Performance Analysis and Comparison of PESQ and 3SQM in Live 3G Mobile Networks

M.Goudarzi and L.Sun

School of Computing, Communications and Electronics,
University of Plymouth, Plymouth, United Kingdom
e-mail: L.Sun@plymouth.ac.uk

Abstract

The ITU-T's Perceptual Evaluation of Speech Quality (PESQ) is an intrusive objective assessment tool that has been widely used in telecommunications and IP networks. 3SQM is an ITU-T standard for single-sided non-intrusive quality measurement. The purpose of this paper is to investigate the accuracy of PESQ and 3SQM in evaluating voice quality over live 3G networks. A testbed was setup based on Asterisk to measure voice quality over 3G mobile networks. 192 voice samples from the ITU-T database were recorded via mobile phones during different times of the weekdays and the results of the objective measurements (in terms of PESQ and 3SQM) were analyzed. A further 30 samples were selected (from 192 recorded ones) and carried out informal subjective tests (with 33 subjects). The correlation of the objective and subjective results was analyzed using a 3rd order polynomial regression method. The results showed that overall PESQ (including PESQ-LQO) measures have a high correlation with subjective assessments whereas 3SQM measurements had a fair correlation. This suggests that PESQ is preferred to use for objective speech quality testing in live 3G networks when compared with 3SQM. It was also found that two individual cases in which 3SQM provided better prediction results than PESQ. It can also be noticed that the quality degradation in these cases is mainly due to loss position. It indicates that PESQ still needs to improve its performance in cases such as different loss locations. 3SQM also showed useful in identifying quality problems in Individual tests. Therefore, it is recommended that a co-existence of both measures when investigating speech quality problems in 3G mobile networks.

Keywords

Speech quality measurement, Subjective, PESQ, 3SQM

1 Introduction

There are two approaches to measuring the speech quality in telecommunication networks: *Subjective* and *Objective*. In subjective listening tests a subject hears a recorded speech processed through different network conditions and rates the quality using an opinion scale, and the MOS is then calculated as an average of all participants' scores. Subjective tests are the most reliable method for obtaining the true measurement of user's perception of voice quality and have good results in terms of correlation to the true speech quality. Traditionally, user's perception of speech quality has been measured by this method which is time-consuming and expensive, and it is impossible to use them to supervise all calls in the network. Hence, they are not suitable for monitoring live networks.

Objective models have been developed to provide machine-based automatic assessment of the speech quality score. These objective measures that can be easily automated and computerized have gained popularity and are becoming broadly used in the industry. Many field tests have shown that objective speech quality measurements can be highly useful in managing cellular networks and have necessary variety of applications in mobile networks such as daily network maintenance, benchmarking and resource management.

Intrusive measurements such as *Perceptual Evaluation of Speech Quality* (PESQ) are generally based on sending stimulus through the system under test and comparing the output signal to the original. Intrusive methods have a number of disadvantages such as consuming network capacity when used for testing live networks. More calls can be assessed if the voice quality is measured through non-intrusive methods based on single sided monitoring, by using the in-service speech signals. This is the basic principle for ITU-T's *Single Sided Speech Quality Measure* (3SQM), developed for non-intrusive voice quality testing. It is based on recommendation (ITU-T P.563). 3SQM is less reliable in terms of the correlation with subjective tests, but as a non-intrusive technique is effective in live networks since single sided measurement will not occupy any network bandwidth and is expected to become more accurate in the near future.

Since PESQ is a more popular tool and has been widely deployed in the industry, many researches have been carried out to investigate the effects of different impairments on the results of PESQ. The effects of packet loss in VoIP networks have been investigated by (Hoene and Enhtuya, 2004). However it only focuses on the impact of packet loss in simulated VoIP environment, which may not properly model the signal characteristics during the normal operation of a mobile network. The performance of PESQ for various audio features and codecs has been studied in the reports by (QUALCOMM, 2008) and (Ditech, 2007). Also a detailed case study of the defects of PESQ time alignment features in the presence of silence gap and speech sample removal or insertion due to packet loss concealment and jitter buffer adjustment in mobile devices has been carried out by (Qiao *et al.*, 2008).

