PESQ and 3SQM measurement of voice quality over live 3G networks

Mohammad Goudarzi, Lingfen Sun, Emmanuel Ifeachor
Centre of Signal Processing and Multimedia Communications
School of Computing, Communications and Electronics, University of Plymouth, UK
Email: {mgoudarzi, lsun, eifeachor}@plymouth.ac.uk

Abstract

The purpose of this paper is to investigate the accuracy of PESQ and 3SQM in predicting voice quality over live 3G networks. We set up an Asterisk-based testbed to measure voice quality over 3G mobile networks. 192 experiments with speech samples from the ITU-T database were recorded via 3G mobile phones and the results of the objective measurements were analyzed. We further selected 30 samples (from 192 recorded ones) and carried out informal subjective tests (with 33 subjects). The results showed that overall, PESQ has a very good correlation with subjective assessments and therefore is a reliable tool for measuring voice quality in mobile networks, whereas 3SQM measurements had a fair correlation. However individual prediction errors showed that both PESQ and 3SQM tend to over-predict the voice quality when subjective MOS score is over 3 and under-predict for lower quality samples. We have also investigated the effect of the talker's gender on the behaviour of objective measurements. The work should help to improve the performance and accuracy of objective speech quality measurement for voice over 3G networks.

Keywords

Speech quality, MOS, PESQ, 3SQM, AMR, GSM, 3G, gender

1 Introduction

Speech quality is the most visible and important aspect of quality of service (QoS) in mobile, telecommunications and VoIP networks. Communications networks are rapidly developing, new services are emerging and 3G cellular networks are widely available. Service providers are challenged with finding the root causes of voice quality problems over their heterogeneous networks and the ability to monitor and design this quality has become a main concern.

There are two approaches to measuring the perceived 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. The Mean Opinion Score (MOS) is a widely used subjective measure of speech quality recommended by ITU-T [1]. It is known to be the most reliable method for measurement of user's perceived quality. However, such methods are 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 in managing cellular networks and have found variety of applications in mobile networks such as daily network monitoring, maintenance, and even resource management. Objective measurements can be classified as intrusive and non-intrusive.

Intrusive objective measurements such as the latest ITU-T's *Perceptual Evaluation of Speech Quality* (PESQ) [2] are generally based on sending reference signal through the system, then comparing the original unprocessed signal with the degraded version at the output. PESQ will then produce a quality score MOS-LQO (Listening Quality Objective) based on the result of this comparison [3]. PESQ is a popular tool and has been widely deployed in the industry. Many studies 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 in [4]. 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 [5] and [6]. 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 [7]. Although PESQ is the state-of-the-art in objective speech quality measurement, many studies have shown that PESQ does not always predict perceived quality accurately. In [7] the inaccuracy of PESQ measurements as a result of improper time-alignment is investigated. Also [6] reports 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, echo cancelling [5] or transcended speech [8].

ITU-T's *Single Sided Speech Quality Measure* (3SQM), is developed for non-intrusive voice quality testing. It is based on recommendation (ITU-T P.563) [9]. 3SQM combines three non-intrusive algorithms and calculates a MOS-like quality score based on 12 different signal-based parameters. The accuracy of 3SQM has received little attention in the literature not only because of its lower correlation with subjective results, but also due to the processing time and power required by the algorithm. 3SQM's correlation coefficient with subjective listening tests is reported around 0.8 by ITU-T recommendation P.563. This suggests that 3SQM is less reliable in terms of the correlation with subjective tests, but as a non-intrusive technique it is more effective in live network monitoring, as single sided measurement can use in-service signal and will not occupy any network bandwidth. Signal-based non-intrusive measures are expected to become more accurate in the near future. Some works have compared the behaviour of 3SQM and PESQ in different network conditions. In [8], the reliability of PESQ and 3SQM methods in measurement of speech quality in case of transcended speech samples was assessed. The results indicated that neither of the models is reliable enough to be used independently in case of multiple coder tandeming. But after regression PESQ results showed a better performance than that of 3SQM. In [10] the impact of different noise types on the quality assessment in VOIP scenarios was investigated. Both PESQ and 3SQM showed a good correlation with subjective results, and 3SQM seemed to have a better conformity when comparing the residual errors. However, since the study looked at the SNR of different noises as the variable and SNR is also a dominant parameter in the 3SQM calculation algorithm, the results were expected to be more in favour of 3SQM than PESQ.

