C-Qual - a validation of PESQ using degradations encountered in forensic and law enforcement audio

Dushyant Sharma[1], Gaston Hilkhuysen[2], Nikolay D. Gaubitch[1], Mike Brookes[1] and Patrick A. Naylor[1]

[1]*Imperial College London, UK*

[2]*University College London, UK*

Correspondence should be addressed to Dushyant Sharma (dushyant.sharma02@ic.ac.uk)

ABSTRACT

Assessment of speech quality of law-enforcement audio recordings is important as degradations introduced by non-ideal recording conditions can reduce the intelligence value of such recordings. Furthermore a model that predicts speech quality could be beneficial for assessing the performance of audio collection and enhancement systems. The Perceptual Evaluation of Speech Quality (PESQ) algorithm (ITU-T P.862) has been validated for degradations common in telecommunications. In this paper we apply PESQ to degradations typically encountered in law-enforcement. Also we present a subjectively labeled database (C-Qual) containing distortions encountered in law enforcement scenarios. Comparing the prediction by PESQ and the observed opinions provided by the listeners shows that PESQ is less suitable for estimating the speech quality in this context.

1. INTRODUCTION

Degradation of audio is commonplace in law enforcement due to the limitations of the collection methods employed, which in many cases are constrained by the requirement to be covert. Degradations encountered in law enforcement include additive noise, non linear effects such as clicking and peak clipping as well as reverberation, coloration and coding artifacts. In addition, the level of distortion may be severe (SNR worse than -5 dB) compared with those encountered in other domains. Also there may be a compounding of various degrading effects.

Speech quality modeling is a useful tool for analyzing the perceived effects of different processing systems employed in law enforcement, including assessment of the performance of a particular audio collection system or an enhancement system for example. In the field of telecommunications, such modeling is employed for the assessment of the quality of service of a particular communication system. The subjective assessment of speech quality is a time consuming and expensive exercise, thus it is advantageous to employ an objective technique. A number of techniques have been proposed in the literature for intrusive speech quality assessment, where the original signal before processing is also available. The current industry standard algorithm (PESQ) has only been validated for the types and levels of degradation encountered in the field of telecommunications [1].

The PESQ algorithm estimates the perceived effects of degradation by comparing the unprocessed signal with the degraded signal using a number of processing blocks, including time alignment, psychoacoustic modeling and distortion mapping. The estimated mean opinion score (MOS) is obtained using parameters previously trained [2]. It is an intrusive algorithm, requiring the clean signal as well as the processed (degraded) signal. This means that PESQ cannot estimate the single ended quality, where only the degraded signal is available. However, it is useful for evaluating the performance of different speech recording, communication and enhancement systems.

PESQ has not been validated in the known literature for use on highly degraded audio such as found in law enforcement. In this study we assess the validity of wideband PESQ in this context. An additional outcome of this study is a new database (C-Qual) containing speech data of the types and levels of distortion encountered in law enforcement, along with a subjectively labeled mean

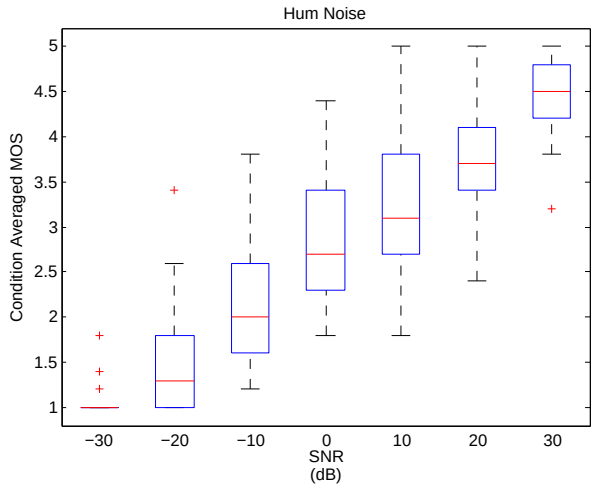

Fig. 1: A linear relationship is observed between subjective MOS and SNR for hum noise. Three outliers are detected for the -30 dB SNR condition.

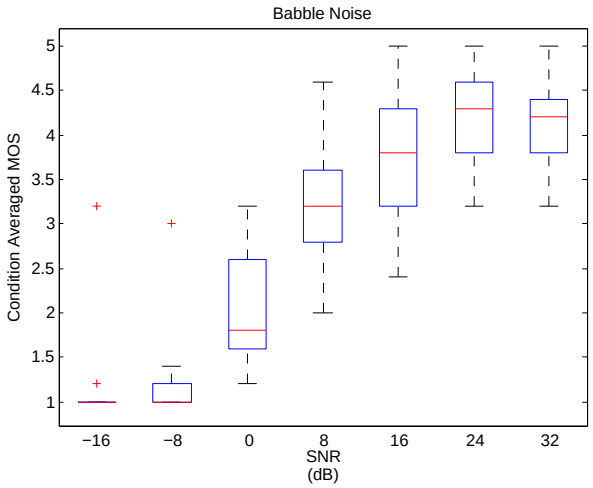

Fig. 2: Relationship between SNR and subjective MOS for babble noise. Two outliers are detected for the -16 dB SNR condition and another one at -8 dB SNR.

opinion score (MOS) for each speech utterance.

