

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

Martín Varela¹, Ian Marsh¹, and Björn Grönvall¹
mvarela@sics.se, ianm@sics.se, bg@sics.se

Swedish Institute of Computer Science (SICS) Kista, Sweden

Abstract. In this paper we study, in a systematic way, how the behavior of PESQ estimations vary with the network loss process. We assess the variability of these estimations with respect to the network conditions and the speech content. We judge the estimation accuracy with subjective tests and the ITU's single-sided measure.

1 Introduction

PESQ [ITU01], 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 quality assessment. In terms of accuracy, i.e. correlation with subjective assessments, it has an advantage over other purely objective quality metrics [Psy01]. While it does perform very well for traditional telephony applications, it has been noted that its performance decreases when used on VoIP scenarios, which exhibit bursty losses [Pen02,Psy01].

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. However, as far as the voice quality itself is concerned, we consider that the dominant degradation factor will be the network losses. For our experiments, we considered G.711 streams with and without packet loss concealment (PLC). To this end, we have created a basic testing framework which helps prepare and carry out tests, both objective and subjective. Our goals within this work is assessing the performance of PESQ in two different VoIP settings, namely wired and wireless networks.

We have studied the performance of PESQ under a variety of both uniform and bursty losses. For the latter case, we have also conducted subjective assessments in order to derive an idea of PESQ's performance in relation to real user tests. The general idea is pre-generate loss sequences, for various distributions and examine the scores given by PESQ. This treats the processing of PESQ as a black box as explained. We have additionally studied how PESQ's results compare to those obtained with the ITU's P.563 single-sided metric [ITU04]. 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. We used G.711 coding, both with and without PLC. 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 used to model network behavior when the loss rate is relatively low [HW99,MCA01]. We performed several tests using uniform loss sequences. The first was to see the PESQ scores as the loss rate was increased. We assessed ten different samples, each with ten different loss sequences for each loss rate considered. We then calculated the average of the 100 PESQ scores obtained, as well as their variance. The uniform loss model was also used to study the variations of PESQ scores observed when a given loss sequence occurs in different positions within a voice segment. Essentially, this means what is the difference in PESQ scores when certain parts of speech are lost due to packet loss.

2.2 Gilbert losses, large loss space

The second loss model used was a simplified version of the Gilbert model [Gil60]. This simplified version is widely used in the literature [SCK00,BFPT99], since it provides an accurate, yet relative easy method of generating bursty loss sequences.

The Gilbert model: In this model the channel has two states shown in Figure 1), one in which the transmission is successful and another in which errors occur. The states 0 and 1 represent a 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 follows:

$$p = \frac{1}{\text{MLBS}} \frac{\text{LR}}{1 - \text{LR}}, \quad q = \frac{1}{\text{MLBS}}. \quad (1)$$

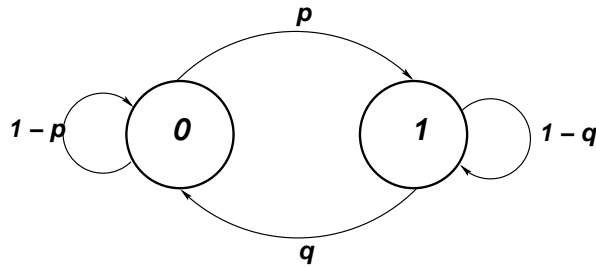


Fig. 1. The simplified Gilbert model. When in state 0, the transmission is error-free. In state 1, a loss occurred. Note that the transition from state 0 to state 1 implies a loss, and that in the opposite direction, it implies that the packet arrived.

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

PESQ standard samples are 8 seconds long. At 20 ms packetisation this corresponds to 400 packets. One problem we found when using the Gilbert model was to generate loss sequences for such a low number of packets. The two state process needs more than 400 packets to reach a steady state distribution. Although we could have generated longer sequences, we decided to keep to the 400 sequence length. This generally induces a difference between the target values of LR and MLBS, and the actual values obtained in the loss strings. This, in turn, adds some variance to the tests. We dealt with this issue when working on the restricted loss space described below.

