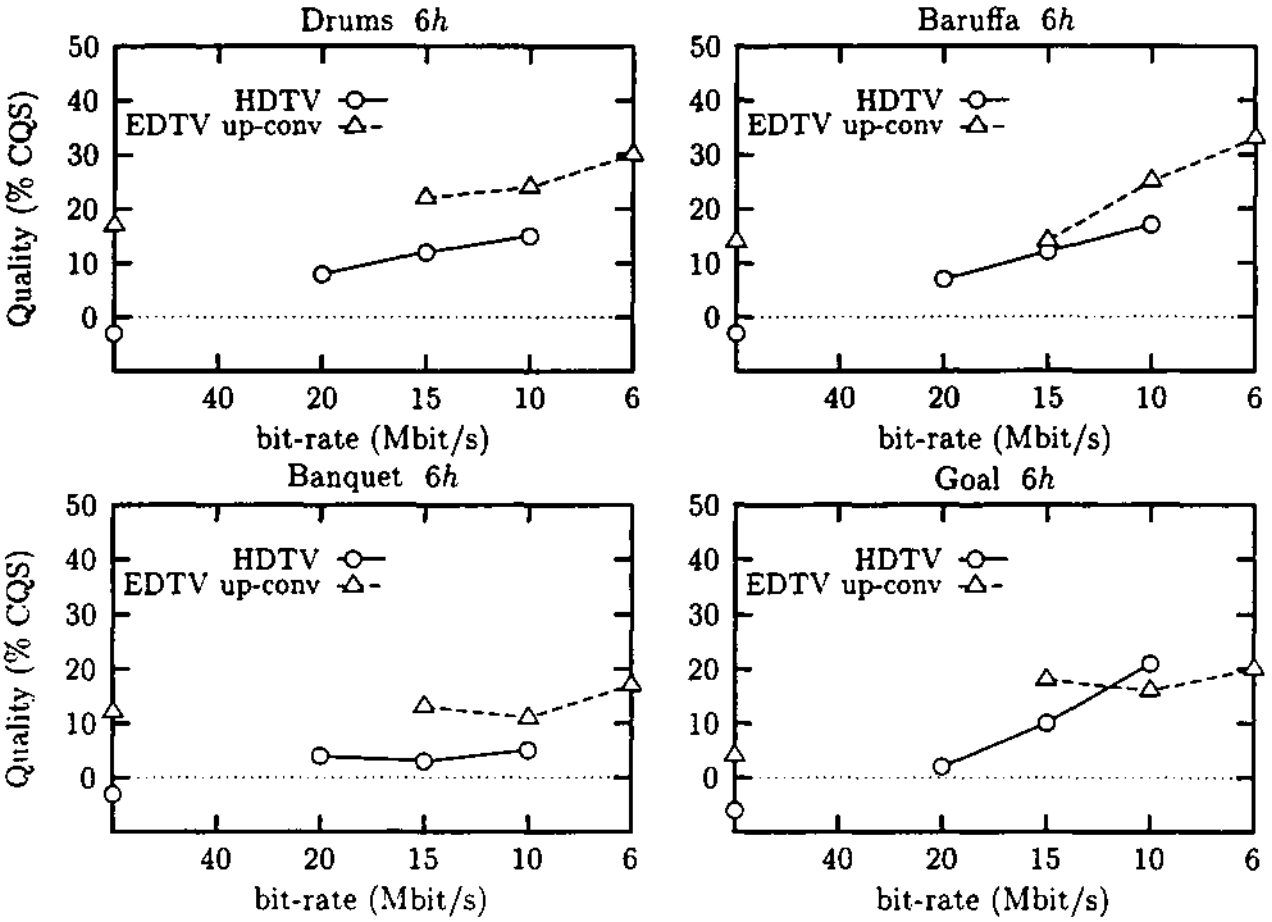


Fig. 3.4. HDTV and Up-converted EDTV. Subjective quality vs. bit-rate. Viewing distance $6h$ from a 38" HDTV monitor (PVD for a 28"/38" range monitor).

At bit-rates lower than 7–8 Mbit/s the overall quality of up-converted EDTV pictures is better than that available using HDTV pictures.

At bit-rates higher than 15–20 Mbit/s HDTV pictures still present a better overall quality than up-converted EDTV pictures.

Taking into account picture quality and costs, and considering that the 38" displays, with present technology, are the larger CRT screens available in the user-homes, it could be not completely evident the advantage to deliver HDTV signals to the home users using a transmission channel, whose capacity is not higher than 20 Mbit/s.

In addition, the EDTV sequences were obtained by down-conversion of HDTV pictures, so they can already present artifacts connected to the conversion process and their spectral density could extend to the high frequencies more than that of EDTV picture shot by EDTV cameras.

Consequently, the MPEG–2 coding of EDTV pictures down-converted by HDTV

ones seems to be more critical than EDTV native pictures.

The adopted up-conversion process selects spatial or temporal interpolator, adaptively on the basis of the motion detected on the picture. Coding noise present on the picture can modify the interpolator selection provoking artifacts on the picture. The artifacts are more or less visible depending on the picture contents.

More sophisticated up-conversion algorithms could reduce, in the future, the impairments connected to the process.

References

[1] M. Ardito. Wide-aspect ratios and production. *ITU/BR Workshop on Enhanced Television Tomorrow's Television,* The Wider Picture. Auckland, New Zealand, 3-5 October 1993.

[2] Investigation on the Preferred Viewing Distance for video images. MOSAIC RACE Project CEC No. R2111 - CEC Deliverable number R2111RAIDSR014.a1, December 1994.

Chapter 4

Subjective Quality Implications of Statistically Multiplexing MPEG–2 Coded TV Signals

Richard Aldridge
University of Essex

The dimensioning of channels to transmit video sources individually can be a significantly inefficient use of networks or spectrum. The use of statistical multiplexing of signals has therefore been seen as a way of increasing the efficiency of signal transmission. However, this brings the problem that on occasions, due to several signals requiring a relatively large capacity simultaneously, the multiplexer will be overloaded and will have to restrict the bit rate of the encoding sources or discard bits before transmission. In either case this will cause the subjective quality of the received signals to degrade, depending on the nature of the material and the type of error correction available in the coding process. The aims of this study, therefore, were to identify the gains that can be made by multiplexing digitally-coded video, and to quantify the likely degradation in subjective quality caused when the multiplexer is overloaded.

4.1 Coding of television extracts

It was decided to use a wide range of video material that may typically be offered by a broadcaster to its customers, and thus trying to ensure a range of scene criticality. It was also decided to code this using MPEG–2, since this has already been commercially accepted as the standard coding algorithm for digital broadcast television. The design of the multiplexer was made as simple as possible, with no source being given any prioritisation for transmission.

After coding each of the selected sequences, each of which was 30 s in duration, a range of statistics per sequence was calculated from which the salient data in Table 4.1 were derived. Each sequence has an equivalent mean bit rate which is shown in the

Sequence	Peak Frame (bits)	Standard Deviation (bits)	Peak-to-Mean Ratio	Mean Bit Rate 30 s Sequence (Mbit/s)
Adverts	191716	29700	3.77	1.27
Cartoon	217998	35345	3.22	1.69
Children's TV	219356	46949	2.67	2.06
Debate	257631	42455	3.49	1.84
Film	190418	38708	2.49	1.91
Football	285819	51219	2.48	2.88
News	310629	55124	4.04	1.92
Opera	330208	82643	3.24	2.55
Pop Music	218841	37584	2.09	2.61
Quiz	265206	48353	4.12	1.61
Soap	184989	35734	2.83	1.64
Weather	190687	39505	3.28	1.45
Total	2863498	–	–	23.41

Table 4.1. Main statistics for each of the MPEG-2 coded 30 s sequences

table. Also, the peak/mean ratio and standard deviation have been included to provide measures of spread of each sequence's statistical characteristics.

It is obvious from this data that if the twelve sources were to be transmitted independently on individual constant bit rate (CBR) channels with no degradation in quality , i.e. no discard of bits, and no time-smoothing buffer, then each channel would have to be dimensioned *at minimum* to accommodate the peak frames from each sequence. Thus the minimum overall capacity required for all twelve CBR channels can be determined from the total of the peak frame sizes, i.e. 2863498 bits, which equates to a channel with a capacity of approximately 71.59 Mbit/s for sequences running at 25 Hz, even though the overall mean bit rate for all twelve 30 s sequences was only about 23.41 Mbit/s.

One point to note here is that the nature of MPEG-2 is such that the coded I frames tend to be far larger in size than the P frames (except perhaps at scene cuts), which in turn are larger than B frames. In this case, with twelve coded 30 s (i.e. 750 frame sequences) the average sizes of frames were about 75 kbits, 35 kbits and 25 kbits for I, P and B frame-types respectively. Thus the I:P:B ratio of sizes for these particular twelve coded sequences was approximately 6:3:2.

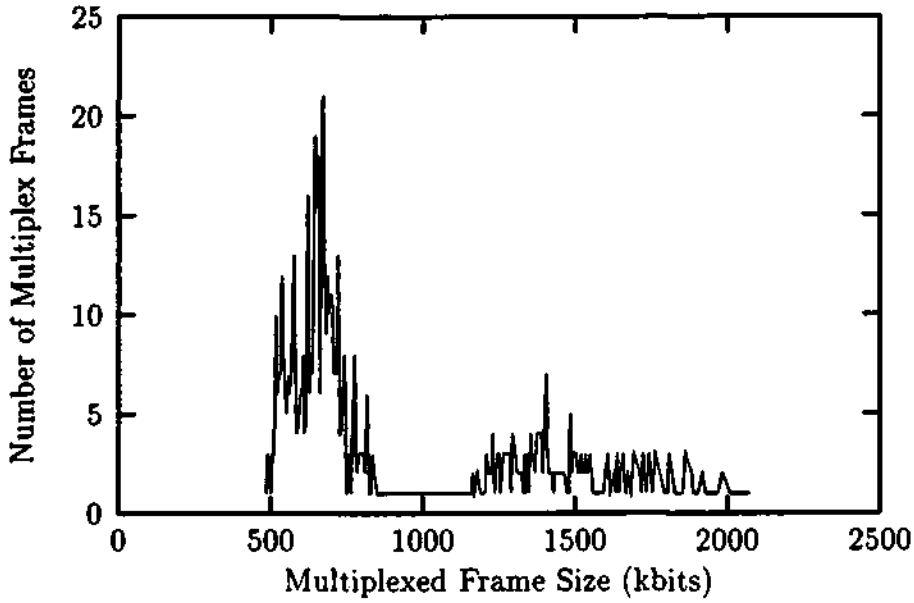


Fig. 4.1. Distribution of multiplexed frame sizes with I frames aligned

4.2 Statistical multiplexing of MPEG–2 coded video

The design of the multiplex was made as simple as possible, where it was assumed that the multiplexed frame unit was divided into slots, each accommodating a certain number of bits (e.g. one byte). Then, each coded video source presented a variable amount of data to the multiplex. Whilst the multiplex was limited by the number of slots which could accommodate data from all sources in one multiplexed frame time interval, it could allocate sources a variable number of slots depending on the amount of data offered by each source. Also, data encoded in the same time interval, in this case a TV frame slot of 40 ms, were multiplexed together in the single resultant data stream. When the multiplex channel capacity was exceeded, in practice feedback from the multiplex to the encoders would increase the quantiser step size. This coarser quantisation would introduce block effects into the picture. However, for simplicity no such feedback was employed here and overflowed bits were just discarded from each source proportionately.

As highlighted earlier, MPEG–2 video coding generally results in I frames which are far larger than P frames that are in turn larger than B frames. If all the coded video sources are presented to the multiplex such that the I frames from each sequence arrive at the multiplex simultaneously, the result is that extremely large number of bits occur at the multiplex at the start of every group of pictures (GOP). This will require a high multiplex channel capacity if no loss is to occur. When simulating the

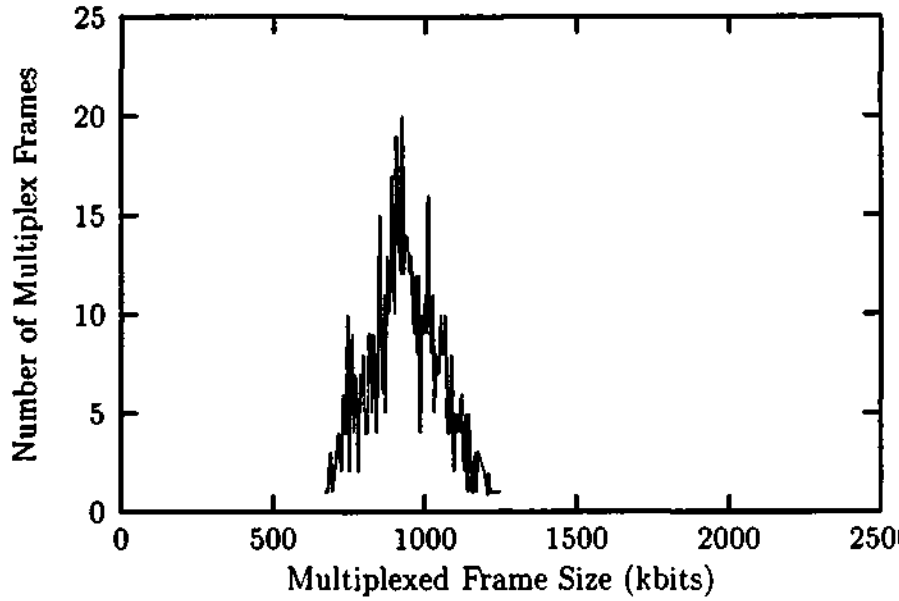


Fig. 4.2. Distribution of multiplexed frame sizes with uniform distribution of I frames between sequences over time

multiplex of the twelve recorded sequences with the I frames aligned in this manner, the distribution of required multiplex frame size in Figure 4.1 was obtained. This shows that the number of bits the multiplex is expected to accommodate in any frame interval could be as large as approximately 2 Mbits, which corresponds to a required channel capacity of 50 Mbit/s (at 25 Hz). The standard deviation of this particular distribution is over 400 kbits. Note that this value is still considerably smaller than the 71.6 Mbit/s required for individual transmission (by a factor of 71.6/50 = 1.43). This is not only because the largest I frames in each of the twelve sequences do not occur simultaneously but also because P frames occasionally need a larger number of bits than I frames. This sometimes occurs due to scene changes.

On average the most likely case is that the I frames are uniformly distributed in time between all the sources and therefore the same number arrive at the multiplex in any frame interval. This was simulated with the twelve coded sequences, and the histogram in Figure 4.2 was obtained. It shows that with this multiplexing method the largest number of bits arriving at the multiplex in any one frame interval was about 1.25 Mbits, which corresponded to a capacity of approximately 31.3 Mbit/s. The distribution itself has a standard deviation of just over 100 kbits in this case. Thus a multiplex channel dimensioned to this figure would transmit all twelve coded sources with no loss of information, yet require less than half the bandwidth which would have been required had they been transmitted on individual channels (31.3 Mbit/s compared to 71.6 Mbit/s, a factor of 71.6/31.3 = 2.3).

When simulating the multiplex of coded sources whose I frames arrive at the

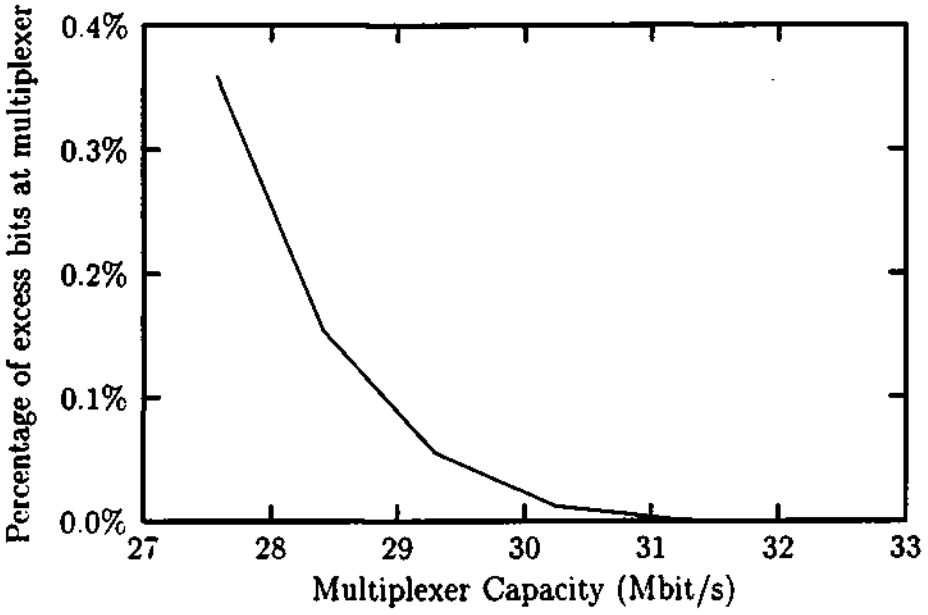


Fig. 4.3. Multiplex overload characteristic

multiplex uniformly distributed over time, the order of sequences will affect the minimum channel capacity required. In our experiments this uniform distribution was achieved by starting the sequences at different points in time. To verify that the order chosen was representative, a sample of 5000 random orders was simulated and it was found that the average minimum required multiplex capacity was 32.3 Mbit/s. Furthermore, 99% of this sample produced minimum multiplex capacities below 35.2 Mbit/s. Hence it would be possible to dimension the multiplex channel to this latter capacity and be virtually certain that overload, and hence picture distortion, would never occur. Note that this is still about half the capacity required for transmitting the sources individually (i.e. 71.6 Mbit/s compared to 35.2 Mbit/s, a factor of $71.6/35.2 = 2.03$).

4.3 Subjective quality implications of multiplexer overload

In practice it is never statistically possible to eliminate overload of the multiplexer, only to minimise the chances with some degree of certainty. As described earlier, when overload occurs picture distortion will be introduced. In practice this will be due to the multiplexer controlling the quantiser step sizes in the coders of the individual video sources.

With the twelve coded sequences used here, the overload characteristic was as shown in Figure 4.3 for the particular order chosen (which was arranged in the op-

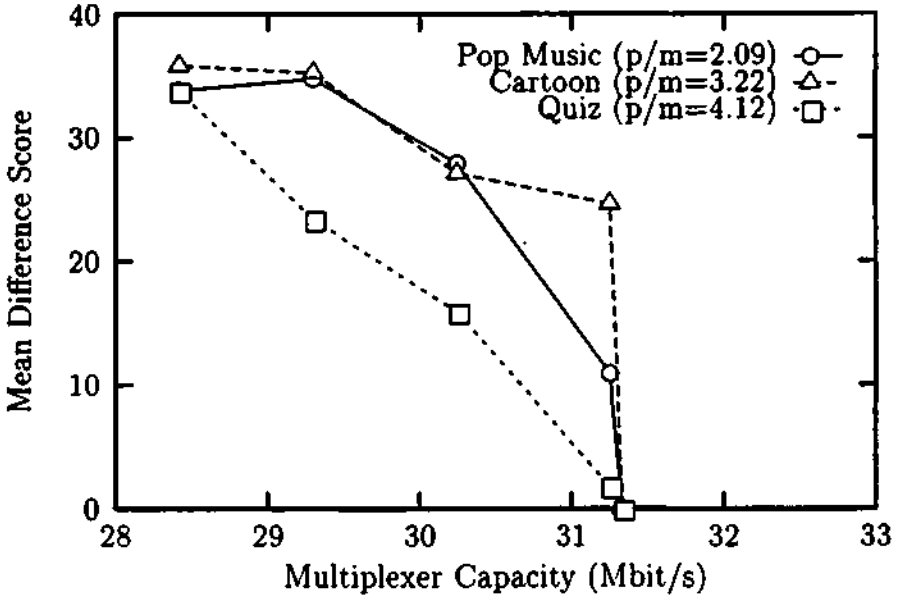


Fig. 4.4. Mean difference score vs. multiplex capacity for the three programme types

timum manner of ensuring all I frames arrive at different points in time).

In order to try and characterise the subjective quality effects of multiplexer overload, excess bits were simply discarded prior to decoding. Hence the results can be regarded as a lower bound to the subjective quality characteristic. Three sequences were selected for subjective quality evaluation under such circumstances, namely *Pop Music*, *Quiz* and *Cartoon*, since they possessed the maximum, minimum and median peak-to-mean ratios (refer to Table 4.1). The results shown in Figures 4.4 and 4.5 were obtained. They show a rapid decline of subjective quality for even the smallest (about 0.1%) discard of excess data. In practice this reduction would not be so severe because the encoders would simply have their bit rates restricted and the decoders would still be able to track the encoder. It must be remembered, however, that an MPEG-2 codec with no error correction or detection was used and no prioritisation of discard was performed by the multiplexer. Nevertheless, these provisional results suggest that it is unlikely that broadcasters would wish to utilise statistical multiplexing much after the point where the multiplexer becomes overloaded. However, it has been shown earlier how the multiplexer may be optimised for MPEG-2 coded video and how channels may be dimensioned with a degree of certainty that virtually prevents any overload from occurring.

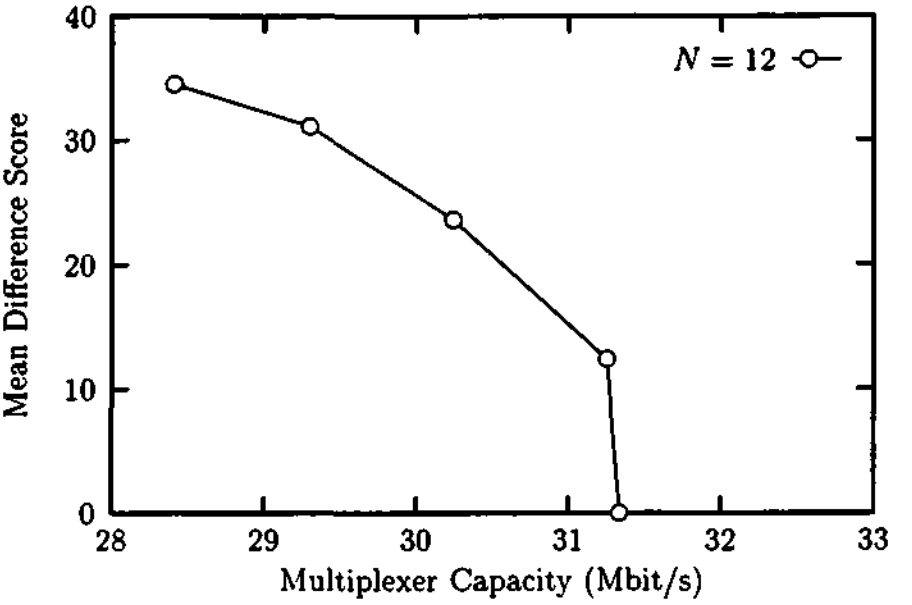


Fig. 4.5. Average mean difference score vs. multiplex capacity

4.4 Exploiting the structure of MPEG–2 for statistical multiplexing

Within each MPEG–2 GOP structure, there are three distinct frame types - I, B and P. The I frames are purely intra-coded and rely on no other frames in order to be reconstructed at the receiver. P frames are predictively-coded from either the previous I or P frames. Predictions for B frames are made from the I or P frames, or from their interpolated values between I and/or P frames. B frames are never used for prediction themselves. Consequently if during transmission a B frame of a sequence is subject to multiplex overload, then no other frames in the sequence will be affected. However, if a P frame is subject to overload it will be affected, as will all frames at the receiver reliant on it for information until the next I frame occurs. When an I frame is subjected to overload, the entire GOP will be affected, as will any B frames between the previous GOP's last P frame and this affected I frame.

A subjective test was carried out to find the effect of overload on different types of MPEG–2 frames. In order to try and conceal such errors, a basic error concealment technique was also tried, where corrupted frames were replaced by the last unimpaired decoded frame (i.e. 'freezing' frames). Results in Figure 4.6 show that distortion in B frames causes the least reduction in subjective quality, as expected. Also, of the two sequences tested, distortion was less visible in the slower moving video sequence. These results therefore suggest that if overload of multiplexing MPEG–2 video ever does occur, then priority should be given to sources whose I and P frames are being

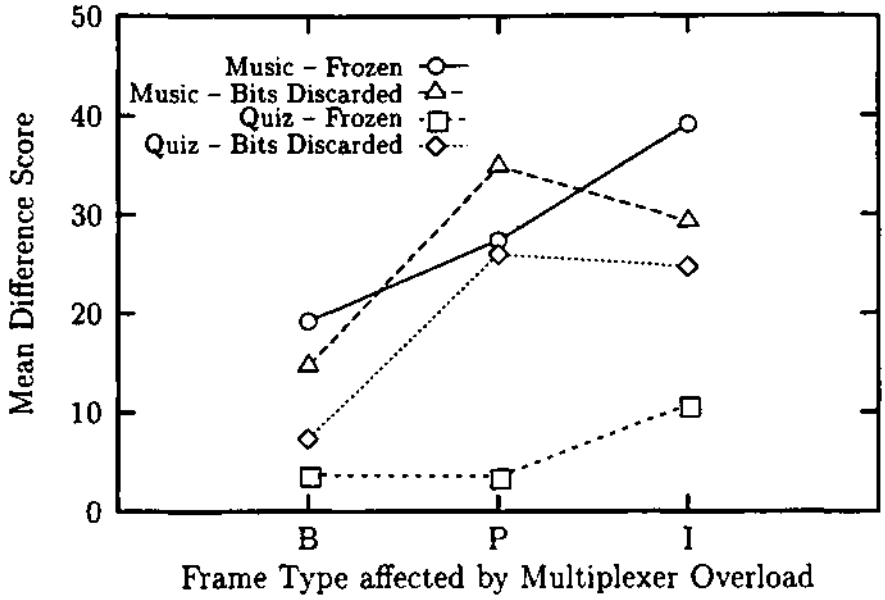


Fig. 4.6. Graph showing effect of overload on different MPEG-2 frame types

multiplexed, rather than those with B frames at the multiplexer. In this way, combined with ensuring a uniform arrival of I frames over time and correct dimensioning of the multiplexed channel, it should be virtually certain that overload never occurs but if it does only a minor reduction in subjective quality occurs. This can all be achieved with a substantial saving in bandwidth.

4.5 Conclusions

It has been shown how statistical multiplexing of MPEG-2 coded video can create significant capacity savings. In the investigation undertaken with twelve assorted MPEG-2 coded television sequences, it was shown that the multiplex capacity needed to be only 35.2 Mbit/s to be 99% certain that no overload occurred and for picture quality to remain constant. This is just less than half the minimum bandwidth (71.6 Mbit/s) that would have been required if the channels had been transmitted individually (a factor of 2.03 = 71.6/35.2). This would be commercially useful to broadcasters where bandwidth limitations exist, e.g. satellite or UHF terrestrial broadcasts, video-on-demand services along telephone cables, etc. However, it must be noted that to optimise these benefits for MPEG-2 coded video, the occurrence of I frames from the different sources should be uniformly distributed over time.

Cases of multiplex overload of these VBR transmissions were examined and found to have a dramatic effect on subjective quality, although it should be noted that the simulation created distortions which would be more severe than is likely to be the

case in practice. This severity of distortion occurs because the decoder cannot track the encoder, i.e. the encoder assumes information has been received by the remote decoder when it has not been. This is in contrast to what would happen with CBR transmission, where the restriction on bandwidth would cause both the encoder and decoder to raise the quantiser step size; although picture quality would be reduced, this degradation would be less severe and would degrade gracefully as the quantiser step size increases since the decoder would still be able to track the encoder.

As well as minimising the occurrence of multiplex overload by ensuring that the MPEG–2 coded video sources uniformly distribute the I frames at the multiplex, it has also been shown how the reduction in subjective quality on the remaining (rare) occasions of overload can be alleviated by the employment of a simple algorithm at the multiplex. Specifically it was demonstrated that at overload it would be advantageous to give priority to I and P frames, and to introduce distortion to B frames only, in order to control the bit rate. The loss of a complete B frame was shown to be better subjectively than the introduction of distortion to an I or P frame. Subjective quality reductions would be restricted to less than one point on the five-point quality scale using such a method. With the addition of a simple error concealment method where frames are frozen to conceal picture distortions, this was demonstrated as having an even greater effect in reducing the impact on subjective quality.

Chapter 5

Recency Effect in the Subjective Assessment of Digitally-coded Television Pictures

Richard Aldridge, Jules Davidoff, Mohammad Ghanbari, David Hands, and Don Pearson
University of Essex

5.1 Introduction

In MPEG-2 coding of television pictures, buffer overflow occurs during motion-intensive scenes; this introduces time-varying levels of coding distortion into the video. In this paper we report on the subjective effect of such distortion on viewers. It is known that standard ITU-R subjective testing methodology [1], which uses 10 s presentation times, is not easily able to deal with such variations in quality [2], [3]. Most of the applications of this methodology have been to situations where there is a single level of quality associated with a fixed set of engineering transmission parameters. However, if MPEG video is coded at a constant bit rate which is low enough to cause visible distortion, the quality of successive 10 s sections can be quite different from each other. At present, there is little reported experimental work concerning such quality variations in digitally-compressed television, although a study has been made of the time-varying distortions that occur in ATM video transmission due to sporadic network overload [4]. This study confirmed that inadequacy of 10 s test sequences for subjective assessment and, in some follow-up tests with longer sequences, quantified the "forgiveness effect" i.e. the preparedness of viewers to forgive or discount the importance of transitory impairments that are embedded in unimpaired video [5].

5.2 Experimental method

To elucidate the effects of time-varying distortion on subjects, test sequences of 30 s were employed. With these longer sequences, it is possible for there to be considerable variation in quality during the presentation of the sequence; relatively good quality video is typically interspersed with peaks of quite visible impairment [6]. We were interested in how subjects formed an overall quality judgement after viewing such a sequence; in particular, we wanted to know how they were influenced by the form of the quality variations over time, including the location of the impairment peaks. We therefore arranged for the peaks to occur either towards the beginning or towards the end of test sequences. Quality was measured using the Double Stimulus Continuous Quality Scale (DSCQS) [1]; in this method, quality is expressed as a Mean Difference Score or MDS (on a 100-point scale) between ratings for the test sequence and for an unimpaired reference sequence. The ratings themselves were recorded using the ITU-R 5-point quality scale (Excellent, Good, Fair, Bad, Poor), so that an MDS of 100 corresponded to the test stimulus being rated at one end of this scale and the reference at the other. The investigations were conducted as part of RACE project MOSAIC (Methods for Optimisation and Subjective Assessment in Image Communication), in which the University of Essex is involved. This project is examining a number of different ways in which current subjective testing methodology can be improved; one such method is described in this paper.

5.3 Measurement of quality variations

Initially two extracts (*100 m Sprint* and *Vending Machine*) from a drama programme called "Exam Conditions" (courtesy of Central TV) were used in the tests. Each sequence had a total duration of 50 s and was coded at two bit rates: 1 and 4 Mbit/s. The 50 s sequences were chosen such that each contained a 10 s central sections in which the distortion was generally less visible. They were then each divided into five 10 s sections, labelled A–E. The quality of each section was measured subjectively when presented in isolation from the other material i.e. as a 10 s stand-alone sequence. 24 non-expert subjects participated.

5.4 Selection of 30 s sequences

Two 30 s sequences, ABC and CDE, were then constructed from each 50 s sequence. This gave two sets of 30 s presentations in which the worst section of impairment tended to occur either at the beginning or at the end. Subjects viewed the whole of the 30 s sequence and were then asked to rate the overall quality. 29 non-expert subjects participated in this experiment.

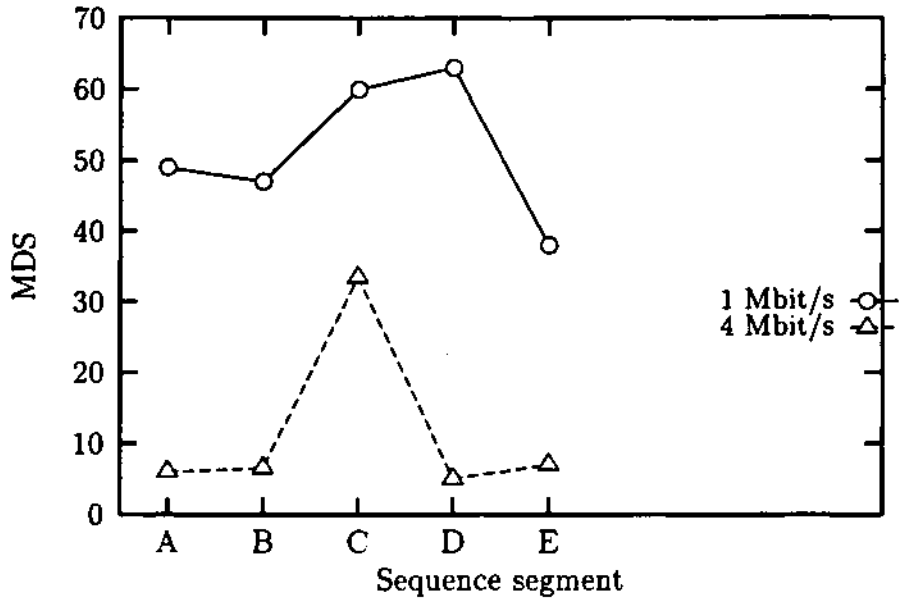


Fig. 5.1. *100 m Sprint:* Mean Difference Scores for the 10 s sections A–E seen alone.

