A systematic study of PESQ's behavior (from a networking perspective)

Martín Varela[*], Ian Marsh and Björn Grönvall
Swedish Institute of Computer Science (SICS)
Kista, Sweden
{mvarela,ianm,bg}@sics.se

June 2, 2006

Abstract

In this paper we study, in a systematic way, how the behavior of PESQ estimations varies with the network loss process. We assess the variability of the estimations with respect to the network conditions and the speech content, and also their accuracy, by comparing the estimates with subjective assessments.

1 Introduction

PESQ [9], the ITU-T's Perceptual Evaluation of Speech Quality is among the most widely used objective voice assessment tools in telecommunications and IP networks. Several commercial offerings incorporate it as a central component for voice over IP (VoIP) quality assessment. In terms of accuracy, i.e. correlation with subjective assessments, it has an advantage over other purely objective quality metrics [12]. While it does perform very well for traditional telephony applications, it has been noted that its performance may decrease when used on VoIP scenarios [11, 12], which present bursty losses.

In this paper we take a systematic, black–box approach to analyzing the performance of PESQ, from a networking perspective. We focus on the impact of the packet loss process. If need be, other network parameters (e.g. delay and jitter) can be considered, and the analysis re-conducted with relatively low effort. However, as far as the one–way voice quality itself is concerned, we consider that the dominant network factor will be the loss process, since jitter can mostly be accounted for as lost packets.

For our experiments, we considered G.711 streams with and without packet loss concealment (PLC). Again, other coding options could be analyzed with relatively low effort. To this end, we have created a basic testing framework which helps prepare and carry out the tests, both objective and subjective. This framework, along with the data obtained from our experiments will be available on-line soon.

Our goals with this work are two–fold. Firstly, we are interested in assessing the performance of PESQ on different VoIP settings, both wired and wireless. Secondly, we are also interested in the development of the testing framework mentioned above.

The results presented herein stem from several series of experiments we have carried out. We have studied the performance of PESQ under a variety of both uniform an Gilbert loss patterns. For the latter case, we have also conducted subjective assessments in order to derive an idea of PESQ's performance

[*] M. Varela's work was carried out during the tenure of an Ercim fellowship.

over a large loss space. While these results are a first approach, they offer insight into what kind of performance we can expect under different scenarios. We have studied how PESQ's results compare to those obtained with the ITU's P.563 single–sided metric [7].

The rest of the paper is organized as follows. Section 2 presents a description of the experiments we carried out. The results we obtained are discussed in Section 3. Finally, we conclude the paper and discuss future work in Section 4.

2 Description of the experiments.

As mentioned above, we have focused our experiments on the behavior of PESQ under different loss processes that can be found on wired and wireless Internet connections. The reasons for concentrating on losses, while ignoring delay aspects are basically two. Concerning the end–to–end delay itself (assuming it constant), it is related to conversational quality, but it does not affect the voice quality itself. Of course, the jitter can have some effect on the perceived voice quality, since packets can arrive at the receiver after their play-out time has expired. As this packets are discarded, and we are not considering interactivity, they can be counted as lost, and thus the effect of jitter can be mapped to packet loss. This is a simplification, as the packet loss models used do not take this into account, but we believe it to be an acceptable one for our purposes.

The experiments we conducted can be classified, according to the scenarios considered, as follows.

1. Uniform losses.
2. Gilbert losses, large loss space.
3. Gilbert losses, restricted loss space

2.1 Uniform losses.

The first loss model we used for our study is that of uniform loss distribution. While this is a very simplistic model, since it assumes no temporal correlation between consecutive losses, it can be (and sometimes is [6, 10]) used to model IP networks' behavior.

We performed several tests using uniform loss patterns. The first application was to see the evolution of PESQ scores and their variability as the loss rate increased. To this end, we assessed ten different samples, each with ten different loss patterns for each loss rate considered. We then calculated the average of the 100 PESQ scores obtained, as well as their variability and bounds.

The uniform loss model was also used to study the variations of PESQ scores observed when a given loss pattern occurs in different "positions" within the voice streams. This, along with the previous tests is useful in predicting how much variability is observable in the quality estimate, given a set of working conditions.

