ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Fakulta elektrotechnická

DIPLOMOVÁ PRÁCE

2008 **Joan Enric Graells Moles**

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Fakulta elektrotechnická

Katedra telekomunikační techniky

METHODOLOGIES OF VIDEO QUALITY ASSESSMENT

červen 2008

Diplomant: Joan Enric Graells Moles
mentor: Ing. Ivan Pravda

Statement

I hereby declare that I am the only author of this paper with the help of my mentor and consultant and I only used the literature mentioned in the paper. I also declare that I have nothing against landing or publishing of my diploma thesis or its part with the approval of the department.

Čestné prohlášení

Prohlašuji, že jsem zadanou diplomovou práci zpracoval sám s přispěním vedoucího práce a konzultanta a používal jsem pouze literaturu v práci uvedenou. Dále prohlašuji, že nemám námitek proti půjčování nebo zveřejňování mé diplomové práce nebo její části se souhlasem katedry.

Datum: 2. 6. 2008

……………..…………………….

podpis diplomanta

Anotace:

Tahle práce je zamněrena na modelování parametrú kvality videosignálú v IP sítích. Pojednáva o hondocení kvality videa. Sou to metódy objektivní a taky subjektivní. Objektivní metódy rozdělujeme na rušivé a nerušivé.

Summary:

This Diploma Project is focus on model of quality parameters for transmission of video signals in IP networks. It consists in a scope about the methods of videos quality assessment. These methods are objective methods and subjective methods. About objective methods we can distinguish between intrusive methods and non-intrusive methods.

INDEX

1 ABBREVIATIONS

VIRIS	Video Reference Impairment System
PSQA	Pseudo-Subjective Quality Assessment Objective Speech Quality
PSQM	Objective Speech Quality Measures
SSIM	Structural Similarity Index Measures
PESQ	Perceptual Evaluation of Speech Quality
ACR	Absolute Category Rating
DCR	Degradation Category Rating
MOS	Mean Opinion Score
DMOS	Degradation Mean Opinion Score
SSCQE	Single-stimulus continuous quality evaluation
DSCQE	Double-stimulus continuous quality evaluation
CCI	Call clarity index
IMMD	In-service Non-intrusive Measurement Device
MSE	Mean Square Error
SNR	Signal to Noise Ratio

MNB	measuring normalizing blocks
EMBSD	the enhanced modified bark spectral distortion
VQM	Video Quality Measurement
TSSDM	Time/Space Structural Distortion Measurement
RISV	Reference Impairment System
SIF	Standard Intermediate Format [picture formats defined in ISO 11172 (MPEG-1): 352 lines × 288 pixels × 25 frames/s and 352 lines × 240 pixels × 30 frames/s]
RNN	The random neural network
OBQ	Output-Based Quality
HSM	*Human Visual System Model*
EC	RACE MOSAIC
DSIS	Double Stimulus Impairment Scale

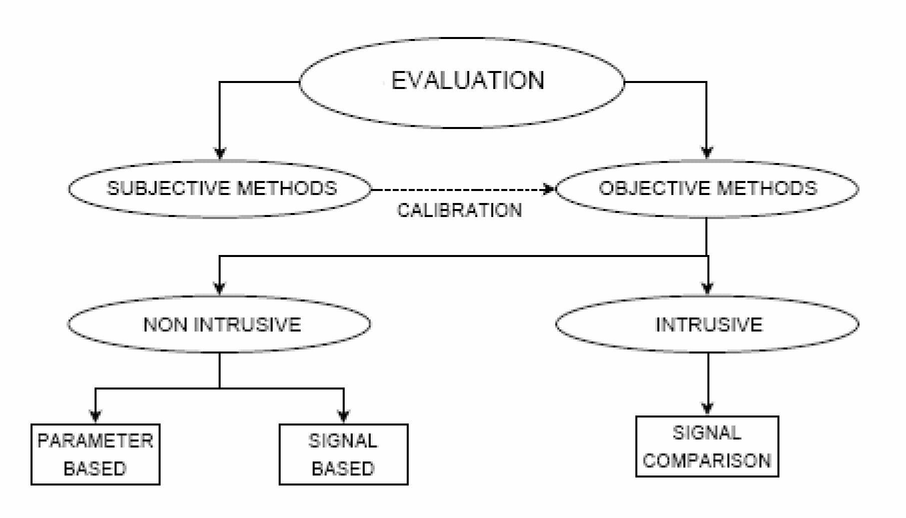
2 INTRODUCTION

Nowadays, the major problem of internet is assured to end-user QoS. This feature has always focused on network parameters (delay, jitter, BER…) however; there are more methods to assess it. We can distinguish between objective methods and subjective methods.

Subjective Methods are used for several purposes: selection of algorithms, ranking of audiovisual system performance and quality level evaluation during an audiovisual connection. All these methods consist in the evaluation of the average opinion that a group of people assign to different audio and video sequences in controlled tests.

Objective methods do not depend of the people; these methods take objective measures of the signal. So, we can say that they are more reliable than subjective methods. There are two kinds of the objective methods intrusive or non intrusive measures. Intrusive put extra data for performing the measures; it means that intrusive measures modified the throughput of the signal. These methods are based on the comparison of two signals, one reference (original) and one distorted (transmitted signal).

Non intrusive methods do not need the reference signal. They use the current signal to compare with the one before, which is also compared with its one before. Depending on the kind of information they use, non intrusive methods can be classified on signal based or parameter based. In the case of signal based methods, they just apply different algorithms to impairment the signal. On the other side the parameter based methods; network features as well as characteristics of the multimedia itself are taken as input. It is necessary a calibration for all different objective methods must have in some sense a calibration phase as their results are not in the same scale as subjective ones.


EVALUATION
SUBJECTIVE METHODS
CALIBRATION
OBJECTIVE METHODS
NON INTRUSIVE
INTRUSIVE
PARAMETER BASED
SIGNAL BASED
SIGNAL COMPARISON

3 OBJECTIVE METHODS

The assessment of perceived quality in multimedia services can be achieved by two different kinds of methodologies, either *subjective* or *objective* ones.

Objective methods do not depend on people, making them really attractive for automating the evaluation process. Objective PQoS measures can be either *intrusive* or *non-intrusive*. In network's context, intrusive means the injection of extra data (audio and/or video streams in multimedia networks, signals from now on) for performing the measure.

Intrusive methods are based on the comparison of two signals, one reference (original) and one distorted (e.g. by the network while transmitted). In general, this comparison is performed either in the time/space domain (simply comparing samples: Mean Square Error (MSE), Signal to Noise Ratio (SNR) or peak signal to noise ratio (PSNR)) or in the *perception domain*, using models of the human senses for improving the results. In this last category we find (for audio assessment) the Perceptual Speech Quality Measure (PSQM), the measuring normalizing blocks (MNB), the enhanced modified bark spectral distortion (EMBSD) and the perceptual evaluation of speech quality (PESQ).

If we focused on video, some of the developed tools are the Structural Similarity Index Measurement (SSIM) and the Institute for Telecommunications Science algorithms, the Video Quality Measurement (VQM) and the Time/Space Structural Distortion Measurement (TSSDM). Later some methods will be introduced.

3.1 NON INTRUSIVE METHODS

3.1.2 Video Reference Impairment System (VIRIS)

Reference Impairment System (RISV) for Video is used to generate the reference conditions necessary in order to characterize the subjective picture quality of video produced by compressed digital video systems. RISV can be also used to simulate the impairments from the compression of video sequences, independent of compression scheme.

RISV be capable of generate the following kinds of the impairments:

1) Artifacts due to conversions between analog and digital signals formats such as blurring and noise.
2) Impairments due to coding and compression such as jerkiness, edge busyness and block distorsion.
3) Artifacts due to transmission channel errors such as block errors.

From the viewer's point of view, the artifacts produced by the RISV should be a good approach of artifacts generated by digital video coding and transmission systems.

RISV have three possible applications:

1) Creating reference conditions in subjective tests of digital video systems to ensure that quality scenes presented to viewers covers the entire range of picture quality;

2) Defining standard video impairment levels that can be used to compare subjective test results; and

3) Quantifying the user-perceived quality of a video system with respect to a known reference.

Video Reference Impairment System (VIRIS) is specific implantation of a RISV, which simulates the impairments of block distortion, blurring, edge busyness, jerkiness and noise. This system is intended as a general video laboratory tool for evaluating the performance of digital video coders. VIRIS (sometimes referred to as VIRIS1, a more recent version of the Video Reference Impairment System) is a software system implemented in the C programming language to manipulate digital video files to introduce simulated coding impairments into a video image. It is designed to operate on SIF images but the method can be applied to other image formats such as CIF, QCIF and CCIR 601 format. VIRIS is in a preliminary stage and is useful only as a general purpose laboratory tool.

VIRIS can simulate the following impairments:

1) Block distortion, is often caused by a too coarse quantization during the compression process which results in a distortion or loss of high frequency components

2) Blurring is the reduction in sharpness of edges and spatial detail in a picture.

3) Edge busyness distortion is caused by too high a quantization level in a block containing both a smooth area and some pels with a significantly different average level.

Edge busyness, there are two different algorithms that were developed to implement the edge busyness impairments:

a) The edge busyness impairment is applied to only the vertical edges of objects;
b) The edge busyness impairment is applied to both the vertical and horizontal edges.

