PESQ Based Speech Intelligibility Measurement

J. G. Beerends[1], R. A. van Buuren[2], J. M. Van Vugt[1], J.A. Verhave[2]

[1] *TNO Information and Communication Technology, P.O. Box 5050, NL-2600 GB, Delft, The Netherlands,*
Email: john.beerends@tno.nl, jeroen.vanvugt@tno.nl

[2] *TNO Human Factors, P.O. Box 23, NL-3769 ZG Soesterberg, The Netherlands,*
Email: ronald.vanbuuren@tno.nl, jan.verhave@tno.nl

Introduction

Several measurement techniques exist to quantify the intelligibility of a speech transmission chain. In the objective domain, the Articulation Index [1] and the Speech Transmission Index STI [2], [3], [4], [5] have been standardized for predicting intelligibility. The STI uses a signal that contains spectro-temporal characteristics similar to natural speech. By comparing intensity fluctuations of degraded and reference signals, the modulation transfer function of the system under test is measured from which the STI can then be calculated. In modern speech transmission, various types of coding are used that may behave differently for speech than for the STI test signal, implying a possible incorrect intelligibility prediction. A more fundamental approach in assessing intelligibility is to take natural speech signals, and derive internal representations of the reference and degraded signals, the difference of which can then be used to estimate speech intelligibility.

In this paper we investigate the idea to use a test signal composed of concatenated CVC (Consonant Vowel Consonant) words in combination with PESQ (Perceptual Evaluation of Speech Quality, ITU-T Rec. P.862 [6], [7], [8]) to measure speech intelligibility. PESQ was originally developed for measuring speech quality [9], [10] and uses a psycho-acoustic model to map the reference and degraded speech signals onto an internal representation. The difference in this representation is then used to calculate a difference disturbance function from which the perceived speech quality can be predicted in terms of Mean Opinion Scores (MOS). Since speech quality and speech intelligibility are closely related (e.g. [14], chapter 5). PESQ should allow for measuring speech intelligibility.

Subjective and objective test data

Subjective testing of speech intelligibility requires a different approach than subjective testing of quality for which PESQ was designed. In quality testing, a subject is asked for a personal opinion while in intelligibility testing his personal opinion is not relevant and only the fact whether or not a speech fragment is correctly reproduced is taken into account. To test whether one can use PESQ to assess intelligibility of degraded speech, a Consonant-Vowel-Consonant (CVC) database was constructed after [13]. This database used 40 conditions of communication channel signal processing to construct degraded speech signals. Distortions that were used included band filtering, peak clipping, reverberation and noise. The database was constructed using two different speech signal approaches.

In the first approach, sets of about 50 nonsense CVC words were embedded in carrier sentences producing speech stimuli of about three minutes. Eight of these speech stimuli, using eight different talkers (both male and female), were passed through a single degradation condition. Subjects were instructed to type the CVC words on a computer terminal and the percentage correctly identified words is taken as the intelligibility score. For each condition, the average score over eight talkers was taken as the intelligibility score for a degradation condition.

The second approach in the test set up was focussed on the objective testing and used a test signal derived from the signal of the first approach. This test signal contains the most important set of CVC words used in the first approach. The set of CVC words used in the test signal was chosen on the basis of completeness of the set of phonemes. As a compromise between speed of measurement, i.e. test signal duration, and accuracy, a length of 20 seconds was chosen. All objective measurements were carried out with this test signal.

PESQ results

The standard PESQ P.862.2 wideband algorithm as defined in [8] was first applied to the 20-second test signals. The correlation between the PESQ P.862.2 MOS values and the CVC scores, averaged over eight different talker signals, was low (r = 0.54), see Figure 1. The first obvious observation is that intelligibility problems mostly occur with MOS values below 2.0 and that a clipping effect is visible due to the fact that PESQ uses a MOS range of 1.0 (=bad quality) to 5.0 (=excellent quality). There is very limited degradation information available below a MOS value of around 1.2. Taking the raw PESQ scores gives a slightly higher but still low correlation with the CVC intelligibility scores (r = 0.59 see Figure 2).

