4th International Conference on

# TELEVISION MEASUREMENTS

# Fourth International Conference on

# TELEVISION MEASUREMENTS

## 20-21 June 1991



*Organised by the*
Electronics Division of the Institution of Electrical Engineers in association with:

*The Institute of Electrical and Electronics Engineers*
*The Royal Television Society*
*The Society of Motion Picture and Television Engineers*

**Venue**
The Casino, Montreux, Switzerland

## Organising Committee



## Corresponding Members



### Author Disclaimer

"While the author and the publisher believe that the information and guidance given in this work are correct, all parties must rely upon their own skill and judgment when making use of it. Neither the author nor the publisher assume any liability to anyone for any loss or damage caused by an error or omission in the work, whether such error or omission is the result of negligence or any other cause. Any and all such liability is disclaimed."

### Copyright and Copying



### Conditions of Acceptance of Advertisements

"The Institution reserves the right to refuse to insert any advertisements (even if ordered and paid for) and/or to make alterations necessary to maintain its standards.

It is not guaranteed that any advertisement will be placed in any specified position or on any specified page unless agreement has been entered into and the agreed surcharge paid.

Every effort will be made to avoid errors but no responsibility will be accepted for any mistakes that may arise in the course of publication of any advertisement. These mistakes may include non-insertion, insertions other than those ordered and errors and omissions within the advertisement.

Notice to cancel any advertisement must be received in writing ten days prior to its next scheduled appearance, otherwise a charge will be made.

No responsibility will be accepted for repetitive errors unless the advertiser's correction has been accepted in respect of that error.

No responsibility will be accepted for loss or damage alleged to arise from errors within advertisement copy, non-appearance of an advertisement or delay in forwarding box number replies.

Advertisers are required to ensure that the content of advertisements conforms with all legislation currently in force affecting such matters. They shall further indemnify the publisher in respect of any claims, costs and expenses that may arise from anything contained within the advertisement and published on their behalf by the Institution of Electrical Engineers.

The placing of an order or contract for insertion of an advertisement in any newspaper or journal published by the Institution of Electrical Engineers whether in writing or by verbal or telephone instructions will be deemed an acceptance of each and all of the above conditions.

Published by the Institution of Electrical Engineers, London, ISBN 0 85296512 5 ISSN 0537-9989

# Contents

**The Institution of Electrical Engineers is not, as a body, responsible for the opinions expressed by individual authors or speakers.**

# Contents

# Contents

*Page No.*

# List of Authors

# RECENT DEVELOPMENTS IN PICTURE QUALITY EVALUATION

David Wood

European Broadcasting Union, Geneva, Switzerland

## SUMMARY

Image quality and impairment evaluation methods continue to develop in the light of experience, and to respond to the demands of new systems. This paper outlines some recent progress in procedures, and particular developments associated with digital system evaluation and HDTV evaluation. The needs of digital television may be met by a new concept termed the 'picture-content failure-characteristics'. The needs of HDTV evaluations include new viewing conditions. Furthermore, a recent challenge in HDTV was to compare the quality potential of alternative HDTV production standards. The paper explains how a real scene was used as a reference in this case.

## INTRODUCTION

Subjective assessments have always been an important element in the development of new image systems. In the end, the most decisive factor in the selection of an algorithm, or process, is what the pictures are perceived to look like. Subjective assessments are the only universal tools we have for quantitatively finding this out.

There is obviously a need for clearly defined and reproducible methods which can be used to obtain reliable results which, in turn, can be compared between laboratories, and with assessment results made at other times.

Traditionally, the CCIR has been one of the groups most active in the development of subjective assessment methods for television. An Interim Working Party, IWP 11/4, was specifically charged with agreeing assessment methods, and made significant progress in the last CCIR Study Period from 1986-1990.

## KEY DEVELOPMENTS IN CCIR STUDIES

To summarize a relatively complex matter and decision tree, the two most effective methods of evaluating picture quality have been found to be the so-called Double Stimulus Continuous Quality Scale method (DSCQS), and the Double Stimulus Impairment Scale method (DSIS).

In the DSCQS method, the assessor is presented with pictures, or sequences, in pairs. One of the pair is the picture after having passed through the process being evaluated. The other of the pair is a reference version of the same picture. This might be, for example, the input to the process being evaluated. The assessor sees the pictures, one after the other, several times, but he is not told which is which. They are simply referred to as 'condition A' and 'condition B'. Finally, when he has a mental measure of the quality associated with each, he is asked to score each on two continuous lines, rather like quality thermometers. As a guide, different parts of the line are associated with each of the five CCIR quality scale descriptors: 'Excellent', 'Good', 'Fair', 'Poor', 'Bad'.

