End-2-end End-2-end evaluation of IP multimedia services, a user-perceived QoS approach

Pedro Casas Hernandez, Pablo Belzarena, Sandrine Vaton

End-2-End Evaluation of IP Multimedia Services, a User Perceived Quality of Service Approach

Pedro Casas
TELECOM Bretagne
Brest, France
Universidad de la República
Montevideo, Uruguay
pedro.casas@telecom-bretagne.eu

Pablo Belzarena
Universidad de la República
Montevideo, Uruguay
belza@fing.edu.uy

Sandrine Vaton
TELECOM Bretagne
Brest, France
sandrine.vaton@telecom-bretagne.eu

***Abstract*—Providing Quality of Service (QoS) has always been an important task for Internet Service Providers. However, the proliferation of new multimedia content services has turned it a vital and challenging issue. The problem with QoS in nowadays Internet is what to measure and how to do it to provide real quality levels to end-users. Recent works in the field have focused on the service consumer, assessing the QoS as perceived by the end-user. This paper addresses the automatic evaluation of the QoS as Perceived by an end-user (PQoS) of a multimedia service. We present a general overview of the PQoS approach, studying the impact of different network and multimedia features on the quality as experienced by human beings. We develop an original software tool that integrates all the aspects related to the automation of the estimation process, using a broad set of PQoS methodologies. To the date and to the best of our knowledge, there is no free software implementation that completely estimates the PQoS for a VoIP and VideoIP service in a real environment. Using this software tool and real subjective tests, we perform an unbiased comparison of the different proposed techniques for video and audio services over IP.**

I. Introduction

In traditional telecommunications, quality of service (QoS) has always been focused on network metrics: packet loss, delay, jitter, available bandwidth, etc. Classical QoS provisioning consists in keeping particular subsets of this performance metrics within certain limits, in order to offer the user reasonable quality levels. The problem with this approach is that in today's Internet, the heterogeneous features of current services make it difficult, sometimes even impossible to clearly identify the relevant set of performance parameters for each case. Even more, the quality experienced by a user of the new multimedia services not only depends on network features but also on higher layers' characteristics [1] (multimedia coding and compression, recovery algorithms, nature of the content, etc...). In this sense, a final user might experience acceptable quality levels even in the presence of severe network degradation. These observations show that rating the quality of the new multimedia services from the network's side may no longer be effective.

The *user perceived quality of service* (PQoS) field addresses this problem, assessing the quality of a service as perceived by the end-user. The assessment of perceived quality in multimedia services can be performed either by *subjective* or *objective* methodologies. Figure 1 presents a general overview of the PQoS evaluation field.

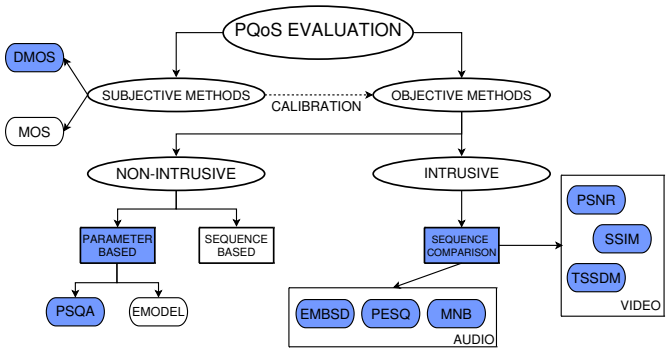


Fig. 1. PQoS Evaluation.

Subjective methods represent the most accurate metric as they present a direct relation with the user's experience. 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. Different recommendations standardize the most used subjective methods in audio [10] and video [11], [12]. The problem with subjective methodologies is their lack of automation (by definition, they involve a group of people for conducting the tests) resulting in an expensive and time consuming approach.

On the other hand, objective methods do not depend on people, making them really attractive to automate the evaluation process. The objective PQoS evaluation can be either *intrusive* or *non-intrusive*. In network's context, intrusive means the injection of extra data (audio and/or video sequences) to perform the measurement. Intrusive methods are based on the comparison of two sequences, a reference sequence (original) and a distorted sequence (i.e. the one modified during network transmission). This comparison is generally performed either in the time/space domain (simple sample comparison: mean square error (MSE), signal to noise ratio (SNR) or peak signal to noise ratio (PSNR) [1]) or in the *perception domain*, using models of the human senses to improve results. In this last category we have (for audio

assessment) the perceptual speech quality measure (PSQM) [16], the measuring normalizing blocks (MNB) [14], the enhanced modified bark spectral distortion (EMBSD) [15] and the perceptual evaluation of speech quality (PESQ) [17], [18]; in the case of video, some of the developed tools are the Structural Similarity Index Measurement (SSIM) [22], [23], [21], the Video Quality Measurement (VQM) [19] and the Time/Space Structural Distortion Measurement (TSSDM) [20]. All these tools provide a measure of the perceptually relevant degradation of the multimedia sequence ([25] presents an interesting validation report of objective models for video quality assessment). Considering their application in real-time assessment (a desirable property in todays' networks), the major problem with objective intrusive methodologies is their inherent need of both sequences, something that may result too restrictive in some network scenarios. In the case of video sequences there is an extra problem, the time and resources consumed by complex methods are generally too high.

Non-intrusive methods present an important advantage, they do not require any extra sequence to perform the estimation. This allows their use in real-time scenarios. Depending on the kind of information they use, non-intrusive methods can be classified as either sequence based or parameter based. In the case of sequence based methods, the assessment is done without any reference sequence, just applying different algorithms to the distorted sequence. These methods are also known as "null reference". In the case of parameter based methods, network features as well as characteristics of the multimedia itself are taken as input. The idea is to conceive a model which allows to map a PQoS relevant set of these parameters into a quality value (as perceived by the end-user). Examples of these features are loss rate, length of loss bursts, delay, jitter (network features), coding, nature of the content (e.g. motion level, language), bit rate, frame rate (multimedia features), etc. The ITU E-Model [8] and the pseudo subjective quality assessment (PSQA) [3], [2], [4] methods fall into this category. The E-Model is an empirical/mathematical set of formulas originally designed for telephony networks planning, and even though it is actually being used in IP networks, results have shown that it is not accurate enough for user perceived quality assessment [7]. The recently introduced PSQA approach uses a statistical learning algorithm (a Random Neuronal Network [5]) to *learn* the relation between network and multimedia features and user perceived quality. The PSQA has already shown promising results in the PQoS field ([2], [3]). The main drawback of parameter-based methods is their strong dependence on subjective tests' results for training (in fact, all different objective methods must have in some sense a calibration phase as their results are not in the same scale as subjective tests' results).

