A metric for single-ended speech quality estimation

Martín Varela *[1]
Ian Marsh*
Björn Grönvall*
Florian Hammer†
* Swedish Institute of Computer Science, Kista, Sweden
† Telecommunications Research Center Vienna (ftw.), Austria
{mvarela,ianm,bg}@sics.se, hammer@ftw.at

***Abstract*— Estimating the quality of VoIP-transmitted speech is the subject of this paper. We present a loss-based technique that allows a single–sided quality approximation of the ITU-T P.862 (PESQ) standard. Our focus is to provide a simple, human-interpretable voice quality metric in real-time at a receiver. Therefore, our quality metric has to offer low computational complexity and need to work without the reference signal needed when using PESQ. The contribution of this work is a real-time PESQ-like quality measure that can be simply implemented in a mobile handset. We verify our solution across a wide range of network conditions and show it to give acceptable estimations of the perceived quality.**

I. Introduction and related work

Packet loss can seriously deteriorate the quality of a conversation within VoIP systems. Humans are well aware of disturbances in the spoken speech caused by losses and either "interpolate" the lost segments, or ask the other person to repeat the last phrase or sentence. From a system perspective, this should be seen as a last resort, it is a user–level request to repeat. It would be preferable if the system could detect poor quality conditions on behalf of the user and initiate preventive measures itself. The goal of this work therefore is for the system to assess the impairment experienced by the human user measured using only the network parameters. A single–sided quality metric can be used in a mobile system to trigger a handover to alternative systems or send reports back to the operator about network conditions.

In this work we estimated the impact of packet loss on perceived speech quality using an ITU standard know as the Perceptual Evaluation of Speech Quality or PESQ for short [1]. Our objective was to find a mapping between network conditions and PESQ scores. We performed these tests for a large number of speech samples and network conditions. With this data the receiver can map the measured loss statistics to our estimated PESQ scores. The paper continues with a brief description of the related work, some background on PESQ and the loss models we considered, the results and some conclusions of this effort.

The ITU-T's P.563 recommendation also is known as a "Single-sided measure" was published in 2004 [2]. It provides signal processing–based, single–sided assessment of narrow–band PSTN voice quality. To the best of our knowledge there is no independent study of its performance, either with respect to subjective scores or relative to other well–known metrics such as PESQ. P.563 is based on purely signal processing techniques. Pseudo–Subjective Quality Assessment (PSQA) is another metric for VoIP quality and is based on neural networks [3]. It can be used with voice, audio and video, and it provides very good correlation with subjective quality scores. PSQA works by learning the relation between parameters which affect the perceived quality of a stream, and hence the quality itself. In order to implement PSQA, the quality–affecting parameters need to be identified and suitable ranges for their values selected. Then, degraded samples are created for several points, called *configurations* in the parameter space, and then assessed by human subjects. Work by Hoene et al. propose a real-time implementation of PESQ known as PESQlite [4]. It reduces the computational complexity of PESQ by using constant length test samples and no time alignment of the degraded samples. It is not clear how the authors deal with not having the reference signal available at the receiver for comparison even given a real-time implementation.

[1]M. Varela's work was carried out at SICS during the first period of an ERCIM fellowship. He is currently with VTT Electronics, in Oulu, Finland.

II. Deriving a PESQ–based single–sided metric

A. PESQ background

Figure 1 shows the functional units of PESQ. A reference speech signal is transmitted through a network. This results in a quality degradation corresponding to the path conditions and chosen coding scheme. PESQ analyzes both the reference and degraded signal and calculates their representation in the perceptual domain based on a psychoacoustic model of the human auditory system. The "difference" between the original and the degraded speech signal is calculated by the quality estimation algorithm and a corresponding MOS score is derived. The evaluation of speech quality using PESQ cannot be performed in real–time since it needs to compare the received signal with the original one which is normally not available. The quality estimation algorithm of PESQ produces values between 1 and 4.5. In all of the experiments we used G.711 coded speech only.