2.2 Gilbert losses, large loss space.

The second loss model used was a simplified version of the Gilbert model [4]. This simplified version is widely used in the literature [14, 1, 15, 2], since it provides an accurate, yet simple way of obtaining loss patterns like those found on the Internet.

2.2.1 On the Gilbert model.

In this model the channel has two states (cf Figure 1), one in which the transmission is perfect and another one in which errors occur. The states 0 and 1 represent packet arrival and loss respectively. We denote by p the probability of a packet being lost given that the previous one arrived. The probability $1-q$ is that of losing a packet given that the previous one was also lost.

The relationship between the parameters in the model and the ones we use in this paper, the loss rate (LR) and the mean loss burst size (MLBS) is as

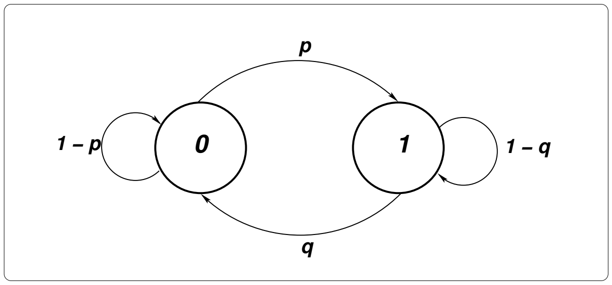

Figure 1: The simplified Gilbert model. When in state 0, the transmission is error–free. In state 1, all packets are lost. Note that the transition from state 0 to state 1 implies a loss, and that in the opposite direction, it implies that the packet is not lost.

follows:

$$p = \frac{1}{\text{MLBS}} \frac{\text{LR}}{1 - \text{LR}}, \qquad q = \frac{1}{\text{MLBS}}. \tag{1}$$

Note that if there are losses (at least one) and if not every transmission is a loss, then MLBS > 1 and $0 < \text{LR} < 1$, leading to $0 < p, q < 1$.

One problem we found when using the Gilbert model to generate loss patterns to be used with standard 8–second speech samples (which correspond to 400 20–ms packets) is that the number of packets needed is too low for the model to be stable. This generally induces a difference between the target values of LR and MLBS, and the actual values obtained in the loss patterns generated. This, in turn, adds some “noise” to the experiment. We dealt with this issue when working on the restricted loss space described below.

2.2.2 On the experiments.

We considered a very large loss space, with loss rates ranging from 0 to 50%, and with mean loss burst sizes ranging from 1 to 10 packets (using 16 intermediate MLBS values). We consider that this loss space covers (and exceeds) all possible loss conditions that can be found when doing VoIP. Moreover, it allows to consider very extreme scenarios, which are unlikely to be found in common usage. However, it also allowed us to push PESQ, and see how it performed under those unusual conditions.

The use of this wide range of combinations allowed us to consider loss patterns commonly found in both wired and wireless networks. In the latter, it is relatively common to experience very bursty losses, even for relatively low loss rates. One downside to using this large space is that some of the combinations are not actually feasible when using 400–packet samples. For those cases, we fed the model with the target LR and MLBS values, and used the resulting strings as they were, accepting that the results in those areas would be noisier than the rest.

For each point of the loss space (816 in total), we generated 10 different patterns, and then processed 20 speech samples through each of them, both with and without PLC. This gave us 400 degraded samples, for which we then calculated PESQ scores. This run implied 426000 PESQ runs, which were timed at about 2 seconds each (using the reference PESQ implementation), which is equivalent to about 180 hours of computing time on a Pentium IV with 1GB RAM.

2.3 Gilbert losses, restricted loss space.

As mentioned previously, using the Gilbert model presents some problems with the large loss space and with the 400–packet speech samples. In order to improve the accuracy of our results, a possible solution would be to use longer speech samples, so that the Gilbert model implementation can converge to the target values. We performed tests to determine how long the samples should be in order for the loss model to converge. The results obtained indicate that between 3000 and 4000 packets would allow for good convergence. This, however, implies very long samples, which would exceed the sample length recommended for PESQ [9].

