Comparison between subjective listening quality and P.862 PESQ score

A. W. Rix

Psytechnics Limited, Fraser House, 23 Museum Street, Ipswich IP1 1HN

antony.rix@psytechnics.com

Abstract

This paper discusses the relationship between subjective listening quality (LQ) mean opinion score (MOS), and objective quality score from the perceptual evaluation of speech quality (PESQ) model defined in ITU-T Recommendation P.862. The causes of variation of MOS between subjective tests, and the methods used in the ITU for comparing subjective and objective speech quality scores, are introduced. The motivation for using a single, average, mapping function is presented. Detailed analysis is given of a proposed mapping known as PESQ-LQ, including performance results for a large database of subjective tests. The results suggest that PESQ-LQ provides a good predictor of MOS for all of the network technologies, and for most of the languages, that were tested.

Keywords

Perceptual evaluation of speech quality, subjective test, mean opinion score

1 Introduction

The output of PESQ [1–3], termed PESQ score, has high correlation with MOS for a wide range of subjective tests spanning many different languages and network types [4]. However, PESQ score was calibrated against an essentially arbitrary objective distortion scale. It was not designed to be on exactly the same scale as MOS, either in general or for any specific subjective test. PESQ score may be between –0.5 and 4.5, while ACR listening quality MOS is on a 1–5 scale [5].

Subjective MOS varies significantly from test to test, depending on the balance of conditions and the individual and cultural preferences of the subjects. This is a very important point that will always limit the generality of any objective quality scale. This variation, and the mapping methods used to account for it, are described in section 2.

By analysing a defined set of subjective tests, it is possible to characterise an average relationship between PESQ and MOS. This can be used to derive a one-to-one functional mapping to compute a MOS-like objective quality score. The mapping offers a number of important benefits, including reducing the mean squared error by removing systematic bias, and facilitating interpretation of the results on the subjective listening quality scale, without any material reduction in the accuracy of the model.

The same concept may be applied to other models, such as PSQM [6–7], and had already been used in the development of PAMS [8]. PESQ was based on an integration of these two models.

Since many end-users are not able to perform per-experiment comparison with their own subjective test data, it is highly desirable to standardise a mapping. A mapping function, termed PESQ-LQ, has been proposed for this purpose [9] and is described in section 3. It is also important that the scope (range of network conditions, and languages) of the mapping is known. Section 4 therefore provides a detailed analysis of PESQ-LQ across a large database of subjective tests.

Work is now under way in ITU-T SG12 to evaluate this and other mappings. The aim is to standardise a single mapping function for use with P.862 in future.

2 Variation in quality scores

2.1 Variation between subjective tests

The most common subjective test method in telecommunications is the five-point absolute category rating (ACR) listening quality (LQ) scale defined in ITU-T P.800 [5]: excellent, good, fair, poor, bad. A one-to-one comparison between subjective MOS from different subjective tests is difficult with tests conducted according to the ACR LQ method. This is because

subjective votes are affected by factors such as the following.

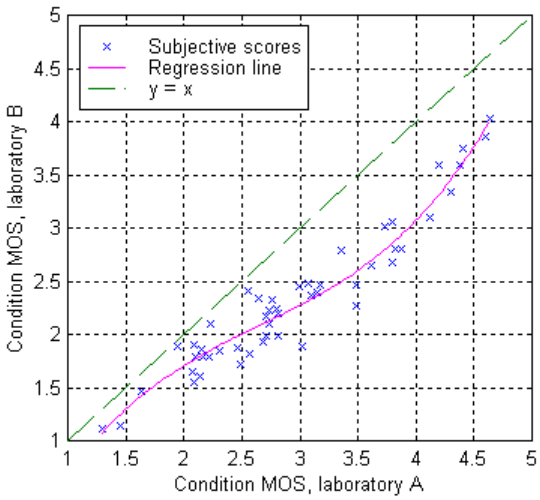
Cultural variation – in different languages and cultures, the meanings of "excellent .. bad" differ. This can have an effect of up to 1.0 MOS when comparing results for the same conditions from different laboratories. This is shown in Figure 1, which shows the scatter plot of conditions, and the 3rd-order polynomial regression line, between results of a test conducted in two different laboratories, each in their own native language and with native speakers as subjects.