B. Loss models

1) Uniform losses: We began by assessing PESQ's sensitivity to small temporal shifts in a loss pattern. A loss pattern

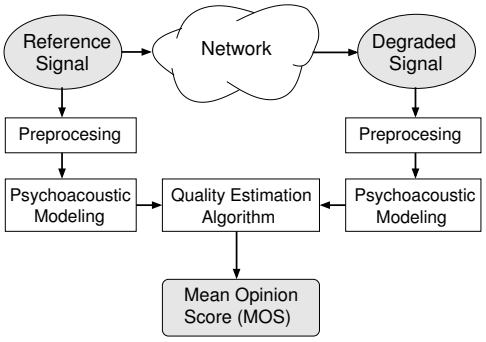

Fig. 1. The PESQ processing structure in block diagram form.

is a series of 1's and 0's representing losses and non-losses respectively. First, we generated loss patterns from a uniform distribution with rates between 1% and 50%. We then applied these loss patterns to standard speech samples from an ITU-T database [5]. The database samples are sequences of 400 packets with a 20ms packetization interval. With the original and degraded loss speech samples, we used PESQ as shown in Figure 1, to assess the effect of each loss pattern. In order to assess the temporal effect of losses on the standard sample, we rotated the loss pattern one packet at a time 400 times, until each position had impacted the standard sample in all 400 possible positions. We observed that in some cases, a relatively small shift in the loss pattern (e.g. ten packets) can produce differences of up to ≈ 0.7 MOS points in the PESQ output (e.g. Figure 2). Interestingly, the maximum variation was also ≈ 0.7 MOS points. It should be noted that the average user can only distinguish quality variations of at least 0.5 points during subjective tests, this means that the variations caused by temporal shifts in the loss patterns would barely be noticeable to most users.

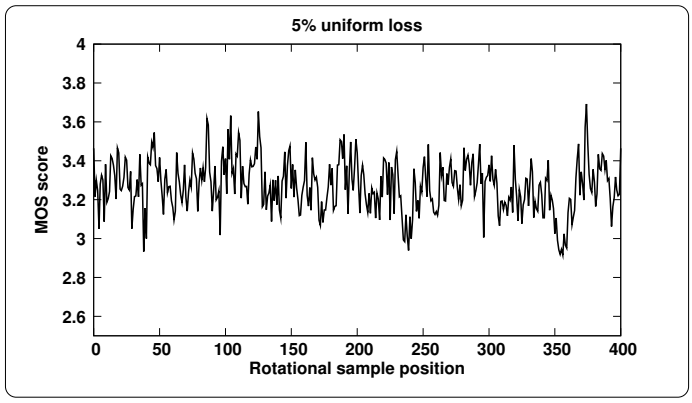

Fig. 2. PESQ scores for 5% uniform losses over a standard sample.

2) Bursty losses: Several models, ranging from loss independence, fixed–size loss bursts and complex models such as. $k^{th}-$order Markov chains have been proposed to model network losses. One of the most widely used is a simplified version of the Gilbert model [6]. In this model the channel has two states, one in which the transmission is loss-free and another in which errors occur. The relationship between the parameters in the Gilbert model (normally known as p for the uncorrelated losses and q as the correlated losses) with the ones we will use in our work is as follows:

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

Where LR is the loss rate and MLBS is the mean loss burst size. Note that if there are losses (at least one) and if not every transmission is a loss, then $\text{MLBS} > 1$ and $0 < \text{LR} < 1$, leading to $0 < p, q < 1$. Similarly, the LR and MLBS in terms of p and q can be written as:

$$\text{LR} = \frac{p}{p+q}, \qquad \text{MLBS} = \frac{1}{q}. \tag{2}$$

So, for example, when using this model, a measured loss rate of 5% with a mean loss burst size of 1.6 packets would translate to $p = 0.0329$ and $q = 0.625$. We chose to use LR and MLBS as they map more intuitively to network conditions.

III. Evaluation technique

We explored the effects of a wide range of losses and the resulting PESQ scores. To obtain sufficient generality in the speech patterns we used 20 different PCM samples, of the same language, taken from the ITU database. These samples were subjected to different loss patterns as follows: We generated losses ranging from 0 to 50% and mean loss burst sizes ranging from one to ten packets. For the loss rates, we chose steps of 1%, and for the mean loss burst sizes we used 16 different values, in the range of one to ten (slightly unevenly distributed). In the whole loss space only some of the loss rate and mean loss burst size combinations make sense in many network scenarios (see Figure 3). For each combination of loss rate and mean loss burst size we used ten different seeds to generate the loss patterns, and we ran each of the 20 samples against each of these ten patterns with PESQ.