The remainder of this paper is organized as follows. In Section II we study the impact of different network and multimedia features on PQoS for VoIP and VideoIP. A detailed description of the PQoS algorithms as well as the measurement methodology of the software tool are presented in section III. The software implementation is described in Section IV, presenting the architecture and design of the developed tool. In Section V we present and analyze the experimental results, describing the test environment and comparing the performance of the selected estimation methods. Finally, section IV concludes this paper.

II. QoS in VoIP and VideoIP

The QoS experienced by the end-user of an audio and video transmission depends on many different features. We can classify them into two categories: *network features* and *multimedia features*. Network features refers to all the objective QoS metrics involved in a multimedia transmission through an IP network: losses, delays, bandwidth, etc. In an attempt to standardize the definition of QoS at the IP layer, the IP Performance Metrics IETF working group has specified many of these network features in several recommendations: one-way-delay [27], packet loss [28], round trip delay [29], loss pattern [30], delay variation (jitter) [31], network capacity [32], etc. Multimedia features includes all the higher layer's features for multimedia transmission (recovery algorithms, de-jitter buffers, etc.) and the specific components of the multimedia itself, like coding, bit-rate, frame-rate, motion level of the video sequences, etc. Delay has a major impact over interactive or real time multimedia applications, such as telephony, video conference, gaming or live transmissions. For example, the ITU-T recommendation G.114 specifies that delays from sender to receiver must be lower than 150ms. to avoid the loss of interaction between end-points in a conversation. Figure 2 presents the different components that contribute to delay from sender to receiver in an end-to-end multimedia transmission: multimedia coding/packing at the source, intermediate buffers, network transmission and decoding/unpacking at destination.

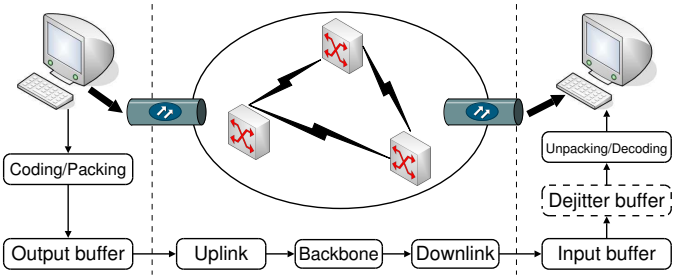


Fig. 2. End-to-end multimedia transmission.

Large delays have another undesirable effect: they reduce the throughput of transmissions over TCP. Even though TCP is not the most suitable transport protocol for real-time multimedia, it is largely used in applications such as radio streaming (Virgin Radio [35], Pandora [36] etc.) and video content delivery (YouTube [34], MSN TV [33], etc).

The quality of a multimedia transmission is also affected by the delay's variation or *jitter*. Audio and video are coded at the source at a given rate, and so packets are expected to arrive at destination at the same rate for an accurate decoding. In [24], the authors show that the effects of packet jitter over the

experienced quality are similar to those of packet loss; this is somehow expected, as packets that do not arrive on time are seen as lost information by the decoder at destination. The effects of jitter can be reduced by using de-jitter buffers at the reception (fig. 2), but this solution has the drawback of increasing the delay between end-points.

Coding is another important feature regarding PQoS. Almost every codec used in audio and video takes advantage of the correlation of the sequence to reduce the bandwidth requirements (compression). The quality obtained with different codecs depends on the compression algorithms they use and on the compression rate. Figure 3, taken from [13], presents the quality (PSNR, section III-B2) of a video sequence as a function of the codec bit-rate. For example, we can see that for the same quality level, the codec bit-rate of a MPEG2 coding is approximately the double of a H.264/AVC coding.

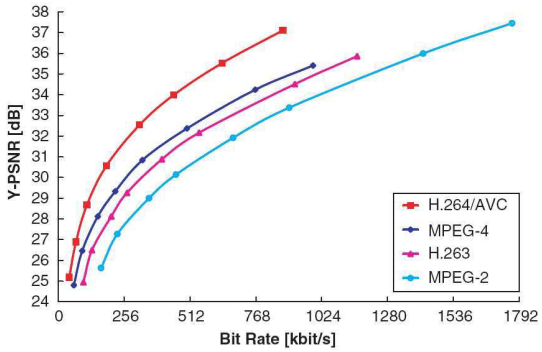


Fig. 3. Performance of different video codecs for different bit-rates.

Codec compression makes that not all transmitted packets have the same importance as regards quality at the receiver side. Indeed, if we take for example an standard MPEG coding, some packets will carry more information than others (I-frames in contrast with P,B-frames [1]) and decoding robustness will directly depend on which packets are lost. This takes us to another important feature that influences quality, the loss pattern: single isolated losses do not have the same impact as consecutive losses.

The effects of losses on the perceived quality of service are highly correlated with the multimedia coding. Figure 4 presents this idea. In 4(a), a video with MPEG1 coding is transmitted over a lossy connection. In 4(b), the same video is transmitted over the same connection, but using a MPEG4 coding. The differences that can be perceived are evident, and they can be easily explained: MPG4 coding uses more information (I-frames [1]) for those parts of the sequence with higher motion levels (in the figure, the motorbike moves faster than the background), whereas MPEG1 does not make any difference between elements, using the same rate for every part of the sequence.

The motion level of a video sequence has also a noticeable impact on PQoS. As we show in the obtained results, video sequences with higher motion levels (action sequences, sport sequences) are more sensitive to network degradation (as

(a) MPG1.

(b) MPEG4.

Fig. 4. Loss influence for different video codecs.

perceived by the end-user) than those with lower activity (like the news).

There are many other features that influence the experience of the end-user in multimedia transmissions, like the ear-to-mouth relation, silence detection, echo (VoIP/PSTN gateway), blocking and blurring, etc. However, we will limit our study to a reduce and relevant set of features: losses (loss rate and mean loss burst length), delay variation (jitter), video bit rate and motion level, and video and audio codec.