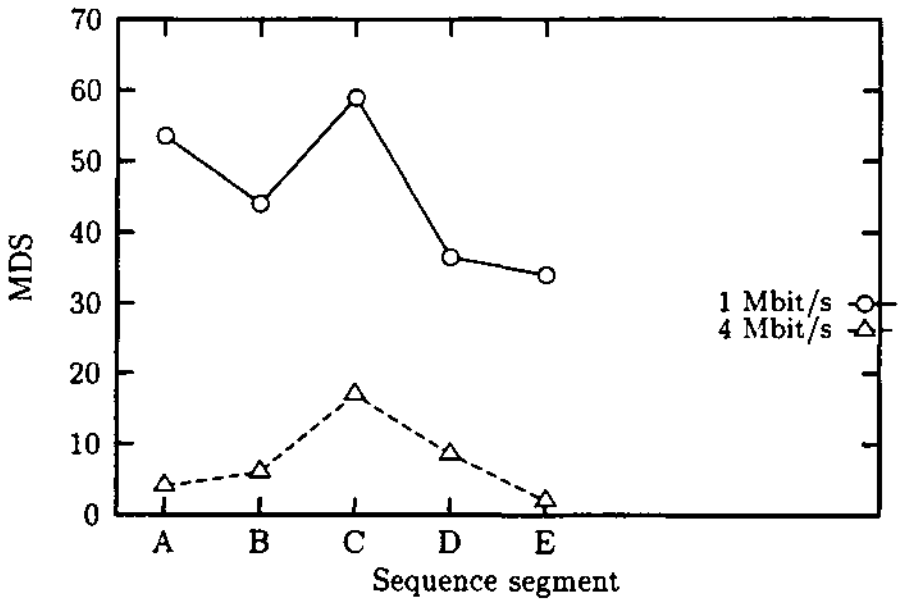


Fig. 5.2. *Vending Machine*: Mean Difference Scores for the 10 s sections A–E seen alone.

5.5 Results

5.5.1 Quality Variations

Figs. 5.1 and 5.2 show the results for the case where subjects viewed the individual 10 s sections A, B, C, D and E separately. It can be seen that at 4 Mbit/s there is

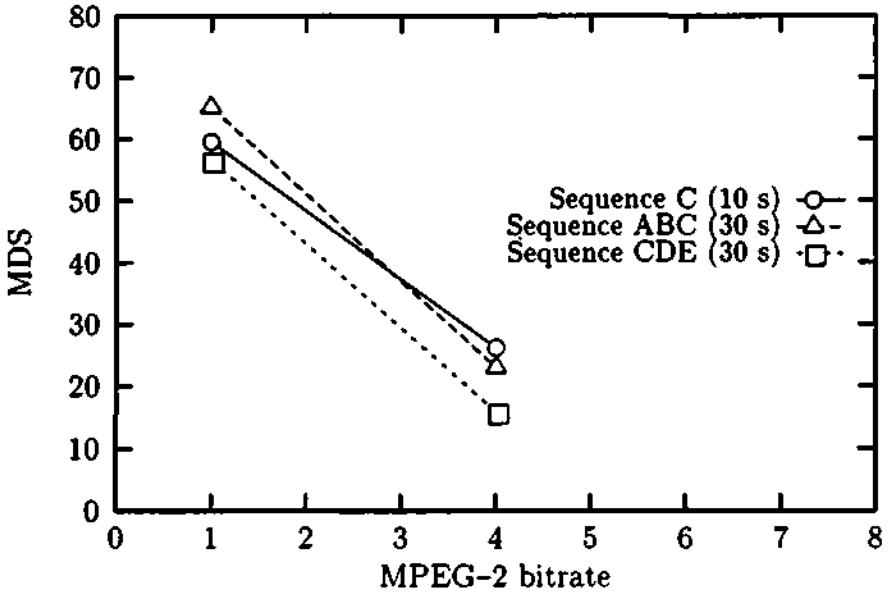


Fig. 5.3. *100 m Sprint:* Mean Difference Scores when the 10 s sections were either seen alone (C), or preceded (ABC) or followed (CDE) by relatively less impaired sections to form a 30 s sequence.

a clear peak in the impairment at the central section C for both scenes; at 1 Mbit/s this peak is less distinct, although the C section is still among the worst in terms of quality. These figures illustrate the fact that temporal variations of quality in coded video are different at different bit rates; if a particular section is chosen as the poorest-quality at one bit rate, it may not be true at another bit rate. Thus the experimental effects associated with seeing the poorest-quality section at the beginning or end of a 30 s presentation should be most evident in the 4 Mbit/s results. 4 Mbit/s is also a more realistic bit rate in terms of the overall quality of transmission at which broadcasters might aim; at 1 Mbit/s the coding distortion is at an unacceptably high level throughout the sequence.

5.5.2 Influence of preceding and following sections of video

Figs. 5.3 and 5.4 show the Mean Difference Scores when the 10 s sections were either seen alone (C), or preceded (ABC) or followed (CDE) by relatively less impaired sections to form a 30 s sequence. When the 10 s C sequence was placed at the end of a 30 s sequence (ABC), ratings were more similar to those for section C when seen alone. The magnitude of the rating difference between seeing the impairment at the start (CDE) and at the end (ABC) is about 10 points on the 100-point MDS scale, or half a scale point on the five-point quality scale.

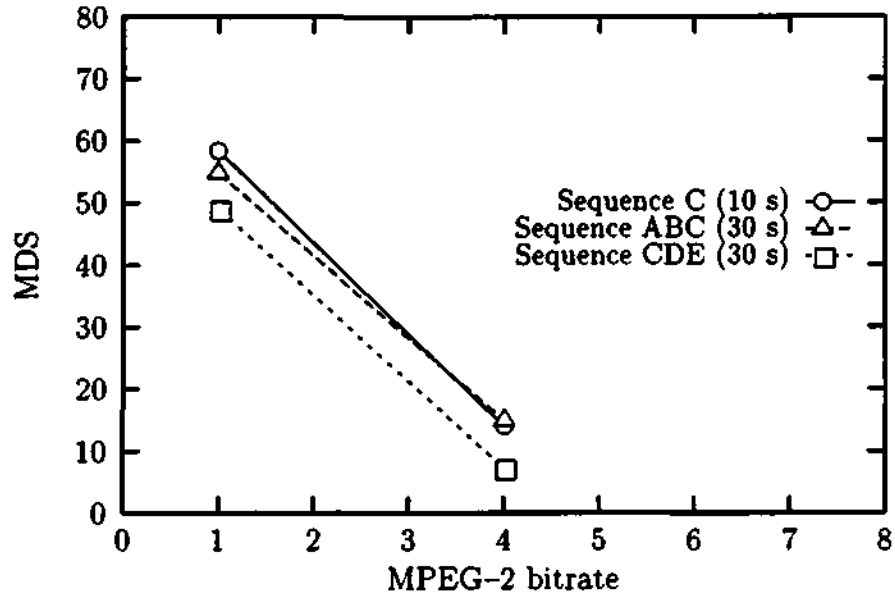


Fig. 5.4. *Vending Machine*: Mean Difference Scores when the 10 s sections were either seen alone (C), or preceded (ABC) or followed (CDE) by relatively less impaired sections to form a 30 s sequence.

5.5.3 Analysis of variance

A 2×2×3 analysis of variance was performed on the data, the factors being scene (*100 m Sprint* and *Vending Machine*) × bit rate (1 Mbit/s and 4 Mbit/s) × sequence (C, ABC and CDE). This found the three main effects of scene, bit rate and sequence all to be highly significant. The effect of bit rate was significant at $p < 0.01$; 4 Mbit/s coded images are clearly much superior in quality to those coded at 1 Mbit/s. *100 m Sprint* was rated significantly worse ($p < 0.01$) than *Vending Machine*. This was attributed to the fact that *100 m Sprint* consisted of higher motion than *Vending Machine* and possessed more scene changes. The effect of sequence type was also significant at $p < 0.01$. In order to investigate the nature of this effect further, a post-hoc analysis using Scheffé test was performed. This showed that there was no significant difference between the 30 s ABC ratings and the 10 s C ratings; however, there was a significant difference between the CDE ratings and the C ratings, as well as between CDE and ABC, both at the $p < 0.01$ level of significance. As in previous experiments [5], when an impaired section of video was followed by a section of relatively high-quality video, subjects tended to "forgive" the bad section by a kind of averaging of quality over the period of the sequence. However, where good-quality video precedes poor-quality video, subjects appear to rate the sequence as a whole on the basis of the poor-quality end section alone, apparently with little regard to the earlier material.

5.6 Discussion

The results indicate that subjects, if asked to give a single quality rating after viewing an impaired video sequence, are strongly influenced by what they see in the last section of the sequence. If this section is of good quality, it tends to raise the rating; if of poor quality, to depress it. It is as if subjects form a weighted temporal average of the instantaneous quality variations that occur during the sequence, maximum weighting being given to the most recently-seen part of the sequence. Impairments occurring 20–30 s prior to the end of the sequence seem to count for little or nothing in the weighting. This *recency* phenomenon appears to fit in with what is known about the length of human working memory. There is evidence to suggest that working memory has a duration of about 20 s and that the rate of decay in working memory is dependent on the amount of information presented, as it has a limited capacity [7]–[9]. Both of these facets of memory can be seen as important in the results, in that the end of sequences are more accessible to memory recall (the recency effect) and may bias the subjects' overall rating. If we apply the working memory model to the present study, then information presented most recently will have a strong memory trace which is readily accessible to memory. Earlier information not only decays naturally in working memory, but in addition is interfered with by subsequent information, leading to a reduction in its availability when an assessment is made.

An interesting point to emerge from this study is that subjects treat spatially-varying and temporally-varying image distortions in different ways. With spatially-varying distortion, there is a tendency to identify and rate the worst-quality sub-area of the image [10]. However, with temporally-varying distortion, subjects do not specifically remember and rate the worst-occurring portion of the video over time; rather they tend to rate that which they have seen most recently.

Because substantial temporal variations of quality can occur in MPEG–2 coded television pictures, and because subjects appear to form a weighted average of this quality over time, the use of standard-length 10 s test sequences poses problems for the interpretation and application of results obtained using them. By employing sequences that are even longer than 30 s, the experimental situation in the laboratory could be brought more into line with viewing situations in the home. This may, however, necessitate the use of single stimulus rather than double-stimulus methodology, since with longer sequences the overall duration of the test may become abnormally long. In addition, the comparative aspect of test and reference sequences, seen in temporal juxtaposition, may be weakened.

This study not only provides evidence for memory mechanisms affecting subjects' judgements, it also draws attention to the need for new experimental methodology which can overcome these cognitive processes.

5.7 Conclusions

Subjects were presented with 30 s television test sequences coded using MPEG-2 into 1 and 4 Mbit/s. The quality of each 10 s section of the 30 s sequences, as well as the overall quality, was assessed using the ITU-R DSCQS method. Substantial variations in quality were observed between the 10 s sections. In forming an overall judgement of the quality of a 30 s sequence, subjects were strongly influenced by what they saw in the last section of the sequence. We concluded that they form a weighted average of the instantaneous quality variations that occur during the sequence, maximum weighting being given to the most recently-seen part of the sequence. This *recency effect* appears to correspond with what is known about the length of human working memory. The magnitude of the rating difference between seeing the impairment at the start and at the end of a 30 s sequence is roughly 10 points on a 100-point MDS scale, equivalent to half a scale point on a five-point ITU-R quality scale.

5.8 Acknowledgements

The authors would like to thank NTL for providing the MPEG-2 coded material and the Independent Television Commission for their constructive assistance and guidance with the overall direction of the project.

References

[1] ITU-R Recommendation 500-5. Method for the Subjective Assessment of the Quality of Television Pictures. September 1992.

[2] N. Lodge. Picture Quality Implications of Low Bit-rate Terrestrial Television. *IEE Colloquium on Terrestrial Television for the UHF Band*, London, January 1992.

[3] N. Lodge and D. Wood. Subjectively Optimising Low Bit-rate Television. Proc. International Broadcasting Convention IBC '94, Amsterdam, September 1994.

[4] C.J. Hughes, M. Ghanbari, D.E. Pearson, V. Seferidis, and J. Xiong. Modelling and Subjective Assessment of Cell Discard in ATM Video. *IEEE Transactions on Image Processing*, Vol. 2, No. 2, 212–222, April 1993.

[5] V. Seferidis, M. Ghanbari, and D. Pearson. Forgiveness Effect in Subjective Assessment of Packet Video. *Electronic Letters*, Vol. 28, No. 21, 2013–2014, October 1993.

[6] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson. Measurement of scene-dependent quality variations in digitally-coded television pictures. *IEE Proceedings on Vision and Signal Processing*, Vol. 142, No. 3, 149–154, 1995.

[7] L.R. Peterson and M.J. Peterson. Short-term memory retention of individual items. *J. of Experimental Psychology*, Vol. 58, 193–198, 1959.

[8] C.D. Wickens. *Engineering Psychology and Human Performance.* Harper Collins, New York, 1991.

[9] A.D. Baddeley. *Human Memory: Theory and Practice.* LEA, Hove, East Sussex, 1990.

[10] J.O. Limb. Distortion Criteria of the Human Viewer. *IEEE Trans. on Systems, Man, and Cybernetics,* Vol. SMC–9, No. 12, 778–793, 1979.

Chapter 6

Continuous Assessment of Perceptual Image Quality

Roelof Hamberg and Huib de Ridder
IPO

Abstract

The present study addresses the question whether subjects are able to assess the perceived quality of an image sequence continuously. To this end, a new method for assessing time-varying perceptual image quality is presented. In this method, subjects continuously indicate the perceived strength of image quality by moving a slider along a graphical scale. The slider's position on this scale is sampled regularly, e.g. every second. In this way, temporal variations in quality can be monitored quantitatively and a means is provided with which differences between, for example, alternative transmission systems can be analyzed in an informative way. The usability of this method is illustrated by two experiments. In the first experiment, subjects assessed the quality of still pictures comprising time-varying degrees of sharpness for a period of 815 seconds. In the second experiment, the time-varying quality of real video was assessed.

6.1 Introduction

For the optimisation of imaging systems, processing algorithms, etc., it is crucial to know how parameter settings affect the perceptual quality of displayed images. This calls for reliable and valid techniques for assessing perceptual image quality. At the moment, there are a number of techniques available if one wants to evaluate the perceived quality of images in an appreciation-oriented setting [1]. Single-stimulus or comparative judgements can be made on a set of still pictures [2, 3, 4]. For moving video the standards of the ITU-R, formerly CCIR, can be applied to 10-second sequences [2]. Either way, only one score is requested for each stimulus.

In the case of moving video, this implicitly urges a subject to weight the quality impression over time [5].

The limitation to short test sequences becomes a problem if one is interested in evaluating new digital systems. They often involve substantial quality variations that are not evenly distributed over time [6]. It is not unusual for severe degradations to appear only once every 10 or 20 minutes, while otherwise the quality is good. The standard methods of the ITU-R are not suited to the evaluation of such long sequences, whereas an evaluation of a pre-selected, particularly poor piece of video does not reflect the overall system's quality either.

Accordingly, the question becomes: how can the perceived quality of such long sequences be evaluated? In the present study, we have concentrated on an intermediate issue, i.e. are people able to assess quality continuously? And if so, what are they doing, and can we model their behaviour in a simple way? To this end, an experiment was designed in which subjects were instructed to continuously assess the quality of temporally degraded image sequences by moving a slider along a graphical scale.

Initially, we wished to concentrate on the measuring method. Therefore, the stimulus material had to be significantly simplified in the first experiment. We chose to limit ourselves to stills which were time-variably blurred. The advantages of this are that stills do not have time variation of the scene content, and that the mapping from Gaussian spatial filtering to perceptual quality can be predicted quantitatively [7]. Additionally, stimuli with a constant degree of blur were included in the experiment. This simplification allowed for a separation of the model prediction into two independent steps, one of which relates the sensorial strength of quality to the physical parameter of blur, while the second one describes the temporal characteristics of continuous scaling.

Subsequently, we wished to extend the method to real video. In that case the relation between the physical image and its perceived quality is not at hand. This meant that in order to check the validity of continuous scaling, perceived quality of the video at any moment in time should be measured independently. In terms of the aforementioned model, only the second step is relevant here. The image material was taken from 'Exam Conditions', an English comedy, which was MPEG-1++ coded at 2 Mbit/s.

6.2 Experiment 1: Static images with time-variable blur

6.2.1 Experimental set-up

The stimulus set was built up as follows. Two digitised stills, containing a harbour and a terrace scene, were blurred with 2D separable binomial filters with lengths of 1 (the original), 21, 41, 61, 81, and 101 pixels, and were subsequently transformed into video sequences of 10 seconds. Four additional sequences were made by time-

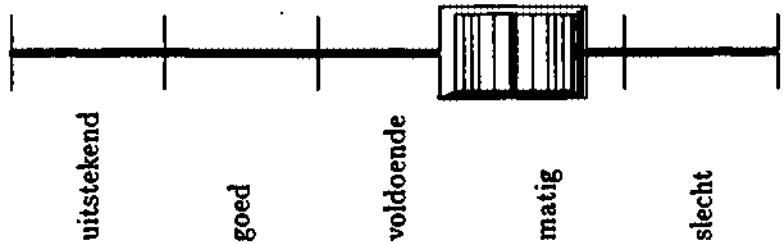


Fig. 6.1. Rotated schematic representation of the slider with Dutch adjectives, which have the following approximate English translations (from top to bottom): excellent, good, sufficient, poor, and bad.

linearly varying the filter length from 3 to 51, 3 to 101, 51 to 3, and 101 to 3 pixels, respectively, over 10 seconds of video. The resulting 10 sequences were shown either alone, or with 20 seconds of the original image before or after the 10 seconds. This yielded $30 - 1 = 29$ different sequences per picture (the 30-second original was not counted twice): 10 sequences of 10 seconds, and 19 sequences of 30 seconds. A random series of 29 sequences was put on a D1 tape, with 5 seconds of gray between each pair of sequences. The two source images were used alternately. The stimulus set was not completely factorial in the sense that each sequence of each source was shown: every possible sequence was shown only once, irrespective of the source image. Before the test session began, the subjects were shown an introductory set of 3 sequences with the original and two filtered versions (filter lengths 51 and 101) of another source image, containing a market street scene.

Each of the three female and four male subjects participated in four identical sessions. The number of subjects in each session varied between 2 and 4. They always sat at a distance of 6 times the picture height from the screen of a Philips 28 inch diagonal, 50 Hz consumer TV set. At this distance the pixel size on the screen is one arc min. The only lighting in the room was moderate lighting behind the TV set, at a level of about 15% of the maximum luminance level in the scenes, which was about 100 cd/m^2.

The instructions were to monitor quality [8] as closely as possible by continuously positioning a slider on a graphical scale. During the gray parts between the stimuli the subjects were asked to give an overall quality score for the previous sequence. The latter data are not used here, but this information is relevant for understanding the subjects' behaviour with respect to the sliders during the gray periods. The positions of the sliders were sampled once a second, with a resolution of 1 mm on a total scale of 100 mm. A schematic representation of the slider is given in Fig. 6.1. The adjectives next to the slider are the Dutch equivalents of the ITU-R Double Stimulus Continuous Quality Scale (DSCQS) method [2].

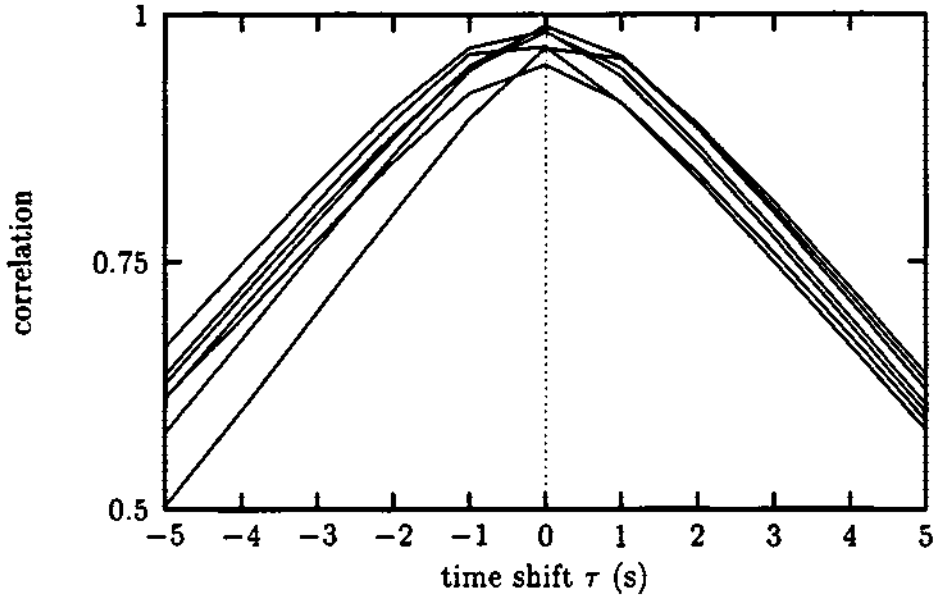


Fig. 6.2. Cross correlation between data of each of the 7 subjects and data averaged over subjects, for time shifts ranging from −5 to +5 seconds.

6.2.2 Results and Discussion

The correlations between the average results of the subjects varied between 0.889 and 0.975, with an average of 0.938. A cluster analysis on the correlation matrix did not yield clearly separated groups. Furthermore, cross correlation between data of each of the seven subjects and data averaged over subjects shows a clear maximum at time shift $\tau = 0$ (see Figure 6.2), meaning that differences between their behaviours in time are negligible.

We therefore decided to take the average over the subjects, after a linear correction for different uses of the scale. The latter transformation, a so-called z-transform [9], identifies the averages and standard deviations of the data sets of different subjects. Subsequently, the averaged data were linearly transformed to the original scale, fitting the original data as well as possible.

A surprising feature of the averaged time-varying score is that the scoring level of the original images remained fairly constant at a few millimeters below 80 throughout the session of 815 seconds. It seems as though the subjects treated the position at the mark between "uitstekend" (excellent) and "goed" (good) as an anchor position.

The two-stage model that was used to interpret the data is represented in Figure 6.3. The stimulus material had been chosen so as to enable the separation of the fitting procedures for the parameters into the two submodels. Furthermore, the use of blurred stills facilitated the identification of the first part of the model, the sensorial stage.

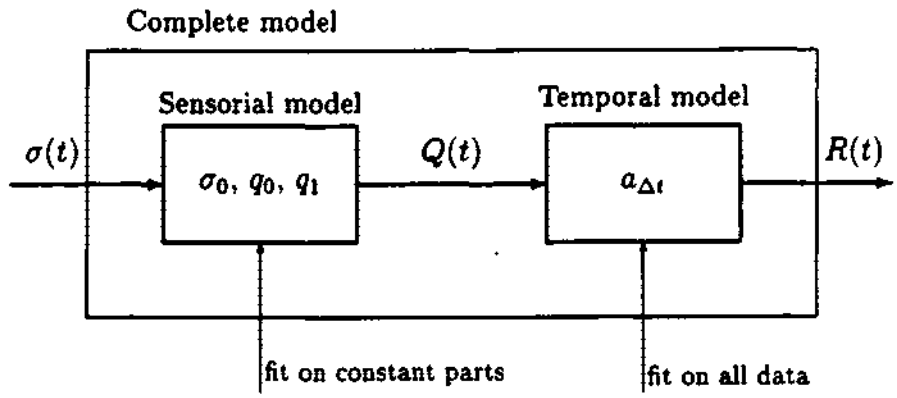


Fig. 6.3. The model used to relate the spread parameter σ of the blurring kernel to the continuous response $R(t)$.

6.2.3 Model step 1

As stated in the introduction, the relation between Gaussian filtering and quality can be described quantitatively [7]. In this case quality is linearly related to the sensorial strength of blur, because only one artifact is present [10, 11, 12]. The relation between the spread σ of a Gaussian blurring kernel and the sensorial strength of blur s is modeled by [7]

$$s(\sigma, \sigma_0) = 1 - \frac{1}{\left[(\sigma/\sigma_0)^2 + 1\right]^{1/4}}, \tag{6.2.1}$$

where σ_0 may be interpreted as the spread of the eye's internal blurring kernel. Eq. 6.2.1 is based on Fechnerian integration of jnd data measured by Watt and Morgan [13]. The relation between a Gaussian blurring kernel's σ and a binomial filter length ℓ is $\sigma = \frac{1}{2}\sqrt{\ell - 1}$ (see e.g. Martens [14]). In theory, the measure $s(\sigma, \sigma_0)$ ranges from 0 ($\sigma = 0$) to 1 ($\sigma = \infty$), whereas we need a measure on a scale from 0 to 100. In order to be able to match the ranges, we performed a non-linear 3-parameter fit on the following relation between the average response R and the theoretical estimator of quality Q

$$R \simeq Q \equiv q_0 - q_1 \times s(\sigma, \sigma_0), \tag{6.2.2}$$

where the parameters were q_0, q_1, and σ_0. The parameter q_0 can be interpreted as the quality score for the original image, whereas q_1 is the range of scores divided by the largest $s(\sigma, \sigma_0)$ in the experiment. As the fit only involves the relation between experiment and theory at stationary instants - no time aspects are involved at this moment - the fitting procedure only employed the constant blurred parts of the stimuli, whereby only data obtained more than 3 seconds after the onset of the constant plateau were taken into account. The fitted parameters had the values $q_0 = 78$ mm, $q_1 = 122$ mm, and $\sigma_0 = 0.85$ arc min ($r = 0.998$). The value of σ_0 is in reasonable agreement with the experimental data of Vos et al. [15] and Blommaert

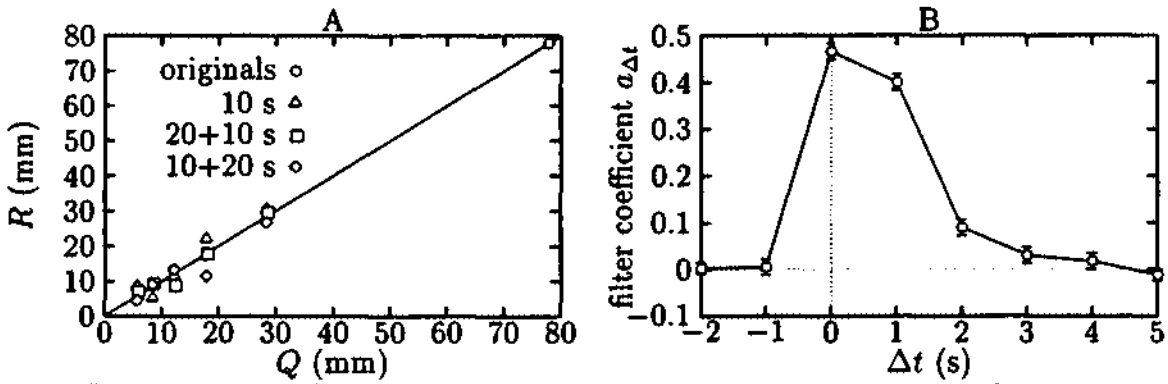


Fig. 6.4. Model fits on the quality data. The parameters q_0, q_1, and σ_0 in graph A were fitted on the constant, blurred parts of the stimuli. The labels indicate the different conditions: 10 s parts are the blurred ones, while 20 s parts before or after represent original material (see experimental set-up). The fitted time-filter coefficients are plotted in graph B.

et al. [16]. In Fig. 6.4A the averaged scores are plotted versus the model values with the above parameters. The different conditions, i.e. with or without 20 seconds of unimpaired video before or after the stimulus itself, are displayed separately, but there are no systematic deviations.

The above-mentioned absence of deviations indicates that the scaling procedure is reliable. This, together with the linearity displayed in Fig. 6.4A, suggests that no contrast effects occurred in the quality ratings. Secondly, the results support the validity of the rating method, because the construction of the sensorial scale is based on jnd data gathered in a completely different context. Thirdly, the data confirm that quality judgements are linearly related to the perceived difference between the original and the impaired image. This has previously been observed with another scaling procedure, viz. numerical category scaling [17, 11, 12].

6.2.4 Model step 2

Another question which has not been addressed up to this point concerns the specific temporal behaviour of the subjects as reflected in the data. Of course, one can argue that there must be a perceptual as well as a motor action delay, but how can this be modeled? To avoid having to introduce complicated submodels for the separate mechanisms influencing temporal behaviour, we just fitted a linear causal filter to minimise the difference between the sensorial strength data Q and the measured data R. This means that the rating R measured at a certain point in time is not only predicted by the sensorial strength Q at that point, but also by past values of Q. The weighting coefficients in this combination were determined via linear regression. In other words, if the sensorial strength data at time t are represented by $Q(t)$, and

the average rating by $R(t)$, the parameters $a_{\Delta t}$ in the relation

$$R(t) \simeq \sum_{\Delta t=-2}^{5} a_{\Delta t} Q(t - \Delta t) \tag{6.2.3}$$

were fitted over all the data in the experiment. The result of this fit is shown in Figure 6.4B. In this fit $Q(t)$ in the gray parts was considered to remain at the same position as the last sample of the previous stimulus, because the subjects had to fill in a form during those moments and were hence unable to move the slider. Causality was supported by the fact that the coefficients $a_{\Delta t}$ for $\Delta t = -1, -2$ s were negligible. In subsequent theoretical calculations only the first four nonnegative coefficients were taken into account, because the first two and last two were not significantly different from zero (see Figure 6.4B).

The resulting time-filter is clearly not a well-known simple function. In retrospect this was to be expected, as the different mechanisms which were proposed to explain the delay combine and most probably differ in nature. Moreover, its shape is influenced by the sample rate. It is to be expected that the real filter coefficient at $\Delta t = 0$ is zero, and that the values of the estimated filter coefficients at $\Delta t = 0, 1$ s must be distributed over time with a maximum around 0.5 s. The theory matches the data adequately, with a correlation of $r = 0.990$. This means that quality can be measured instantaneously, with little time-weighted averaging; in any case in a very consistent and reliable manner.

In order to better validate the last remark, we show those parts of the averaged scores where the 10-second parts of the sequences had time-varying blur. The three conditions (10 seconds alone, with 20 seconds unimpaired before, with 20 seconds after) are shown next to each other in separate panels. The central 10 seconds were labeled 0 to 10. The resulting graphs are shown in Fig. 6.5. From these graphs it is clear that the two-stage model presented above explains the data quite well. Note that the quality level of the previous stimulus is clearly visible in the graphs; the curves with different starting positions merge after approximately 2 seconds. This value reflects the properties of the fitted time-filter.

6.2.5 Conclusion

With this experiment we have shown that it is possible to measure quality continuously in a consistent and reliable way. Leaving the specific modeling aspects of fitting the data to a descriptive theory out of consideration, we believe that a quite promising and general method for assessing image quality has been found. With its great advantages this new method meets the present demands of developments in advanced digital coding techniques, i.e. quality variations can be monitored quantitatively and can then be used to analyze differences between alternatives in a much more informative way. Of course, this statement has to be validated in experiments in which real video material is applied.

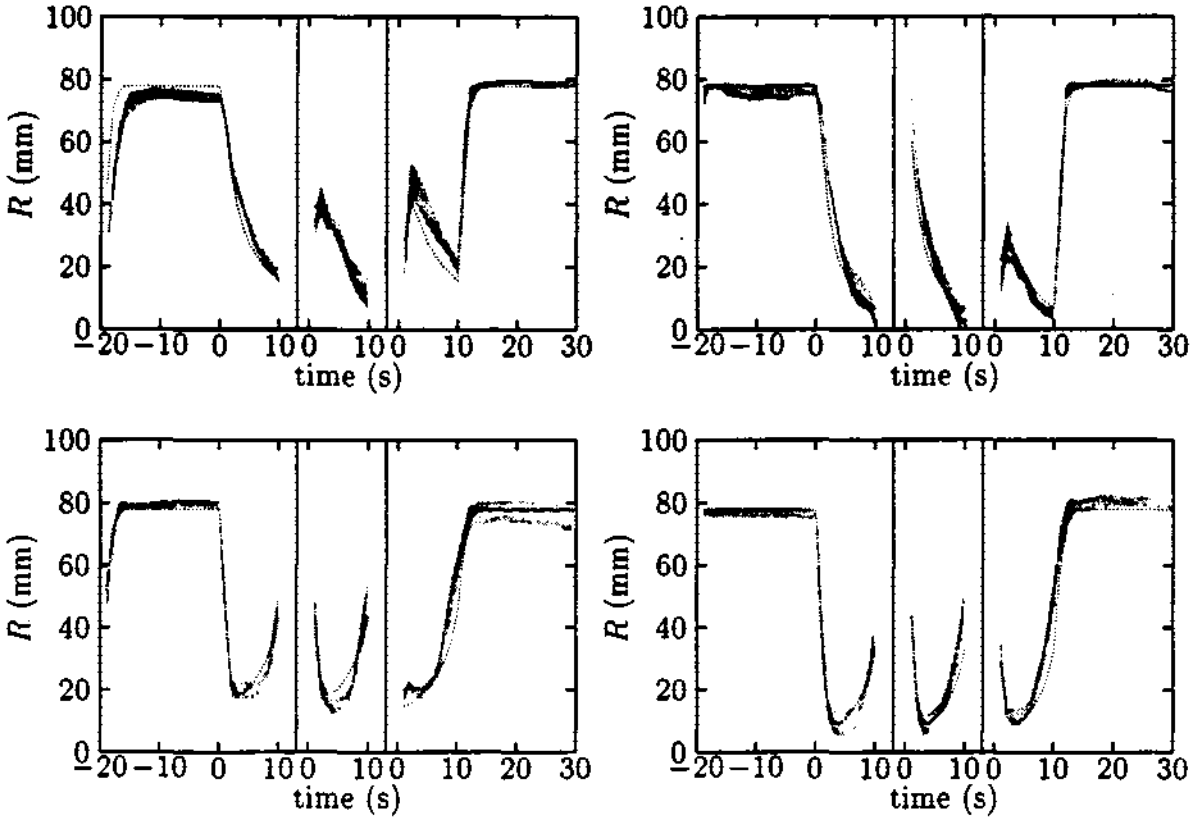


Fig. 6.5. Averaged responses R and their theoretical counterparts as a function of time for the stimuli with changing blur levels. The shaded areas indicate the confidence interval of 2 times the standard error of the mean (68%). Dotted lines are model fits. Note the adequately observed onset effects, in the two upper graphs near −20, and in all graphs near 0. The starting levels were assumed to be the same as the last sampled level of the previous stimulus. The three different conditions were separated as 20+10 s, 10 s, and 10+20 s (see caption of Figure 6.4).