Therefore, in order to use the standard 8–second

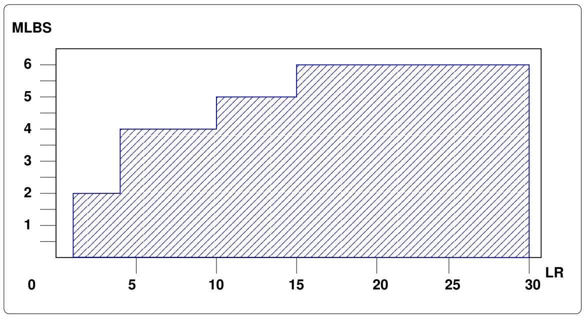

Figure 2: The restricted loss space considered. Note that some combinations which are relatively common on wireless networks, like low loss rate and high burstiness, had to be dropped in order to obtain more accurate loss patterns.

samples and improve the accuracy of our measurements, we had to

1. do away with the unfeasible LR and MLBS combinations, and
2. find a way to obtain 400–packet loss patterns which are accurate with respect to the target LR and MLBS values.

In order to eliminate the unfeasible loss conditions, we simply restricted the loss space, so that all LR and MLBS combinations would be feasible (cf Figure 2). We also reduced the maximum loss rate and mean loss burst sizes to 30% and 6 packets respectively. As for the accuracy problem in the generated loss patterns, we needed to obtain several different ones for each point in the loss space. In order to do this, we chose from a vast pool of seeds for the random number generator, created traces, and kept only those which were accurate enough. Unfortunately, given the random nature of the model, this implied a brute–force approach.

We also run preliminary tests to determine whether variations on the speech samples induced more variability on PESQ results than variations on the loss pattern, or vice–versa. The results obtained indicate that the both parameters imply similar variability in the PESQ scores. We therefore used equal numbers of speech samples and loss patterns (15 each) for the last batch of experiments. Considering both PLC and non–PLC codings, we ended up 450 PESQ scores for each (LR, MLBS) point in the loss space.

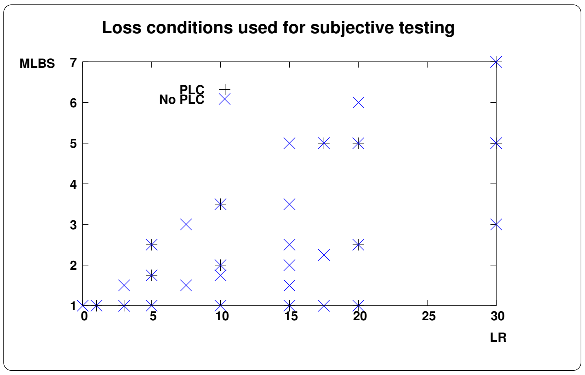

Figure 3: Distribution of the loss conditions considered for the subjective tests.

2.4 Subjective assessment.

In order to determine how the accuracy of PESQ's assessments varies with the network conditions, it is necessary to compare them with subjective assessments. We have, to this end, carried out an ITU P.800–based [8] subjective assessment campaign. This campaign, while small in scale, provides a good view of the relation between subjective scores and PESQ estimates over the loss space considered, and in some adjacent points.

We had 42 4–sample groups assessed, providing a good coverage of the restricted loss space described above. Of those 42 groups, 29 corresponded to samples without PLC, and the remaining 13 did use PLC. Figure 3 shows the loss configurations considered during the test.

We had 11 subjects assess the 168 speech samples, preceded by a series of warm–up samples, which included original–quality (i.e. non–degraded) ones.

The samples and groups were randomly named and the groups were randomly sorted, so as to avoid any bias during the tests. The tests were driven by a script, and the subjects wrote down their assessments in paper forms. The grading scale used was a 9–point one, and the results were later mapped into a 5–point scale for comparison with PESQ's output. Test times varied between about 30 and 45 minutes, and the test instructions suggested a mid–test rest of 5 to 10 minutes. The scores obtained were then statistically screened, and none of the subjects had to be dropped.

3 Experimental results.

In this Section we summarize and explain the main results we obtained from each of the experimental setups described above.

3.1 Results for the uniform loss scenarios.

Using a uniform loss model gave us data to analyze the PESQ results in only one dimension, namely the loss rate. The results obtained (cf Figure 4) seem to indicate that PESQ is over-estimating the perceived quality of the samples, especially for the higher loss rates (e.g. it predicts toll quality at 17% losses in the PLC case). This was also observed later on when analyzing the data from the subjective assessment campaign and the results given by PESQ (cf Section 3.4). These results can be improved by using PESQ-LQ [13].