2. METHODOLOGY

In order to perform the validation, a database of degraded speech was constructed and subjectively scored to obtain the mean opining score for each audio example. The levels of degradation were adjusted to give a balanced distribution of PESQ scores. The estimated PESQ scores were then compared with the subjective scores to validate the performance of PESQ in the context of law enforcement.

2.1. Database design

Subjective quality scores were obtained from 24 subjects, who were native speakers of British English. In addition, the listener selection process required subjects to have a non-technical background to ensure them to be naïve to the effects of the degradations presented. All subjects had hearing thresholds of 20 dB HL or below at octave frequencies ranging from 125 to 8000 Hz.

The English subset of the ITU-T P.23 [3] database was used for the speech material, consisting of a pair of utterances with a small pause between the utterances. This ensures that our database can form an extension to the P.23 database. Two male and two female speakers were included in the stimuli. The active level of the speech was adjusted to give all files the same level using the ITU-T P.56 [4] method, before further processing. The level of degradation was selected to give a uniform distribution of PESQ scores. Six classes of degradations were included:

1. Additive noise: car, babble and hum noise were included at seven levels of the signal-to-noise ratio (SNR). Car and babble noise were added to the clean speech at -16, -8, 0, 8, 16, 24 and 32 dB SNR and hum noise was added at -30, -20, -20, 0, 10, 20 and 30 dB SNR.

2. Reverberation: two room impulse responses (RIRs) were included with different microphone to source distances. The Multichannel Acoustic Reverberation Database at York (MARDY) RIR [5] was included with 3 distances and the 'Imperial Office' RIR with 2 distances. The reverberation time (T_{60}) for the MARDY and Imperial rooms was calculated [6] to be 1.35 and 1.02 s respectively.

3. Coloration: in this study we implement spectral coloration using shelf filters. Three types of shelf responses are included, including an anti-clockwise spectral tilt.

4. Peak clipping: symmetric hard clipping was applied at four levels (-4, -8, -12 and -16 dB relative to the speech level).

5. Clicks: five levels of temporal erasures were included. The 20 ms erasures occurred with a frequency of 2, 7, 16, 150 and 440 clicks.

6. Modulated noise reference unit (MNRU) [7]: six levels of amplitude modulation are applied to the speech. These were included to validate the results from this study with results obtained in the P.23 database.

The task for the subjects was to listen to the stimuli and give a mean opinion score (MOS) on a scale from 1 to 5, based on the ITU-T P.800 [8] protocol. The subjects were asked to score the overall effect on a single dimension.

The experiment setup used for the database follows closely the ITU-T supplement P.23 Experiment 1 design. We include five sessions per listener (instead of the four in the P.23 setup), each containing 44 audio examples and the first being a practice session. Since results from the practice session do not differ from four sessions presented subsequently, the practice session was included in the data-analysis. A randomization of stimuli presentation order within sessions and between subjects was applied. All stimuli were presented at 60 dB SPL (fixed speech level) and listening tests were conducted in a sound-proof booth.

3. RESULTS

We present the reliability of the subjective scores obtained by the listening tests and the effect of the degradations on the MOS. In addition, we compare the performance of the PESQ algorithm using the correlation coefficient between PESQ scores and subjective quality scores. The Pearson correlation coefficient R is defined as:

$$R = \frac{\sum_i (\hat{Q}_i - \mu_{\hat{Q}})(Q_i - \mu_Q)}{\sqrt{\sum_i (\hat{Q}_i - \mu_{\hat{Q}})^2 \sum_i (Q_i - \mu_Q)^2}}, \qquad (1)$$

where $\hat{Q}$ is the estimated speech quality (also known as MOS-LQO) and Q is the subjective speech quality (also known as MOS-LQS).

3.1. Reliability of data

Principal Component Analysis (PCA) was used for assessing the inter-subject and the intra-subject reliabilities [9]. High inter-subject and intra-subject reliabilities (0.91 and 0.93 respectively) indicated that the naïve listeners gave reliable responses and these may be used to judge the measurement error on the mean opinion scores obtained in the listening tests. In addition, no learning effect was found in the analysis of the listening tests. In comparison, the P.23 English subset was found to have an intra-subject reliability of 0.89.