4) Noise, there are two types of noise produced by VIRIS, quantization noise and signal correlated noise.

a) Quantization noise is a noise sometimes created in the quantization step of the compression process.

b) Signal correlated, noise is a term utilized to describe the appearance resulting from the combination of edge busyness and mosquito noise impairments as seen in compressed video.

5) Jerkiness is defined as motion, originally smooth and continuous, perceived as a series of distinct "snapshots".

Calculation of the peak signal-to-noise ratio (PSNR)

PSNR give us the amount of peak error. We can say that PSNR estimate the error or deviation between two images with the same content but different compression or address. If we compare video sequence, then we refer to the frames. PSNR values oscillate between 20 and 50. PSNR is usually expressed in terms of the logarithmic decibel scale. PSNR is most commonly used as a measure of quality of reconstruction in image compression.
Currently each of the simulated impairments in VIRIS is objectively characterized by calculating the Peak Signal-To-Noise Ratio (PSNR) over each processed frame and the average PSNR over all the frames of a processed picture sequence. The unweighted PSNR is one measure for assessing the distortion of the processed sequence. For each processed frame, *k*, the RMS noise, *N rmsk*, is computed as:

$$N_{rms_k} = \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\left[U_{ij} - I_{ij}\right]^2}{N \cdot M}}$$

Where:

Uij luminance value of unimpaired pel at row *i* and column *j* location
Iij luminance value of impaired pel at row *i* and column *j* location
N 240 for SIF image
M 352 for SIF image
k *k*th frame

To calculate the PSNR over a sequence of *K* frames, the per-frame average noise, *Nrmsk*, is first determined in Equation (I.3-1). Next, the average noise, *Nrms*, across the sequence of *K* frames is calculated in Equation (I.3-2) as follows:

$$N_{rms} = \frac{1}{K} \sum_{k=1}^{K} N_{rms_k}$$

Finally, the PSNR is calculated using Equation (I.3-3) as follows:

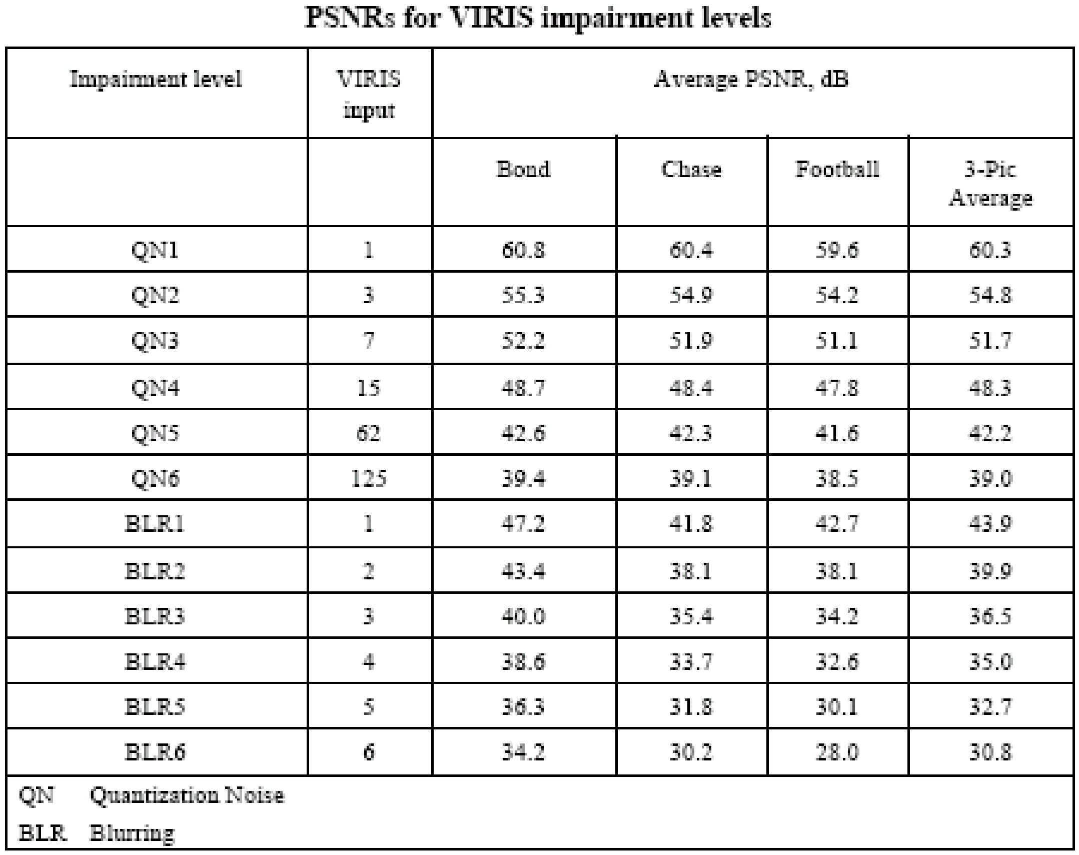
$$PSNR = 20 * \log_{10} \frac{S_p}{N_{rms}}$$

Where:

Sp Equals the number of levels to which the luminance intensity is quantized. The system on which VIRIS operates quantizes the luminance pels to 8 bits, or *Sp* = 255. Because the unweighted PSNR as described above is loosely correlated to the human visual system, a more accurate objective measure may be required. This is a subject for future study.

<u>PSNR applications in VIRIS</u>

VIRIS test conditions for the noise and blurring impairments, each picture sequence was processed by VIRIS 12 times (2 impairments x 6 impairment levels). The following table shows the PSNR calculated by VIRIS for each of the three pictures and the average PSNR total for each impairment level.



PSNRs for VIRIS impairment levels

Impairment level	VIRIS input	Average PSNR, dB			
		Bond	Chase	Football	3-Pic Average
QN1	1	60.8	60.4	59.6	60.3
QN2	3	55.3	54.9	54.2	54.8
QN3	7	52.2	51.9	51.1	51.7
QN4	15	48.7	48.4	47.8	48.3
QN5	62	42.6	42.3	41.6	42.2
QN6	125	39.4	39.1	38.5	39.0
BLR1	1	47.2	41.8	42.7	43.9
BLR2	2	43.4	38.1	38.1	39.9
BLR3	3	40.0	35.4	34.2	36.5
BLR4	4	38.6	33.7	32.6	35.0
BLR5	5	36.3	31.8	30.1	32.7
BLR6	6	34.2	30.2	28.0	30.8
QN Quantization Noise BLR Blurring					

We can see the variation across the picture sequence for blurring and noise impairments. The PSNR variation of the blurring is on the order of 1 of 7 dB depending of the impairment level. The PSNR variation across the picture sequence of the noise impairment is considerably less, on the order of 1 dB.

The VIRIS input works the following way:

The data file controls the level of impairments added to a video sequence. There are essentially six impairment level control parameters to operate VIRIS:

1) The block distortion level, input as a whole number and represents 0.1 per cent of the 1320 total blocks to be changed (result is rounded to nearest whole number). For example, a block distortion level of 10 results in 13 blocks that are impaired (10 ´ 0.001 ´ 1320 = 13.2 which rounds to 13).

2) The quantization noise level, input as a whole number and representing 0.001 per cent of the 84 480 total luminance pel values (rounded to nearest whole number) to be changed. For example, a noise level of 10 results in a change in the luminance value of 8 pels (10 ´ 0.00001 ´ 84 480 = 8.4 which rounds to 8).

3) The signal correlated noise level, input as a whole number, which represents the range of luminance values by which a pel can be altered. For instance, a level of 10 indicates that the luminance value of pels classified as edges can be randomly altered by a range from -10 to 10 luminance levels from its original value.

4) The blurring level, input as whole numbers between 0 and 6 with 0 indicating no blurring. The numbers 1 to 6 select low-pass filters with cut-off frequencies of 1.5, 1.0, 0.75, 0.5, 0.375 and 0.25 MHz.

5) The edge busyness echo displacement which consists of a whole number between 0 and 3, with 0 selecting no edge busyness simulation and 1, 2 or 3 selecting 0.5, 0.75 and 0.375 msecs displacements.

6) The edge busyness echo amplitude level. The input data item is an integer number between -30 and -1 and represents the filter tap coefficient value for the particular echo displacement selected.

Plots of the PSNR, averaged across the three pictures, versus the input to VIRIS for the two impairment levels are shown in Figures 1 and 2. An exponential function fitted to the data is also shown on each of the plots to provide a pathway from PSNR to VIRIS input for each of the impairments.

