EVALUATION OF VOICE QUALITY IN 3G MOBILE NETWORKS

A thesis submitted to the University of Plymouth in partial fulfilment of the requirements for the degree of Master of Science

Project supervisor: Dr. Lingfen Sun

Mohammad Goudarzi

September 2008

School of Computing, Communications and Electronics
Faculty of Technology
University of Plymouth

Declaration

This is to certify that the candidate, Mohammad Goudarzi carried out the work submitted herewith

Candidate's Signature:

Mohammad Goudarzi _____ Date: 30/9/2008

Supervisor's Signature:

Dr. Lingfen Sun _____ Date: 30/9/2008

Copyright & Legal Notice

Acknowledgements

I wish to extend my warmest thanks and appreciation to those who have helped me during my thesis work.

My sincere thanks to my supervisor, Dr. Lingfen Sun, for the enthusiasm, inspiration, and all her support and guidance from the start to the end of this MSc project.

Mr. Zizhi Qiao (William) and Dr. Zhuoqun Li (Wood) from Motorola; thanks for helping out with the equipment and many discussions on how to get the system up and running.

My family, on whose constant love and support I have relied throughout my time at the University of Plymouth. I am grateful to my parents for creating an environment in which following this path seemed so natural. Without them none of this would have been even possible.

And I would like to thank the subjects who provided me with the experimental data that I regard as so important.

Abstract

The ultimate judge of Speech quality in mobile networks is the end-user. It is essential for network operators to consider user's needs in their network's technical standards. Two main approaches for measuring the speech quality are *Subjective* and *Objective*. Subjective tests are more accurate but often expensive and time consuming, and cannot be used for continuous measurement or simultaneous measurement in live networks. Objective measurements have been developed to estimate the opinion score of the speech quality.

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 and is the central component of speech quality assessment in many companies. 3SQM is an ITU-T standard for single-sided non-intrusive quality measurement.

In this research project, the speech quality in 3G mobile networks is evaluated by setting up a testbed platform based on Asterisk open source PBX to mediate between 3G mobile network and quality measurement equipment. Using over 200 speech samples, the performance of GSM and AMR codecs has been investigated using objective measurement tools, as well as the effect of other parameters such as the gender of the talker, time of the call and the mobile operator. To examine the accuracy of the objective tests, an informal subjective test was carried out with 33 subjects, and the correlation of the results was analyzed using a 3rd order polynomial regression method.

In all the experiments, perceived quality of the AMR encoded speech samples are higher than that of the GSM codec. Almost none of the GSM encoded samples in live recordings graded over 3.5. The results also showed gender dependency of the speech quality measurements. Female talkers tend to have a meaningfully lower objective mean opinion scores (MOS).

In terms of accuracy, the results of the informal subjective quality test shows that in general, PESQ and PESQ-LQO measures have a high correlation with subjective assessments whereas 3SQM measurements had a fair correlation. According to the results, PESQ can be used reliably for objective speech quality testing in live 3G networks. 3SQM as non-intrusive test method could not supersede intrusive analysis as expected. However, individual cases in which 3SQM performed better than PESQ were found. Also, 3SQM showed useful in identifying quality in individual tests and - as a non-intrusive measurement - has many advantages in live networks. Therefore, we recommend a co-existence of both measures when investigating speech quality in 3G mobile networks.

Table of Contents

List of Figures

List of Tables

1. Introduction

Speech quality is the most visible and important aspect of quality of service (QoS) in mobile, telecommunications and VoIP networks. Therefore, the ability to monitor and design this quality has become a main concern. Speech quality or Voice quality (often used interchangeably) refers to the comprehensibility of a speaker's voice as perceived by a listener.

Voice quality measurement (VQM) is a relatively new discipline in telecommunications networks. By measuring the speech quality, end-user's perspective can be added to traditional network management evaluation of VOIP, voice and telephony services.

Traditionally, user's perception of speech quality has been measured using subjective listening tests in which a subject hears a recorded speech processed through different network conditions and rates the quality using an opinion scale. Subjective listening 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. Nonetheless, they are time-consuming and expensive, they only measure on test calls and it is impossible to use them to supervise all calls in the network. Hence, they are not suitable for monitoring live networks.

As a result of major developments in market competition and rising quality of service- in importance - in the telecommunications industry during the past three decades, the area of research has been developed to estimate the quality of calls using objective methods. These objective measures that can be easily automated and computerized are becoming broadly used in the last two decades.

