Quality Assessment of Interactive Voice Applications ⋆

Ana Paula Couto da Silva [a], Martín Varela [b,*], Edmundo de Souza e Silva [a], Rosa M.M. Leão [a] and Gerardo Rubino [c]

[a] *Federal University of Rio de Janeiro/COPPE/PESC/CS Departament Rio de Janeiro, Brazil*

[b] *VTT – Technical Research Centre of Finland, Oulu, Finland*

[c] *INRIA/IRISA, Campus de Beaulieu, Rennes, France*

Abstract

The conversational quality of a VoIP communication is dependent on several factors such as the coding process used, the network conditions and the type of error correction or concealment employed. Furthermore, the quality perceived by the users is also dependent on the characteristics of the conversation itself. Assessing this kind of communication is a very difficult problem, and most of the studies available in the literature simplify the issue by restricting the analysis to only one or two parameters. However, the number of potentially affecting factors is typically higher, and their joint effect on quality is complex. In this paper we study the combined effects of bit rate, forward error correction, loss rate, loss distribution, delay and jitter on the perceived conversational quality. In order to achieve this we use the Pseudo-Subjective Quality Assessment (PSQA) technique, which allows us to obtain accurate, subjective–like assessments, in real time if necessary. Our contributions are thus two–fold: firstly, we offer a detailed analysis of the impact of these parameters and their interactions on the perceived conversational quality. Secondly, we show how the PSQA methodology can be used to provide accurate conversatinal quality estimations.

Key words: Pseudo-Subjective Quality Assessment (PSQA), Quality of Service (QoS), Voice over Internet Protocol (VoIP) applications.

1 Introduction

The number of multimedia applications running over the Internet has been steadily increasing lately. Voice over IP [1], IP telephony [2], audio streaming, videoconferencing, are becoming commonplace. Nevertheless there are still major challenges to be overcome in order to provide the end user with acceptable levels of service quality, especially over connections with long propagation delays, great differences in channel speeds (from high backbone speeds to comparatively low *last mile* speeds) and physical transmission media (e.g. from fiber cables to wireless). One of the major issues these applications face is to maximize the perceived QoS (this is often referred to as "Quality of Experience", or QoE) for drastically varying network states.

Since quality is not guaranteed in the current Internet, it is important for the QoS to be constantly monitored so that the applications can take the proper actions needed to maintain it over some minimum level. Therefore, it is essential to determine what are the parameters that mostly influence the user's perception of the QoS and to understand their combined effects on quality from the user's perspective.

In this paper we are concerned with the conversational quality of a voice over IP session. We show that the conversational quality, as perceived by the end user, depends on a complex combination of several parameters. We present a method that can map the values of such a parameter set into a single numerical score that is close to that a human observer would give to an interactive conversation session in a *subjective test*.

Subjective quality assessment methods measure the perceived quality from the user's perspective and as such embody the subjective sensations inherent to humans. For interactive multimedia streams, the ITU–T P.800 [3] and ITU–T P.920 [4] recommendations give guidelines on how subjective assessment should be performed, define the environmental setup and provide information

⋆ A.P.C.da Silva was supported by a fellowship from CAPES during her visit at IRISA and CNPq. E. de Souza e Silva and Rosa M.M. Leão were supported in part by grants from CNPq and FAPERJ. M. Varela's work was partly carried out during the tenure of an ERCIM fellowship both at VTT and at the Swedish Institute of Computer Science in Kista, Sweden.
* Corresponding author: VTT Technical Research Centre of Finland, PL 1100, 90571 Oulu, Finland. Tel: +358 20 722 2495. Fax: +358 20 722 2320.

Email addresses: anapaula@land.ufrj.br (Ana Paula Couto da Silva), martin.varela@vtt.fi (Martín Varela), edmundo@land.ufrj.br (Edmundo de Souza e Silva), rosam@land.ufrj.br (Rosa M.M. Leão), rubino@irisa.fr (Gerardo Rubino).

on the kinds of tasks that the test subjects should perform. As with other subjective assessment techniques, the result of these kinds of tests is a Mean Opinion Score (MOS), which gives a numeric expression of subjective quality. Normally, subjective tests involve a relatively large group of subjects who (in the case of conversational quality assessment) carry out a conversation (normally based on some tasks) over the system to be tested, and then grade the quality as they perceived it. The ITU recommendations suggest a 5–point scale, which spans from bad to very good quality.

Although subjective assessment has served as the basis for analyzing many aspects of speech quality, this kind of tests are very costly to perform, and require very stringent requirements (not generally available in most laboratories) in order to be in full compliance with the standards. They are therefore not desirable for very large–scale tests or tests that need to be carried out regularly. Also, given their nature, they are obviously not suitable for real-time operation.

Therefore, significant efforts have been devoted to the development of *objective* quality assessment technologies. Most objective metrics propose different methods to compare the received sample against the *original* one. While these metrics lower the cost of quality assessment, their correlation with subjective scores can sometimes be low, mainly when networking parameters are taken into account [5–7]. Furthermore, these metrics are geared toward listening quality, not conversational quality. This is an important issue, since listening quality only concerns the actual sound quality of a speech stream, whereas the conversational quality refers to the overall quality of the conversation, including interactivity aspects. Also, since the original signal is generally needed in order to perform the assessment, most objective tests are not suitable for use in real time. Real–time quality assessment is essential for instance if the application is to perform some form of dynamic quality control, e.g. by changing encoding or redundancy parameters to optimize quality when network conditions worsen.

Almost every objective assessment tool currently available deals only with listening quality. Some of the most well known metrics for listening quality assessment of VoIP are Signal-to-Noise Ratio (SNR), Segmental SNR, Perceptual Speech Quality Measure (PSQM and PSQM+) [8,9], Measuring Normalizing Blocks (MNB) [6], ITU E–model [10] and PESQ [11]. Currently, PESQ, and to some extent the E–model, are the most widely used in the industry. From the metrics mentioned above, PESQ provides the best correlation with subjective scores for listening quality, but it is not suitable for conversational quality assessment. The ITU–T P.563 algorithm [12] can provide a single–sided estimation of listening quality, but it does not correlate very well with subjective scores.

The E–model provides an estimation of the conversational quality based on several network, environment and coding parameters, and thus does not need the original signal. This tool, however, was designed for planning purposes and not for quality assessment (although it is often used to this end), and in many cases its results do not correlate well with human perception [13].