Individual variation – our own personal experience also influences how we vote. As subjective tests tend to use a relatively small number of subjects (typically 24–32), systematic individual variations can leave a residual variation. Assuming that the standard deviation of individual variations is 0.5, with 24 subjects the 95% confidence interval for the mean quality score is on the order of 0.2 MOS.

Balance of conditions – the absolute rating method means that subjects adapt to some extent to the range of conditions in a test. A large proportion of conditions that are poor to bad, means that the best conditions are likely to be rated as excellent, as they are clearly distinct. Conversely, if there are fewer bad conditions, it is possible that the subjects may only rate the best conditions as good, as they are harder to distinguish. This can account for variations of up to 1.0 MOS between tests conducted at the same laboratory, which could not fully be explained by individual variations.

This makes it impossible to directly compare results from one subjective test with another; some form of mapping between the MOS results is required.

Figure 1: Cultural variation in MOS



2.2 Comparison between objective and subjective quality

As discussed above, it is unreasonable to expect results from different subjective tests to be identical. However, if the tests are well-designed and consistent, the ordering should be preserved (within experimental error) and the relationship between the two should be monotonic. A monotonic mapping function can therefore be applied to the results of one test to put it on exactly the same scale as another.

The same is true for comparing objective quality scores with subjective MOS, as objective perceptual models are generally calibrated against some arbitrary scale which is unlikely to be the same as MOS. In the case of P.862, PESQ score was intended to be used with a per-experiment mapping, and it was not designed to match any particular subjective test.

The mapping function used in ITU-T evaluation of objective models is a monotonic 3rd-order polynomial. This is applied, for each subjective test, to map the objective score onto the subjective score. It is then possible to calculate correlation coefficient and residual errors. Usually the process is performed per condition, reducing material dependence, but it can also be applied per file.

This process is illustrated by the following example, a subjective test on the performance of fixed and mobile networks with errors, noise and noise suppression. Figure 2 shows a scatter plot between subjective MOS and PESQ score, with the monotonic 3rd-order polynomial fit with minimum mean squared error. The PESQ score is mapped by this polynomial to give a prediction of subjective quality, shown in Figure 3.

Figure 2: Mapping between objective and subjective MOS – Raw scores and monotonic polynomial fit

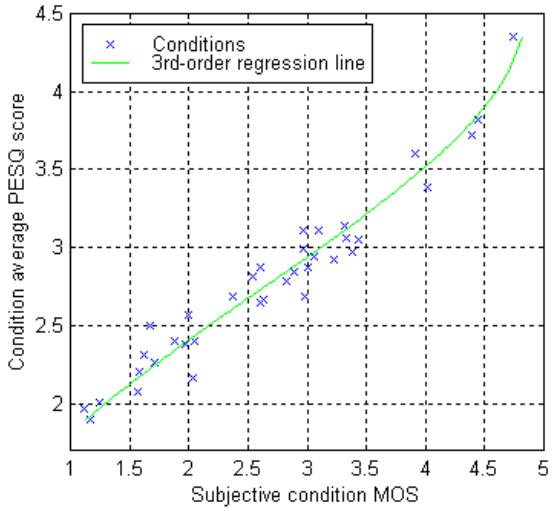


Figure 3: Mapping between objective and subjective MOS – Scores after application of mapping