Seeing how much the variability in the results increases with the loss rate can help decide under which conditions the use of PESQ is appropriate for a given application.

We also studied the variation of PESQ scores as the same loss pattern was shifted in time with respect to the speech sample, as well as the variations due to having different loss patterns with the same loss rate degrade a given sample. The maximum variations we found in these cases were in the order

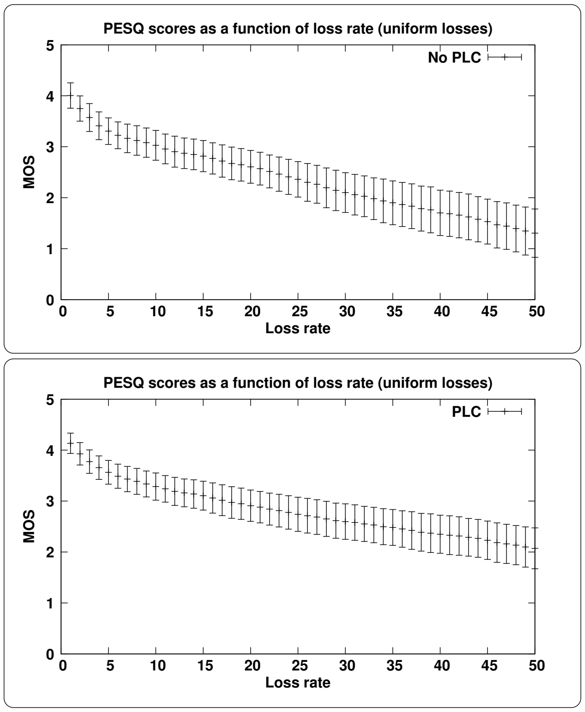

Figure 4: PESQ scores as a function of the loss rate using a uniform loss model. Note that the estimations remain quite high even for very high loss rates. Also, the variability in the estimations is slightly higher when PLC is not used, although in both cases is relatively small.

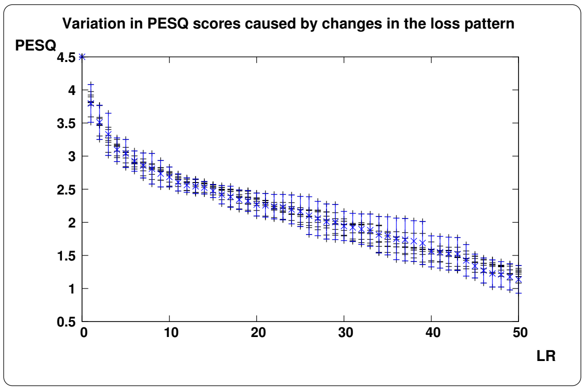

Figure 5: Example of the variation of PESQ scores for 10 different loss patterns applied to the same original sample. Note that the variations are generally small, with maximum variations of about 0.7 MOS points. Although it is not feasible to study all the possible patterns, all the results we obtained were consistent, and similar to this one.

of 0.7 MOS points. Interestingly, this sometimes happened within just a 10–packet (200ms) shift in the loss pattern. Most of the time, however, the scores were very similar, irrespective of the changes to the loss pattern. Figure 5 shows how the PESQ scores vary for 10 different loss patterns applied to the same original sample.

3.2 Results for the large Gilbert loss space.

In this section we discuss the main findings from the experiments run on the large Gilbert loss space. Figures 6 and 7 show the median PESQ scores calculated over the whole loss space, with and without PLC respectively. We can observe how the quality drops, as expected, with both the LR and the MLBS. Also, it is clear that the while the LR is the dominant parameter, a bursty loss process can seriously impair the quality as per PESQ assessments. The use of PLC allows for a smoother quality degradation in both dimensions, which is especially noticeable at low loss rates. In the non–PLC case, the drop in quality over the first 10 to 20% LR is noticeably more steep than when PLC is used.

Also interesting is the fact that the quality decreases more steeply when the LR values are low, and then the degradation is less pronounced.