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. This loss space covers, and probably exceeds, most possible loss conditions that can be found for VoIP traffic. Considering all these combinations allowed us to consider loss sequences 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 space is that some of the combinations are not actually feasible when using 400-packet samples for the reasons stated above.

For each point of the loss space (816 in total), we generated 10 different sequences, and then processed 20 speech samples both with and without PLC. This gave us 400 degraded samples, for which we then calculated PESQ scores. This run implied 426000 PESQ executions, which needs about two seconds per execution. The total time for such an experiment was about 180 hours of computing time, using a Pentium IV with 1GB RAM as reference.

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 (short) 400 packet 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 [ITU01]. Therefore, in order to use the standard 8-second samples and improve the accuracy of our measurements, the next sections will discuss:

- Remove infeasible LR and MLBS combinations
- Obtain more accurate 400 packet loss sequences

In order to eliminate the unfeasible loss conditions, we simply restricted the loss space, so that all LR and MLBS combinations would be feasible, see 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 sequences, we needed to obtain several different sequences for each point in the loss space. In order to do this, we chose from a large pool of seeds for the random number generator, created sequences which were close enough for our purposes. We used a brute-force approach, however it would have been possible to generate such sequences with variation reduction techniques such as antithetic variables for example.

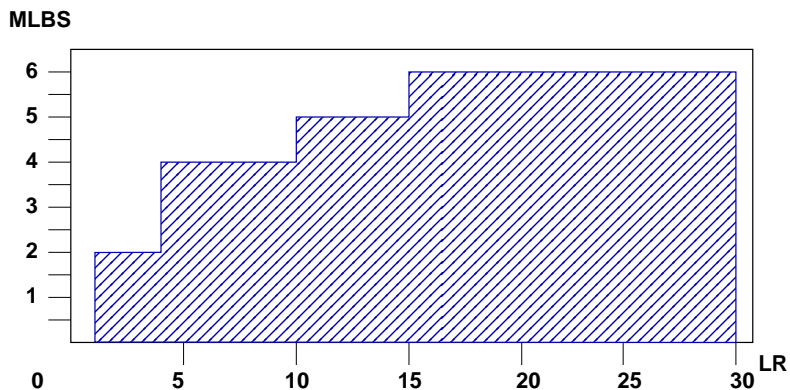


Fig. 2. The restricted loss space considered. Note that some combinations which are relatively common on wireless networks, like low loss rate and high burstiness, were removed to generate more accurate loss sequences.

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 (e.g. for hearing problems) none of the subjects had to be discarded.

3 Experimental results

In this Section we summarise the main results obtained from the experimental descriptions above.

3.1 Results for the uniform loss scenarios

Using a uniform loss model gave us the data to analyse the PESQ in terms of loss rate only. The results obtained Figure 4 indicate that PESQ is over-estimating the perceived quality of the samples, especially for the higher loss rates. This was also observed later when analyzing the data from the subjective assessment tests and the results given by PESQ (see Section 3.4). These results can be improved by using PESQ-LQ [Rix03]. Seeing how much the variability in the results increases with the loss rate can help us decide under which conditions the use of PESQ is appropriate for a given telephony application.

We also studied the variation of PESQ scores as the same loss sequence was shifted in time with respect to the speech sample. We also studied the variance due to having different loss sequences with the same loss rate degrade a given sample. The maximum variations we found in these cases were in the order of 0.7 MOS points, which indicate that they should be noticeable by the average listener. Interestingly, this happened within just a 10-packet (200ms) shift in the loss sequence. Most of the time, however, the scores were very similar, irrespective of the changes to the loss sequence. Figure 5 shows how the PESQ scores vary for 10 different loss sequences 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 falls, as expected, with both increasing LR and MLBS. Also, it is clear that while the LR is the dominant parameter, a bursty loss process can seriously impair the quality from a PESQ perspective. Naturally, the use of PLC allows for a smoother quality degradation for both loss types 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 becomes less pronounced. It also appears that the curve for the median PESQ as a function of LR (for fixed MLBS