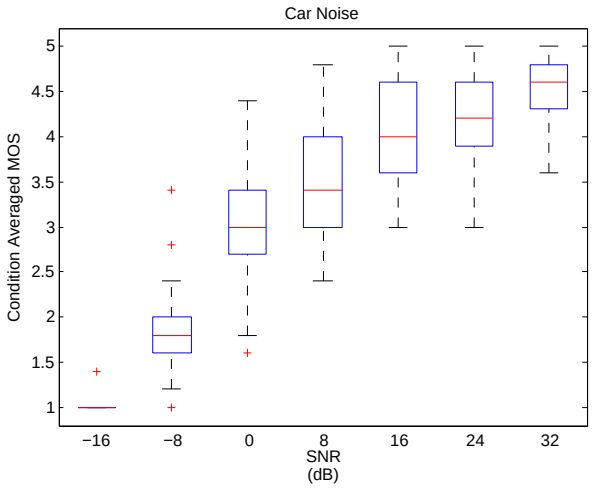

Fig. 3: Relationship between SNR and subjective MOS for car noise. Five outliers are detected over the -16 to 0 dB SNR range.

3.2. Effect of degradation on subjective quality scores

The results from the listening tests highlight the relationship between the perceived quality of speech for each degradation condition (level and type of degradation). The box-plots present the median (central line in the box), the 25th and 75th percentiles are represented as the limits of the box and the dashed lines present the extreme data points and outliers are plotted as a '+'.

Figure. 1 shows the linear relationship observed for hum noise with SNRs ranging from -30 to 30 dB. In the case of babble noise, a linear relationship is observed between SNR and subjective quality in the -8 to 16 dB range (Fig. 2). The overall relationship for babble noise resembles a psychometric curve observed in the context of intelligibility testing. A similar relationship is observed for car noise (Fig. 3).

The effect of various levels of peak clipping on subjective quality is illustrated in Fig. 4, where a linear relationship is discovered. A similar relationship is shown to exist between the effects of clicking (temporal erasures) and subjective MOS (Fig. 5). The effect of reverberation is shown in Fig. 6. It is clear that the perceived effect

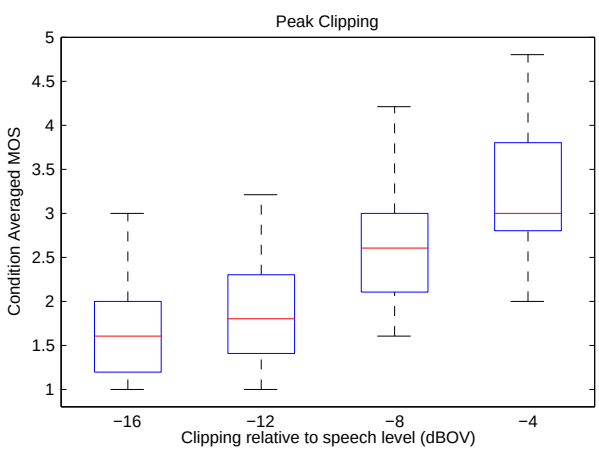

Fig. 4: Relationship between subjective MOS and different levels of peak clipping. No outliers are found for all conditions.

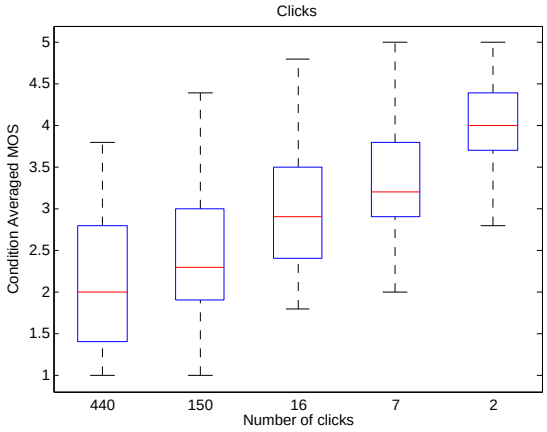

Fig. 5: A linear relationship is observed between the number and durations of clicks and subjective MOS.

of reverberation on quality for the MARDY RIR is a linear relationship with the source to microphone distance. Figure. 7 shows the condition averaged MOS obtained for the MNRU conditions. A comparison of the scores obtained by our database and the P.23 suggests that the MOS range obtained in this study has been compressed when compared with the P.23 MNRU conditions. A reason for this discrepancy could be the higher levels and types of distortions been tested in our stimuli.