It should also be noted that neither of these two models 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, round trip delay, sidetone, echo, and other impairments related to two-way interaction are not reflected in the quality scores given by these models [2]. Also [11] states that although the data fitting procedure used recommended by ITU-T minimizes prediction errors that the PESQ algorithm should not be penalized for, it also may minimize prediction errors that PESQ should be penalized for, which might result in overoptimistic measures of accuracy for objective measurements.

It has been established in many works such as [11-14] that the accuracy of objective measurements is highly dependent on the nature of the network connection and the system under test. The sensitivity of objective model may be dissimilar to human subjects in different system conditions and their performance may therefore vary according to the network condition. Signal-based objective methods are now being widely used in many aspects of UMTS networks ranging from setting service levels and design [15] to resource management and daily maintenance [16]. Also many studies have used PESQ for evaluating the quality of various codecs [17] and voice applications in UMTS networks [18]. Considering that current UMTS networks commonly use methods that may not be known to the algorithms (e.g. noise suppression and echo cancelling) that can affect the performance of these signal based algorithms, it is necessary to evaluate the usability of such methods in estimating the "per call" and "overall" quality in a real 3G mobile environment.

We focus on the behaviour of PESQ and 3SQM models in 3G mobile network, using our speech quality testbed and two public 3G (UMTS) networks. We set up a testbed for evaluating speech quality in a live 3G mobile environment and tested over 200 mobile to PSTN calls. We further assessed the accuracy of the objective results by comparing some of the results with subjective assessments. The results showed that overall, PESQ have a very good correlation with subjective assessments and therefore is a reliable tool for measuring overall quality in mobile networks, whereas 3SQM measurements had a fair correlation.

The rest of the paper is structured as follows. In Section 2, the quality measurement platform and the methods of objective and subjective measurements are presented. The experimental results are presented and discussed in section 3 and 4. Section 5 concludes the paper and suggests some future investigations.

2 Test platform and experimental scenario

2.1 Objective speech quality test platform

In order to objectively measure user perceived speech quality in a mobile conversation we set up a speech quality test platform. As can be seen in Figure 1 the aim of the test platform is to provide the necessary network connectivity to objectively measure user perceived quality in a 3G mobile network. The platform consists of a voice server based on Asterisk, connected to the mobile network. To be able to receive and make 3G voice and video calls and handle them in the asterisk dialplan, we have used the H324m library which adds 3G call support to Asterisk and allows it to bridge UMTS 3G mobile calls via an ISDN interface. We also installed the AMR-NB codec on top of the H324m stack to enable AMR support for 3G calls.

To perform play and record operations from/to the mobile handset, the handset needs to be connected to the monitoring PC. An electrical cable is required to replace the air interface of the handset so that instead of hearing the sample from the earpiece and playing the sample from the microphone, the samples are played and recorded directly through the soundcard.

The sound card used for playing and recording speech samples needs to be a high quality soundcard to avoid unwanted noise, gaps and other distortions introduced by normal soundcards. Also the software used must be reliable and tested to ensure that they do not introduce unwanted distortions to the speech samples. The cable used to connect the mobile handset to the soundcard was also loop-tested with the software and soundcard to make sure that no distortions are introduced by the hardware or the software when playing or recording the speech samples.

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 any selected channel and recorded from the mobile on the quality test 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. In the experiments described in this research paper, the calls from the 3G handset to the PSTN line were directed to the monitoring PC - where the degraded speech was be recorded and the quality score was be objectively measured using PESQ and 3SQM. The same recorded sample speech files were later used in the subjective measurement as described in section 0.

2.2 Reference speech signals

ITU-T P862.3 provides guidance and considerations for the source materials used in objective 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.

In our experiments, we have used 16 British English speech samples (8 male and 8 female), from ITU-T P.50 database [19] for all the objective measurements. Samples are each 6-8*s* in length as seen in Table 1.