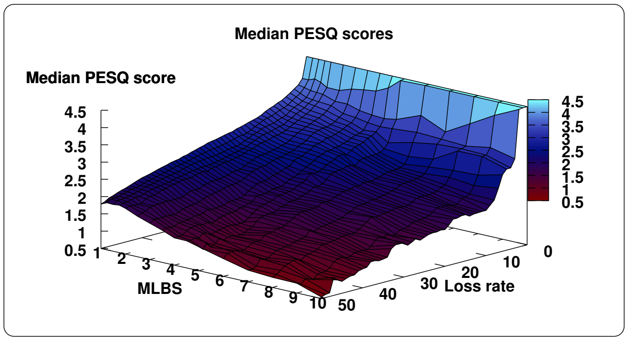

Figure 6: Median PESQ scores over the complete loss space considered, with PLC. The median was calculated over 200 PESQ scores for each (LR,MLBS) combination.

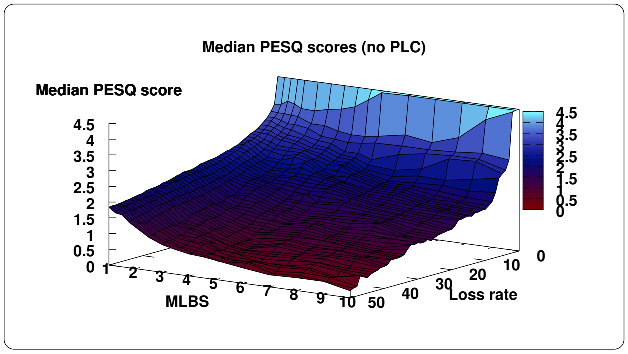

Figure 7: Median PESQ scores over the complete loss space considered, without PLC. The median was calculated as in the PLC case. Note the steeper descent of the quality as the loss rate increases when no PLC is used.

3.3 Results for the restricted Gilbert loss space.

As mentioned in Section 2.3, covering the whole loss space implies a certain decrease in the accuracy of the results obtained. To remedy this, we have studied a more restricted loss space, and increased the accuracy of the Gilbert model's output. The results obtained present a more accurate view of PESQ's behavior as the network conditions change. An interesting first result, is that the overall variability in the estimations is significantly reduced.

In Figure 8 we can compare the absolute deviations of the estimations over both the large and the restricted loss space. The accuracy of the estimations is much more even for the latter case, especially when network conditions degrade.

Figures 9 and 10 show plots of median PESQ scores as a function of LR for two values of MLBS. Interestingly, it would seem that not using PLC induces a greater variability in the results. We still do not know the reason for this. However, the variability is small in most cases. This hints that the median can be a relatively good approximation for the PESQ scores of the 225 samples considered for each point. We've also calculated interquartile ranges, and found them to be small too. These results, along with the ones described below, are currently being used in the development of a single–sided, loss–based quality estimation metric, which approximates PESQ's behavior.

3.4 Comparison with subjective scores.

Although the subjective campaign we carried out was relatively small, it does provide some insight on the actual accuracy of the PESQ assessments as the loss conditions vary. Figure 11 shows the MOS value obtained for each sample, along with their standard deviation.

The overall correlation of PESQ and subjective scores was 0.867, which is a similar value to the one reported in [12]. The scatter plot in Figure 12. This plot suggests that the performance of PESQ, in terms of correlation with subjective scores, remains relatively stable even when the network conditions degrade. Correlation coefficients for each subset used in the figure were of 0.751 and 0.733, respectively.

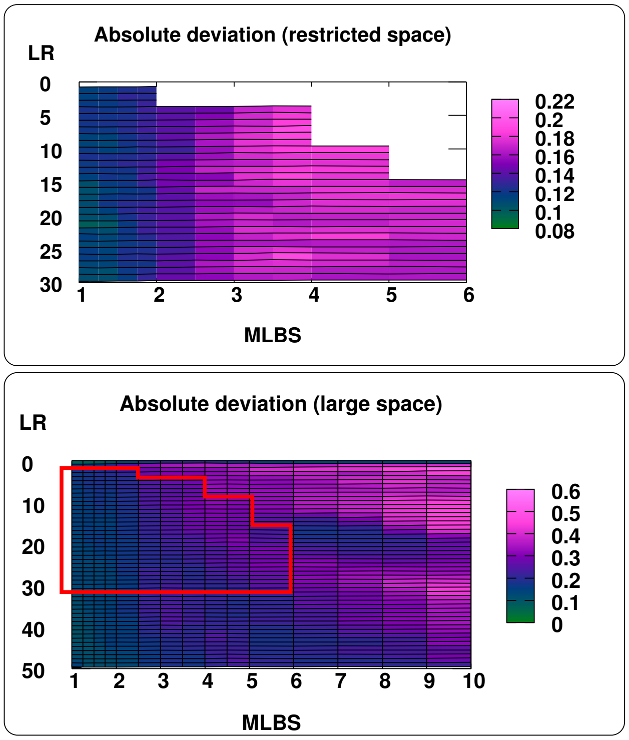

Figure 8: Absolute deviation of PESQ scores at each point of the loss space. The red outline in the large space indicates the restricted space. Note how the variability of the results has decreased.