The method has been shown to work very well, and its particular benefit is that the results seem to be relatively independent of the range of impairments available for the tests. That is to say, if only modestly impaired pictures are used, or are available, in the test, there is a little or no tendency for assessors to mark them down unduly.

The method has been used in the selection process that led to the particular 34Mbit/s 4:2:2 codec and 140Mbit/s 4:2:2 codec, now being standardized by the CCIR. Furthermore, it was used for the internal selection processes in the Eureka 95 project to develop an HDTV DBS system, and it is being used by the FCC Advisory Committee in North America, which is choosing an HDTV terrestrial emission system. In short, virtually all major television development projects in the world have used or will use this method to establish basic picture quality.

In the second method, the DSIS method, the assessor is also presented with pictures in pairs. One is the picture to be evaluated, and the other is the reference. The first of the pair is always the reference, and the assessor is told that this is the case. After he has seen the test picture, he is asked for his opinion of the perceptibility and annoyance associated with any impairments seen in the test picture, compared to the reference. He is asked to score his opinion using the CCIR five grade impairment scale.

This method is faster and easier to use than the DSCQS method, but it is 'context sensitive'. This means that the method can only be used in cases where a full range of impairments is available, all the way down to 'very annoying'. However, it is ideally suited to the assessment of failure-characteristics. These are curves which show the impact of varying amounts of a transmission impairment, such as noise, delay, etc. on the picture. Failure characteristics, furthermore, tend to follow an S-shaped curve, which means that curve fitting, can be used with a 'logistic' function, to further improve confidence in the results.

This method has been used for evaluating the failure-characteristics of the proposed 34Mbit/s and 140Mbit/s codecs standard, and will be used for future evaluations by the

FCC Advisory Committee on terrestrial HDTV, and by the EBU, to evaluate the failure characteristics of the HDMAC system.

Having established the two key assessment methods, the CCIR has also looked at points of detail in the application of the methods. Currently, the technological frontiers in television are in two major areas; digital systems and high-definition television.

## THE ASSESSMENT OF HIGH-DEFINITION TELEVISION

The concept of high-definition television is not uniquely defined. It began ten years ago to be seen as a system which would essentially provide a window on the real world, that is, a system where further increases in definition would have no value; the point where the eye becomes saturated with picture detail.

In practice, however, the eye's response has a relatively slow roll-off, and systems meeting thresholds of perception are well beyond practical realization today. It was necessary to decide on the prospective viewing conditions for high-definition television, and provide, at that distance, a reasonable compromise between saturation of picture detail and practical equipment, in terms of the technical characteristics of the system.

The EBU has previously performed assessments of conventional television at both 4 times and 6 times picture height, and traditionally, 6H is considered the design viewing distance for conventional 625-line television. Most home viewing is actually done at 8H-9H, but it is not unreasonable to design a broadcast system around the most stringent bound of home viewing.

For HDTV, early studies by NHK suggested that a screen size of about $1m^2$ was the optimum for home viewing. The home HDTV screen needs to be relatively large to reap the benefits, such as the sense of involvement, of HDTV; but there is a practical ceiling on screen sizes because of modern room size. A further complication is that different types of programme material lead to different natural viewing distances. For sports events, where there is a lot of action, viewers prefers to sit further back than, for example, for drama.

After a relatively lengthy discussion, it was agreed that the viewing distance for HDTV subjective assessments should be three times picture height.

It might be mentioned here that 3H is, however, seen as only the lower bound for evaluations of HDTV sound systems. In this case it is more important to make the evaluations at a distance that will actually be used in the home. This may be expected to be (by interpolation), say, 4H. The sound system required for HDTV, to create the appropriate multi-dimensional sound field, would be significantly different at 3H and 4H.

The CCIR has proposed a complete set of viewing conditions for HDTV, which include viewing distance, screen size, and all other elements needed to allow reproducible and appropriate evaluations. There is also a substantial body of text giving guidance on the choice of methodology and test pictures and sequences.