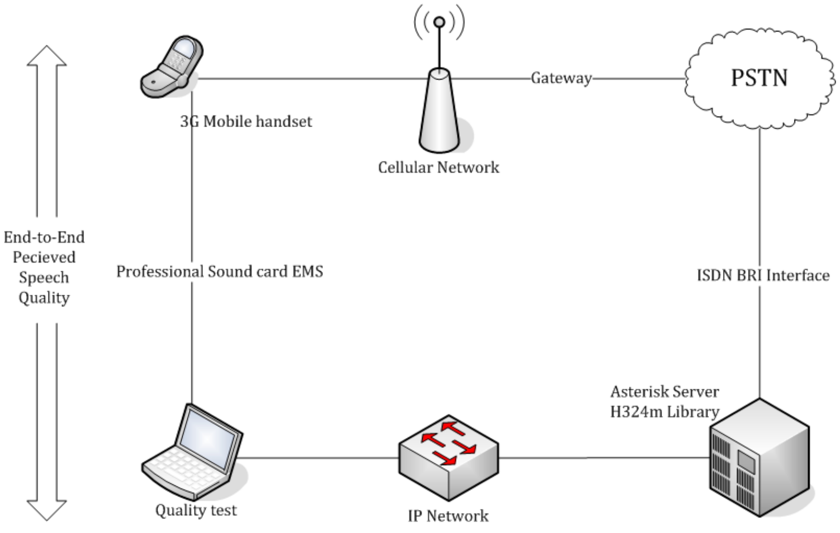

Figure 1- Objective speech quality test platform

Table 1-Length of the reference speech samples

Reference	Length	Reference	Length
Female1	7.2s	Male1	5.4s
Female2	7.3s	Male2	7.0s
Female3	7.2s	Male3	8.3s
Female4	7.2s	Male4	7.9s
Female5	7.6s	Male5	10s
Female6	7.2s	Male6	8.3s
Female7	6.5s	Male7	7.1s
Female8	8.1s	Male8	6.6s

2.3 Objective testing

The reference signals described in section 0 are used for transmission through the live 3G mobile network. Sampling rate of 8000 was used for all the speech samples in the study. GSM and AMR voice codecs were employed as the main voice coders used in mobile networks. Live recordings were made during *week days* and at three different times of the day. Overall over 200 samples were recorded in 6 different times for each codec.

For PESQ measurements, the raw PESQ score for speech quality was first measured and then converted to achieve the MOS-LQO (listening quality objective) using equation (1).

$$LQO = 0.999 + \frac{4.999 - 0.999}{1 + e^{(-1.4945 \times PESQ\ + 4.6607)}} \qquad (1)$$

2.4 Subjective testing

The subjective test conducted in this research project was an informal subjective test meaning that the test was carried out in an un-controlled environment and using standard quality audio facilities.

33 participants completed the informal subjective test. The subjects were all eligible according to ITU-T P.800 recommendation as none of them had been involved with the works connected to assessment of voice quality, and had not participated in any other subjective tests for the past 6 months. 39% of the subjects were female and 61% were male. Basic information gathered from the participants showed that the majority of the subjects aged between 21 and 30 years old. Also 3 out of 33 used speakers to listen the samples and the rest used earphones to complete the experiment.

As the purpose of the subjective test was to investigate the accuracy of PESQ and 3SQM results, the main criteria in selecting speech samples used in the informal subjective test was the difference between the PESQ and 3SQM predicted MOS of the samples. The samples that had the highest difference between their MOS-LQO and 3SQM scores were selected and used as the source material for this subjective test. Also, in order to further investigate the gender effect discussed in later, samples were selected from both genders (12 male, 18 female). Overall 30 samples (17 GSM and 13 AMR) with different recording time and conditions were used from 192 previously used for objective tests for the subjective experiment.

For the test procedure, 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) [1]. Instruction sheets and score sheets were compiled based on the guidelines provided by ITU-T P.800. Score sheets with instructions and voice files were sent out to the subjects; and the results were collected by e-mail.

In order to compare the results of objective and subjective tests, a regression mapping on a per study basis should be applied before comparing objective and subjective results [2]. This regression mapping compensates for the systematic differences between subjective and objective results. The mapping function that ITU-T has accepted for the evaluation of objective models (both PESQ and 3SQM) is a monotonically constrained 3rd order polynomial. This is applied, for each subjective test, to map the objective score onto the subjective score. The mapping should be constrained to be monotonic across the range of the data [20] or it may not preserve the ordering of the objective scores. It is then possible to calculate correlation coefficient and residual errors. The process is usually performed per condition, but it can also be applied per file. [21]

To express the correlation between the subjective and objective methods we used the person's correlation coefficient calculated by (2).

$$r = \frac{\sum(x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (2)$$

To further analyze the differences between the results of tests we used residual errors (3) described in more detail in the following section.

$$e_i = x_i - y_i \qquad (3)$$

3 Experimental results

In this section we present and explain experimental results for objective and subjective assessments and their comparison is discussed in more details.

3.1 Objective assessment results

To investigate the quality degradation after encoding/decoding of the speech samples, we first conducted an experiment using GSM and AMR encoder/decoder and studied the results. The average quality scores and standard deviations for AMR and GSM codecs are presented in Table 2.