III. Quality Assessment

A. Subjective Evaluation

In this kind of test, a group of people rates the quality of several distorted sequences (audio or video). There are mainly two categories for these tests, depending on whether a reference sequence is included or not in the evaluation. When there is no reference sequence, people evaluate only the distorted sequences and grade its quality, according to a quality scale like I(a); the output of this test is known as the Mean Opinion Score (MOS). In the case of audio, this test is known as an Absolute Category Rating (ACR) test; in video, the test is referenced as a Single Stimulus (SS) test. When the reference sequence is included in the test, people compare the original sequence with the distorted sequence and then grade the perceived degradation, according to the quality scale I(b). The output of this test is known as the Degradation Mean

Opinion Score (DMOS). In audio, this test is known as a Degradation Category Rating (DCR) test; in video, the test is called Double Stimulus Impairment Scale (DSIS).

Score	Sequence Quality
5	Excellent
4	Good
3	Regular
2	Bad
1	Awful

(a) MOS Quality Scale

Score	Sequence Degradation
5	Imperceptible
4	Perceptible, not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

(b) DMOS Quality Scale

TABLE I
DIFFERENT QUALITY SCALES.

There are many other variations of subjective tests, all of them are defined in the ITU recommendations [10] (audio) and [11], [12] (video).

B. Objective Evaluation - Intrusive Methods

In an intrusive evaluation of PQoS, a multimedia sequence is transmitted through the communication system under study. The obtained distorted sequence is compared with the original sequence to measure the degradation suffered during transmission. As we stated before, two kind of comparisons can be performed: direct rough sample comparison (like SNR) are very simple to implement but they are poorly correlated with subjective tests. The comparison can also be done by considering a model of human perception to improve the results. In this case, the sequences are transformed into a perception domain and then compared, considering only the perceptually relevant distortion.

*1) **Audio Methods**:* three different methods where analyzed and implemented in the software tool: the Enhanced Modified Bark Spectral Distortion (EMBSD), the Perceptual Evaluation of Speech Quality (PESQ-ITU P.862), and Measuring Normalizing Blocks (MNB). These algorithms perform the comparison in the perception domain. Three psychoacoustic concepts are considered in all of them: the *Critical Bands*, *Loudness* and *Masking*. The *Critical Bands* are based on the ability of a human to distinguish between different tones. In low frequencies, a few hertz are enough to recognize two different tones, whereas in high frequencies this threshold increases to hundreds of hertz. The auditory system is modeled as a filter bank of band-pass filters. The *Loudness* considers the *perceived intensity* of a sound. For example, a sinusoidal signal of 40 dB at 50 Hz is equally perceived (in terms of strength) as a sinusoidal signal of 0 dB at 1 KHz. The perception of loudness is related to both the intensity and duration of a sound (the auditory system integrates intensity over a certain time window). The *Masking* concept represents the psychoacoustic effect that occurs when the presence of a sound does not allow the perception of another. A typical example of masking can be found in the city, when two people can not hear each other because of traffic noise. Briefly, the auditory threshold is modified by the presence of a sound.

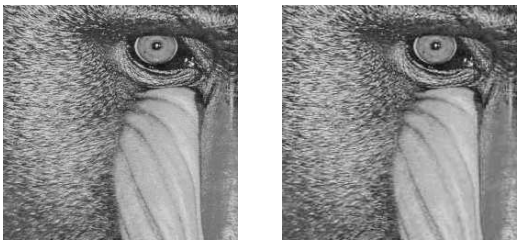
*2) **Video Methods**:* the considered algorithms for PQoS evaluation in video algorithms differ in what their consider as relevant to the human perception.

a) Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR): the MSE and PSNR algorithms are the simplest methods to compare two sequences. They do not take into account any perceptual feature, they just provide a raw pixel comparison between frames of a video sequence. The MSE and PSNR are defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \quad (1)$$

$$PSNR = 10Log_{10}(\frac{L^2}{MSE}) \quad (2)$$

where n is the number of pixels in the image or video, x_i and y_i are the i-th pixel of the original and distorted image respectively, and L is the range of possible values for the pixels (i.e. the pixel's dynamic range). These quality assessment methods have been the most used because of their mathematical simplicity. However, they have been criticized due to their poor correlation with subjective methods. Figure 5 makes clear this drawback. In both figure 5(a) and figure 5(b), the original image (on the left) is compared against the distorted image (on the right). Both groups of pictures have almost the same PSNR value, but the differences in the first group (5(a)) are almost unnoticeable, whereas in the second group (5(b)) the are really evident.



(a) Mandrill, PSNR = 159

(b) Pepper, PSNR = 160

Fig. 5. PSNR as a measure of perceived difference between images.

b) Time/Space Structural Distortion Measurement - TSSDM: the target of this algorithm is to measure changes in the spacial activity, considering certain spatial-temporal (ST) regions of the original and distorted videos for the comparison. The basic metric for the comparison is the gradient module of each ST region (it represents a measure of the spatial activity). The main advantage of this technique is that it can be used with reduced reference information, as the comparison is only performed in the selected ST regions.

c) Structural Similarity Index Measurement-SSIM: a new philosophy in the design of quality metrics was introduced in [22], [23], [21]: "the main function of the human visual system is to extract structural information of the viewing field, and the human visual system is highly adapted for this purpose". These works propose that the measurement of the structural distortion is a good approximation to the perceived distortion. According to [21], structural information is the feature that represents the structure of the objects, independently of the luminance level and contrast of the image.

C. Objective Evaluation - Non-intrusive Methods

a) E-Model: the relation between different network/multimedia features and the speech quality has been quantified by the E-Model [7], introduced by ITU-T as a planning tool for telephony services. However, this tool presents a serious drawback: it assumes that individual quality features such as loudness, delay, talker echo and speech distortion have mutually independent effects on the perceived quality, which is not the case.

b) PSQA: as mentioned before, the PSQA method is based on the Random Neuronal Network (RNN) model. The results of subjective tests (DMOS) depends basically on the network features (losses, delay, jitter) and the multimedia features (codec, bit rate, nature of content). If it is possible to model the relation between these parameters and the subjective DMOS, we can approximate the DMOS by measuring these objective features. The RNN are a supervised learning machine, that uses a set of couples *network/multimedia features-DMOS* in a learning stage to build an approximation to this model. After this stage, the knowledge of the state of the network and the features of the multimedia are enough to predict the DMOS. The learning of the RNN consists in minimizing a cost function that penalizes the difference between predicted values and real DMOS subjective tests' results.