6.3 Experiment 2: Real video with time-varying quality

6.3.1 Experimental set-up

The image material was taken from 'Exam Conditions', a 25-minute English comedy programme with little spoken text and no background music. In the actual tests, no sound was present. The programme was coded using an upgraded MPEG-1 coder at a bitrate of 2.0 Mbit/s. One segment of 50 seconds was selected according to the sensitivity for coding of the middle 10 seconds. This segment was embedded in the first 4 minutes of the programme. In experiment 2a the subjects were instructed

to monitor the quality of the 4 min sequence as closely as possible by continuously positioning a slider on a graphical scale. The positions of the sliders were sampled once a second, with a resolution of 1 mm on a total scale of 100 mm. A schematic representation of the slider is given in Fig. 6.1.

Preparatory to experiment 2b the 50-second fragment was divided into 77 segments with lengths varying between 0.44 and 0.72 seconds. The lengths varied, because we wanted to avoid the situation that scene cuts would be present within the segments. In the course of experiment 2b a segment (chosen at random from the 77 segments under test) was copied from a D1 tape to a video hard-disk system, from which the segment was subsequently shown in a palindromical way as long as it would take to copy the next segment off the tape to the hard-disk system, which was in the order of 20 seconds. At the end of each display the scores were read by computer from the same sliders as presented before.

In experiment 2a seven subjects performed the task 4 times, whereas 12 subjects participated in experiment 2b; only 2 subjects participated in both experiments. The number of subjects in each session varied between 2 and 4 in experiment 2a, and was equal to 4 in experiment 2b. The subjects were seated at a distance of 6 times the picture height from the screen of a Philips 28" 50 Hz consumer TV set. The only lighting in the room was moderate lighting behind the TV set, at a level of about 15% of the maximum luminance level in the scenes, which was about 100 cd/m^2. However, the most important fact here is that the conditions were equal in both experiments 2a and 2b.

An additional experiment 2c was done, in which – amongst other tasks – continuous quality scores were sampled twice a second for 3 subparts of the 50-second sequence, i.e., the first 30 seconds, the middle 10 seconds, and the last 30 seconds. The objective of this experiment was to investigate the relation between continuous quality scores and scores, which are given after viewing a sequence. This objective differs from the current one, but the results of the continuous quality scores can and will be compared to the results of experiment 2a. In this experiment 24 subjects participated, some of which also had participated in one of the experiments 2a and 2b. The viewing conditions were identical to the ones presented for experiments 2a and 2b.

6.3.2 Results and Discussion

The correlation between the averaged continuous results of each subject and the total average varied between 0.710 and 0.920 with an average of 0.838. The overall results of experiments 2a and 2c are plotted in Figure 6.6. The scales were linearly transformed before averaging, in order to correct for different uses of the scale (see section 6.2.2).

The data plotted in Figure 6.6 only guarantee the stability of the method, because the curves run parallel, but the validation should involve the 'real' perceived quality at any moment in the 50 s sequence. This ingredient is provided by experiment

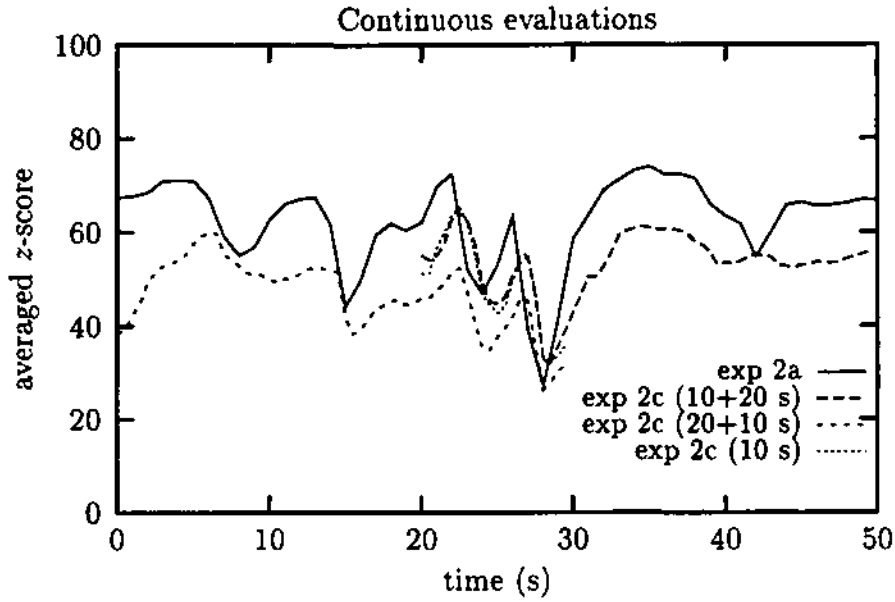


Fig. 6.6. The averaged results of continuous quality assessments over different experimental conditions. In experiment 2a, the 50 seconds are extracted from a 4 min sequence, whereas in experiment 2c the first 30 s, the last 30 s, and the middle 10 s were all evaluated separately.

2b, the results of which are plotted in Figure 6.7. It is obvious that the variations in scores are much larger than in the two other experiments. Moreover, the severe drops in quality in the middle occur at an earlier moment in time than the corresponding dips in Figure 6.6.

In order to relate the data sets of experiments 2a and 2c on the one hand and the data set of experiment 2b on the other hand, we took the same approach as in experiment 1, shown in Figure 6.4, the only difference being that in the current case only the second part of the model is relevant. The following regression has been carried out:

$$R(t) \simeq \sum_{\Delta t=-1}^{10} a_{\Delta t} Q(t - \Delta t) \tag{6.3.1}$$

In this relation the increments on Δt were 0.5 s near zero, and 1 s for $\Delta t > 6$ s. The time filters were separately fitted for experiment 2a and the averaged results of experiments 2a and 2c. The results are plotted in Figure 6.8.

In this figure it is apparent that a plain delay of 1–1.5 s is present, together with a exponentially decaying tail which extends to about 8 s in the past. We explain the first observation by a motoric delay, whereas the second one can be understood as a kind of resistance behaviour of the subjects of changing their scores in time. The quite large filter coefficient at $\Delta t = -0.5$ s can be explained either by synchronisation errors between the experiments and/or by instability of the fitting procedure; the

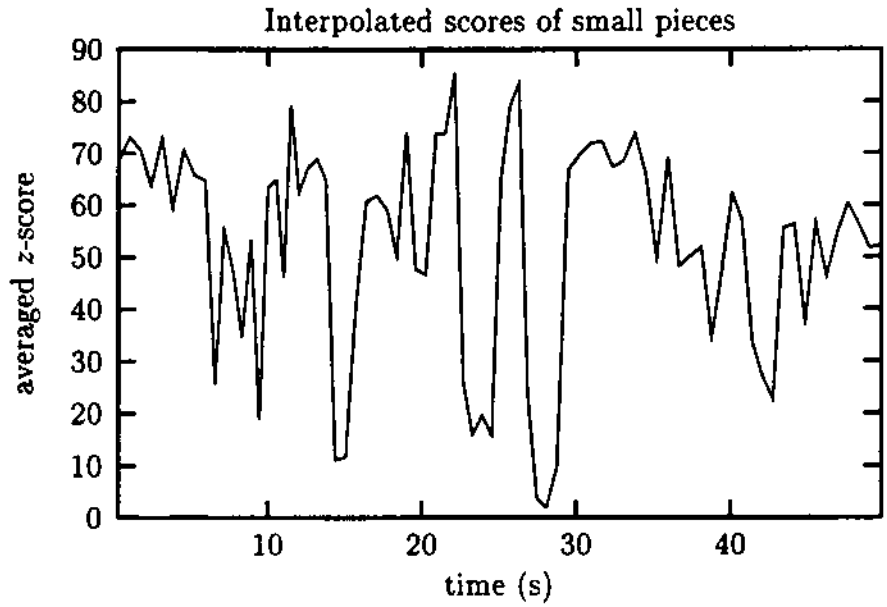


Fig. 6.7. The averaged results of quality assessments of small pieces of video, interpolated to get a continuous curve.

latter argument also applies to the nonzero coefficient at $\Delta t = 10$ s.

The comparison between the time filters of experiment 1 and experiment 2 yields no remarkable differences. Only the length of the tail is longer in the present case. Therefore, we conclude that the method of continuous evaluation can be applied in a simplified situation (static images) as well as in the complex case (real video) with the side remark that the length of the tail of the time filter is dependent on the degree of complexity of the scoring task.

The correlation of the filtered results of experiment 2b with the averaged results of experiments 2a and 2c is 0.818, whereas the fit for experiment 2a yields a correlation of 0.868. The first fit is shown in Figure 6.9. Because the filter has a length of approximately 8 s, the first 8 s in this graph are subject to initialisation errors.

The graph shows a deviation of the two curves in the last 10 s. This may be explained by the content of the video. In those seconds the sequence consists of a boy walking down an empty corridor. Coding errors are present in the background, while the boy is not affected by coding. In the continuous case, the subjects stated that they hadn't noticed the artifacts in the background. However, if one has to judge the quality of a shuttled short sequence which is observed for about 20 s, one can imagine that the artifacts will be included in the judgement. In other words, subjects have time to inspect the whole image in experiment 2b, whereas in the continuous cases the focal point is determined by an attention process. For practical end-user application of the evaluation continuous scaling seems thus more relevant than scaling of short pieces.

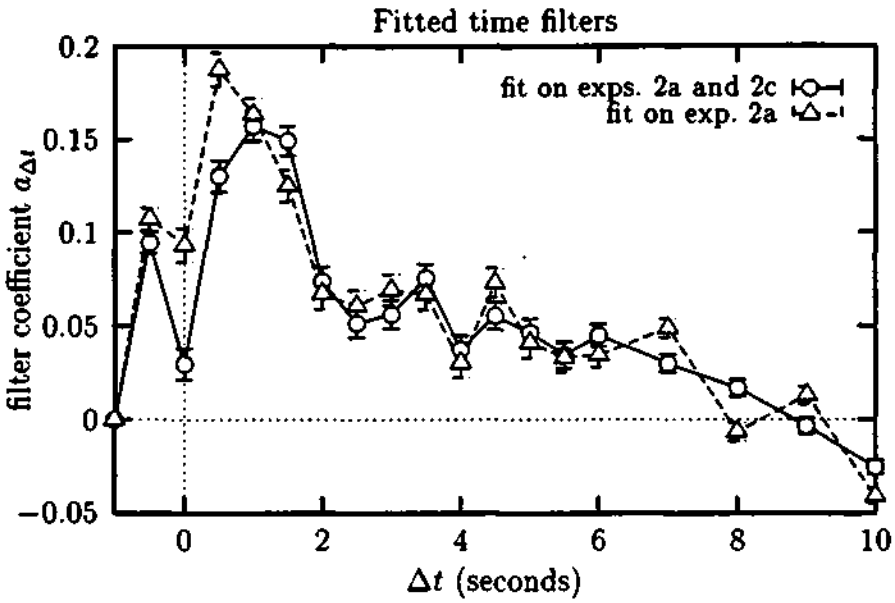


Fig. 6.8. The fitted time filters. Although the resolution in Δt is different from Figure 6.4B, the filters clearly yield comparable results. Apparently, a plain delay of 1–1.5 s is combined with a tail which extends to about 8 s.

6.3.3 Conclusion

With this experiment we have shown that it is possible to measure quality of video sequences continuously in a consistent and reliable way. With its great advantages this new method meets the present demands of developments in advanced digital coding techniques, i.e. quality variations can be monitored quantitatively and can then be used to analyze differences between alternatives in a much more informative way. How this method can be used to explain how overall quality of a transmission service or coding algorithm is related to the quality at any moment in time is still subject to research.

References

[1] J.A.J. Roufs. Perceptual image quality: concept and measurement. *Philips J. Res.*, 47:35–62, 1992.

[2] CCIR Recommendation 500-5. Method for the Subjective Assessment of the Quality of Television Pictures. 1992.

[3] J.W. Allnatt. *Transmitted-picture assessment.* John Wiley and Sons, New York, 1983.

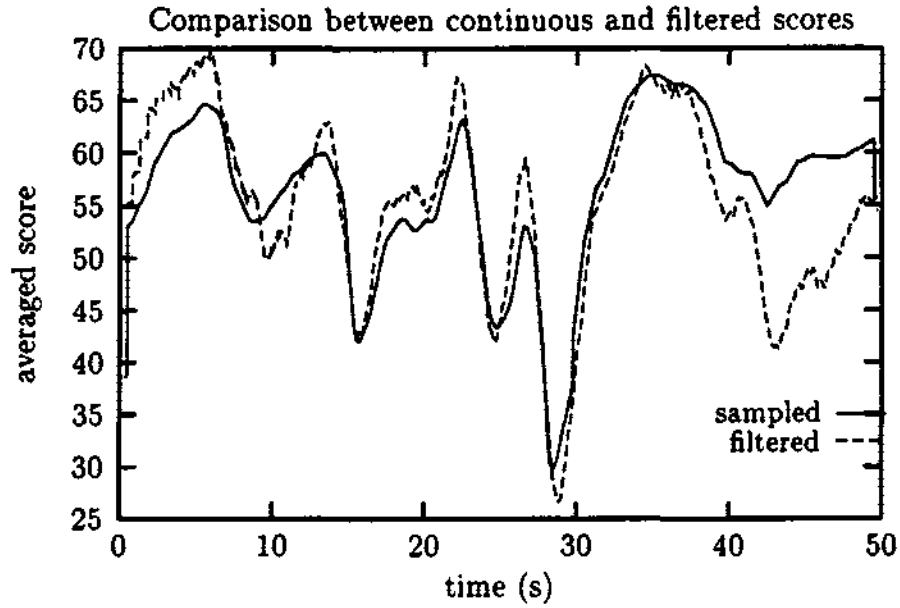


Fig. 6.9. A comparison between the continuous scores and the model prediction calculated from the scaled quality of short pieces by filtering.

[4] H. de Ridder and G.M.M. Majoor. Numerical category scaling: an efficient method for assessing digital image coding impairments. In B.E. Rogowitz and J.P. Allebach, editors, *Human Vision and Electronic Imaging: Models, Methods, and Applications*, pages 65–77. SPIE, 1249, 1990.

[5] V. Seferidis, M. Ghanbari, and D.E. Pearson. Forgiveness effect in subjective assessment of packet video. *Electronics Letters*, 28(21):2013–2014, 1992.

[6] N.K. Lodge. *Interpolative Coding Methods for the Digital Transmission of Conventional and High Definition Television.* PhD thesis, Heriot-Watt University, Edinburgh, 1993.

[7] M.R.M. Nijenhuis. *Sampling and interpolation of static images: a perceptual view.* PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1993.

[8] Since perceived blur and image quality have been shown to be linearly related [7], this experiment probably would have had the same results in the case subjects would have been asked to rate blur.

[9] C. Chatfield. *Statistics for technology.* Chapman and Hall, London, 3rd edition, 1983.

[10] J.H.D.M. Westerink. *Perceived sharpness in static and moving images.* PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1991.

[11] H. de Ridder. Minkowski-metrics as a combination rule for digital-image-coding impairments. In B.E. Rogowitz, editor, *Human Vision, Visual Processing, and Digital Display III*, pages 16–26. SPIE, 1666, 1992.

[12] M.R.M. Nijenhuis and F.J.J. Blommaert. Perceptual error measure for sampled and interpolated imagery. In *Eurodisplay '93*, pages 135–138, 1993.

[13] R.J. Watt and M.J. Morgan. The recognition and representation of edge blur: evidence for spatial primitives in human vision. *Vision Research*, 23:1465–1477, 1983.

[14] J.B. Martens. The Hermite Transform – Theory. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 38:1595–1606, 1990.

[15] J.J. Vos, J. Walraven, and A. van Meeteren. Light profiles of the foveal image of a point source. *Vision Research*, 16:215–219, 1976.

[16] F.J.J. Blommaert, H.G.M. Heijnen, and J.A.J. Roufs. Point spread functions and detail detection. *Spatial vision*, 2(2):99–115, 1987.

[17] A.J. Ahumada Jr. Computational Image-Quality Metrics: A Review. In *SID 93 Digest*, pages 305–308, 1993.

Chapter 7

SSCQE 'MOSAIC' Single Stimulus Continuous Quality Evaluation Associated Hardware and Software

Jean-Pierre Evain
EBU

7.1 Introduction

A great wealth of know-how and experience has been developed throughout the world in subjective assessment methodology. Subjective evaluation of picture and service quality has been often used in Europe. A limited number of laboratories, with high expertise, has regularly been involved in this work. Those experts, part of them MOSAIC partners, have largely contributed in the standardisation of the existing ITU-R methods. Results of international assessment campaigns, e.g. such as undertaken by the EBU or more recently by MPEG, have also been submitted to the ITU and published.

With the new digital techniques, there is an increasing need in subjective evaluations of bit rate compressed pictures and services. New test procedures (particularly the SSCQE), input/output data formats, and statistical process procedures have been defined by the MOSAIC partners.

7.2 General description

In addition to being the test driver (including hardware interfaces), the MOSAIC prototype software allows unified processing of data issued in a common format, e.g. from different European laboratories. The software deals with the genuine methodology developed in the project (e.g. SSCQE). For convenience, the DSCQS and DSIS methods will be included as described in ITU-R BT 500. As new methodolo-

gies could possibly be established later, even outside the framework of the project, it was decided to use a flexible object-oriented programming language to allow future extensions.

The software is considered as the central element of the subjective evaluation procedure. It is organised according to the different test phases: preparation, test performance, data processing and results presentation. The object-oriented core prototype software has been designed to be compatible with all types of laboratory peripherals (vote terminals, VCRs, etc.). Each laboratory will be encouraged to create the adequate 'software objects' (interfaces) to adapt the core MOSAIC software to its own environment and to deliver data in the formats presented in this document. These formats have been designed to leave all degrees of freedom in the definition of most of the parameters, also for undefined future specific needs.

7.3 Technical description

7.3.1 Short description of a subjective test framework

Performing a test can be done considering four different stages (see Figure 7.1):

- Test preparation (initialisation of the different test parameters);
- Test driver (laboratory hardware configuration, test performing and data gathering);
- Results preparation (data processing and preliminary presentation of results);
- Results presentation (elaborated matrices, EXCEL graphics, final report).

7.3.2 Input interface and data formats

Laboratories which have already been involved in international subjective evaluation procedures have all had the bad experience of receiving data files in various formats not directly compatible with their local processing tools. It is not rare that each laboratory has its own format and/or editor (non-ASCII formats from different word processing systems, data sheets, etc.). The format conversion is not only time consuming, but also a permanent source of errors. File and data structures, candidates for standardisation, have therefore been proposed by MOSAIC.

7.3.2.1 Test initialisation file

The test configuration is described in an initialisation file (e.g. TESTNAME.INI) in a Windows' TIF format. The structure of this file is described in Table 7.1.

{TEST}
Test and associated tape identification. The test initialisation file will be common to all the configuration files, data files and result files. It also contains all the information needed to edit the test tape. The environment of edition of the Master tape is also recorded as part of the test history.

{ALGORITHMS}
In this section, the different algorithms under test are identified including the reference as such (e.g. PAL, SECAM, MPEG profiles and levels at various bit rates, 4:2:2 source, REFERENCE). This information can be hidden to perform blind tests.

{SEQUENCES}
In this section, the different sequences used along the test are identified.

{CELL}
A CELL describes the timing elements of a basic test presentation. Description of the voting time slot is also given in this section.

{VTYPE}
The voting scale is defined by its minimum, maximum and step. If the step is very small in comparison to the voting then the scale is considered as continuous. A number of labels (grades/numerical adjectives or adjectives) can be defined in relation to a value or a voting range.

The man/machine interface, i.e. the voting peripheral, is different from one laboratory to another. The VTYPE format will be the reference for normalisation in each laboratory.

{VTYPE (observer vote)}
Description of additional observer vote types in complement to the default observer VTYPE. The observer vote are expected to be delivered only in an ASCII format.

{VTYPE (external complementary data)}
The possibility is offered to collect objective data in addition to the observer votes. A format is proposed to take this data into account in a BINARY or ASCII format.

{CELLx}
Each CELL corresponds to a pair (Sequence, Algorithm).

The information contained in the initialisation file is necessary and sufficient to edit a test tape. If the section ALGORITHMS is empty, it is possible to use the tape to perform blind tests. The Master INI file is needed to process test data and to deliver non-blind results.

7.3.2.2 Input data file format

The input data structure has also been defined. This format is designed to be compatible with usual text editors. This feature is particularly important if part of data is directly delivered on score sheets which need to be stored electronically before further processing.

The main features of the proposed data structure in Table 7.2 are as follows:

{RESULTS}
This section contains the data file identification including a reference to the test master initialisation file.

{PRIVATE}
This section should allow each laboratory to store private data. This field can be used for R&D purposes (e.g. studies on new methodologies).

{OBSERVERS}
Each observer's identification is required to apply efficiently the rejection criteria as described in ITU-R BT 500.

{COMPLEMENTARY DATA}
Additional data can be objective parameters (e.g. buffer occupancy, detection of shot changes) or complementary subjective parameters.

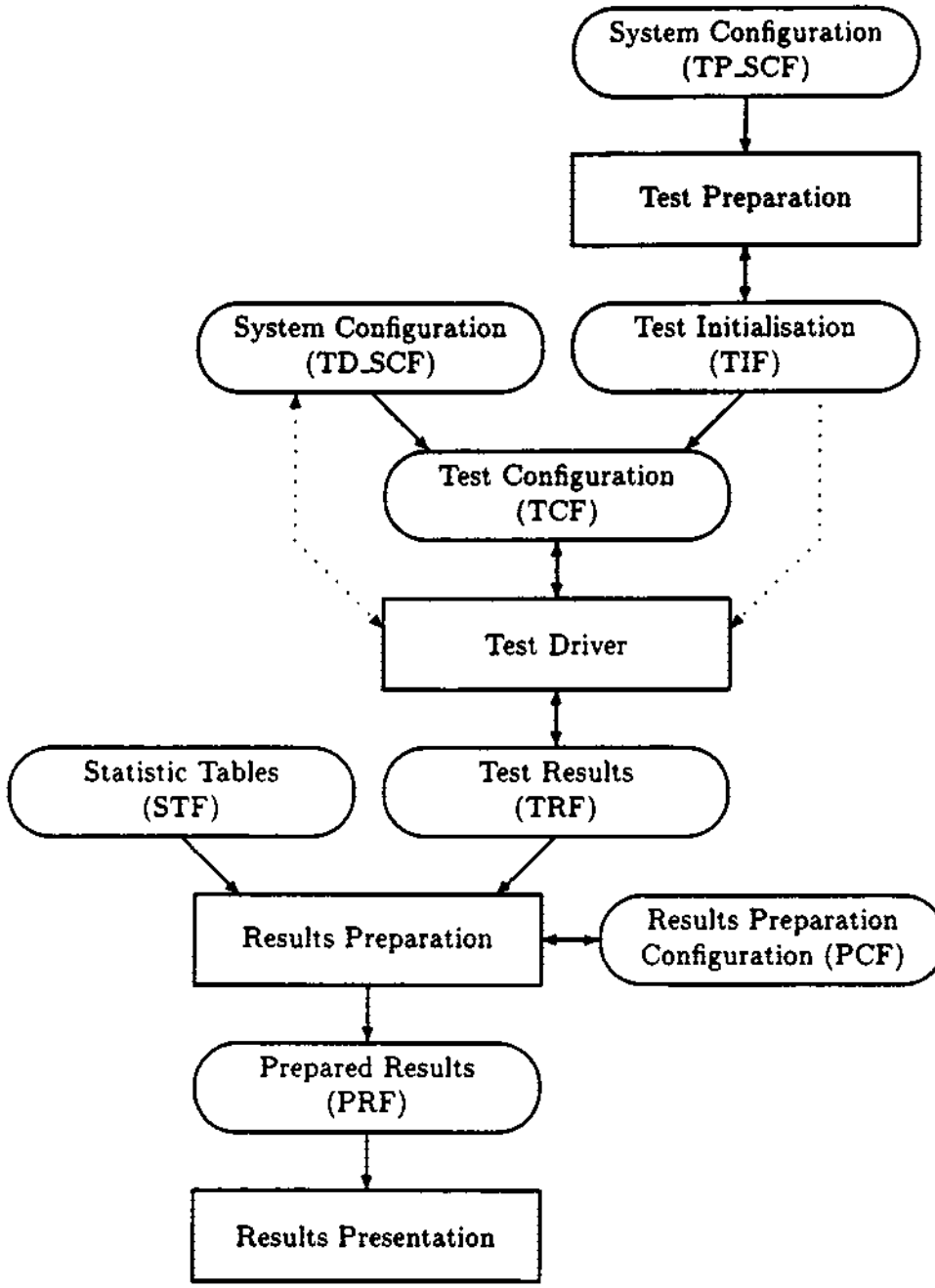
{DATA}
Data in a raw format with observer votes in line is arranged according to the presentation order. Each presentation is identified by the CELL number. It is possible to run blind tests using an initialisation file with an empty ALGORITHM section.

Looking at the data structure proposed in Table 7.2, it appears clearly that the use of an initialisation file simplifies dramatically the format requirements of the input data file.

7.3.3 Processing and output data formats

In the course of a subjective evaluation to assess the performance of a video communication system, a large amount of data is collected. These data, in the form of observers' score sheets, or hopefully in future in their electronic equivalent according to the formats proposed in this document, must be condensed by statistical techniques to yield results in numerical and/or graphical form which summarise the performance of the system/service under test.

In order to achieve maximum flexibility in data processing, it is assumed that each parameter can be selected independently (e.g. suppression of sequences, algorithms, observers). Each parameter has been identified and can be addressed separately through software filters (e.g. laboratories, sessions, viewing distances, vote types).



TP_SCF : Test Preparation, System Configuration File
TD_SCF : Test Driver, System Configuration File
TIF : Test Initialisation File
TCF : Test Configuration File
STF : Statistical Tables File
TRF : Test Results (data) File
PCF : Preparation Configuration File
PRF : Processed Results (data) File

Fig. 7.1. General test framework and file identification

{TEST}	
; Commentary	
Includes = <Filename>, ...	
Test Name = <Name>	
Test Type = <DSCQS/DSIS/...>	
; Master Tape ID	Date, type, tape format, equipment ID (VCR type), laboratory.
Timecode start = <Timecode>	
Number of CELLs = <value>	See {CELL} and {CELL(x)} hereafter.
Viewing Distances = <value>,...	
{ALGORITHMS}	In case of a blind test, [ALGORITHMS] will be empty.
Number of Algorithms = a	
A(1) = < name> (, REF)	REF (optional): In a DSCQS test, the reference is an algorithm as such in order to compare the reference to itself, taking into account a presentation order A-B/B-A.
...	
A(a) = <name>	
{SEQUENCES}	
Number of Sequences = σ	
SE(1) = <name>	
...	
SE(σ) = <name>	
{CELL}	
Number of Video Segments = s	
S(1) = <Start>, <End>	
...	
S(s) = <Start>, <End>	
Number of 'grey' segments = g	Optional.
G(1) = <Start>, <End>	
...	
G(g) = <Start>, <End>	
Number of vote segments = v	Default <Number of votes> value = 1.
V(1) = <Start>, <End>	If the INI file contains only the 'default' {VTYPE} then
(, <Number of votes>)	<Vtype> does not need to be specified here.
(, <Vtype>...)	
V(1,1) = <Video Segment> (,...)	Link definition between this vote segment and the video segment(s).
...	
V(1,nv) = <Video Segment> (,...)	
...	
V(v) = <Start>, <End>	Different <Vote Type> can be defined simultaneously corresponding to separate voting segments
(, <Number of votes>)	
(, <Vtype>...)	
V(1,1) = <Video Segment> (,...)	
...	
V(1,nv) = <Video Segment> (,...)	
{VTYPE}	{VTYPE}: Example of default Observer Vote Type to be used if no other {VTYPE<Name>} in the INI file
(Type = Observer)	Observer: keyword to identify 'vote data' in opposition to 'objective data'.
Min = <value>	Default <Step> value = 1.
Max = <value>	
(Step = <value>)	
Number of Labels = ℓ	
L(1) = <Label Name>, <Position>	The Label Name is a grade or an adjective (chain of characters).
(, <Min>, <Max>)	
...	
L(ℓ) = <Label Name>, <Position>	
(, <Min>, <Max>)	

to be continued on page 77

{VTYPE<Obsvote1>}	{VTYPE<obsvote1>}: new custom VTYPE definition.
Type = Observer	'Observer Type': description of an observer vote.
Min = <value>	
Max = <value>	More than one {VTYPE<name>} structure of Observer
Step = <value>	Type can be defined in the same INI file.
Number of Labels = ℓ	
L(1) = <Label Name>, <Position> (,	
<Min>, <Max>)	
...	
L(ℓ) = <Label Name>, <Position> (,	
<Min>, <Max>)	
{VTYPE<Extdata1>}	{VTYPE<Extdata1>}: new custom VTYPE definition.
Type = Other Data	'Other Data Type': description of a data format for e.g.
Min = <value>	objective data collected in parallel to the observer votes.
Max = <value>	
Format = <ASCII/BIN>	
Step = <value>	More than one {VTYPE<name>} structure of Other Data
Number of Labels = ℓ	Type can be defined in the same INI file.
L(1) = <Label Name>, <Position> (,	
<Min>, <Max>)	
...	
L(ℓ) = <Label Name>, <Position> (,	
<Min>, <Max>)	
{CELL(1)}	(x) = number in the random series from 1 to C.
(Status = <DUMMY/ACTIVE>)	In case of a blind test [CELL(x)] will not be defined.
S(1) = <Sequence name>, <Algorithm name>	
...	Status: optional, default is ACTIVE
S(s) = <Sequence name>, <Algorithm name>	
(G(1) = <level>)	
...	
(G(g) = <level>)	
...	
{CELL(C)}	
(Status = <DUMMY/ACTIVE>)	
S(1) = <Sequence name>, <Algorithm name>	
...	
S(s) = <Sequence name>, <Algorithm name>	
(G(1) = <level>)	
...	
(G(g) = <level>)	

Table 7.1. Structure of a Test Initialisation File

A Processing Configuration File (PCF) can be edited to filter new data according to a preliminary defined processing scheme.

7.3.3.1 Generation of random series

In single stimulus or double stimulus tests, presentations (sequences, algorithms) must be made in a random order (for instance derived from Graeco-Latin squares). The test order should be arranged so that any effects on the grading of tiredness or adaptation are balanced out from session to session. Some of the presentations can be repeated from session to session to check coherence.

{RESULTS}	
; Commentary	
Includes = <filename>,...	This file includes at least the test initialisation file (TIF).
Test name = <name>	
Laboratory name = <labname>	
Test Date = <dd/mm/yy>	
;Test configuration	Type of display, etc.
Data format = <DATA/CELL>	Default is CELL.
{PRIVATE <Laboratory Name>}	The possibility is offered to store local data.
...	
{OBSERVERS}	
Number of observers = *o*	
O(1) First Name = <name>	
O(1) Name = <name>	
O(1) Sex = <M/F>	
...	
O(1) Vdistance = <Distance>	
...	
O(*o*) First Name = <name>	
O(*o*) Name = <name>	
O(*o*) Sex = <M/F>	
...	
O(*o*) Vdistance = <Distance>	
{OTHER DATA}	
Number Data = *d*	Complementary data other than votes e.g. objective parameters such as buffer occupancy or a Boolean flag on scene change detection.
D(1) Name = <name>, <Type>	
...	
D(*d*) Name = <name>, <Type>	
{DATA}	BINARY format.
...	
{CELL(1)}	ASCII format.
V(1,1) = <vote (O1)>,...<vote (O*o*)>, <vote (D1)>,...,<vote (D*d*)>	(*x*): number in the test random series from 1 to C.
...	
V(1,*nv*) = <vote (O1)>,...<vote (O*o*)>, <vote (D1)>,...,<vote (D*d*)>	V(*v*, *nv*): subjective <vote O*o* > votes and objective <vote D*d* > 'votes' collected during voting segment '*v*' and corresponding to vote number '*nv*' in this segment.
V(*v*,1) = <vote (O1)>,...<vote (O*o*)>, <vote (D1)>,...,<vote (D*d*)>	
...	
V(*v*,*nv*) = <vote (O1)>,...<vote (O*o*)>, <vote (D1)>,...,<vote (D*d*)>	
{CELL(*C*)}	
V(1,1) = <vote (O1)>,...<vote (O*o*)>, <vote (D1)>,..., <vote (D*d*)>	
etc.	

Table 7.2. Input Data file Format

7.3.3.2 Statistical analysis for SSCQE

The main parameters to take into account in a subjective evaluation statistical analysis have been identified:

- J = number of different algorithms including the reference as such (REF);
- K = number of sequences;
- N = number of observers;
- M = number of votes;
- R = number of repetition (e.g. A-B, B-A, etc.);
- $v_{j,k,n,r}$ = raw vote (algorithm j, sequence k, observer n, repetition r);
- V_{sequ} = Variance of votes for the sequences ;
- V_{algo} = Variance of votes for the algorithms;
- V_{obs} = Variance of votes for the observers;
- $V_{\text{sequ,algo}}$ = Second order variance for the sequences and algorithms;
- $\overline{x}_{j,k}$ = mean of votes for one sequence and one algorithm;
- $\overline{x}_k$ = overall mean of votes for one sequence;
- $\overline{x}_j$ = overall mean of votes for one algorithm;
- $\overline{\overline{x}}$ = grand mean;
- ϕ = degree of freedom;

A basic statistical analysis includes generally:

- Arithmetic Mean
- Standard Deviation
- Confidence Interval
- Student Analysis
- ANOVA
- Rejection criteria to be applied to the observers

Mean, Standard Deviation and Confidence Interval are the main relative or absolute statistical values to be actually stored in the output data file.

In SSCQE, this preliminary analysis is achieved after specific data processing. It is also completed with specific statistical analysis.

In SSCQE, a continuous scale is used and scores are collected continuously (using e.g. sliders) along the duration of each presentation. In this case the amount of data can be very high (up to 2 votes per second). It is also possible to gather additional data relative to each voting instant: scene change detection, buffer occupancy or another entropy criteria dependent on the scene content and on the process under test, complementary statistical data such as score peak detection, etc.