There has, in addition, been a concentration of effort in another particular area of HDTV studies, that of classification of HDTV emission (or broadcast) systems. The evaluation of ideas in HDTV broadcast systems has been such that when studies began, systems were developed which might typically be used for the 24-27MHz bandwidths available for FM satellite broadcasting in the 12GHz band. These were termed 'narrow RF band HDTV emission systems'. Subsequently, studies moved to a new generation of emission systems which did not have virtually any quality compromises, and were intended to provide HDTV studio quality to the home. These systems would call upon higher broadcast bands (around 20GHz) where there are no apriori restrictions on channel width. These were termed 'wide RF band HDTV emission system'. In recent times, a third class of systems, which can be used in the AM terrestrial broadcast channels (6-8MHz) is being examined. Although formally untitled by the CCIR, they could be expected to provide similar quality to the narrow RF band emission systems mentioned above.

The studies on wide band systems cover a range of bit-rates and bandwidths, and therefore, efforts have been made to categorize the systems by the quality they provide, rather that by any particular bit-rate or bandwidth. In essence this amounts to defining what 'transparency' to the studio standard means.

Ideally, characterisations of systems would use a 'picture content failure characteristics' type approach, as described in the next section of this paper on digital systems. However, in the short term, the proposal has been made that such systems should be evaluated using critical test material. These would be test sequences fully occupied with complex motion, having a spatial/temporal frequency content which exercises the available bandwidth, and a wide variation of colour and textural content. The criteria proposed for wide RF band systems is that, using the double stimulus continuous quality scale methods, the mean difference should be less than 12%, compared to the studio standard. This is about half a CCIR-grade. It has not been possible to arrive at an agreed quality criteria for narrow RF band systems, but, for example, a 25% mean difference could be appropriate here.

## THE ASSESSMENT OF DIGITAL SYSTEMS

It is clear that, in the fullness of time, all parts of the broadcasting chain: production, point-to-point transmission, and emission, will be digital. The advantages are essentially the possibility of readily maintaining high quality throughout the chain, once captured at the source.

In the early 1980s, one of the key factors which influenced the choice of parameter values for the single world-wide digital 625/525 studio standard was a series of subjective assessments of quality versus luminance and colour difference bandwidth and bit-rate (the system is a simple PCM system). A DSCQS approach was used (historically for the first time) and established that a luminance sampling-frequency of more than 12MHz was needed for picture quality transparency, at 4H and 6H, and a colour-difference sampling-frequency of more than 4MHz was needed. For picture processing however, a higher sampling-frequency for colour-difference signals, of greater than 6MHz, proved necessary. It was partly on the basis on these test results that the parameter values of CCIR Recommendation 601 were chosen.

The critical role of subjective evaluations continued with the development of digital codecs for the point-to-point transmission of 4:2:2 signals. However, it has begun to be clear in the last few years that the world of digital codecs calls for a more sophisticated method of establishing quality acceptability.

It has always been true that the quality achieved by a given system has been related to picture content, and indeed the CCIR assessment methods specify that test material should be 'critical but not unduly so' for the system under test. However, the influence of picture content on picture quality has become even more marked with current bit-rate reduction systems. The commonly used "hybrid-DCT" codecs, for example, use a buffer control system to ensure that the final data rate is constant. The techniques used to reduce information rate, which amounts to approximating the original signal, are only called into play when the buffer is full. Thus, the systems can pass, without impairment, pictures up to a point where the total picture entropy exceeds a certain value. This point, where the coefficient truncation, etc. begins, varies with the output bit-rate.

The other somewhat new dimension for digital systems is that the onset of impairments, or failure characteristic, tends to be more abrupt or sharp with digital systems compared to analogue systems.

To be able to fully appreciate the behaviour of a digital codec, investigations are currently being made into what is termed a 'picture-content failure-characteristic' (PCFC).

To define the PCFC, we begin with the assumption that the quality available from a given codec system is influenced, in the most dramatic way, by a particular property of the source. This is termed the criticallity index. This might be, for example, the conditional entropy of the prediction errors.
The first step in the method is to delineate the relationship between picture quality and the criticallity index, by subjective assessments.

The second step is to evaluate the relationship between total programme time and the criticallity index. This is the probability distribution of crticallity. By taking random samples from a representative range of programme material it should be possible to arrive at the curve needed. This will show us how often pictures of given criticallities occur, and thus help us to decide how important it is to have achieve, impairment free, pictures with a particular criticallity index.

For our third step we marry together the first two steps to establish the probability distribution of quality. This tells us how likely a given quality grade is to occur, over a long period of programme time.