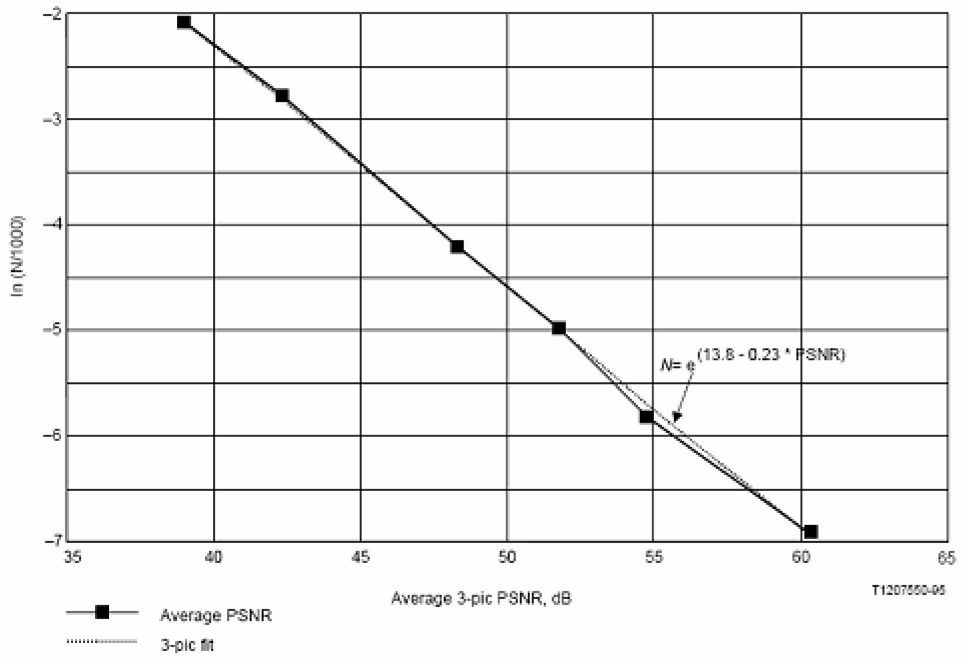


FIGURE 1

Quantization noise PSNR vs QN data input to VIRIS

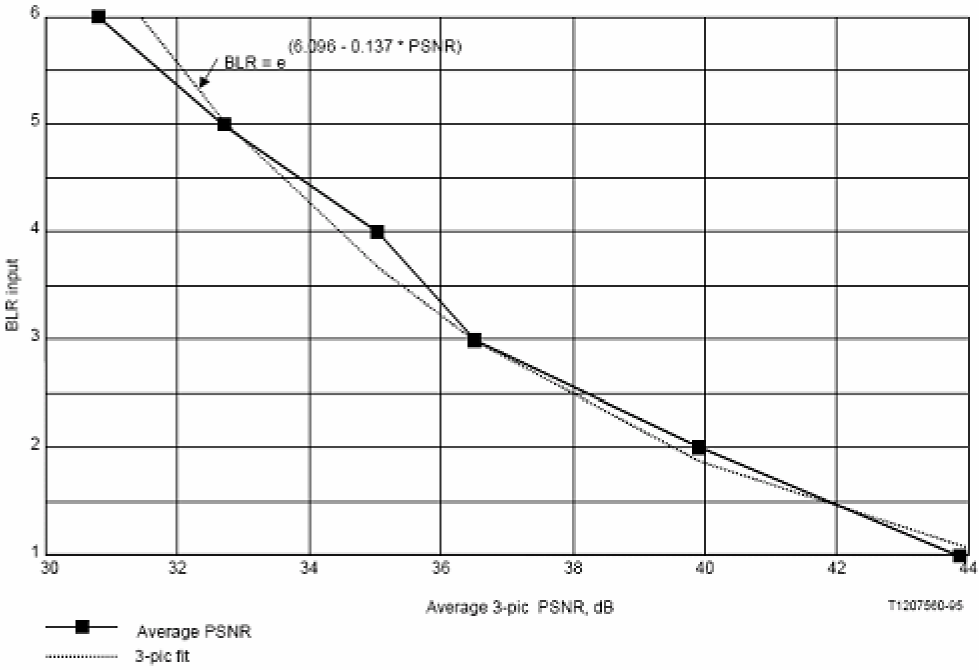


FIGURE 2

Blurring PSNR vs BLR data input to VIRIS

3.1.2 Pseudo-Subjective Quality Assessment (PSQA)

This non intrusive method allows to measure the quality of a video flow (or audio, or multimedia). Can be unidirectional or bidirectional, perceived by the user, accurately and efficiently (in particular, in real time if needed). To provide a quantitative assessment PSQA needs to use some standard range (usually, a MOS-like evaluation). PSQA builds the quality as a function of two types of parameters: source-oriented parameters (parameters associated with the source, the stream, the codec used…), and network-oriented parameters (mainly, those associated to the possible losses, also delays, jitter…).

To implement PSQA, three main steps must be followed:

1) A set of (a priori) quality-affecting parameters must be selected;
2) A (set of) subjective tests session(s) must be performed, and
3) A RNN must be chosen and then trained and validated.

Let us briefly describe them in more detail:

PSQA works by learning how humans react to the communication from the quality point of view, through a set of selected variables. These must be *measurable* (at a low cost) parameters expected to have a significant impact on the perceived quality. Their selection largely depends on the target application.

An important thing to consider when choosing the parameters, is that using more parameters means that more subjective tests need to be carried out in order to train the RNN, and this puts practical limits (in terms of cost, mostly) to the parameter choice. The implementer needs therefore to prioritize those parameters that in his experience are likely to have the biggest impact on quality.

It should be noted that some parameters are best represented by random variables while others are not. For those that are not seen as random variables a range of possible values must be selected for the tests. For the random variables, a distribution must be selected and then the range of values for the parameters that characterize the selected distribution.

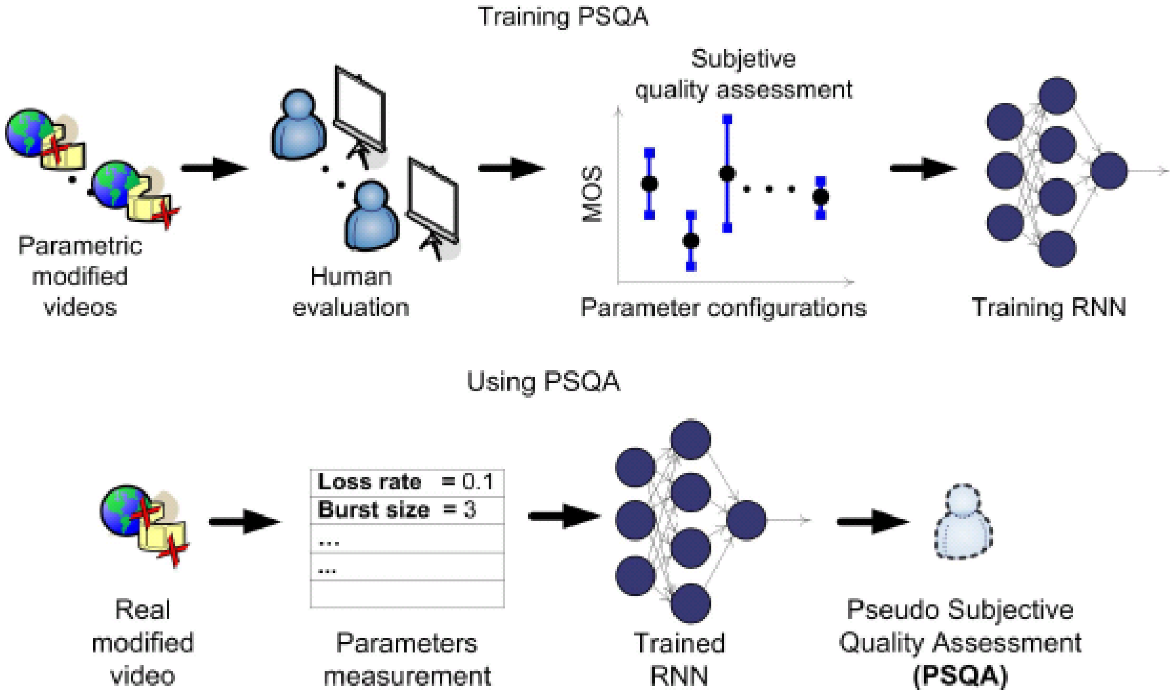
For step 2), we need a Video tool and a module that is capable of emulating network conditions according to the parameters chosen (e.g. packet loss and delay). A panel of human subjects is paired, establishing an interactive video connection for each pair. Then, we select different values combinations of selected variables (called *configurations*), and for each of them we emulate the corresponding network conditions. As the number of possible parameter configurations is typically large, only a subset of them are used during the subjective tests, and thus to train the RNN. The RNN's ability to generalize is then exploited by PSQA to provide accurate MOS estimations for the rest of the parameter space.

The human subjects evaluate the quality of a video in those conditions and using many pairs of subjects for each of the selected configurations, we obtain a MOS value. Each subject assigns a conversational quality score to each conversation session, from a predefined quality scale [M_{min}, M_{max}]. The parameter values for a configuration must not be known to the subjects and they should not establish any relation between the quality their perceive and the corresponding parameters values.

After performing a screening and statistical analysis in order to remove the grading of the individuals who might have given unreliable results, the average of the scores given by the remaining subjects to each configuration is computed.

After step 2) we have a database (actually a table) associating the values defining each configuration with the corresponding Mean Opinion Score (MOS). Step 3) consists of finding a real function of the selected parameters that provides a value close to the MOS given by the panel of observers. For this purpose, our RNN works as any standard Neural Network: a part of the data is used for training, the rest for validating the network.

