A User Perspective for End-To-End Quality of Service Evaluation in Multimedia Networks

Pedro Casas
Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
pcasas@fing.edu.uy

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

Ignacio Irigaray
Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
irigaray@fing.edu.uy

Diego Guerra
Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
dguerra@fing.edu.uy

ABSTRACT

Despite a large literature in Quality of Service (QoS) evaluation, end-user QoS provisioning remains an open research field. There is no general agreement about what to measure and how to do it in order to ensure real quality levels. Even more, new heterogeneous multimedia applications have redefined the problem, turning many previous implementations no longer appropriate for current scenario.

This paper addresses the problem of QoS assessment of a multimedia service over IP as perceived by humans, applying statistical learning techniques. We describe two end-to-end performance evaluation methodologies, the former based on Perceived QoS (PQoS) and the latter based on functional nonparametric regression. By merging them we build an improved system for end-to-end PQoS evaluation which allows analysing and better understanding the tradeoffs between different proposed techniques in the field.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: [measurement techniques, modelling techniques, performance attributes] ; C.2.3 [**Communication Networks**]: Network Operations—*network monitoring* ; I.2.6 [**Artificial Intelligence**]: Learning—*connectionism and neural nets, parameter learning* ; G.3 [**Probability and Statistics**]: [correlation and regression analysis]

General Terms

Performance, Measurement, Algorithms, Human Factors.

LANC '07, 10-11 October 2007, San José, Costa Rica

Keywords

Network performance evaluation, perceived quality of service (PQoS), statistical learning, multimedia services.

1. INTRODUCTION

QoS evaluation has always been focused on network parameters: latency, packet loss, packet jitter, available bandwidth, etc. Traditional methods apply several measurement techniques to estimate different combinations of these parameters, assuming a direct connection between them and quality levels. However, two major issues generally underestimated arise when considering end-to-end multimedia services evaluation. Firstly, almost all estimation techniques are highly *invasive*, as they rely on active measurements which distort the network during the analysis. Secondly, the quality experienced by a user of new multimedia services not only depends on network parameters but also on higher layer characteristics [2] (multimedia coding and compression, recovery algorithms, content nature, etc.), making it difficult to clearly identify the relevant set of performance parameters for each case. We propose two end-to-end performance evaluation techniques to tackle both problems. The former uses light probe traffic for the estimation, applying an adaptive learning algorithm based on functional regression. The latter considers the *user perceived quality of service* (PQoS) perspective, assessing the quality of a service as perceived by end-users. Both techniques are combined to achieve a novel and integral non-intrusive system for end-to-end user PQoS evaluation.

1.1 QoS evaluation based on functional regression

In [3], we develop a non-intrusive technique for QoS evaluation, using active end-to-end measurements based on light probe traffic. This technique was implemented in **MetroNet**, an end-to-end performance evaluation system for multimedia networks [4]. We consider a single, bidirectional network path between the user and the applications' server. Different multimedia applications present different stochastic characteristics, depending on their *content* (audio, video, high-medium-low bit rate, motion level, coding, etc.). We classify them into different multimedia categories, according to

their content (for each category i we assume a representative sequence M_t^i). We focus the study on the bottleneck link l_{bn}, assuming fixed capacity C and buffer size B. The performance index Y (latency, packet loss, packet jitter, etc.) depends on the stochastic characteristics of the user's multimedia we want to evaluate (M_t^i), the stochastic process of the cross-traffic that shares l_{bn}'s buffer with M_t^i (T_t), C and B: $Y = F(T_t, M_t^i, C, B)$. Considering that C and B remain constant during the evaluation, and as M_t^i is a known process, we can consider that Y depends on T_t through another function Φ: $Y = \Phi(T_t) + \epsilon$ (ϵ is a random, centred and independent process which represents the model error). This relation presents two difficult problems: Φ is unknown, and the cross-traffic process T_t is dependent and non-stationary, thus difficult to estimate. However, we propose a simple, two steps methodology to overcome both problems. The