It may well be that a given type of system would always have the same basis for the criticallity index, and thus we could establish how varying the bit-rate would effect picture quality; or, more properly, the proportion of overall programme time that would achieve a given quality. A threshold quality might be taken to be, for example, impairment grade 4.5, as an estimate of the 'just perceptible' point.

We could also conceivably arrive at a three dimensional surface which can be used to characterize the performance of a particular codec algorithm family, the axis being probability of occurrence, quality grade, and bit-rate.
Such ideas as these are at an early stage, and in order to proceed we need to encourage one or more laboratories with the appropriate resources to analysis the content of a large number of samples from different types of programme. However, at least one laboratory has announced his intention to do so, and we can be optimistic that the concept of picture-content failure-characteristic will be developed.

## THE MOSCOW GROUP EVALUATIONS

In 1988, an international group of individuals with a special interest in HDTV was formed to try to make comparative evaluations of the proposed HDTV studio standards then before the CCIR, as candidates for a single world-wide standard. The Soviet state broadcaster Gostelradio, offered to host the first series of tests, and the name 'Moscow Group' has been used for this group since, as a consequence.

The group hopes to evaluate the performance of the alternative HDTV scanning proposals, as far as they affect all parts of the HDTV broadcasting chain. As a first step, however, the group turned to an examination of the basic quality available with the two HDTV production standards proposed; 1125/60/2:1 and 1250/50/1:1.

After negotiations with manufacturers it was finally possible to assemble two sets of HDTV production equipment in Moscow at the same time (1125/60/2:1 and 1250/50/2:1), in early 1990, and the first series of tests were successfully completed.

Before the tests were done, however, there were several difficult methodological problems to be resolved.

The method which produces the most reliable and stable results for the assessment of small impairments is the 'double-stimulus continuous quality scale method' mentioned earlier. This method seemed therefore the appropriate one to use in this case. Comparative results can be achieved with this method by indirect comparison. That is, each system is used in an evaluation with the same reference, and afterwards the results obtained from the two systems can be compared.

The major difficulty faced here was to implement the DSCQS reference system, because the test conditions themselves give the highest available quality electronically generated pictures. The idea arose in the Moscow Group of solving this problem by using the real scene as the reference. The idea might work, it was thought, if the real scene could be made to be identical, in terms of size and colour-balance, to the system being evaluated. The double-stimulous method requires that the assessor is unable to tell which is the reference and which is the test condition, apart from by the picture quality itself.

The RAI Research Laboratories in Turin, Italy created and disguarded a large number of designs for test rigs that might perform the task, and eventually a system was devised and built. In essence, the assessor sees through a small window either the HDTV monitor screen; or, his line of sight is intercepted by a traveling mirror, that presents him with a reflected image of the real object. These give the two required conditions, reference and test.

Initial tests were made in 1989 with still pictures, which showed the system to be workable, although an arrangement had to be developed to change illumination level and colour temperature for the two conditions. The apparatus, furthermore, can only be used by one assessor at any one time, as he has to be kept un-aware of the mechanical changes being made, and he has therefore to be contained in an enclosed booth.

Initial trials with a small set containing toys etc. as test source (a 'diorama') brought up a further difficulty, in that the real scene had, in this case, a depth element that was clearly lacking in the off-screen display. In the original test rig both the screen and the real set were arranged to be at three times picture height (3H) from the assessor, the CCIR-recommended viewing distance for HDTV assessments. The problem was partially solved by moving the real set (but not the assessment monitor) to 6H, re-framing the camera, and using relatively flat lighting of the set. When all these measures were taken, the depth effect was sufficiently weak to allow meaningful tests.

When the tests were being designed, we believed that if there were to be important differences in the systems, they would be associated with the vertical-temporal domain. The systems have different numbers of lines/frame and frames/second. In essence, we needed to focus on the particular quality factors which would be affected by these differences. In order to explore this domain, two laboratories; IRT, and NHK, agreed to make rigs for moving panels or posters in the vertical plane at a controlled speed. This proved particularly successful, because the assessments not only targeted precisely the relevant quality factors, they also gave no depth perception difficulties.

In summary, we used three kinds of test pictures for the tests: large static detailed colour prints, large vertically moving detailed colour-prints, and small sets containing toys or models.

As well as making indirect comparison tests, as an experiment, we also made a series of direct comparisons. Here we used two identical monitors, one for each scanning system, and the mirror assembly was used to switch the assessors line of sight from one to the other. Our objective here was to examine the extent to which the context effect would influence the magnitude of the perceived differences in quality between the systems.