Once the RNN has been trained, the validation process ensures that it is able to provide accurate results in a generic environment, and not only for the cases considered during training. The validation itself is simple; it consists of comparing the results given by the RNN to the actual MOS values for a set of configurations which was not used during the training phase. This also provides us with a measure of the quality assessment performance (e.g. in terms of correlation with subjective scores for previously unknown parameter configurations).



3.1.3 In-service Non intrusive Measurement Device (INMD)

1 Introduction

INMD stands for In-service Non-intrusive Measurement Device. It is a passive voice quality monitoring method based on ITU-T P.561. Two types of measurements are covered by INMD:

1) speech and noise characterization;
2) echo characterization.

Sage's current implementation of INMD on the 960 platform focuses only on echo characterization. More specifically, once the presence of echo is detected, Sage's INMD in 960 will report in real time the detected echo level and echo delay. A graphical snapshot of the reference and echo signals is also displayed as further visual confirmation. If the monitored DS1 are PRI-ISDN lines, then the source and destination phone numbers associated with the monitored DS0 channel are also presented.

2 Test Configurations

When using INMD, a user should connect Sage's 960 to the network as shown in Figure 10.

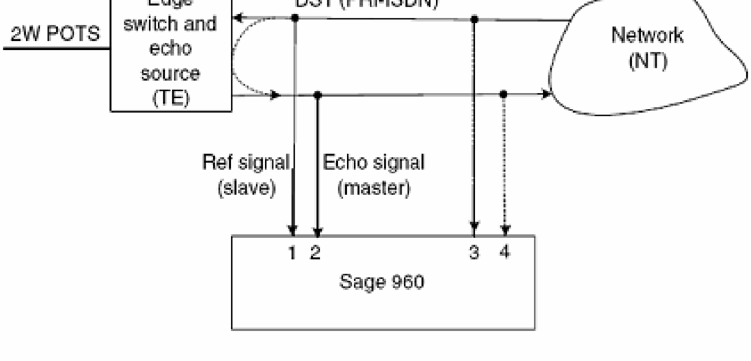


Figure 10: Connection diagram for running INMD on Sage's 960

It Has 4 DS1 spans, numbered as PCM1, PCM2, PCM3 and PCM4. Internally, PCM1 and PCM2 are a pair controlled by one embedded processor, whereas PCM3 and PCM4 are another pair controlled by another embedded processor. Since INMD entails precise relative delay measurement between two input signals, Sage's 960 must be configured in the following ways:

1) Two DS1 spans are used for INMD, although only the receiving ports need to be connected. The two DS1 spans must be either the PCM1 and PCM2 pair, or the PCM3 and PCM4 pair. Other combinations such as PCM1 and PCM4 or PCM2 and PCM3 etc are not allowed.

2) The chosen DS1 pair must be configured to "DUAL MONITOR" mode. When performing span configuration through 960's GUI, one only needs to (and must) configure the first span (PCM1 or PCM3) of each DS1 pair. The partner DS1 span (PCM2 or PCM4) will be configured automatically for you. A user configuration on the second DS1 span (PCM2 or PCM4) has no effect. In "DUAL MONITOR" mode, the internal software will always set the DS1 clock source to "EXTERNAL LOOP CLOCK" regardless what the user selects on the GUI.

3) The INMD on 960 includes two seemingly separate tests that in practice must be treated as an atomic pair. The 960 GUI names the tests as "INMD-SLAVE" and "INMD-MASTER". The "INMD-SLAVE" must be run on a DS0 channel of the first DS1 span (PCM1 or PCM3) that is physically connected to the incoming reference signal. The "INMD-MASTER" must be run on the same DS0 channel of the second DS1 span (PCM2 or PCM4) that is physically connected to the incoming echo signal. The measurement results are only available at the "INMD-MASTER" side.

4) If the DS1 lines being monitored are PRI-ISDN lines, then the D-channel must be specified correctly when performing the span configuration in order for

INMD to intercept the source and destination phone numbers associated with the DS0 channel(s) being monitored. If one is also interested in decoding all the ISDN call messages, one should also specify the TE and NT modes correctly. As in "TERMINATE" mode, the 960 DS1 span should be set to match the incoming NT or TE mode. For example, as shown in Figure 1, if the signal goes into the receiving port of PCM1 is from the NT equipment, and then 960's PCM1 must be set to TE mode. Internally, the PCM2 will automatically be set to NT mode to match the signal from the TE equipment.

3 Operation principles

Simply speaking, INMD detects echo by principle of cross-correlation. More specifically, as implemented in Sage's 960, both "INMD-SLAVE" and "INMD-MASTER" have an internal signal analyzing window of 256ms long (2048 samples at 8000Hz sampling rate). The presence of echo is declared when all of the following conditions are met:

1) The echo signal side ("INMD-MASTER" side) analyzing window captured some signal (*e*(*n*)) whose power level (*Pe*) is greater than -60 dBm.
2) The reference signal side ("INMD-SLAVE" side) analyzing window also captured some signal (*r*(*n*)) whose power level (*Pr*) is greater than *Pe*.

3) Circular cross-correlation is performed between *r*(*n*) and *e*(*n*):

$$cor(n) = \sum_{m=0}^{2047} r(m)e((m+n)\%2048), n = 0, 1, 2, \ldots, 2047$$

Where:

% Represents modular (remainder) calculation. In actual implementation, the circular correlation is obtained through two forward FFTs and one inverse FFT. The FFT-based approach is far more efficient than the direct brute-force computation. An example of the signal *r*(*n*), *e*(*n*) and their circular-cross-correlation is shown in Figure 11. In Figure 11, the correlation trace has been normalized as

$$cor_{normalized}(n) = \frac{cor(n)}{\sqrt{\sum r^2(n) \sum e^2(n)}}$$

.

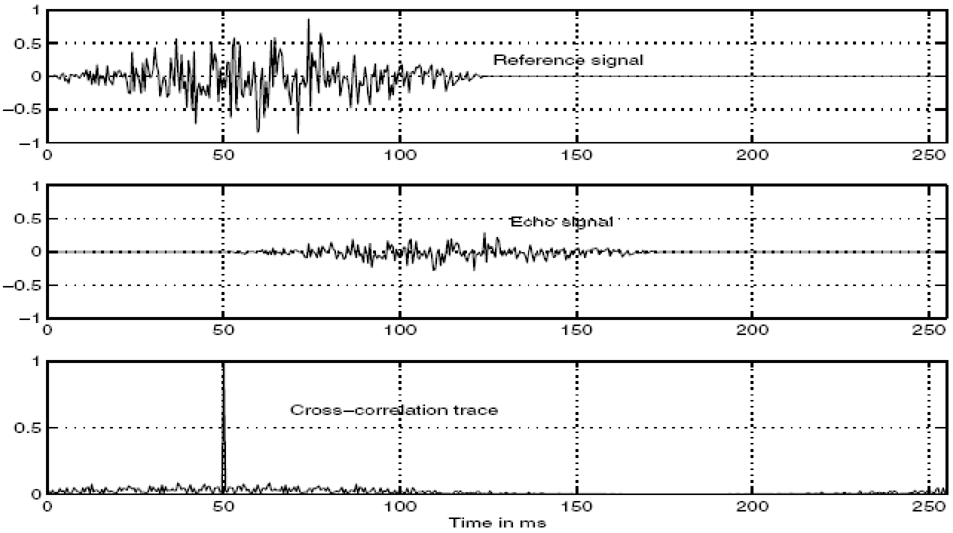


Figure 11: An exemplary reference signal, echo signal and their normalized cross correlation trace. In this example, the echo path is flat. Echo delay is 50 ms and echo level is -10 dB.

4) Check to see if *r*(*n*) and *e*(*n*) are narrow-band tone signals. If yes (they are tone signals), then ignore the captured signals, and restart capturing new signals again. Mathematically speaking, the cross-correlation *cor*(*n*) will truly resemble the echo impulse response *h*(*n*) only if the reference signal *r*(*n*) is “white noise” that has flat spectrum in frequency domain. If *r*(*n*) is a “highly-colored” narrow-band tone signal, then the obtained *cor*(*n*) does not resemble the echo impulse response *h*(*n*) because the excitation signal *r*(*n*) did not excite all aspects (all spectrums) of the “system” (echo path). More specifically, the echo delay measurement will be affected. This is a classical signal processing problem, not a problem specific to Sage’s implementation. If *r*(*n*) and *e*(*n*) are determined to be valid voice-like complex signals, then the INMD algorithm will proceed to the following steps.