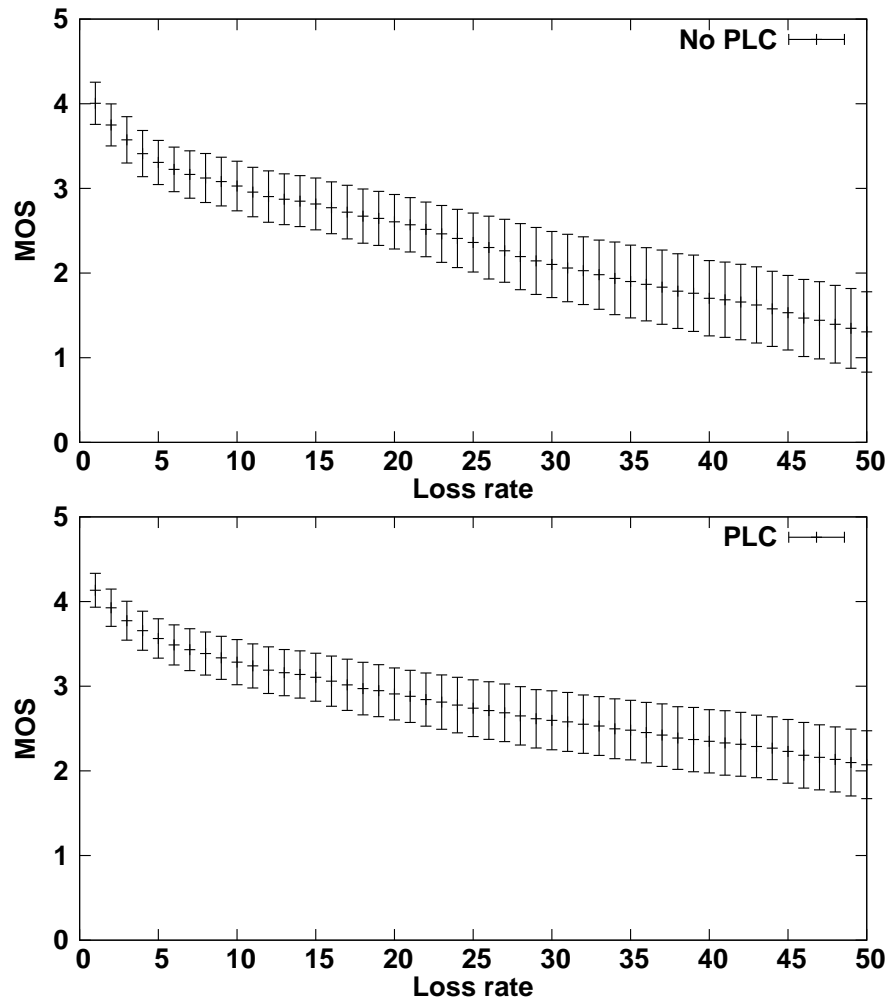


Fig. 4. PESQ scores as a function of the loss rate using a uniform loss model. Note that the estimates remain high even for very high loss rates. Also, the variability in the estimates is slightly higher when PLC is not used, although in both cases is relatively small.

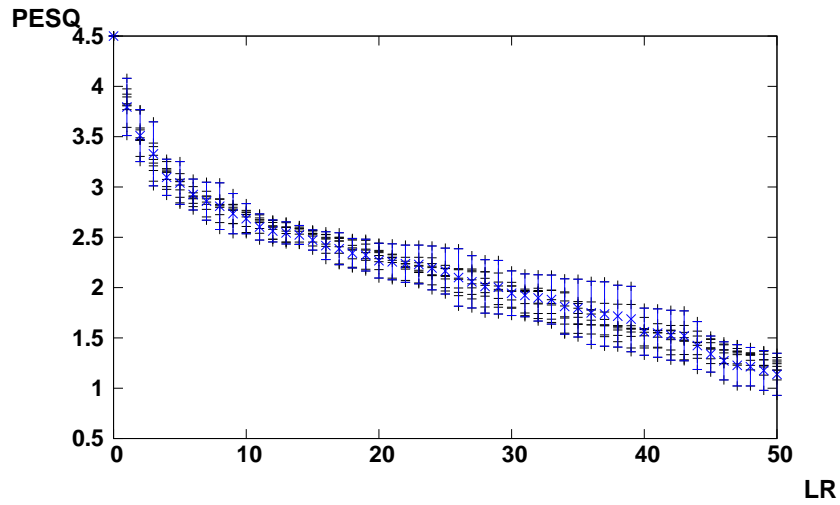


Fig. 5. Example of the variation of PESQ scores for 10 different loss sequences applied to the same original sample. Note that the variations is generally small, however a maximum variations of about 0.7 MOS points can be observed.

values) is composed of two roughly linear segments, as can be seen in Figures 9 and 10.