PSQA (Pseudo–Subjective Quality Assessment, described in Section 3) [14–16] is a technique based on merging subjective assessment with a statistical learning tool (a Random Neural Network, or RNN [17]), which allows to produce subjective–like quality estimations based on measurable network and application parameters. The main advantage of this approach is that it provides results very close to actual MOS values, while being cheap (in terms of cost, time, and computational resources), and suitable for real–time applications.

We have used PSQA in previous work [15] for speech quality assessment of VoIP transmitted over the Internet. In that paper, we explored the approach in analyzing one–way communications. In this work, we present the analysis of conversational quality, thus also considering the interactivity aspects of the call, instead of focusing on the listening quality. Our first contribution is to show that the PSQA approach can be successfully used in a conversational quality assessment context. This was not obvious because in two–way flows new parameters related to interactivity need to be considered. Then, our main contribution is a detailed analysis of conversational quality in an echo–free VoIP environment, as a function of many different parameters, as well as an extension of our assessment methodology to provide real–time conversational quality assessment. To the best of our knowledge, there are no previous results providing a comprehensive view of the combined effects of all the parameters considered herein on the perceived quality of a VoIP session, nor any other automated tool capable of providing real–time conversational quality assessment.

The rest of the paper is organized as follows. Section 2 briefly surveys the current literature on the subject. Our quality assessment methodology is discussed in Section 3. The experiments we performed are described in Section 4 and our main results are presented in Section 5. In Section 6 we summarize our conclusions.

2 Related works

There is a large body of literature on the impact of quality–affecting parameters – such as the codec employed, the redundancy, the packetization interval (PI), the network loss rate (LR), the mean loss burst size (MLBS), the one-way delay and jitter on the quality of multimedia applications. Most of these

works are related to streaming, i.e. one–way, applications (e.g. video, or listening quality for VoIP) and are focused on the effect of one or the combined effect of at most two quality–affecting parameters. In [18], the effect of both LR and PI on speech quality is presented. The study is based on an automatic speech recognition system which is in turn based on hidden Markov models instead of the usual subjective quality tests. A similar Neural Network–based study (itself based on a more primitive incarnation of PSQA) is presented in [19], using data obtained from PESQ. The effect of packet loss on several speech codecs is evaluated in [20]. Other works [21] aim at studying the performance of speech codecs.

In [22] the authors describe the effects of transmission delay on speech quality in a telephone conversation for a traditional Public Switched Telephone Network (PSTN). Their main goal is to obtain subjective assessments using a set of tasks which are representative of different types of conversations. They conclude that long round–trip transmission delays in the range of 500ms give considerable difficulties to subscribers. This is consistent with other works, notably the ITU-T G.114 [23] recommendation, which defines similar delay thresholds for interactive speech. However, the accuracy of the G.114 recommendation has been questioned in [24], and given the results we have obtained (cf Section 5), we agree with this questioning. Other works investigating pure delay effects on conversational quality [25,26] also present results which contradict with previous "common knowledge" results. Some intriguing results on the combined effects of echo and delay and their interaction can be found on [27]. In that work, the authors present results are based on independent tests performed in three separate laboratories, in which the presence of echo might make longer one-way delays desirable under certain circumstances, which also contradicts "common knowledge". These findings also indicate that the current state of the art is lacking when it comes to understanding the interactions between the different parameters which affect the perceived quality.

The authors of [28] discuss the factors affecting real-time multimedia QoS. They propose the joint use of an n-state extended Gilbert model and an inter-loss distance (ILD) metric to characterize loss burstiness. However, no subjective assessments were performed in order to study how loss burstiness is related to perceptual quality.

In [29] measurements collected over backbones of major Internet Service Providers were used for assessing the perceived quality of telephone calls based on the E–model [10]. The effects in the voice quality due to echo and encoding are also studied.

The work in [30] describes the state of the art of perceptual QoS assessment methodologies for VoIP systems. Past and current activities of the ITU for the objective quality assessment are discussed as well as perceptual QoS as-

sessment methodologies for the next generation multimedia communications systems. [31] reviews the development of perceptually-motivated models for quality assessment of speech transmission / storage systems.

An effort toward improving the accuracy of the E–model can be found in [32]. The authors studied the performance of the E–model for estimating quality considering speech distortion, delay, talker echo and loudness. A new model based on the E–model is proposed for improving its accuracy as compared to MOS values.

Our previous work [15] studies the listening quality of VoIP streams as a function of several network and coding parameters (codec used, packetization interval, etc). This work was also based on the PSQA methodology described in the next section. As we only considered unidirectional streams and assessed listening quality only, interactivity–affecting factors such as delay and jitter were not taken into account.

Finally, some new methodologies for subjective testing are currently being developed [33,34]. In these papers, the authors propose new methodologies for assessing the perceived quality via its impact on users' performance. These methodologies aim at providing testing scenarios and metrics which may better reflect the way in which the quality of a conversation affects the end users.

3 Overview of PSQA for conversational quality assessment

In this section, we briefly describe our technique to automatically assess the quality of an VoIP conversation carried over the Internet. As mentioned previously, PSQA allows to analyze, in real–time, how a set of parameters actually affect the perceived quality. It can help to develop quality–driven control mechanisms, in order to improve the perceived quality of a voice stream, or to keep it within certain bounds in order to help control resource allocation if need be.

To implement PSQA, three main steps must be followed: (a) a set of (a priori) quality–affecting parameters must be selected; (b) a (set of) subjective tests session(s) must be performed, and (c) 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. In general, network parameters such as those related to the loss process,

delay and jitter tend to be an obvious choice, but one could also consider other parameters, such as MAC–layer scheduling (see for instance [35]), Diff-Serv marking schemes for intra–flow packet prioritization [36], etc. At the application level, even more parameters can be considered, depending on the application itself. 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 (e.g. the delay) while others are not (e.g. the FEC algorithm). 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 (b), we need a VoIP 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 are paired and an interactive VoIP connection is established for each pair. Then, we select different combinations of values of the selected variables (we call them *configurations*), and for each of them we emulate the corresponding network conditions. As the number of possible parameter configurations is tipically 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. More details on the configurations selected for this study can be found in the following Section.

Each chosen configuration has then a concrete value for each parameter. For instance, we will choose a bit rate of 11Kbps, packet loss rate of 2% and a mean delay of 110ms, etc., and we will establish a VoIP session between two subjects where the network operates with 2% of losses and the chosen mean delay, the VoIP tool uses a 11Kbps bit rate, etc.