SSCQE specific data processing and statistical analysis can be summarised as follows:

- Data pre-processing (e.g. z-transform);

  The test sequence is divided into a multiplicity of sub-sequences (of e.g. 10") which votes are filtered to obtain a unique grade per 10" time slot. This allows

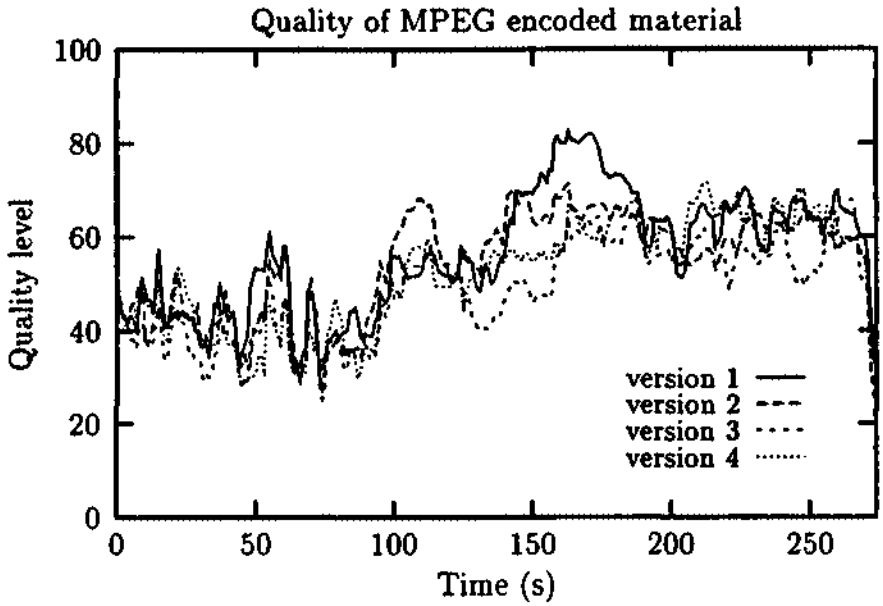


Fig. 7.2. Example of pre-processed data issued from single stimulus tests with continuous voting

to reduce drastically the amount of data, but it also encompasses different psychological effects bound the subjective evaluation (recency effects, etc.). The z-transform is one of these different possible linear filters.

Figure 7.2 shows an example of pre-processed data after z-transform.

- Statistical analysis of the pre-filtered data is then possible to compare different elements under test such as different processes (e.g. encoders, decoders at different bit rates, different transmission conditions) or different types of programmes.

  In addition to the basic statistical analysis over the number of observers, the histograms of the averages and standard deviations is processed and displayed.

- According to the results of the SSCQE statistical analysis, either it is immediately possible to draw a conclusion on the evaluation, or these results can be used to select 10" abstracts of the original long viewing test sequence to perform complementary DSCQS (Double Stimulus Continuous Quality Scale) or DSIS (Double Stimulus Impairment Scale) subjective evaluations.

7.3.4 Output data format and presentation of results

7.3.4.1 Output data format (numerical results)

The Numerical results will be stored according to the frame described in Table 7.3.

{TEST}
Includes = <TIF Filename>,...
Test Name = <Name>
Test Date = <Date>
Processing Laboratory Name = <Lab Name>
Result Sources = <Lab Name 1> (,...)
Number of algorithms considered = <value>
Number of sequences considered = <value>
Number of Repetition considered = <value>
Number of observers = <value>
Number of observers rejected = <value>
Viewing distance(s) considered = <value 1> (,<value 2>)

	Sequence 1	...	Sequence 'J'	Sequence 1 to 'J'	*reserved*
Reference if any	mean, SD, CI		mean, SD, CI	mean, SD, CI	
Algorithm 1	"		"		
...	"		"		
Algorithm 'K'	"		"		
Algo. 1 to 'K' + ref	"		"		

Table 7.3. Output data format for numerical results

7.3.4.2 Presentation of graphical results

The result data file is compatible with EXCEL, and a software call to EXCEL is automatically done for custom edition of complex graphs. Representative examples of graphs used for results presentation are given in Figures 7.3–7.6.

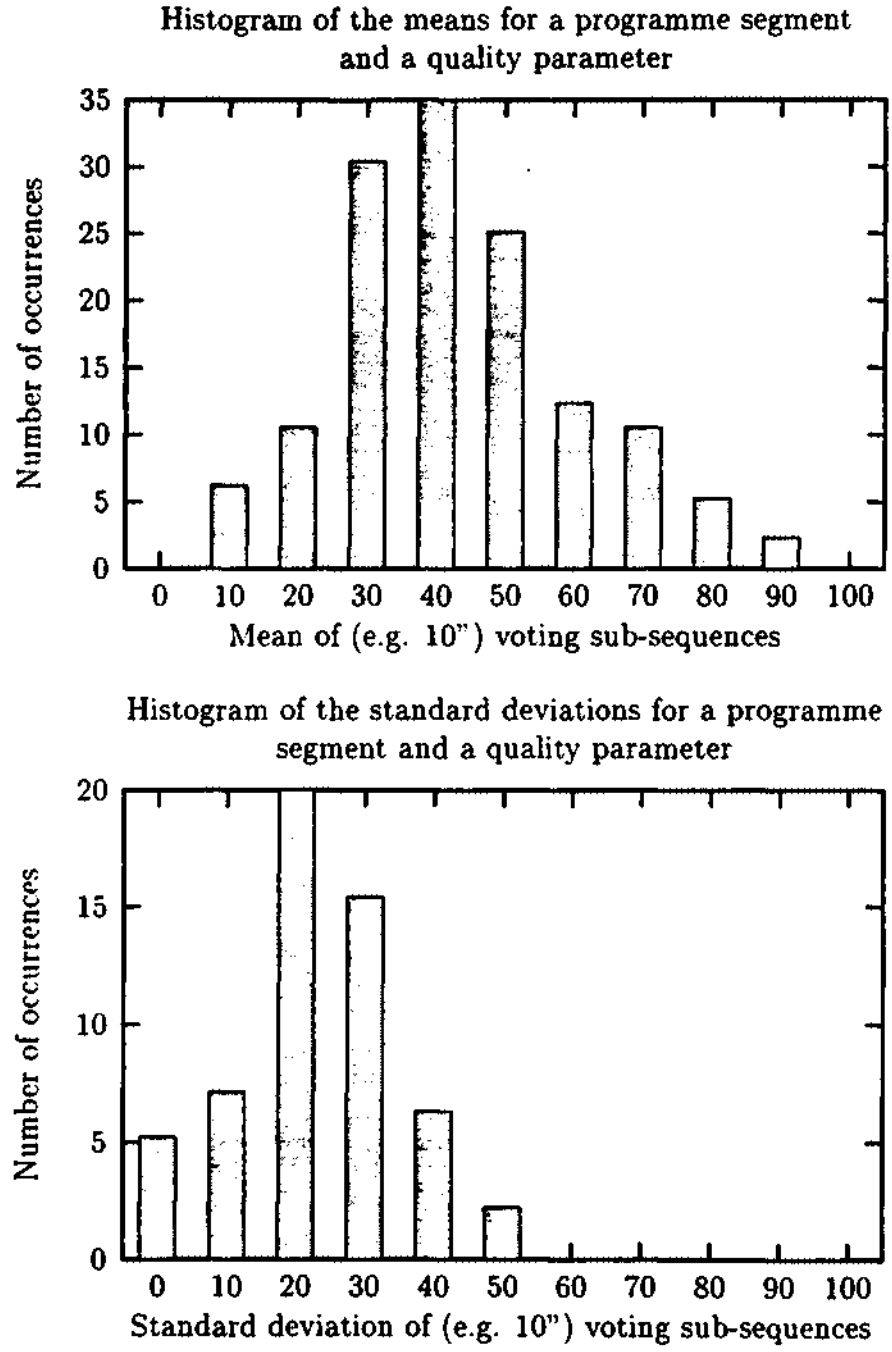


Fig. 7.3. SSCQE: Statistical distribution of occurrences (e.g. on 10" sub-sequences)

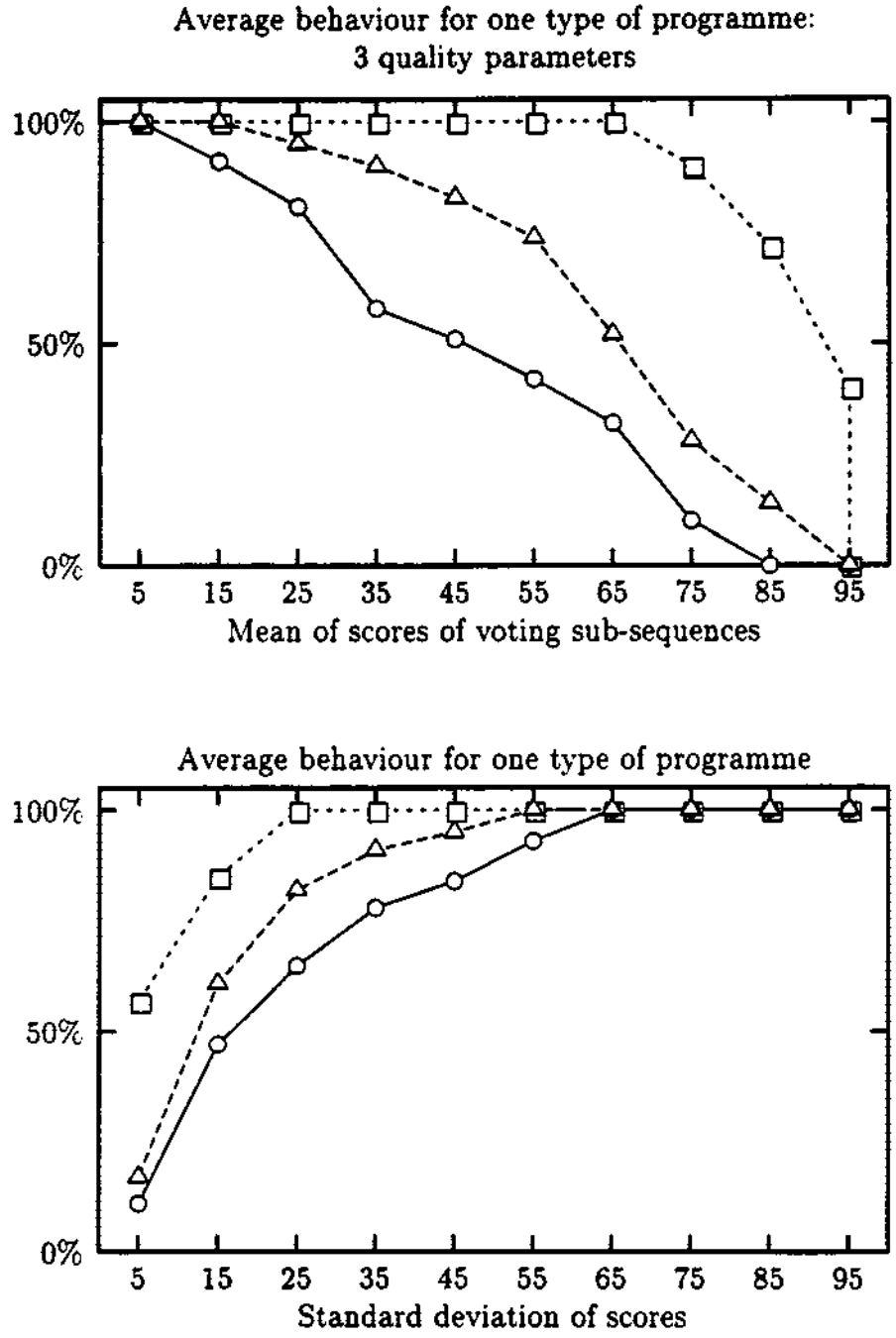


Fig. 7.4. SSCQE: Global annoyance characteristics calculated from the statistical distributions shown in Figure 7.3

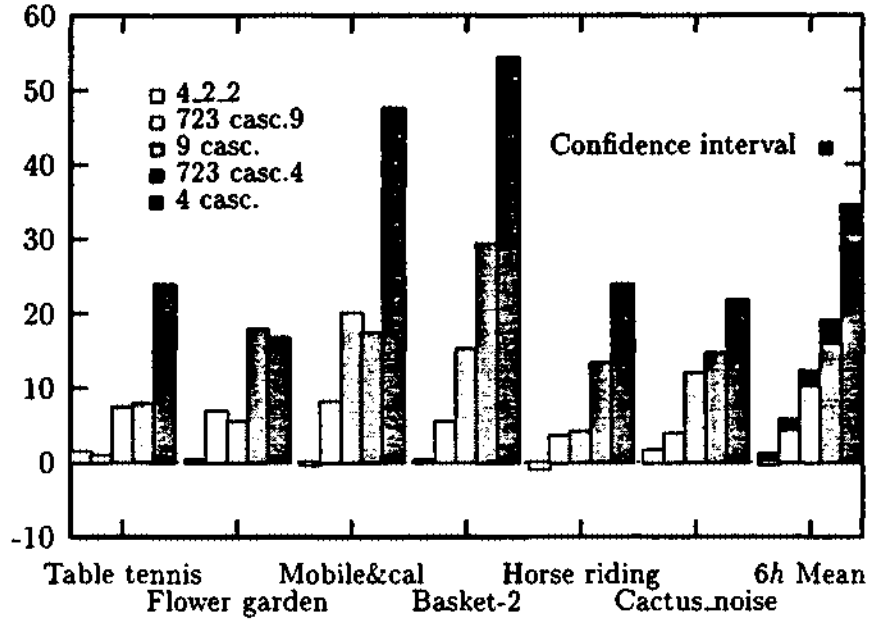


Fig. 7.5. DSCQS: MPEG-2 tests - Bar chart based on differential means

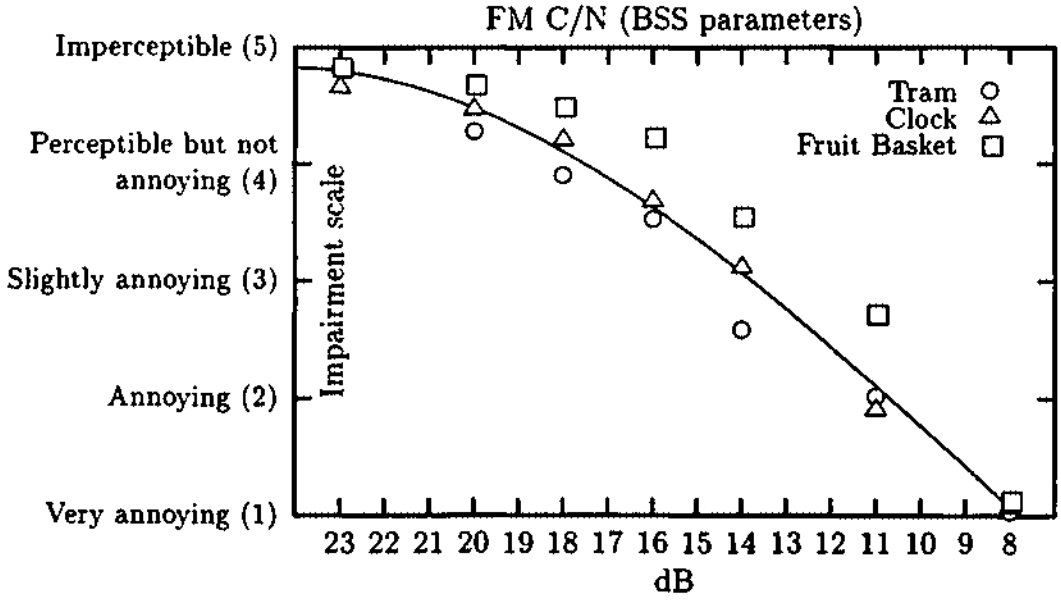


Fig. 7.6. DSIS: Example of noise failure characteristic (absolute mean / dispersion)

Chapter 8

Managing Subjective Tests to Evaluate the Quality of Images The IQ++ Platform

François Ziserman
ATLANTIDE

8.1 Introduction

This contribution deals with the IQ++ project. It first gives an overview of the IQ++ platform, and then, describes the format used to represent the data. Finally, the contribution outlines the programmes that are part of the tests' platform.

8.2 Overview

The IQ++ platform allows to drive subjective tests for the evaluation of quality of images. It is made of 3 programmes:

1. Test Preparation
2. Test Driver
3. Results Processing

All these programmes use a single data format: the IQF format (see Figure 8.1).

8.3 Data Format

The IQF format defines the structure of the data files created and processed during tests. As it is shown in Figure 8.2, each programme only uses a subset of the data.

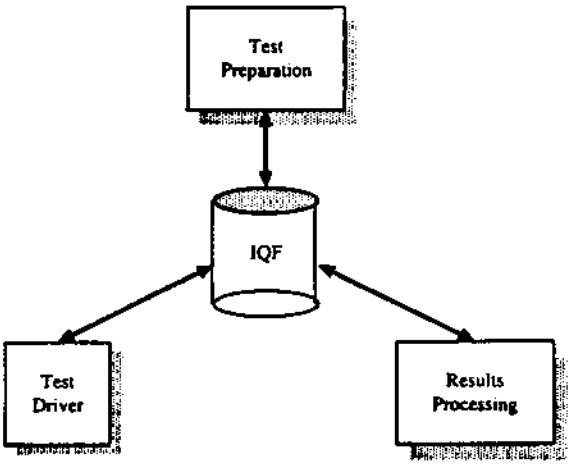


Fig. 8.1. Overview of the IQ++ platform

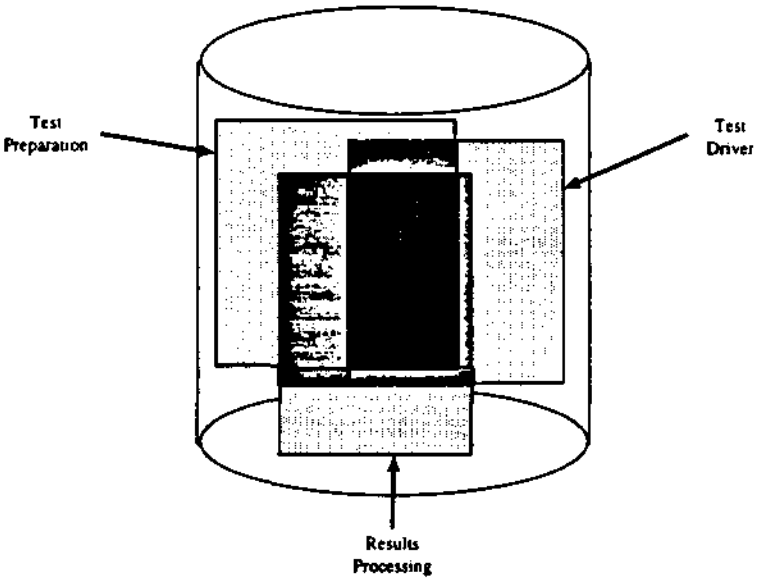


Fig. 8.2. The IQF data format

The subsets have of course several areas in common (represented as dark areas in the figure). They correspond to the data exchanged between programmes during a test. As an example, the data defined by the TestPreparation programme for a test will be accessed by the TestDriver programme when realising it.

The specific areas correspond to the programme private data. For instance, the configuration of the room where the tests occur, is useful for the TestDriver pro-

gramme, but has no relevance for the other two programmes. So, this kind of information is only managed by the concerned programme and not seen by the others.

It was important for this format to be human-readable, since the data have to be edited without any specific software. This led to the choice of the Windows .INI format. However, it is also possible to store some data in a much more reduced format when they can become very large. This is the case for example for the results of sliding votes. For these kinds of binary data, the results are kept in separate files.

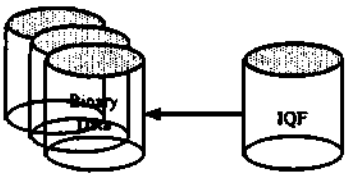


Fig. 8.3. Additional binary files

In Figure 8.3, the arrow between the files represents an enclosing binding: the human-readable data are put in the IQF file and are supplemented with the binary data stored in one or more files.

It is important to notice that the IQF format is very general, thus allowing to cover all current testing methods. It is also likely that the format might easily be adapted to new methods.

8.4 Test Preparation

This programme is used to prepare the tests. These tests can correspond to existing one like DSIS or DSCQS, but the programme also let you define you own kinds of tests (MyDSIS for example). Moreover, the programme can be run to create tests that don't fit in the two cases above.

In addition, the TestPreparation programme generates file using the SELO format, thus allowing to drive video tape recorders compatible with this format. A reduced object model of the data used by the TestPreparation programme is given in Figure 8.4.

8.5 Test Driver

The TestDriver programme receives in input a test defined by the TestPreparation programme and processes it. It can also takes an already run test, for example to achieve it because it was interrupted, or to play it again (entirely or partially). A model of the TestDriver programme is given in Figure 8.5.

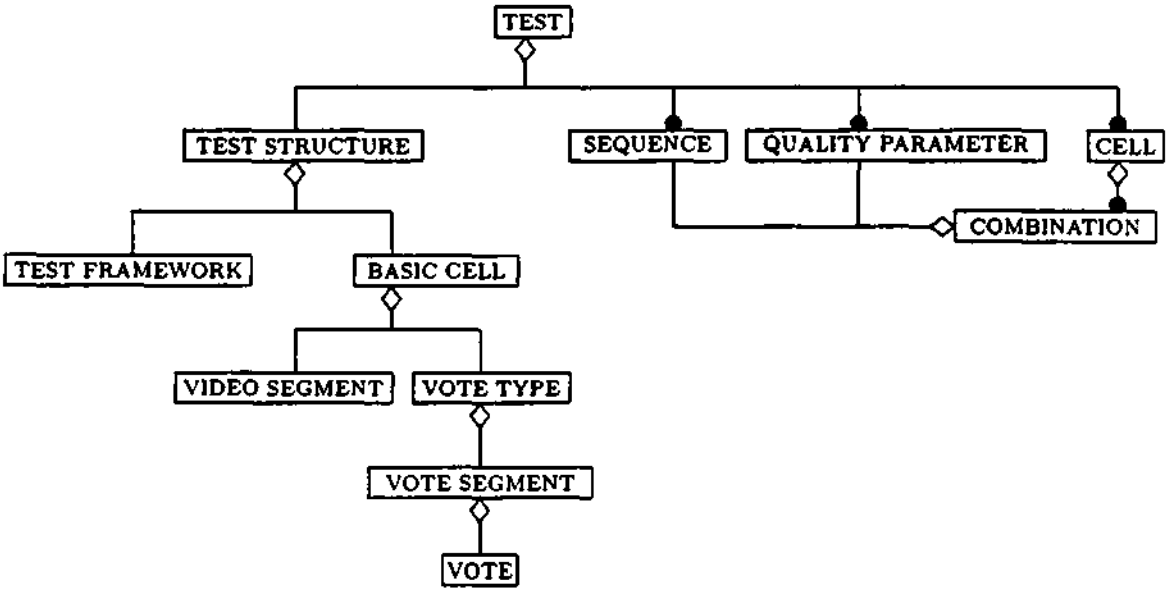


Fig. 8.4. Reduced object model of the data used by the TestPreparation programme

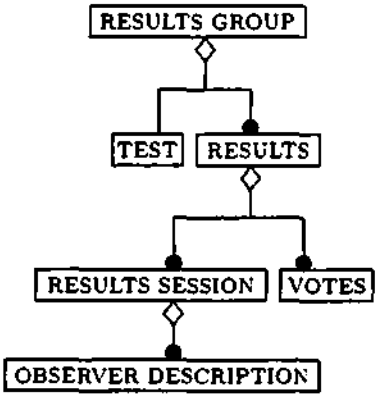


Fig. 8.5. Reduced object model of the data used by the TestDriver programme

8.6 Results Processing

This programme is used to process the results of the tests from a statistical point of view. The results are displayed in graphical form with an external viewer, like the Excel spreadsheet. A model of the ResultsProcessing programme is given in Figure 8.6.

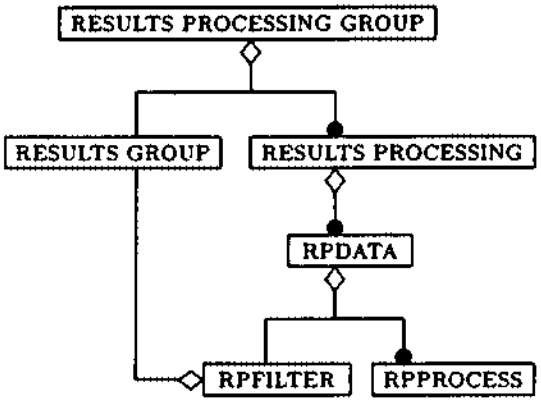


Fig. 8.6. Reduced object model of the data used by the ResultsProcessing programme

Chapter 9

Operational Monitoring of PAL Broadcast Television Quality

Derek Hawthorne
ITC

9.1 Introduction

In the United Kingdom, the Broadcasting Act 1990 requires the ITC to include conditions in its licences for terrestrial PAL television broadcasting services which are appropriate for ensuring high standards of technical quality and reliability. These standards have been published in the ITC Technical Performance Code and the ITC Handbook of Technical Standards for Television Programme Production [1, 2]. These documents are concerned with the setting of objective measurement targets and contain performance figures and tolerance limits for the main elements of video and audio equipment and signal paths used in television production. Programmes broadcast by ITC licensees must also meet subjective quality levels for vision and sound in order to ensure that overall quality criteria will be met.

The International Telecommunications Union Radiocommunications Sector (ITU-R) has established a set of scales and methods for overall subjective picture and sound quality assessment which is based on laboratory style methodology. This methodology is normally applied to the subjective quality assessment of still pictures or short sequences containing various levels of one or two impairments. In many cases tests are structured using identical unimpaired reference sequences for comparison.

The descriptive quality scale recommended by the ITU-R for use in these tests is a five point scale as follows:

Grade 5 - EXCELLENT
Grade 4 - GOOD
Grade 3 - FAIR
Grade 2 - POOR
Grade 1 - BAD

Broadcast engineers often assess the quality of programme material containing multiple distortions at different levels which may vary throughout a programme. The overall quality assessment for a complete programme has no precise theoretical or mathematical origin. In practice quality assessments are made at numerous times throughout the programme and a quality opinion is deduced leading to the award of an overall subjective quality grade. It is customary to apply the recommendations of the ITU-R when carrying out this type of 'operational grading' but in practice this only extends to the use of quality descriptions and standardisation of viewing and listening conditions. There is often no reference to an opinion sample and consistency is achieved by the consensus of skilled engineering staff.

Television transmission centres now operate with fewer specialist engineering staff and a quality control culture which relies on detecting defects during transmission is no longer appropriate when over 80% of all programmes transmitted in the UK have been acquired or independently produced. The ITC has produced a videotape which can be used by technical and non-technical staff as a quality assessment tool. Quality levels together with vision and sound impairments are described using programme sequences and a production style which is easy to follow.

This paper describes how the ITC videotape called 'An Introduction To 5 Point Grading' was made and concentrates on the difficult task of methodology design, selection criteria for sequences and assessment. The essential stages in this work were as follows:

- Outline test methodology
- Location and selection of programme material
- Engineering quality assessment
- Selection of sequences for sound & vision test tapes
- Non-expert assessment
- Analysis of results
- Selection of final sequences
- Produce ITC tape for distribution

9.2 Test methodology

It was the ITC's intention to draw on recommendations of the ITU-R in so far as these might be applied to actual programme sequences containing multiple distortions. From the outset it was our objective to regard the work as impartial and to produce a videotape containing quality reference examples based on the assessment of actual programme material by viewing panels made up of non-experts.

A number of important questions were considered before a test methodology could be designed:

- How many programme examples will be needed?
- Programme copyright issues?
- What criteria will determine which example is selected for assessment?
- What technical defects should be included?
- Where can assessors representative of viewers be found in sufficient numbers with access to suitable viewing and listening rooms?
- What can the tests tell us about the assessors?
- What factors in the assessment procedure might condition assessors and how can they be removed?
- What scoring method should be used, how will results be ranked and what results can be expected?
- How will sequences be chosen for the final tape?
- Will sound and vision quality be assessed together or separately?
- What will be the approach towards monophonic and stereophonic sound?
- What tape format will be used to acquire and display the test material for assessment?
- What is the expected programme duration?
- What delivery format will be used for the final tape?

Work has already been carried out by Bronwen L. Jones et al. [3] on the use of graphic scaling methods to determine the perceived intervals between various qualitative descriptive terms. Assessors can be pre-conditioned to assume a linear relationship between quality terms whereas previous work has demonstrated that the ITU-R terms are not always perceived as occupying equal increments on a linear scale. At the start of each viewing session and before any pre-recorded material was presented each assessor was asked to look at 15 sheets of paper collated at random and each containing a single quality category scale term and a 180 mm long scale (see Figure 9.1) for scoring from the following list:

Awful, Good[1], Not usable, Bad[1], Ideal, OK, Excellent[1], Inferior, Passable, Fair[1], Marginal, Poor[1], Fine, Not quite passable, Superior.

[1]ITU-R(CCIR) 5 point scale terms.

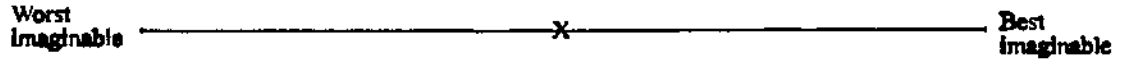


Fig. 9.1. Graphical scale for quality terms

The end points descriptions 'Worst Imaginable' and 'Best Imaginable' are boundless and are not intended to constrain the assessor. Scores are measured distances from the left hand edge. Assessors were told to use the same scoring method for vision or sound sequences which would follow.

9.3 Programme material resources

UK broadcasters maintain records of operational programme quality grades awarded on transmission which report technical defects. Thames Television allowed IBA and ITC engineering staff to use its extensive programme library and transmission logs in order to select suitable programme excerpts for this project. For operational grading purposes technical staff depart from the recommended 5 step scale by introducing $\frac{1}{2}$ grade levels to produce a scale with 9 steps. It should be noted that the ITC has distinguished the strict interpretation of the ITU-R five point quality scale and recommended methodology from what is common operational practice by referring to 'Star Grades' throughout this work.

Approximately 100 programme titles were initially selected based on quality grading reports from transmission logs and ITC engineering assessments. Programme material assessed below 'FAIR' grade (3 star) would not normally be transmitted and selections were made from material classified as technically exempt or rejected.

9.4 Assessment testing

Previous work [4] has shown that psychometric testing can be expected to produce scores which fit a normal statistical distribution. Furthermore, the ITU-R assessment procedures require, within certain tolerance, the average grade of the material assessed to be Grade 3, or 'FAIR'.

A normal distribution of test sequences which includes $\frac{1}{2}$ grade steps can be achieved with 54 examples where twelve sequences represent 'FAIR' quality and 'BAD' and 'EXCELLENT' quality are represented each by two sequences.

Variations between experts and non-expert opinions can arise. However results from previous work have shown that non-expert assessments can be reliable if the opinion sample is sufficiently large. It is the scores of non-experts which have determined which sequences have been selected for the ITC tape. Programme sequences may contain subject interest or artistic interest which can bias individual opinions and these particular sequences were avoided. Some preference was given towards

sequences which contained colour and sound which assessors could relate to reality e.g. human flesh tones, speech and outdoor scenes.

A single stimulus and double stimulus method of testing are usually carried out by non-expert observers with normal eyesight and hearing who are not concerned with programme quality assessments in their normal work. A Single stimulus method of testing presents a vision or sound sequence only once for assessment. Identical unimpaired reference sequences were not used in this work as is the current practice for operational quality grading.

9.5 Viewing and listening conditions

Viewing conditions conforming to ITU-R Recommendation 500-4 were used for vision assessments involving 5 assessors per session with mixed viewing distances of 4 and 6 times picture height in each session. The assessment of vision and monophonic programme excerpts was made in the same room which was acoustically treated to control reverberation. Loudspeaker power levels were adjusted for a peak power programme reference level of 75 dB at the listening position. This level was necessary to discriminate sound impairments at low level. Our tests confirmed that female assessors find sound levels higher than this uncomfortable.

Separate sessions were arranged for vision and sound assessments consisting of groups of 5 non-expert staff. A total of 60 non-experts participated in vision assessments and 85 in sound assessments. Separate sound sessions were set up using monitoring loudspeakers and headphones. Each assessment session lasted about 35 minutes. Stereophonic material was not assessed alongside monophonic material in the same session.

Operational quality assessments are made on vision and sound when both are present. It was anticipated that non-expert assessors might be influenced by sound quality when assessing vision or vice versa. Sequences did not have the same sound and vision quality. Since actual programme material was used for tests it was thought that some sound should be present to convey information during scene and shot changes. Conversely vision continuity was thought to be desirable during sound assessments. During vision assessments the level of programme sound was reduced by 15 dB. During sound quality assessments pictures were presented in monochrome with a reduced luminance level of 200 mV. A vision and sound tape with randomly ordered sequences having a normal distribution of quality levels was compiled for the assessment tests.

9.6 Pre-conditioning of assessors

Instructions were given at the start of each assessment before the 54 sequences were assessed. The marking system was explained and three examples of picture and sound quality were shown. The examples were loosely described as average quality

and below and above average. These examples were taken from a previous quality demonstration tape produced by the Independent Broadcasting Authority in order to maintain continuity with previous standards. The following statements were used to describe the examples shown: "The picture, example A, is of about average quality. The general colour is reasonable although there is some lack of sharpness in the hair. There are also some random flashes (Noise) and colour variations." This picture, example B, is of below average quality. The general colouring may not seem correct. The black and white areas of the picture show very little detail and the skin colour is poor, the picture lacks a general sharpness and overall clarity." "This picture, example C, is of above average quality. The colours may seem correct with plenty of detail in the light and dark areas and on the faces. There are no random flashing effects visible." At the start of sound sessions assessors listened to three examples and their attention was loosely drawn to coloration, hollow thin sound, noise, poor tone, unnatural speech quality, sibilance, balance and background noise.