Table 2- PESQ and 3SQM scores after encoding and decoding

		PESQ-LQO	3SQM
GSM	MOS	3.61	2.94
	STDEV	0.31	0.46
AMR	MOS	4.09	3.31
	STDEV	0.11	0.61

As it can be seen in Figure 2, the quality score of objective tests for all female talkers are lower than that of male talkers in both AMR and GSM experiments. Generally, higher average frequency of the female voice can cause a lower quality in the encoding process of most codecs. Similar results are reported in [22] for GSM codec and also examined for more codecs in [23]. If the gender of the talker only affects the codec functionality, the results of the samples after transmission through live network should be similar for male and female speech samples.

The live recording experiments were carried out using GSM and AMR encoded samples during week days at different times of the day to take account of different network conditions in the normal operation of a live network. The results obtained by objective methods (PESQ and 3SQM) are summarized in Table 3.

Overall, AMR codec (12.2 Kbit/s bit rate) showed higher quality scores when compared to GSM codec results. PESQ results had a higher average quality score and lower variations. 3SQM had higher standard deviation which shows higher variations in its quality scores. 3SQM also had a relatively better average quality in case of AMR experiments.

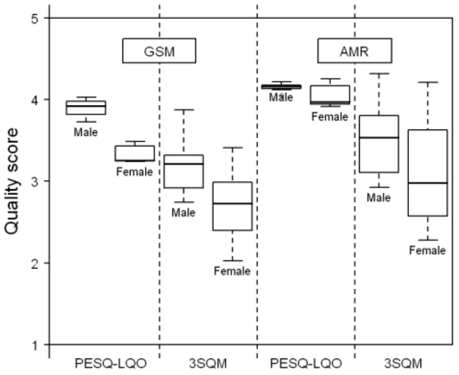

Figure 2-Comparison of objective measures for pure GSM and AMR codec experiments

Table 3-MOS and standard deviations of objective measurement results

		PESQ-LQO	3SQM
GSM	MOS	2.88	2.73
	STDEV	0.37	0.58
AMR	MOS	3.20	3.42
	STDEV	0.37	0.55
ALL	MOS	3.14	3.07
	STDEV	0.28	0.66

Figure 3 compares the quality scores obtained by PESQ and 3SQM from live GSM experiments. For the first 4 conditions, there is an agreement between PESQ and 3SQM in terms of comparing the overall quality of the network. Although 3SQM results are lower than the PESQ results, when PESQ shows that the quality is lower, so does 3SQM. But for the last two conditions, the results of PESQ show that the average quality is going down and then up, whereas 3SQM results show a raise and then a drop in the quality. These contradictory behaviours in case of different network conditions indicate possible over- or under-estimations of the perceived quality in certain network conditions.

The results of objective measurements for AMR experiments also show more of such disagreements as can be seen in Figure 4. Such disagreements between 3SQM and PESQ results are interesting since they represent conditions that one of the algorithms can lead to wrong decisions regarding the overall quality of a network condition especially if used for benchmarking in a live network. The accuracy of the models should be investigated by comparing the results with subjective quality measurements.

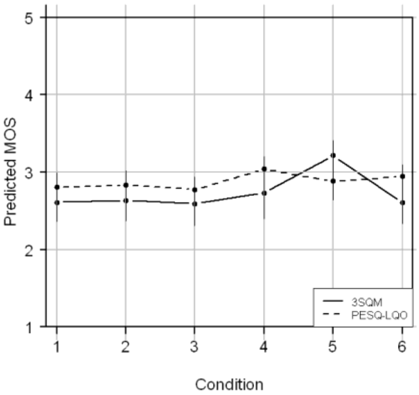

Figure 3-Behaviour of PESQ and 3SQM in live GSM experiments

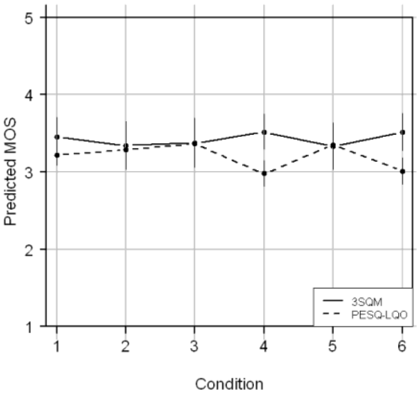

Figure 4- Behaviour of PESQ and 3SQM in live AMR experiments

Further, we compared the PESQ and 3SQM samples for male and female samples separately. The results obtained by PESQ from live GSM recordings are shown in Figure 5. For all testing conditions the male samples have a higher average quality score than female samples. The average quality score obtained by PESQ was 2.88 (3.10 for 2.65male and for female talkers) which shows a drop of around 0.7 due to transmission (Not considering the effect of the encoding/decoding). In case of PESQ, the behaviour towards the gender of the samples remained similar to the results of objective measurements after encoding the samples to GSM.

