Production Systems and Information Engineering
Volume 5 (2009), pp. 139-150.

QUALITY ASSESSMENT OF MOTION PICTURE TRANSMISSION OVER DIGITAL CHANNELS

ATTILA K. VARGA
University of Miskolc, Hungary
Department of Automation
varga.avarga@uni-miskolc.hu

DÉNES DALMI
University of Miskolc, Hungary
Department of Automation

dalmi2@mazsola.iit.uni-miskolc.hu

[Received January 2009 and accepted April 2009]

Abstract. Digital technology forms our environment day by day and we can enjoy its advantages in more and more fields everyday. As a matter of fact television and computer networking technology take over the leading role. The real error analysis of digital channels is the best way for modeling the behaviour of motion picture transmission over digital channels.

In this paper digital transmission has been focused on considering such factors as noises, disturbances and movement as well as the mathematical modeling of digital transfer channels has been shown. The paper first presents the most important standardized subjective quality assessment methods described in the ITU-R BT.500 recommendation. We briefly summarise why these subjective tests are so important. Finally, we discuss the implementation of the new subjective video quality measurement related to impaired digital quality television programs. Our aim is to improve these subjective picture quality assessment methods to get sophisticated results, which correlate better with objective picture quality test results.

Keywords: digital broadcasting, digital cable television, head-end, digital channels, transport stream, subjective test, video quality analysis

1. Aims and Scope of the Paper

Several articles and studies have investigated the quality of telecommunication transfer [1]. ITU-T Recommendation P.800 describes methods and procedures for conducting subjective evaluations of telecommunication transmission quality. ITU-T recommendation [2] gives an objective method for determining voice quality, described in P.862 Recommendation, which is known as "Perceptual Evaluation of Speech Quality" (PESQ).

Although we cannot find a recommendation or standard for the objective quality measurment of video transfer, but subjective measuring algorithm exist [3]. The development of picture quality analysis algorithms available today started with still image models which were later enhanced to also cover motion pictures. The measurement paradigm is to assess degradations of a decoded video sequence output from the network in comparsion to the original reference picture.

The main objective of this paper is to present what kind of assessment tests have been used for examining the quality of digital channels and to describe the standards and subjective methods we used for determining the sources of the errors in the transfer channels.

2. Introduction

In 2004, the Department of Automation (University of Miskolc, Hungary) won a three-year project (GVOP) in the field of Digital Broadcasting the aim of which is to develop software and hardware modules for Digital Cable Television (CATV) head-ends in cooperation with one of the most famous and world-wide known Hungarian Digital CATV components manufacturing company called CableWorld Ltd. A team formed by students and a group of the staff of the Department carried out the developments at the University, mainly the software developments and the installation of the head-end were performed in the laboratory of the Department of Automation.

For the past few years we have dealt with subjective and objective picture quality measurements of digital television streams in the Digital Television Laboratory of the Department of Automation. After we had analysed the results of our subjective tests and drawn the conclusions, we started new subjective quality measurements, which focus on the video quality of digital television streams, so-called transport streams having different bit-rates.

Compression methods for digital television use different compression algorithms. Quality measurements are used to find the best compression method. There are two main categories of comparison methods: the objective video quality evaluation method based on mathematical calculations and the subjective video quality evaluation methods based on tests performed by the audience.

Digital television streams are compressed according to the MPEG-2 or MPEG-4 standards. Nowadays digital television broadcasting systems often use statistical multiplexers. In statistical multiplexing, the communication channel is divided into an appropriate number of variable bit-rate digital channels or data streams. Our goal is to determine the lowest bit-rate, which has still acceptable quality. This bit-

rate would be used in statistical multiplexers as the minimum bit-rate. Consequently, we use these quality measurements in order to find the compression parameters, which still result in acceptable video quality.

3. Subjective Motion Picture Quality Assessment Methods

In this section we would like to introduce the most common subjective quality assessment methods of the digital television picture [2].