The human subjects evaluate the quality of a conversation in those conditions and using many pairs of subjects for each of the selected configurations, we obtain a MOS value. The methodology we use for this assessment is described in the ITU-T P.800 and P.920 recommendations. 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 which might have given unreliable results [37], the average of the scores given by the remaining subjects to each configuration is computed. See Section 4 for more details on the experimental setup.

After step (b) we have a database (actually a table) associating the values defining each configuration with the corresponding MOS. Step (c) 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).

The RNN model has been chosen over other statistical estimators for its very good generalization capabilities. A comparison of the performance of PSQA when implemented with RNN and other tools such as Artificial Neural Networks and Bayesian classifiers can be found in [38]. The reader may also refer to [15,14,39] for more details on the RNN model, and how it is used in PSQA.

4 Experiment description

Assessing the quality of interactive VoIP streams is a much more difficult task than that of assessing the quality of one–way streams. Not only more parameters need to be considered, such as the delay and jitter, but other factors, such as the interactivity "level" should be taken into account as well. Another example is the intelligibility problems that may arise due to double–talk. The methodology for carrying subjective assessment of interactive multimedia is specified in the ITU-T recommendations P.800 [3] and P.920 [4], which provide the definition of the environment and test setup, and the interactive tasks that to be used, respectively.

We considered six parameters which affect the perceived quality. Four of them relate to the network state and the remaining two concern the encoding schemes used. The network parameters are the loss rate (LR), the mean size of loss bursts (MLBS), the mean one–way delay and jitter (as a fraction of the delay). The LR is defined as usually: the ratio of lost to sent packets. The MLBS is the average number of packets in a loss event, and defines the

"burstiness" of the loss process.

Some parameters, such as packet loss and delay, should be considered as random variables, and they are a bit more complex to handle than other variables since we need to choose a proper model or distribution for them. The parametrization of the models should be simple to keep the overall computational complexity of the method low.

In the case of packet loss, several mathematical models can be found in the literature to represent the packet loss process in the Internet [40–43]. Many authors argue that relatively simple models, such as a two–state homogeneous Markov chain, provide a good approximation of the packet loss processes [44,41]. Therefore, we chose a simplified Gilbert model which is a two–state Markov model with two degrees of freedom for our tests. We choose this over the original Gilbert [45] model since it requires only two parameters instead of the three needed for the original one. These model parameters can be easily derived from the target values of the LR and MLBS (the reader may refer to [14] for a detailed explanation of this mapping).

The mean one–way delay refers to the network delay (we did not consider processing delays since they can be seen as constant, and on the hardware we used, they are mostly negligible). Since we considered a very wide range of delays, we considered jitter values proportional to them (which prevents having incongruent combinations of delay and jitter). These jitter values were then fed into the network emulator module, which used them as a base to calculate the actual delay of each packet, including the jitter.

The network emulator we used was the NetEm [46] Linux kernel module. NetEm represents the delay by a constant value with random increments/decrements. In this model it is possible to define a correlation value between two consecutive delay samples. The correlation approximates a temporal dependency by limiting the value of the next sample within a given interval centered at the current sample. The jitter definition used by NetEm is the expected value of the absolute difference between actual packet delay and the mean delay. Since we consider the jitter values to be dependent on the mean delay $\overline{D}$, we normalize its value by $\overline{D}$. In other words, if D is the random variable that takes values from the set of actual packet delays in the voice stream, then we define jitter as $E[|D - \overline{D}|]/\overline{D}$.

Several real Internet delay and jitter traces were collected using the Traffic Engineering module of the Tangram–II tool [47]. They were collected during several days in March 2005 at three different times per day between the University of Massachusetts (USA) and the Federal University of Rio de Janeiro (Brazil). These traces were used to validate the model implemented by NetEm. We found that the delay and jitter obtained from the emulator were statisti-

cally similar to those from the real traces.

For the encoding, we considered the FEC scheme proposed in [48] and the bit rate. We used the Speex codec for our experiments (cf [49] for a more detailed discussion about the codec). Speex is a free, open source codec which is widely available and is being currently used in several VoIP applications, such as Linphone [50], Gnomemeeting [51], and the software–based Asterisk PBX [52]. It also offers robust, high quality speech coding even at very low bit rates, which makes it an attractive option for any IP–based telephony application. We found these characteristics to be a compelling reason to use Speex in our experiments, even though other codecs such as G.711, G.723 or G.729 are more widely used in IP telephony.

We considered an echo–free environment (which is a normal situation for users using an all–IP network and headsets), and thus did not consider echo impairments in our tests.

Table 1 shows the ranges used for the parameters considered. In this table, the FEC scheme 1:2 is capable of correcting loss bursts of size 1 and scheme 1:2::3:6 may correct long loss bursts, up to size 4, depending on the loss pattern in the stream. The FEC scheme used is based on XOR operations among packet groups of different lengths. This scheme works by dividing the packets into windows of a given length l, and dividing each window into s non–overlapping subsets. An XOR is then performed among packets in each subset, and the results are piggy–backed on the first s packets of the next window (these configurations are noted as $s : l$). More than one of these schemes can be overlapped in order to obtain better protection against losses. A detailed analysis of the performance of this scheme can be found in [48].

The VoIP tool employed is called VivaVoz, and was developed at the Federal University of Rio de Janeiro [53]. VivaVoz is an open source tool and it was adapted to shape it in accordance with the testing conditions required (e.g. eliminating visual clues of the bit rate or FEC level used).

As interactive testing requires a live network, we needed to recreate the desired network conditions for each configuration in our testing environment. To this end, we used a Linux host acting as a router between the two PCs hosts running VivaVoz. Using a production network or a test network with added background traffic does not allow for precise control of the network parameters, which is needed for this kind of tests. We therefore used the NetEm [46] Linux kernel module in order to recreate network conditions similar to those found in the Internet, introducing controlled delay, loss probability, mean loss burst size and jitter. We made some modifications to NetEm in order to enable it to generate losses according to the Gilbert model, instead of just generating independent losses.

Table 1
Network and encoding parameters used for the subjective tests.