5) Locate the peak on |*cor*(*n*)|, and record the index *ix* that corresponds to the peak. Then re-compute the "linear" cross-correlation at the following 8 points (within 1 ms span):

$$R(n) = \sum_{m=0}^{2047-ix-n} r(m)e(ix+m+n), n = 0,1,2,3,4,5,6,7$$

Then calculate the following ratio:

$$ratio = \frac{\sum_{n=0}^{7} R^2(n)}{\sum_{n=0}^{2047-ix} r^2(n) \sum_{n=ix}^{2047} e^2(n)}$$

If this ratio is greater than 0.36, then the presence of an echo is detected, and *ix/*8 is the echo delay in ms. The echo level is computed as:

$$EchoLevel_{dB} = 10\log 10(\frac{\sum_{n=ix}^{2047} e^2(n)}{\sum_{n=0}^{2047-ix} r^2(n)})$$

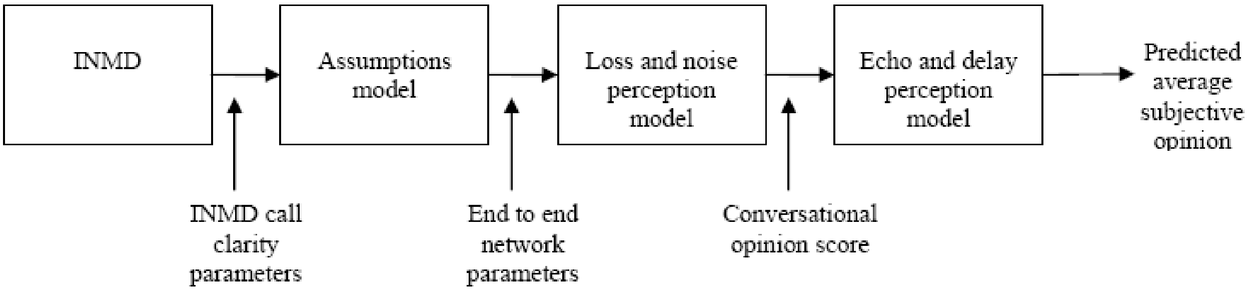
The reasons for performing the linear cross-correlation at 8 points and later summing up their total "contribution" is to account for the dispersion effect that is always present on a real echo path.

6) The delay adjusted signals *r*(*n*) and *e*(*n* + *ix*) are sent to the display in linear scale. Since the analyzing window size is set to 256ms, an echo with delay longer than 256ms will not be detected. The presence of echo is detected and reported within 256ms.

3.1.4 Call Clarity Index (CCI)

The call clarity index or the CCI model was developed by British Telecom for interpreting INMD measurement to predict voice quality on a call to call basis. It supervises call performance by searching for and ranking mismatches in delay, echo and noise in the network.

The CCI method combines the measured data into a single call clarity index, CCI using a model of human perception that is calibrated against subjective quality scores. The index is expressed in a conversational quality scale and represents the mean opinion of the tested connection. The CCI algorithm can be loaded onto test equipment or deployed in network components like echo cancellers, voice quality enhancers and switching platforms.


INMD
Assumptions model
Loss and noise perception model
Echo and delay perception model
Predicted average subjective opinion
INMD call clarity parameters
End to end network parameters
Conversational opinion score

The model of CCI contains of three steps:

1) The assumptions model: This first step makes estimates about parameters related to the network and users that cannot be measured by the INMD. With these assumptions a complete model of the tested system can be made.

2) The loss and noise model: This block takes human perception factors like frequency selectivity and noise on the connection into account. This stage also considers effects of side tone, noise masking and room noise. The output of this step is a single score representing conversational speech quality.

3) Echo and delay: The next part of the model uses complex mathematical calculations to add the effects of echo and delay into a final output value.

Multiple CCI values should always be used to assure an accurate statistical averaging. The mean value can then be expected to represent all degradations in a call correctly.

3.2 INTRUVISE METHODS

3.2.1 Perceptual Speech Quality Measures (PSQM)

An ideal objective speech quality measure would be able to assess the quality of distorted or degraded speech by simply observing a small portion of the speech in question, with no access to the original speech. One attempt to implement such an objective speech quality measure was the Output-Based Quality measure [Jin and Kubicheck, 1996]. To arrive at an estimate of the distortion using the output speech alone, the OBQ needs to construct an internal reference database capable of covering a wide range of human speech variations. It is a particularly challenging problem to construct such a complete reference database. The performance of OBQ was unreliable both for vocoders and for various adverse conditions such as channel noise and Gaussian noise.

Current objective speech quality measures base their estimates on both the original and the distorted speech even though the primary goal of these measures is to estimate Mean Opinion Score (MOS) test scores where the original speech is not provided.

Although there are various types of objective speech quality measures, they all share a basic structure composed of two components as shown in Figure 3.

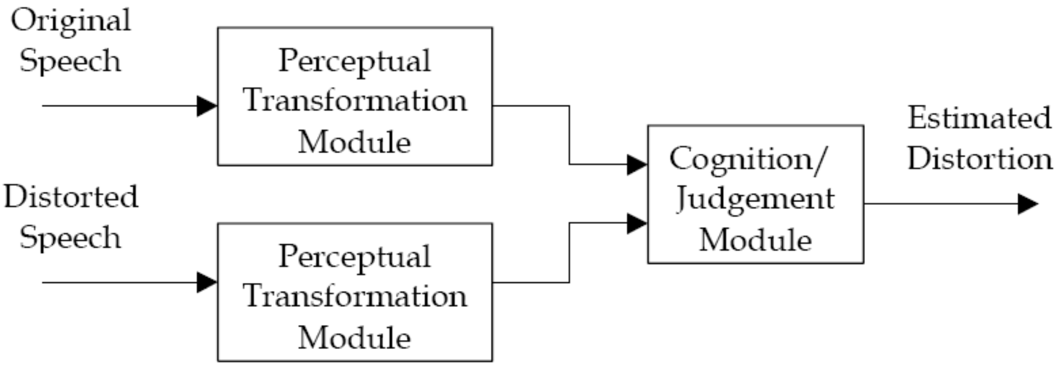

Original Speech
Perceptual Transformation Module
Distorted Speech
Perceptual Transformation Module
Cognition/ Judgement Module
Estimated Distortion

Figure 3: Basic Structure of Objective Speech Quality Measures.

The first component is called the perceptual transformation module. In this module, the speech signal is transformed into a perceptually relevant domain such as temporal, spectral, or loudness domain. The choice of domain differs from measure to measure. Current objective measures use psychoacoustic models, and their performance has been greatly improved compared to the previous measures that did not incorporate psychoacoustic responses. The second component is called the cognition/judgment module. This module models listeners' cognition and judgment of speech quality in the subjective test.

After the original and the distorted speech are converted into a perceptually relevant domain, through the perceptual transformation module, the cognition/judgment module compares the two perceptually transformed signals in order to generate an estimated distortion. Some measures use a simple cognition/judgment module like average Euclidean (kind of function) distance while others use a complex one such as an artificial neural network or fuzzy logic.

Recently, researchers in this field have been focusing on this module because they realize that a simple distance metric cannot cover the wide range of distortions encountered in modern voice communication systems. The potential benefits of including this module are not yet fully understood.

Objective speech quality measures can be classified according to the perceptual domain transformation module being used, and these are: time domain measures, spectral domain measures, and perceptual domain measures.

Perceptual Speech Quality Measures (PSQM) is an objective approach to measure the quality of a telephone call and is based on ITU standard P. 861. PSQM defines an algorithm through which a computer can derive scores based on levels of distortions to a sound file between the sent and received audio tracks. PSQM, which uses an inverted scale from MOS, provides a reasonably close correlation to MOS, with the limitation that PSQM was not originally designed for packet telephony networks and therefore only partially accounts for packet loss and jitter. Despite PSQM's limitations, it remains the method of choice because it is quantitative, repeatable and scalable. Thus, it is a preferred scale for use in conjunction with certain embodiments.

3.2.2 Perceptual Evaluation of Speech Quality (PESQ)

The basic idea behind the PESQ algorithm is the same as the one used in the development of the PSQM algorithm. Figure 4 gives an overview of the basic philosophy used in PESQ. A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the output of the device under test with the input, using alignment information as derived from the time signals in the time alignment module.

In PESQ the original and degraded signals are mapped onto an internal representation using a perceptual model. The difference in this representation is used by a cognitive model to predict the perceived speech quality of the degraded signal. This perceived listening quality is expressed in terms of Mean Opinion Score, an average quality score over a large set of subjects. Most of the subjective experiments used in the development of PESQ used the ACR (Absolute Category Rating). In these types of experiments subjects do not get a reference speech signal to judge the quality and some types of distortion, like missing words, sometimes go unnoticed in such experiments. Experiments in which this missing word phenomenon was clear were used only to a small extent in the optimization of PESQ. In these cases a lower correlation between subjective and objective results is likely.

An essential difference with the PSQM method is that the time alignment, necessary for the correct comparison of the matching parts of original and degraded, is an integrated part of the new standard.

The internal representations, that are used by the PESQ cognitive model to predict the perceived speech quality, are calculated on the basis of signal representations that use the psychophysical equivalents of frequency and intensity. This idea was also used in the PSQM method; however the psycho-acoustic parameters used in the mapping are now more in line with literature. A minor disappointment is that the psychoacoustic model that is used in PESQ still has no correct modeling of masking caused by smearing in the time-frequency plane. Although masking models were implemented and tested in several stages of the development it never improved correlations between subjective and objective scores. This counterintuitive result was already presented in and the first ideas towards incorporating masking into a speech quality model are given in. A final solution to this problem is still under study.

The most important difference, besides the inclusion of a perceptual time alignment, between PSQM and PESQ is found in the cognitive part of the model. In PSQM two major cognitive effects are modeled in order to get high correlations between objective and subjective scores: asymmetry and different weighting of distortions during speech and silence.

.