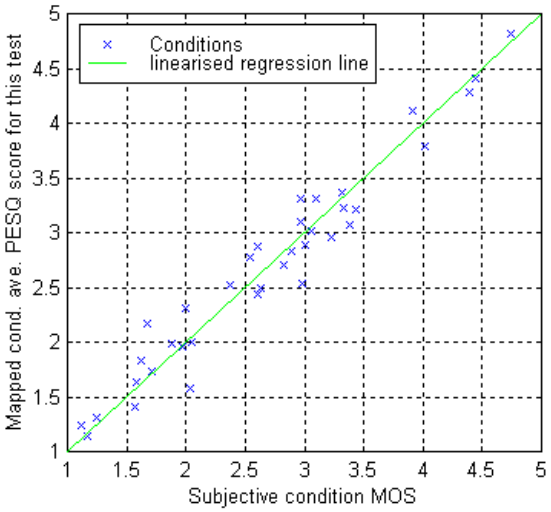
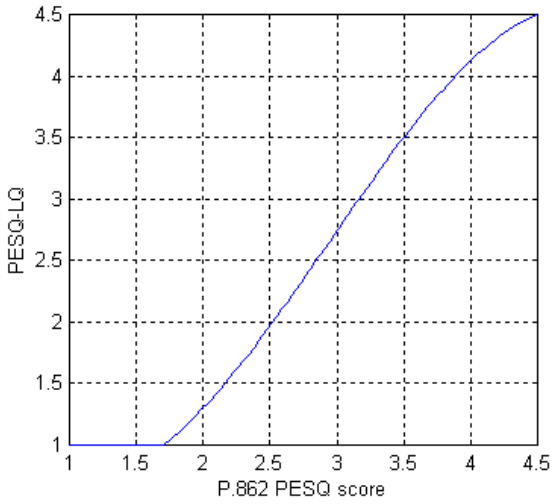


Figure 4: Mapping from PESQ score to PESQ-LQ



3 Overview of PESQ-LQ

Most users of objective quality measurement systems do not have access to subjective tests, and are not able to perform an analysis similar to that described in the previous section. Furthermore, the variation between subjective tests is often seen as confusing and undesirable. It would be preferable if the objective quality score was on an "average" MOS scale, independent of language or network type.

A mapping function from P.862 PESQ score to an average P.800 ACR LQ MOS scale was proposed in [9]. This is known as PESQ-LQ. The author has stated the following aims in designing PESQ-LQ:

- to produce values in the range [1, 4.5]
- to reduce the need for users to perform analysis such as set out in the previous section
- to give results that are close to average MOS over a large corpus of clean speech subjective tests
- to be applicable to a range of network types (fixed, mobile, VoIP)
- to be applicable to a number of different languages/countries.

The maximum value of 4.5 was chosen because this is the maximum quality achieved, for a clear, undistorted condition, in a typical ACR LQ test. For the same reason the maximum PESQ score was set to 4.5 when P.862 was developed.

PESQ-LQ is defined as follows, where x is the P.862 PESQ score and y is the corresponding PESQ-LQ:

$$y = \begin{cases} 1.0, & x \le 1.7 \\ -0.157268\,x^3 + 1.386609\,x^2 - 2.504699\,x + 2.023345, & x > 1.7 \end{cases}$$

The PESQ-LQ mapping passes through the points (x, y): (1.7, 1.0), (3.5, 3.5), (4.5, 4.5). In particular, the maximum remains at 4.5. The function is invertible for $x \ge 1.7$ using Cardano's method.

The function form of PESQ-LQ is shown in Figure 4.

4 Analysis of the performance of PESQ-LQ

The only way to verify if PESQ-LQ, or any other similar mapping, achieves its aims is to analyse its performance against a large database of subjective tests covering a wide range of technologies and languages. This section describes the performance criteria, datasets and results of this analysis.

In all cases the mapping is applied per file, although the statistics are calculated per condition. This is equivalent to using the mapped quality score in place of the original PESQ score. In addition, the performance of P.862 PESQ score was evaluated on the same basis.

4.1 Datasets

Table 1 and Table 2 list the datasets that were used. This database contains a large number of subjective tests conducted on a wide range of network technologies.

Table 1 lists the datasets by network type; the Clean, Mobile, Fixed and VoIP datasets were all subsets of the

Table 1: Network type datasets

Dataset	Number of experiments	Description
Overall	43	All P.800 ACR listening quality subjective tests
Clean	29	All clean speech (no background noise) tests
Mobile	19	Tests of mobile network conditions (both with and without background noise)
Fixed	9	Tests of fixed network conditions (both with and without background noise)
VoIP	10	Tests of multi-type and VoIP network conditions (both with and without background noise)

Table 2: Language datasets

Dataset	Number of experiments	Language, country
British	22	British English, United Kingdom
French	4	French, France
American	3	American English, USA and Canada
Japanese	2	Japanese, Japan
German	4	German, Germany
Dutch	3	Dutch, the Netherlands

Overall dataset. About half of the tests cover mobile network conditions; the remainder are split between fixed and VoIP technologies. About half of the tests were conducted in British English, and half are in other languages, including French, American, Japanese, German and Dutch.

