Extension of ITU-T Recommendation P.862 PESQ towards Measuring Speech Intelligibility with Vocoders

John G. Beerends
TNO Telecom, P.O. Box 5050, 2600 GB Delft, The Netherlands
Telephone: +316 51612563

E-mail: j.g.beerends@telecom.tno.nl

Sander van Wijngaarden, Ronald van Buuren
TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
Telephone: +31 346 356 330

E-mail: vanwijngaarden@tm.tno.nl, vanbuuren@tm.tno.nl

ABSTRACT

ITU-T recommendation P.862 PESQ was developed for assessing speech quality. The basic idea in PESQ is to compare a reference speech signal with the degraded signal through a psycho-acoustic model and a model of human quality comparison (cognitive model). Within NATO, testing of low bit rate speech codecs is more focused on speech intelligibility than on speech quality. Although the cognitive model of PESQ was designed to represent the quality judging process, it was already found that, for specific applications, PESQ can also predict intelligibility. In this paper PESQ is validated for assessing speech intelligibility with low bit rate vocoders. The results show that improvements in PESQ are necessary in order to obtain high correlations between objective and subjective intelligibility scores.

1 INTRODUCTION

ITU-T recommendation P.862 PESQ [1], [2], [3] was developed for assessing speech quality and was never validated in terms of speech intelligibility. Although there is a relation between speech quality and speech intelligibility, it is not clear that PESQ can be used to predict intelligibility. One should be aware of the fact that one can improve speech quality while decreasing intelligibility and reversely increase intelligibility while decreasing the speech quality. An example is the trend to improve the end-to-end perceived speech quality by using noise suppression systems that tend to improve the quality regarding the background noise but that tend to decrease the speech intelligibility.

This paper gives the results of a validation and extension of ITU-T Rec. P.862 on speech intelligibility. The database that was used results from a NATO speech intelligibility test on vocoders/noise suppressors and uses CVC (Consonant Vowel Consonant) intelligibility score. This data base consisted of long speech files (about 3 minutes long) containing 50 CVC (Consonant Vowel Consonant) words embedded in a carrier sentence. The intelligibility score is defined as the percentage of words for which both consonants and the vowel were identified correctly. This score is calculated per vocoder/noise condition and was measured as an average over four speech files (2 male and 2 female talkers). Twelve different noise conditions were used to assess the quality of 9 vocoders/noise suppressors. Bit rates of the codecs were between 1 and 5 kbit/s. For each condition 8 different sentences were used to assess the overall speech intelligibility score. Signals were presented diotically to listeners seated in a sound-treated room over wideband headphones.

Beerends, J.G.; van Wijngaarden, S.; van Buuren, R. (2005) Extension of ITU-T Recommendation P.862 PESQ towards Measuring Speech Intelligibility with Vocoders. In *New Directions for Improving Audio Effectiveness* (pp. 10-1 – 10-6). Meeting Proceedings RTO-MP-HFM-123, Paper 10. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.aps.

2 PESQ P.862 VALIDATION RESULTS

Applying ITU-T Recommendation P.862 to the input and output speech samples of the database produces the results as given in Figure 1. Although the correlation between the objective PESQ MOS score and the CVC intelligibility score is quite high, r = 0.86, it is not up to a level that reliable intelligibility prediction can be made.

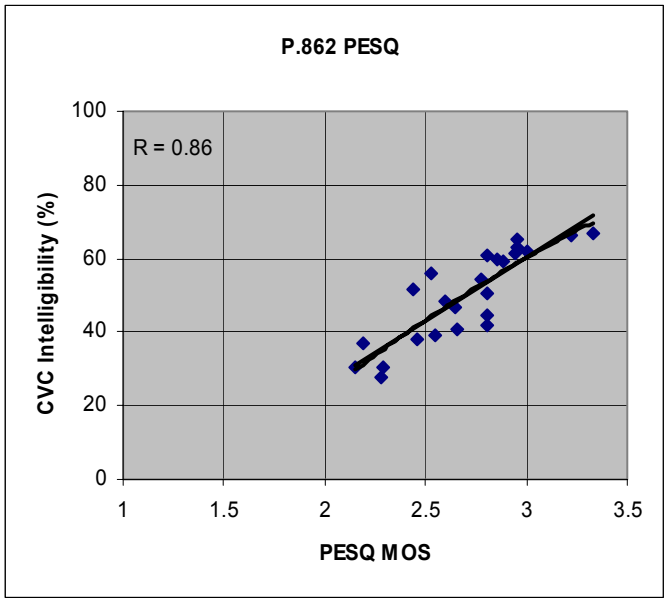

Figure 1: Relation between CVC Intelligibility and the P.862 PESQ Mean Opinion Score for the NATO speech intelligibility database.