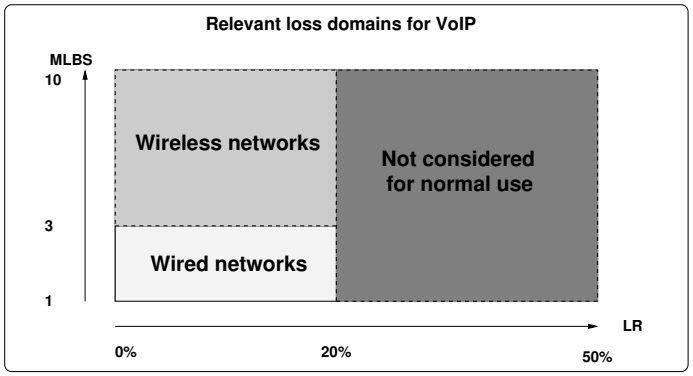

Fig. 3. The loss rate and mean loss burst size values that are most likely to be found in both wired and wireless networks.

IV. Results

In this section we present the main findings from our investigations. For the uniform losses, Figure 4 shows the median and the standard deviation for each of the loss rates

in the uniformly distributed loss case. As stated earlier we considered losses up to 50% although in practice only up to about 20% loss is tolerable with G.711 coded voice (and using packet loss concealment).

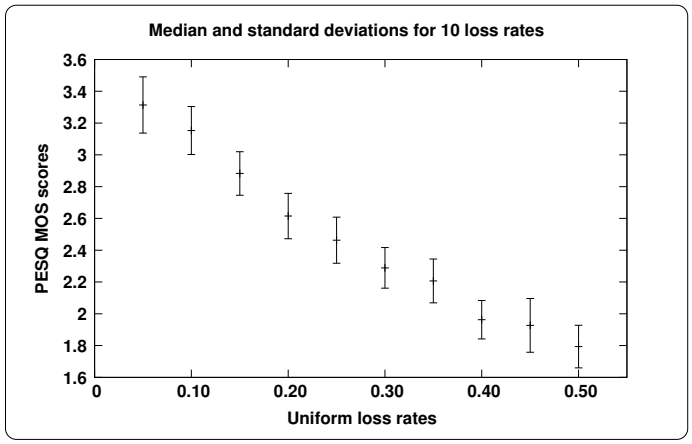

Fig. 4. PESQ scores for uniform loss rates of 5% to 50%.

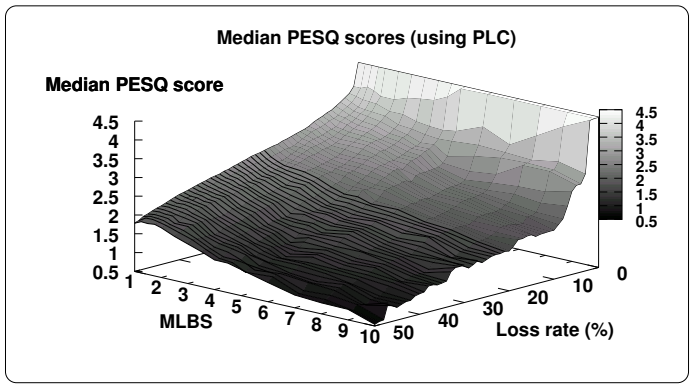

Fig. 5. Median PESQ scores over the complete loss space considered, with PLC. The median was calculated over 200 PESQ scores for each (LR, MLBS) combination.

For the bursty losses Figure 5 shows the median PESQ scores calculated over the whole loss space. Note we have taken the median of the PESQ scores to obtain an estimate of all the scores. We can observe how the quality drops, as expected, with both the LR and the MLBS. Also, it is clear that while the LR is the dominant factor, a bursty loss process can seriously impair perceived quality. The quality decreases more markedly if the LR values are low whilst the degradation changes less for higher LR's. Shown in two dimensions is the loss rate for a non-bursty channel (Figure 6) and a bursty one (Figure 7). The absolute deviations are also given. In most cases the absolute deviation is small, this indicates that the median is indeed a good approximation for the PESQ scores for the 200 samples considered for each point (20 samples and 10 random seeds for the loss generation).

A. Discussion

We consider that an approximation within approximately 0.5 MOS points of a sample's actual PESQ score as reasonable.