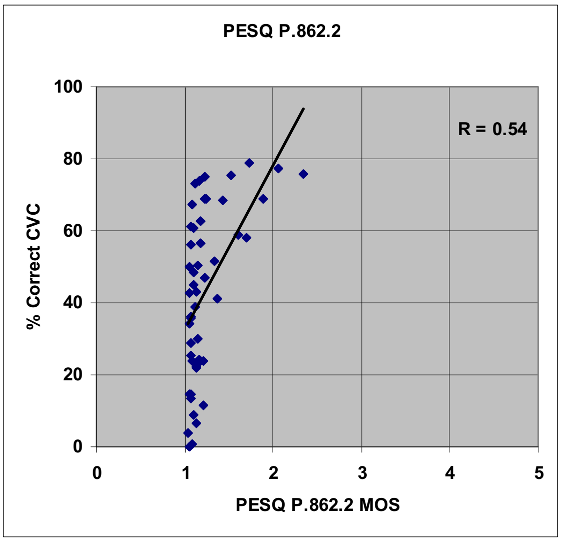

Figure 1: PESQ P.862.2 results for the CVC intelligibility database.

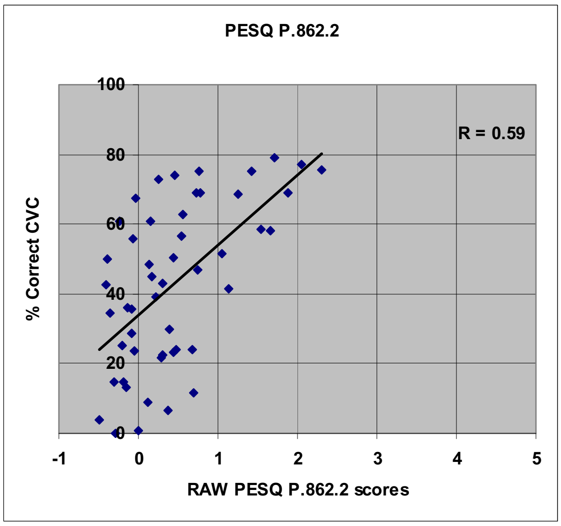

Figure 2: Raw PESQ P.862.2 results for the CVC intelligibility database.

Retraining PESQ

Investigating the reason for the low correlation showed that not only the disturbance introduced by the link under test is important, but also the variance behaviour over the time-frequency plane. All places in the time-frequency plane where there are “holes” in the disturbance can be used for understanding speech. This effect can be taken into account by using an averaging over the time-frequency plane that also takes into account this variance behaviour. As explained in [10] the integration over the time-frequency plane of the Disturbance D(time, freq) in PESQ takes place in three steps:

- first, the frequency axis is integrated using an Lp norm to obtain the loudness of the disturbance in a single frame
- next, these loudnesses are integrated over 30 frames using an using an Lp=q norm (speech spurt integration)
- and finally, there is an integration over the speech spurts over the whole speech file using an Lp=r norm

The Lp powers of Dpqr in each integration (see [10]) can be optimized by taking into account the variance behaviour by calculating differences between averages with a different pqr (the standard variance is equal to $(L_2)^2$ - $(L_1)^2$). For maximum correlation the following time-frequency plane, Dpqr (time, freq). integration to CVC score (in pseudo code) was found:

```
$x = 0.864*D_{522} - D_{124}$;
IF ($x > 0.0$) $x = 0.0$;
CVC score = $95.0 + 1.55x - 1.36x^2$ ;
IF (CVC score $< 0.0$) CVC score = $0.0$;
```

A few things can be observed from the mapping. First is that the asymmetric disturbance DA (see [10]) is no longer used in the calculation of the model output. Second is that our mapping saturates at 95% correct CVC scores, caused by imperfect behaviour of the subjects which have difficulty in getting 100% correct CVC scores.