Parameter	**Values**
Loss rate	0%...60%
Mean loss burst size	1...5
Mean one–way delay	0ms...600ms
Jitter (as a fraction of the delay)	0%...40%
Bit rate (Speex codec)	2.4Kbps...24.8Kbps
FEC (media–independent)	off, 1:2, 1:2::3:6

As stated on Section 3, the six parameters considered, and their possible values, yield many possible configurations which should be assessed. In order to perform the subjective tests themselves, we need to choose a subset of parameter configurations so as to have a reasonable number of sequences to assess but, at the same time, to have enough data to train and validate the RNN. It is worth noting that the total number of configurations used should not be very large, since long subjective tests result on the subjects' fatigue and influence the results obtained, decreasing their accuracy. Based on our previous experience with the listening quality tests, we selected 120 configurations among all the possible ones (this number represents a good compromise between the total data for training and validating the RNN and the length of the subjective tests). Of these, 48 correspond to *border conditions*, in which one or more parameters are at extreme values. The rest were randomly chosen inside the configurations space, presented on Table 1. The random selection had a bias toward what we consider normal operating conditions, namely: low loss rates and mean loss burst sizes, low delays and jitter values, and medium to high bit rates. Note, however, that there is no optimal way of choosing the configurations to be used for testing, since the actual operating conditions for VoIP applications are too variable.

The campaigns were performed at INRIA/France and at UFRJ/Brazil with 12 subjects who had previous experience with VoIP use. This restricts the scope of our study to this kind of demographics. For each configuration, two tasks were performed, the idea behind the different tasks being to have conversations at different levels of interactivity. These variations on interactivity may have an impact on the perceived effects of delay on the quality. However, since we do not dispose of an accurate way of quantifying the interactivity, we cannot (yet) consider it as an input to PSQA. By including different levels of interactivity in the subjective tests, however, this parameter is folded into each configuration, and included into the overall assessment. The following tasks were considered:

- counting up to 20 as quickly as possible while alternating speakers, and
- a conversation, which was based on a scenario, on a picture, or a free topic.

For each set of tasks and the respective configuration, subjects gave an overall conversational quality, based on their perception. These grades range from 1 to 5, with the rating scheme given in Table 2.

Table 2
5-point assessment scale.

Numerical Score	**Corresponding Quality**
5	Very Good
4	Good
3	Fair (Toll Quality)
2	Poor
1	Bad

For each configuration, the MOS value was calculated (following the guidelines in [3,4].) As a last remark related to the testing process, as subjects' fatigue has an important impact on the quality of the assessment and on the accuracy of the results, the 120 configurations were divided into 4 sets of 30 configurations each. These tests were then done over 4 sessions, so as to avoid fatigue.

The MOS values obtained, along with their corresponding configurations were used to calibrate the PSQA tool as described in Section 3.

4.1 On PSQA's Performance

Once we had the subjective assessment results, we trained several RNNs with different architectures, and compared their performance in terms of how well they could estimate the subjective scores. This performance evaluation is done with configurations previously unknown to the RNN, so that their ability to generalize is validated. From the 120 configurations considered in our tests, 100 were used to train the RNN and 20 to validate it. The RNN training performance was measured against the validation data set, in order to make sure that the network was generalizing properly, and that no over–training was affecting the results.

Previous results on the assessment of unidirectional voice streams [16] led us to use a very simple 2–layer RNN architecture for estimating the quality

of the streams. This is useful, since it provides a very simple closed–form expression for the perceived quality as a function of the parameters used, and the performance of the simple RNN was very similar to that of more complex ones. We repeated the same procedure for this work but the simplest RNN architecture was not able to capture the conversational quality as accurately as more complex ones. Thus, a 3–layer RNN was used in the present study. Therefore, the results presented herein are those from a 3–layer feed–forward RNN with 6 input, 13 hidden and one output neurons. Figure 1 presents a comparison of the performance of both architectures used for 20 validation configurations. Despite of the accuracy differences, the simple RNN was still able to capture the behavior of the perceived quality when affected by the different parameters.

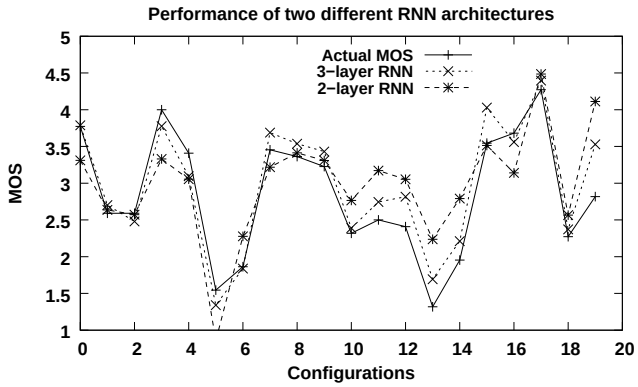

Fig. 1. Actual and estimated MOS values for two different RNN architectures. The results shown are for validation data. The 3–layer RNN has a correlation coefficient of 0.95 with subjective scores, while the 2–layer RNN has a lower correlation of 0.76.

The 3–layer RNN produced a 0.95 correlation coefficient with the validation data (with a MSE of 0.002). These results are very good, especially considering that they outperform most objective metrics for listening quality, while performing in the more complex interactive context.

As a last comment related to the RNN performance, we can state that the RNN generated after the training and validation phases is general enough to consider a large range of values of the network's parameters.

5 Results

In this section we describe the results obtained from our experiments. We do so by showing how the different parameters affect the perceived conversational quality, according to PSQA estimations. We use PSQA in order to be able to cover the whole parameter space, which would not be feasible only with subjective assessments.

Some of the results obtained are quite different of what we had expected, especially with respect to the influence of the delay, jitter and mean loss burst size. The relative impact of losses and delay is particularly interesting, since it opens new possibilities for quality control, which are usually not considered due to delay constraints. Knowing exactly how each of this parameters affects the conversational quality can allow for the appropriate trade–offs to be made between loss concealment/correction and bounding delay, with a better quality as a result.

5.1 Loss rate, delay and FEC scheme

We begin by studying the MOS of an interactive conversation when the loss rate, the delay and FEC scheme vary with the remaining parameters kept constant. To better present the MOS behavior obtained from several combinations of these parameters, we look at the perceived quality as a function of delay, for several loss rates and as a function of loss rate, for several delay values, with and without FEC.

Figures 2 and 3 show the variation of the quality as a function of the delay for five different loss rates, with and without a FEC, respectively. We can observe that for a loss rate of 0%, the MOS drops over approximately 0.6 points when the one–way delay varies from 0 ms to 600 ms. This drop in quality begins to be noticeable for some users. However, as the loss rate increases, the impact of the delay diminishes significantly. The conversational quality is then roughly insensitive with respect to the delay at higher loss rates. When the loss rate is greater than 7% or 10%, the impact of the delay on the perceived quality is almost not noticeable in the ranges considered. This means that the loss rate has a significantly higher impact on conversational quality than the delay. This result implies that the thresholds stated in the ITU-T Recommendation G.114 [23] are not accurate at least for VoIP scenarios, being too restrictive. Other studies [24,25] have also questioned the traditional belief that delays over 150 or 200ms result in an important decrease in conversational quality. The figures also show that the use of FEC produced a significant improvement on the perceived quality, as expected.