Table 2 lists the subsets of the Overall dataset for different languages that were also analysed, where at least two subjective experiments were available for the given language. Note that most of these datasets are small because it is difficult to get access to subjective tests from different laboratories.

4.2 Performance measures

The following measures of the accuracy of the objective model were used for this comparison.

Correlation

P.862 section 7 recommends that the performance of an objective model should be measured using the correlation coefficient, after monotonic 3rd-order polynomial regression is applied to the per condition objective scores [1]. Regression is performed to eliminate any per-experiment variation.

The mean and worst-case correlation for each dataset are reported.

Raw RMS residual error

This is a measure of the accuracy of the mapped quality score as a predictor of MOS with no per-experiment mapping. This measures both prediction errors (due to errors by the model, or random variability in the votes) as well as any remaining systematic variation in votes between experiments.

Mean residual error

This is a measure of systematic mean deviation between the mapped quality score and MOS with no per-experiment mapping. A positive value indicates that the model gives too high a score.

4.3 Results

4.3.1 Network type

Table 3 shows the results for each of the performance measures, calculated for PESQ score and the PESQ-LQ mapping, across the different network type datasets.

4.3.2 Language

Table 4 shows the results for each of the performance measures, calculated for PESQ score and the PESQ-LQ mapping, across the different language datasets.

4.3.3 Scatter plots by language

Figure 5 to Figure 10 provide scatter plots for PESQ score and the PESQ-LQ mapping compared to condition MOS, for the language datasets.

Table 3: Results by network type

Dataset	Measure	PESQ	PESQLQ
Overall	Mean correlation	0.944	0.945
Clean	Mean correlation	0.955	0.956
Mobile	Mean correlation	0.962	0.963
Fixed	Mean correlation	0.948	0.950
VoIP	Mean correlation	0.931	0.933

Dataset	Measure	PESQ	PESQLQ
Overall	Min correlation	0.810	0.818
Clean	Min correlation	0.902	0.901
Mobile	Min correlation	0.905	0.908
Fixed	Min correlation	0.908	0.908
VoIP	Min correlation	0.837	0.839

Dataset	Measure	PESQ	PESQLQ
Overall	Raw RMSE	0.470	0.405
Clean	Raw RMSE	0.445	0.383
Mobile	Raw RMSE	0.376	0.336
Fixed	Raw RMSE	0.490	0.422
VoIP	Raw RMSE	0.540	0.457

Dataset	Measure	PESQ	PESQLQ
Overall	Mean error	0.204	-0.022
Clean	Mean error	0.193	-0.030
Mobile	Mean error	0.129	-0.115
Fixed	Mean error	0.234	0.005
VoIP	Mean error	0.249	0.035

Table 4: Results by language

Dataset	Measure	PESQ	PESQLQ
British	Mean correlation	0.957	0.958
French	Mean correlation	0.917	0.919
American	Mean correlation	0.941	0.943
Japanese	Mean correlation	0.951	0.951
German	Mean correlation	0.958	0.958
Dutch	Mean correlation	0.877	0.881

Dataset	Measure	PESQ	PESQLQ
British	Min correlation	0.905	0.908
French	Min correlation	0.837	0.839
American	Min correlation	0.926	0.927
Japanese	Min correlation	0.945	0.944
German	Min correlation	0.902	0.901
Dutch	Min correlation	0.819	0.819

Dataset	Measure	PESQ	PESQLQ
British	Raw RMSE	0.419	0.365
French	Raw RMSE	0.470	0.480
American	Raw RMSE	0.348	0.371
Japanese	Raw RMSE	0.647	0.417
German	Raw RMSE	0.529	0.407
Dutch	Raw RMSE	0.640	0.561

Dataset	Measure	PESQ	PESQLQ
British	Mean error	0.183	-0.042
French	Mean error	0.003	-0.187
American	Mean error	0.120	-0.112
Japanese	Mean error	0.597	0.350
German	Mean error	0.281	0.034
Dutch	Mean error	0.294	0.080