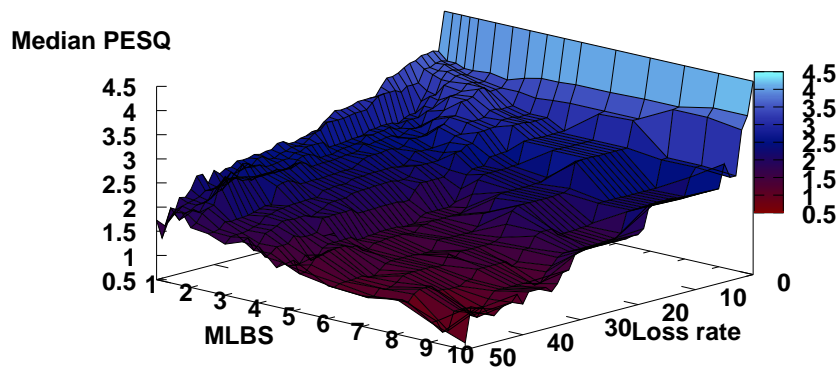


Fig. 6. Median PESQ scores over the complete loss space considered, with PLC. The median was calculated over 200 PESQ scores for each (LR,MLBS) combination.

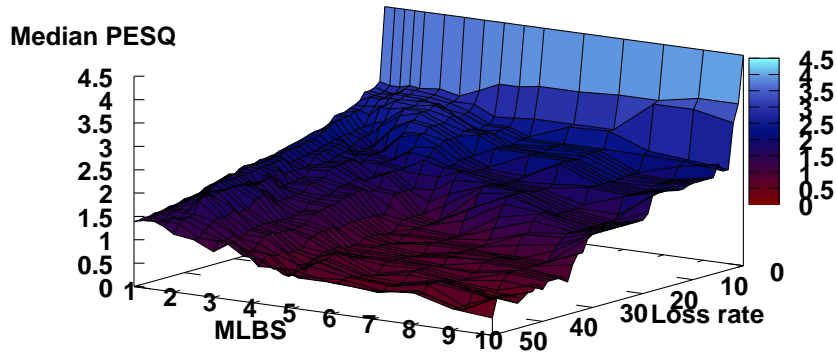


Fig. 7. Median PESQ scores over the complete loss space considered, without PLC. The median was calculated as in the PLC case. Note the steeper descent of the quality as the loss rate increases when no PLC is used.

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, as above especially when network conditions degrade.

Figures 9 and 10 show plots of the median PESQ scores as a function of LR, with the absolute deviation also plotted. Interestingly, it would seem that not using PLC induces a greater variation into the results. We still do not know the reason for this. However, the absolute deviation 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 also found them to be small.

3.4 Comparison with subjective scores

Although the subjective campaign we carried out was relatively small, it does provide some insight into 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 deviations.

The overall correlation of PESQ and subjective scores was 0.867, which is a similar value to the one reported in [Psy01]. The scatter plot in Figure 12 suggests that the performance of PESQ, in terms of correlation with subjective scores,