Figure 4 shows the effects of the loss rate for several one–way delay values. We can observe that the differences in the resulting quality from different delay values decrease as the loss rate increases.

Figure 5 consolidates the results presented above, showing how the quality evolves with both the loss rate and the one–way delay. It is easy to see that the impact of a very high delay when no losses occur is roughly the same as that of a 5% loss rate when the delay is low. Furthermore, when the loss rate increases

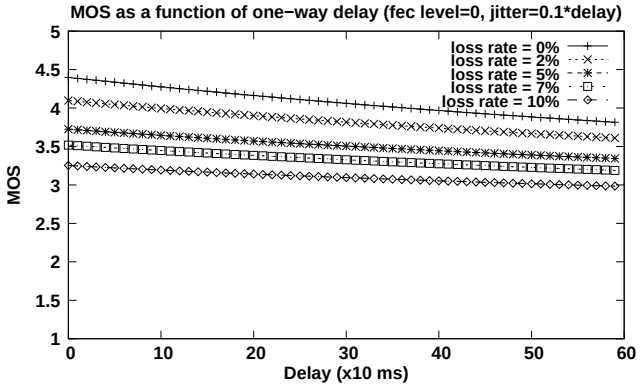

Fig. 2. Perceived quality as a function of delay for some loss rates (without FEC). Jitter was 10% of the delay. Note how the impact of the delay diminishes as the loss rate increases.

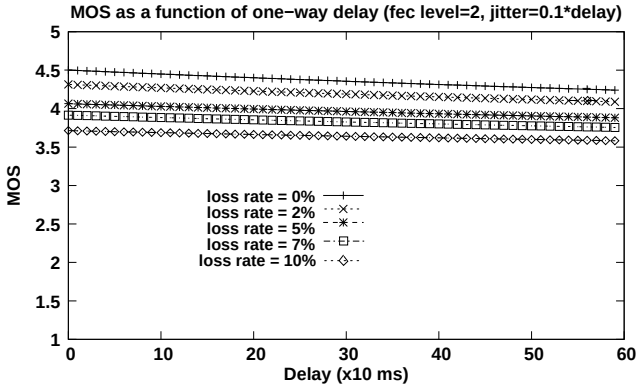

Fig. 3. Perceived quality as a function of delay for some loss rates (with FEC). Jitter was 10% of the delay. Again, we observe that the impact of the delay diminishes as the loss rate increases.

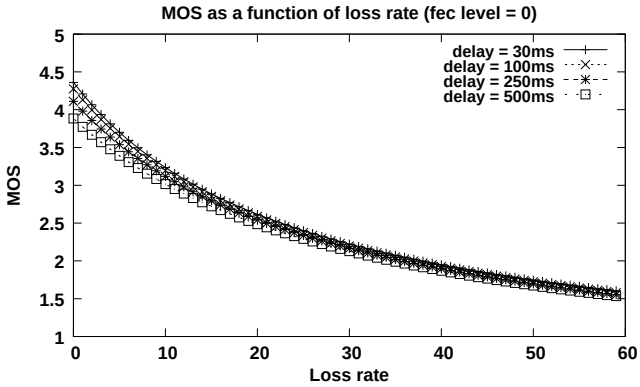

Fig. 4. Perceived quality as a function of loss rate for some delay values (without FEC).

beyond 10%, the impact of delay diminishes significantly. This insight allows to consider new FEC or loss concealment schemes which, although requiring an increase of the mouth–to–ear delay, cope better with losses than current schemes. This might be useful for instance in wireless contexts, which tend to

be prone to high loss rates.

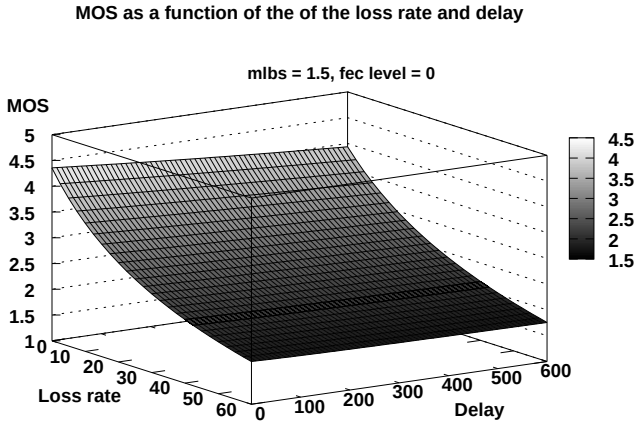

Fig. 5. Perceived quality as a function of loss rate and delay without FEC. Note how the relative impact of the loss rate is much higher than that of the one–way delay.

The FEC scheme used has a very significant impact on the perceived quality. Its use results in acceptable qualities (note that in the MOS scale, a score of 3 is considered acceptable) even at high loss rates. For example, in Figure 6 we can observe that when using FEC, the quality remains acceptable at up to approximately 22% losses. On the other hand, without FEC the quality is only acceptable up to a loss rate of about 10%. The use of this FEC scheme, combined with the good performance of the Speex loss concealment algorithm provides very good results even when the network conditions are severely degraded.

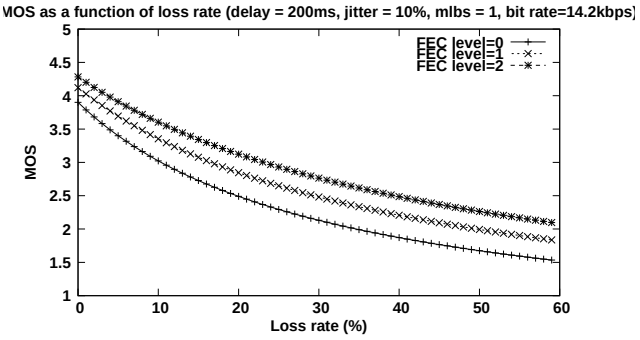

Fig. 6. Perceived quality as a function of loss rate, for three different FEC settings. The use of FEC allows for a very significant increase in the loss rate to be tolerated while maintaining acceptable quality levels.

5.2 Loss rate, Mean loss burst size

In this subsection we evaluate the MOS behavior as a function of the mean loss burst size (MLBS) and loss rate. Quite surprisingly, for the MLBS ranges

considered, the perceived conversational quality was mostly independent of the MLBS value. While there are some slight variations, they are small enough not to be perceptible by the average user.