International recommendations for subjective quality assessment of television picture consist of specifications of how to perform many different types of subjective tests. Subjective assessment methods are used to establish the performance of television systems. Measurements are therefore applied, which more directly anticipate the reactions of those who might view the tested systems. In this regard, it is understood that it may not be possible to fully characterize the system performance by objective means. Consequently, it is necessary to supplement objective measurements with subjective measurements.

In the course of a typical subjective quality test, a number of non-expert observers are selected, tested for their visual capabilities, shown a series of test scenes for about 10 to 30 minutes in a controlled environment and asked to score the quality of the scenes in one of a variety of manners.

In general, there are two types of subjective assessments. First, there are assessments that bring about the performance of systems under optimum conditions. These are usually called quality assessments. Second, there are assessments that create the ability of systems to retain quality under non-optimum conditions associated with the transmission or emission called impairment assessments. Some of these test methods are double-stimulus where viewers rate the quality or the change in quality between two video streams (reference and impaired). Others are single-stimulus where viewers rate the quality of just one video stream (the impaired one). These methods will be later described.

In a modern television system, however, the picture quality is not constant over time due to the compression streams. In the case of statistical multiplexing, the picture quality is a function of the complexity of the program material and the continuous operation of the transmission system. The selection of the assessment method is affected by a number of procedural elements. These are the viewing conditions, the choice of observers, the scaling method to score the opinions, the reference conditions, the signal sources for the test scenes, the timing of the presentation of the various test scenes, the selection of a range of test scenes and the analysis of the resulting scores.

A description of the various subjective measurement methods provides some insight in the following sections.

3.1. Double-stimulus Impairment Scale Method

The double-stimulus Impairment Scale (DSIS) is a subjective assessment method when observers are shown multiple reference scenes and degraded scene pairs. The reference scene is always shown first. Scoring is on an overall impression scale of impairment.

Table 1. Five-grade scale recommended by ITU

Five-grade scale	
Quality	**Impairment**
5 Excellent	5 Imperceptible
4 Good	4 Perceptible, but not annoying
3 Fair	3 Slightly annoying
2 Poor	2 Annoying
1 Bad	1 Very annoying

This scale is commonly known as the 5-point scale, where 5 equals the imperceptible level of impairment and 1 shows the very annoying level as shown in Table 1.

3.2 Double-stimulus Continuous Quality-scale Method

In case of the Double-stimulus Continuous Quality-scale (DSCQS) method, observers are shown multiple sequence pairs with the reference and degraded sequences randomly first. Scoring is on a continuous quality scale from excellent to bad where each sequence of the pair is separately rated but in reference to the other sequence in the pair. Analysis is based on the difference in rating for each pair rather than the absolute values.

3.3. Single-stimulus Methods

Multiple separate scenes are shown in the Single-stimulus methods. There are two approaches: SS with no repetition of test scenes and SSMR where the test scenes are repeated multiple times. Three different scoring methods are used. The adjectival scoring method has a 5-grade impairment scale, and half-grades may be

allowed. The numerical scoring method has an 11-grade numerical scale, useful if a reference is not available. And finally there is Non-categorical scoring, where assessors can score in a continuous scale with no numbers or a large range.

3.4. Stimulus-comparison Method

The stimulus-comparison method is usually implemented with two well-matched monitors but may be done with one. The differences between sequence pairs are scored in two different ways: the adjectival scale is 7-grade, +3 to -3 scale labelled: much better, better, slightly better, the same, slightly worse, worse, and much worse, while the Non-categorical is a continuous scale with no numbers or a relation number either in absolute terms or related to a standard pair.

3.5. Single Stimulus Continuous Quality Evaluation

The Single Stimulus Continuous Quality Evaluation (SSCQE) is performed with a program, as opposed to separate test scenes, which is continuously evaluated over a long period of 10 to 20 minutes. Data are taken from a continuous scale every few seconds. Scoring is a distribution of the amount of time a particular score is given. This method relates well to the time variant qualities of new compressed systems. However, it tends to have a significant content of program quality in addition to the picture quality [4].