Speech will remain one of the most important services in third generation mobile networks. Customers are now able to choose their service provider by comparing the price and the quality of service offered by the operator. It is absolutely vital that service operators can predict the quality from a customer's perspective in order to optimize their service and maintain their networks. The challenge is to enhance speech quality while simultaneously optimizing the efficiency of the network to provide customers with a robust, reliable and affordable service.

1.1. Motivation

Due to rapid changes in user expectation, 2G networks do not satisfy today's wireless needs by the today's users. More and more mobile telecommunications networks are being upgraded to use 3G technologies. The ultimate judge of speech quality in mobile networks is the end-user. Thus, it is essential for network operators to feature the user's needs in their network's technical standards. Also, measurement of speech quality perceived by the user has many constructive applications in 3G networks such as testing speech and channel codecs, signal processing algorithms and handsets through to entire network In 3G planning, procurement, optimization, network monitoring, upgrades and network operation. Objective speech quality measurement 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.

Evaluation of speech quality has been subject of extensive research especially in the last decades. At the present time, not all the parameters that can affect the perceived speech quality are completely studied in live environments and some of them may not be fully understood. Even the good measurement methods with high correlations with subjective methods such as PESQ have shown to be inaccurate in certain network conditions and cannot be used reliably in every network condition. While there is a widespread belief that intrusive methods have a better performance in most network conditions, the behaviour of both intrusive and non-intrusive measurements methods needs to be more investigated and compared. Future studies in this area will most probably focus on developing new models and incorporating new parameters and algorithms to the existing models and further analyzing live network traffic to achieve more accurate measurements of the perceptual speech quality. An accurate understanding of the strengths and imperfections in the current quality measurement methods may help to optimize the design and development of more accurate algorithms. 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 goal of this thesis is to investigate the speech quality in a live 3G mobile environment by building up a quality test platform for 3G and using objective methods, namely Perceptual Evaluation of Speech Quality (PESQ) and Single Sided Speech Quality Measure (3SQM).

We have chosen PESQ since it is one of the most widely deployed objective measurement techniques used in the industry. Also 3SQM is the ITU-T's standard for non-intrusive measurement of voice quality in telephone networks and its performance has not been investigated by many researches in live 3G mobile networks compared to PESQ. Another purpose is to assess the accuracy of each objective method using subjective measurement results. For comparison and evaluation purposes, some of the test cases were tested by conducting an informal subjective test to obtain subjective opinion scores. The results of this research can contribute to the results of other researches on voice quality measurement. Improving the accuracy of current speech quality measurement techniques and/or designing new quality prediction models remains as a challenging task for future work.

1.2. Project aim and objectives

The aim of the project is to enhance and develop Asterisk-based 3G test platform and to evaluate voice quality for voice calls over 3G mobile networks.

The objectives of this project are:

- Obtain up-to-date knowledge on voice quality assessment for 3G networks.
- Set up and enhance voice quality test platform for 3G network, based on Asterisk open source package to transfer the calls from 3G network to the quality test equipment.
- Make live recordings over 3G network, measure the quality of the calls using PESQ and 3SQM and Analyze data to investigate the relationships between voice quality and relevant network parameters.
- Conduct an informal subjective test in order to investigate the accuracy of objective measures.

This research will extend work done in this area and contribute to other ongoing researches on Voice and Video quality measurement at the University of Plymouth.

1.3. Thesis structure

This thesis is divided into three major parts:

Chapter 2 provides an outline of the current literature in speech quality measurement techniques related to this research. In this section, a number of parameters that affect the overall quality perceived by the end user are discussed. The main ideas and basic principals of objective and subjective speech quality measurement are presented. Additionally, technical aspects of speech quality measurements and the objective models are discussed in detail in this chapter.

Chapter 3 and 4 are devoted to the approaches and the research methodology used by the author for carrying out the experiments in this research. Chapter 3 provides detailed, step-by-step specifications and instructions of the testbed platform built up for undertaking the quality tests in this research project. Chapter 4 is aimed to look deeper into the experimental design and how the experiments are carried out, how the samples are selected and what methods will be used for analyzing the results.