This result is quite different from previous results we had obtained for listening quality. The PSQA results, however, correlate well with subjective scores, and the behavior of the network emulator used is as expected. A likely explanation for this independence of the quality with respect to loss burst sizes lies in the good performance of the Speex's packet loss concealment mechanism. One should also note that, for the same loss rate, as the MLBS increases the user observes a smaller number of loss bursts. Probably the decrease in the number of loss bursts (together with the Speex's packet loss concealment algorithm) counteracts the increase in the MLBS in the scenarios studied. Figure 7 shows the results obtained with 100ms one–way delay, and 10ms jitter.

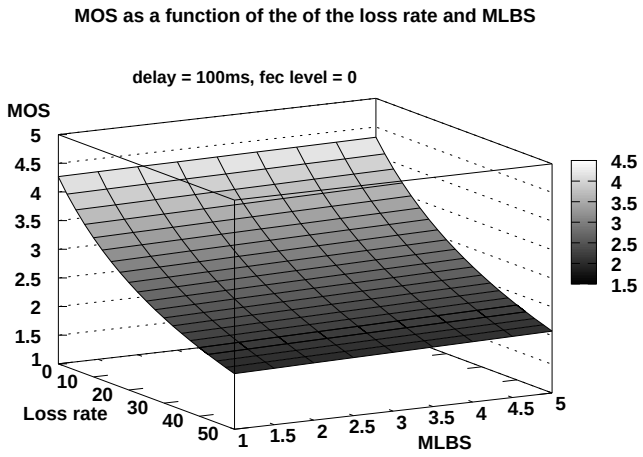

Fig. 7. Perceived quality as a function of loss rate and MLBS values (no FEC), for a delay of 100 ms with 10% jitter.

5.2.1 Jitter

Similar to what was observed for the delay, the impact of the jitter on the conversational quality was not significant. For lower loss rates and high delay values, some variation of the perceived quality is visible, but it is not very noticeable for the average user. It should noted that VivaVoz has a static play-out buffer. Figures 8 and 9 show the variation of the perceived quality as a function of delay and jitter, for loss rates of 0% and 5% respectively.

5.3 Bit rate

We considered eight different bit rates for our experiments, ranging from 2.4Kbps to 24.8Kbps. As expected, the perceived quality varies very signif-

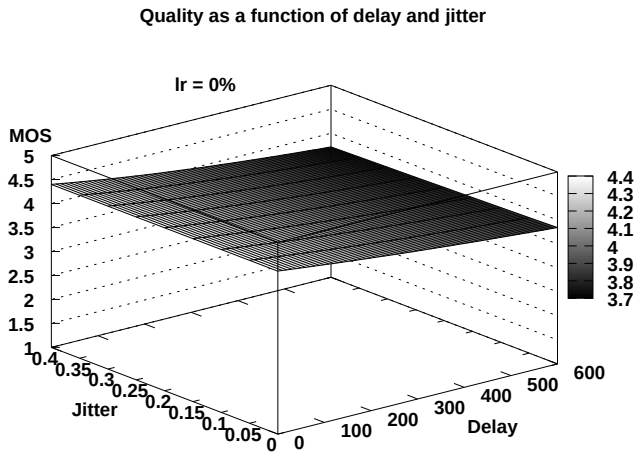

Fig. 8. Perceived quality as a function of the delay and jitter (loss rate is 0%). Notice that the impact of jitter, while visible at very high delay values, is minimal.

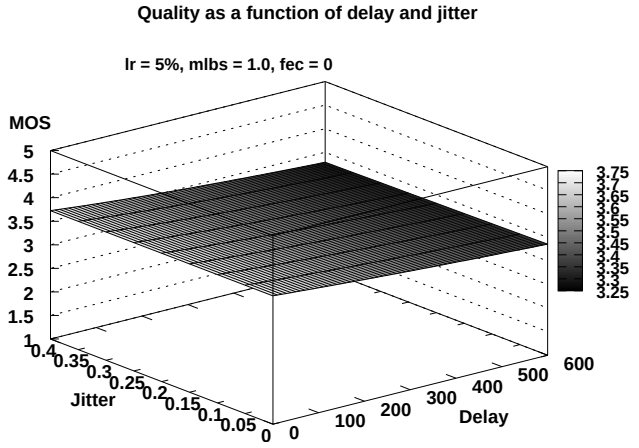

Fig. 9. Perceived quality as a function of the delay and jitter (loss rate is 5%, and MLBS is 1 packet). In this case, the impact of jitter is even lower.

icantly with the bit rate used. However, acceptable qualities are attainable throughout the entire bit rate range for losses of up to 5% when using FEC. The higher bit rates are able to offer acceptable quality even when the loss rate is very high. The "sweet spot" for the bit rate seems to be located at either 11.2 or 14.2Kbps, since these bit rates offer a very good compromise of quality and bandwidth consumption. Figures 10 and 11 show the perceived quality as a function of the bit rate for several loss rate values, with and without FEC respectively. If FEC is not employed, the bit rate should be set at least at 6Kbps in order to obtain acceptable quality values when the loss rate is about 5%.

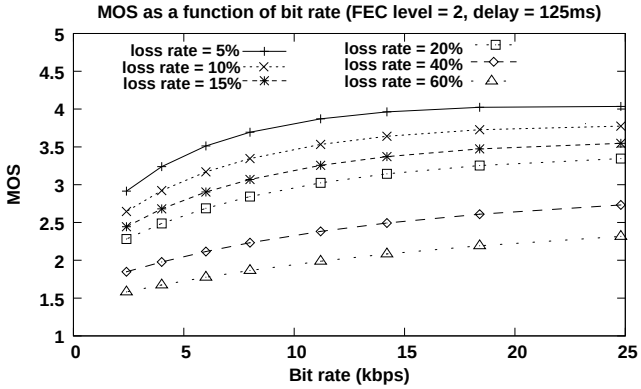

Fig. 10. Perceived quality as a function of the bit rate, for several loss rates. Delay is 125ms, and FEC is being used. Note that all the bit rates provide acceptable qualities up to 5% losses.