3 EXTENDING PESQ P.862 TOWARDS MEASURING INTELLIGIBILITY

PESQ P.862 was developed for telephone band filtered speech signals [4] while the processing of the NATO speech signals involved no telephone band filtering in either the input to the system under test nor in the listening of the speech files in the subjective evaluation. The input files thus contained frequencies up to about 4000 Hz, half the sampling rate used in the experiments, and down to about 70-100 Hz, the natural lower limit of the human voice. The spectral balance tends to show that a microphone has been used that suffers from a low frequency boost to the proximity effect [5]. The output files in general showed almost no filtering effects although it would have been beneficial to filter the incoming speech below 300 Hz in order to get an optimized bit allocation in the speech encoder. Figure 2 gives a typical example of an input output frequency spectrum over a time span of about 2 seconds. It also shows significant amounts of distortion in the lower part of the frequency spectrum, between 50 and 300 Hz, a region that in normal telephone connections is filtered severely in order to optimize the signal to noise ratio (headroom), the perceived timbre and to minimize self masking as occurs when speech is played on a louder than natural level. Note that in a normal telephone connections the play back level is up to about 20 dB louder than a live conversation [6], making the self masking effect significant.

This analyses shows that a first straight forward idea for improving the correlation between PESQ scores and intelligibility scores is to extend the processed frequency range from the currently used IRS receive telephone band filtering (300-3400 Hz) to a wide band filtering (50-7000 Hz). However one should realise that PESQ was never developed to correctly model the loudness perception in the frequency bands

between 50 and 300 Hz. A simple removal of the input IRS filtering as used in PESQ only showed a marginally improved correlation and clearly other improvements need to be implemented in order to get higher correlations. Two ideas related to the spectral balance problem that should improve correlation are the introduction of upper slope frequency domain masking and the loudness growth in the lower frequency region (below 300 Hz). From these ideas only the loudness growth gave a significant improvement and was used in the PESQ intelligibility extension.

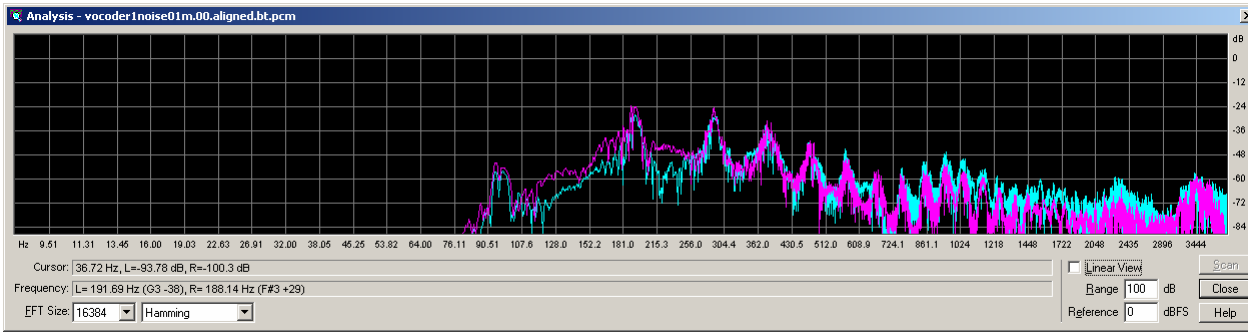

Figure 2: Typical example of the input and output power spectral density of a 2 second fragment as used in the NATO speech intelligibility database. In the lower part of the spectrum (below 300 Hz) the distorted vocoder output is higher than the input, while for the upper part of the spectrum (above 800 Hz) the reverse is true. from the view point of both quality and intelligibility this is a non optimal frequency distribution.

In order to further improve correlation one can take inspiration from the standard manner of measuring intelligibility using either the Articulation Index [7], [8] or the Speech Transmission Index [9], [10]. The AI is more or less a classical signal to noise approach, while the STI uses a modulation degradation calculation. Both of these effects are taken into account in the calculation of the PESQ MOS which is based on a frame by frame (32 ms length) bark power spectrum difference between the time aligned input and output speech signal [3]. Noise will result straight forward in a frame by frame bark power spectrum difference. Modulation differences between input and output will predominantly be seen as a global simultaneous change in all bark spectral frequency components. The effective frame length is in the order of 20 ms thus allowing to see modulation differences up to about 25 Hz. A major difference between the STI approach and the PESQ approach is that these global spectral changes only provide a global modulation difference and do not allow for a modulation frequency decomposition. An advantage is that the impact of each frequency component on the modulation is automatically weighted with its effective loudness. Because extreme low modulation frequencies (<0.1 Hz) are only perceived as a temporal change in volume these modulations are compensated for in PESQ by an adaptive frame by frame power scaling with a time constant of about four frames.