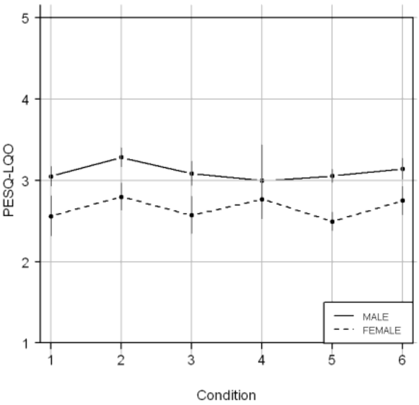

Figure 5- PESQ-LQO quality score for male and female talkers in GSM experiments

However the results of 3SQM for live GSM experiments shown in Figure 6, show systematic higher quality scores for female samples. This shows that the gender of the talker has had an effect the results of 3SQM algorithm after being recorded through live mobile network. This is rather an interesting result since the higher quality score for male samples was considered to be due to the way that GSM codec functions. But this results show that it can also be dependent on the measurement algorithm and not only the codec in case of 3SQM measurements.

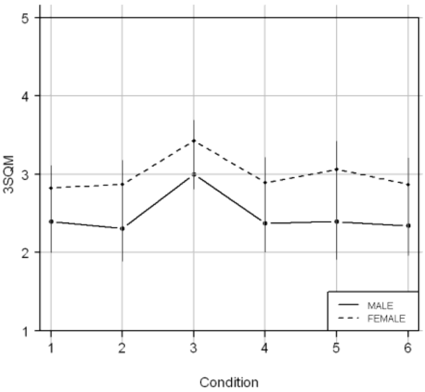

Figure 6- 3SQM quality score for male and female talkers in live GSM experiments

Similar results were obtained from experiments with AMR encoded samples. The box and whiskers plot in Figure 7 compares the average quality scores given by PESQ and 3SQM for AMR experiments. The average PESQ score for male samples is approximately close to the maximum score observed for female samples. On the other hand 3SQM gives better quality for female samples whereas PESQ shows that male samples had better perceived quality. In order to establish the accuracy of 3SQM and PESQ towards the talker's gender, these results need to be compared with the results of the subjective measurements.

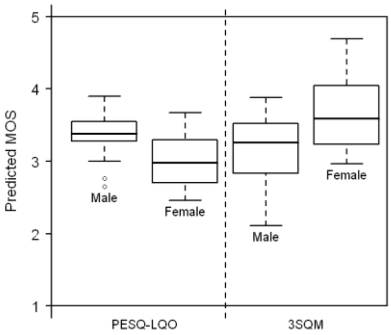

Figure 7- PESQ and 3SQM results by gender for live AMR experiments

3.2 Subjective assessment results

The MOS scores from subjective experiments are calculated by taking the average opinion scores given by the participants for each file. Table 4 summarises the results of the informal subjective test and the respective quality score of the samples produced by the PESQ and 3SQM models.

Table 4-Results of subjective test

	PESQ-LQO	3SQM	MOS-LQS
MOS	2.93	3.16	3.30
STDEV	0.35	1.22	0.67

The standard deviation for subjective results ranges between 0.7 to 1.01 and less than 1 for most for most of the cases. This indicates that the individual results differ quite significantly from subject to subject. Nevertheless, people have quite different opinions and expectations when it comes to the quality, hence such variations in subjective results were expected. Preliminary comparisons showed that overall there is an agreement between the results of subjective and objective tests. In general when the quality perceived by subjects goes up or down, so do the PESQ and 3SQM scores.

3.3 Comparison with subjective results

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(as recommended in ITU-T P.862). The closeness of fit between the objective and the subjective scores may be measured and analysed after the mapping function has been applied. The results obtained from our subjective listening tests are compared with the PESQ-LQO results after 3rd order regression in Figure 8.