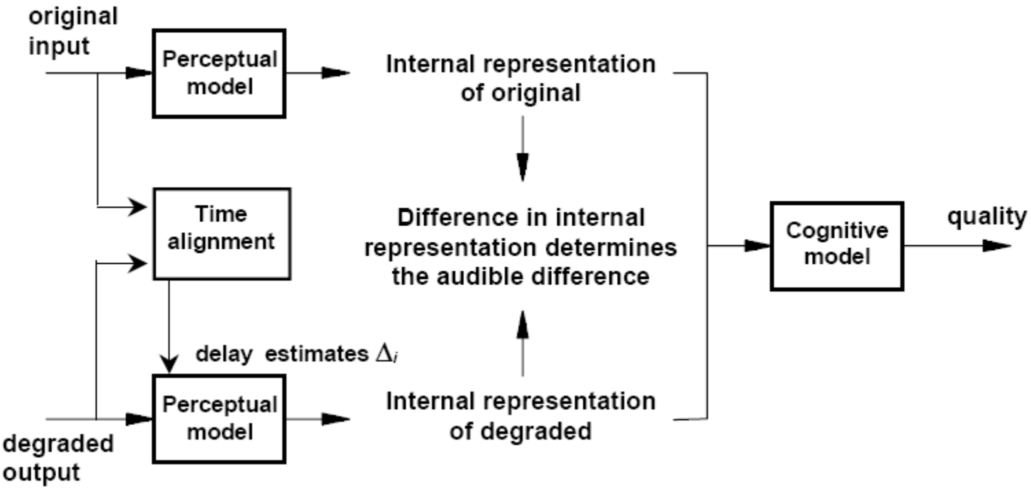

original input
Perceptual model
Internal representation of original
Time alignment
Difference in internal representation determines the audible difference
Cognitive model
quality
delay estimates Δ_i
Perceptual model
Internal representation of degraded
degraded output

Figure 4: Basic overview of the basic philosophy used in PESQ

3.2.3 The Structural Similarity Index (SSIM)

The most fundamental principle underlying structural approaches to image quality assessment is that the HVS is highly adapted to extract structural information from the visual scene, and therefore a measurement of structural similarity (or distortion) should provide a good approximation to perceptual image quality. Depending on how *structural information* and *structural distortion* are defined, there may be different ways to develop image quality assessment algorithms. The structural similarity (SSIM) index is a specific implementation from the perspective of image formation.

To understand the intuition of the SSIM index method, first we have to examine the image space described. In a reference image (original “Einstein” image) is represented as a vector in the image space. Any image distortion can be interpreted as adding a distortion vector to the central reference image vector. In particular, the distortion vectors with the same length define an equal-mean squared error (MSE) hyper sphere in the image space.

However, as shown in the Einstein figure, images that reside on the same hyper sphere may have dramatically different visual quality. This implies that the length of a distortion vector does not suffice as a useful image quality measure, and that the directions of these vectors have more important perceptual meanings.

Recall that the luminance of the surface of an object being observed is the product of the illumination and the reflectance, but the structures of the objects in the scene are independent of the illumination. Consequently, we wish to separate the influence of illumination from the remaining information that represents object structures. Intuitively, the major impact of illumination change in the image is the variation of the average local luminance and contrast, and such variation should not have a strong effect on perceived image quality. This is confirmed by Einstein image, where the images with only luminance or contrast changes have much better quality than the other images with severe “structural” distortions.

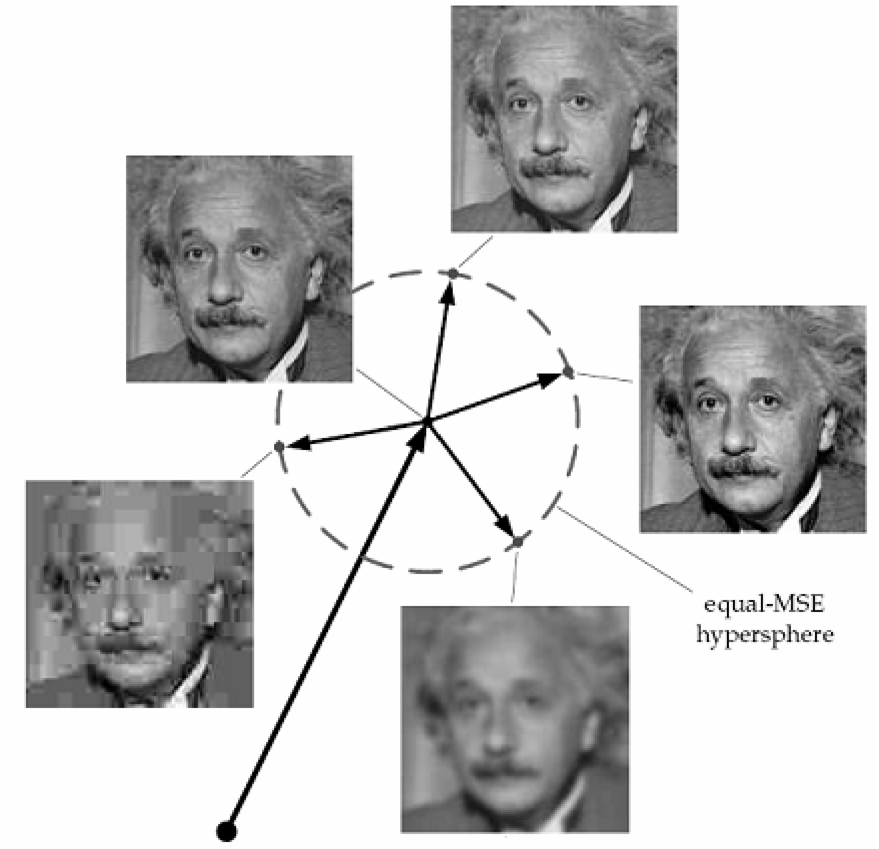


Image Quality Assessment Using a Structural Similarity Index

The SSIM indices measure the structural similarity between two image signals. If one of the image signals is regarded as of perfect quality, then the SSIM index can be viewed as an indication of the quality of the other image signal being compared. When applying the SSIM index approach to large-size images, it is useful to compute it locally rather than globally.

The reason is manifold. First, statistical features of images are usually spatially no stationary. Second, image distortions, which may or may not depend on the local image statistics, may also vary across space. Third, due to the non-uniform retinal sampling feature of the HVS, at typical viewing distances, only a local area in the image can be perceived with high resolution by the human observer at one time instance.

Finally, localized quality measurement can provide a spatially varying quality map of the image, which delivers more information about the quality degradation of the image. Such a quality map can be used in different ways. It can be employed to indicate the quality variations across the image. It can also be used to control image quality for space-variant image processing systems, e.g., region-of-interest image coding.

4 Subjective Evaluation

In this kind of test a group of people is asked for the quality of a group of sequences (audio or video in our case). There are mainly two types of tests in this category, the Absolute Category Rating (ACR) and the Degradation Category Rating (DCR). In a DCR test, people compare the original sequence with the distorted one and then score the perceived degradation. The output of this test is the Degradation Mean Opinion Score (DMOS). In an ACR test, people evaluates only the distorted sequence and scores its quality; in this case the output is the Mean Opinion Score (MOS).

The EC project, RACE MOSAIC, was set up to find ways of overcoming specific digital picture quality issues (e.g. content-dependent encoding performance, codec cascading and dynamic statistical multiplexing). A new methodology has been designed to allow subjective assessment of both picture and service quality, in conditions that are closer to the actual home environment. New method – known as Single Stimulus Continuous Quality Evaluation and, more particularly, “SSCQE Stage 1” which was recently introduced in ITU-R Recommendation BT.500-7. The double-stimulus DSCQE methodology – recently studied in the EC project, ACTS TAPESTRIES – is an adaptation of SSCQE. DSCQE has been proposed to the MPEG-4 group to address the specific issue of error-robustness evaluation, and is briefly described here.

4.1 Absolute Category Rating (ACR)

The Absolute Category Rating method is a category judgment where the test sequences are presented one at a time and are rated independently on a category scale. (This method is also called Single Stimulus Method.) The method specifies that after each presentation the subjects are asked to evaluate the quality of the sequence shown.

The time pattern for the stimulus presentation can be illustrated by Figure 5. The voting time should be less than or equal to 10 s, depending upon the voting mechanism used. The presentation time may be reduced or increased according to the content of the test material.