Figure 5: British

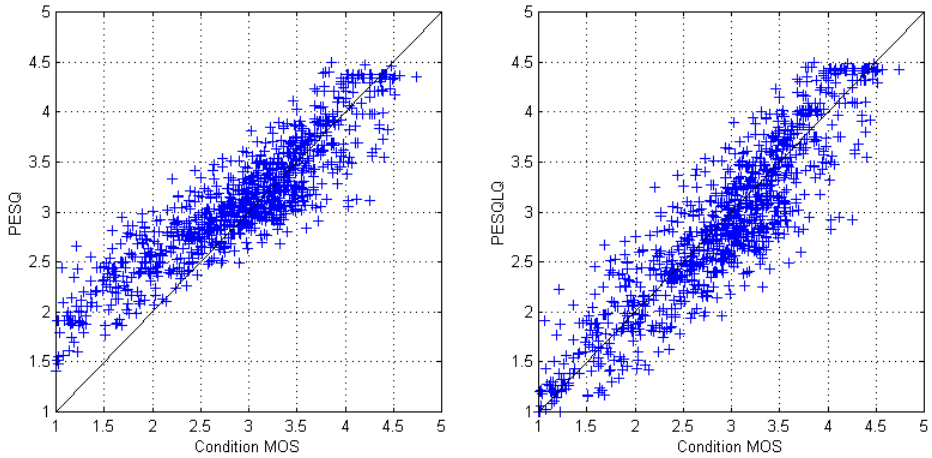


Figure 6: French

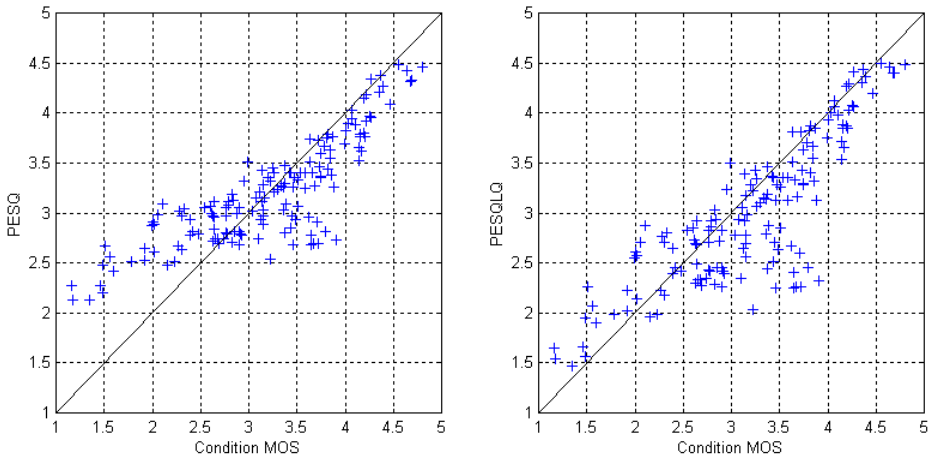


Figure 7: American

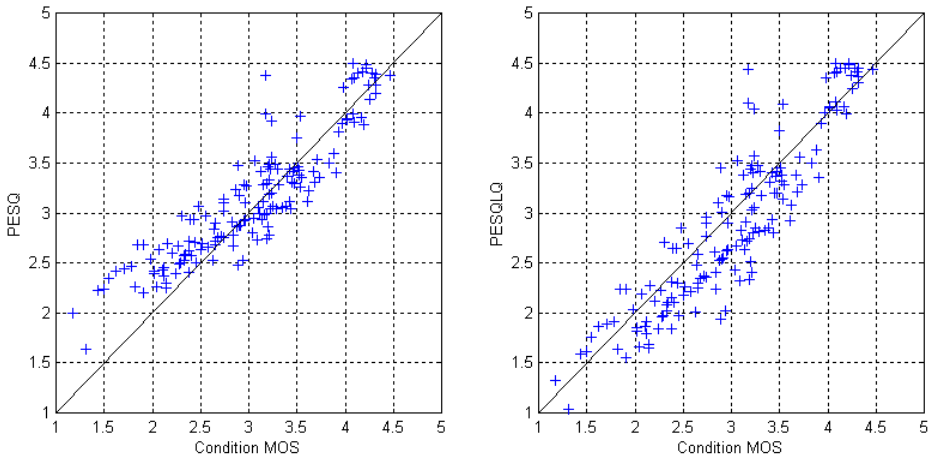