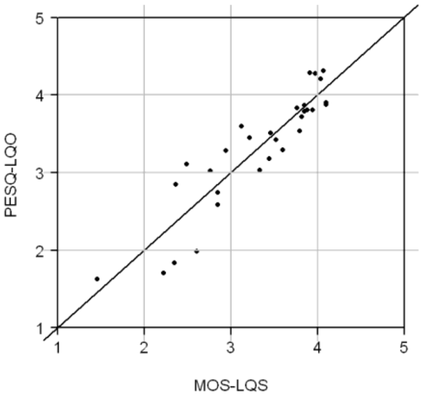

Figure 8- Subjective results (MOS-LQS) versus mapped (3rd order regression) PESQ results

For 3SQM tests, attempts to use 3rd or 2nd order polynomial regression resulted in non-monotonic mapping. Therefore 1st order regression was used instead for scale fitting of 3SQM test results. The results obtained from our subjective listening tests are compared with the PESQ-LQO results after 1st order regression in Figure 9.

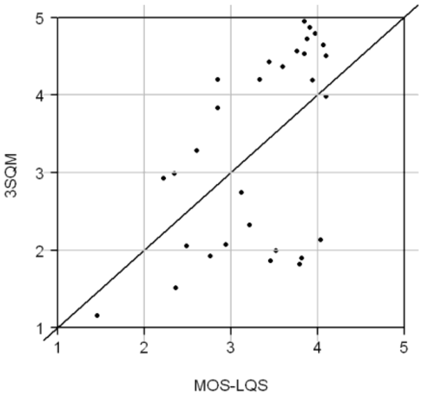

Figure 9- Subjective results (MOS-LQS) versus mapped (1st order regression) 3SQM results

The correlation coefficient, Root Mean Square Error mean residual error and the maximum/minimum prediction errors for subjective and objective methods are presented in Table 5.

Table 5-Performance measures of objective models before and after polynomial regression

Measure	PESQ-LQO	3SQM
Before regression		
Correlation	0.618	0.5193
Max. error	0.687	1.301
Min error	-1.071	-2.302
Mean error	-0.391	-0.045
RMSE	0.658	1.035
After regression		
Correlation	0.922	0.552
Max. error	0.441	1.181
Min error	-0.426	-2.127
Mean error	4.80e-12	6.01e-12
RMSE	0.276	0.999

The correlation coefficient after the 3rd order polynomial regression for PESQ is very good and shows good correspondence among subjective and objective results. 3SQM has a fair correlation which shows less conformity with the subjective tests. 3SQM also shows wider differences between the predicted and the subjective MOS. Nevertheless, even in case of PESQ, when comparing the results of individual files, differences ranged between -1.071 and 0.687 among the results of objective and subjective assessments are noticeable (Especially considering this is a 5 point scale). This is because the Pearson's correlation coefficient measures the similarity of the curve shape and the accuracy for individual components is not taken into consideration. The differences can be more analysed using residual errors. Residual error is the difference between the objectively predicted MOS and the MOS achieved from the subjective tests [11].

Figure 10 shows the residual errors from PESQ tests by MOS. Each sample is also labelled with its respective talker's gender on the plot. If the objective method accurately predicted the quality score, the residual error would be close to zero and all the residuals would fall around the horizontal axis in the scatter plot. However, it can be seen that PESQ scores differed from MOS scores by more than ± 0.2. Only 5 out of 30 of the samples have an absolute residual error less than 0.2. The PESQ scores are more under-estimated than over-estimated (over 75% under-estimated). As can be seen by the trend of residual errors in, PESQ shows a tendency to under-predict the quality score for high quality samples and for low quality samples the MOS is over-predicted.

The residual values by MOS for 3SQM are shown in Figure 11. The 3SQM scores are more over-estimated compared to PESQ (over 60% over-estimated). As it can be seen by the distribution of the male and female samples in the plot, there is a dependency between the talker's gender and the prediction errors of the 3SQM algorithm. Almost all of the Male samples are underestimated and all the female samples are over-predicted. The areas that 3SQM predicts the subjective MOS more accurately are the areas where residual errors are closer to zero. Thus, 3SQM has estimated the quality more accurately for female samples of higher quality and for male samples of lower quality (MOS-LQS). 3SQM results also show the over- and under-estimation behaviour within each gender group.

In both PESQ and 3SQM cases, it can be seen that the voice quality difference between male and female voice is systematic and gender has a significant effect on the quality of speech measured by both objective models. The female and male samples are clearly separated and for PESQ results the separation is more visible in lower qualities.