4. Statistical Multiplexing

The flexibility of the MPEG-2 coding system provides the opportunity to broadcast digital television streams, which have more or less bit-rates. Everybody knows that the picture contains more information and has better quality when the rate of the stream, which transmits the compressed picture, is higher. In case of still or slowly moving picture sequences, which do not contain fine details, there is a limit, above which there is no use increasing the data rate, the picture, which has good quality, cannot be better at the receiver side. The change of the picture content and the moving of picture elements increase the amount of information to be transferred. Consequently, to observe the video quality, the data rate must be raised.

The creation of data rate depending on the picture content only makes sense when we can utilize the unused data rate range. In different transmission networks, where more TV programmes can be simultaneously transmitted, in the spaces, which become vacant, one or more TV programmes can be delivered if we can control the resulting data rate.

Statistical multiplexing means that at the transmitter site we compress the data stream with content-dependent data rate; however, we should meet the requirements that the resulting data rate cannot be higher than a predefined value. It is also important to determine a predefined order with which we ensure how much data rate will be allocated to the given programme in case of a large bit-rate demand at the same time [3].

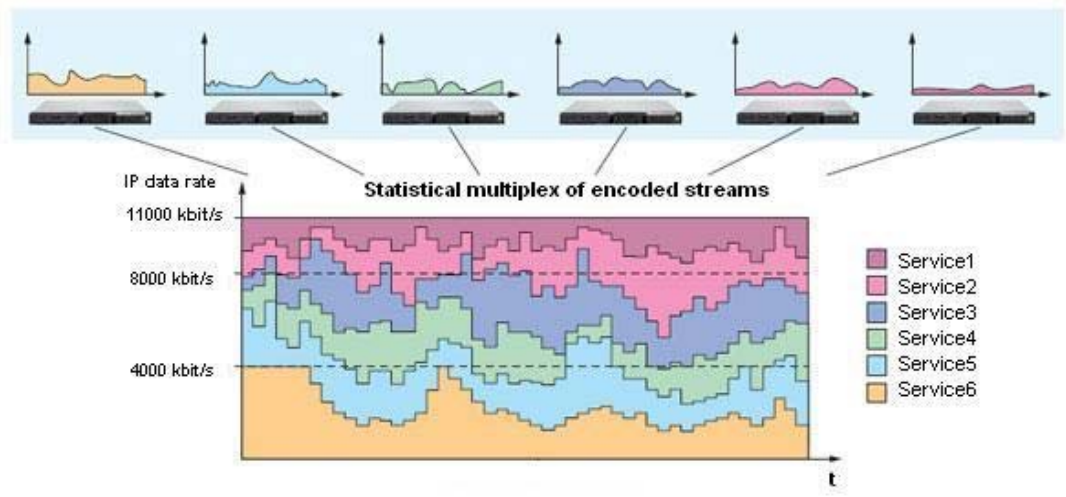

Figure 1. Statistical multiplexing

Figure 1. shows how the statistical multiplex works, so the digital television streams which are coming from different locations (e.g. studios) with variable bit-rates are added in one statistical multiplex stream.

With subjective quality measurements of digital TV streams, the minimum level of bit-rate and other coding parameters, such as GOP (Group of Pictures) size and structure, as well as video picture parameters like brightness, contrast, saturation, can be determined. Nowadays there is a significant demand for these subjective results.

5. Subjective Video Quality Measurements

Measuring the quality of digital transfer channels can be carried out by using subjective methods described above. The main idea of measuring subjective video quality is the same as in the Mean Opinion Score (MOS) for audio. Many parameters of the viewing conditions can influence the results, such as room illumination, display type, brightness, contrast, resolution, viewing distance, and the age and educational level of experts. There are an enormous number of ways of

showing video sequences to experts and to record their opinion. A few of them have been standardized.

These methods can be used for several different purposes including, but not limited to, selection of algorithms, ranking of audiovisual system performance and evaluation of the quality level during an audiovisual connection. Source might be a TV0 type signal given in ITU-R Recommendation BT.601-5 [5]. We can test audio channel (without video), or video channel (without voice) and audiovisual channel (voice and video).