Figure 8: Japanese

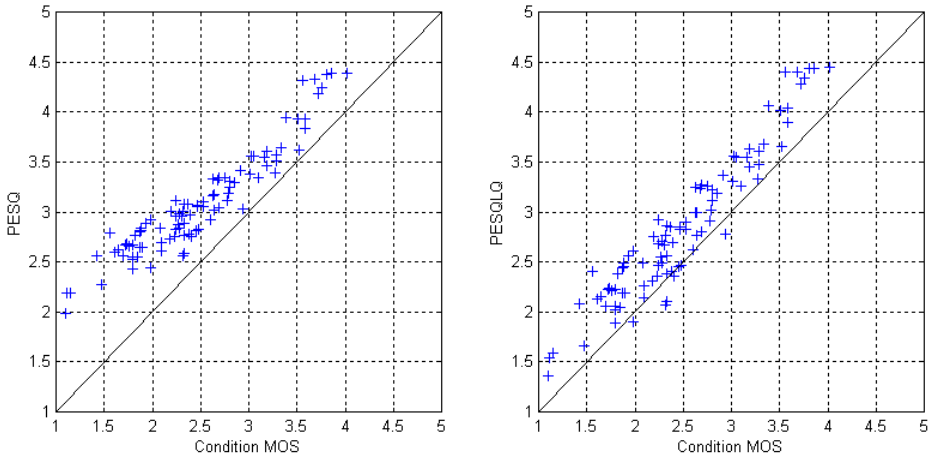


Figure 9: German

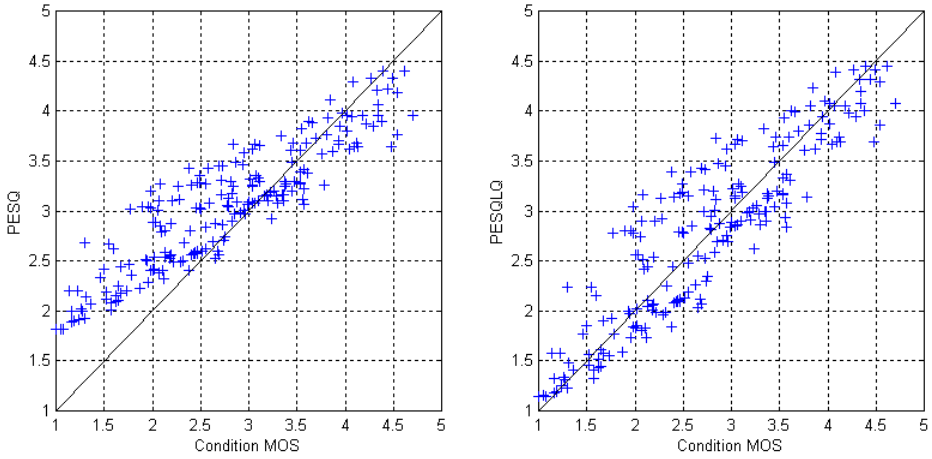
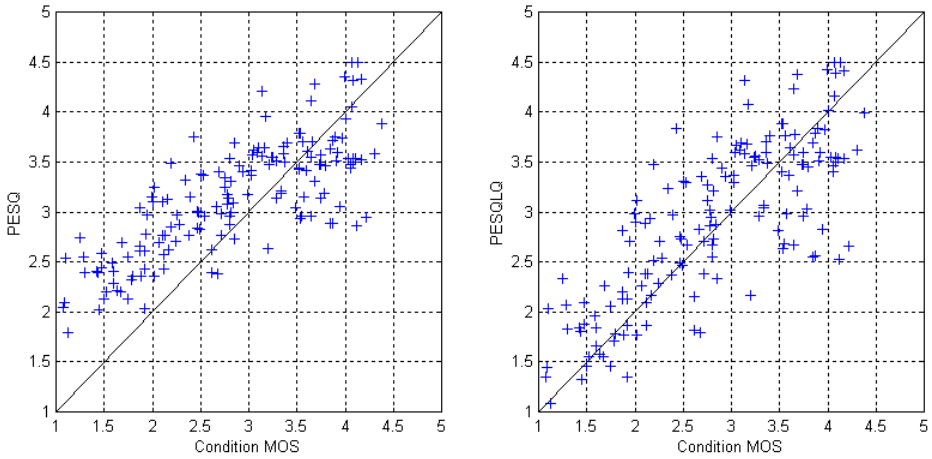