3.3. Performance of PESQ

In this section we discuss the performance of wideband PESQ for the degradations encountered in law enforcement. The correlation of objective quality scores with subjective MOS obtained from the listening experiment was used to validate the performance of wideband PESQ. Overall, PESQ scores were found to give a correlation of 0.82 with subjective quality scores for our database. In the context of telecommunications, PESQ correlates 0.92 with the English subset of the P.23 database.

In addition to the overall correlation of PESQ scores with subjective MOS, we present the correlation with different classes of degradations in table. 1. It is clear PESQ correlates well for the additive noise conditions, giving a correlation of 0.94. However, for the nonlinear degradations such as reverberation and coloration, PESQ has a poor performance, resulting in a correlation of 0.17 with subjective MOS.

Degradation class	Correlation
Additive noise	0.94
Non linear distortions	0.17
MNRU	0.97

Table 1: Correlation of condition averaged PESQ scores with subjective MOS for three classes of degradations.

4. CONCLUSIONS

PESQ correlates well with degradations encountered in the field of telecommunications, however, for the level and types of degradations present in law enforcement audio, PESQ has a relatively poor overall correlation of 0.82 with subjective opinion scores. In particular, a low correlation of 0.17 is observed for non linear degradations. However, it was found that for the additive noise conditions, PESQ has a good correlation with subjective MOS of 0.94. This limits the use of wideband PESQ, in its current form for application in the context of law enforcement audio.

In order to validate PESQ, the C-Qual database of subjectively labeled mean opinion scores has been collected, which promises to be a useful tool for evaluating the performance of speech processing and collection systems deployed in the field of law enforcement and forensics audio research.

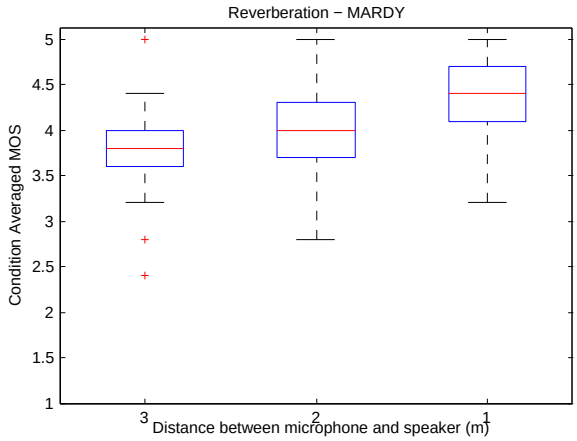

Fig. 6: Subjective MOS for reverberation using three distance for the MARDY RIR.

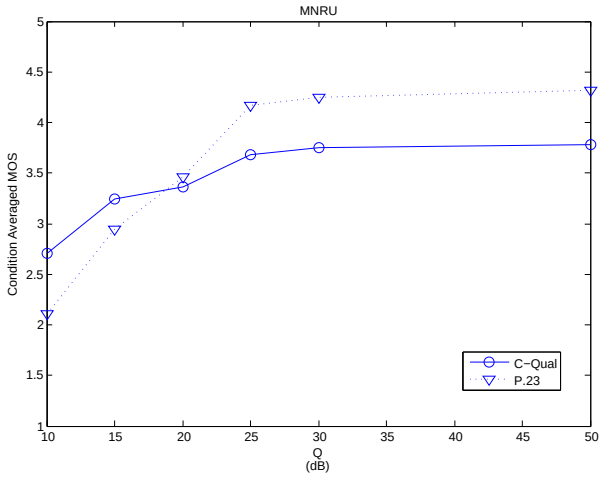

Fig. 7: The relationship between MNRU and MOS for C-Qual and P.23 databases.

5. REFERENCES

[1] A. Rix, J. Beerends, M. Hollier and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs", Proc IEEE Intl Conf Acoustics, Speech and Signal Processing, vol.2, pp. 749-752, 2001.

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

[3] ITU-T, "ITU-T coded-speech database", ITU-T Supplement P.Sup23, February 1998.

[4] ITU-T, "Objective Measurement of Active Speech Level", ITU-T Recommendation P.56, March 1993.

[5] J. Wen, N. D. Gaubitch, E. Habets, T. Myatt and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database", Proc Int. Workshop Acoust. Echo Noise Control, Paris, France, September 2006.

[6] M. R. Schroeder, "New method of measuring reverberation time", J. Acoust. Soc. Amer., vol. 37, pp. 409412, 1965.

[7] ITU-T, "Modulated Noise Reference Unit (MNRU)", ITU-T Recommendation P.810, 1996.

[8] ITU-T, "Methods for subjective determination of transmission quality", ITU-T Recommendation P.800, August 1996.

[9] N. J. Versfeld, J. M. Festen and T. Houtgast, "Preference judgements of artificial processed and hearing-aid transduced speech", J. Acoust. Soc. Amer., vol. 106, pp. 1566-1578, 1999.