In this section we would like to describe our previous subjective picture quality measurements, and then we would like to go into details about our new measurements.

5.1 Short Presentation of Previous Quality Tests

We have previously executed three different types of subjective picture quality tests of digital television pictures coming from different digital television channels. We used a wide screen LCD television for the experiment, whose screen could be separated into two parts. We chose three different digital television channels: satellite, cable and terrestrial. We selected three different programs: m2, Duna and Autonómia, which can be freely received in Hungary. The observers were undergraduates and one test session consisted of 5-15 of them. In the first test, observers rated the still pictures one after the other. In the second one, picture sequences were displayed on the two separate screens, so students had to evaluate the picture quality simultaneously. Finally, in the last test, observers assessed the quality of short motion picture sequences.

The evaluation was created by taking into account three aspects: sharpness, naturalness and subjective order. Therefore, observers had to determine an order between pictures A and B. They could note the results in an evaluation form. Test sessions took about 20-30 minutes. One test session comprised 8-12 pairs of 10-second pictures, covered the possible combination of different sources, such as satellite vs. cable. Between pictures there was a 10-second interval for the evaluation. Before the test pictures there was a mid-grey picture as mentioned in the ITU standard. We evaluated the test results by counting the scores of the observers in the different categories. In the serial subjective test of still pictures, we collected 216 scores, according to which the cable system was given most of the scores in each category. In the serial test of motion pictures, we obtained a varied result, from the 243 scores gathered, the terrestrial system dominated in the sharpness category, while the satellite system was given most of the votes in the naturalness and the subjective order categories [5].

Drawing the conclusions, we can make some important remarks. First of all, we should create some teaching methods for video assessment, so that the non-expert observers could prepare for assessing the quality. It is very important to teach the observers what they should pay attention to before the real test, because it greatly influences the test results. The experiment leader should explain and demonstrate the evaluation categories (naturalness, sharpness, saturation, hue, etc.), the typical errors, which can occur in the digital video streams, and naturally the essential information about the subjective quality assessment (number of test sequences, the duration of the scoring period, the scoring scale, etc.). In our opinion, by using a well-implemented teaching method, the fidelity of the subjective quality assessment can be improved.

Another important point is to select and record the test material in an appropriate way. In our previous subjective quality measurements it was a serious problem that the test sequences were recorded after the error correction on the receiver side and not at the end of the transmission channel before the error correction. In the new subjective quality assessment, it was also a difficult task how to record test samples with various bit-rates. We provide the related information in the following section.

We should also consider the laboratory circumstances (the distance between the screen and the observers, the resolution and other parameters of the television set, etc.). The ITU recommendation has good criteria to establish the appropriate laboratory environment; however, it has financial implication.

Finally, we should find a better way to record the scores of the observers, because so far they have filled a voting form. We had to evaluate thousands of scoring papers, which resulted in mistakes. Consequently, a subjective quality assessment application is developed in order to help our work.

5.2. Subjective Quality Measurement

As mentioned previously, our purpose is to conduct some subjective video quality tests of digital television streams, which have various bit-rates.

Subjective Quality Assessment Supporter Application

For these measurements we have developed an application in Java environment, which provides a graphical interface in order to assess the digital television video easily.