When comparing the actual estimates, it is easy to see that, even as the correlation remains relatively high, there are variations in its behavior with respect to the subjective scores. In Figure 13 we can see that PESQ is over–estimating the quality when the loss conditions are light. As the losses become more bursty, PESQ's estimations drop faster than the actual MOS, so it under–estimates for the

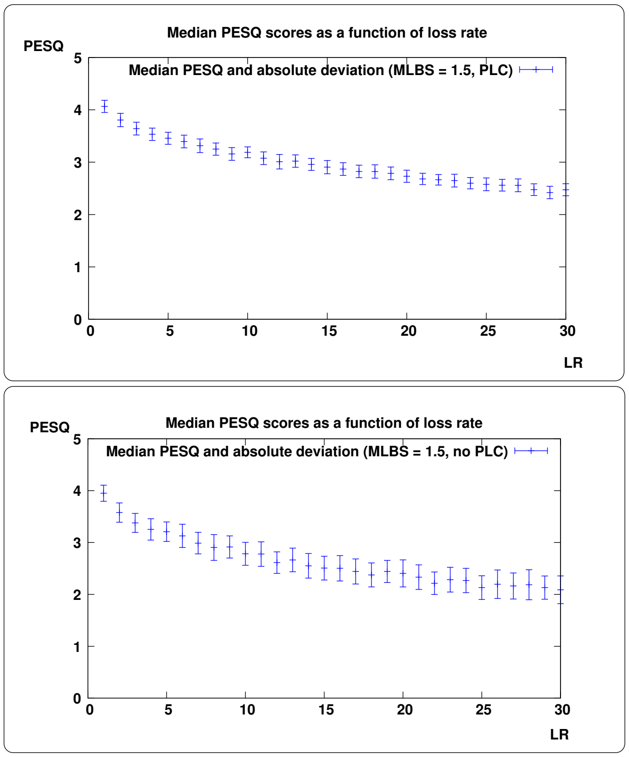

Figure 9: Median PESQ scores and absolute deviation as a function of loss rate. MLBS = 1.5 packets, and both PLC and non PLC cases are shown.

high–burstiness regions of the loss space. The best estimations correspond to moderately bursty losses, which is a good thing, since these are among the most commonly found.

We expect that with further subjective data, it should be possible to derive compensation mechanisms, similar to PESQ-LQ, but based on the network conditions, so that more accurate estimations could be obtained.

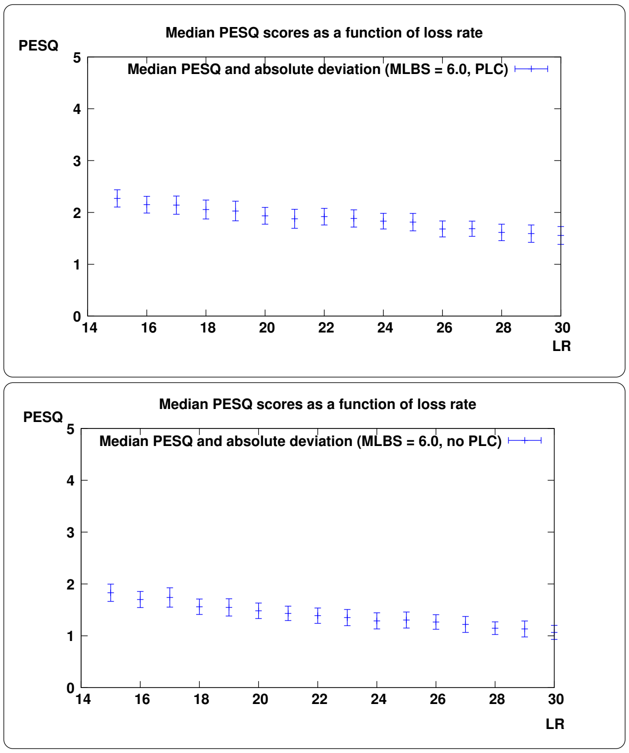

Figure 10: Median PESQ scores and absolute deviation as a function of loss rate. MLBS = 6.0 packets, and both PLC and non PLC cases are shown.