Chapter 5 and 6 present the results of the objective and subjective experiments conducted, discussion and analysis of the results, and finally the conclusion of the research project as well as the expected future work. Chapter 5 provides the significant findings of the research along with their related discussion. Ultimately chapter 6 presents the conclusions of this research as well as the limitations of the work and the suggestions for future work.

2. Literature Review

2.1. Background

In telecommunications, Quality of service (QoS) is considered to be divided into three components(Möller, 2000). The main component is the speech or voice communication quality related to a two-way conversation over the telecommunications network. The second component is the service-related influences also referred to as "service performance", which includes service support, a part of service operability and service security. The third part, which is the necessary terminal equipment performance, is separated from service performance because service can sometimes be accessed from different terminals. Speech quality is user-directed and corresponds to a major component of the overall communication quality perceived by the user. The question is which feature results in acceptability of the service by the user.

Quality can be defined as the result of the judgment of a perceived constitution of an entity with regard to its desired constitution. The perceived constitution contains the totality of the features of an entity. For the perceiving person it is "a characteristic of the identity of the entity"(Möller, 2000). In terms of voice communication systems, quality means the overall customer's perception of the service and Voice quality measurement(VQM) means the measurement of the customer's experience of the service(Mahdi, 2007). Therefore, the most accurate method of measuring the speech quality would be to actually ask the customers. However, this is purely hypothetical. In practice, there are two main types of voice quality tests: *Subjective* and *Objective*.

2.1.1. Parameters Affecting Speech Quality

2.1.1.1. Speech codecs

The source encoding functions transform the user's information stream into digital format. The aim of a source encoder is to encode the traffic into the smallest number of bits and minimize the number of bits which will be sent over the air interface(Korhonen, 2001).

Speech coding has its most important applications in mobile and Voice over IP. AMR (Adaptive Multi-Rate) is the mandatory speech codec selected in 3GPP (3rd Generation Partnership Project) for 3G mobile networks.

2.1.1.1.1. Narrowband AMR

Adaptive multi-rate (AMR) speech codec designed to operate on narrowband audio signals (300-3400 Hz). It is based on an A-CELP technology (Algebraic Code Excited Linear Prediction). AMR codec supports eight different variable coding rates (range from 4.75 to 12.20 kbps) which enable it to change the trade-off between bit-rate and speech quality every 20 ms. In addition to that, the AMR codec is provisioned with a voice activity detector (VAD) and comfort noise scheme for discontinuous transmission(Barrett and Rix, 2002).
AMR-NB was originally developed for GSM to provide the best possible coding based on the radio link quality. The AMR narrowband codec was adopted by 3GPP as default speech codec for various services such as audio component of low-bit rate streaming content (release 4), audio component of circuit-switched H.324 multimedia (release 99), and the audio component of packet switched multimedia (release 5).

2.1.1.1.2. Wideband AMR

AMR-WB has been a major step towards quality improvement. It is the extension of AMR concept to wideband signals (50-7000 Hz). It supports variable coding rates(ranging from 6.60 to 23.86 kbps), voice activity detection (VAD), Discontinuous transmission (DTX), and Comfort Noise Generation (CNG) and is capable of changing mode every 20 ms. Due to its audio bandwidth extension to 7 KHz, which results in improved intelligibility and naturalness of speech, the subjective perceived speech quality is significantly superior to of AMR-NB (Mullner *et al.*, 2007).
Wideband AMR codec is the default codec for wideband telephony services and the audio component of packet-switched conversational and streamed multimedia services. It has also been standardized by 3GPP for use in GSM, EDGE and 3G applications. AMR-WB is mandatory for many wireless services in 3GPP such as multimedia messaging(MMS), packet-switched streaming service (PSS), IMS messaging and presence, multimedia broadcast/multicast service (MBMS)(Varga *et al.*, 2006).

Compared to GSM codecs such as GSM-FER and GSM-BR, AMR includes a flexible solution by adopting the relation between speech coding and channel coding to the channel conditions. Furthermore, AMR can generally offer a large gain in quality. Speech quality gain can be traded for higher system capacity at the same quality level. (Corbun *et al.*, 1998; Uvliden *et al.*, 1998). The flexibility of AMR codec makes it a major candidate for future applications in 3G cellular systems as well as internet and VoIP applications.