The results produced a few surprises. In our initials trials using static pictures we found that, in one or two cases, assessors actually gave the HDTV pictures a slightly higher score than the real pictures. This turned out to be due, in some cases, to the skin texture in the poster. The HDTV picture had fewer skin blemishes, and thus seemed more attractive to the assessor than reality. There is a lesson here to be learned in the selection of test material.

In the main tests, the real pictures constantly received a mean score of about 91 on a 100 point DSCQS scale, independent of the picture or sequence used. The 1250/50/2:1 pictures scored about 69 with a range of 65-74, depending on the particular sequence. The 1125/60/2:1 pictures scored about 81 with a range 78-85, depending on the particular sequence. In summary, there was a relatively systematic difference in favour of the 1125/60/2:1 system of about 12 points. This is about a half grade on the classical 5 point quality scale.

We might have expected the 1250/50/2:1 system to have a slight lead on static pictures (because of the greater number of lines), and the 1125/60/2:1 system to have a lead on the vertically moving sequences (because of the greater number of motion phases). In practice however, I believe that large area flicker had a swamping effect on all other artifacts. The impairment introduced by large-area flicker when the eye is not desensitized to it is dramatic, particular with a 50Hz display rate.

The monitor peak screen luminance was about 70cd/m², the value for maximum resolution of the monitors. This is not an unduly high value, and in fact, domestic televisions normally run at 150-250cd/m².

For the tests we used a 38" HDTV CRT monitor, which we found to be the best available, from the point of view of size and resolution. We used the same monitor type for the evaluation of both the 50Hz and 60Hz systems, to avoid differences in colorimetry influencing the result. The HDTV cameras used were arranged to have the same kind of tube. Nevertheless, there were some differences in noise structure and colorimetry, which may have had a small influence on the result.

The results of the direct comparison normally gave the same rank order as the indirect comparisons, but the magnitude of the difference between the systems narrowed considerably to less than 5 points, and overall, the 95% confidence intervals of the results overlapped. In one case, a static picture with much fine lettering, the 50Hz system even came out ahead.

Clearly there is a complex interaction of different impairments involved, coupled with an element of flicker desensitization when no completely flicker-free picture is available.

To separate the different impairments we would like to repeat the assessments using upconverted displays. This should allow us to assess the true value of the additional motion phases and lines. However, even this will not be the complete answer, because the upconversion process itself may introduce artifacts to complicate the issue. At the moment no upconverted monitor systems are available, and we are continuing the evaluations with an examination of the impact on picture processing capacity of the two systems. The hope is to evaluate the impact of the alternative scanning systems on what is considered the most demanding processing system, colour-matte. Furthermore, this will be done with and without the inclusion of a bit-rate reduction system, of the kind which may be found on a studio-to-studio contribution link.

## CONCLUSIONS

There continues to be a case to review and improve subjective assessment methods; but two particular methods, the DSCQS and DSIS methods, have proved to be the best currently available for the development of new systems.

In the two major image system technical frontiers today, high-definition television, and digital systems, the need for new viewing conditions for assessments, and new methods of characterizing system performance have been met. The concept of picture-quality failure characteristics is particularly important.

Another interesting recent development has been the use of a real scene as a reference for picture quality assessment as part of a comparative evaluation of HDTV production systems.

## ACKNOWLEDGMENTS

The opinions given in this paper are those of the author, but the work described was undertaken by Members of CCIR IWP 11/4, the Moscow Group, the EBU Technical Committee, its Working Parties, Sub-groups, and Specialist Groups.

# FACILITY FOR SUBJECTIVE EVALUATION OF ADVANCED TELEVISION IN NORTH AMERICA

B Caron

Communications Research Centre, Canada

## INTRODUCTION

Canada and the United States hope to introduce an advanced television (ATV) system soon. However the North American way to attain this objective differs from those adopted in other places in the world.

In Europe and Japan, transmission of ATV by satellite is favored and many organizations are collaborating to create this new satellite television system. In North America, the emphasis is on terrestrial transmission and many organizations are competing to have their ATV systems adopted as the ATV standard. In the first case a central authority is necessary to co-ordinate the development of the systems. In the second case it is necessary to evaluate the different systems proposed and to select the best one.

In the United States, the Federal Communication Commission (FCC) plans to establish an ATV standard by 1993. To select a new system, the Commission has created an Advisory Committee on ATV. Proposals were requested in 1988 and more than twenty were received. Some of them were eliminated when the FCC ruled that a new ATV system must use the present 6 MHz channel allocation and that NTSC transmissions must be maintained. The remaining systems were thus divided into two types: simulcast and NTSC compatible.