Third is that when we have a flat distribution (i.e. $D_{124} = D_{522}$) the mapping to CVC score becomes $90.0 - 0.21D - 0.025D^2$, showing a normal behaviour of the intelligibility on the (always positive) difference D. In general, when we have a sharp frequency distortion, D_{522} dominates over D_{124} and there is little impact on intelligibility. This effect is well known, a tone can be very disturbing but the impact on intelligibility is marginal. Finally we see that for the frequency integration the contribution of the Lp power with the positive part (+ $0.864*D_{522}$) will often be larger than the negative part (- D_{124}) due to the fact that the D_5 integration is mostly larger or equal to the D_1 integration (value range is mostly above 1.0). However, when this occurs, the distribution of the distortion over the frequency axis is such that large parts of the spectrum are available for understanding speech. This is reflected in the pseudo code where positive values of the mapping to x are set to zero and the resulting intelligibility score is 95% correct. Applying this model to the intelligibility database gives an excellent correlation (see Figure. 3).

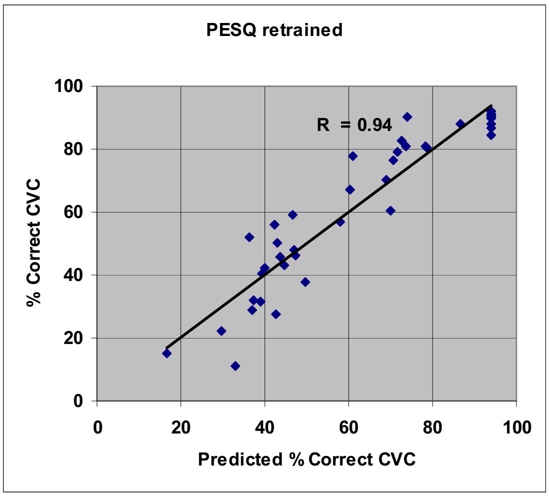

Figure 3: Results of the retrained PESQ model.

Conclusions and recommendations

The results clearly show that PESQ in its standard P.862.2 form cannot be used to measure intelligibility. Even when the raw PESQ output values are used, the correlation between the measured CVC scores and the PESQ model output is too low for practical use (r = 0.59).

However, re-training PESQ in such a way that the variance of the disturbance over the time-frequency plane is taken into account provides a method with a good correlation between model output and speech intelligibility (r = 0.94). This re-training only uses a different integration of the time-frequency plane of the Disturbance D(time, freq) to obtain an intelligibility score.

The new re-trained PESQ model only needs to process a short speech test signal of about 20 seconds that contains the set of most relevant CVC words used in the subjective CVC intelligibility test. The subjective CVC test that was used in the development of the method used carrier sentences, and it is expected that for sentence intelligibility the method will provide high correlations when using exactly the same speech sentences as used in the sentence intelligibility test.

It is clear that the results in this paper have been obtained by a retraining and that a validation on a larger set of intelligibility databases is necessary.

References

[1] 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.).

[2] 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.).

[3] ANSI S3.5, "Methods for Calculation of the Speech Intelligibility Index", 1997.

[4] ISO 9921, "Assessment of Speech Communication", 2003.

[5] IEC 60268-16, Sound system equipment, Part 16: "Objective Rating of Speech Intelligibility by Speech Transmission Index ", 2003.

[6] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs," International Telecommunication Union, Geneva, Switzerland (2001 Feb.).

[7] ITU-T Rec. P.862.1, "Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO," Geneva, Switzerland (2003 Nov.).

[8] ITU-T Rec. P.862.2, "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," Geneva, Switzerland (2005 Nov.).

[9] 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.).

[10] 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.) .

[11] J. G. Beerends, S. J. van Wijngaarden, R. A. van Buuren, "Extension of ITU-T Recommendation P.862 PESQ Towards Measuring Speech Intelligibility with Vocoders," NATO Human Factors and Medicine Panel Symposium on New Directions For Improving Audio Effectiveness, Amersfoort 11-13 April 2005.

[12] E. Zwicker and R. Feldtkeller, "Das Ohr als Nachrichtenempfänger," S. Hirzel Verlag, Stuttgart (1967).

[13] H.J.M. Steeneken, "Subjective phoneme, word, and sentence intelligibility measures," *in*: "On measuring and predicting speech intelligibility," Ph.D. thesis, ISBN 90-6743-209-1, pp. 37-75 (1992).

[14] R.A. van Buuren, "Speech Intelligibility and Sound Quality in Hearing Aids," Pd.D. thesis, VU Amsterdam, 1997