9.7 Presentation of test material

Vision assessment sessions each containing 52 sequences with an introduction and preliminaries lasting 35 minutes were structured as follows: Sequence caption on mid grey field - 3 seconds, Test sequence 20 seconds with sound level reduced 15 dB, scoring period - caption on mid grey field 10 seconds. Sound assessment sessions each contained 52 sequences, introduction and preliminaries lasting 35 minutes were structured as follows: Sequence caption on mid grey field - 5 seconds, test sequence 20 seconds displaying monochrome picture level of 200 mV, scoring period - caption on mid grey field 10 seconds.

9.8 Results of assessments

Scores for quality descriptive terms, vision quality assessments and sound assessments with and without headphones were checked for validity and results confirmed the normality of data with some positive and negative skew at the upper and lower boundaries. Scores for the 15 quality descriptive terms based on 99 opinions were ranked according to mean and standard deviation to give the results as shown in Figure 9.2.

These results confirm that the quality terms used for operational grading do not fit a convenient straight line with equal intervals between each term. In addition non-experts have considered 'POOR' and 'BAD' to mean much the same thing although there seems little doubt as to where 'FAIR' should be on the quality scale. The dilemma we are now faced with is whether to relate the results of picture and sound assessments to the non-linear interpretation of the ITU-R descriptive terms, apply a strict linear interpretation or compromise a best fit.

The ITC has set a minimum quality threshold of 'FAIR' technical quality which

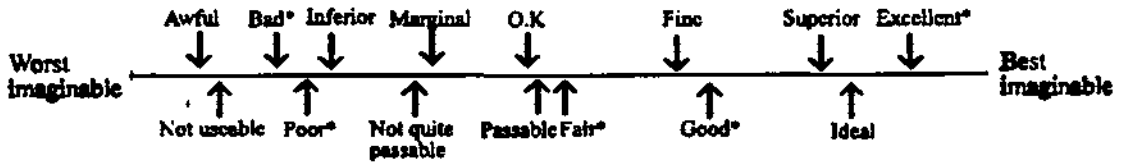


Fig. 9.2. Quality terms on the graphical scale, ranked according to mean and standard deviations. ITU-R 5 point scale terms are marked (*).

Vision sequences: sample size 59					Sound sequences: sample size 39			
Test No.	Mean Score	Std. Deviation	Star Grade		Test No.	Mean Score	Std. Deviation	Star Grade
V48	151	14	5.0	Excellent	S36	143	32	5.0
V26	138	26	4.5		S46	132	32	4.5
V38	121	28	4.0	Good	S22	115	34	4.0
V39	99	26	3.5		S40	99	37	3.5
V28	89	25	3.0	Fair	S39	85	36	3.0
V54	76	30	2.5		S43	69	31	2.5
V7	57	30	2.0	Poor	S6	59	34	2.0
V43	45	31	1.5		S7	40	27	1.5
V16	31	18	1.0	Bad	S9	33	32	1.0

Table 9.1. Star Grades

its licensees are expected to meet unless programme material is classified as technically exempt. The ITC and its licensees are particularly concerned to see threshold quality levels demonstrated which are above and below 'FAIR'. Applying the non-linear descriptive scale would not be appropriate since 'POOR' and 'BAD' quality excerpts might be indistinguishable. Calculated standard deviation and assessment of the dispersion profiles for each programme excerpt showed that there was considerable overlap in perceived quality levels. A best fit compromise was adopted to draw the Star Grade points as shown in Table 9.1. However this work has been well documented to allow staff involved in assessments to use the reported mean scores as benchmarks for self assessment.

The illustrations in Figs. 9.3 and 9.4 show the dispersion of scores for the grades 1, 3 and 5 star sequences.

Results for sound assessments using loudspeakers and headphones were very similar. It should be noted that there is greater dispersion in the data which suggests that assessors found sound quality more difficult to judge consistently. There is scope in future work to look at alternative test methodologies, the quality implications of

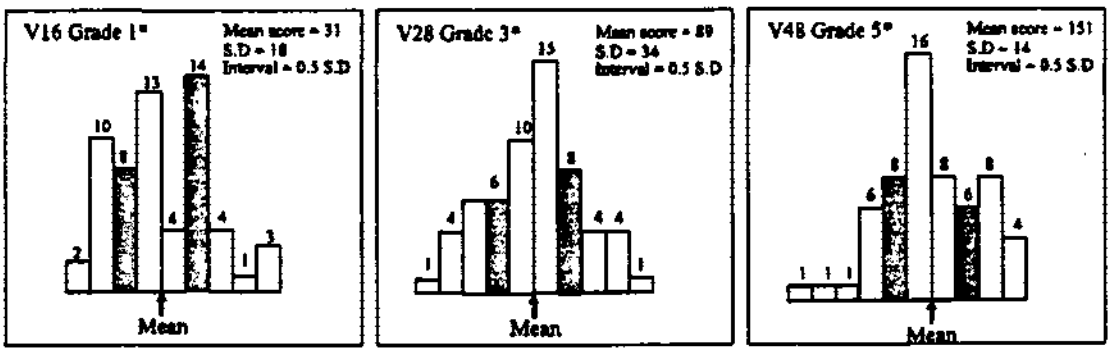


Fig. 9.3. Frequency dispersion vision sequences

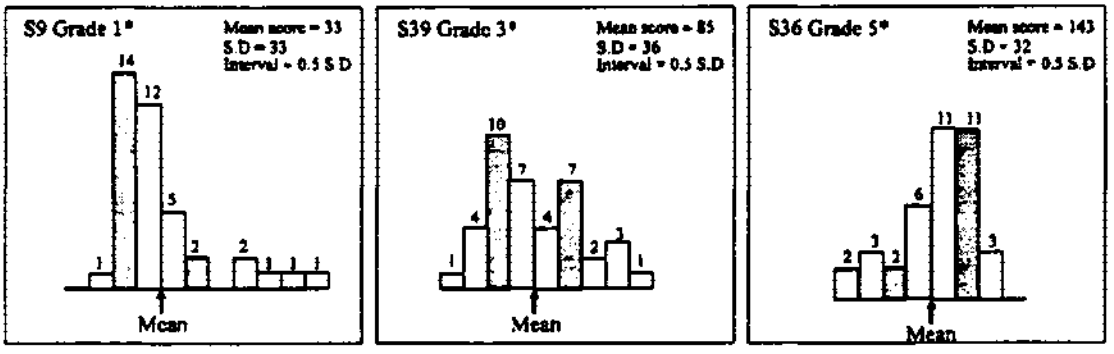


Fig. 9.4. Frequency dispersion sound sequences

impaired sound and vision together and the use of experts for assessment work.

9.9 Production and delivery format

The ITC demonstration tape will be of interest to all professional television production and broadcasting organisations. A popular broadcast quality tape format was chosen which had acceptable first generation performance to demonstrate programme excerpts during the assessment sessions. Digital videotape recorders were used to acquire programme material from the Thames Television library and to record excerpts from live studio programmes for the sound and vision demonstration tapes. Digital post production was used to produce the final ITC programme tape containing the selected excerpts and all copies of the ITC quality grading tape will be produced from a digital master tape. In this way programme sequences featured on the tape will be at the same quality level as those used in the original assessment.

The quality grading tape which the ITC has produced is intended represent qual-

ity in the UK PAL television system. The definition of 'EXCELLENT' quality is 'The best which the PAL system can provide'. In the same way that it is unfair to present mixed monophonic and stereo sound in the same session it is also unfair to present component and PAL sources together. All vision test sequences presented for demonstration had been PAL encoded.

Extensive use was made of Digital Video Effects and mixing to highlight problems on test sequences and whilst engineering opinions are given, these were not necessarily in the minds of the non-expert assessment group when sequences were being scored.

9.10 Summary

The making of a subjective quality grading tape to demonstrate quality levels requires considerable care thought and control at each stage. Statistical methods are involved in sampling opinions and individual opinion will vary. However the ITC tape will be useful as a calibration tool for assessing programmes and exploring the opinions of others.

9.11 Acknowledgements

The ITC wishes to acknowledge the considerable facility support provided by Thames Television and ITC colleagues.

References

[1] ITU Recommendations of the CCIR 1990 Volume X1 - Part 1, Broadcasting Service (Television) Rec. 500-4: Method for the subjective Assessment of the Quality of Television Pictures.

[2] ITU Reports of the CCIR 1990 Annex to Volume X1 - Part 1, Broadcasting Service (Television) Report 1082-1: Studies towards the Unification of Picture Assessment Methodology.

[3] Bronwen L. Jones and Pamela RMcmanus. Graphic Scaling of Quality Terms. *SMPTE Journal*, November 1986.

[4] John Allnatt. *Transmitted-picture Assessment.* Wiley.

Chapter 10

Contextual Effects in Sharpness Judgements

Huib de Ridder
IPO

10.1 Introduction

Scaling is one of the most efficient methods for assessing perceptual image quality and its underlying dimensions (sharpness, brightness, colourfulness etc.) [1]. Experiments with simple stimuli such as squares, circles and dot patterns have shown, however, that the outcome of a scaling experiment is susceptible to contextual effects [2, 3]. That is, the response to a stimulus depends not only on the stimulus itself but also on the other stimuli to be judged in a session. Contextual effects due to stimulus spacing or frequency of occurrence of stimuli have been found to have a substantial influence on the results of a single stimulus (or 'direct') scaling experiment [3].

The present study examined whether contextual effects are also present when complex stimuli, in particular images of natural scenes, are evaluated. To this end, two experiments were carried out measuring the possible influence of stimulus spacing on the evaluation of perceived sharpness for the following three scaling techniques: single stimulus scaling, double stimulus scaling and a scaling procedure based on difference judgements. Note that these techniques represent the three kinds of evaluation methods recommended by the International Telecommunication Union (ITU), formerly CCIR [4]. The first experiment concerned the use of three different response scales during single stimulus scaling. The second experiment was a direct comparison between the three above-mentioned scaling techniques using a 10-point numerical category scale only. In both experiments the stimulus set consisted of blurred versions of one static image of a terrace scene.

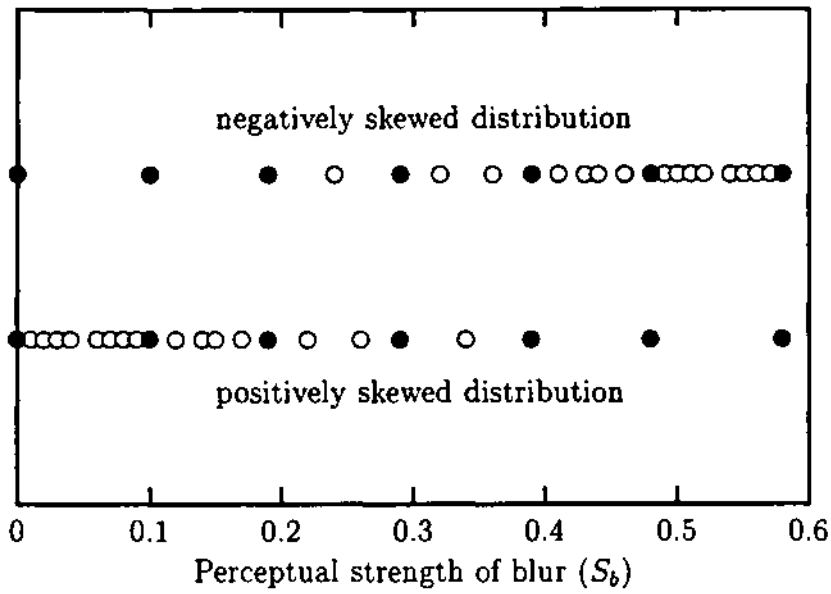


Fig. 10.1. Schematic representation of the negatively and positively skewed stimulus sets.

10.2 Experiment 1: Single stimulus scaling with different response scales

10.2.1 Method

10.2.1.1 Subjects

Eleven subjects in the age from 22 to 28 yrs participated in the first experiment. They had normal or corrected-to-normal vision. Their visual acuity measured with the aid of a Landolt chart at a distance of 5 m varied between 1.0 and 2.5. Two subjects took part in all experimental conditions. The other subjects participated in one condition only.

10.2.1.2 Stimuli

The sharpness of the static image of a terrace scene was manipulated using a Gould deAnza Image Processing System IP8400. The video signal obtained by scanning the slide of this scene was digitised with 8 bits/pixel on a grid of 512×512 pixels. During the experiments, however, only the central part of the scene was displayed (476×471 pixels). Blurred versions of the original image were generated by low-pass filtering the original with the aid of a 2D separable binomial filter. The resulting perceptual strength of blur is related to the standard deviation of the corresponding Gaussian

kernel (σ, expressed in pixels) by the following equation [5]:

$$S_b = 1 - ((\sigma/\sigma_0)^2 + 1)^{-0.25}, \tag{10.2.1}$$

where S_b denotes the perceptual strength of blur and σ_0 can be interpreted as the standard deviation of the eye's internal blurring kernel. In the present study, σ_0 was fixed at a value of 0.73.

There were two sets of blurred images which had seven images in common. The perceptual strength of blur S_b of these seven images ranged from 0 (the original image; $\sigma = 0$) to 0.58 ($\sigma = 4.07$) in regular steps of about 0.1. Accordingly, these images were evenly distributed with respect to their perceived (un)sharpness (Fig. 10.1, filled symbols). A stimulus set with a negatively skewed distribution was created by adding 15 comparatively unsharp images (Fig. 10.1, open symbols in upper row). Similarly, a stimulus set with a positively skewed distribution was created by adding 15 comparatively sharp images (Fig. 10.1, open symbols in lower row).

10.2.1.3 Procedure

The black-and-white images were displayed on a 70 Hz Barco CCID7351B CRT monitor placed in a dark room in front of a dimly lit 'white' background. The monitor was corrected such that the screen luminance was linearly related to the optical density of the original slide. The images were presented for five seconds after which a 9 cd/m^2 adaptation field appeared on the screen. Viewing conditions were in accordance with ITU Recommendation 500 [4]. The subjects viewed the monitor at a distance of about 1.5 m. At this distance, the pixel size is about 1 min of arc. During a session, the subjects assessed the sharpness of either the positively or the negatively skewed stimulus set. The three experimental conditions were determined by the rating scale the subjects had to use. In the first condition the subjects were instructed to judge sharpness on a 10-point numerical scale ranging from 1 (lowest sharpness) to 10 (highest sharpness). In the second condition this scale was replaced by a 5-point numerical scale ranging from 1 (lowest sharpness) to 5 (highest sharpness). In the final condition subjects rated sharpness on a 5-point adjectival scale using the Dutch equivalents of the ITU recommended descriptors as labels of the five categories.

10.2.1.4 Data analysis

The possible influence of the stimulus spacings described in section 10.2.1.3 on the sharpness judgements was analyzed with the aid of Parducci's range-frequency model [6, 7]. This model states that subjects tend to cover the perceptual range under investigation by the whole response scale and at the same time try to use each category an equal number of times. This implies that category judgements are a compromise between two principles, namely, a *range principle* postulating that each stimulus is judged in relation to the extreme stimuli that form the stimulus range and a *frequency principle* postulating that the same number of stimuli is assigned to each category.

Judgement $J_{i,c}$ of stimulus i in context c is assumed to be the weighted sum of these two principles, or

$$J_{i,c} = w * R_{i,c} + (1 - w) * F_{i,c}, \tag{10.2.2}$$

where $J_{i,c}$ is the category judgement linearly transformed to a scale running from 0 to 1. Range value $R_{i,c}$ is described by the following equation:

$$R_{i,c} = (R_{i,c} - R_{min,c})/(R_{max,c} - R_{min,c}), \tag{10.2.3}$$

in which $R_{min,c}$ and $R_{max,c}$ are the range values for the extreme stimuli. The frequency value $F_{i,c}$ of stimulus i in context c is related to the rank $r_{i,c}$ of this stimulus, or

$$F_{i,c} = (r_{i,c} - 1)/(N_c - 1), \tag{10.2.4}$$

N_c being the total number of stimuli in context c.

In the present study the extreme stimuli were always the same. Hence, the range values are independent of the stimulus spacing. Furthermore, the frequency values are based on the perceptual strength of blur, whereas the subjects evaluated the sharpness of the images. Taking this into account, the weighting factor w, reflecting the influence of context effects, can easily be derived from eqs.(10.2.2), (10.2.3) and (10.2.4). This results in the following expression:

$$w = 1 - (J_{i,neg} - J_{i,pos})/(F_{i,pos} - F_{i,neg}), \tag{10.2.5}$$

where *pos* and *neg* represent the positively and negatively skewed stimulus sets respectively. If w equals one, then the judgements are independent of the stimulus spacing. If w is less than one, a clear context effect is present. For simple stimuli, the value of w varies around 0.5 [3].

10.2.2 Results and discussion

No systematic subjective differences were found in the experimental data. Accordingly, it was decided to limit the application of Parducci's model to the data averaged across the subjects. The stimuli in the middle of the blur range were consistently judged sharper when they appeared in the negatively skewed stimulus set than when they were part of the positively skewed stimulus set. This trend was observed with all three response scales. The observed difference is in accordance with Parducci's model. To quantify this phenomenon, eq. (10.2.5) was fitted to the averaged data. The resulting values were: 0.76 for the 5-point numerical scale, 0.83 for the 5-point adjectival scale, and 0.73 for the 10-point numerical scale (Fig. 10.2, left-hand panel). These values deviate significantly from one, implying that there is a context effect. At the same time they are significantly higher than 0.5, suggesting that the contextual effects observed with natural images will not be as strong as those observed with simple stimuli.

Figure 10.2 also shows that the magnitude of the context effect does not depend on the kind of response scale employed. This conclusion can be generalised to graphical

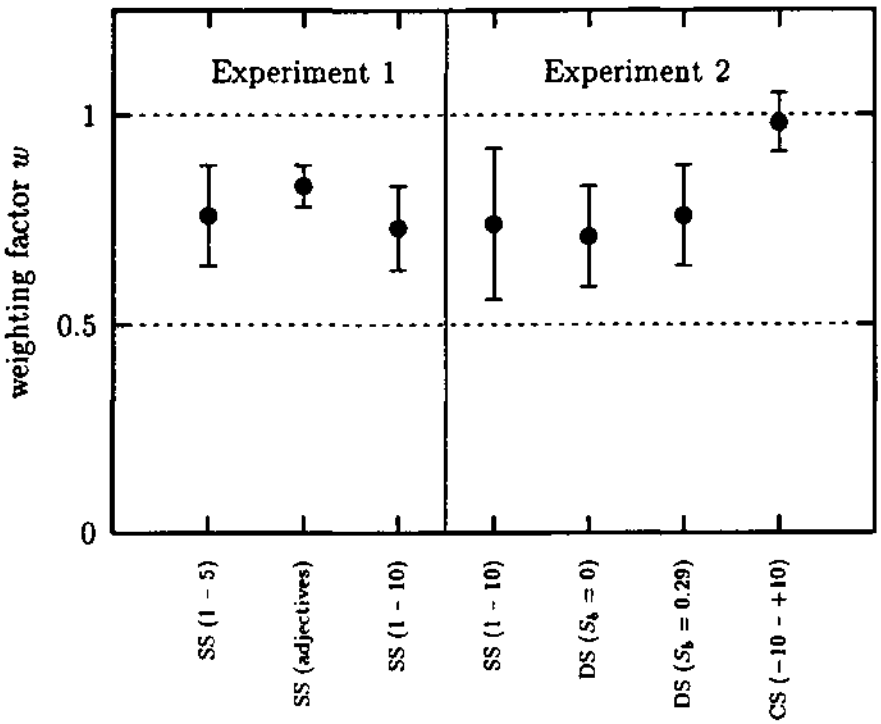


Fig. 10.2. The estimated values of weighting factor w. Vertical bars indicate the 95% confidence intervals. Left-hand panel: Experiment 1; SS = single stimulus scaling. Right-hand panel: Experiment 2; SS = single stimulus scaling, DS = double stimulus scaling, CS = comparison scaling.

scaling. Recently, Schifferstein and Frijters [8] found no differences between a 7-point numerical scale and a line scale during the assessment of the sweetness of various solutions of sucrose. The finding that the number of categories has no effect on the influence of the stimulus spacing agrees with experimental data gathered with simple stimuli [2, 3]. These studies also demonstrated, however, that this insensitivity to the number of response categories relies on the kind of context effect involved. For example, the contextual effects due to the varying degree of occurrence of the stimuli were strongly influenced by the number of categories [3]. It is not known whether this also holds when complex stimuli like natural images are evaluated.

10.3 Experiment 2: Single stimulus, double stimulus, and comparison scaling

10.3.1 Method

10.3.1.1 Subjects

Ten subjects in the age from 20 to 28 yrs participated in the second experiment. They had normal or corrected-to-normal vision. Their visual acuity measured with the aid of a Landolt chart at a distance of 5 m varied between 1.5 and 2.5. Four subjects took part in both the single stimulus scaling experiment and the comparison scaling experiment. The other subjects carried out the double stimulus scaling experiment.

10.3.1.2 Stimuli and procedure

The two stimulus sets and the viewing conditions were identical to the ones in the first experiment. The single stimulus scaling experiment was an exact replication of the one described in Section 10.2. Sharpness was rated on the 10-point numerical category scale only. For the double stimulus as well as the comparison scaling experiment reference images had to be introduced. In these experiments each trial consisted of a test and a reference image that were displayed sequentially with an interval of two seconds between the two five-seconds presentations. After each trial the subjects had to rate the sharpness of the two images on two separate 10-point numerical scales in the case of the double stimulus scaling and the difference between the perceived sharpnesses on a single scale ranging from -10 (the first image is much sharper than the second one) to 10 (the second image is much sharper than the first one) in the case of the comparison scaling. Before the results obtained by the double stimulus method were analyzed, the ratings for the test and the reference image were always subtracted. Two reference images, viz. the original with $S_b = 0$ and a mildly blurred image with $S_b = 0.29$, were used in the double stimulus scaling experiment. During a session only one of these reference images was shown. Three images with $S_b = 0, S_b = 0.29$ and $S_b = 0.58$ were used as references in the comparison scaling experiment. Per session the subjects assessed the difference in sharpness between the 22 images of either the positively or negatively skewed stimulus set and each of these three reference images.

10.3.2 Results and discussion

The right-hand panel of Figure 10.2 denotes the results of fitting eq. (10.2.5) to the experimental data averaged across the subjects. The most remarkable result is that the weighting factor w obtained for the comparison scaling is almost one, implying that for this scaling procedure the influence of the stimulus spacing is negligible. A similar result has been reported for simple stimuli provided a single perceptual dimension is involved in the judgements [2, 9]. The other conclusion that can be drawn

from Figure 10.2 is that there are no significant differences between the single and the double stimulus method with respect to their sensitivity to contextual effects due to stimulus spacing. Apparently, the double stimulus method behaved like a single stimulus method despite the fact that the differences in the sharpness judgements were used for the data analysis. The value of weighting factor w was 0.71 and 0.76 for the single stimulus and double stimulus method respectively.

10.4 Conclusion

In practice, the composition of the stimulus set to be evaluated on, for example, image quality is often fixed and cannot be manipulated. Accordingly, contextual effects due to stimulus spacing may seriously threaten the reliability and, particularly, the validity of the outcome of such a quality assessment. The present study showed that these effects are negligible when a method based on difference judgements is used. Other methods based on a single stimulus or a double stimulus procedure have been found to be sensitive to contextual effects due to stimulus spacing. Fortunately, the contextual effects observed with a natural image were found to be significantly smaller than the ones usually encountered with simple stimuli such as squares, circles or dot patterns.

10.5 Acknowledgements

I like to thank Johan Vereijken for generating the images and Christine Döring en Minke Apontoweil for running the first and second experiment respectively.

References

[1] J.A.J. Roufs. Perceptual image quality: Concept and measurement. *Philips Journal of Research*, Vol. 47, 35-62, 1992.

[2] B.A. Mellers and M.H. Birnbaum. Loci of contextual effects in judgement. *Journal of Experimental Psychology; Human Perception and Performance*, Vol. 8, 582-601, 1982.

[3] A. Parducci and D.H. Wedell. The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 12, 496-516, 1986.

[4] ITU-R Recommendation 500-5. Method for the subjective assessment of the quality of television pictures. 1992.

[5] M.R.M. Nijenhuis. *Sampling and interpolation of static images: A perceptual view.* PhD Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1993.

[6] A. Parducci. Category judgement: A range-frequency model. *Psychological Review*, Vol. 72, 407-418, 1965.

[7] A. Parducci and L.F. Perrett. Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology Monograph*, Vol. 89, 427-452, 1971.

[8] H.N.J. Schifferstein and J.E.R. Frijters. Contextual effects on judgements of sweetness intensity. *Perception & Psychophysics*, Vol. 52, 243-255, 1992.

[9] H.N.J. Schifferstein. Contextual effects in difference judgements. *Perception & Psychophysics*, Vol. 57, 56-70, 1995.

Chapter 11

The Influence of the Home Viewing Environment on the Measurement of Quality of Service of Digital TV Broadcasting

Thierry Alpert
CCETT

One of the main purpose of the RACE MOSAIC project was the improvement of the subjective evaluation procedures in the specific context of digital systems for TV broadcasting operations. The main drawbacks of the standardised methods are linked to the context-related apparitions of artifacts in the displayed digital images. In the current protocols, the viewing time of video sequence under evaluation is generally limited to 10 seconds which is obviously not enough for the observer to have a representative judgement of what could happen in the real service. Digital artifacts are strongly dependent on the spatial and temporal content of the source image. This is true for the compression schemes but also concerning the error resilience behaviour of digital transmission systems. Until now there was a big difficulty to choose representative video sequences, or at least to conclude about their representativity. The probability of occurrence of the impairments was not a controlled feature.

Another important motivation is the tentative to approach nearly the quality of service by placing the observers in a situation close to the real consumption of TV. It is somehow open to criticism to display and show source signals to observers and ask for a relative comparison of quality to this source quality. The source quality had never been and will never be broadcast so, it can be objected that these kinds of assessments are too severe.

Having settled the problem, a mean to reach the needed information concerning the general public was to organise a survey. This survey was defined to get the material conditions of TV watching but also made in order to evaluate the quality perceived by the general public as a function of the type of programme, the viewing conditions and the equipment level. These results could also be interesting to give

some orientations in the design of new TV services, particularly on what is concerned by the quality.

11.1 The survey

This work was based on observation of a panel of 500 television viewers who were non-representative of the TV viewing population as a whole, i.e. the sample is intentionally biased. 300 people come from a home with cable TV and 200 have neither cable TV nor satellite equipment. The advantage of such a sample is that information can be collected on two distinct populations (one which functions, in theory, in an environment that is closer to the future situation, and the other which functions in the current terrestrial broadcasting situation). Numbers in both groups were sufficiently high to produce significant results allowing a comparison of the various responses.

The company Médiamétrie was in charge of this survey for the CCETT. This explains why the interviewers had to be trained by the CCETT's experts on the TV quality parameters and the way to correctly measure subjectively or objectively these different relevant parameters.

The practical methodology applied was the following :

- preliminary sample recruitment by telephone,
- visit by investigators to homes provided with materials allowing measurement of the physical characteristics of the television and describing the environment of the TV set (lighting, observation distance),
- the people interviewed were at least 15 years old,
- only one single person per household was interviewed,
- the survey took place in several towns of France from January to June 1994.

Each observation was composed of six phases. Three phases of questioning the individuals: perception of current image and sound quality, expectations with regard to quality of reception and interest for future services and equipment. There was also three phases of objective observations of the households. The physical characteristics of the main TV set were measured (brightness, colour, geometric distortions, audible sound). The investigators also assessed the conditions of channel reception with a characterisation of visual and sound problems. This was done as objectively as possible, they were trained too at CCETT. Finally, they had to state the usual conditions of viewing and modes of signal reception.

11.2 Results about the physical description

Concerning the physical TV set observation, in 62% of cases the overall physical quality of the TV sets is good. Good quality spatial definition applies to 89% of

TV sets, photometric quality to 78% and geometric quality to 51%. Colorimetric quality is less satisfactory, only 37% of the TV sets present a good measured quality. The indices of correlation between the different measured qualities clearly shows an independence between the four intermediary qualities, and an impact of each one on the general calculated quality reaching 0.6 for the photometric quality which constitutes the main factor. The three other parameters are of equal importance with a smaller value of 0.4. The photometric quality was estimated by taking into account the ambient lighting, the white luminance of the TV set, the black one and the number of visible bars on a "pluge" signal. The photometric quality is of course dependent on the environment. The observed sizes of the TV sets were very spread from 33 cm to 82 cm in diagonal with the modus at 51 cm, the average being 55 cm.

Concerning now the parameters which are independent of the TV set, it was observed that, at least during the day, the lighting in front of the screen was in average 500 lux, the sensor being parallel to the screen. The absolute observation distances were spread between 1 m and 6 m with an average value of 3 m, the distribution being rather concentrated between 1.5 and 3.5 m. When this distance is measured in number of screen height (as usual in image quality) to be screen size independent, the average observed is around 9 but with a big variance. The distances where more than 10% of the panel can be found are 6, 7, 8, 9 and 10 times the picture height.

11.3 Analysis of the quality perception

The first quality related result which is essential for the validity of the other results is that interviewees appear to show a good understanding of the notion of quality regarding the picture and associated sound, when it is well explained. The interviewers' assessments are coherent with those of the people being surveyed.

Almost half the individuals questioned perceive a very good image quality on all of the channels studied (TF1, France 2, France 3, Canal+ and M6). Furthermore the perceived quality is better on cable networks than on terrestrial broadcasting. When looking for other variables having an influence on the perception of quality, it was found that the observation distance and the geographic situation were the main explanatory features. It can be mentioned that the ambient lighting is not so important, neither the consumption habits or the socio-demographic variables.

When it is asked about the quality on the video recorder, a markedly inferior score is obtained. On bought or rented pre-recorded cassettes the scores are again worse.

Concerning the analysis of the quality by type of programme, on the whole, individuals judge the quality to be very satisfactory. However, the types of programme offering the best quality are advertisements, news and sport programmes, with about 60% of people very satisfied. Then come the variety and entertainment programmes, game shows and documentary and magazine programmes, for which 55% are very satisfied. Lastly come youth programmes, telefilms, films and series for which the percentage of satisfied individuals ranges from 54 to 48%, which is still a decent level

of satisfaction! Whether it concerns image or sound individuals connected to cable network always appear more satisfied than those who are not.

There was a risk that the people make a confusion between the perceived quality and the interest in the programme content. This relation and bias certainly exist but they are not too large, the perceived quality and the consumption frequency of types of programmes are clearly independent. The fact of being either a large or small volume consumer of a specific type of programme does not lead to a particular perception of quality. So there is in this survey a global understanding of the meaning of the perceived quality, which was not really obvious when designing the questionnaire.

Therefore we can say that the consumption frequency of a type of programme depends on its content and not on the perceived qualities of image and sound, at least until a certain threshold of impairment is reached.

11.4 Analysis of the quality expectation

Analysing the quality expectations following the factor type of programme, it was found that the most demanding types were the films, then documentaries, magazine programmes, news and TV films. It is necessary to note the agreement of results between the high expectation of quality for films and the relative dissatisfaction concerning the current quality of this type of programme. Then come sport programmes, series and variety entertainment programmes. Lastly we notice reduced expectations for youth programmes, game shows and advertisements, which is easily understandable!

According to consumption habits, higher levels of expectation were observed for the large volume consumers and that for each type of programme.

Another observation is that the quality expectations don't vary neither according to the conditions of current perceived quality nor according to the screen size and the observation distance. Only the type of programme factor shows significant differences in the quality expectation.

11.5 Implication on TV quality assessments

As it was seen that both the quality perception and the quality expectations were content dependent in terms of type of programmes, it could be recommended that this fact should be taken into account by including excerpts of different programmes in the subjective tests. Long viewing sequences including audio could be also better because it will provide a simulation of real watching behaviour.

It was also observed that the general public is very sensitive to the degradation of the image quality but not so demanding, the current quality with good reception conditions is relatively satisfying. That is why a continuous evaluation of the test sequences with a protocol including no high quality reference could be more realistic in the case of an experiment aiming at the measurement of a quality of service.

Finally, a standard home context is certainly preferable rather than a ITU Rec. 500 viewing environment when results close to what could be the general public opinion are expected. Of course, for assessments of which the purpose is specifically the ranking of several quality parameters (algorithms, codecs, ...) the classical methods are more appropriate as they are certainly more sensitive though they are less realistic.

11.6 General conclusion

The big amount of data provided by this survey was processed in order to extract the main information about the quality perception and its relevant parameters.

The conclusions of this study have given some ideas about the general public expectations in terms of TV quality. It has also provided elements for the design of a standard home environment in the case of TV viewing and it has given some guidelines for the enhancement of the current subjective quality test methods.

The introduction of new TV services like thematic channels or tele-shopping needs a new subjective protocol, able to measure the quality of service on longer viewing sequences, representative of programmes contents including viewing conditions simulating a home environment.

Chapter 12

Preferred Viewing Distance and Display Parameters

Mauricio Ardito, Massimo Gunetti, and Massimo Visca
RAI

Abstract

An investigation, based on subjective tests according to ITU Recommendation 500, was carried out to correlate the TV Contrast-Luminance-Definition parameters to the subjective quality perceived by the viewers. The results show that high definition picture (HDTV) with reduced contrast has a subjective quality comparable to that of pictures with standard definition (SDTV) and high contrast. The luminance level also plays a role. Tests were carried out at the HDTV Design Viewing Distance: 3 times the screen height (DVD = $3h$). A test using the Preferred Viewing Distance (PVD) was also carried out. PVD is a subjective parameter, function of screen height. For the 38" HDTV, 16:9 screen used in this test, PVD was $5.2h$. The results show that the design of new displays for HDTV should guarantee large screen with adequate contrast and luminance in order that the higher definition contribution not be made useless.

12.1 Introduction

The standards of digital coding and compression of the pictures make it near and realistic the aim of the transmission of High Definition TV programmes (HDTV) to the users.

The overall chain of shooting, post production, coding, transmission, reception, decoding will be conceived to transfer to the user a better picture quality with a definition higher that the current one, but also without echoes, flickers and transmission noise; also the multichannel digital audio will contribute to implement a television which, in its whole, could be defined as High Quality.