The simulcast systems would use channels presently free to transmit signals uncompatible with NTSC. The regular NTSC transmission will continue as in the past.

The NTSC compatible signal is compatible with current NTSC receivers and will be transmitted on the channels presently occupied by NTSC signals. Improved video will be available on Advanced Television sets and the regular NTSC video will be displayed on regular television sets. The compatible system may not improve quality as much as simulcast systems could, but its cost may be quite a bit lower. Among the 6 proposals which will be tested, 5 are simulcast systems and only 1 is NTSC compatible. They are listed in Table 1.

To evaluate these proposals the Advisory Committee of the FCC created different subcommittees and working parties open to representatives of the industry.

Broadcasters have a major interest in these groups. Cable operators are also very concerned as the introduction of an ATV system will be much easier if the same ATV receivers can be used both for cable and over-the-air reception. Canada also supports the adoption of a single ATV standard for North America.

To demonstrate their support these organizations provided the FCC Advisory Committee with the necessary testing facilities. The Advanced Television Test Center (ATTC) created by the broadcasters, is responsible for the objective tests, as well as for the terrestrial (over-the-air)

TABLE 1 - List of the ATV systems to be assessed in North America

| Type | System | Proponent |
|---|---|---|
| NTSC compatible | Advanced Compatible Television (ACTV) | David Sarnoff Research Centre, etc. |
| Simulcast | Narrow Muse<br>DigiCipher<br>Spectrum Compatible HDTV<br>Simulcast HDTV<br>Channel Compatible HDTV | NHK<br>General Instrument & MIT<br>Zenith<br>N.A. Philips, etc...<br>MIT & General Instruments |

transmission impairment tests and the production of all the test material recordings.

Cable Television Laboratories, created by the National Cable Association, will complete all the tests related to cable impairments.

Finally the Advanced Television Evaluation Laboratory created in Canada is to complete the subjective assessments with non-expert observers.

The Advisory Committee's System Subcommittee's Working Party 2 (SS-WP2) on ATV Evaluation and Testing is responsible for conceiving and supervising all the tests necessary to evaluate the different proposals. The results of these tests, along with the studies completed by other Working Parties on spectrum, cost, etc., will be used by SS-WP4 on System Standards to prepare a recommendation to the FCC.

ATV EVALUATION

The task of SS-WP2 was made quite challenging by some peculiarities of the ATV evaluation process.

First, for many reasons, it was decided early to test the different proposed systems in a sequential manner: they will not be compared directly to each other. Instead the comparison will be made indirectly by comparing each system to a common reference. The 1125-line studio standard was selected as the common reference because of the availability of equipment. It will operate at 59.94 Hz for all ATV systems using this field rate. For the ATV system operating at 60 Hz, the field rate of the reference will be 60 Hz. It was also decided that the comparison will be made on the same display instead of a side-by-side comparison with 2 displays.

Another problem is the fact that not all the systems proposed used the same scanning standard. Four different scan rates are used as listed in Table 2.

TABLE 2 - Proposed scan rates

| |
|---|
| 525/59.94/1:1 |
| 1050/59.94/2:1 |
| 787.5/59.94/1:1 |
| 1125/60/2:1 |

To complete the assessment, it is necessary to create the same video test materials using all the different scan rates. It is also useful to record the output picture for archival purposes or to be able to complete some tests later on using the recorded material. Video multi-standard cameras, recorders and displays are thus required.

No single camera could operate correctly at all the scan rates. It was necessary to use different cameras.

The recording problem was solved by the invention of the "format converter" at the ATTC. This device is an interface to the Sony HDD-1000 High-Definition Digital Video Tape Recorder. It can convert the pictures of any of the above scan rates to a stream of data which can be recorded on the DVTR. It also converts the recorded data stream back to the original pictures.

With this device and multi-standard cameras, it is possible to shoot and record the same scenes in all the different formats. At the testing laboratories, the tapes in the appropriate formats are played back as required to test a system. The output of the decoder of the system under test can be recorded using the format converter. Finally, the recordings of the results could be played back any time using the format converter. There is no need for an ATV decoder to be available.

A multi-standard rear-projector was selected as the common display. It has a 16 to 9 aspect ratio and a diagonal dimension of 65 inches (165 cm).

Another problem is the need to design the test procedures before detailed information on the systems to be tested are known. A large part of the test procedure was written even before some of the systems to be tested were conceived.