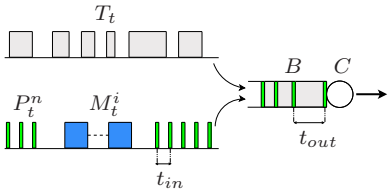

Figure 1: QoS evaluation process.

first step consists in *learning* the function Φ (figure 1). We send a train of probe packets P_t^n followed by a representative sequence of the evaluated application. P_t^n consists of n small, constant-length packets with inter-departure time t_{in}. We measure the inter-arrival time of these packets at destination (t_{out}) and we obtain a time-series $\Psi_{t_{out}}$ strongly correlated with cross traffic T_t. We also measure the performance index Y directly over the transmitted sequence M_t^i. This procedure is repeated several times during the *learning step*, obtaining a sequence of pairs (X_j, Y_j), where X_j represents the empirical distribution function (*edf*) computed from time-series $\Psi_{t_{out}}^j$ at iteration step j. Function Φ is finally estimated from sequence (X_j, Y_j), applying functional regression techniques (we consider a generalization of the Nadaraya-Watson estimator).

The second step consists in applying the estimated *quality function* $\hat{\Phi}$ for QoS evaluation. By only sending the probe packets P_t^n and computing the *edf* of $\Psi_{t_{out}}$, we obtain an estimation $\hat{Y}$ of the performance index. The major advantage of this technique is therefore the use of light probe traffic for the evaluation process.

1.2 Perceived QoS assessment

The assessment of perceived quality in multimedia services can be achieved by either *subjective* or *objective* methodologies. Figure 2 presents a general overview of PQoS evaluation. Subjective methods present a direct connection with user's experience. They consist in the evaluation of the average opinion that a group of people gives on different audio and video sequences in controlled tests. Different recommendations standardize the most used subjective methods in audio and video. Among them, the MOS (Mean Opinion Score) and DMOS (Degradation MOS) are by far the most applied. The problem with subjective methodologies is their high cost of implementation.

Objective methods do not depend on people, making them attractive for automatic evaluation. Objective PQoS measurements can be either *intrusive* or *non-intrusive*. Intrusive methods are based on the comparison of two sequences or ***signals***, the reference (original) and the distorted (e.g. during transmission). This comparison is performed either in the time/space domain (Mean Square Error (MSE), Signal to Noise Ratio (SNR) or Peak SNR (PSNR)) or in the *perceptual domain*, applying models of the human senses for performance improvement. In this 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); in the case of video, we have the Structural Similarity Index Measurement (SSIM) and the Time/Space Structural Distortion Measurement (TSSDM) as examples. The major drawback of objective intrusive methodologies is their inherent need of both signals. In the case of video there is an extra problem, the time and resources consumed by complex methods are generally high.

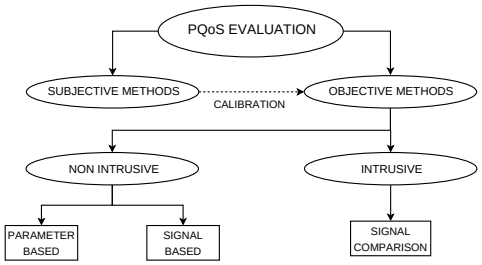

Figure 2: PQoS Evaluation.

Non-intrusive methods do not require any extra sequence, allowing their use in real-time scenarios. They can be classified as either *signal based* or *parameter based*. In the case of signal based methods, the assessment is done without any reference signal, applying complex algorithms to the distorted signal. In the case of parameter based methods, network features (loss probability, loss length, delay, jitter, etc.) and characteristics of the multimedia itself (coding, bit rate, frame rate, nature of the content, etc.) are taken as input data. The idea is to define a mapping function between a PQoS relevant set of these parameters and a quality value as perceived by the user. The recently introduced Pseudo Subjective Quality Assessment (PSQA) methodology uses a statistical learning algorithm (Random Neuronal Networks, RNN) to *learn* the mapping between parameters and user perceived quality. The main drawback of parameter-based methods is their strong dependence on subjective tests' results for calibration/training. A whole in depth description of presented algorithms and techniques is provided in [2].