Even if it may be expected that in a first phase, the films are the main and economic resource for the HDTV programmes, it cannot be ignored that the production cost of programmes specific and original for the HDTV will be surely high; consequently, it would be a considerable economic and production effort, but not useful and destined not to reach the final user if the terminal chain elements do not guarantee the complete quality performance.

In the current technological phase, the HDTV chain element deemed the weakest for the cost quality ratio could be the consumer display.

Today the professional HDTV displays with CRT (Cathode Ray Tube) technology have reached a high quality level, spreading in different application domains: from the graphic design to the professional cinema, from the computer art to the medicine.

HDTV, in order to offer the viewer the necessary involvement, and to enable him to appreciate the given resolution, requires very large display [7]: hence the encumbering size of conventional monitors would be difficulty acceptable in a domestic environment.

Looking at this consideration, quite strong is the stimulus towards the research of technologies which enable to implement space-saving displays: the so-called flat screen or projections screens; unfortunately the different technologies under development, be they at liquid crystals, plasma, projection, have not yet attained the same quality level of the CRT systems.

Another important parameter is the viewing distance. In fact, HDTV standards were designed with the assumption that the distance between the observer and the screen was fixed at the value defined Design Viewing Distance. For HDTV the DVD is equal to 3 times the height of the screen, i.e. DVD = $3h$. From this distance the viewer is able to appreciate the improvement due to higher definition. Vice versa, if the observer is left free to locate in front of the screen he tends to choose a viewing distance that is a function of the screen height h. This distance will be called Preferred Viewing Distance (PVD). This paper describes an experiment carried out in order to measure the PVD and its consequence on subjective assessment results.

Hence the present investigation aims at evaluating the interaction of contrast-luminance and viewing distance on the subjective quality of HDTV pictures. As basic condition is assumed that the display has a satisfying response to the definition provided by the video source.

The following of the paper is divided into three sections.

Section 12.2 recalls the concept of DVD and summarises the method used to assess the values related to the new concept of PVD.

Section 12.3 summarises the methodology used in order to correlate the objective contrast-luminance supplied by the display to the subjective picture quality. Subjective evaluations have been assessed considering both DVD and PVD.

Section 12.4 infers some conclusions taking into account results of the previous Section 12.3.

12.2 Influence of Viewing distance

The viewing distance is normally expressed by the ratio between the actual distance from the observer to the screen, and the height of the screen. So, $3h$, $4.5h$ or $6h$ viewing distance corresponds to 3, 4.5 or 6 times the used picture height, respectively.

12.2.1 Design Viewing Distance (DVD)

Design Viewing Distance strictly depends on the design parameters and on the specific use of the equipment. According to the ITU standards [9], DVD is the distance to be adopted for subjective quality measurements and for HDTV it is equal to 3 times the height of the screen. (DVD = $3h$).

12.2.2 Preferred Viewing Distance (PVD)

The aim of this part of the investigation was to answer to the following question: "Apart from the Design Viewing Distance (DVD) and neutralising as much as possible the environment parameters, which is the viewing distance chosen by a sample of viewers, who have been left free to decide, for a long and comfortable observation?"

The answer to this question defines the Preferred Viewing Distance (PVD).

For this purpose a series of tests was planned, during which each single observer, sitting in front of a video screen, was asked to choose the most comfortable viewing distance for a long observation.

As it is reasonable to think that Preferred Viewing Distance is a function of different parameters such as screen size, picture resolution and the degree of movement contained in the picture, the planned test took into account these three parameters.

Screen size - In order to assess the influence of the screen size, the investigation was conducted with 16:9 viewing system, from 12" to 160" large screens in HDTV standards. The following equipment was used: 12" to 38" CRT monitors, an HDTV 54" rear projector and a projector which allowed to get pictures up to 160" by varying the distance from the screen.

The viewing and environmental conditions of the investigation are summarised in Table 12.1.

Picture Resolution - In order to assess the influence of picture resolution it was planned to repeat the test on different picture standards: HDTV (1440 pixels 16:9 format) and SDTV (720 pixels 16:9 format) sequences were considered.

The tests were carried out with HDTV pictures recorded on the Quadriga system; the pictures were scaled to SDTV by means of a down-converter.

Since only 16:9 HDTV viewing systems have been used, the display of SDTV sequences required an up-conversion process.

Screen technology	Screen diagonal [Inch]	Screen luminance [Nit]	Environmental luminance	Room dimension [m]	Maximum viewing distance
CRT	12" - 38"	70	ITU Rec. 500	5.40 × 8	12h with 38"
Rear projector	54"	70	ITU Rec. 500	5.40 × 8	9h with 54"
Projector	Up to 160"	25	Darkened	7.5 × 12	5.5h with 160"

Table 12.1. Viewing conditions for the different monitors used in the test

Title	Moving or still	Typical of
Country skiing	Moving	Fast panoramic
Arcade	Moving	Slow panoramic
Noel	Moving	Nearly still
Peter and the Wolf	Still	High contrast
Boy and Toys	Still	High brightness
Kiel Harbour	Still	High detail

Table 12.2. Features of the sequences used in the test

Degree of movement - A further parameter considered was the degree of movement of the pictures, so the investigation was carried out with still and moving pictures separately.

Table 12.2 recalls the main features of the sequences chosen for both tests.

The still pictures were edited in sequence cyclically repeated, with a total duration of 30 minutes. The same was done for the moving pictures.

12.2.3 Performing of the test

Several tests were conducted, each aiming at assessing a specific combination of the parameters under test.

The sample of observers taking part in each test ranged from 15 to 20.

The tests were carried out introducing into the viewing room one observer at a time; in front of the screen a wheeled swivel chair was placed at half the maximum useful viewing distance.

The observer was asked to choose by subsequent approximations, without any hurry, the viewing distance he felt more suitable and comfortable for a long observation of the pictures.

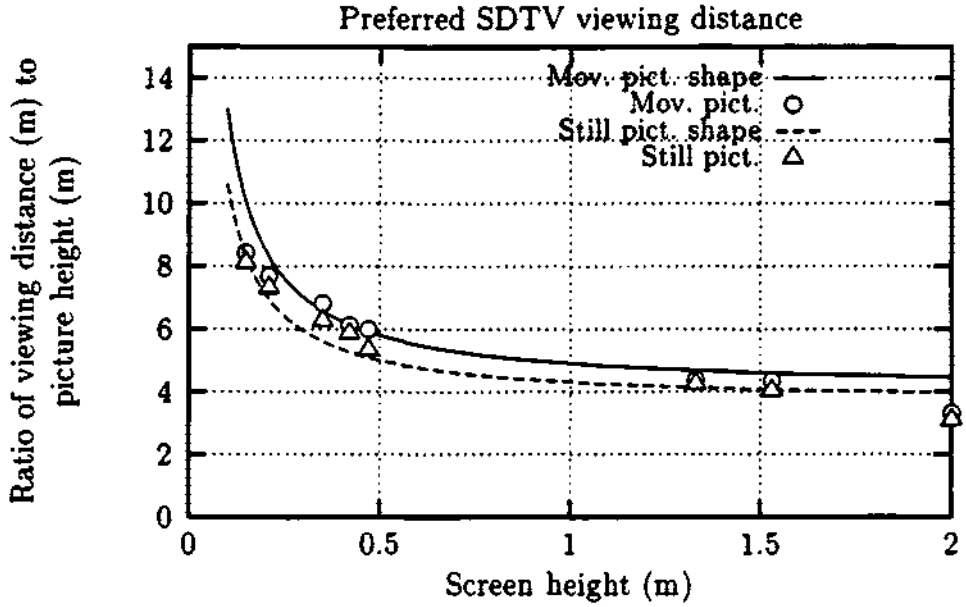


Fig. 12.1. Preferred Viewing Distance for SDTV

At the end of each observation, the distance between the on duty observer's eye and the screen was measured. PVD was then computed on the viewer sample as average value and other statistical parameters.

12.2.4 Analysis of results

The diagram of Fig. 12.1 shows the Preferred Viewing Distance for SDTV still and moving pictures in function of the screen height.

The diagram of Fig. 12.2 shows the value of the Preferred Viewing Distance for HDTV pictures in function of the screen height considering separately moving and still sequences.

The PVD curves for HDTV and SDTV are compared in the diagram of Fig. 12.3.

All the curves which interpolate and approximate the measured values show an hyperbolic shape, decreasing as screen size increases; this applies to moving and still pictures, both in HDTV and in SDTV.

From these results we can conclude that definition and movement degree have little influence on the Preferred Viewing Distance, being the size of the screen (i.e. the screen height "h") the most important factor.

The analysis of the results shows also that PVD tends towards $3h$, corresponding to the Design Viewing Distance for HDTV, only for very large screen, independently of the picture standard.

Results of this investigation agree with other results achieved in other works on the same subject [2].

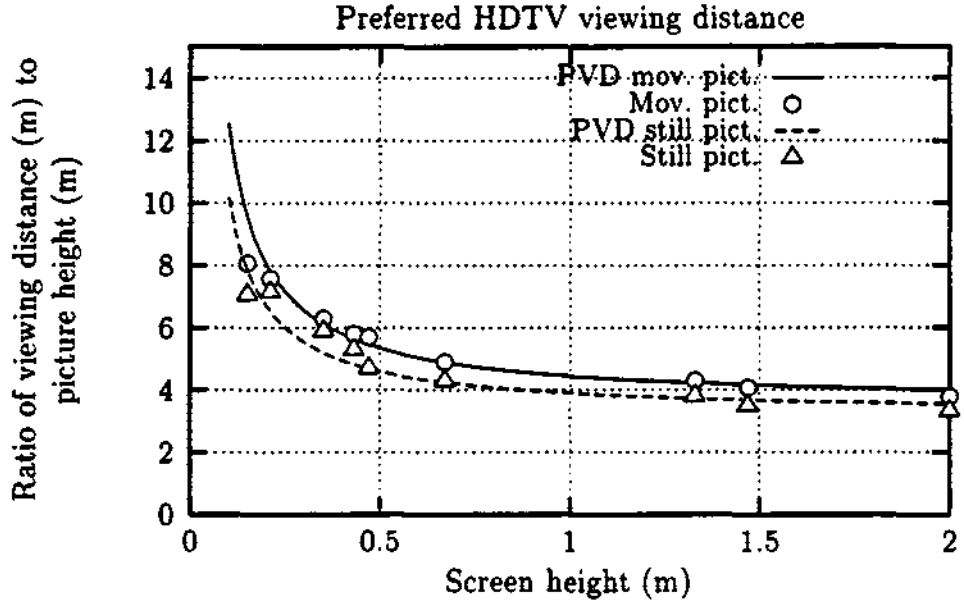


Fig. 12.2. Preferred Viewing Distance for HDTV

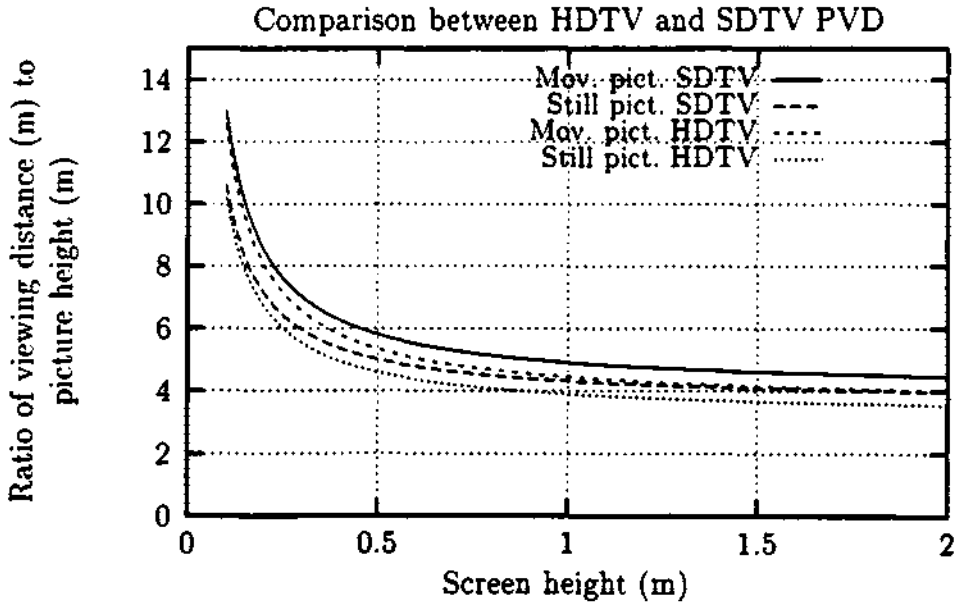


Fig. 12.3. Comparison between Preferred Viewing Distance of SDTV and HDTV

Screen height seems to be the most influential parameter in the definition of PVD.

12.3 Influence of contrast and luminance

The aim of this part of the experiment was to assess the relationship between contrast-luminance and subjective quality of HDTV picture by means of subjective tests [8, 9].

The experiment was divided into four steps.

The first was a preparatory step and it consisted of the measurement of contrast and luminance for the two 38" 16:9 monitors to be used to carry out subjective assessments.

In the second step some selected sequences were processed in order to change their contrast-luminance features in a controlled way. In the third and fourth steps subjective assessments were carried out on contrast and luminance taking into account the previously defined DVD and PVD.

The analysis on contrast and luminance was performed separately.

12.3.1 Preparatory phase of the experiments

The preparatory phase of the experiment intended to determine the specific characteristics of the display to be used for the subjective tests.

The first scope was to correlate the contrast ratio (C) provided by that display in a controlled lighting environment, to the electrical parameters of the video signal. The monitors used for the experiment were two Sony HDTV 38", 16:9 format, with environmental lighting according to ITU Rec. 500. The initial procedure step foresaw to align brightness and contrast of the monitors with the help of the HDTV PLUGE signal [5]. The white peak level was adjusted at 70 cd/m^2.

Hence the contrast ratio provided by the monitors was measured, by using the methodology described in [6] and the relevant test picture showed in Fig. 12.4. The "Contrast Ratio", briefly indicated as "contrast", was expressed as ratio between luminance of the white spot and luminance of the black spots inside the test picture. The measurements were carried out with a high-sensitivity telephoto meter.

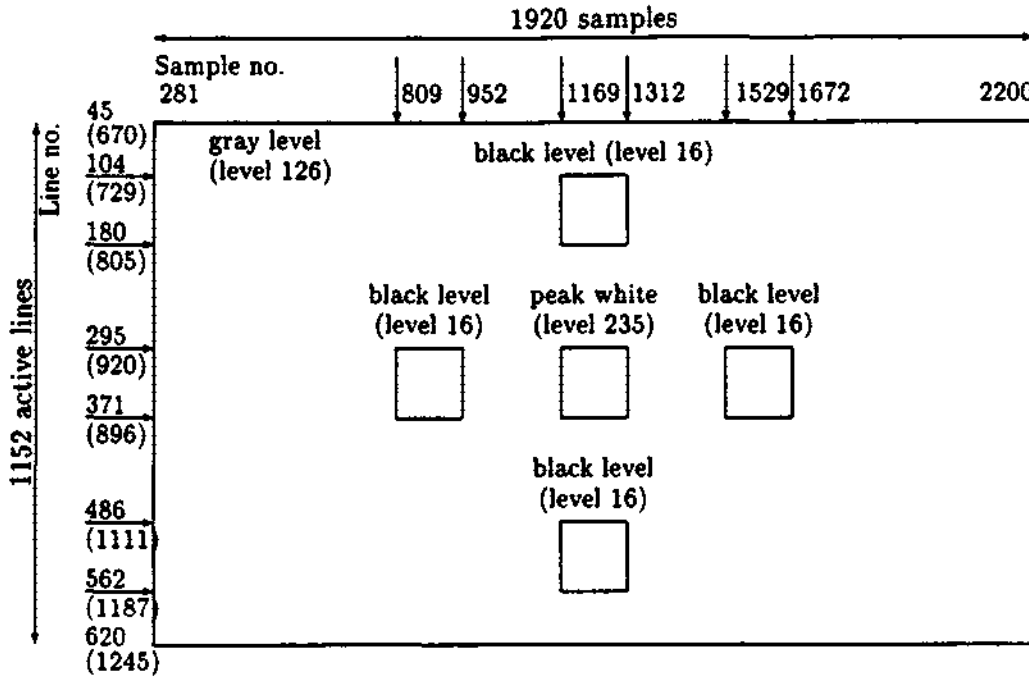
This method permitted to measure the monitors basic contrast, comprehending both the CRT Gamma effect and the environmental luminance effect.

Some modified versions of the signal of Fig. 12.4 [6] were prepared. In the various versions the original 0 mV black pedestal level varied from 50 mV to 300 mV, with 50 mV steps; the white peak level remained steady at 700 mV, whilst the grey intermediate level was adjusted, time by time, at 50% of the white/black difference as per the standard version.

The contrast provided by each one of the modified test pictures was measured, without changing the previously made adjustments of the monitors. This made it possible to correlate the video signal pedestal level to the contrast effectively perceivable on the tested monitors. Results obtained are collected in Table 12.3.

A procedure similar to the previous one allowed to determine the luminance characteristic of the used CRT.

This time too, the test picture of Fig. 12.4 [6] was used and modified by varying the white peak value from 700 mV to 400 mV by 50 mV steps. This time the pedestal



() indicates the second field

Fig. 12.4. Test pattern used for the measurement of contrast ratio

Table 12.3

Black Pedestal [mV]	Contrast Ratio
0	182
50	139
100	79
150	39
200	22
250	13
300	9

Table 12.4

White-peak level [mV]	Luminance cd/m^2	Contrast Ratio
700	70 (100%)	183 (100%)
650	56 (80%)	177 (97%)
600	45 (64%)	172 (94%)
550	35 (50%)	168 (92%)
500	27 (39%)	168 (92%)
450	20 (29%)	160 (87%)
400	16 (23%)	163 (89%)

Table 12.3. Contrast Ratio versus the black pedestal level

Table 12.4. Luminance and Contrast versus the White Peak Level

level was maintained constant at 0 mV, while the grey intermediate level was adjusted, time by time, at 50% of white/black difference as in the standard version. Results obtained are collected in Table 12.4.

As can be observed analysing the results in Table 12.4, the white peak level

variation only slightly influences the contrast: a strong reduction to 23% of the white peak level has little influence on the contrast, quite steady at 89%.

This can be explained considering that the picture black areas have an unwanted emission due to the brightness diffusion of the picture clearer areas. The reduction of the white peak luminance level proportionally reduces the diffusion and the unwanted emission of the black areas.

It is evident that the parameters measured with such a method directly depend on the care, but even on the personal evaluation, the operator has given in aligning the monitors. In fact, the PLUGE signal requires a subjective evaluation for the black level pre-set; the white peak luminance measurements are hence carried out with a luminance meter and a high sensitivity telephoto meter.

This initial procedure allowed the following two findings:

- Firstly, to relate the pedestal levels of the video signal to the contrast really reproduced on the monitor. The video peak level was kept constant, hence the video peak luminance was constant.

- Secondly to relate the white-peak levels of the video signal to the luminance really reproduced on the monitor. The pedestal level was kept constant, and the contrast ratio was roughly constant.

Consequently it was possible to quantify the contrast and the luminance **really perceivable by the viewer** under the different conditions. The monitors were the same used for the subsequent tests and the parameters, measured according to [5], comprehended the effects of the environmental brightness and of the CRT Gamma.

12.3.2 Sequences selection and processing

Some HDTV sequences were selected to be submitted to the electrical levels treatment, so as to reproduce, with real images, the conditions examined in the preliminary phase on the test patterns. The sequences were selected from Annex 4 of ITU Rec. 710-2 [3] and are reported in Table 12.5.

In this specific application the sequences were selected according to their definition and contrast features: namely, pictures having a good detailed content and wholly exploiting the contrast dynamics, were chosen. A particular importance was given to the fact that the signal had the black level well aligned to 0 Volt, with significant luminance peaks not exceeding 0.7 volt, in order to wholly exploit the electrical dynamics of the signal.

The selected sequences were processed to modify the contrast or luminance according to the parameters defined in the preliminary procedure of Section 12.3.1. The pictures, originally available on the digital Quadriga system for HDTV, were dubbed, together with the reference signal consisting of colours bars, on a BTS analogue HDTV VTR. The signal processing was done by using such an apparatus as analogue HDTV source and a BTS analogue HDTV mixer for the processing; the

HDTV selected sequences
Singer (SIN)
Noel (NL)
Geranium (GER)
Olympic ceremony (OCE)
Tennis (TEN)
Winter Olympics (SKI)
Tram (TR)

Table 12.5. HDTV Sequences selected for the test (ITU Rec. 710 Library)

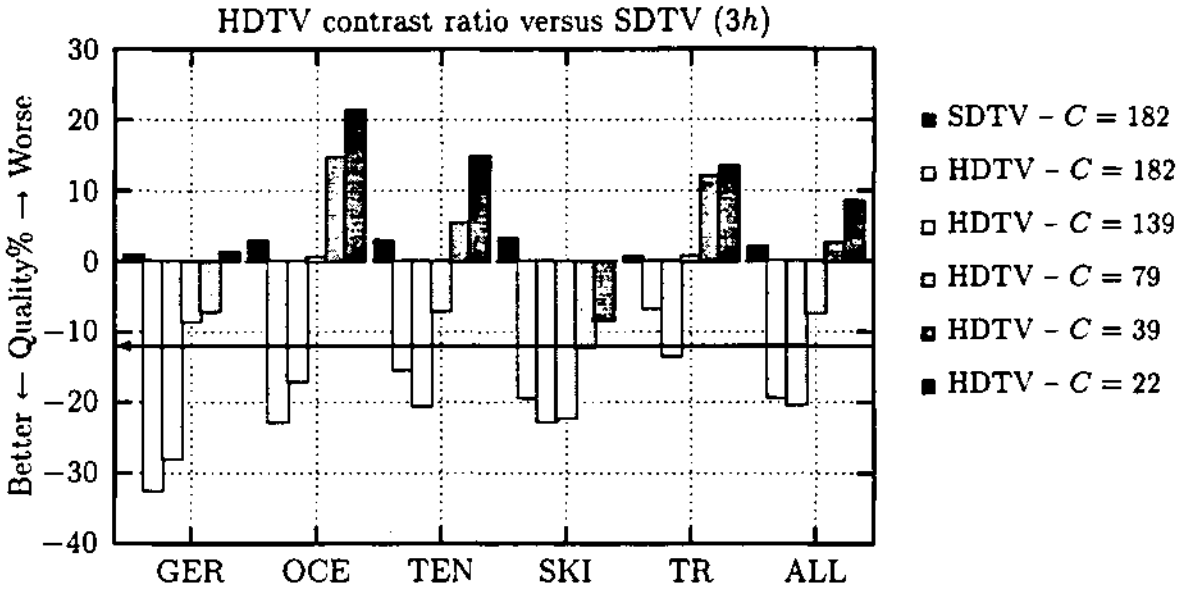


Fig. 12.5. Results of contrast subjective assessments carried out at DVD

processed material was recorded on a Quadriga system based on four Sony digital D1 recorders.

The video peak and pedestal levels were set by the mixer, using the colours bars as reference for the white and black levels.

To make the test both the HDTV pictures and the corresponding down converted pictures were necessary; the down-converted pictures were obtained by using a Snell & Wilcox down-converter. A subsequent up-conversion process allowed to record the SDTV images on the Quadriga system, jointly to the HDTV material.

12.3.3 Results of contrast test performed at DVD

The subjective assessments were carried out using the Double Stimulus Continuous Quality Scale method [8, 9].

In the first test the reference was the SDTV picture quality with full contrast and compared with the corresponding HDTV pictures having different contrast levels. This test allowed to assess the contrast contribution to the overall quality for the HDTV pictures. Results are shown in Fig. 12.5.

The references (SDTV with full contrast ratio) for the single sequences and for the general average, labelled ALL, are well aligned, with a maximum error around 3%. The trend is rather regular for all the sequences and the better quality of the high-contrast HDTV with respect to the reference SDTV is evident. The average improvement on all the sequences (ALL) is 19%, similar value to the results already noted from other investigations.

The subjective HDTV quality clearly decreases when the contrast ratio decreases. The arrow in Fig. 12.5 individuates the so-called "Transparency Threshold", conventionally fixed at 12%, with respect to the reference; for differences included within this value the subjective quality of the system under examination is considered equivalent. On the average of the used sequences (ALL), the "Transparency Threshold" falls between HDTV $C = 139$ and HDTV $C = 79$. The intermediate value corresponds to a $C = 109$, equal to 60% of the maximum contrast, that is to say $C = 182$.

Therefore, on the basis of Fig. 12.5, for the average of the used sequences (ALL), it is possible to state that the subjective quality of the SDTV pictures with full contrast is equivalent to that of the HDTV pictures with a contrast at 60% with respect to the full value. A loss of contrast jeopardises the quality difference between SDTV and HDTV.

12.3.4 Results of contrast test performed at PVD

This second test tends to verify the subjective quality of HDTV pictures at different contrast levels when the Preferred Viewing Distance is adopted instead of the Design Viewing Distance.

For the 38" HDTV 16:9 screens used in the test, with HDTV moving images, it was found PVD is equal to $5.2h$. As PVD is a function of the screen height the results of this second test are specifically linked to the 38" screen size used for the test.

Results of this test are given in Fig. 12.6.

The references (SDTV $C = 182$) for the single sequences and for the general average, labelled ALL, are well aligned, with errors lower than 3.5%. The trend is rather regular for all the sequences and the gradual reduction of the HDTV quality when contrast decreases is quite evident. The subjective quality of HDTV with full contrast is rated not so better than the SDTV reference and results well below the "Transparency Threshold". This means that, at the Preferred Viewing Distance for the used 38" monitors, the observer is not able to appreciate the contribution due to HDTV; the dominant parameter under this condition is the contrast.

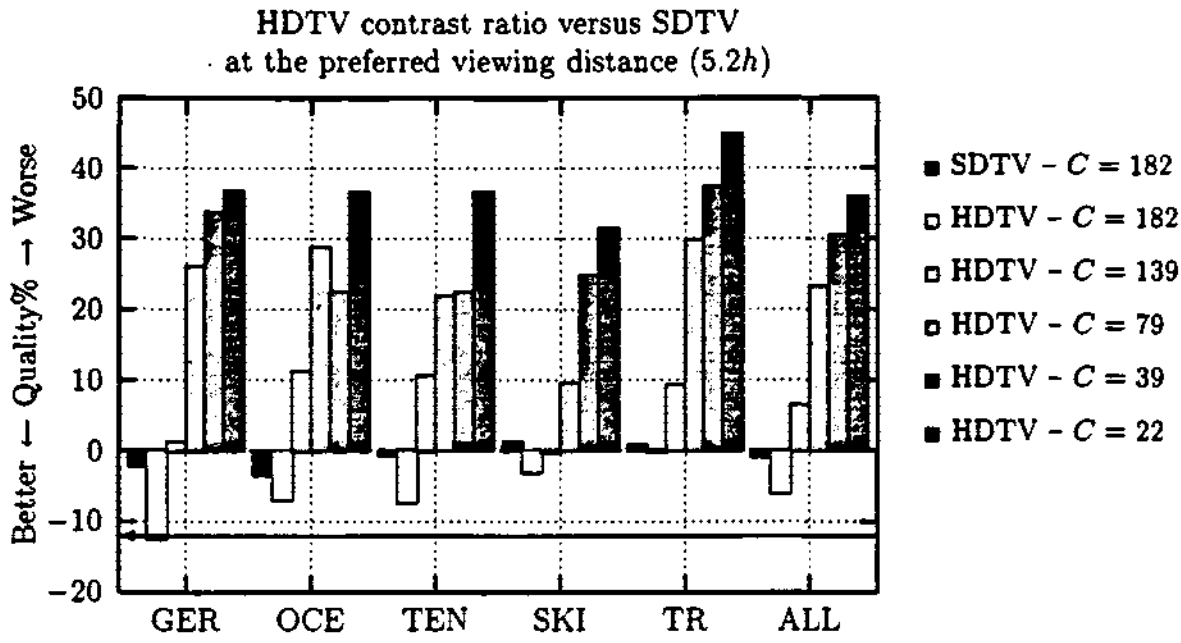


Fig. 12.6. Results of contrast subjective assessments carried out at PVD

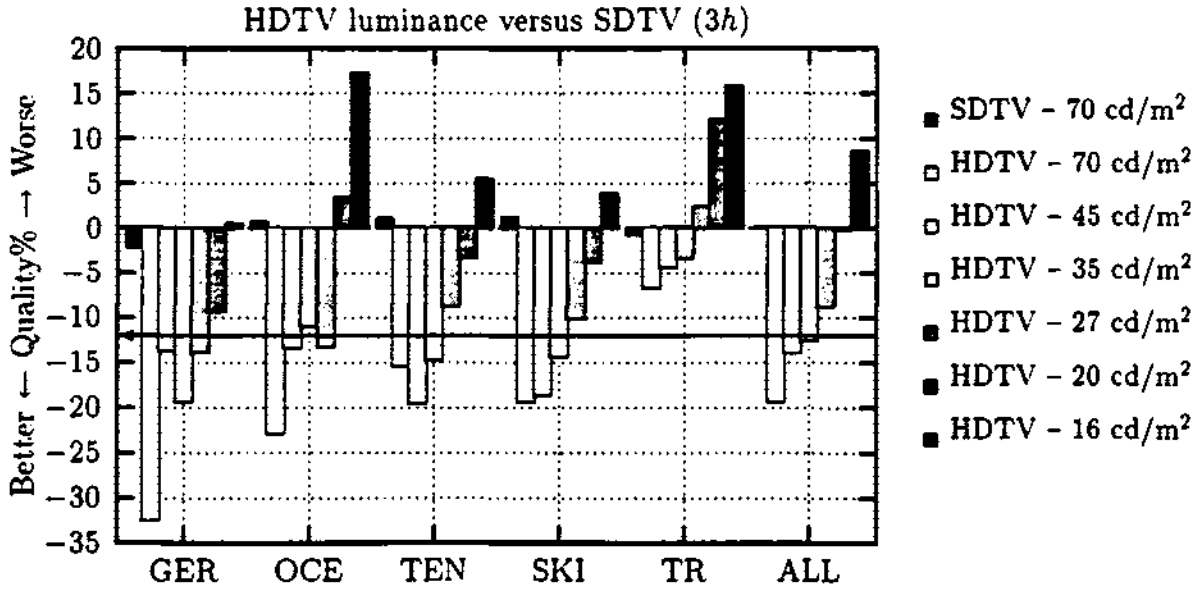


Fig. 12.7. Results of luminance subjective assessments carried out at DVD

On the basis of the results of this test, considering the average of the used sequences (ALL), it is possible to assess that the subjective quality of SDTV and HDTV pictures with full contrast is completely equivalent when Preferred Viewing Distance is taken into account.

12.3.5 Results of luminance test performed at DVD

In this test the reference was the SDTV picture quality with full peak-luminance, compared with the corresponding HDTV pictures having different peak-luminance levels. This test allowed to assess the luminance contribution to the overall quality for the HDTV pictures. Results are shown in Fig. 12.7.

The tested range did not included the HDTV - 70 cd/m^2 peak-luminance case, which in the experiment structure, exactly corresponds to the HDTV full contrast (HDTV $C = 182$) condition.

Both the cases correspond to a regular HDTV signal with full luminance and contrast; this permitted to complete the graph of Fig. 12.7 (HDTV - 70 cd/m^2 values) using the HDTV $C = 182$ data of Fig. 12.5. Both data are referred to DVD.

The graph analysis shows that the references (SDTV with full luminance) for the single sequences and for the general average, labelled ALL, are well aligned, with a maximum error around 2%.

The average on all sequences (ALL) has values and trend rather regular: the HDTV improvement starts from the usual value near 20% and gradually becomes 8% worst than the SDTV when the peak-luminance decreases.

Some sequences show a less regular trend. Three hypothesis have been made:

- A residual error on the subjective measurement.

- A change in the image luminance may alter the "meaning" of the image itself and modify the scores. In the conventional TV and Cinema language, a low luminance image means "evening" or "night". In the absence of a known quality reference, not provided by the DSCQS method, the judgement might be distorted.

- The reduction of the peak-luminance level (and consequently of the average picture luminance APL) might improve the displayed resolution of the image: even with the adopted professional HDTV monitors. A better resolution could give better scores.

The last hypothesis seems to be the more effective.

The arrow in Fig. 12.7 individuates the 12% "Transparency Threshold", for differences included within this value the subjective quality of the system under examination is considered equivalent to the reference. On the average of the used sequences (ALL), the "Transparency Threshold" falls near the HDTV - 35 cd/m^2 value.

Therefore, on the basis of Fig. 12.7, for the average of the used sequences (ALL), it is possible to state that the subjective quality of the SDTV pictures with full luminance (70 cd/m^2) is equivalent to that of the HDTV pictures with 35 cd/m^2. That is to say that a HDTV image with 50% of reduction in the peak-luminance has no quality difference with respect to SDTV.