Although PESQ is state-of-the-art in terms of the objective prediction of perceived quality and is claimed to have the highest correlation with the subjective measurements, by looking at a number of published case studies and reports, it can be seen that there is still work to be done in the area of objective quality measurement. PESQ does not always predict perceived quality in live network accurately, as a result of improper time-alignment as reported by (Qiao *et al.*, 2008). (Ditech, 2007) also reports that there are significant known limitations to the PESQ algorithm with regards to its time alignment and psychoacoustics model. Furthermore PESQ has not been validated for many methods commonly used in live networks to enhance the quality such as noise suppression or echo cancelling (QUALCOMM, 2008); Packet loss concealment and adaptive Jitter buffer are also examples of such methods.

It should also be noted that neither PESQ nor 3SQM algorithms provides a comprehensive evaluation of transmission quality, and only the effects of one-way speech distortion and noise on speech quality are measured using these objective

methods. Factors such as loudness loss, delay, sidetone, echo, and other impairments related to two-way interaction are not reflected in the quality scores given by these models (ITU-T, 2001).

Further research can be done in the area to pinpoint the flaws and strengths of each objective model that can help to further improve the accuracy of the model or may lead to the development of new, more accurate objective measurement techniques.

Moreover, assessing how and under which conditions these methods may be more accurate, and comparing the accuracy of each algorithm under real live mobile environments is an essential issue, in order to improve the performance of speech quality measurement techniques.

The main objective of this paper is to investigate and evaluate the accuracy of PESQ and 3SQM objective measurement models in a live wireless 3G mobile network. Objective testing was carried out through our testbed platform and the correlation of the results with subjective votes was analyzed. The Individual cases in which PESQ did not predict accurate quality scores were also analyzed.

The remainder of the paper is structured as follows. In Section 2, the quality measurement platform and the methods of objective and subjective measurements are presented. In Section 3, the correlation of the objective and subjective results, plus two individual cases has been investigated. The conclusion is presented in Section 4.

2 Methodology

2.1 Objective measurement test platform

In order to objectively measure user perceived speech quality in a live telecommunication conversation, a speech quality test platform was set up to mediate between calls from 3G mobile network and the quality test equipment. As can be seen in Figure 1, the platform consists of a voice server based on Asterisk connected to the mobile network via an ISDN interface, 3G mobile handsets, monitor PC and a local pc, all networked together in an IP environment.

Using the quality test platform, calls placed or received via the voice server could be recorded on the server, or the sample speech files could be played back on the channel and recorded from the mobile on the monitor pc using the microphone and line-in of the soundcard connected to the mobile handset. Alternatively, calls could be forwarded to a SIP client for experiments with SIP, which is out of the scope of this paper.

(ITU-I P862.3) provides guidance and considerations for the source materials that will be used in speech quality tests. Reference speech should contain pairs of sentences separated by silence. It is also recommended that the reference speech should include a few continuous utterances rather than many short utterances of speech such as rapid counting. ITU-T P.862 also suggests that signals of 8-12*s* long should be used for the experiments.