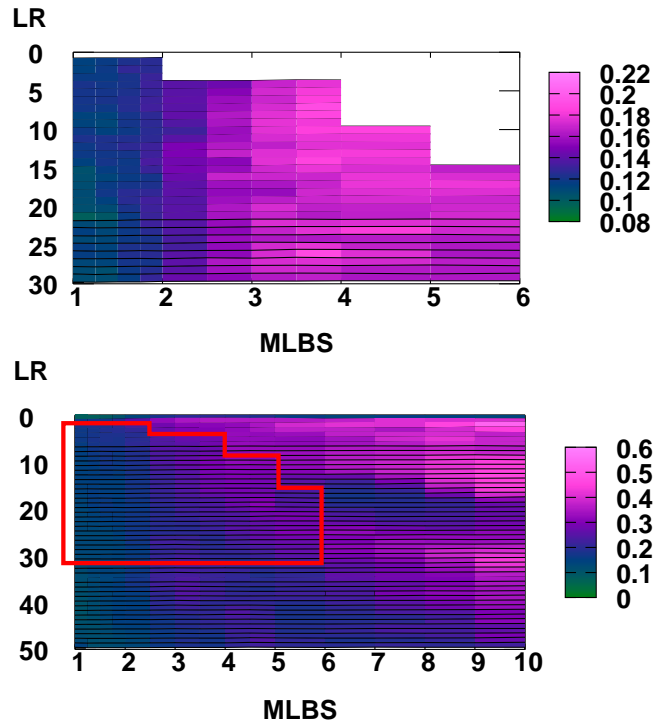


Fig. 8. Absolute deviation of PESQ scores at each point of the loss space. The red outline in the large space indicates the restricted space. Note how the variability of the results has decreased.

remains relatively stable even when the network conditions degrade. Correlation coefficients for each subset were of 0.751 and 0.733, respectively.

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 losses are small. As the losses become more bursty, PESQ's estimations drop faster than the actual MOS, so therefore PESQ underestimates for the highly bursty losses. The best estimations correspond to moderately bursty losses.

3.5 An informal performance comparison of PESQ and P.563

We performed a short comparison on the performances of PESQ and the P.563 single-sided assessment technique, in order to obtain an idea of how reasonable the P.563 estimations were. We believe that, although both metrics work under different conditions, access to the reference signal should provide better estimates of the quality. The P.563 estimations of the degraded samples used

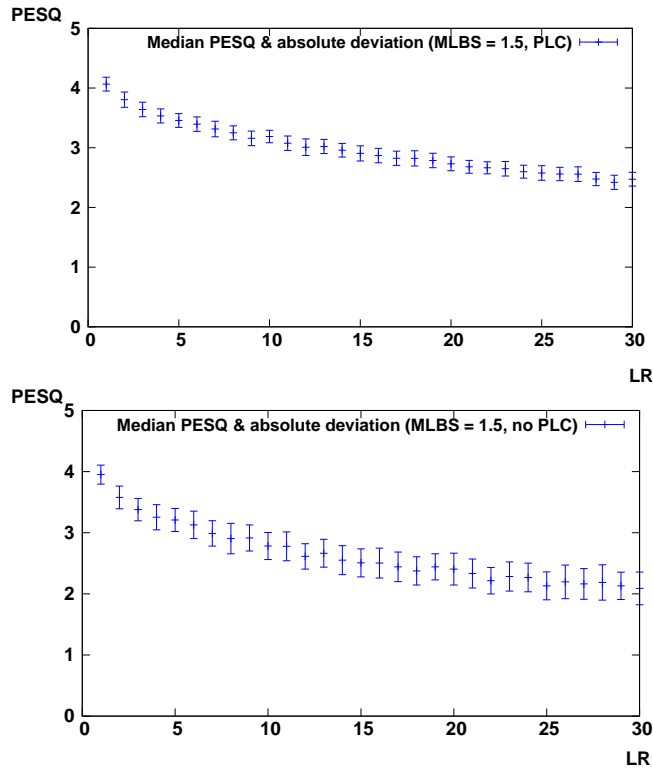


Fig. 9. Median PESQ scores and absolute deviation as a function of loss rate. MLBS = 1.5 packets, and both PLC and non PLC cases are shown.

for the subjective tests presented quite different results to that of PESQ. The single-sided metric underestimated the quality under low losses, and gradually approached the MOS values as the loss rate and burstiness increased (slightly overestimating 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 0.795. While not insignificant it is worth while to use PESQ where possible.

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 what kind of adjustments need to be made when the accuracy degrades.