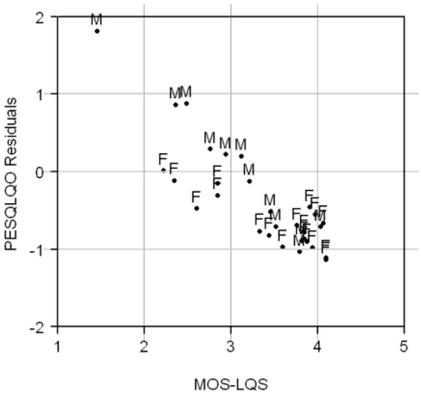

Figure 10- PESQ residual errors by subjective MOS

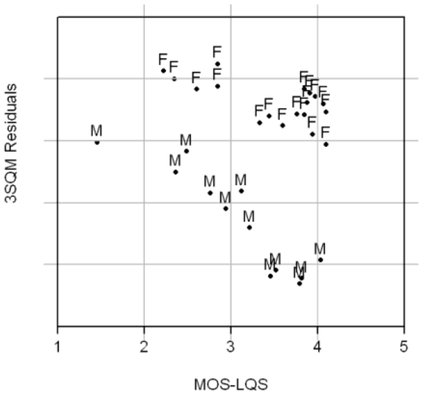

Figure 11- 3SQM residual errors by subjective MOS

Figure 12 shows the comparison of subjective and objective results for male and female samples. The average results obtained by PESQ were very close to the subjective results only for male samples, and under-predicted for female samples. The difference between subjective and PESQ results is approximately 0.8 MOS for female samples. On the contrary, 3SQM scores for male samples are under-estimated and the results for female samples are closer to subjective results.

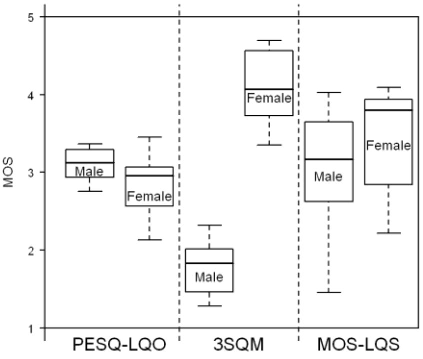

Figure 12- Comparison of subjective and objective results for male and female samples

Overall, the correlation between the results of our objective and subjective results shows that PESQ is a reliable tool for testing the quality and identifying the potential quality of service problems in 3G mobile networks. 3SQM had a lower correlation coefficient, thus a lower conformity to the subjective test results.

Further, the residual errors show cases with side difference between subjective and objective results. Wider differences were observed in 3SQM results. Results obtained from PESQ experiments had a very good correlation and generally prediction of overall quality of system, but still from the residual point of view, prediction errors of over 1 are noticeable.

Also, the results of our experiments before and after being transmission through live network for make and females show that the gender of the talker affects the results of objective measurements. We have shown that the gender of the speech sample affects the encoder and the quality scores drop after the encoding to GSM and AMR. However 3SQM shows a different behaviour with regards to the gender in live testing conditions. We suspect that gender can act as a co-variable in certain conditions in live networks, influencing the results of objective measurements. We leave this as a point for future work as exhaustive investigation is required to validate this hypothesis.

4 Conclusions and future work

This paper investigated the voice quality in a real live 3G mobile environment. The main goal of the study was to analyze the behaviour of two ITU-T’s objective measurement techniques (3SQM and PESQ) and their reliability in predicting the speech quality in 3G mobile networks. The comparison between the results of our objective and subjective tests shows that PESQ is a good and reliable tool for testing the quality and identifying the potential quality of service problems in 3G mobile networks. The results of our experiment showed a good correlation of objective measurements and subjective tests results, hence a good tests conformity.

However, the Pearson’s correlation coefficient is based on the similarity of the curves and does not reflect the differences in the results of individual test components. Also the polynomial regression applied to fit the objective results into the subjective conditions will conceal some prediction errors can cause over optimistic correlation results without reflecting the low accuracy. A comparison of individual samples using residual errors showed significant differences between objective and subjective test results. From the residual error point of view, PESQ showed a tendency to under predict the quality when the perceived quality is high and over predict the quality when the quality was lower than 3 (MOS score).

We also investigated the influence of talker’s gender on the results of objective measurements. The expected lower quality of female voice transmission was observed after encoding and decoding the samples. However we observed contradictory results after live recordings especially in case of 3SQM tests. All 3SQM results are over and under-predicted for male and female samples, respectively. This shows that the gender dependency is not only due to the codec functionality and may depend on the measurement algorithm as well.

