An extension of PESQ for assessing the quality of speech degraded by severe time clipping and linear frequency response distortions

John G. Beerends[1], Jeroen M. van Vugt[1]

[1] *TNO Telecom, P.O. Box 5050, 2600 GB, Delft, The Netherlands, Email: j.g.beerends@telecom.tno.nl*

Abstract

PESQ (ITU-T recommendation P.862) has been established as a world standard for assessing speech quality. This paper shows its limitations when it is used for assessing the quality of speech that has been degraded by large amounts of time clipping and linear frequency distortions. A significant improvement in correlation could be obtained without degrading the PESQ performance on other types of distortion.

Introduction

In the development of PESQ (ITU-T recommendation P.862 [1]) a large set of speech quality data was used in both the training and the validation. The focus of the data was on distortions as can be found in real world applications but the data originated from recordings in the electric domain, thus containing little amounts of linear frequency response distortion. Within PESQ these lineair distortions are compensated up to differences of 20 dB. Currently PESQ P.862 is being extended towards the acoustic domain (P.AAM=Acoustic Assessment Model [2]) where lineair frequency distortions can reach values above 20 dB for which PESQ cannot correctly estimate the impact on the perceived speech quality. Furthermore, already during the ITU validation of PESQ, it was noticed that for severe amounts of time clipping (above 25%) PESQ also fails to correctly predict the impact on the perceived quality [1]. This paper deals with the development of an extension to PESQ that can cope with both types of distortions. Although the amount of distortions that are used in this paper are only seldom found under normal operational conditions one would like an objective method to provide reasonable correlations in order to be able to assess systems under extreme conditions.

Subjective experiment

A subjective experiment was set up to contain severe amounts of linear frequency response distortions and packet loss (time clipping). A simple database approach was taken, using two voices, one male and one female, with a length of 35 s, sampled at 8 kHz, 16 bits, mono. Eleven different linear frequency response distortions (***F***) were generated by applying an FFT filter to the speech signal. Bandwidth limitations, single and multiple peaks, single and multiple dips, a frequency slope and wild frequency response distortions were used. Frequency response distortions with variations up to ±30 dB were used. Twelve different time clipping distortions (***T***) were used. They were constructed by introducing a regular packet loss between 1 and 50 % with packet sizes between 3 and 96 ms. All samples representing a lost packet were set to zero, meaning no packet loss concealment was used.

Combining the time distortions with the frequency distortions results in a product space ***T***⊗***F*** of 132 conditions that had to be assessed. All 132 speech files were scaled to the same Active Speech Level of –30 dBov, according to ITU-T P.56. In order to have a controlled reproduction environment a sealed Hi-Fi headphone reproduction was used that produces minimal linear (and non-linear) distortion. Note that this implies that the telephone band filter as used in PESQ P.862 has to be removed in order to have a 1-1 relation between the acoustic signal as present in the ear of the subject and the signal as used by PESQ. The signal was only put on one ear of the headphone. Only a small set of experts, trained to behave as an average naive listener, was used in the subjective assessment.

PESQ and P.AAM Results

As stated one cannot apply PESQ directly to the recorded signals. The PESQ results on the ***T***⊗***F*** database that are given in Figure 1 were obtained by removing the IRS filter from PESQ. It is clear that even this context adapted PESQ gives far too low correlation on the dataset.

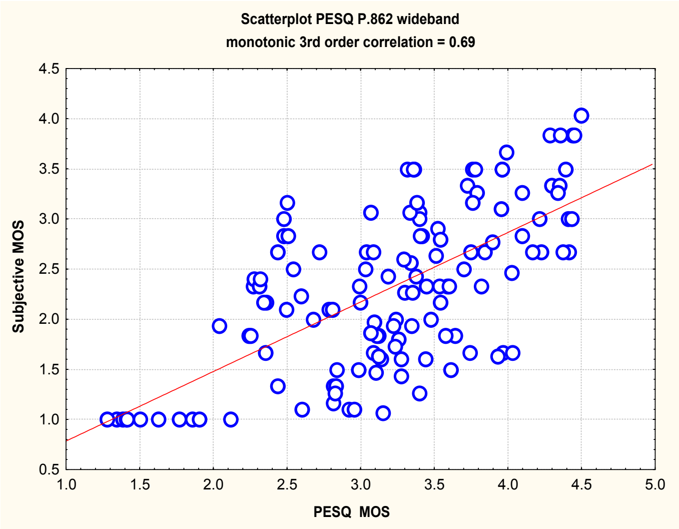

Figure 1: PESQ results on the *T*⊗*F* database containing severe amounts of time clipping and severe linear frequency response distortions.

P.AAM was developed for end-to-end acoustic measurements where the signal that is presented to the model is an exact copy of the signal that is presented to the objective measurement algorithm. Thus P.AAM can be used on the ***T***⊗***F*** database. Furthermore, during the development of P.AAM, linear, acoustic domain frequency response distortions were taken into account and it was thus expected that both P.AAM models that were submitted for standardization would give a significantly higher correlation. The results as given in Figures 2 and 3 show no improvement in correlation. The reason for this low correlation is that P.AAM was trained with distortions as found under normal operational conditions.