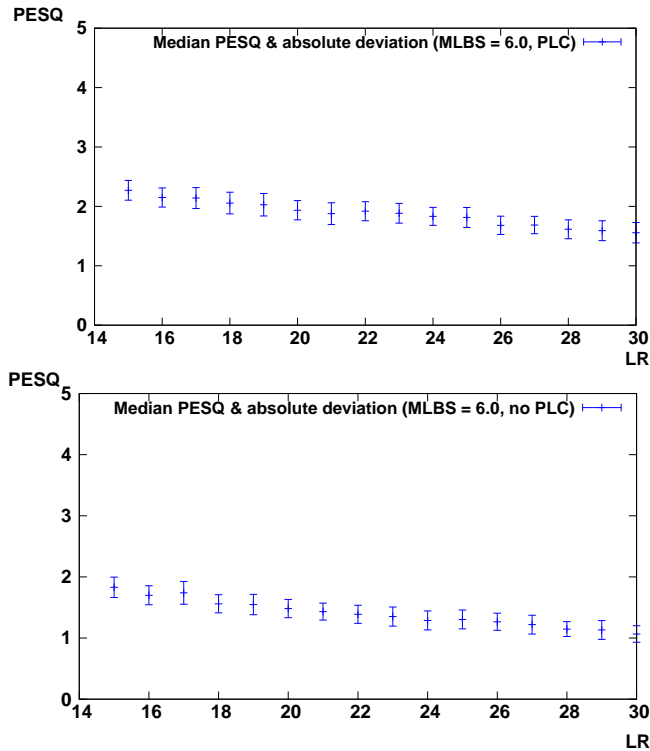


Fig. 10. Median PESQ scores and absolute deviation as a function of loss rate. MLBS = 6.0 packets, and both PLC and non PLC cases are shown.

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 analysed the accuracy of PESQ as the network conditions change, by means of comparison with subjective scores. In particular, it seems that PESQ maintains reasonable correlation with subjective scores even when the network conditions are poor. Also, the deviations it exhibits from the subjective scores seem systematic, which suggests that a simple compensation factor might be found (for instance, derived from the network conditions) and used to further 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 PESQ provides more accurate quality estimates. As stated this seems natural if the signal processing has access to the original samples.

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

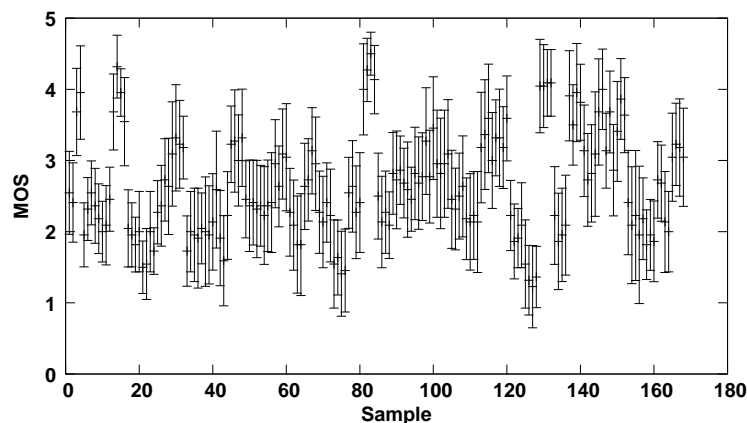


Fig. 11. MOS values and their respective standard deviations for all the samples tested.

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

- [BFPT99] 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.
- [Gil60] E. Gilbert. Capacity of a Burst-loss Channel. *Bell Systems Technical Journal*, 5(39), September 1960.
- [HW99] 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.
- [ITU96] ITU-T Recommendation P.800. Methods for Subjective Determination of Transmission Quality, August 1996.
- [ITU01] 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.
- [ITU04] ITU-T Recommendation P.563. Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications, May 2004.
- [MCA01] 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.
- [Pen02] 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.
- [Psy01] Psytechnics Ltd. PESQ: an Introduction. <http://www.psytechnics.com>, September 2001.