Figure 10: Dutch



4.4 Analysis of results

4.4.1 Effect of network type

PESQ-LQ makes little change to the mean or worst-case correlation coefficient – the largest difference is an improvement of 0.8%.

PESQ-LQ reduces raw RMS error on all network types tested. On the Overall dataset, raw RMS error is reduced by 0.065 compared to PESQ score. On the Mobile dataset, raw RMS error is reduced by 0.040.

The remaining residual RMS error is largely due to variations between subjective tests: for comparison, the RMS error per condition for PESQ score with a per-experiment mapping, on the Overall dataset, is 0.25.

PESQ-LQ reduces the mean error between objective score and MOS for all network types tested. For the Overall, Clean, Fixed and VoIP datasets the mean error for PESQ-LQ is smaller than 0.05. However for the Mobile dataset the mean error is –0.115 for PESQ-LQ, compared to +0.129 for PESQ score.

4.4.2 Effect of language and country

PESQ-LQ reduces raw RMS error for four of the six languages tested. However, it slightly increases raw RMS error for the French and American datasets, by 0.01 and 0.023 respectively.

PESQ-LQ reduces the mean error for five of the six languages tested. The exception is French, where PESQ-LQ is about 0.19 too low while PESQ score is on average almost correct for this language. Note however that PESQ-LQ is still 0.35 too high for the Japanese dataset, compared to 0.60 for PESQ score.

Note that, with the exception of British English, the datasets for most of these languages were small. Further study by laboratories with subjective test data and access to an implementation of P.862 is therefore desirable.

5 Conclusions

A mapping from P.862 PESQ score to an average MOS scale can significantly reduce the raw RMS error, when compared to many subjective tests without using per-experiment mapping.

The PESQ-LQ scale described in [6] appears to give good results across a wide range of network conditions and languages.

This suggests that PESQ-LQ is a good candidate for a universal mapping between P.862 PESQ score and average MOS.

However MOS does vary significantly between subjective tests, in particular between different languages and host countries. In consequence, a good mapping that predicts well on average may nevertheless give scores that are consistently too high for some languages and too low for others.

6 References

[1] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.* ITU-T Rec. P.862, February 2001.

[2] A. W. Rix, M. P. Hollier, A. P. Hekstra and J. G. Beerends. "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I – Time-delay

compensation." *Journal of the Audio Engineering Society*, 50 (10), 755-764, October 2002.

[3] J. G. Beerends, A. P. Hekstra, A. W. Rix and M. P. Hollier. "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II – Psychoacoustic model." *Journal of the Audio Engineering Society*, 50 (10), 765-778, October 2002.

[4] A. W. Rix, A. P. Hekstra, J. G. Beerends and M. P. Hollier. "Performance of the integrated KPN/BT objective speech quality assessment model." ITU-T SG12 COM12-D136, May 2000.

[5] *Methods for subjective determination of transmission quality.* ITU-T Rec. P.800, August 1996.

[6] J. G. Beerends and J. A. Stemerdink. "A Perceptual Speech Quality Measure based on a psychoacoustic sound representation." *Journal of the Audio Engineering Society*, 42 (3), 115–123, March 1994.

[7] *Objective quality measurement of telephone-band (300–3400 Hz) speech codecs*. ITU-T Rec. P.861, February 1998.

[8] A. W. Rix and M. P. Hollier. "The perceptual analysis measurement system for robust end-to-end speech quality assessment." *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* (3), 1515-1518, June 2000.

[9] A. W. Rix. "A new PESQ-LQ scale to assist comparison between P.862 PESQ score and subjective MOS." ITU-T SG12 COM12-D86, May 2002.