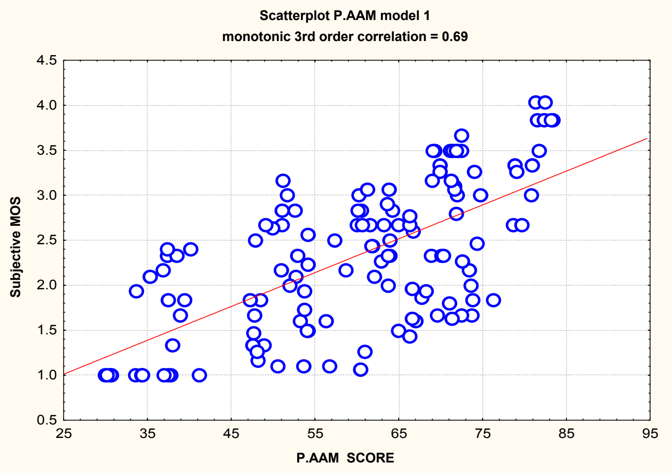

Figure 2: P.AAM model 1 results on the *T⊗F* database containing severe amounts of time clipping and severe linear frequency response distortions.

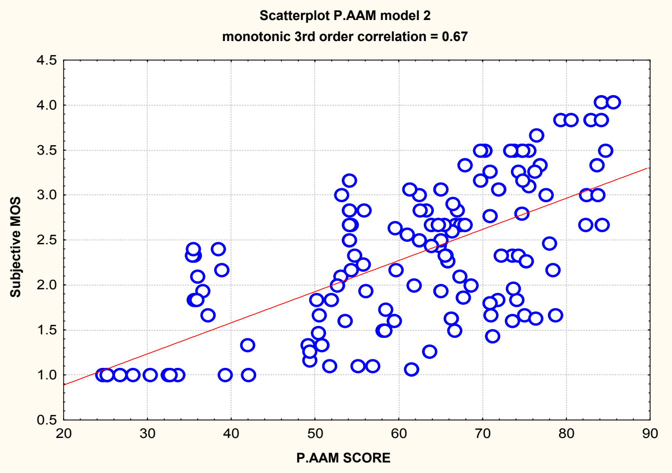

Figure 3: P.AAM model 2 results on the *T⊗F* database containing severe amounts of time clipping and severe linear frequency response distortions.

Improving PESQ/P.AAM

From the results in the previous section it is clear that both PESQ and P.AAM need further improvement in order to be able to assess quality over a very wide range of distortions. The main difficulty lies in maintaining high correlations over the databases with which PESQ and P.AAM were developed. Several ideas were exploited, the main idea being that for the wideband listening situation, i.e. where the speech signals are still narrowband (sampling of 8 kHz) but there is no telephone band filter used, the quality assessment modelling should be different from the situation where telephone band filtering is used. The exact manner that upper and lower frequency bands (below 200 Hz and above 3000 Hz) are used in the speech quality assessment is difficult to model and should be optimised for both listening situations seperately. This leads to a "split" model approach where the listening situation (telephone band, wide band) is used to set a number of parameters in the quality assessment model. The P.AAM models that were submitted to the ITU use a "single" model approach.

A second idea is the introduction of separate quality indicators to be able to quantify certain types of special distortions. This idea was already used in the development of P.AAM, especially in model 2 that uses a binaural approach. The main problem of this approach is stability over all possible distortions.

The combination of the two ideas were used in the development of an enhanced PESQ, called PESQ+, that shows a minor improvement in overall quality when assessing databases used in the development and validation of both PESQ and P.AAM, and a significant improvement in the *T⊗F* database (see Figure 4). The final correlation on this *T⊗F* database is still not up to a level that one would require for accurate prediction of the perceived quality (>0.9) and the method is still under development. Exact details of the improvements will be made available in future standardization of speech quality measurement methods. It should be noted that when the amount of training data is limited the correlation that can be obtained for the *T⊗F* database goes above 0.95.

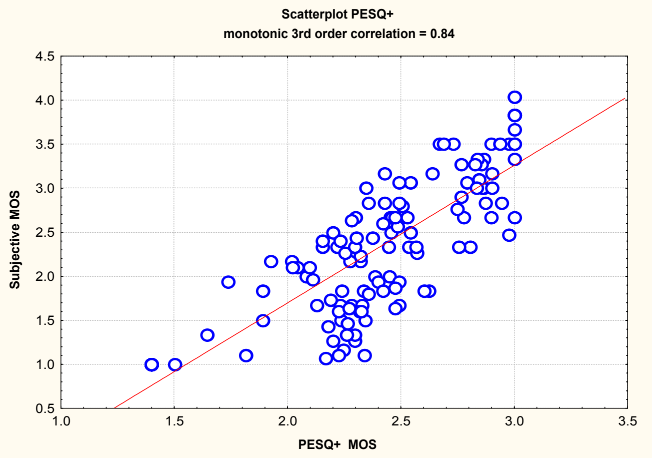

Figure 4: PESQ+ results on the *T⊗F* database containing severe amounts of time clipping and severe linear frequency response distortions.

References

[1] ITU-T Rec. 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," ITU, Geneva, Switzerland (2001 Feb.).
[2] T. Goldstein, J. G. Beerends, H. Klaus and C. Schmidmer, "Draft ITU-T Recommendation P.AAM," COM 12-64 to ITU-T Study Group 12 (Sep. 2003).

Acknowledgement

This work was supported by the Dutch Ministry of Economic Affairs via its agency Senter (project EQUANET) and KPN Mobile The Netherlands.