The time constant with which volume changes (i.e. low frequency modulations) are processed has a significant impact on the way that modulation differences are taken into account in PESQ and it was one of the further parameters that was changed in order to get a better correlation with the speech intelligibility.

Further significant improvements could be made by implementing the following ideas (in order of importance):

- Re-optimizing the integration over the time frequency distortion plane.
- Re-optimized to take into account the difference in impact of noise in the silent parts when comparing speech on intelligibility with speech quality. The impact of additive (background)

noise is taken into account in P.862 in an early stage of the perceptual mapping by applying a local scaling that reduces the noise level.

- Re-optimizing the linear frequency compensation. In P.862 PESQ linear frequency distortion are simply disregarded below a level of 20 dB, and distortions above 20 dB are reduced by 20 dB. This simple processing is replaced by a more advanced compensation that partly compensates all distortions, including distortions below 20 dB.

For the NATO database these improvements increased the correlation from 0.86 to 0.95. The results are given in Figure 3 where the PESQ MOS scale is replaced by an intelligibility score that is based on a compromise intelligibility scale with values that lie between a percentage CVC intelligibility and a sentence intelligibility score [11].

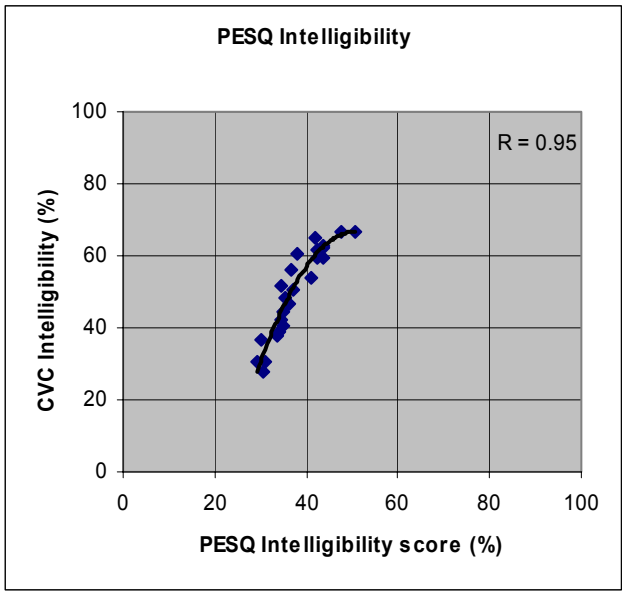

Figure 3: Relation between CVC Intelligibility and the P.862 PESQ Intelligibility score for the NATO speech intelligibility database.

4 CONCLUSIONS

The results show that PESQ P.862 [1], [2], [3] provides acceptable results when used to predict intelligibility. Significant improvements can be formulated that increase the correlation between objective and subjective intelligibility scores from 0.86 for PESQ up to 0.95. Further validations are necessary in order to see if the improvements that are implemented can cope with a wide range of distortions.

5 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," International Telecommunication Union, Geneva, Switzerland (2001 Feb.).

[2] A. W. Rix, M. P. Hollier, A. P. Hekstra and J. G. Beerends, *"PESQ, the new ITU standard for objective measurement of perceived speech quality, Part 1 - Time alignment,"* J. Audio Eng. Soc., vol. 50, pp. 755-764 (2002 Oct.).

[3] J. G. Beerends, A. P. Hekstra, A. W. Rix and M. P. Hollier, *"PESQ, the new ITU standard for objective measurement of perceived speech quality, Part II - Perceptual model,"* J. Audio Eng. Soc., vol. 50, pp. 765-778 (2002 Oct.).

[4] ITU-T Rec. P.48, "Specification for an Intermediate Reference System," International Telecommunication Union, Geneva, Switzerland (1989).

[5] D. Josephson, "A Brief Tutorial on proximity Effect", contribution to the AES 107th Convention," preprint 5058, Sep. 1999.

[6] ITU-T, "Handbook on Telephonometry," International Telecommunication Union, Geneva, Switzerland (1992).

[7] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, pp. 90-118 (1947 Jan.).

[8] K. D. Kryter, "Methods for the calculation and use of the Articulation Index," *J. Acoust. Soc. Am.*, vol. 34, pp. 1689-1697 (1962 Nov.).

[9] H.J.M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, pp. 318-326 (1980 Jan.).

[10] H.J.M. Steeneken, "On measuring and predicting speech intelligibility," PhD University of Amsterdam (1992).

[11] J. G. Beerends, E. Larsen, N. Lyer, J. M. van Vugt, *"Measurement of speech intelligibility based on the PESQ approach,"* Proceedings of the Workshop Measurement of Speech and Audio Quality in Networks (MESAQIN), Prague, Czech Republic, June 2004 (equivalent to TNO Telecom publication 33366).

ACKNOWLEDGEMENT

The authors wish to thank NATO NC3A for making available the speech intelligibility database used in this paper.