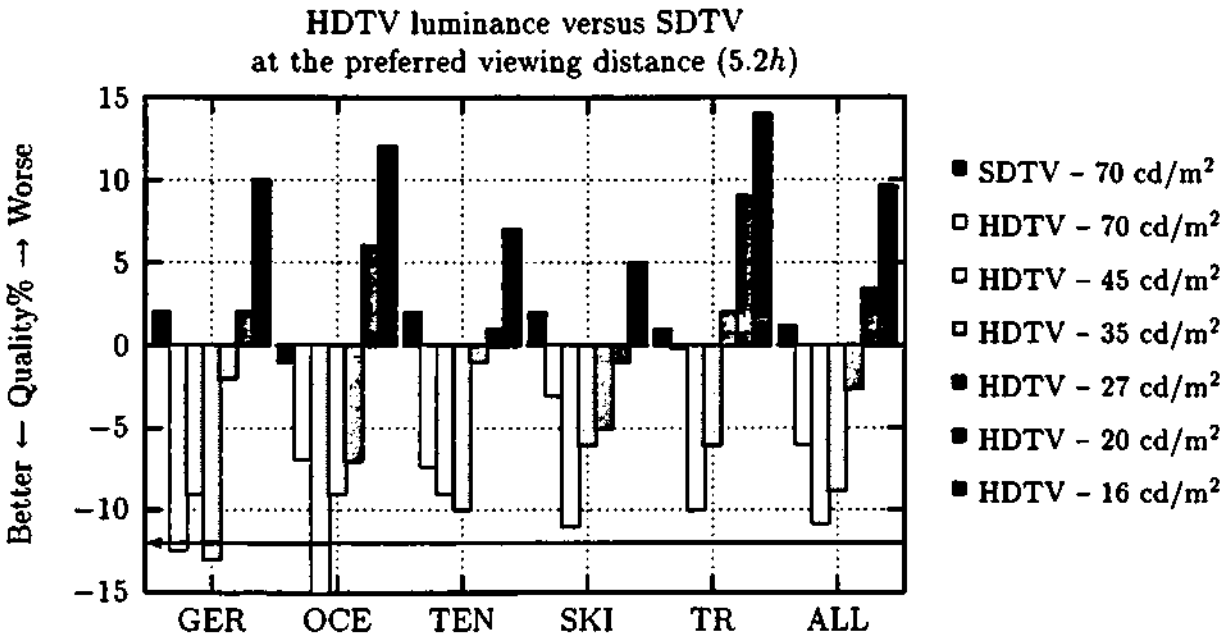


Fig. 12.8. Results of luminance subjective assessments carried out at PVD

12.3.6 Results of luminance test performed at PVD

This last test tends to verify the subjective quality of HDTV pictures at different luminance levels when the Preferred Viewing Distance is adopted, instead of the Design Viewing Distance.

For the HDTV 16:9 38" screens used in the test, with HDTV moving images, was found PVD is equal to $5.2h$. As PVD is a function of the screen height the results of this test are specifically linked to the 38" screen size used in the test.

Results of this test are given in Fig. 12.8.

The tested range did not included the HDTV - 70 cd/m^2 peak-luminance case, which in the experiment structure, exactly corresponds to the HDTV full contrast (HDTV $C = 182$) condition.

Both these cases correspond to a regular HDTV signal with full luminance and contrast; this permitted to complete the graph of Fig. 12.8 (HDTV - 70 cd/m^2 values) using the HDTV $C = 182$ data of Fig. 12.6. Both data are referred to PVD.

In Fig. 12.8, the references (SDTV - 70 cd/m^2) are well aligned with errors lower than 3%.

The HDTV 70 cd/m^2 results are the same as for HDTV $C = 182$ in Fig. 12.6 (as it was said the two conditions are exactly the same): just a little better than the SDTV reference; moreover: other assessed values are quite under the 12% Transparency Threshold, with a very specific trend.

As the HDTV peak-luminance decreases to a 45 cd/m^2 value, the HDTV sub-

jective quality tends to reach 12% of improvement, then the HDTV quality decreases gradually becoming worse than the reference.

This fact enforces the idea, previously examined in Paragraph 12.3.5, that the reduction of the peak-luminance level (and consequently of the average picture luminance APL) might improve the displayed resolution of the image.

In any case, the resolution gives little contribution to the quality that, in most cases, remains under the 12% Transparency Threshold. This means a not appreciable difference of HDTV with respect to the SDTV reference, when the 38" screen PVD is used.

The peak-luminance reduction seems to influence the overall quality less than the contrast.

12.4 Conclusion

The tests carried out confirm that HDTV and SDTV images will maintain their quality if displayed with an adequate contrast ratio level. The luminance has proved to be as well meaningful for the overall quality.

In other words the quality contribution of the higher definition can result totally compromised if it is not supported by appropriate display performance.

Results of the first test, carried out at the Design Viewing Distance, show that in the case of HDTV pictures, a $C = 109$ (60% of $C = 182$) makes the quality difference between HDTV and SDTV not perceivable.

A second test at the Preferred Viewing Distance was carried out. For the 38" HDTV 16:9 screen used in this test, with HDTV moving images, PVD is 5.2, almost twice the DVD.

The results prove that, with 38" screen used in this investigation, the quality contribution, due to higher definition, is completely lost.

In this second test, the subjective quality of the HDTV pictures with full contrast is exactly equivalent to that of the SDTV pictures with full contrast, moreover it rapidly decreases reducing the HDTV picture contrast.

Similar conclusions can be obtained from the third and fourth test on luminance.

From the gathered results it can be concluded that in order to give the user the complete HDTV quality it is necessary to respect the following conditions:

1. **Large viewing screens are necessary. In fact this is the condition beyond which the Preferred Viewing Distance is near the Design Viewing Distance ($3h$).**

2. **In order to maintain a difference of subjective quality between HDTV and SDTV, screens having contrast-luminance parameters comparable to those of the current CRTs for the conventional television are necessary.**

The adopted reference parameters, luminance equal to 70 cd/m^2 and environmental lighting according to ITU Rec. 500, might not seem to represent the viewing in real domestic environment, where usually higher luminance and environmental lighting are adopted.

However, the current technology does not yet make available professional HDTV monitors which give performance, in terms of luminance and contrast, comparable to those adopted for the domestic television. The 70 cd/m^2 luminance level for an HDTV monitor is quite a limit.

References

[1] Investigation on the Preferred Viewing Distance for HDTV Programmes. ITU Doc. 11A/10.

[2] A.M. Lund. The Influence of Video Image Size and Resolution on Viewing-Distance Preference. *SMPTE Journal*, May 1993.

[3] M. Ardito. Studies of the Influence of Display Size and Picture Brightness on the Preferred Viewing Distance for HDTV. *SMPTE Journal*, Vol. 103 No. 8, August 1994.

[4] N.E. Tanton and M.A. Stone. HDTV Displays: Subjective Effects of Scanning Standards and Domestic Picture Size. BBC Report RD, September 1989.

[5] Specifications and Alignment Procedures for Setting of Brightness and Contrast of Displays. ITU Doc. 11/186-E, December 1993.

[6] Specification of a Signal for Measurement of the Contrast Ratio of Displays. ITU Doc. 11/188 E, December 1993.

[7] Investigation on the Preferred Viewing Distance for Video Images. MOSAIC RACE Project CEC R2111 - CEC Deliverable No. R2111RAIDSR014.a1, December 1994.

[8] Method for the Subjective Assessment of the Quality of Television Pictures. ITU Rec. 500–5, September 1992.

[9] Subjective Assessments Methods for Image Quality in High Definition Television. ITU Question ITU-R 211/11, Draft Revision of ITU Rec. BT. 710–1, Doc. 11/194-E.

[10] W. Russel Neuman and Michael A. Kriss. Variables in the Viewing Experience: Preliminary Findings in a Study of Contrast, Resolution and Luminance in Advanced Imaging Systems. MIT Media Lab, June 1991.

[11] J.H.D.M. Westerink and J.A.J. Roufs. Subjective Image Quality as a Function of Viewing Distance, Resolution and Picture Size. *SMPTE Journal*, February 1989.

[12] R. Hamberg. Physical and Perceptual Contrast Issues of CRT Image Reproductions. IPO Report 888, 1993.

[13] J.A.J. Roufs. Brightness, Contrast and Sharpness, Interactive Factors in Perceptual Image Quality. *SPIE Vol. 1077, Human vision, Visual Processing and Digital Display*, 66-72, 1989.

[14] J.A.J. Roufs, V.I.F. Koselka, and A.A.A.M. van Tongeren. Effective Gamma, Global Brightness Contrast and Perceptual Image Quality. *IPO Annual Progress Report* 28, 47-52, 1993.

Chapter 13

The Influence of Audio on Perceived Picture Quality and Subjective Audio-video Delay Tolerance

Samuel Rihs

13.1 The influence of audio on perceived picture quality

When a consumer uses an audio-visual communication system of any kind, his overall perception of the picture quality derives not just from the image, but also from the accompanying sound. However, the importance of such an influence is not very clear. Subjective quality assessments of e.g. digital coding systems are generally carried out separately for the sound and vision path. Consequently, the possible mutual biasing of sound and vision is completely ignored in such assessments.

Within the framework of the MOSAIC project, it was decided to carry out some specific experiments in order to clarify these interrelations.

The objective of an initial experiment was, to find out whether and how much a traditional "soundless" subjective picture quality assessment could be biased by an additional sound programme. Considering the generally higher robustness of the sound path against transmission errors, it was decided to use an unimpaired, high quality (CD) sound programme in combination with picture material of different qualities. Therefore we used an already existing D1 video tape recording with a quite amusing content (humouristic tale) of about 25 minutes continuous duration. The tape was partitioned and re-recorded in 30 sequences of 50 seconds each. Three different picture qualities have been used, uncompressed D1-reference material and data reduced material of 4 Mbit/s and 2 Mbit/s. Each quality appeared randomly but with the same total amount of 10 times on the tape. Every 50 seconds sequence was followed by a 10 seconds pause for the assessment (grey screen with an inserted written instruction). The stereo sound-track of the videotape was a mixing of the original live sound with an appropriate hi-fi background music programme of varying content.

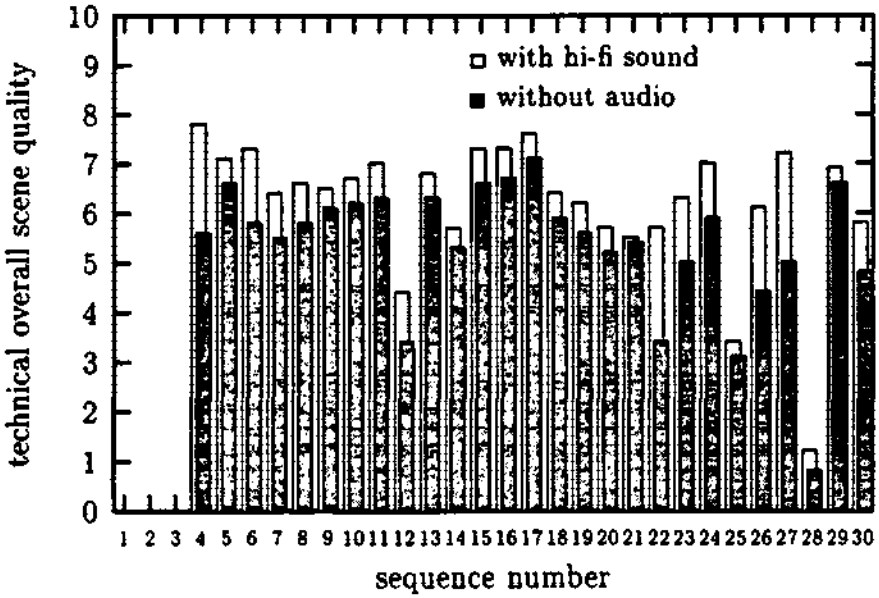


Fig. 13.1. The influence of audio on overall quality. Experiment with test tape "Exam Conditions"; the mean scores in order of presentation are displayed, and the number of votes per sequence is 24.

The tape was presented to a first group of observers without any accompanying sound. Later on, a second group of observers watched the same programme but in combination with the mentioned hi-fi stereo sound-track. All participants were asked to rate the technical overall scene quality immediately after every 50 second sequence (single stimulus method). A quality scale with steps from "0" to "10" but without any adjectival references was applied.

The detailed mean scores of the experiment are shown in Figure 13.1.

As can be seen from Figure 13.1, independently of the picture coding quality or picture content, all scenes were assessed with a somewhat better score for the technical scene quality when the sound-track was additionally switched on!

So, the general conclusion of this experiment can be drawn:

Results of subjective picture quality assessments are generally getting better when the scenes under test are accompanied by a good quality sound programme.

Or in other words:

Subjects ability to detect picture impairments is lowered if a good quality sound programme is added.

Surprisingly, it seems not to be very important what sort of scene or what kind of sound is implicated. The simple fact, that one more of our senses is activated seems to explain the effect.

It is even possible that visible picture quality losses e.g. due to coding artifacts could be subjectively compensated by the covering influence of additional sound.

Consequently, a high quality sound channel combined with a picture channel of reduced quality could be a reasonable solution for a bandwidth saving audio-visual communication system.

The mentioned experiment indicate the positive effect of good quality audio on subjective picture quality assessments. As a consequence, it can be expected that in practice, subjective qualities of picture coding standards (as e.g. MPEG-2) will be somewhat better than assessed, since in real application the picture will generally appear in combination with good quality audio.

The subjective assessment of unimpaired, good quality audio-visual programme material in comparison with impaired one was the subject of some complementary experiments. Therefore, particular impairments of different levels (e.g. low bitrate coding of different qualities) have been introduced in the audio- or video path of several short TV sequences. The sequences were selected such that the modality most important for understanding the material was biased towards the audio, the video or both audio and video.

The participants had to assess the unimpaired and the impaired *"overall transmission quality"* of each scene (assessment method according to ITU-R, Rec. 500, DSCQS).

The interpretation of the resulting quality scores was not always obvious. As the level of an audio impairment may be quite different from that of a video impairment, cross-comparison of impairment levels in audio and video is rather difficult. Nevertheless, some general conclusions could be drawn:

- *Both, audio and video quality is important for subjective assessment.*
- *An audio impairment is more detrimental to quality assessment than a video impairment.*

These studies indicate the complexity of the role audio occupies within a television system. Good quality audio can reduce the effect of impaired video compared with a no-audio situation. Poor quality audio, however, has a pronounced negative impact on quality evaluation, to a greater extent than poor quality video.

These two findings would suggest that *priority should be given to protecting the audio signal in any audio-visual system.*

13.2 Subjective audio-video delay tolerance

An other, special case of audio influence is the relative timing error of sound and vision. It is a well-known problem in the field of audio-visual transmission services. However, increasing digitalisation and therewith the introduction of new data reducing systems have even enhanced the problem.

In the literature we can find diverging recommendations for *"The maximum acceptable relative timing error of sound and vision"*, e.g.: ITU-R Rec. 717 is suggesting a maximum value of 20 ms (sound ahead of vision) and 40 ms (vision ahead of sound)

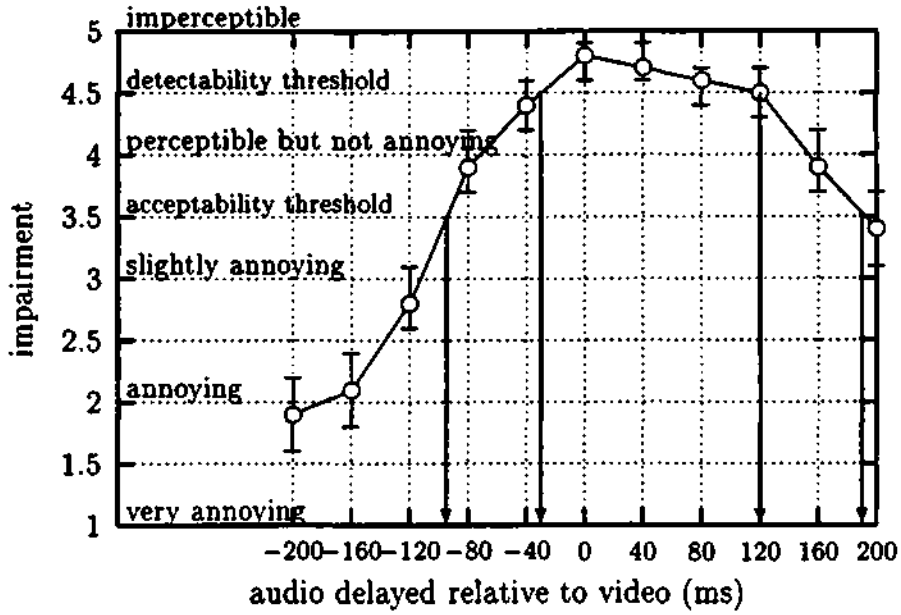


Fig. 13.2. Audio-video delay tolerance. Experiment with ITU 5 grade impairment scale; the mean scores of 18 assessors are displayed, and every point includes 54 votes. The error bars indicate the 95% confidence interval.

for the *"international exchange of programmes"*. EBU R37 on the other hand is recommending maximum 40 ms (sound ahead of vision) and 60 ms (vision ahead of sound) *"at any feeding point to broadcasting transmitters"*. So, depending on the application, the same problem seems to result in different requirements.

However, independently of the application, there must be a certain human *"overall tolerance for the relative audio-video timing error"*.

By carrying out different subjective assessments on the subject matter, it was tried to find out a realistic and practical value for this overall tolerance.

In Figure 13.2, the mean scores of one of the experiments are shown.

If we look at the graph, the following general conclusions can be drawn:

- *An average observer would tolerate overall audio-video timing errors of about 90 ms advanced sound to about 180 ms delayed sound.*
- *The detectability thresholds are situated at about 40 ms advanced sound and at about 120 ms delayed sound.*

The detailed results and conclusions of these experiments were presented at the October 94 session of ITU-R Working Party 11A in Geneva. Furthermore, the results have been discussed in the 11A joint special rapporteur group on *"relative timing of television sound and vision signals"*.

It was decided to carry out additional subjective assessments with groups of observers at different locations and with different languages.

The goal is to propose a single value for *"the overall tolerance of sound-vision timing difference"* at the December 95 session of ITU-R Working Party 11A.

Chapter 14

An Error Model for Digital Broadcast Television Channels

Richard Aldridge
University of Essex

Impairment of picture quality occurs on broadcast television due to interference on the transmission channel. When digital broadcast television is employed this interference causes bit errors. Due to the nature of digital video compression techniques these errors can have effects which affect several consecutive frames, even after the channel interference has ceased. This will result in digital video codec systems exhibiting various failure characteristics which are visibly distinct from analogue television systems, and dependent upon the type of coding algorithm employed. It is therefore necessary to examine the subjective effects of channel errors, in order to determine the choice of codec to use for a given channel and to optimise network planning.

To perform these tests it is possible although not always practical to use a real channel. However, the use of a real channel would lack repeatability between test sessions, (e.g. due to atmospheric conditions, alternative routing, etc.) and does not provide sufficient flexibility and control for the investigation, i.e. the ability to set the bit error rate, the type of distortion, etc.. Consequently it is preferable to simulate the effects of error behaviour by implementing a statistical model.

14.1 Proposed model

One important phenomenon about digital channels is that errors tend to occur in bursts. The reason for this is not clear but its source is probably a combination of a range of factors, e.g. electrical interference from inductive devices, fading characteristics of propagation, impulsive electrical transients, etc.. By far the simplest model proposed in the past which takes account of this burstiness is that presented by Gilbert [1].

The basis of the Gilbert model is relatively straightforward, comprising a two-

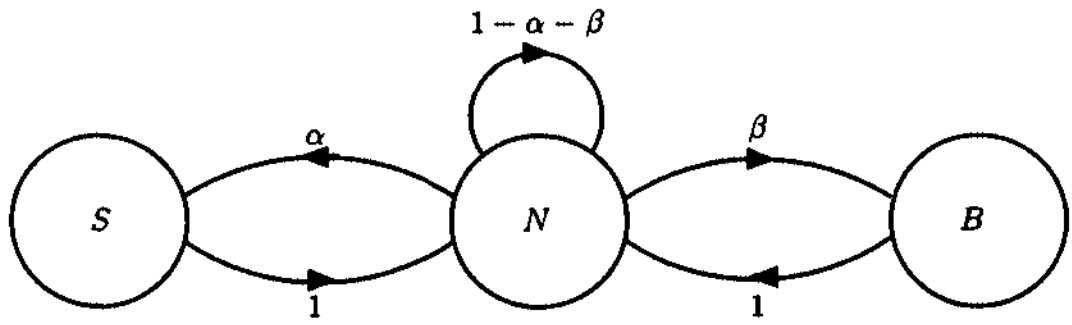


Fig. 14.1. Proposed model represented by Markov chain

state Markov process. The process is either in a error-generating (or 'bad') state or a non-error-generating (or 'good') state. Bursts and gaps are created randomly and both burst and gap lengths possess exponential distributions. However, experimental observations have already rejected a geometric distribution for error burst lengths, [2], and have suggested in some cases that burst lengths are Poisson distributed which can be approximated by using the Neyman Type A model [3]. The Neyman model was originally utilised in the area of bacteria growth but has since been successfully adapted for modelling errors on certain digital links [4, 5].

The proposed solution therefore takes heed of both these models. It assumes that error burst lengths are Poisson-distributed and error-free gaps, are by the nature of the model, geometrically distributed. A distinction is made between single bit errors. Hence the term burst will be used from here on to denote errors of two or more consecutive bits. Gaps can be of any length, between single bit or burst errors.

14.2 Markov representation of the proposed model

The proposed model can be represented by a three-state Markov process as illustrated in Figure 14.1. The three states are defined as state N which has no error, state S which has a single bit error and state B which has a burst of errors. The transition from state N to S occurs with probability α. Since there is only one error possible in this state the probability of returning to state N is 1. Similarly, the transition from state N to B occurs with probability β. After any burst the probability of returning to state N must also be 1.

14.3 Calculation of parameters

If the state probabilities for the Markov process of Figure 14.1 are P_n, P_s and P_b respectively, then it is obvious that:

$$P_n + P_s + P_b = 1 \tag{14.3.1}$$

Since at equilibrium the following is true:

$$P_s = \alpha \cdot P_n \tag{14.3.2}$$

and

$$P_b = \beta \cdot P_n \tag{14.3.3}$$

then substituting (14.3.2) and (14.3.3) into (14.3.1) gives:

$$P_n = \frac{1}{1 + \alpha + \beta} \tag{14.3.4}$$

If we assume the error burst lengths are to be Poisson distributed, the probability of having a burst of k errors, P_k, is given by:

$$P_k = \frac{\lambda^k}{k!} e^{-\lambda} \qquad (k \geq 2) \tag{14.3.5}$$

Normally λ would be the mean of this distribution, (i.e. the mean burst length) for the range $(0 \leq k \leq \infty)$. However, since we are in state B, P_k is zero for $k = 0$ or 1 as state B only accommodates bursts of at least two bits in length. Thus the mean is shifted up from λ. In other words when we are in state B the modified mean burst length λ' refers to the mean of all bursts 2 bits or greater in length. Subsequently we must use a normalised form of this Poisson distribution to obtain a normalised set of probabilities P'_k as follows:

$$P'_k = P_k \left\{ \sum_{k=0}^{k=\infty} P_k \right\} / \left\{ \sum_{k=2}^{k=\infty} P_k \right\} \tag{14.3.6}$$

which yields a modified Poisson distribution:

$$P'_k = \frac{1}{1 - e^{-\lambda}(1 + \lambda)} \cdot \frac{\lambda^k e^{-\lambda}}{k!} \qquad (k \geq 2) \tag{14.3.7}$$

where

$$\sum_{k=2}^{k=\infty} P'_k = 1 \tag{14.3.8}$$

This allows calculation of the modified mean λ' as:

$$\lambda' = \frac{\lambda(1 - e^{-\lambda})}{1 - e^{-\lambda}(1 + \lambda)} \tag{14.3.9}$$

Now since on average there is a single error in state S and λ' errors in state B, then the overall mean bit error rate P_e is given by:

$$P_e = P_s + \lambda' \cdot P_b$$

Again substituting for P_s and P_b from (14.3.2) and (14.3.3) gives:

$$P_e = \alpha \cdot P_n + \lambda' \cdot \beta \cdot P_n$$

and substituting for P_n from (14.3.4) provides:

$$P_e = \frac{\alpha + \beta \cdot \lambda'}{1 + \alpha + \beta} \tag{14.3.10}$$

Consider now the process for the error-free gaps from Figure 14.1. Given we are in state N, the probability of having a gap length k, G_k, is equal to having $k-1$ consecutive 'non-errors' followed by a single or burst error, so:

$$G_k = (1 - \alpha - \beta)^{k-1} \cdot (\alpha + \beta) \tag{14.3.11}$$

Thus the distribution of gap lengths is inherently geometric from this Markov model and the mean gap length of this distribution, m_g, can easily be shown to be:

$$m_g = \frac{1}{\alpha + \beta} \tag{14.3.12}$$

Hence from (14.3.10) and (14.3.12) it is possible to express α or β in terms of P_e, m_g and λ'. Thus knowing α and β, P_n can be evaluated from (14.3.4), as can P_s and P_b from (14.3.2) and (14.3.3). Therefore only three parameters need to be found to simulate the model in full - P_e, m_g and λ' - all of which can be measured from live channels.

14.4 Conclusions

A model has been proposed which can model the errors on typical digital transmission channels. In particular the typical bursty nature of errors is accommodated. Single bit errors are treated separately from error bursts, should their behaviour need to be treated differently. The model then only needs three measured parameters for simulation.

References

[1] E.N. Gilbert. Capacity of a Burst-Noise Channel. *Bell System Technical Journal*, September 1960, 1253-1265.

[2] J. Bito, J-R. Ohm, and P. Noll. A Simple Model for the Loss Process in the Cell Stream of Variable Bit Rate Video Sources. Institut für Fernmeldetechnik, Technische Universität Berlin, 1993.

[3] J. Neyman. On a New Class of "Contagious" Distributions Applicable in Entomology and Bacteriology. *Annal. Math Statistics*, 1939, Vol. 10, 33-57.

[4] D. Becam, P. Brigant, R. Cohen, and J. Szpirglas. Testing Neyman's Model for Error Performance of 2 and 140 Mbit/s Line Sections. *Proceedings IEEE International Conference on Communications 1984*, Vol. 3, 1362-1365.

[5] D. Becam and P. Brigant. Poisson and Neyman Models Applied to Errored Secondson Digital Transmission. *Proceedings IEEE International Conference on Communications 1986*, Vol. 2, 1199-1203.

Chapter 15

Acceptability of Recovery Time for Channel Hopping and Transmission Breaks

Éric Bourguignat and Thierry Alpert
CCETT

The recovery time is clearly defined in ITU-R Report 1211 as the "time taken by the system to recover from gross errors causing a complete loss of signal to the decoder". The "gross errors" are not the single reason for temporary signal loss. The channel hopping is certainly more frequent and induce the same type of artifact, specially if it corresponds to a switch between two programmes which are not in the same transmission channel. The user requirement, specifying the maximum acceptable recovery time, is reported for contribution and primary distribution in the new ITU-R Recommendation 800: 160 ms when the interruption is longer than 50 ms. On the one hand, none of the codecs formally tested by the ITU-R were able to fulfill this specification and, on the other hand, there is no clear, technical or psychophysical, justification of this value.

In order to clarify the EBU position in this context, a Special Rapporteur has been appointed by the sub Group V1. Due to the need for original experiments based on new procedures and to obtain statistically significant results with several laboratories, a working group has been set up by the Special Rapporteur in collaboration with the MOSAIC RACE project to collect all the related information and to organise appropriate studies.

As explained subsequently, the acceptability of the impairment induced by the recovery time could not be measured with standard subjective procedures as those described in ITU-R texts. Besides, the range of recovery time parameters around which the measures should be set up was unknown. As a consequence, the study was carried out in a recursive way in order to solve at a given step the problems discovered during the former step when trying new procedures or impairments characteristics.

After a short analysis of the behaviour of the current equipment, a preliminary

TV-Set	"hopping behaviour"	Delay in seconds
ITT NOKIA Digivision 6381 PIP	Black, then "cut in"	0.3...1.0
JVC AV 25	Black, then fade in"	fix 0.9
JVC AV 25F1EG	Black, then fade in"	fix 1.0
LOEWE 63 SAT	Black, then "cut in"	0.5...1.0
PANASONIC VHS-Rec NV-FS 100	Black short, then vertical "tumble in"	0.3...0.5
PHILIPS 25ML 8300	Black, then "cut in"	fix 0.9
SONY KV A 2S21	Black, then fade in"	fix 0.8
SONY KV X 2151	Black, then fade in"	fix 0.8
SONY KV FX29 TD (100 Hz)	Black, then fade in"	fix 0.8
SONY KV MD14D	Black, then "cut in"	0.3...0.8
THOMSON TD 5550 PG5	Black short, then vertical "tumble in"	0.3...0.5

Table 15.1. 'Hopping' delays.

experiment was organised to explore the subject. The conclusions allowed a first series of measures in the different laboratories to evaluate the acceptability of recovery time in case of transmission breaks and channel hopping. Owing to the results, an acceptability threshold was defined for the recovery time in the context of channel hopping. On the contrary, it was still not possible to conclude in the transmission breaks context, so, a new experiment was decided and defined which was carried out by four laboratories. The final accurate conclusions required nevertheless a last test organised in two laboratories to explore in details the significant part of the results of the former series.

15.1 Objective measurements

15.1.1 Measures on analogue systems

Before beginning the work on digital Codecs, the Swiss PTT laboratory has started its study by measuring the channel hopping delay on different today used home TV. All of them produced channel hopping delays significantly higher than the 160 ms, the next channel being displayed after a switch to a black screen during the change. Complete results are provided in Table 15.1.

It has to be noted that these "hopping" delays are well accepted by the general pubic.

15.1.2 Measures on digital Codecs

These measures were made at VURT with two types of 34 Mbit/s Codecs from Thomson and RE. Breaks with different durations (20 ms, 0.76 s and 1.6 s) were introduced at the transmission level (before decoding) or at the video level (before coding) and the recovery times were measured on the digital video output. The results

TV Codec	R. F. Breaks	Video breaks
Thomson-CSF TER 8520/21	2.86	3.06
RE Technology 3400 - ETSI / R.T. = 1.6 s	1.62	4.15
RE Technology 3400 - ETSI / R.T. = 0.35 s	0.5	

Table 15.2. Recovery time on 34 Mbit/s Codecs.

are provided in Table 15.2, in each box, the reported recovery time represents the mean over several measures made with the three break durations. The means could be made because there was no correlation between break duration and recovery time. The Codec from RE Technology was provided with two recovery time adjustments, so, both were measured.

Measures were also made at Swiss PTT on the Thomson equipment. The results are more or less similar since the induced recovery time is less than 2 seconds for the R.F. case and less than 5.5 seconds for the Video case.

Even if different recovery times were measured depending on the location of the break in the transmission chain and the engineering of the equipment, it is nevertheless clear that none of the Codecs were able to comply with the ITU-R Recommendation. The task of the Special Rapporteur was mainly concerned with the R.F. breaks, the video breaks having very different temporal characteristics since they appear in the studio.

15.1.3 Digital processing involved in Recovery time

After a complete loss of synchronisation, a series of step are necessary to recover the correct picture at the output of a digital picture transmission. Each step corresponds to a specific processing:

- Recovery of the transmission synchronisation
- Multiplex
- Access control
- Recovery of the video clock
- Synchronisation of the FEC
- Recovery of the video data

A complete analysis of each item of this list could give an idea of the impact of each one on the recovery time. It was not a task of the group but a quick reading of the list makes obvious the possible marginal influence of the picture coding algorithm. As a matter of fact, the other aspects like recovery of the transmission synchronisation or access control may be time consuming enough to exceed the now well known 160 ms.

15.1.4 Comments

The objective measurements were certainly not the main task of the group but they bring information improving the context of the question.

First of all, the situation is not so clear with the analogue TV service. The ITU-R Recommendation is not acceptable for these current systems but the users seem to accept the channel hopping time that they produce. It is therefore probably possible to increase the current Recommended value without any disturbance on the service.

Many steps of the transmission chain being implied in the specification of the recovery time, the influence of the video coding should not be considered independently, except if the influence of the others is accurately quantified. This is confirmed by the behaviour of the analogue systems which include several processing needed for the digital ones.

The available digital Codecs are not as far from the Recommendation as it was generally thought. In fact, a special adjustment of one of the Codec yields a recovery time value (0.5 s) which is close enough to the Recommended one to hope that a solution satisfying both perceptual and hardware aspects may be found.

15.2 Preliminary study at CCETT

The first study carried out at CCETT aimed at measuring the acceptability of the recovery time for channel hopping. The channel hopping was chosen in a first step because it is, a priori, more frequent than the transmission break.

15.2.1 Description of the experiments

The experiments were using ITU-R procedures: ratio scaling and double stimulus continuous quality scale. A DCT video Codec was used for the three experiments in order to take into account the possible influence of the picture quality of the Codec but there was neither access control nor complete multiplex and the synchronisation was never lost.

In the first experiment, using the "ratio scaling" procedure, the channel hopping between two programmes was interactive but the duration of the recovery time was random. There was one session per observer. During the recovery, the frozen last frame was displayed. Observers had to assess the quality of the channel hopping.

The second experiment was based on a recorded channel hopping with the same psychophysical procedure. With this procedure, the channel hopping happened at the same time for all observers.

The last experiment aimed at comparing the perception of impaired channel hopping for different behaviours of the displayed picture during the event. The test was still based on recorded tape and the Double Stimulus Continuous Quality Scale procedure was applied. Four behaviours were compared: grey frame, frozen frame, frozen field and uncorrelated still picture.

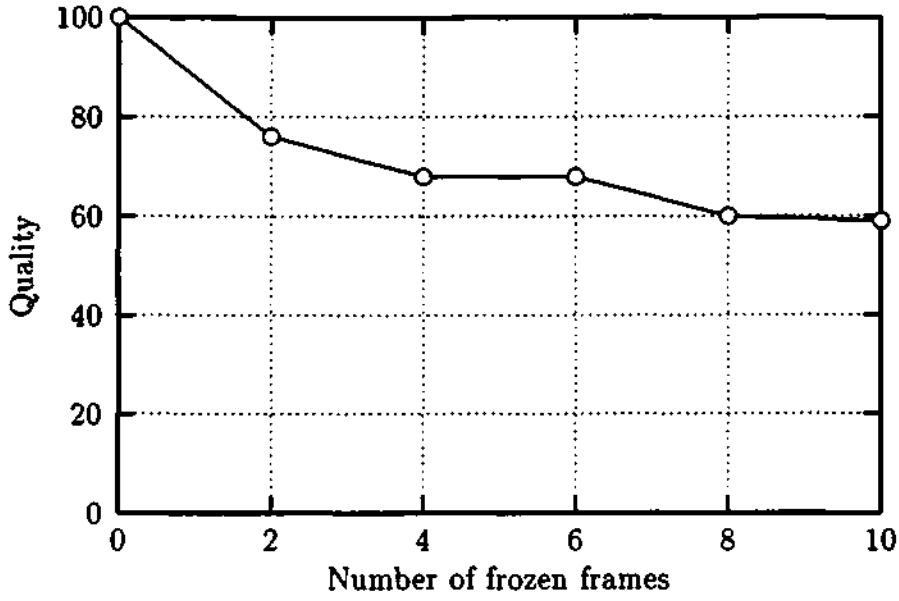


Fig. 15.1. Perception of the channel hopping time, CCETT.