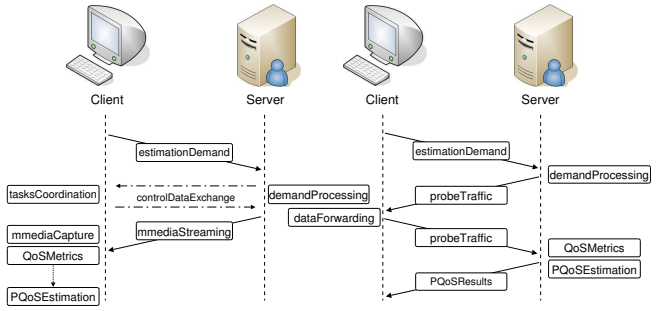
D. Objective IP level QoS metrics estimation

The estimation of the objective QoS metrics at the IP layer is conducted between the end-points of the connection. The considered network features are packet loss (loss rate and mean loss burst length), packet delay and packet jitter. Delay's estimation can be conducted either in a single way (one-way-delay, owd from now on) or for the round trip (RTT), depending on weather the end-point devices are synchronized or not; in the general case, the time synchronization provided by the NTP (Network Time Protocol) protocol is not accurate enough to provide a good estimation of the owd, but GPS time synchronization is becoming usual as prices tend to drop, so time synchronization can be ensured in many different network scenarios.

The estimation can be achieved in different ways, depending on the kind of PQoS evaluation to be performed. In the case of a non-intrusive evaluation, the estimation can be conducted by active measurements, using probing traffic of similar characteristics to the service under evaluation (basically mean traffic size and packets' inter-departure time). This is for a simple reason: network QoS features are not in fact an own characteristic of the network but of the user's traffic as well (e.g. delay will not be the same for a radio transmission of BW_{radio} bit-rate and a high quality video streaming service of BW_{video} bit-rate if connection's available bandwidth is between BW_{radio} and BW_{video}). Said in other words, user's traffic itself directly influences quality, so it must be taken into account for the evaluation. In the case of an intrusive evaluation, the PQoS analysis is performed by sequences' comparison after the streaming of the multimedia and there is no need to estimate the network features. However, we implement a simple methodology [6] that takes advantage of this multimedia transmission to estimate the network features, using the information provided by the RTP and RTCP protocols.

E. Measurement Methodology

The developed software tool integrates both intrusive and non-intrusive objective estimation methods. The aim of this tool is not only to perform an automatic PQoS estimation but also to compare the performance of the different approaches and algorithms. The implemented algorithms were PESQ, EMBSD, MNB, and PSQA in the case of audio, and MSE, PSNR, SSIM, TSSDM and PSQA for video.



(a) Intrusive PQoS estimation. (b) Non-intrusive PQoS estimation.

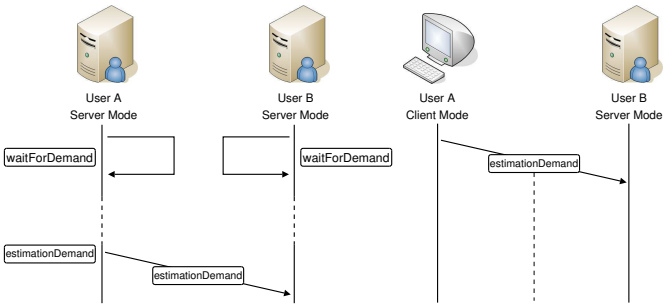
Fig. 6. Measurement methodology.

The PQoS evaluation is performed between the end-points involved in the service under evaluation. Figure 6 presents a brief summary of the measurement methodology. The client begins the measurement by sending an estimation demand to the server. Depending on the type of algorithm selected by the client, the server will either transmit a reference sequence of similar characteristics to the actual service (intrusive methods), or begin a connection's features estimation using active measurements (non-intrusive methods). If the selected algorithm is intrusive, the client stores the sequence transmitted by the server and performs the PQoS estimation by comparing the reference and the transmitted sequence (both the client and the server have the same reference sequences). The software tool allows to specify a features' based estimation at the same time, in which case the RTP and RTCP headers of the multimedia transmission are analyzed to gather the network features. In the case of non-intrusive methods, the server uses the estimated

network features (loss rate, jitter, mean loss burst length) and the corresponding multimedia features (coding, bit-rate, frame rate and motion level) as input to perform the estimation. Tasks' synchronization between end-points is achieved by a specially developed communication protocol.

IV. Software Implementation

The PQoS estimation software tool was designed to be used in both end-points of the service at the same time. The architecture foresees a symmetrical operation, in which both end points can play either the client or the server role (considering the classical client/server paradigm, where the client asks for some service and the server responds to his demands).



(a) Before a PQoS estimation demand. (b) After the demand.

Fig. 7. Symmetrical architecture.

Figure 7 explains this concept of symmetry. In the very beginning, both end-points act as servers, waiting for a PQoS evaluation demand from the opposite side. When one of both machines decides to perform an estimation, the scheme changes to a traditional client/server architecture as the previously discussed (figure 6). The main advantage of this symmetrical architecture is the ability that both end-points acquire to process and generate information, saving transmission and operation time.

A. Software Design

During the software design phase, special attention was directed to the modularity of the tool. The key idea was to conceive a reusable and easily to improve/modify design. The final implementation consist of five independent software modules (each of them can be used isolated from the rest, in any other application). Figure 8 presents a general overview of these modules. The ***System Manager*** is the software's brain. It manages the connection establishment and data exchange between end-points as well as the interaction between the rest of the different modules. It is basically composed of 3 sub-modules: a *client*, a *server* and the *manager* itself. The ***PQoS Algorithms*** module is the most important module of the system, as it implements the different estimation algorithms so far discussed. The ***Multimedia*** module supplies the audio and video sequences for intrusive PQoS estimation. It consists of an audio streaming platform (implemented with the Java Media Framework toolbox, [37]), a video streaming platform (implemented with the Video Lan Client project, [38]) and a reference sequences' database. The ***QoS Metrics Estimation*** module is responsible for the network features estimation. Finally, the ***GUI*** module implements the graphical user interface to easily interact with the tool.