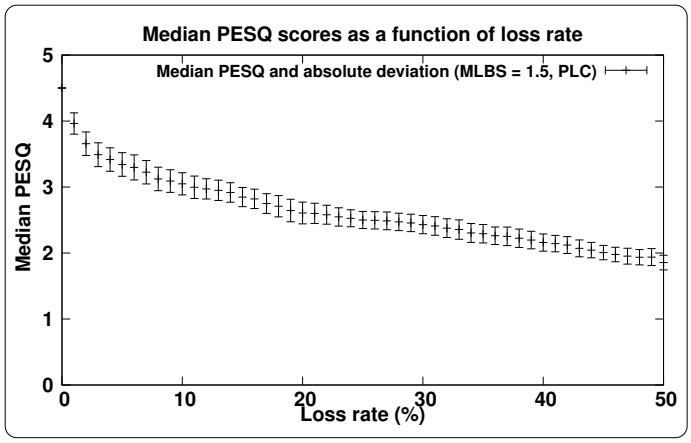

Fig. 6. Median PESQ scores and absolute deviation as a function of loss rate (MLBS = 1.5 packets).

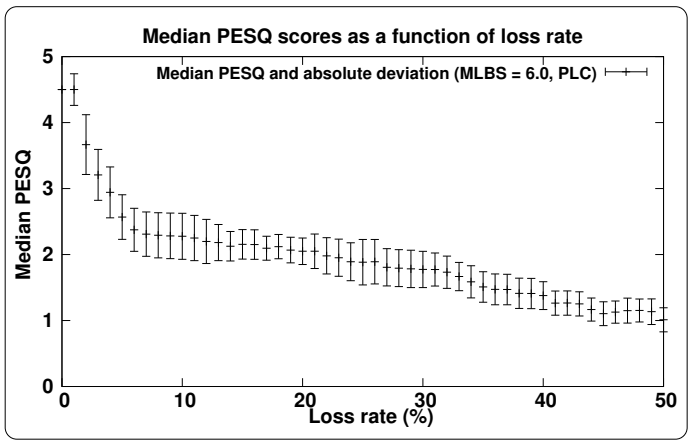

Fig. 7. Median PESQ scores and absolute deviation as a function of loss rate (MLBS = 6.0 packets).

In order to test the choice of the median as a such a measure, we calculated for each point of the loss space, the fraction of samples that are outside of this range at each point in the loss space. The results are shown in Figure 8. For MLBS values of up to 2 packets, the approximations include 90% of the points throughout the whole LR range. For MLBS values of up to four packets, 80% of the points lie within this range. In general, we observed that for the combinations of LR's and MLBS most likely to be found during VoIP usage (i.e. low LR and MLBS values), the estimation to be reasonable. There is more variability in the PESQ scores, and hence less accuracy, in the approximation, for combinations of LR and MLBS outside of the ranges we considered. This can be observed in the region where the LR is low and the MLBS is high.

In order to investigate this somewhat further we took a look at four chosen points in the loss space, these are given in Table 9. We can see the median is reasonably close to the mean, indicating that the data is not particularly skewed. We can also observe that the higher burst sizes (7.0) are positively skewed, this indicates that the higher portion of the tail of the distribution is longer than the lower end of the tail. For the lower burst sizes (2.5) the skews are negative, indicating the

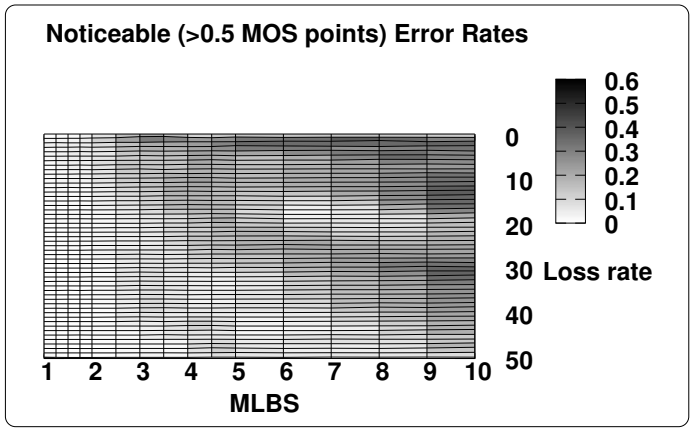

Fig. 8. Noticeable error rates for our metric. The error rates were computed as the fraction of points lying at more of 0.5 MOS points from the median value. Note that for *sensible* (LR,MLBS) combinations, accuracy is mostly well over 90% for the samples used at each point.

LR	MLBS	median	mean	var.	stan.dev	skew
12	2.5	2.81	2.78	0.07	0.27	-0.63
12	7.0	2.18	2.32	0.30	0.55	1.38
37	2.5	2.01	2.00	0.05	0.23	-0.63
37	7.0	1.35	1.39	0.10	0.31	0.85

Fig. 9. Some statistical measures for four sample points in our loss space

reverse case for the distribution of losses.