15.2.2 Results

The results of the first experiment, relatively unstable, indicate that the 25% of the scale (ITU-R acceptability criterion for error performance), for the difference between the "ideal" channel hopping and the impaired one, are reached as soon as the number of frozen frames is higher than two. It was nevertheless difficult to analyse the results due to the unstability related, notably, to the content of the sequences just before the channel hopping.

The results of the second experiment, Figure 15.1, with imposed simulated channel hopping, are very close to those of the first one and the relationship between the motion in the sequence before the channel hopping and the rates becomes obvious.

The result of the third experiment is a ranking order of different behaviours of the displayed picture during the channel hopping: frozen fields, frozen frame, grey frame and, the worse, the uncorrelated picture. Nevertheless, only the uncorrelated picture is noted significantly different of the others (Student analysis).

15.2.3 Comments

Using the current ITU-R procedures, inducing a comparison, implicit or explicit, between a reference and some impaired version of the event and defining the acceptable value on the basis of the ITU-R usual criterion, the channel hopping delay is almost never acceptable. The observers build their assessment on a visibility threshold basis and not on a subjective perception one. As soon as the recovery time be-

comes visible it is unacceptable. It has been assumed by the group that the normal conditions in which the channel hopping is made at home should induce different behaviours and results. The upper results, yielding an acceptability threshold at about 80 ms, cannot be accepted without confirmation since they are so far from the current values on analogue systems which seem to be accepted.

In addition with new experiments on channel hopping, first experiments on transmission break acceptability have to be set up. The main difference between both events is the unpredictable aspect of the transmission breaks.

15.3 First series of measurements organised by the group

For this first series of experiments, each laboratory organised its own work freely, after discussion within the group. Due to the lack of experience of the members for this type of measurement, it was too early to define any common procedure.

15.3.1 Experiments at IRT

The study was oriented towards the perception of the recovery time after transmission breaks. A first expert viewing confirmed that the display of a frozen picture during the recovery time is preferred rather than a grey one, the experiment was therefore carried out on this basis.

The experiment was built from a DPCM 140 Mbit/s transmission through a mutable glass fibre. Five different critical five seconds sequences were used for a ratio scaling test during which observers had to assess the quality of the sequences when a transmission break, associated with different recovery times, was introduced.

The results, Figure 15.2, indicate as soon as the recovery time reaches one frame (40 ms), the difference with the reference (without any interruption) is about 40% of the scale. Taking into account the ITU-R criterion, less than 25% of the scale for error performances, 40 ms is already not acceptable in the conditions of the test. This result confirm the former ones from CCETT.

15.3.2 Experiments at VURT

After a short expert viewing confirming a clear preference for the frozen picture to the grey or black ones for displaying during the recovery time, the subjective perception of the transmission breaks was measured.

The study began with the measurement of the "quality" of a single transmission break. Due to the available equipment, a clean black picture was displayed during the break and the simulated recovery time. Expert observers used a five grade quality categorical scale in a double stimulus procedure experiment to assess the quality of the presented recovery time introduced in 30 s. sequences.

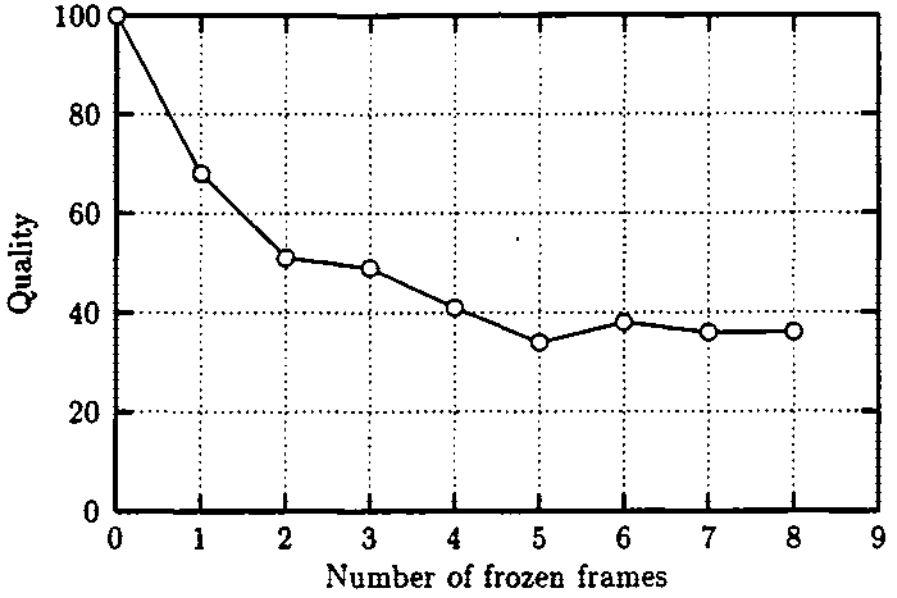


Fig. 15.2. Perception of channel hopping, IRT

The results, Figure 15.3, indicate a 100 ms recovery time corresponds to the 25% of the scale. The difference with former data is probably due to the fact that the observers were experts.

The second part of the study aimed at measuring the subjective perception of several recovery times at different break frequencies. The experimental arrangement was the same as for the single transmission break except for the sequence duration which was one minute. Using again the same 25% criterion, the results provide the subjectively acceptable period (AcP) between two breaks for a given recovery time (RT), for example:

- RT = 20 ms ⇒ AcP = 5 s
- RT = 200 ms ⇒ AcP = 15 s

Taking into account the sequences duration, it is assumed the acceptability measurement of periods longer than 15 s cannot be considered as valid.

15.3.3 Experiments at Swiss PTT R&D

It is now the channel hopping time which is considered. In the following experiment, audio and video were interrupted simultaneously when a channel hopping was requested by the observers. A black or a 50% grey picture was displayed during the channel hopping time the duration of which being under the control of the observers.

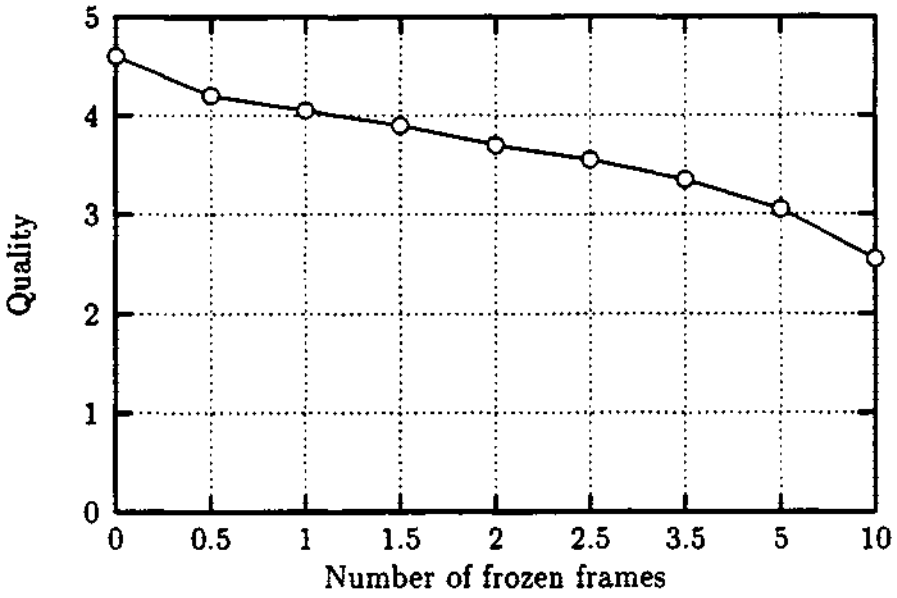


Fig. 15.3. Results from VURT

The experiment ended when the observer, one at once, decided he had found the upper acceptable limit of the channel hopping time. The mean result over the 27 participants is 800 ms.

The last experiment aimed at measuring the acceptable channel hopping time when the observer is not informed of the actual goal of the test. In a session, the observer was therefore given the task of selecting his preferred programme among several ones. Then, some questions were asked him and, among them, one was inquiring about the comfort of the equipment. The hopping time was then increased at each session and the acceptable value was the last one before he complained of the "hopping" response.

The mean result over the 21 participants of the second experiment is 890 ms, as reported on Table 15.3, and the grand mean over both experiments is 840 ms for the acceptable channel hopping time.

15.3.4 Experiments at CCETT

The experiments also addressed the channel hopping time. The first experiment was the same as the second one in Swiss PTT laboratory, with an hidden actual objective. The mean result over the 19 participants is 658 ms. Complete results are reported in Table 15.4.

In the second experiment, observer was requested to switch through several programmes and, then, he was requested to use an 11 grades categorical scale in order to evaluate the quality of the hopping response. The annoyance seems to begin beyond

"Channel hopping" experiment

Number of observers at a given accepted time	Accepted time
2	500
2	600
1	700
6	800
6	1000
4	1200
Mean value:	890 ms
Standard deviation:	230 ms
95% confidence interval:	790 - 990 ms

Table 15.3. Swiss PTT results

"Channel hopping" experiment

Number of observers at a given accepted time	Accepted time
2	400
4	500
5	600
3	700
3	800
1	1000
1	1200
Mean value:	658 ms
Standard deviation:	201 ms
95% confidence interval:	568 - 748 ms

Table 15.4. CCETT results

500 ms but the slope of the curve is not hard enough to conclude. The middle of the curve is reached for about 650 ms (see Figure 15.4).

In a last experiment, the task was the same as the former one except for the scale which was a graphic one with the criterion "acceptable" reported in the middle of it. The mean result, corresponding to the rate 5 and the adjective "acceptable", over the 35 participants is 650 ms. The curve based on the results of the last two experiments

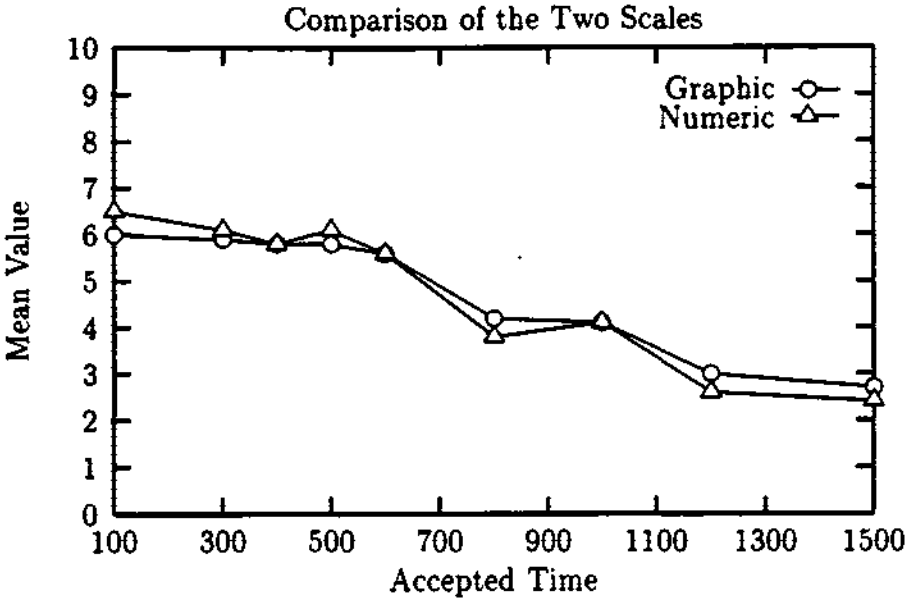


Fig. 15.4. Comparison of CCETT results

are then almost identical (see Figure 15.4).

15.3.5 Comments

Concerning the channel hopping, a conclusion can be drawn from the experiments described above. The results from Swiss PTT R&D and CCETT indicate the channel hopping time, corresponding to the global recovery time of the receiving equipment, may be higher than 160 ms. The assumed acceptability threshold of the recovery time, in case of channel hopping, has been calculated from both similar tests in Swiss PTT and CCETT experiments (second test at Swiss PTT and first test at CCETT). The results of these tests were confirmed by the two last experiments at CCETT. The statistical criterion selected to specify this threshold is the probability (95%) to satisfy 75% of the observers. In order to calculate this value, we need to know the rate corresponding to the acceptability threshold of the impairment. If we make a comparison between the two last experiments at CCETT, it can be noted that the behaviour of the numerical and the graphic scaling are very similar (see Figure 15.4). This leads to an extra result, which is that, at least for this type of impairment, note 5 on the 0 to 10 scale corresponds to the "acceptable" criterion. **On this basis and using the statistical criterion described upper, the acceptability threshold for channel hopping time is 550 ms.**

On the contrary, the conclusion on the recovery time limit in case of transmission break may not be drawn easily. The quality measurements addressing a single event in a short sequence induce a very restrictive answer since the one frame interruption

(40 ms) is already unacceptable. Nevertheless, on operational networks, the major disturbance frequency is so small that the results based on short sequences cannot lead to realistic conclusions. The VURT experiment addressing the problem for longer sequences and several events is therefore specially interesting and has to be completed to reach realistic figures, at least several minutes between two transmission breaks. These VURT results may already be interpreted as a good tolerance to the long recovery time when the event frequency decreases, we can even foresee that for several minutes periods the acceptable recovery time could be higher for the transmission breaks than for the channel hopping.

15.4 First common experiment

To conclude on the transmission break aspect, it was proposed to organise a new experiment in order to measure the transmission break frequency beyond which the recovery time corresponding to the maximum channel hopping time (550 ms) became subjectively unacceptable. Assuming this threshold frequency should be reasonable, from the VURT results, the recovery time tolerance would be specified by means of two numbers: maximum channel hopping time and maximum transmission break frequency.

15.4.1 Description of the procedure

During the discussion about the organisation of this new test, the dependence of the video recovery time acceptability on the audio recovery time arose. To evaluate the relative effect, IRT accepted to provide all implied laboratories with a D1 tape containing long sequences with a given video interruption (520 ms) and different audio ones. The conclusion of these expert sessions was that the audio influence on the global perception of the break is quickly higher than the video one. It has therefore been decided to carry out the next experiment with and without audio breaks in order to take into account its possible influence. The audio breaks were nevertheless kept short (40 ms) in order to avoid complete masking of the video effect by the audio one.

The experiment was finally organised as follows. Long (5 min) sequences were extracted from 30 min programmes in order to avoid observers' worry. Then, six break frequencies were applied to the sequences (720, 360, 180, 60, 12 and 0 per hour). As said above, audio breaks duration was 40 ms and video one 560 ms (14 frames). At the end of each sequences, the observers had to assess the global quality of the sequence on an 11 grades quality scale. Identical instructions were read to the observers at the beginning of the session.

	Freq	VURT 15-20 obs.	IRT 13 obs.		SWISS PTT 18 obs.		CCETT 15-25 obs.		Grand mean	sem
		Ballet	Hockey	PTT	Hockey	PTT	Concert	Reporting		
	0			7.6 (0.48)		7.7 (0.39)	8.4 (0.25)	8.6 (0.23)	8.08	0.34
A	12	8.5 (0.17)					8.5 (0.23)	8.0 (0.18)	8.30	0.19
	60	8.3 (0.17)					7.3 (0.19)	8.0 (0.25)	7.88	0.20
	180	7.9 (0.16)					7.0 (0.28)	8.0 (0.22)	7.64	0.22
	360	7.5 (0.24)					6.6 (0.29)	7.5 (0.26)	7.21	0.26
	720	7.6 (0.26)					4.0 (0.37)	7.3 (0.21)	6.28	0.28
V	12	7.6 (0.24)	4.7 (0.43)		6.6 (0.51)		7.7 (0.22)	8.0 (0.21)	6.91	0.32
	60	6.5 (0.38)	2.7 (0.38)		5.6 (0.46)		6.1 (0.29)	6.6 (0.23)	5.49	0.35
	180	5.4 (0.39)	3.1 (0.40)		4.9 (0.36)		4.8 (0.26)	5.0 (0.27)	4.64	0.34
	360	5.1 (0.42)		2.2 (0.43)		3.2 (0.46)	3.9 (0.22)	3.3 (0.34)	3.54	0.37
	720	4.0 (0.54)		2.1 (0.43)		2.6 (0.41)	2.0 (0.26)	2.1 (0.28)	2.56	0.38
A + V	12	8.2 (0.29)		6.5 (0.24)		7.0 (0.36)	7.1 (0.25)	7.5 (0.25)	7.27	0.28
	60	7.2 (0.33)	3.2 (0.41)		5.5 (0.46)		5.2 (0.25)	6.3 (0.23)	5.48	0.28
	180	6.0 (0.47)		3.4 (0.45)		3.4 (0.20)	3.9 (0.30)	4.6 (0.39)	4.27	0.36
	360	4.5 (0.44)	1.5 (0.31)		2.7 (0.41)		3.5 (0.24)	3.3 (0.41)	3.10	0.36
	720	4.1 (0.53)		2.1 (0.60)		1.8 (0.31)	1.7 (0.25)	2.0 (0.32)	2.35	0.40

Table 15.5. Perception of break frequencies, first common experiment.

15.4.2 Results

Four laboratories were involved in the experiment. The results are reported in Table 15.5 and Figure 15.5. The table contains in each box the mean over the observers and the error on the mean ($s/\sqrt{N}$, N being the number of observations and s the standard deviation).

In the Figure 15.5, the break frequencies are reported as the number of breaks per hour and the quality on the 11 grades scale basis. Different sequences were used in each laboratories in order to smooth the scene content effect. The coherence between results of all laboratories being nevertheless relatively good, it has therefore been possible to use the grand mean to build the figure.

The influence of the audio part of the breaks is very small (less than 5% of the scale) for the break duration we used (40 ms). In the DAB context, the audio recovery time is considered as unacceptable beyond 24 ms but the relationship with video is not known. It may therefore be concluded that, knowing the maximum value assumed for the acceptability of the audio breaks alone, within this range its effect on the video break perception is negligible.

Starting from these results, two points were then under discussion in the group. The first one concerns the lack of data in the interesting part of the curve and the second one the criterion to use to define the maximum frequency acceptable value. Because of these questions, it has been decided to organise a second common experiment before deciding on the threshold.

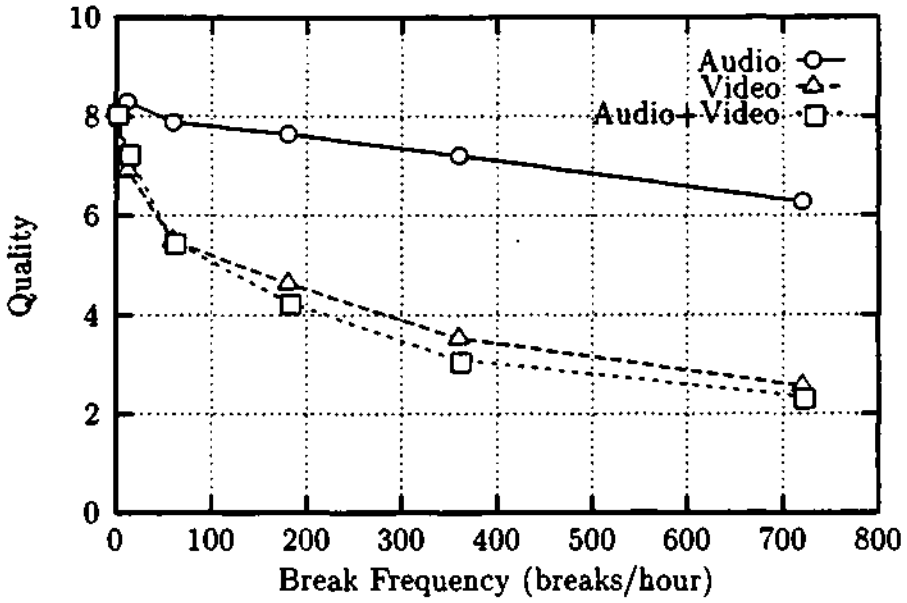


Fig. 15.5. Perception of break frequencies, first common experiment.

15.5 Second common experiment

This last test had to be organised to increase the number of data in the critical part of the curve providing the perceived quality as a function of the transmission break frequency. These needed data corresponding obviously to low break frequencies, the sequences duration had to be significantly increased. It appeared that 30 min were necessary, the test session duration reaching four hours. In that context, the ITU-R viewing conditions were no more acceptable, we had to introduce observers in normal home TV watching conditions.

15.5.1 Description of the procedure

The observers were invited to participate in the test for a complete day, two hours in the morning and two in the afternoon. The test conditions were similar to those Recommended by ITU-R except for the room lighting which might be increased to about 50 Lux, avoiding reflection in the screen, and the quality of the seats which is improved in order to simulate normal long TV watching conditions in a living room.

During a session, two hours, a normal recorded TV programme was displayed to the observers. The technical quality of the programme had been checked in order to avoid bias due to the unpredictable disturbances induced by the standard processing of a normal TV transmission chain. Several sub-programmes were prepared in advance and the observers, four per session, were requested to built their own two hours programme at the beginning of the session by adding some sub-programmes. Each 30

Break Freq. (Breaks/hour)	Swiss PTT (18 observers)	CCETT (16 observers)	Mean (34 observers)	Error on the mean
0	7.9 (0.23)	8.6 (0.43)	8.26	0.17
2	7.8 (0.25)	8.0 (0.53)	7.88	0.18
6	6.9 (0.33)	7.9 (0.46)	7.38	0.23
10	7.0 (0.29)	7.3 (0.60)	7.12	0.21
20	6.1 (0.26)	7.5 (0.70)	6.74	0.24
40	5.6 (0.34)	6.9 (0.66)	6.21	0.26
80	4.8 (0.32)	5.2 (0.57)	4.97	0.22

Table 15.6. Breaks frequency perception.

minutes, the observers were requested to vote on the quality of the past 30 minutes.

The 550 ms transmission breaks were randomly introduced in the video programme the audio part of the programme being unchanged. In that case, the place of the breaks was different in each session and the results for one frequency should be an average over different picture contents during the break. The impairments were built from one dummy sequence and 7 break frequencies: 0, 2, 6, 10, 20, 40 and 80 breaks per hour. The dummy presentation, at the beginning of the first session, was used to focus the attention of the observers on the impairments: frozen pictures. On this basis, four break frequencies were displayed during one session and the eight frequencies were included in the one day test.

The scale on which the observers report their votes each 30 min was the 11 grades scale without any items. A text was proposed for the presentation of the test at the beginning of the first session.

15.5.2 Results

Only two laboratories were finally involved in this long test. The global results are reported in Table 15.6 and Figure 15.6. Since the number of different sequences is too great, it is not possible to report the results for each one. In Table 15.6, the results for both laboratories are provided in addition with the mean. Due to the different sequences used, there is a difference between both series of results but this difference being constant, it is possible to use the average over series in order to provide the final evaluation of the acceptability threshold. This average is reported on the Figure 15.6.

On the basis of these data, an acceptability threshold has to be defined. To process it, the same criterion as for the channel hopping acceptability is used. In order to obtain at least 75% of the viewers judging acceptable (note of five) a break frequency and that with a confidence probability of 95%, we use the following formula, assuming that every distribution is Gaussian:

$$\text{Probability}\{\text{acceptability} < \text{mean(distribution)} + t \cdot \text{std(distribution)}\} = F(t)$$

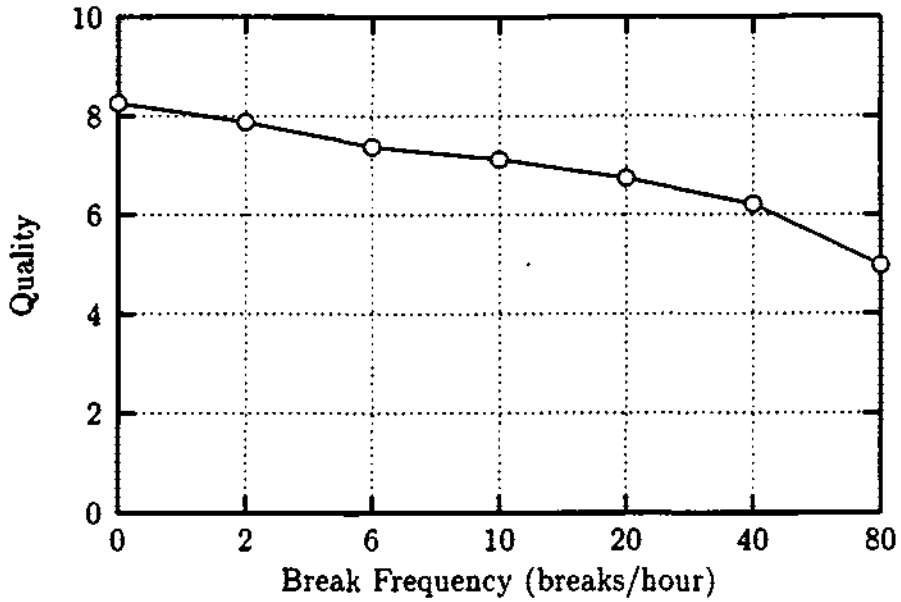


Fig. 15.6. Perception of break frequencies, second experiment

where if $F(t)$ is chosen at 75% then $t = 0.67$, (obviously if $F(t)$ is put at 50%, t is equal to 0).

Assuming that 5 corresponds to the mean acceptability threshold, the "mean (distribution)" is then equal to 5 and the standard deviation of the distribution being processed from the data, it is possible to calculate the limit value of the note complying with the criterion. The mean of the distribution at the note 5 and also the standard deviation are maximised (the most critical case) by taking into account the respective confidence intervals.

This leads to an acceptability threshold of 12 breaks per hour in the first common experiment and 30 breaks per hour in the second one. This processing assures that 75% of the viewers are satisfied with at worse these values.

On Figure 15.7, the results of both common experiments are reported in order to compare the break perception in a laboratory context, 5 min sequences and half an hour sessions, with those of a normal TV watching context at home, normal TV programmes and two hours sessions. The data could be completed but a crossing of the two curves is already assumed. When the break frequency is relatively low, the normal TV conditions induce a higher acceptability of the impairment, on the contrary, the break frequency increasing, the acceptability is higher on short sequences. This result can be interpreted as an increase of the annoyance induced by the strong impairment when it is repeated during a long period of time while the annoyance is masked by the interest for the programme when it is not too strong.

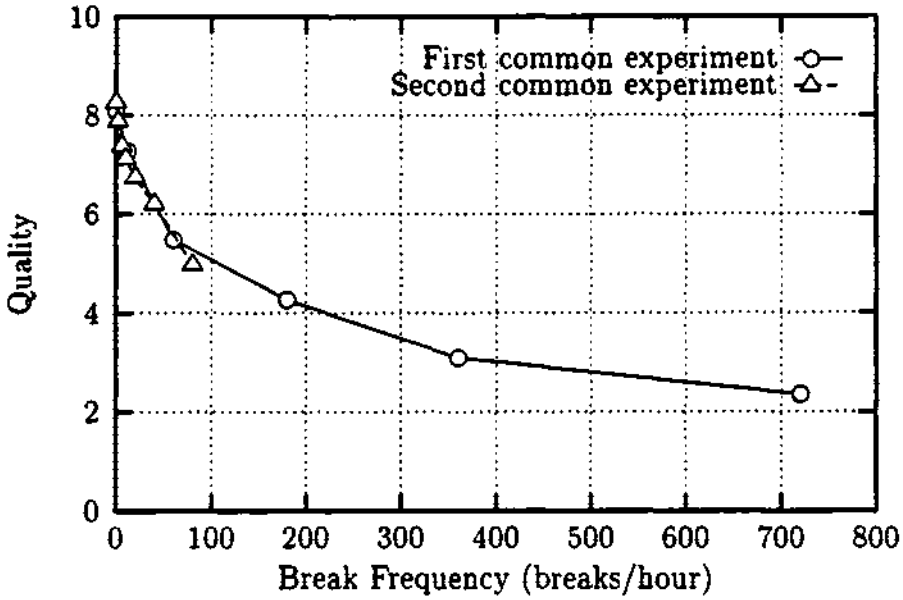


Fig. 15.7. Perception of break frequencies, cumulative results.

15.6 Conclusion of the study

After a first exploratory step, the study was carried out in two main steps, the first one to define the acceptability threshold of the recovery time for the channel hopping and the second one for the same type of threshold in case of transmission breaks. The data on channel hopping allow the calculation of the acceptable recovery time duration while the data in the transmission break environment provide the maximum acceptable number of break of this duration per hour.

The assumed value for the channel hopping acceptability is 550 ms. The calculation of this value, in order to get an acceptability for 75% of the observers, is strongly dependent on the rate selected to define the "acceptability" on the scale. The data provided in this study and some former data concerning the evaluation of the word "acceptable" on a subjective scale lead to the conclusion that the accuracy on this aspect yields an uncertainty on the threshold of ±50 ms. Then, the reliability of the results are dependent on the observers used for the experiment. This uncertainty has been overcome by the high level of satisfaction of the sampling of observers assumed for the threshold calculation (75%). **Consequently, 550 ms may be consider as a reliable value for the subjective acceptability of the channel hopping recovery time.**

The relative possible error is greater for the transmission break recovery time. Depending on the test conditions, the acceptability threshold varies between 12 per hour for the short sequences and 30 per hour for the normal TV programmes. The data obtained in the last experiment with TV programmes are certainly more re-

liable, due to the number of data in the interesting frequency range. Nevertheless, considering that it is quite easy to comply with these values in an operational context, it is propose to keep a small margin beyond the acceptability. **Consequently, 20 transmission breaks of 550 ms per hour, is considered as a reliable value for the subjective acceptability of the transmission breaks.**

Addresses

- Richard Aldridge, University of Essex, Dep. of Electronic Systems Engineering, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom.
  Fax +44 1 206 872 900, e-mail: aldrr@essex.ac.uk

- Thierry Alpert, Centre Commun d'Études de Télédiffussion et Télécommunications (CCETT), 4 rue du Clos Courtel, BP 59, 35512 Cesson-Sévigné, France.
  Tel. +33 99 124 111, Fax +33 99 124 098, e-mail: alpert@ccett.fr

- Éric Bourguignat, Centre Commun d'Études de Télédiffussion et Télécommunications (CCETT), 4 rue du Clos Courtel, BP 59, 35512 Cesson-Sévigné, France.
  Tel. +33 99 124 054, Fax +33 99 124 098, e-mail: bourgui@ccett.fr

- Jules Davidoff, University of Essex, Dep. of Psychology, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom.
  Tel. +44 1 206 873 822, Fax +44 1 206 873 590 (The Graduate Secretary)
  e-mail: janetd@essex.ac.uk (The Graduate Secretary)

- Jean-Pierre Evain, European Broadcasting Union (EBU), Ancienne Route 17A, Geneva, CH-1218, Switzerland.
  Tel. +41 22 7172 734, Fax +41 22 7172 710

- Paul Gardiner, Independent Television Commission (ITC), Kings Worthy Court, Kings Worthy, Winchester, Hampshire, SO23 7QA, United Kingdom.
  Tel. +44 1 962 848 634, Fax +44 1 962 886 109
  e-mail: gardiner@itc.demon.co.uk

- Mohammad Ghanbari, University of Essex, Dep. of Electronic Systems Engineering, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom.
  Tel. +44 1 206 872 434, Fax +44 1 206 872 900, e-mail: ghan@essex.ac.uk

- Massimo Gunetti, RAI Radiotelevisione Italiana, Corso Giambone 68, Torino, I-10135, Italy.
  Tel. +39 11 810 3179, Fax +39 11 619 3779, e-mail: gunetti@crrai.it

- Roelof Hamberg, Instituut voor Perceptie Onderzoek (IPO), Den Dolech 2, 5612 AZ Eindhoven, The Netherlands.

Tel. +31 40 773 884, Fax +31 40 773 876
e-mail: hamberg@natlab.research.philips.com

- David Hands, University of Essex, Dep. of Electronic Systems Engineering, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom.
  Fax +44 1 206 872 900, e-mail: jdavid@essex.ac.uk

- Derek Hawthorne, Independent Television Commission (ITC), Kings Worthy Court, Kings Worthy, Winchester, Hampshire, SO23 7QA, United Kingdom.
  Fax +44 1 962 886 109

- Nick Lodge, Independent Television Commission (ITC), Kings Worthy Court, Kings Worthy, Winchester, Hampshire, SO23 7QA, United Kingdom.
  Tel. +44 1 962 848 634, Fax +44 1 962 886 109
  e-mail: lodge@itc.demon.co.uk

- Don Pearson, University of Essex, Dep. of Electronic Systems Engineering, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom.
  Tel. +44 1 206 872 865, Fax +44 1 206 872 900, e-mail: dep@essex.ac.uk

- Huib de Ridder, Instituut voor Perceptie Onderzoek (IPO), Den Dolech 2, 5612 AZ Eindhoven, The Netherlands.
  Tel. +31 40 773 805, Fax +31 40 773 876
  e-mail: ridder@natlab.research.philips.com

- Samuel Rihs, Swiss Telecom PTT, R&D, FE 416, CH–3000 Bern 29, Switzerland.
  Tel. +41 313 383 639, Fax +41 313 386 823, e-mail: rihs_s@vptt.ch

- Mario Stroppiana, RAI Radiotelevisione Italiana, Corso Giambone 68, Torino, I–10135, Italy.
  Tel. +39 11 810 3118, Fax +39 11 619 3779, e-mail: stroppiana@crrai.it

- David Wood, European Broadcasting Union (EBU), Ancienne Route 17A, Geneva, CH-1218, Switzerland.
  Tel. +41 22 7172 731, Fax +41 22 7172 462
  e-mail: david.wood@itu.ch, wood@pax.eunet.ch

- François Ziserman, Atlantide, 80 Av. des Buttes de Coësmes, 35700 Rennes, France.
  Tel. +33 99 124 094, Fax +33 99 384 267