In [1] we develop an end-to-end PQoS evaluation tool, the PQoSET (PQoS Evaluation Toolbox), which includes all intermediate steps for PQoS estimation in multimedia services (live streaming, network state estimation, multimedia capture, etc.), implementing the different algorithms described above for video and audio evaluation.

2. AN IMPROVED SYSTEM FOR QOS EVALUATION

Each proposed technique partially solves the initial problem: the former is a low network loading estimation method,

but it is focused on network parameters estimation and so, results can be misaligned with user's experience. The latter is focused on quality as perceived by the end user, but the estimation process is generally *intrusive*. We propose an improved evaluation system by merging both estimation methodologies into one single performance evaluation tool. The target is to obtain the least-intrusive PQoS estimation system, using current implementations. Different integration procedures are described below.

2.1 PSQA embedded in MetroNet system

The easiest integration procedure is to directly embed the PSQA methodologies into MetroNet's system. This is in fact the current implementation. We use the already trained RNNs for audio and video quality assessment. This presents in fact an important problem: these RNN were trained using a particular network parameters' estimation methodology. This method uses probe traffic of similar characteristics to the multimedia service under evaluation to compute losses, jitter, delay, bursty losses, etc. However, MetroNet uses light probe traffic with different characteristics during the QoS estimation stage, so estimations obtained from this probe traffic can not be directly used with current RNN training.

A possible solution to avoid training the RNN once again would be to use the network parameters estimation achieved with MetroNet as input for the RNN. Firstly, network parameters are estimated with MetroNet system. Using these estimations and current multimedia features as inputs to the RNN, the estimated PQoS is computed. The obvious problem of this solution is the error propagation that results from two consecutive estimation processes.

2.2 Functional regression with PQoS as performance metric

PQoS can be used as performance index Y in 1.1. Instead of delay, loss probabilities or any other objective performance index, subjective quality can be computed over the transmitted multimedia sequences. During the training step, the final user rates the transmitted sequence quality according to a MOS scale. The whole evaluation procedure remains unchanged. However, transmission methods must allow the user to watch or listen to the transmitted sequences in order to assess it. For this purpose, a subjective evaluation module has been developed and integrated into MetroNet system. The system sends a train of probe packets, followed by the live streaming of a short sequence with similar characteristics to those specified by the user (codec, bit rate, motion-level in video, etc.). The user then rates the transmitted sequence, according to the experienced quality. This score is recorded by the system into the *user's history* for training purposes. Once the system has been trained, the user obtains an estimation of the PQoS he would get at any time without actually transmitting any multimedia sequence.

3. EXPERIMENTS AND RESULTS

We present the evaluation of both individual systems. The combined tool is a prototype and its validation is still in progress. The PQoSET is evaluated over a 150 samples dataset, each of them consisting in the set of parameters used during the transmission of a multimedia sequence and its corresponding subjective evaluation [2]; in the case of MetroNet, we conduct some experiments over the Internet.

Method	MAE		CF	
EMBSD	0.59	-	0.76	-
PESQ	**0.43**	**0.11**	**0.88**	**0.93**
MNB	0.68	-	0.65	-
PSQA	**0.45**	**0.16**	**0.83**	**0.86**

Table 1: Mean Absolute Error (MAE) and Correlation Factor (CF) (*left.* training, *right.* validation).

3.1 PQoSET Evaluation

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 ones (subjective tests). Intrusive methods' results are not in the same scale as DMOS values (they are correlated with human assessment but each one uses its own scale), so a calibration phase is conducted before the comparison. As regards non-intrusive algorithms, the system must be trained. In both cases we split the dataset in a *training dataset* and a *validation dataset*. With the first set we calibrate/train the intrusive/non-intrusive methods, with the second we perform the validation. In the case of video, we consider 70% of samples for training and 30% for validation. In audio, the relation is $80\% - 20\%$.