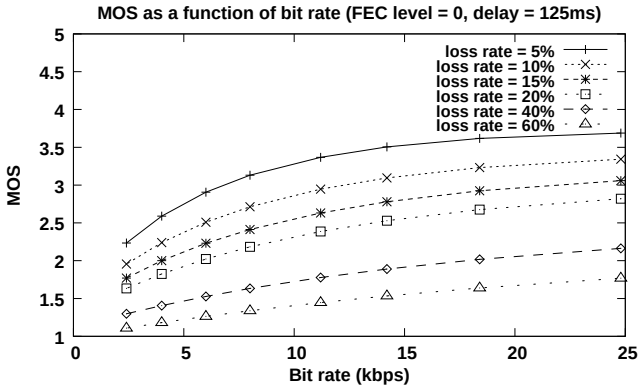

Fig. 11. Perceived quality as a function of the bit rate, for several loss rates. Delay is 125ms, and FEC is not being used. In this case, at least 6Kbps are required to obtain acceptable quality levels at 5% loss rate.

6 Conclusions

In this paper we have presented a detailed analysis of conversational quality as perceived by the users, and studied the impact of six different parameters on quality. The application–level parameters we considered are the bit rate and the FEC scheme used. At the network level, we studied the loss rate, the mean size of loss bursts, the one–way delay, and the delay jitter. To the best of our knowledge, this is the first study of conversational quality covering such a large parameter space.

For the previously described analysis, we have extended the capabilities of PSQA in order to be able to assess the QoE in conversational voice communications, and we have shown that the general approach can be used under these, more complex, conditions. We have then used PSQA to understand how the different parameters affect the perceived quality. The PSQA estimations have a correlation coefficient of 0.95 with subjective scores (this correlation

was of course calculated against validation data, to ensure the generality of the results).

Our main conclusions are that, in the scenarios investigated, the main parameters driving the conversational quality are the packet loss rate, the bit rate used, and the forward error correction mechanism. As for delay and jitter, while they slightly affect quality, their impact is low, and subordinated to that of the loss process. The mean loss burst size does not seem to play a significant role in this context, likely due to the good performance of the FEC mechanism used and the loss concealment algorithm implemented in the codec.

In addition to providing a good understanding of how the conversational quality is defined, the results we have obtained (in the form of a trained RNN) are useful for performing real–time assessment of conversational quality. This can be used, for example, to dynamically control the application (or network, if some QoS mechanism is available) parameters as the conversation takes place, in order to obtain the best possible perceived quality at all times. Since PSQA allows to estimate the effects of parameter changes on quality, it is possible to optimize the perceived quality for any given situation. The reader can refer to [14] for two simple, yet effective, example control algorithms for improving listening quality in VoIP streams.

Future work in this area will include a more detailed view of the interactivity itself. To this end, we might add the level of interactivity (for instance, as defined in [54]) as a quality–affecting parameter. This would imply defining appropriate conversational "tasks" for the test subjects to perform. Another item which is worthy of study is the combined effects of echo and delay, which is relevant in VoIP–PSTN hybrid environments, and also in speaker-phone configurations. Also, another interesting perspective is the study of conversational quality in multi–party calls.

7 Acknowledgements

The authors would like to thank the colleagues from the LAND Laboratory at Federal University of Rio de Janeiro who colaborated for performing the subjective tests. We would also like to thank Ian Marsh for his valuable comments on the draft versions.

References

[1] A. Cray, Voice Over IP: Hear's How, Data Communications International 27 (5) (1998) 44–59.

[2] B. Ahlgren, A. Andersson, O. Hagsand, I. Marsh, Dimensioning Links for IP Telephony, Tech. Rep. T2000–09, Swedish Institute of Computer Science (SICS) (2000).

[3] ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality (1996).
URL `http://www.itu.int/`

[4] ITU-T Recommendation P.920, Interactive Test Methods for Audiovisual Communications (2000).
URL `http://www.itu.int/`

[5] W. Yang, Enhanced Modified Bark Spectral Distortion (EMBSD): an Objective Speech Quality Measrure Based on Audible Distortion and Cognition Model, Ph.D. thesis, Temple University Graduate Board (may 1999).

[6] S. Voran, Estimation of Perceived Speech Quality Using Measuring Normalizing Blocks, in: IEEE Workshop on Speech Coding For Telecommunications Proceeding, Pocono Manor, PA, USA, 1997, pp. 83–84.

[7] A. Rix, Advances in Objective Quality Assessment of Speech over Analogue and Packet-based Networks, in: the IEEE Data Compression Colloquium, London, UK, 1999, pp. 10/1–10/8.

[8] J. Beerends, J. Stemerdink, A Perceptual Speech Quality Measure Based on a Psychoacoustic Sound Representation, Journal of Audio Eng. Soc. 42 (1994) 115–123.

[9] J. Beerends, Improvement of the P.861 Perceptual Speech Quality Measure, ITU-T SG12 COM-34E (dec 1997).
URL `http://www.itu.int/`

[10] ITU-T Recommendation G.107, The E-model, a Computational Model for Use in Transmission Planning.
URL `http://www.itu.int/`

[11] ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (Pesq), an Objective Method for End-To-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs (2001).

[12] I.-T. R. P.563, Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications (2004).

[13] T. A. Hall, Objective Speech Quality Measures for Internet Telephony, in: Voice over IP (VoIP) Technology, Proceedings of SPIE, Vol. 4522, Denver, CO, USA, 2001, pp. 128–136.

[14] M. Varela, Évaluation Pseudo-Subjective de la Qualité d'un Flux Multimédia et ses Applications au Contrôle , Ph.D. thesis, INRIA/IRISA - Universit de Rennes 1 (2005).

[15] G. Rubino, M. Varela, S. Mohamed, Performance Evaluation of Real-time Speech through a Packet Network: a Random Neural Networks-based Approach, Performance Evaluation 57 (2) (2004) 141–162.

[16] G. Rubino, M. Varela, A New Approach for the Prediction of end-to-end Performance of Multimedia Streams, in: Proceedings of the First International Conference on Quantitative Evaluation of Systems (QEST'04), 2004.

[17] E. Gelenbe, Random Neural Networks with Negative and Positive Signals and Product Form Solution, Neural Computation 1 (4) (1989) 502–511.

[18] J. Hooper, M. Russell, Objective Quality Analysis of a Voice Over Internet Protocol System, IEEE Electronics Letters 36 (22) (2000) 1900 –1901.

[19] L. Sun, E. Ifeachor, Perceived Speech Quality Prediction for Voice over IP-based Networks, in: Proceedings of IEEE ICC'02, New York, USA, 2002, pp. 2573–2577.

[20] A. Choi, A. Constantinides, Effect of Packet Loss on 3 Toll Quality Speech Coders, in: Second IEE National Conference on Telecommunications, York, UK, 1989, pp. 380–385.