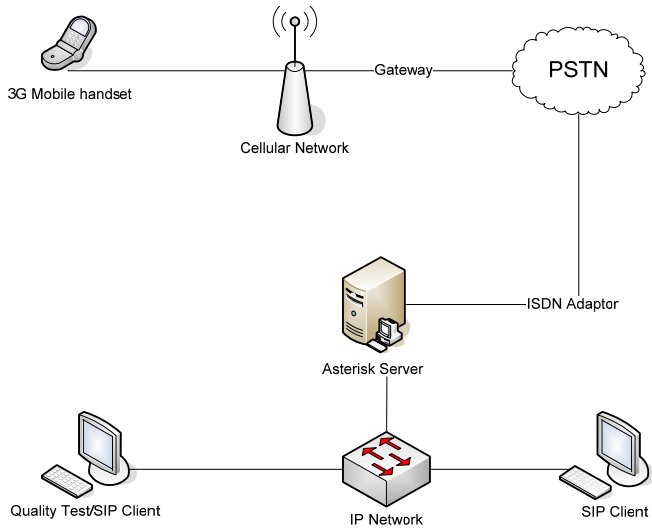

Figure 1- Testbed platform for speech quality measurement

16 British English speech samples (8 male and 8 female), from (ITU-T P.50) database were used for all the subjective and objective measurements. Samples are each 6-8*s* in length. All the speech samples were converted to 8000 sampling rate. GSM and AMR codecs were employed as the main voice codecs used in mobile networks. 192 recording were made during *week days* and at three different times of the day. Figure 2 illustrates the block diagram of the experiments with PESQ algorithm.

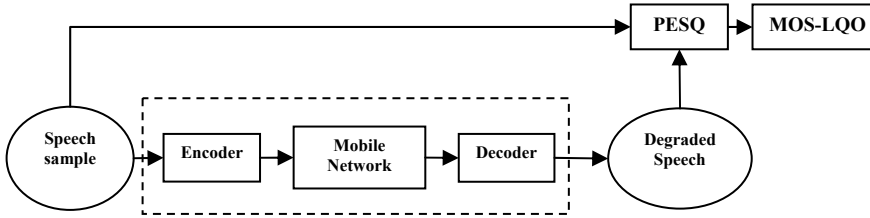

Figure 2: PESQ speech quality evaluation setup

2.2 Informal subjective test for evaluating the accuracy of objective models

The purpose of conducting an informal subjective quality test was to validate the applicability of objective algorithms for assessing the quality of the speech samples. The main criteria in selecting the speech samples used in the subjective test, was the difference between the PESQ and 3SQM of the results. 30 samples with the highest difference in their PESQ and 3SQM results were selected from 192 previously recorded samples, some female and some male (12 male, 18 female), 17 of which were GSM encoded samples and 13 were AMR-encoded samples. The subjective test material involved 8 samples from ITU-T P.50 database with different recording time and conditions. Efforts have been made for the test to conform to the ITU-T

standards for subjective evaluation of voice quality in telephone networks (ITU-T P.800) in terms of the test procedure and eligibility of the participants. Score sheets with instructions and voice files were sent out to colleagues and friends; and the results were collected by e-mail from the subjects. 33 subjects completed the subjective test.

In order to scale the objective scores onto the same scale as the subjective votes, relationship between PESQ and 3SQM scores and subjective MOS is modelled using a monotonic 3rd order polynomial mapping function. The closeness of fit between the objective and the subjective scores may be measured and analysed after the mapping function has been applied.

3 Experimental results

Upon receiving all the score sheets from the subjects, the average of subjective scores given by the participants was calculated for each file to achieve the MOS-LQS. The standard deviation for subjective results ranges between 0.7 to 1.01 and less than 1 for most of the cases. This indicates that the individual results differ quite significantly from subject to subject. Table compares the average results of PESQ, PESQ-LQO and 3SQM with the results of the informal subjective test. Preliminary comparisons between the subjective and objective results showed that in general, 3SQM results have a higher variation.

Codec	PESQMOS	PESQ-LQO	3SQM	MOS-LQS
GSM	3.007941	2.842765	2.406109	2.889483
AMR	3.162538	3.063692	4.164516	3.843823
ALL	3.074933	2.938500	3.168085	3.30303

Table 1: average quality scores of objective and subjective experiments

3.1 Differences between PESQ and 3SQM measurements

Although by looking at the overall results of the objective and subjective tests, the results of both objective methods are linked with the subjective results, some differences were found between the scoring of PESQ and 3SQM. Within both groups of samples (AMR and GSM) gender of the talker showed to have an effect on the perceived speech quality. For PESQ algorithm, MOS score for male talkers tends to be higher than that of female talkers. This result is more consistent with the literature. One reason could be higher average frequency of the female voice which causes a lower quality in the encoding process (Holub and Street, 2004). Also (Lingfen and Ifeacher, 2001) reported the effect of the gender on PESQ results. However, experiments with 3SQM algorithm showed opposite results, which mean relatively better MOS scores for female samples. Figure 3 shows the gender effect on the PESQ and 3SQM scores for GSM samples. The average PESQ score for male samples is more than 3, whereas in 3SQM results male samples have an average of around 2.6. This results shows that there is a difference in the perceptual model of these two algorithms.