3.5 An informal performance comparison of PESQ and P.563.

We performed an informal comparison on the performances of PESQ and the P.563 single–sided assessment technique, in order to obtain an idea of how good the P.563 estimations were. We believe that, although both metrics are conceived to work under different conditions, the comparison remains interesting.

The P.563 estimations of the degraded samples used for the subjective campaign present a behavior quite different to that of PESQ's. The single–sided metric under–estimated under light loss conditions, and gradually approached the MOS values as the

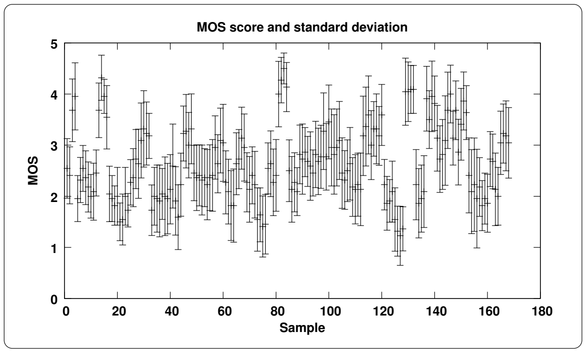

Figure 11: MOS values and their respective standard deviations for all the samples tested. (Note that the samples are not ordered in this plot.)

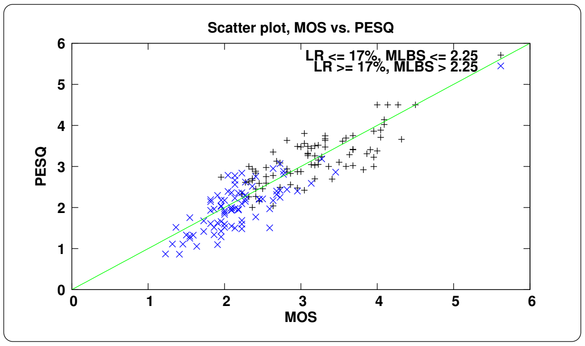

Figure 12: Scatter plot: PESQ scores vs MOS values.

loss rate and burstiness increased (slightly over–estimating for very bursty losses). This can be seen in Figure 14.

In terms of correlation with the subjective scores, P.563 did not provide results as good as PESQ's. The overall correlation was of 0.795. While this is not a bad result, it seems to be a bit close to the performance of the E–model [5, 3], at a much higher computational cost. Using the reference implementation, each 8–second sample took about 3 seconds to assess on a Pentium IV with 1GB RAM. Then again, these results are not conclusive, and are intended only as a preliminary assessment.

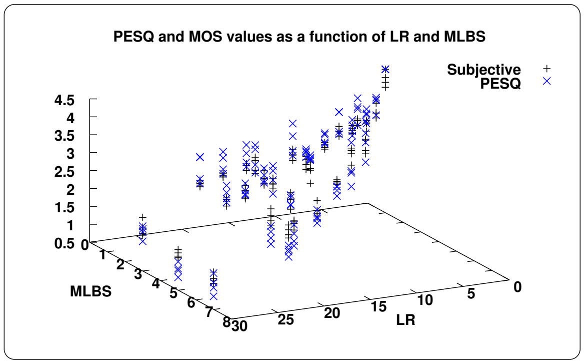

Figure 13: PESQ scores and MOS as a function of the loss rate and the mean loss burst size. We can see that PESQ over–estimates the quality when the burstiness is low, and under–estimates it when the losses are bursty. The best estimations are those corresponding to moderately bursty losses.

4 Conclusions and future work.

In this paper we have presented a systematic study of the behavior of PESQ as the network loss conditions vary. The main goals of this study are to gain a better understanding of the circumstances under which PESQ is able to provide accurate assessments, and to also understand which kind of adjustments need to be made when the accuracy degrades.