3.1.1 Audio analysis

Table 1 presents the training/calibration MAE values and Correlation Factor (CF) between real and estimated DMOS for the implemented algorithms. It is clear that PESQ and PSQA present the best performance (in both cases, table 1 also shows the values obtained with the validation dataset). Compared with the other intrusive algorithms, PESQ has a major advantage: it includes a temporal re-synchronization phase that allows an accurate signal comparison (in the presence of data losses, a direct signal comparison without synchronization may result in very poor performance). It is important to recall that PESQ is the actual ITU recommendation for voice perceived quality assessment [2]. Figure 3 presents the results obtained with PSQA and PESQ in the validation dataset. There is an important difference in MAE values between training and validation datasets; this shows the possible presence of outliers among the samples. Nevertheless, the obtained results make clear the advantages of these algorithms.

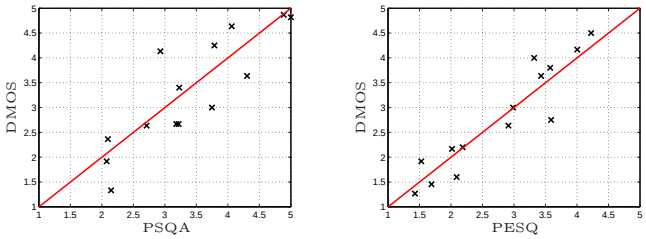

Figure 3: Audio evaluation - validation set.

3.1.2 *Video analysis*

In video analysis, PSQA presents the highest performance, and not only because of the smallest error value, but mainly because of the time involved in the estimation. Table 2 summarizes these observations. Figure 4 shows the different

Method	MAE	ACT (seconds)
SSIM	0.60	> 600
PSNR	0.48	≈ 20
PSQA	**0.40**	$\approx$ **1**
TSSDM	0.53	> 1200

Table 2: MAE (validation set) and Average Computing Time (ACT).

algorithms along with their respective fit curves, considering all samples (training and validation). In the case of PSQA, a straight line Subjective DMOS = PSQA is plotted to see the quality of the results. There are no standardized meth-

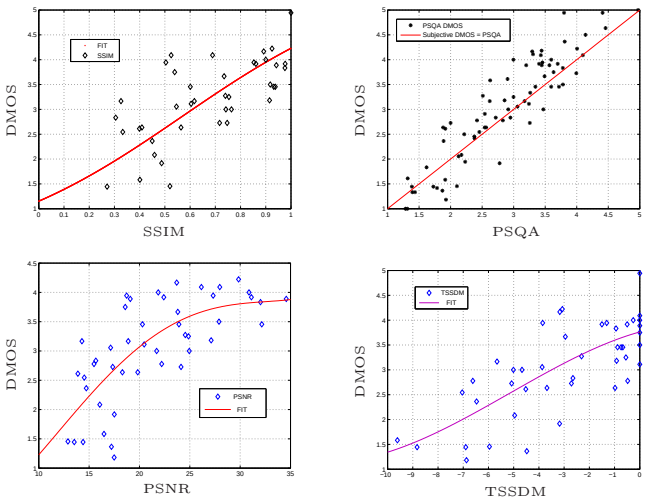

Figure 4: Video evaluation, different algorithms.

ods for perceived quality assessment in video transmissions, something that shows that PQoS for video is still an unsolved problem. The intrusive methods presented in this work suffer from the same synchronization problem previously described. On the other hand, we can appreciate how the PSQA algorithm is capable of identifying the implicit relation between perceived quality and performance parameters.