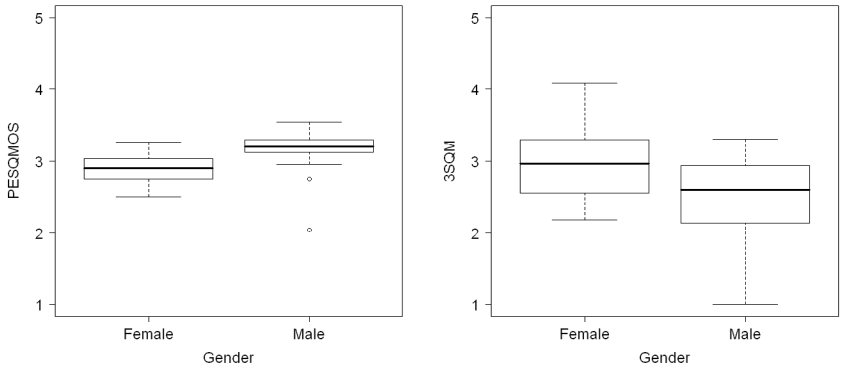

Figure 3: PESQ and 3SQM scores for male and female GSM samples

Also individual cases that showed a major difference between PESQ and 3SQM results were studied. There were 2 cases in which PESQ generated a significantly less accurate results compared to their respective 3SQM results. Also in 5 cases 3SQM results were not adequately accurate and PESQ predicted the quality score more accurately.

Figure 4: Inaccurate PESQ result , loss positions in the degraded speech

By comparing the wave forms of the original and degraded speech files for these individual cases, some of the low PESQ MOS scores are clearly the result of packet loss or bad signal conditions. The waveform for one of the GSM samples with the highest difference between its 3SQM and PESQ quality scores is shown in Figure 4. As indicated in the figure, many parts of the signal, particularly at the beginning of the recording are lost. Also listening tests resulted in a low score (=1.455) for this sample which is reasonable because the beginning of the speech is not intelligible. Therefore regarding it as a "bad" sample with a score of 1-1.5 is reasonable. However PESQ gave a score of 3.302 to this sample (3SQM=1.432). These results show that the position of loss has an effect on the quality score give by PESQ and also the subjects.

On the other hand, for 5 other samples the reason for such low scores is not so obvious. The same sample recorded in another time, showed to have a fairly well waveform did not seem to have a very low MOS score. However, 3SQM score was 1.43 and PESQ score was 3.078 and subjective MOS was 3.455.

3.2 Correlation of objective and subjective measurements

As can be seen in Table 2, correlation of subjective results with the results from objective tests shows the accuracy of each objective measure in predicting the quality of the speech samples.

	PESQ	PESQ-LQO	3SQM
Correlation	0.9433	0.8911	0.5193
Max. difference	0.412	0.591	1.319
Min difference	-0.500	-0.697	-2.054
RMSE	0.237	0.343	1.108

Table 2: statistical summary of objective vs. subjective results

Figure 5 shows the scatter plots of the objective versus subjective results before and after mapping. The correlation coefficient for PESQ results after the mapping is 0.9433, which shows a high degree of correlation between objective and subjective results. PESQ-LQO scores also had a good correlation (correlation coefficient= 0.8911). 3SQM, however, had a lower correlation of 0.5193, which shows a lower level of correlation between 3SQM results and the informal subjective test results.