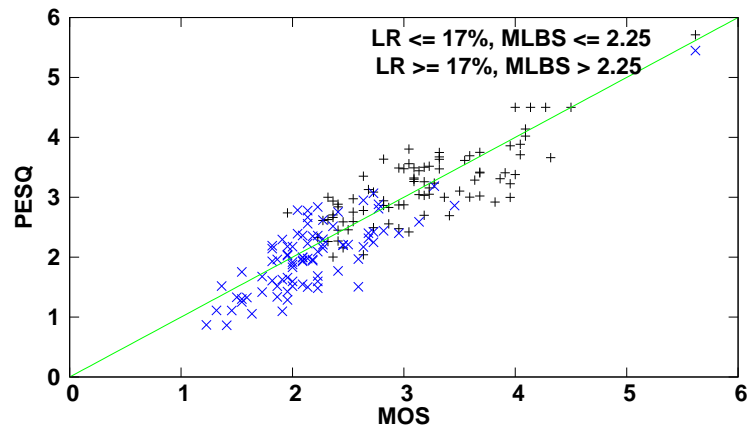


Fig. 12. PESQ scores vs MOS values.

- [Rix03] 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.
- [SCK00] 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.

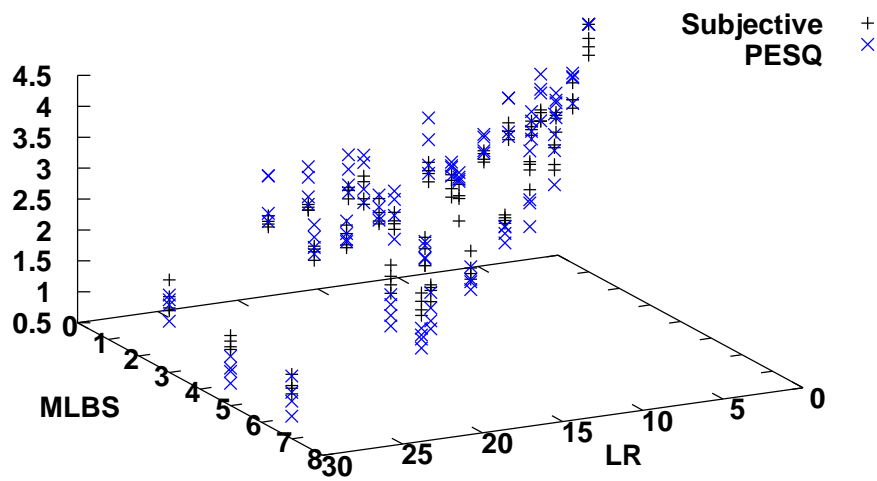


Fig. 13. PESQ scores and MOS as a function of the loss rate and the mean loss burst size. We can see that PESQ overestimates the quality when the burstiness is low, and underestimates it when the losses are bursty. The best estimations are those corresponding to moderately bursty losses.

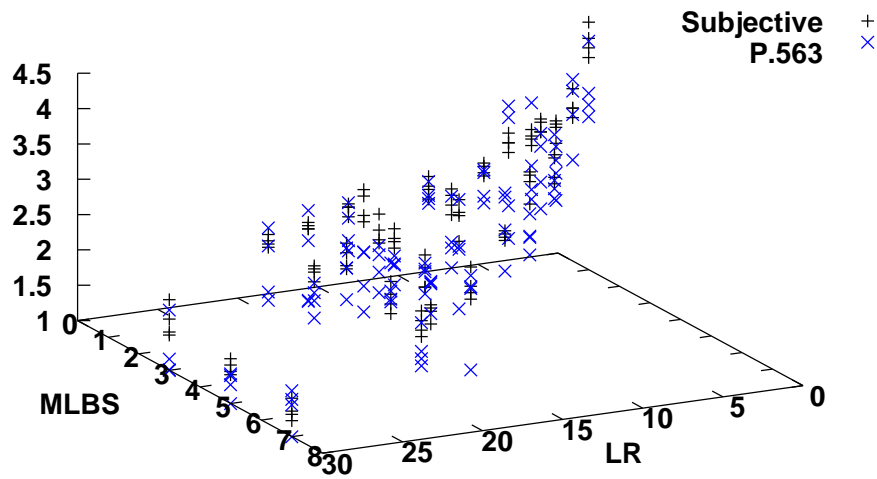


Fig. 14. P.563 scores and MOS as a function of the loss rate and the mean loss burst size.