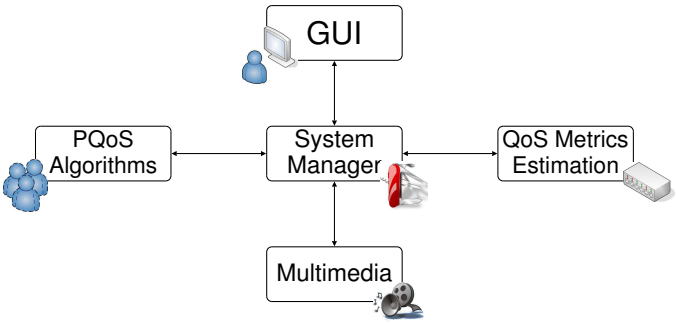


Fig. 8. Software components.

Figure 9 presents a high level diagram of the software's architecture. Given the different restrictions and characteristics of each module (flexibility, portability, time efficiency and accuracy, etc.), different programming languages were used in the implementation. Higher layer implementations where mostly developed in Java (J2RE), while lower layer programming (C and C++) was used in all critical-time applications (e.g. PQoS intrusive algorithms, multimedia coding, time reference, etc.). The interaction between languages was achieved by using the Java Native Interface (JNI) library, a versatile set of Java classes and methods which aloud the communication between native (C/C++) and portable software (J2RE).

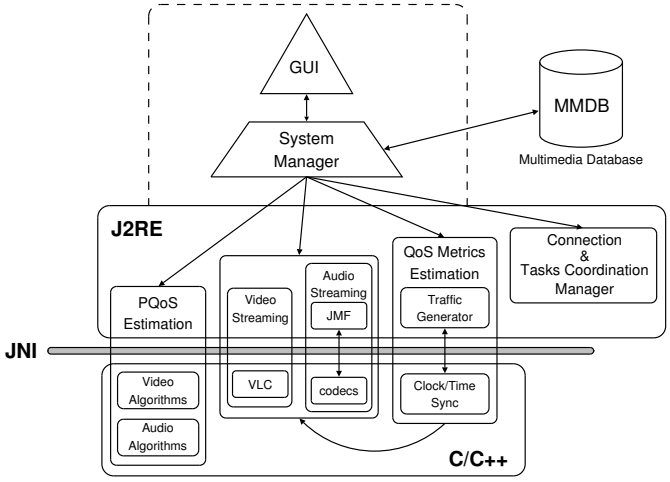


Fig. 9. Software architecture.

V. Experimental Evaluation and Results

A. The Test Bed

In order to perform the subjective tests, calibrate the objective methods and evaluate the performance of the different approaches we developed a simple test bed which allows to emulate network conditions in a controlled fashion [1]. This

test bed is composed of two end point machines (server/client) connected through an intermediate router that simulates losses, delay and jitter. Figure 10 presents this testbed configuration.

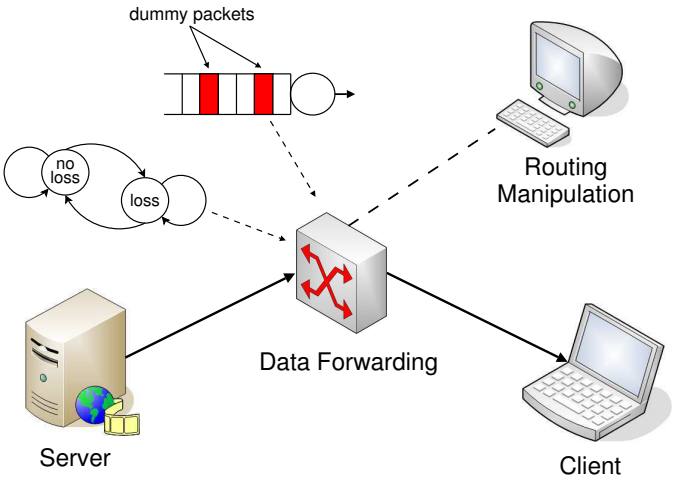


Fig. 10. Evaluation testbed.

Packet losses in an IP network are rarely independent and they generally occur in bursts, due to network congestion. The simplest model to represent this behavior was proposed in [26], using a simple Markov model: the simplified Gilbert loss model. The Gilbert loss model consists in a two states Markov chain, where the state 0 corresponds to a received packet at destination and the state 1 to a lost packet. In figure 11, p represents the probability of loosing a packet given the last packet arrived correctly, and q is the probability of a correct transmission given the last packet was lost. This simple model allows to simulate losses in bursts, as the fate of a packet depends on the result of the last transmission. Given a connection loss rate, we can modify p and q in order to obtain different loss patterns. Jitter and delay are controlled by direct manipulation of buffers and output capacity of the ethernet routers' interface. Dummy packets are inserted in the output buffer to generate jitter, and the buffer size and output capacity are varied to produce forwarding delays.

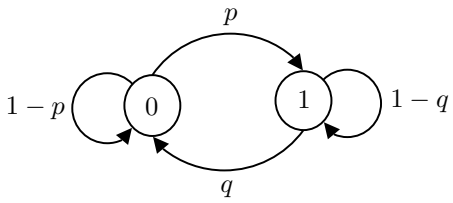


Fig. 11. Gilbert loss model.

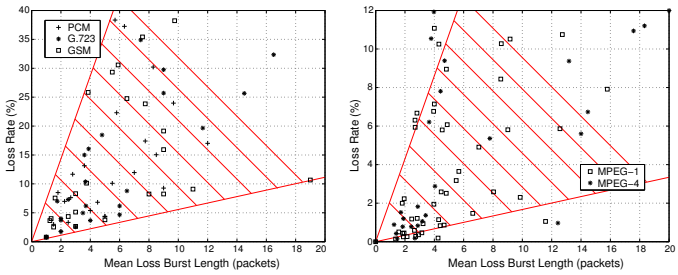
The multimedia sequences' sets consist of 75 original-distorted couples for video and 72 couples for audio. The reference video sequences where chosen according to the reference [11], [12] (40 short sequences of 10-30 seconds) and classified by coding (MPEG1 and MPEG4) and motion level (low, medium and high); this last classification was subjectively conducted, even though it would be interesting to use the codec's motion estimation to have an objective classification. In audio, 24 short sequences where recorded and coded with three different codecs (PCM, GSM and G.723). Each reference sequence was transmitted through the test bed, setting different values for the router's parameters in order to cover the most suitable network features in an Internet like scenario. The generated sequences were then used in the subjective tests (as described in section II), obtaining a final data set of the form:

$$\{sc_j, (p_0, f_1, .., f_i, .., f_n), \text{DMOS}\}, \quad (3)$$

where sc_j is the j-th original-distorted sequence couple, f_i is the value of the i-th feature (e.g. loss rate, mean loss burst length, jitter, codec, motion level, etc..) and DMOS is the corresponding subjective test result. Finally, part of the data set was used to train the PSQA learning algorithm and calibrate the objective intrusive methods, using the remaining data for validation.