We have analyzed the variability of PESQ scores under several different conditions, and found it to be relatively small, which opens the door for performing PESQ–like, single–sided estimations of the quality of a voice stream. We've also analyzed the accuracy of PESQ as the network conditions change, by means of comparison with subjective scores. Although this can be considered work in progress, as more subjective data is needed to provide a more comprehensive view, the results obtained up to date are interesting. In particular, it seems that PESQ maintains a reasonable correlation with subjective

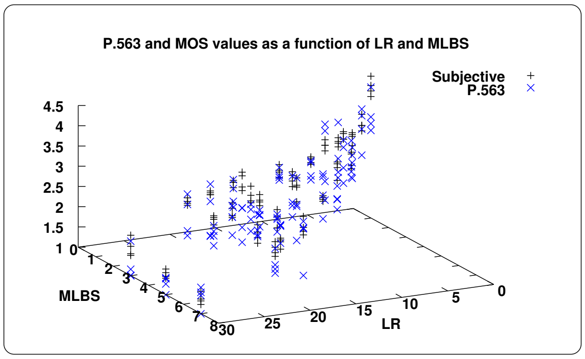

Figure 14: P.563 scores and MOS as a function of the loss rate and the mean loss burst size.

scores even when the network conditions are bad. Also, the deviations it presents from the subjective scores seem systematic, which suggest that a simple compensation factor might be found (for instance, derived from the network conditions) and used to improve the results.

An informal performance comparison has been performed between PESQ and the P.563 single–sided metric, and with the data available, the results indicate that the former provides more accurate quality estimates. A more in–depth study of P.563's performance is part of our research plans.

As for possible research directions in this area, we consider that more subjective assessments similar to the ones presented here would greatly improve our understanding of PESQ, and probably allow for improvements to be made, as mentioned above. We are also working on the development of loss–based single–sided metric based on PESQ, to be used in real–time environments.

References

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

[2] J-C. Bolot, S. Fosse-Parisis, and D.F. Towsley. Adaptive FEC–Based Error Control for Internet Telephony. In *Proceedings of INFOCOM '99*, pages 1453–1460, New York, NY, USA, March 1999.

[3] A. Estepa, R. Estepa, and J. Vozmediano. On the Suitability of the E-Model to VoIP Networks. In *Seventh International Symposium on Computers and Communications, IEEE ISCC 2002*, pages 511–516, Taormina, Italy, July 2002.

[4] E. Gilbert. Capacity of a Burst–loss Channel. *Bell Systems Technical Journal*, 5(39), September 1960.

[5] T. A. Hall. Objective Speech Quality Measures for Internet Telephony. In *Voice over IP (VoIP) Technology, Proceedings of SPIE*, volume 4522, pages 128–136, Denver, CO, USA, August 2001.

[6] D. Hands and M. Wilkins. A Study of the Impact of Network Loss and Burst Size on Video Streaming Quality and Acceptability. In *Interactive Distributed Multimedia Systems and Telecommunication Services Workshop*, October 1999.

[7] ITU-T Recommendation P.563. Single–ended Method for Objective Speech Quality Assessment in Narrow–band Telephony Applications, May 2004.

[8] ITU-T Recommendation P.800. Methods for Subjective Determination of Transmission Quality, August 1996.

[9] 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.

[10] S. Mohamed, F. Cervantes, and H. Afifi. Integrating Networks Measurements and Speech Quality Subjective Scores for Control Purposes. In *Proceedings of IEEE INFOCOM'01*, pages 641–649, Anchorage, AK, USA, April 2001.

[11] S. Pennock. Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm. In *Measurement of Speech and Audio Quality in Networks Line Workshop, MESAQIN '02*, January 2002.

[12] Psytechnics Ltd. PESQ: an Introduction. `http://www.psytechnics.com`, September 2001.

[13] Antony W. Rix. Comparison between subjective listening quality and P.862 PESQ score. In *Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN'03)*, Prague, Czech Republic, May 2003.

[14] H. Sanneck, G. Carle, and 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*, pages 177–187, San Jose, CA, January 2000.

[15] M. Yajnik, S. Moon, J.F. Kurose, and D.F. Towsley. Measurement and Modeling of the Temporal Dependence in Packet Loss. In *Proccedings of IEEE INFOCOM '99*, pages 345–352, 1999.