[21] D. Kirby, K. Warren, K. Watanabe, Report on the Formal Subjective Listening Tests of MPEG-2 NBC multichannel audio coding, in: ISO/IEC JTC1/SC29/WG11/N1419, 1996.

[22] N. Kitawaki, K. Itoh, Pure Delay Effects on Speech Quality in Telecommunications, IEEE Journal on selected Areas in Communications 9 (4) (1991) 586–593.

[23] ITU-T Recommendation G.114, One-way Transmission Time (2003).
URL `http://www.itu.int/`

[24] J. James, C. Bing, L. Garrison, Implementing VoIP: A Voice Transmission Performance Progress Report, IEEE Communications Magazine 42 (7) (2004) 36–41.

[25] F. Hammer, Quality Aspects of Packet-Based Interactive Speech Communication, Ph.D. thesis, Graz University of Technology, Austria (2006).

[26] V. Guégin, V. Gautier-Turbin, V. Barriac, L. B.-J. R., , L. Gros, G. Faucon, Study of the Relationship Between Subjective Conversational Quality and Talking, Listening and Interaction Qualities: Towards and Objective Model for the Conversational Quality, in: Fourth International Conference on Measurement of Audio and Video Quality in Networks, MESAQIN '05, 2005.

[27] J. Holub, O. Tomiska, Non-monotonicity in Perceived Quality of Delayed Talker Echo, in: Sixth International Conference on Measurement of Audio and Video Quality in Networks, MESAQIN '07, 2007.

[28] W. Jiang, H. Schulzrinne, Modeling of Packet Loss and Delay and their Effect on Real-Time Multimedia Service Quality, in: Proceedings of NOSSDAV, 2000.

[29] A. P. Markopoulou, F. A. Tobagi, M. J. Karam, Assessing the quality of voice communications over internet backbones, IEEE/ACM Trans. Netw. 11 (5) (2003) 747–760.

[30] A. Takahashi, H. Yoshino, N. Kitawaki, Perceptual QoS Assessment Technologies for VoIP, IEEE Communications Magazine 42 (7) (2004) 28–34.

[31] A. W. Rix, Perceptual Speech Quality Assessment - A Review, in: Proceedings of ICASSP'04, Vol. III, 2004, pp. 1056–1059.

[32] A. Takahashi, Opnion Model for Estimating Conversational Quality of VoIP, in: Proceedings of ICASSP'04, Vol. III, 2004, pp. 1072–1075.

[33] L. Gros, N. Chateau, V. Durin, Speech Quality: Beyond the MOS Score, in: Fifth International Conference on Measurement of Audio and Video Quality in Networks, MESAQIN '06, 2006.

[34] L. Gros, N. Chateau, V. Durin, Paradigms for Evaluation of Speech Quality's Impact on Users' Behaviour, in: Sixth International Conference on Measurement of Audio and Video Quality in Networks, MESAQIN '07, 2007.

[35] K. Singh, J. Orozco, D. Ros, G. Rubino, Streaming of H.264 Video over HSDPA: Impact of MAC–Layer Schedulers on User–Perceived Quality, Tech. Rep. RR–2007002–RSM, ENST Bretagne (Apr. 2007).

[36] J. Orozco, Quality of Service Management of Multimedia Flows Over DiffServ IP Networks, Ph.D. thesis, INRIA/IRISA, univ. Rennes I, Rennes, France (Mar. 2005).

[37] ITU-R Recommendation BT.500-10, Methodology for the Subjective Assessment of the quality of Television Pictures, in: International Telecommunication Union, 2000.
URL `http://www.itu.int/`

[38] G. Rubino, P. Tirilly, M. Varela, Evaluating Users' Satisfaction in Packet Networks Using Random Neural Networks, in: Proceedings of the XVI International Conference on Artificial Neural Networks, ICANN'06, Athens, Greece, 2006.

[39] G. Rubino, Quantifying the Quality of Audio and Video Transmissions over the Internet: the PSQA Approach, *Design and Operations of Communication Networks: A Review of Wired and Wireless Modelling and Management Challenges* – Imperial College Press, 2005.

[40] J.-C. Bolot, H. Crépin, Analysis and Control of Audio Packet Loss over Packet-Switched Networks, in: LNCS 1018 - Network and Operating System Support for Digital Audio and Video - Fifth International Workshop, NOSSDAV'95, 1995, pp. 163–174.

[41] H. Sanneck, G. Carle, R. Koodli, A Framework Model for Packet Loss Metrics Based on Loss Runlengths, in: Proceedings of the SPIA/ACM SIGMM Multimedia Computing and Networking Conference, San Jose, CA, 2000, pp. 177–187.

[42] K. Salamatian, S. Vaton, Hidden Markov modeling for network communication channels, in: Proceedings of the ACM SIGMETRICS, 2001, pp. 92–101.

[43] F. S. Filho, E. de Souza e Silva, Modeling the short-term dynamics of packet losses, (to appear) Performance Evaluation Review.

[44] J.-C. Bolot, S. Fosse-Parisis, D. Towsley, Adaptive FEC–Based Error Control for Internet Telephony, in: Proceedings of INFOCOM '99, New York, NY, USA, 1999, pp. 1453–1460.

[45] E. Gilbert, Capacity of a Burst–loss Channel, Bell Systems Technical Journal 5 (39).

[46] S.Hemminger, Netem website, http://developer.osdl.org/shemminger/netem/.

[47] E. de Souza e Silva, R. M. M. L. ao, The Tangram-II Environment, in: LNCS - Computer Peformance Evaluation - Modelling Techniques and Tools, Vol. 1786, Springer, 2000, pp. 366–369.

[48] D. R. Figueiredo, E. de Souza e Silva, Efficient Mechanisms for Recovering Voice Packets in the Internet, in: Globecom'99, Vol. 3, 1999, pp. 1830–1837.

[49] J.-M. Valin, Speex website, http://www.speex.org.

[50] Linphone: La télephonie sous linux, http://www.linphone.org.

[51] Ekiga (gnomemeeting), http://www.gnomemeeting.org.

[52] Asterisk: The open source pbx, http://www.asterisk.org.

[53] LAND/COPPE/UFRJ, Vivavoz website, http://www.land.ufrj.br.

[54] F. Hammer, P. Reichl, Hot discussions and Frosty Dialogues: Towards a Temperature Metric for Conversational Interactivity, in: 8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004), Jeju Island, Korea, 2004.