B. Validation

In order to verify the accuracy of our single–sided metric, we performed validation runs with test samples and random patterns not previously used in our tests. We performed these tests within the practical loss space (Figure 3) with loss rates ranging from 0% to 20% and MLBS values of up to 2.5. We also performed validation runs for all of MLBS values in order to see what the actual performance of the metric would be in networks that exhibit bursty losses (e.g. wireless networks).

The validation results gave a lower correlation between the pre–calculated median and the new PESQ scores than when comparing our metric to the PESQ scores used to calculate the median. For the loss ranges stated above, over five 100–sample runs, we obtained an average 64.5% of scores for which the difference with the median was less than 0.5 of a MOS point. Of those runs, the worst case (49% accuracy) was over a loss space of up to 20% LR and a 3 packet MLBS. The best case (75% accuracy) was for a smaller space ranging up to 15% loss rate, and 2.5 packet MLBS. A significant fraction of the PESQ scores lies at less than 0.6 MOS points from the median. The average accuracy is 81.6%, the worst case is 63%, and the best case is 86%. Figure 10 shows the CDF for the absolute errors for one hundred validation points (for LRs up to 20% and MLBS up to 2.5 packets).

For the full range of MLBS values, and loss rates up to 15%, the worst accuracy was 52%, the average 54%. With a 0.6 MOS point tolerance these results become 60% and 64.5% respectively. So the conclusion is for a restricted loss space we are within a suitable operating range and for the all loss rates

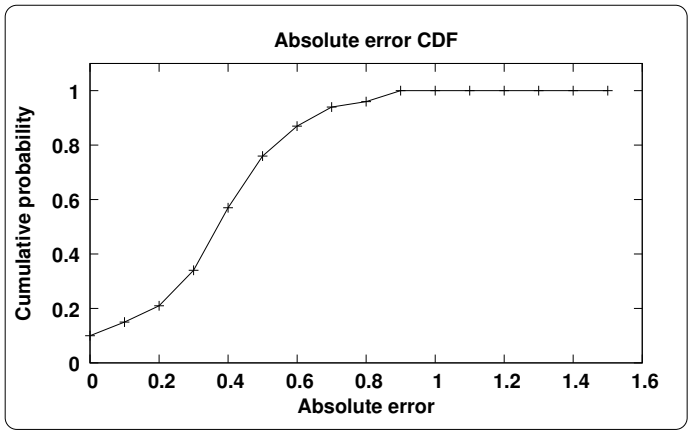

Fig. 10. Cumulative density function for the observed absolute errors in a 100–point validation. Loss rates ranged from 0 to 20% and MLBS values ranged from 1 to 2.5 packets.

we might need to increase the tolerance a little by 0.1 of a MOS point.

V. Conclusions

In this paper we propose a single–sided voice quality metric that gives results close to those given by PESQ. Our solution does not need the complex processing of PESQ nor does it require a reference signal at the receiver. In order to achieve this goal we have studied PESQ scores as a function of network losses. We propose to use the median of the PESQ scores in order to map the network losses to a quality estimation. We have used G.711 in these tests, however any codec could be estimated with this method and the G.729 and iLBC codecs are available to us. Briefly these codecs give more robustness to bursty losses resulting in "shallower" curves. Implementation of the results could be done either by approximating the function as shown in the figures, or by using a table of the loss values and PESQ scores. Thus, it would be required to measure the LR and MLBS at the receiver and then use a mapping to the PESQ scores or perform a table lookup. Our solution for single-sided quality estimation is suitable for real–time use on any portable communication device, in particular those which are do not possess substantial processing power.

References

[1] 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. [Online]. Available: http://www.itu.int/

[2] ITU-T Recommendation P.563, "Single–ended Method for Objective Speech Quality Assessment in Narrow–band Telephony Applications," May 2004. [Online]. Available: http://www.itu.int/

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

[4] C. Hoene, "Internet Telephony over Wireless Links," Ph.D. dissertation, Technical University of Berlin, Germany, Dec. 2005.

[5] International Telecommunication Union, "ITU-T coded-speech database," *ITU-T Series P, Supplement 23*, Feb. 1998. [Online]. Available: http://www.itu.int/

[6] E. Gilbert, "Capacity of a Burst–loss Channel," *Bell Systems Technical Journal*, vol. 5, no. 39, Sept. 1960.