3.2 MetroNet Evaluation

We present the performance evaluation of a video transmission, considering two performance indices: *delay* and *loss probability*. We compare the estimation computed by MetroNet system with the real values, directly obtained from the video transmission. The evaluation is conducted from a standard ADSL home connection of 512*kpps*. In order to appreciate the learning step, the evaluation process presented in 1.1 is slightly modified. We consider different *tests*, each of them represents an iteration of the learning step. After single test i, a sample $\left(X_i, Y_i^1, Y_i^2\right)$ is obtained, where X_i is the *edf* introduced in 1.1 and Y_i^1, Y_i^2 are the real measured delay and loss probability (loss percentage) respectively. Using the first $(i-1)$th samples, quality functions $\hat{\Phi}_{i-1}^1$ and $\hat{\Phi}_{i-1}^2$ are computed. If X_i is "far" from previous X_{i-1}, X_{i-2} $\ldots X_1$ (we use the L^1 norm as distance between *edfs*), estimated delay $(\hat{Y}_i^1)$ and loss probability $(\hat{Y}_i^2)$ are set to 0. In other words, if X_i does not belong to previous learning space, current estimated function $\hat{\Phi}_i$ is useless for the evaluation. Otherwise, $\hat{Y}_i^1 = \hat{\Phi}_{i-1}^1(X_i)$ and $\hat{Y}_i^2 = \hat{\Phi}_{i-1}^2(X_i)$. Figure 5 presents the comparative results

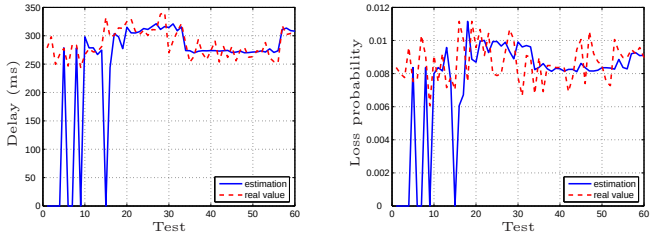

Figure 5: Performance evaluation for a video transmission; *left.* delay, *right.* loss probability.

for 60 consecutive tests (delay on the left, loss probability on the right). As expected, first estimations show a lack of representative data. After the tenth test, the estimation begins to be consistent. At the fourteenth test we start a peer-to-peer connection from the client. This new application modifies the traffic (both delay and losses increase), thus estimation 15 goes back to 0. After test 17 the estimation tracks quite well the real values. It is interesting to note that following estimations remain stable, even after turning off the peer-to-peer application at test 35, and turning it on again at test 55. This shows that previous learning step was good enough to conduct the estimation.

4. CONCLUSIONS

In this paper we have addressed the end-user Quality of Service problem from different points of view. Based on the identification of two important problems in end-to-end quality assessment for multimedia services, we introduced two different methodologies that partially solve them. Different experiences were conducted with two complete systems that implement each of these methodologies, not only for validation purposes but also for comparing the goodness of different algorithms. In light of the obtained results, we introduced an enhanced performance evaluation system based on the integration of previous implementations.

The integrated system is still a prototype and more experiences should be carry out in a general Internet like environment to validate the implementation and to continue with the PQoS study. The problem of QoS evaluation from the user's perspective represents a complex and active subject. We believe that this system for light end-to-end PQoS evaluation will provide interesting keys for future research in the field.

5. ACKNOWLEDGMENTS

This research was partially supported by the *Programa de Desarrollo Tecnólgico* (PDT), grant S/C/OP/46/03.

6. REFERENCES

[1] P. Casas, D. Guerra and I. Irigaray, User Perceived Quality of Service in Multimedia Networks: a Software Implementation, *MVD Telcom 2006*, 2006.

[2] P. Casas, D. Guerra and I. Irigaray, Perceived Quality of Service in Voice and Video IP services, Technical Report, UDELAR, 2005.

[3] L. Aspirot, P. Belzarena, G. Perera and B. Bazzano, End To End Quality of Service Prediction Based on Functional Regression, *HET-NET's '05*, 2005.

[4] P. Belzarena, V. González Barbone, F. Larroca and P. Casas, MetroNet: Software para medición de Calidad de Servicio en Voz y Video, *CITA '06*, 2006.