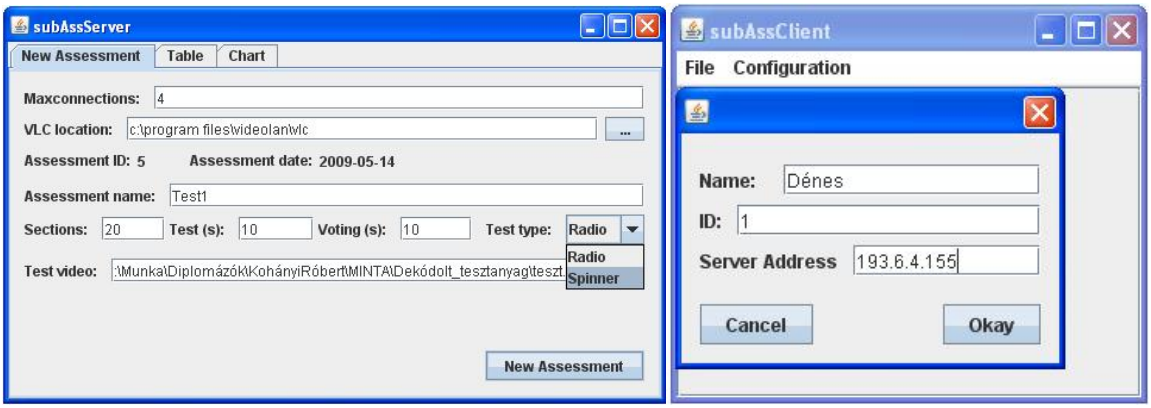

Figure 2. Subjective quality assessment software

The program has two parts: the server and the client, which can be seen in Figure 2. The experiment leader, who conducts the measurement, can configurate or customize the subjective quality test on the *New Assessment* tab in the server software. First, the *Maxconnections* field has to be set, which determines the number of observers. Then, the experiment leader should give the path of the VLC location. If it is well configurated, then after the start of the new assessment, the VLC media player will display the test sequences. The assessment name and date are automatically set by the program. In the following steps the experiment leader should give the name of the assessment, set the number of sections in the test session, configurate the duration of one test sequence and the scoring period in seconds and select the type of the test scale, which can be a 5-grage scale recommended by ITU as it is shown in *Table 1.* or a spinner, which is a 100-grade continuous scale. Finally, the path of the test material has to be set.

The observers should run the client program and set some parameters, such as the name, the unique ID and the IP of the computer on which the server application runs.

When the experiment leader starts the measurement, which can be automatic or manual, the voting screen will automatically appear on the client screen and the observers will have a defined amount of time to score the quality. The client software sends the scores to the server application, which stores them in its database. When the subjective measurement is finished, the experiment leader can evaluate the results in a table or in a chart. The table contains the assessment ID, the assessment name and date, the assessor ID and name, the section number and the quality score. With SQL commands, the experiment leader can create some queries in order to filter the huge amount of data. In the chart, the results of a given assessment can be seen, where the two axes are the number of sections and the mean value of the scores given by the observers.

Recording the Test Material

Our first task was to record digital television video samples, which have different bit-rates. *Fig.3.* presents the environment how we recorded the test material.

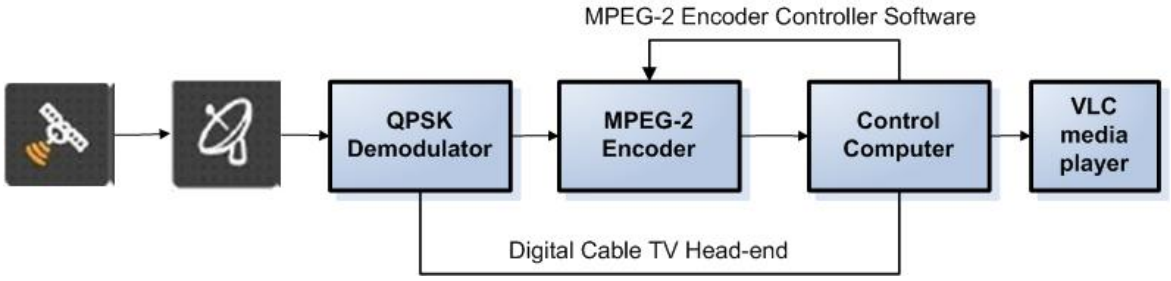

Figure 3. Environment for Recording the Test Material

In the Digital Television Laboratory we used the Digital Cable TV Head-end, which contains special hardware devices developed by CableWorld Ltd. The QPSK demodulator is used to receive the digital transport streams broadcasted via satellite channel. The demodulated transport stream is then sent to the MPEG-2 Encoder. With the MPEG-2 Encoder Controller application running on the Control Computer, the coding parameters and the bit-rates of the transport stream could be configurated. In the final step, this encoded transport stream was displayed with the VLC media player. We used this media player to record video samples.