B. Subjective Tests' Results

Figure 12 shows the distribution of the data-set samples (both audio and video) considering loss rate, mean loss burst length and codec.



(a) Audio samples (b) Video samples

Fig. 12. Distribution of the data-set.

In order to obtain good learning and calibration results, data samples must extensively cover the *inputs' space*, particularly in those values that are more usual or where the quality discrimination is more difficult. In the case of audio transmissions we consider a broader variation for loss rate than in the video transmission, given the fact that even under sever loss conditions an audio transmission can be perceived as acceptable by the end-user (we confirm this observations in the obtained results, see section V-C3).

	Mean DMOS	Mean Variance
Audio	3.04	0.36
Video	3.03	0.25

TABLE II
STATISTICAL CHARACTERISTICS OF THE SUBJECTIVE TESTS.

Table II presents the subjective DMOS tests' results for both audio and video, using the quality scale I(b). According to

the ITU recommendations for audio [10] and video [11], the subjective tests must be designed so that the average result is in the middle of the quality scale (in order to avoid biased results). These recommendations also specified the procedure to remove outliers from the results.

C. Evaluation of the Different Techniques

In order to compare the performance of the different algorithms we use a traditional error estimator, the mean absolute error (MAE), between estimated values (algorithms) and real values (subjective tests). Intrusive methods' results are not in the same scale as DMOS values (they are correlated with human assessment but each algorithm uses its own scale) so a calibration phase is conducted before the comparison. As regards non-intrusive algorithms (we will only consider PSQA in the evaluation, the E-Model has already shown quite poor performance [1], [7]), the system must be trained before using it. In both cases we split the previous data set in a *training data set* and a *validation data set*. With the first set we calibrate/train the intrusive/non-intrusive methods, with the second we do the validation. In the case of video, we consider 70% of samples for training and 30% for validation. In audio, the relation is $80\% - 20\%$ (we consider a bigger training set to overcome some weaknesses of the audio data set, see [1] for discussion).

1) Audio analysis: in table III(a) we present the actual MAE values for all the audio algorithms in the training set, according to the quality DMOS scale I(b). A graphical comparison of the algorithms' performance in the training date set is provided in figure 13.

Method	MAE
EMBSD	0.59
PESQ	**0.43**
PSQA	0.45
MNB	0.68

(a) Training data set

Method	CF
PESQ	**0.93**
PSQA	0.86

(b) Validation data set

TABLE III
MEAN ABSOLUTE ERROR (MAE) AND CORRELATION FACTOR (CF).

The results obtained in audio quality assessment presents the PESQ intrusive method as the most accurate. Compared with the other intrusive methods, PESQ has a major advantage: it includes a temporal re-synchronization algorithm that allows a correct sequence comparison. In the presence of data losses, a direct sequence comparison may result in very poor performance (worst results are obtained as losses occur closer to the beginning, see [1]). It is important to recall that PESQ is the actual ITU recommendation for voice perceived quality assessment [17]. The performance of the non-intrusive PSQA algorithm in the training set is very close to the obtained with PESQ, something that results quite interesting. Indeed, this result shows that the complex psychoacoustic model proposed by the different *perception domain* algorithms (PESQ, MNB, EMBSD) can be approximated with a Random Neuronal Network, something a priori not easy given the number of features that affects the perceived quality.

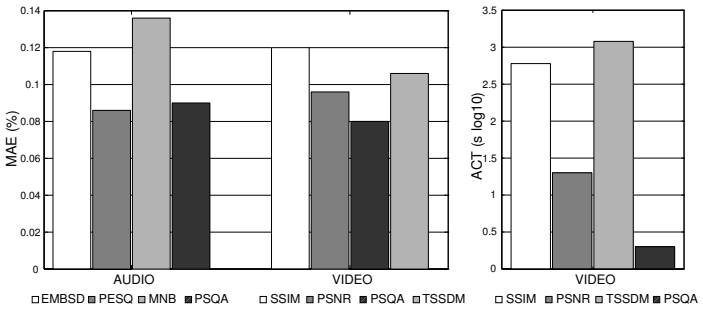
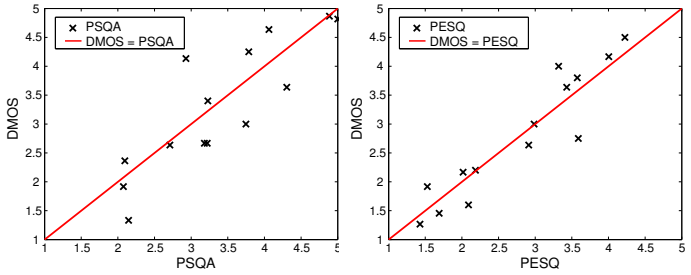


Fig. 13. PQoS in audio and video, performance evaluation of the different algorithms in the training set.

Figure 14 shows the results obtained with PSQA (left) and PESQ (right) with the validation data set. Both approaches present a strong correlation with subjective tests' results. These results confirm that the training of the RNN model was accurate enough to reproduce the good performance with an unknown data-set. Table III(b) presents the Correlation Factor (CF) between real and estimated DMOS for both PESQ and PSQA in the validation set (a value close to 1 indicates high linear correlation).

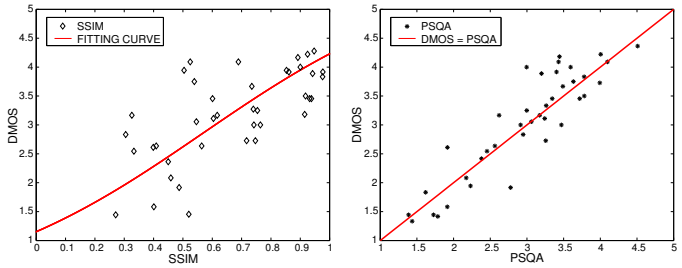


(a) DMOS vs PSQA (b) DMOS vs PESQ

Fig. 14. PQoS in audio, performance evaluation of PESQ and PSQA in the validation set.