In future, we will investigate more into the impact of other impairments in 3G networks such as packet loss, echo and delay. Statistical study of the influence of talker’s gender on the results of objective algorithms is also an interesting topic.

ACKNOWLEDGMENT

The work reported here is supported in part by the EU FP7 ADAMANTIUM project (contract No. 214751).

We are grateful to Dr. Zizhi Qiao and Dr. Zhuoqun Li from Motorola for the help on the setup of Asterisk-based testbed for measuring voice quality over 3G networks.

5 References

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

[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 Recommendation P.862, February 2001.

[3] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs" in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01), 2001. , 2001, pp. 749-752 vol.2.

[4] C. Hoene and Enhtuya, Dulamsuren-Lalla, "Predicting Performance of PESQ in Case of Single Frame Losses" in Proceeding of MESAQIN, Prague,CZ, 2004.

[5] ITU-T study group 12, "PESQ Limitations for EVRC Family of Narrowband and Wideband Speech Codecs", QUALCOMM Inc., 2008.

[6] Ditech Networks, "Limitations of PESQ for Measuring Voice Quality in Mobile and VoIP Networks", http://www.ditechnetworks.com, 2007.

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

[8] J. Holub and L. Blaskova, "Transcoded speech contemporary objective quality measurements reliability" in Wireless Telecommunications Symposium, 2008. WTS 2008, pp. 106-109.

[9] ITU-T, "Single-ended method for objective speech quality assessment in narrow-band telephony applications", in ITU-T Recommendation P.563, May 2004.

[10] Z. Becvar, L. Novak, J. Zelenka, M. Brada, and P. Slepicka, "Impact of Additional Noise on Subjective and Objective Quality Assessment in VoIP" in proceedings of MMSP 2007. IEEE 9th Workshop on Multimedia Signal Processing, 2007, pp. 39-42.

[11] S. Pennock, "Accuracy of the Perceptual Evaluation of Speech Quality (PESQ) algorithm." in MESAQIN, Prague (Czech Republic), 2002.

[12] L. Sun and E. C. Ifeachor, "Subjective and Objective Speech Quality Evaluation under Bursty Losses" in On-line Workshop Measurement of Speech and Audio Quality in Networks (MESAQIN 2002), Prague, Czech Republic, 2002.

[13] P. Paglierani and D. Petri, "Uncertainty Evaluation of Objective Speech Quality Measurement in VoIP Systems”, IEEE Transactions on Instrumentation and Measurement, vol. 58, pp. 46-51, 2009.

[14] T. H. Falk and C. Wai-Yip, "Hybrid Signal-and-Link-Parametric Speech Quality Measurement for VoIP Communications", IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, pp. 1579-1589, 2008.

[15] P. A. Barrett and A. W. Rix, "Applications of speech quality measurement for 3G" in Third International Conference on 3G Mobile Communication Technologies, 2002(Conf. Publ. No. 489), 2002, pp. 250-255.

[16] B. Rohani, J. Hans, and Z. rgen, "Application of a perceptual speech quality metric in power control of UMTS", in Proceedings of the 2nd ACM international workshop on Quality of service & security for wireless and mobile networks Terromolinos, Spain: ACM, 2006.

[17] B. Rohani and H. J. Zepernick, "An efficient method for perceptual evaluation of speech quality in UMTS" in proceedings of Systems Communications conference 2005, pp. 185-190.

[18] T. Hoßfeld and A. Binzenhöfer, "Analysis of Skype VoIP traffic in UMTS: End-to-end QoS and QoE measurements", Computer Networks, vol. 52, pp. 650-666, 2008.

[19] ITU-T, "Artificial voices", in ITU-T Recommendation P.50, September 1999.

[20] Malden electronics Ltd., "Speech Quality Assessment Background Information for DSLA and MultiDSLA Users", 2004. http://files.teraquant.com/malden/Speech%20Quality%20Assessment.pdf.

[21] Psytechnics, "Comparison between subjective listening quality and P.862 PESQ score-WP", http://www.psytechnics.com/site/sections/downloads/whitepapers.php, 2003.

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

[23] Ľ. Blašková, J. Holub, M. Street, F. Szczucki, and O. Tomíška1, "Objective and Subjective Degradations of Transcoded Voice for Heterogeneous Radio Networks Interoperability", 2008, ftp://ftp.rta.nato.int/PubFullText/RTO/MP/RTO-MP-IST-083/MP-IST-083-21.doc.