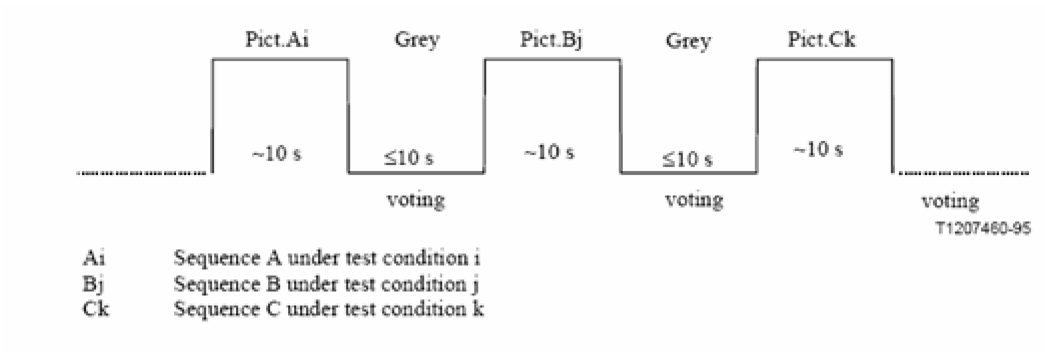

Pict.Ai
Grey
Pict.Bj
Grey
Pict.Ck
~10 s
≤10 s
~10 s
≤10 s
~10 s
voting
voting
voting
T1207460-95
Ai Sequence A under test condition i
Bj Sequence B under test condition j
Ck Sequence C under test condition k

Figure 5: Stimulus Presentation in the ACR method

The following five-level scale for rating overall quality should be used:

5 Excellent
4 Good
3 Fair
2 Poor
1 Bad

If higher discriminative power is required, a nine-level scale may be used. Such dimensions may be useful for obtaining more information on different perceptual quality factors when the overall quality rating is nearly equal for certain systems under test, although the systems are clearly perceived as different. For the ACR method, the necessary number of replications is obtained by repeating the same test conditions at different points of time in the test.

4.2 Degradation Category Rating (DCR)

The Degradation Category Rating implies that the test sequences are presented in pairs: the first stimulus presented in each pair is always the source reference, while the second stimulus is the same source presented through one of the systems under test. (This method is also called the Double Stimulus Impairment Scale method.) When reduced picture formats are used (e.g. CIF, QCIF, SIF), it could be useful to display the reference and the test sequence simultaneously on the same monitor. The time pattern for the stimulus presentation can be illustrated by Figure 6. The voting time should be less than or equal to 10 s, depending upon the voting mechanism used. The presentation time may be reduced or increased according to the content of the test material.

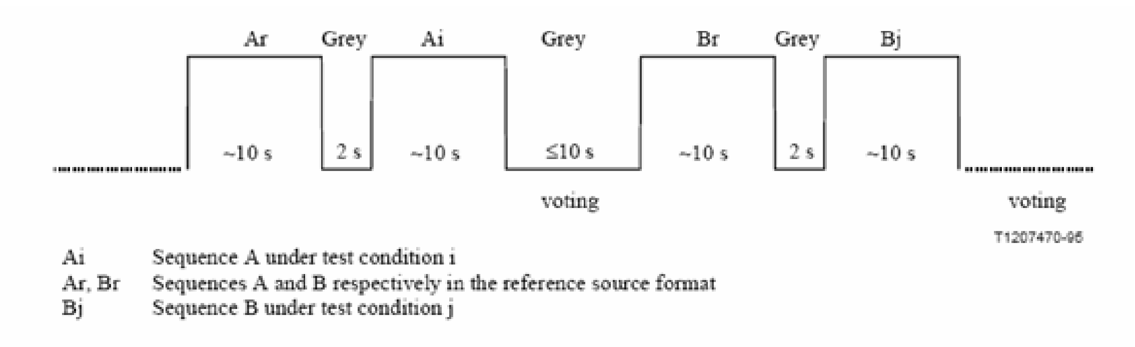

Ar
Grey
Ai
Grey
Br
Grey
Bj
~10 s
2 s
~10 s
≤10 s
~10 s
2 s
~10 s
voting
voting
T1207470-96
Ai Sequence A under test condition i
Ar, Br Sequences A and B respectively in the reference source format
Bj Sequence B under test condition j

Figure 6: Stimulus Presentation in the DCR method

In this case the subjects are asked to rate the impairment of the second stimulus in relation to the reference.

The following five-level scale for rating the impairment should be used:

5 Imperceptible
4 Perceptible but not annoying
3 Slightly annoying
2 Annoying
1 Very annoying

The necessary number of replications is obtained for the DCR method by repeating the same test conditions at different points of time in the test.

4.3 Mean Opinion Score (MOS)

In voice and video communication, quality usually dictates whether the experience is a good or bad one. Besides the qualitative description we hear, like 'quite good' or 'very bad', there is a numerical method of expressing voice and video quality. It is called Mean Opinion Score (MOS). MOS gives a numerical indication of the perceived quality of the media received after being transmitted and eventually compressed using codecs.

MOS is expressed in one number, from 1 to 5, 1 being the worst and 5 the best. MOS is quite subjective, as it is based figures that result from what is perceived by people during tests. However, there are software applications that measure MOS on networks, as we see below. Taken in whole numbers, the numbers are quite easy to grade.

MOS and Corresponding Speech Quality

Rating	Speech Quality
5	Excelent
4	Good
3	Fair
2	Poor
1	Bad

5) Perfect. Like face-to-face conversation or radio reception.
4) Fair. Imperfections can be perceived, but sound still clear. This is (supposedly) the range for cell phones.
3) Annoying.
2) Very annoying. Nearly impossible to communicate.
1) Impossible to communicate

The values do not need to be whole numbers. Certain thresholds and limits are often expressed in decimal values from this MOS spectrum. For instance, a value of 4.0 to 4.5 is referred to as high quality and causes complete satisfaction. This is the normal value of PSTN and many VoIP services aim at it, often with success. Values below 3.5 are termed unacceptable by many users.

A certain number of people are sat and are made to hear some audio. Each one of them gives a rating from within 1 to 5. Then an arithmetic mean (average) is calculated, giving the Mean Opinion Score. When conducting MOS test, there are certain phrases that are recommended to be used by the ITU-T. They are:

1) You will have to be very quiet.
2) There was nothing to be seen.
3) They worshipped wooden idols.
4) I want a minute with the inspector.
5) Did he need any money?

MOS can simply be used to compare between VoIP services and providers. But more importantly, they are used to assess the work of codecs, which compress audio and video to save on bandwidth utilization but with a certain amount of drop in quality. MOS tests are then made for codecs in a certain environment.

There are however certain other factors that affect the quality of audio and video transferred. These factors are not supposed to be accounted for in MOS values, so when determining the MOS for a certain codec, service or network, it is important that all the other factors are favorable to the maximum for a good quality, for MOS values are assumed to be obtained under ideal conditions.

Since the manual/human MOS tests are quite subjective and less than productive in many ways, there are nowadays a number of software tools that carry out automated MOS testing in a VoIP deployment. Although they lack the human touch, the good thing with these tests is that they take into account all the network dependency conditions that could influence voice quality.

4.4 Degradation Mean Opinion Score (DMOS)

In the DMOS, listeners are asked to rate annoyance or degradation level by comparing the speech utterance being tested to the original (reference). So, it is classified as the Degradation Category Rating (DCR) method. The DMOS provides greater sensitivity than the MOS, in evaluating speech quality, because the reference speech is provided. Since the degradation level may depend on the amount of distortion as well as distortion type, it would be difficult to compare different types of distortions in the DMOS test. The following Table describes the five DMOS scores and their corresponding degradation levels.

DMOS and Corresponding Degradation Levels:

Rating	Degradation Level
5	Inaudible
4	Audible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Thorpe and Shelton (1993) compared the MOS with the DMOS in estimating the performance of eight codecs with dynamic background noise [Thorpe and Shelton, 1993]. According to their results, the DMOS technique can be a good choice where the MOS scores show a floor (or ceiling) effect compressing the range. However, the DMOS scores may not provide an estimate of the absolute acceptability of the voice quality for the user.

4.5 Single-stimulus continuous quality evaluation (SSCQE)

It was originally designed to perform time efficient subjective quality evaluations of digital services, in conditions near to the home environment. It also overcomes most of the difficulties encountered when using conventional double stimulus methodologies to assess the picture quality of digital systems. The use of high levels of compression, to varying limits, results in artifacts which are neither regular nor consistent.

The MOSAIC Consortium therefore proposed to use test sequences longer than the 10-second sequences of, for example, the Double-Stimulus Continuous Quality Scale (DSCQS) and the Double Stimulus Impairment Scale (DSIS) methods of ITU-R Recommendation BT.500-7. The use of longer test sequences raised new issues such as how long each sequence should be, and what the voting procedure should be in relation to the behavior of the observer.

Different studies were undertaken to evaluate the *recency* and *forgiveness effects* of the observer, by inserting artifacts at different positions within sequences of varying lengths, and collecting one quality grading at the end of each presentation. The results showed that the reporting time and the human memory processes (beyond 10- to 15-second timeslots) play an extremely important role. Different tests were performed to confirm that the observers could assess the picture and service quality accurately over sequences of 30 to 60 minutes. A continuous quality evaluation mechanism was carefully considered. It was thought that this approach would solve the problem of quasi-random appearances of content-dependent artifacts, bearing in mind the recency and forgiveness effects.

The maximum frequency of vote acquisition was determined (two votes-per-second) using the results of preliminary studies on the recovery time. Continuous quality evaluation was also found to be closer to the real home environment where programme *zapping* allows an immediate sanction over quality. The continuous evaluation is performed using a sliding device where the observer moves the knob in one direction to show appreciation of the picture quality and in the other direction to indicate concern about it.

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

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

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

The three stages of SSCQE

SSCQE is foreseen as a three-stage method but only “stage 1” has so far been introduced in ITU-R Recommendation BT.500-7. *Stage 1* consists of performing the single-stimulus continuous quality evaluation, and collecting data on the instantaneous grading from the slider device used by each observer (we can see that in the Figure 7). Self-consistent processing is already possible at this stage, resulting in a cumulative distribution of quality variations with time.