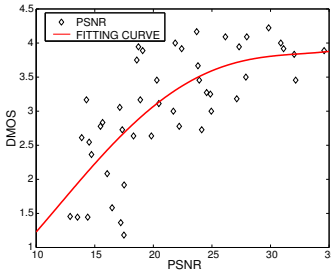
2) Video analysis: in video analysis, PSQA is clearly the best method, and not only because of the smallest error value, but mainly because of the time involved in the estimation. Table IV summarizes these observations, presenting the error values for the training set and the mean time involved in the computation of the PQoS estimation (a graphical comparison of these values is provided in figure 13). Figure 15 shows the different algorithms along with their respective fit curves (in the case of PSQA a straight line Subjective DMOS = PSQA is plotted to see the quality of the results). Figure 15(c) confirms our previous observation with respect to the PSNR misadjustment for PQoS evaluation. Indeed, the same value of PSNR corresponds to many different quality perceptions.

In the case of video there are no standardized methods for perceived quality assessment, something that shows that PQoS for video is still a very difficult problem. The intrusive methods presented in this work suffer from the same synchronization

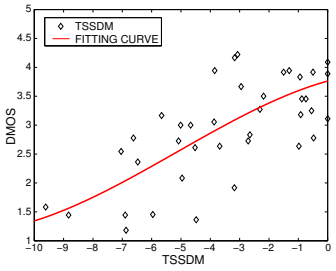


(a) DMOS vs SSIM

(b) DMOS vs PSQA



(c) DMOS vs PSNR



(d) DMOS vs TSSDM

Fig. 15. PQoS in video, performance evaluation of the different algorithms in the training set.

problem as in the audio case. However, the performance obtained by PSQA shows that a priori the problem can be solved.

Method	MAE	ACT (seconds)
SSIM	0.60	> 600
PSNR	0.48	≈ 20
PSQA	**0.40**	$\approx$ **2**
TSSDM	0.53	> 1200

TABLE IV
MAE AND AVERAGE COMPUTING TIME (ACT).

To conclude with video analysis, we show in figure 16 the results obtained by PSQA in the validation data set. As in the audio case, the RNN captures somehow the complex relation between perceived quality and network/multimedia features.

3) Analysis of the influence of through PSQA: an interesting advantage of objective parameter based algorithms is the possibility to analyze the influence of different features on PQoS. Figure 18 presents the influence of voice codec (a) and video motion level (b) on perceived quality as a function of loss rate. As expected in audio, losses in the case of *G.711* coding (pure PCM, higher bit rate, no predictive model seriously affected by losses) are less annoying. In the case of video, the evaluation confirms our initial observation about the influence of motion level on PQoS: video sequences with higher motion levels present a faster decrease of perceived quality with respect to packet losses than those with lower activity levels.

Finally, figure 17 evidences the influence of the loss rate and the mean loss burst length on (a) voice and (b) video

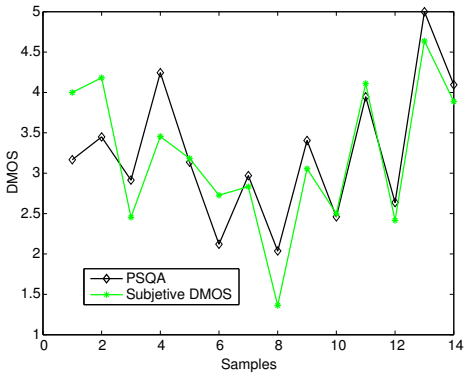
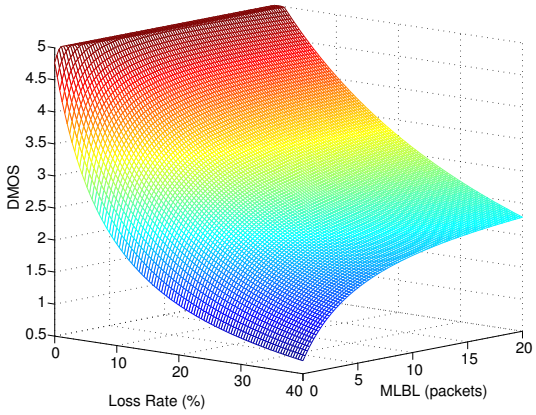
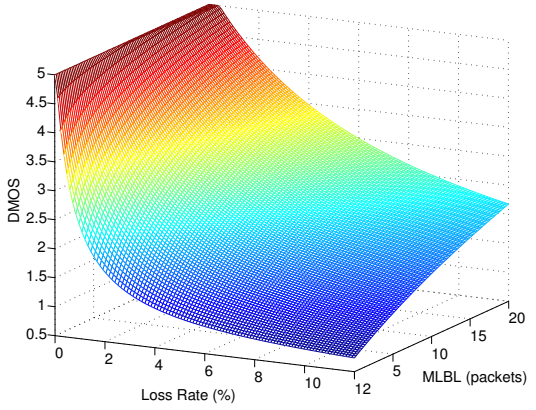


Fig. 16. Subjective DMOS and PSQA - validation data set.



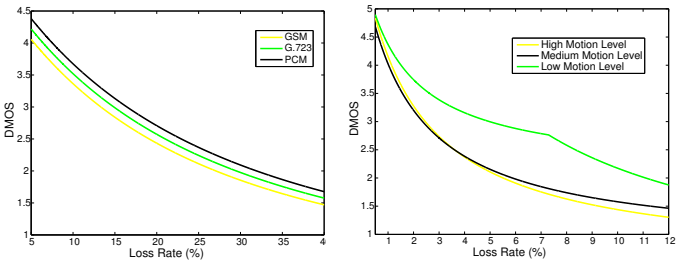
(a) Audio Evaluation (codec PCM)



(b) Video Evaluation (codec MPEG4)

Fig. 17. DMOS vs loss rate and mean loss burst length.

perceived quality. The first interesting observation is that audio perceived quality is less sensitive than video perceived quality to lost information. A possible explanation to this phenomenon is that our visual system is more developed than the auditory system, which makes that our response to visual impairments is naturally more touchy. The second element that may draw the reader's attention is that in both cases, the perceived quality



(a) Different Audio Codecs (b) Different motion levels

Fig. 18. DMOS vs loss rate (MLBL = 5 packets).

monotonically increases with the mean loss burst length, meaning that apparently we prefer concentrated losses to those that are spread over the sequences.