With these problems in mind, three groups of tests were prepared for the evaluation of the ATV systems. First, there are the objective tests, which include such measurements as spectrum bandwidth and luminance or chrominance resolution. These test procedures are described in detail in (1).

Second, a great number of subjective tests are used to evaluate the subjective quality of each system and its robustness to impairments. These test procedures are described in (2). Two real-time channel simulators will be used in the laboratories to create impairments to the signals. One will create impairments typical of terrestrial over-the-air transmission as described in (1). The second one will generate impairments found on cable networks. They are described in (3).

Finally field tests will be performed on the most promising system(s) to confirm the results obtained in the laboratories. The field test plan is under preparation.

The subjective assessments, because of their importance, will now be discussed in more detail.

## SUBJECTIVE ASSESSMENTS

Many of the parameters of each proposed ATV system must be evaluated subjectively. The necessary tests are divided into 2 categories: Picture Quality rating and Transmission Impairment rating.

### Picture Quality Rating

The picture quality rating will compare the basic quality of each ATV system. Fifteen different picture quality attributes such as static and dynamic resolution, motion rendition and dynamic range will be evaluated.

Four different quality rating assessments will be done:

(1) The ATV Basic Quality rating where the system under test is compared with a source reference program. In this case, the ATV signal is modulated and demodulated but no other impairments are added. The reference is always the 1125-line source. This test will provide information about the relative quality of the different systems when there is no added transmission channel impairments.

(2) The ATV Cable Quality rating in which the pictures of the system under test are impaired by a laboratory cable network simulator. In this case the picture at the output of the cable network is compared with the ATV picture transmitted on an unimpaired channel to evaluate the degradation of the picture quality.

(3) The ATV Fibre Quality rating in which the pictures of the system under test are impaired by a laboratory cable network which includes a fiber optic portion. Again the reference is the unimpaired modulated and demodulated ATV signal.

(4) The NTSC Reception Quality rating which is done only for the NTSC compatible ATV system. It is used to evaluate the quality of the ATV transmission when received by conventional NTSC television receivers. In this case, the ATV signal received by the NTSC receiver is compared with a standard NTSC transmission of the same program.

At least four quality rating sessions, described below, will be required to complete each of the 4 quality rating assessments presented above.

A session begins with 4 instruction trials which are used to instruct the viewers about the assessment procedure. They are followed by 5 warm-up trials, not identified as such to the viewers. Finally, the 46 real trials are presented in a random fashion. Twenty-three different sequences are presented twice in each quality rating session. Fourteen of them contain motion.

In each trial, the test and the reference sequences are displayed twice alternatively on the same display for 10 seconds each. Each sequence is identified orally by A and B and is to be graded by the viewers.

The display is blanked to gray for 3 seconds between each sequence and there is a 5 second gray period at the end of a trial for people to write down their evaluation.

There is a maximum of 5 viewers per session. To ensure the validity of the assessment a minimum of 20 different viewers is required for each quality rating assessment.

The non-expert viewers will be recruited from a local university. Each one will participate in only one session. They will be paid, in order to motivate them to complete the assessment. Prior to the assessment, they will be tested to confirm that they have normal color vision and visual acuity.

Figure 1 illustrates the set-up used in Canada for the subjective rating by non-expert viewers. The tapes containing the subjective test material described above are played back. The digital output of the DVTR is converted to the original ATV format using the format converter. The sequences are presented to the viewers on the multi-scan rear-projector. Between each sequence, a control signal switches the scan format from the 1125-line raster to the raster of the ATV system under test.

The viewers are seated at a distance of 3 times the picture height (3H). The projector is surrounded by a light wall. This light wall fills the front of the room. It has a 6500°K color temperature and its luminance is equal to 15% of the projector's peak luminance.

The quality rating for the NTSC compatible system is done the same way. In this case, however, the large rear-projector is replaced by a 35" (90 cm) consumer type television set which is more appropriate for NTSC viewing. The NTSC signal is recorded and played back on a conventional NTSC digital video tape recorder (D2). The viewers are seated at a distance of 6H instead of 3H for the ATV tests.

In both cases, the viewers will write their opinion on two continuous 5-grade A-B scales marked bad, poor, fair, good and excellent. They are free to mark anywhere on the scale. The result of this assessment will be a grade for each of the 15 picture attributes.

Impairment Rating

Once the basic quality of an ATV system has been established, it will be necessary to estimate its robustness to impairments. Each system will be tested for the most important impairments normally encountered for over-the-air and coaxial cable transmission. They are listed in Table 3.