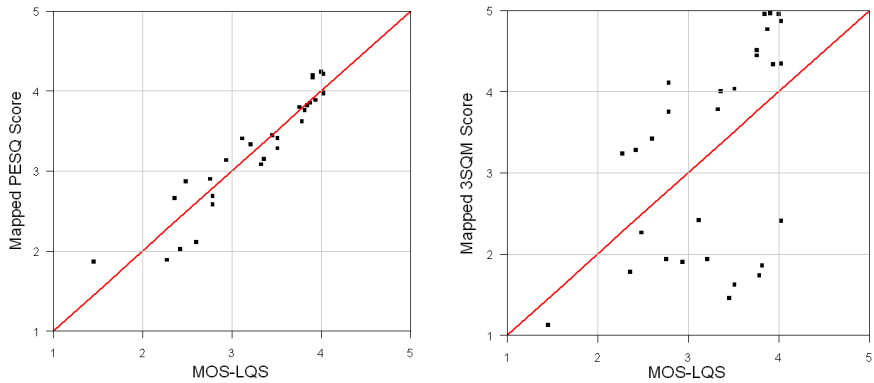

Figure 5: Mapped PESQ and 3SQM results vs. subjective MOS

4 Conclusions

The present study evaluated two objective measures commonly used for evaluating speech quality. The test conditions included real live mobile environments and GSM and AMR codecs as the main voice codecs used in mobile networks have been employed. Around 200 recording were made during weekdays and at different times of the day.

The comparison between subjective and objective results showed that PESQ and PESQ-LQO measures have a good correlation with the subjective results and according to our results PESQ can be used reliably for objective speech quality measurements in live 3G networks. Though in individual cases 3SQM showed to

have a more accurate prediction. In two cases PESQ algorithm seemed to be less accurate compared to 3SQM, one of which is mainly due to the loss position.

3SQM as non-intrusive test method could not supersede intrusive analysis as expected; since it lacks the information from the reference signal. However, it showed useful in identifying quality problems in individual tests and as a non-intrusive measurement has advantages in live telecommunication networks. Therefore, it is recommended that a co-existence of both measures when investigating speech quality in 3G mobile networks.

5 References

Ditech. (2007) *Limitations of PESQ for Measuring Voice Quality in Mobile and VoIP Networks*. http://www.ditechnetworks.com/learningcenter/whitepapers/WP_PESQ_Limitations.pdf, (Accessed: 18/7/2008).

Hoene, C. and Enhtuya, D-L (2004) *Predicting Performance of PESQ in Case of Single Frame Losses*. Proceeding of MESAQIN 2004. Prague,CZ.

Holub, J. and Street, M. D. (2004) *Impact of end to end encryption on GSM speech transmission quality - a case study*. Secure Mobile Communications Forum: Exploring the Technical Challenges in Secure GSM and WLAN, 2004. The 2nd IEE (Ref. No. 2004/10660). pp 6/1-6/4.

ITU-T (1999) Artificial voices *ITU-T Recommendation P.50,September 1999.*

ITU-T (2001) 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 Recommendation P.862,February 2001.*

ITU-T (2004) Single-ended method for objective speech quality assessment in narrow-band telephony applications *ITU-T Recommendation P.563,May 2004*.

ITU-T (2007) Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2 *ITU-T Recommendation P.862.3,November 2007.*

Qiao, Z., Sun, L. and Ifeachor, E. (2008) Case Study of PESQ Performance in Live Wireless Mobile VoIP Environment. IEEE PIMRC 2008. Cannes, France.

QUALCOMM. (2008) *PESQ Limitations for EVRC Family of Narrowband and Wideband Speech Codecs:* http://www.qualcomm.com/common/documents/white_papers/PESQ_Limitations_Rev_C_Jan_08.pdf, (Accessed: 1/7/2008).

Sun, L. and Ifeacher, E.C. (2001) *Perceived Speech Quality Prediction for Voice over IP-based Networks*. Proceedings of the 2nd IP-Telephony Workshop (IPTEL '01). Columbia University, New York.