The problem was that we could not record test samples with various bit-rates continuously; it was the fault of the VLC media player. Therefore, we recorded 10-second video samples and concatenated them into one test video sequence, which could be later used for the subjective quality measurements. However, we have not found appropriate MPEG-2 editor software yet, with which we can concatenate the split sections without re-encoding them. So it is a problem, which needs to be solved in the future.

6. Evaluating the Subjective Quality Assessment

We established a quality assessment environment in our laboratory. We created a computer network with 9-12 personal and one server computers. Observers used the personal computers to run the client application. On the server machine the experiment leader ran the server application and conducted the subjective quality test. One test session took about 10-20 minutes, because the observers needed to concentrate hard during the quality assessment.

Table 2. Five-grade scale recommended by ITU

Seq. N.	Bit-rate (kbps)	1. Measurement (0-5)	2. Measurement (0-100)
1.	8000	2.75	39.75
2.	992	1.25	4.75
3.	1504	3.75	51.50
4.	4000	4.50	73.25
5.	1104	1.50	8.25
6.	1600	2.50	39.25
7.	2608	5.00	87.75
8.	3504	4.25	79.75
9.	3008	3.75	69
10.	2800	4.50	67.75
11.	1904	3.50	33.50
12.	1200	2.25	21
13.	6000	3.50	76
14.	1312	1.00	7.75
15.	4512	4.00	67.75
16.	1408	2.25	22.25
17.	2400	3.25	65
18.	5008	4.00	73
19.	2000	4.25	75.50

So far we have only a few number of test results as described in *Table 2.*We used test material including 19 sections with different bit-rates. In the first and the second measurements the mean of the quality scores can be seen. The difference between the two measurements is the scoring scale, which was used for the test. It can be seen that the video sequence, which has a higher bit-rate, was given better quality scores, but there are discrepancies in the test results. It is important to mention that this result is not representative because the number of assessors who were involved in our assessment is less than 10.

To give a significant result we need to repeat this measurement with a large number of observers. According to our assumption, the lowest bit-rate which has still acceptable quality is about 1500 Kbit/s. However, it will be our future work to verify it.

7. Conclusions

An important issue in choosing a test method is the basic difference between the methods that use explicit references and methods that do not use any explicit reference. If we want to determine the quality of an audio-visual transfer channel, then we can do it using a subjective measuring algorithm.
The accuracy of perceptual objective test methods can be verified by comparison with subjective video quality tests. However, subjective testing can be both time-consuming and costly. In order to achieve statistically relevant results a huge test population must be evaluated.
The possible number of subjects in a viewing and listening test (as well as in usability tests on terminals or services) is between 6 and 40. Four is the absolute minimum for statistical reasons, while there is rarely any point in going beyond 40.
The actual number in a specific test should really depend on the required validity and the need to generalize from a sample to a larger population.
In general, at least 15 subjects should participate in the experiment. They should not be directly involved either in picture or audio quality evaluation as part of their work and should not be experienced assessors.
In case of 1500-4000 scores the result might be inconsistent. If the number of observers less than 15, e.g. 6 then we expect binomial distribution of voting, and if the number of observers more than 15, then the distribution is normal.

REFERENCES

[1] ITU-R Recommendation BT.500-11: *Methodology for the subjective assessment of the quality of television pictures,* International Telecommunication Union, Geneva, Switzerland, pp. 2-24. 2002.
[2] BEERENDS J. G.: *Audio Quality Determination Based on Perceptual Measurement Techniques, Applications of Digital Signal Processing to Audio and Acoustics*, Ed. M. Kahrs and K. Brandenburg, Kluwer Academic Publishers, 1998.
[3] ITU-T REC. P.862: *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, ITU Recommendation, 2001.
[4] ITU-T REC. P.910: *Audiovisual quality in multimedia services, Subjective video quality assessment methods for multimedia applications*, ITU Recommendaton, 1999.
[5] ITU-R REC. BT.601-5: *Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios*, ITU Recommendation, 1995.
[6] ITU-T REC. G-114.: *One-way transmission time*, ITU Recommendation, 2003.