Stage 1 is particularly suited to the requirements of comparison tests. The *Stage 2* option is available to extract 10-second sub-sequences from the original test material to perform complementary DSCQS or DSIS tests. An example might be those subsequences which correspond to the different percentiles of the cumulative distribution obtained at stage 1. Stage 2 can also be used to calibrate the stage 1 results, using the existing adjectival scales given in ITU-R Recommendation BT.500-7. Under *Stage 3*, further developments are currently being considered in TAPESTRIES to apply an overall weighting function (modeling the human memory processes, i.e. the recency and forgiveness effects) in order to arrive at a global *average* for the perceived quality of the sequence being tested.

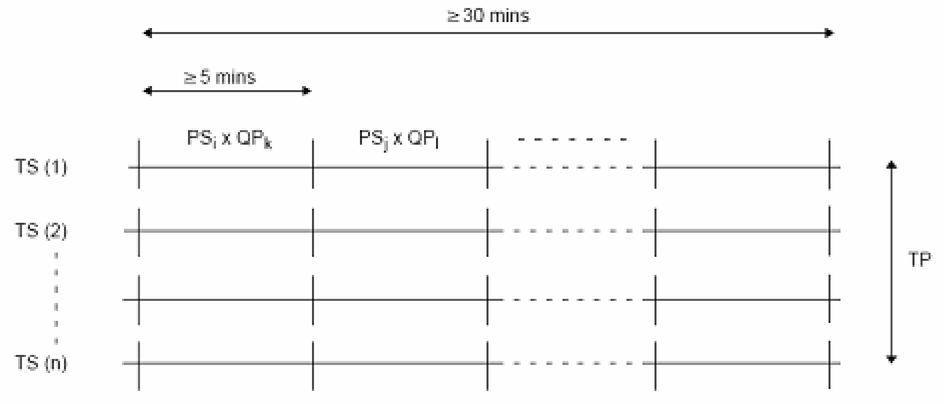


Figure 7
SSCQE – Stage 1
protocol.

4.6 Double-stimulus Continuous Quality Evaluation (DSCQE)

The introduction of digital audio-visual services needed a new subjective protocol which is able to measure the quality of service on longer viewing sequences, representative of video contents and statistical error occurrences. The SSCQE method fits this requirement as regards digital TV services.

In the case of applications like surveillance, it becomes important to assess not only the basic quality of the images but also the fidelity of the information transmitted. For that reason, it was proposed to adapt the SSCQE method to introduce simultaneous double visual stimuli while still performing continuous quality evaluation.

When performing a DSCQE test, the observers watch two displays. One shows the encoded decoded video without any transmission errors (i.e. the reference, or source material). The other shows the same video material after alteration by transmission errors. The observers assess the quality by direct comparison, evaluating the fidelity of the video information by moving the slider of a handheld voting device. An example of data obtained after averaging the votes from the different observers is given in Figure 8. An example of DSCQE results, after data-processing, is given in Figure 9.

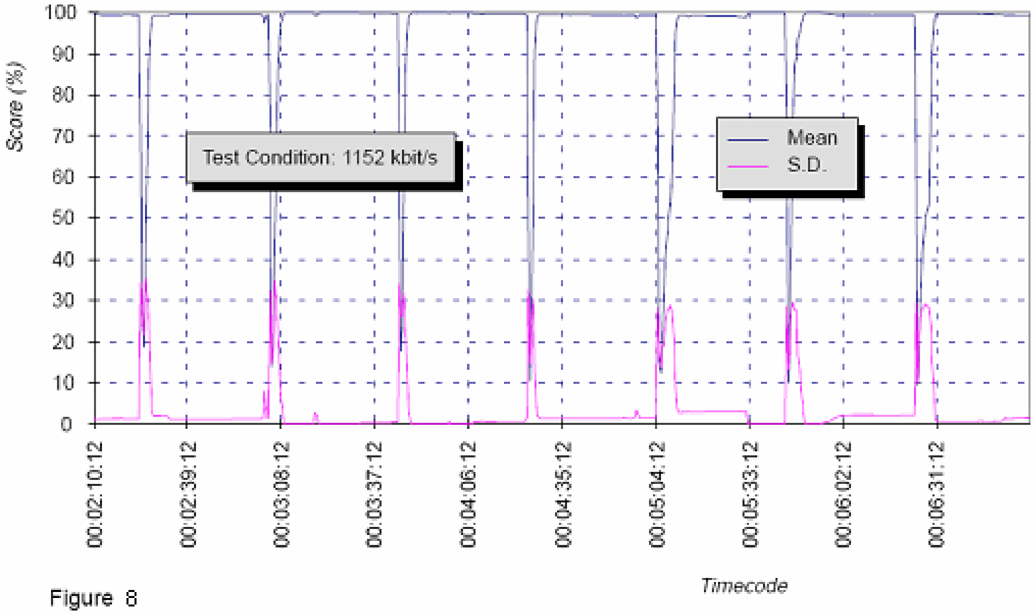


Figure 8
DSCQE – Results after averaging. MPEG-4 error-robustness tests.

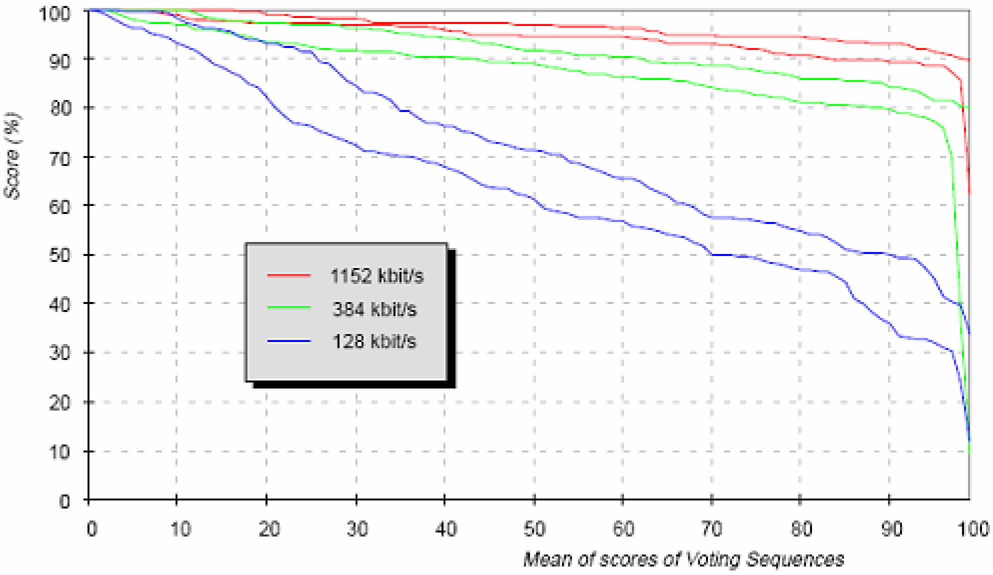


Figure 9
DSCQE – Example of results presentation after real test data-processing. MPEG-4 error-robustness tests.

5 Conclusions

The problem of the subjective is their lack of automation, they involve a group of people for conducting the tests, resulting expensive and time consuming approach. Objective methods do not depend on people, making them attractive for automatic evaluation.

The major drawback of objective intrusive methodologies is their inherent need of both signals (reference signal and impairment signal) . In we focused on video there is an extra problem, the time and resources consuming by complex methods are generally high and about signals there is an extra problem. Objective non-intrusive methods present an important advantage, they do not require any extra signal for performing the estimation, which allows them to be used in real-time scenarios but non-intrusive methods parameter-based major problem have a strong dependence on subjective tests results calibration training.

Pseudo-Subjective Quality Assessment (PSQA) is a method which combines objective and subjective methods. PSQA needs to ask for the quality of a group of sequences like the subjective methods and on the other side, that method do not depend on people, making this really attractive for automating the evaluation process like objective methods.

6 BIBLIOGRAPHY

[1] ITU-T Rec.930 “Principles of a reference impairment system for video”, International Telecommunication Union, 1996 (VIRIS).

[2] H. Cancela, P. Rodríguez-Bocca, and G. Rubino “Video Quality Assurance in Multi-Source Streaming Techniques”, 2007 (PSQA).

[3] Wonho Yang “enhanced modified bark spectral distortion: an objective speech quality measure based on audible distortion and cognition model”, 1999 (PSQM) (DMOS).

[4] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier “Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment”, 2001 (PESQ), (MOS).

[5] Zhou Wang Alan C. Bovik Eero P. Simoncelli “Structural Approaches to Image Quality Assessment”, 2005 (SSIM).

[6] ITU-T Rec.910 “Subjective video quality assessment methods for multimedia applications”, 1999 (ACR), (DCR).

[7] Th. Alpert, J.-P. Evain “Subjective quality evaluation – The SSCQE and DSCQE methodologies”, 1997 (SSCQE), (DSCQE).

[8] Renshou Dai, “Information in Sage’s P.561 INMDTest”, 2004 (IMMD)

[9] Floriano De Rango, Mauro Tropea, Peppino Fazio, Salvatore Marano “Overview on VoIP: Subjective and Objective Measurement Methods”, 2006 (CCI)