VI. Conclusion

In this paper we addressed the Quality of Service evaluation problem from the end-user perspective. Different methodologies were introduced for quality assessment in multimedia services in IP networks. We developed and described an original software tool for automatic PQoS evaluation. The main advantages of this tool are the combination of a broad set of the different methodologies that have been proposed to the date, the integration of all the aspects related to the automation of the estimation process and its modular design. We use the software tool to compare the performance of the most relevant perceived quality evaluation methods in the literature and we present experimental results in a real simulation test bed. The PQoS evaluation tool is completely free and it is available from the authors (please contact us).

Acknowledgments

This work was partially supported by the *"Programa de Desarrollo Tecnológico"* (PDT), grant S/C/OP/46/03. The authors would like to thank Diego Guerra and Ignacio Irigaray for their participation in the development and evaluation, as well as Martín Varela for fruitful discussion.

References

[1] P. Casas, D. Guerra and I. Irigaray, *"Perceived Quality of Service in VoIP and Video IP services"*, Technical Report, http://iie.fing.edu.uy/investigacion/grupos/artes/pqos/pqos.pdf, Universidad de la República, 2005.

[2] S. Mohamed and G. Rubino, *"A Study of Real-Time Packet Video Quality Using Random Neuronal Networks"*, IEEE Trans. on Circuits and Systems for Video Technology, 12, 1071-1083, 2002.

[3] S. Mohamed, G. Rubino and and M Varela, *"Performance evaluation of real-time speech through a packet network: a Random Neural Networks-based approach"*, Performance Evaluation, 57, 141-162, 2004.

[4] G. Rubino, M. Varela and J.M. Bonnin *"Controlling Multimedia QoS in the Future Home Network Using the PSQA Metric"*, The Computer Journal, 2006.

[5] E. Gelenbe, *"Random Neural Networks with Negative and Positive Signals and Product Form Solution"*, Neural Computation, 1, 502-511, 1989.

[6] R. Garroppo, S. Giordano, F. Oppedisano and G. Procissi, *"A Receiver Side Approach For Real Time Monitoring of IP Performance Metrics"*, Proc. of the EuroFGI Workshop on IP QoS and Traffic Control, 169-176, 2007.

[7] T.A. Hall, *"Objective Speech Quality Measures for Internet Telephony"*, Voice over IP (VoIP) Technology, Proc. of SPIE, 128-136, 2001.

[8] International Telecommunication Union, *"The E-Model, A Computational Model for Use in Transmission Planning"*, rec. ITU-T G.107, 2005.

[9] International Telecommunication Union, *"Transmissions impairments due to speech processing"*, rec. ITU-T G.113, 2002.

[10] International Telecommunication Union, *"Methods for subjective determination of transmission quality"*, rec. ITU-P P.800, 1996.

[11] International Telecommunication Union, *"Methodology for the subjective assessment of the quality of television pictures"*, rec. ITU-R BT.500-11, 2002.

[12] International Telecommunication Union, *"Subjective video quality assessment methods for multimedia applications"*, rec. ITU-T P.910, 1999.

[13] J. Ostermann et al, *"Video coding with H.264/AVC: Tools, Performance, and Complexity"*, IEEE Circuits and Systems Magazine, 2004.

[14] S. Voran, *"Objective Estimation of Perceived Speech Quality – Part I:Development of the Measuring Normalizing Block Technique"*, IEEE Trans. on Speech and Audio Processing, 7 (4), 1999.

[15] W. Yang, *"Enhanced Modified Bark Spectral Distortion (EMBSD): An Objective Speech Quality Measure Based on Audible Distortion and Cognition Model"*, Dissertation, Temple University, Philadelphia, USA, 1999.

[16] International Telecommunication Union, *"Objective quality measurement of telephone-band (300-3400 Hz) speech codecs"*, rec. ITU-T P.861, 1998.

[17] International Telecommunication Union, *"Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs"*, rec. ITU-T P.862, 2001.

[18] J. Beerends , A. Hekstra , A. Rix and M. Hollier *"Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II — Psychoacoustic model"*, 1998.

[19] S. Voran, *"The Development Of Objective Video Quality Measures That Emulate Human Perception"*, IEEE GLOBECOM, 1776-1781, 1991.

[20] *"Spatial-Temporal Distortion Metrics for In-Service Quality Monitoring of Any Digital Video System"*, SPIE International Symposium on Voice, Video, and Data Communications, 1999.

[21] Z. Wang, L. Lu and A. C. Bovik, *"Video Quality Assessment on Structural Distortion Measurement"*, Signal Processing: Image Communication, 19 (2), 121-132, 2004.

[22] Z. Wang, *"Rate scalable foveated image and video communications"*, PhD thesis, Dept. of ECE, The University of Texas at Austin, 2001.

[23] Z. Wang, L. Lu and A. C. Bovik, *"Why is image quality assessment so difficult?"*, Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing, 2002.

[24] M. Claypool and J. Tanner, *"The Effects of Jitter on the Perceptual Quality of Video"*, Proc. ACM Multimedia, (2), 115-118, 1999.

[25] VQEG, *"Final report from the video quality experts group on the validation of objective models of video quality assessment"*, 2000, **http://www.vqeg.org/.**

[26] J.C. Bolot, *"End-to-end frame delay and loss behaviour in the Internet"*, Proc. ACM SIGCOMM, 289-298, 1993. 2000.

[27] G. Almes, S. Kalidindi and M. Zekauskas, *"A One-way Delay Metric for IPPM"*, RFC 2679, 1999.

[28] G. Almes, S. Kalidindi and M. Zekauskas, *"A One-way Packet Loss Metric for IPPM"*, RFC 2680, 1999.

[29] G. Almes, S. Kalidindi and M. Zekauskas, *"A Round-trip Delay Metric for IPPM"*, RFC 2681, 1999.

[30] R. Koodlim and R. Ravikanth, *"One-way Loss Pattern Sample Metrics"*, RFC 3357, 2002.

[31] C. Demichelis and P. Chimento, *"IP Packet Delay Variation Metric for IPPM"*, RFC 3393, 2002.

[32] P. Chimento and J. Ishac, *"Defining Network Capacity"*, RFC 5136, 2008.

[33] MSN TV website, **http://msntv.com**

[34] YouTube website, **http://www.youtube.com/**

[35] Virgin Radio Online website, **http://www.virginradio.co.uk**

[36] Pandora Radio website, **http://www.pandora.com**

[37] Multimedia in Java Applications, the Java Media Framework website, **http://java.sun.com/products/java-media/jmf**

[38] VideoLAN project website, **http://www.videolan.org**