Impairment rating tests are to be completed for only 1 impairment introduced at a time. Combinations of impairments will not be tested at this time.

For the interference impairments, the rating tests will be repeated for 3 different levels of desired signal so that nonlinear distortion effects are included.

It would not be appropriate to complete the subjective rating of all the impairments with non-expert observers. Only the impairments in the upper part of Table 3 will be rated by non-expert observers. The effects of all the impairments will be,however, observed and commented at ATTC by 5 video experts.

In preparation for the subjective impairment rating by non-expert observers, the video experts will also establish six representative levels of impairment for which the rating test material will be recorded. The set-up to do so is illustrated in Figure 2. The source picture is played back on the DVTR. Its digital output is converted to the original ATV raster by the format converter. The prototype ATV encoder of the system under test processes the source picture to produce a modulated signal. This signal is impaired by the channel simulator. The impaired signal, once demodulated and decoded, is displayed on the rear-projector to the expert viewers who make their evaluation. Finally, the impaired signal is recorded with the format converter. Appropriate recordings are sent to Canada for the subjective impairment rating by non-expert viewers. This subjective rating is complete with the equipment set-up of Figure 1.

Some of the impairment ratings are performed to evaluate the interference that an ATV transmission would create to regular NTSC reception. This is why, as shown in figure 2, an NTSC signal is sometimes transmitted through the channel simulator along with the ATV signal. In this case, the signal at the output of the channel simulator is fed to a bank of 24 representative NTSC television sets. Based on the experts' observations, the baseband output of one of these 24 receivers is recorded on a digital video tape recorder (D2) for the impairment rating by non-experts.

Each impairment rating session with non-expert observers begins with some instructions and warm-up trials.

Each session is made of trials covering the 6 different levels of an impairment for 3 different video sequences. Each trial

TABLE 3 - Impairments to be tested

| | |
|---|---|
| By Non-Experts and Experts | Noise to ATV<br>Co-Channel and Adjacent Channel Interference<br>- ATV to ATV<br>- NTSC to ATV<br>- ATV to NTSC<br>UHF Taboo ATV to NTSC<br>Cable Third-Order Intermodulation |
| By Experts Only | Discrete Frequency Interference<br>Impulse Noise<br>Multipath<br>Airplane Flutter<br>Cable Inter-Carrier Phase Modulation<br>Cable Hum and Low Frequency Noise |

contains two 10 second sequences, one is unimpaired and the other one is impaired. There is a 3 second gray period between the 2 sequences and a 5 second gray period after the second one for the viewer to write his rating.

The viewers rate the impaired sequence on a 5-grade scale: imperceptible, perceptible but not annoying, slightly annoying, annoying and very annoying. Their choice is restricted to these 5 grades.

The results of this assessment will be a grade for the 6 different levels of each impairment.

OTHER TESTS

On top of the assessments described above objective and subjective audio tests will be completed. A few tests will also be made to assess some characteristics or parameters specific to a particular ATV system. These tests will be defined when the detailed information on each system becomes available, 90 days before the beginning of its assessment in the laboratory.

Once all the systems have been assessed in the laboratory, SS-WP4 on System Standards will select one or two systems for field tests. At the same time it is also possible that some supplementary laboratory tests will be performed to evaluate factors which may not have been covered in the first laboratory tests.

SCHEDULE AND DIFFICULTIES

The following schedule was established at the beginning of 1991:

May 1991: Beginning of Laboratory Tests (2 months per system)

June 1992: End of Laboratory Tests

Summer 1992: Field Tests

Sept. 1992: Recommendation to FCC

2nd Quarter 1993: New Standard

Many problems, however, may affect this schedule. For example, delays or incomplete results can be expected as the ATV system prototype as well as some of the test procedures are unproven. It is also possible that developments or proposals made after the beginning of the laboratory tests may significantly improve the potential performance of ATV systems and make testing of new systems necessary.

It is hoped however that all these difficulties could be overcome and that North Americans will soon enjoy a new era of television.

REFERENCES

1. "ATV Test Procedures. Objective and Transmission Tests", SS/WP2-0189, 28 Sept. 1990.

2. ATV Test Procedures Manual. Cable Television Transmission Tests. SS/WP2-0357, 14 Aug. 1990.

3. "ATV Test Procedures. Video Subjective Tests", SS/WP2-0390, 28 Sept. 1990.