2.1.1.2. Radio Transmission Errors

Transmission errors can dramatically degrade the speech quality delivered by a radio system. Although mechanisms such as forward error coding are used to minimize the effect of transmission errors, it is noticeable that the performance of such mechanisms is highly dependent on the detailed burst characteristics of the errors on the radio channel(Barrett and Rix, 2002). Hence, measurement of speech quality from simple link measures such as mean error rate or frame erasures can not be easily achieved.

2.1.1.3. Mobile Device Design

Some performance aspects of the terminal such as send and receive loudness ratings (SLR and RLR), terminal coupling loss (TCL) and frequency response of the send and receive paths, and noise and RF pick-up affect the conversational quality experienced by the user of the device. The performance of a handset is largely dependant on its physical design. For this reason, manufacturers are now incorporating signal processing into devices which themselves can introduce new issues such as the unpredictable performance of such signal processing algorithms in different conditions(Barrett and Rix, 2002).
Barret *et al.* describe the concept of *Handset testing*. This type of test is to assess the effect of signal processing in the terminal and audio interface to the user as well as the acoustic echo path of the human body by using a head-and-torso simulator (HATS).The test signal will be played through the HATS mouth and recorded from a microphone in the HATS ear. Using this method makes it possible to measure the quality of the network for conversation by combining the results of this test with the simulation of the echo and noise of the network(Barrett and Rix, 2002). The perceptual effect of the echo, delay, speech levels and speech quality will be combined to achieve a conversation quality score(Rix *et al.*, 1999).

2.1.2. Subjective versus Objective Methods

In a typical subjective listening test, recordings processed through different network conditions will be heard by subjects and will be rated using a simple opinion scale such as ITU-T (International Telecommunication Union-Telecommunication Standardization Sector) listening quality scale. MOS score is the arithmetic mean of all the rating registered by the subjects, and can range from 1 to 5.

Table 1- Subjective listening-only test opinion scale

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Objective measures that are based on mathematical algorithms - and can be easily automated -are being extensively used over the past two decades and in most cases as to supplement subjective test results(Mahdi, 2007). Several objective MOS measures have been developed in recent years, namely PAMS (Perceptual Analysis Measurement System).PSQM(Perceptual Speech Quality Measure), PESQ (Perceptual Evaluation of Speech Quality) and 3SQM(Single Sided Speech Quality Measure).

2.1.3. Standardisation of Speech quality measurement techniques

Several objective measures have been proposed for estimating the quality score of the speech using computational models. Among them are ITU-T standard recommendations adopted for measuring speech quality in telephone networks such as PSQM, PESQ and 3SQM. Objective measurement techniques can be divided into to main groups: *intrusive* and non-intrusive:

2.1.3.1. Intrusive Methods

An intrusive test is generally based on sending stimulus through the system under test and comparing the output signal to the original. A test signal -typically a natural speech recording of around 8 seconds or more- is passed through the network. The receiving signal will then be

processed using an algorithm such as PESQ (ITU-T recommendation P.862 as the standard algorithm for intrusive testing) which outputs quality score (estimation of MOS) and some other diagnostic information for further investigation.
Intrusive methods have a number of disadvantages. They consume network capacity when used for testing live networks. More calls can be assessed if the voice quality can be assessed through non-intrusive methods by using the in-service speech signals.

PAMS: Perceptual Analysis Measurement System is an objective measurement algorithm designed for robust end-to-end speech quality assessment(Rix and Hollier, 2000). PAMS is designed for intrusive assessment and is being used successuly in many different applications.

PSQM: Perceptual Speech Quality Measure is an algorithm defined in ITU Recommendation P.861 that objectively evaluates and quantifies voice quality of speech codecs. It was primarily standardized to assess the speech codecs mostly used in mobile networks as other services like VoIP was not yet a topic. As a consequence of later developments in VoIP applications, the requirement for measurement techniques changed significantly and the measurement algorithms had to deal with much higher distortions than before. PSQM was revised to overcome new problems such as burst errors and varying delay which resulted in the development of other versions like PSQM+ and PSQM/IP. PSQM shows a good performance in terms of relation to actual speech quality. But Like other speech-based methods, it is based on test calls and can not be used for constant monitoring and optimisation of the network. PSQM is not recommended by ITU-T for degraded cellular conditions and distortions such as handovers and bursts of frame erasures are generally out of scope for PSQM. The ITU-T has withdrawn P.861 and replaced it with P.862 (PESQ) which contains an improved speech assessment algorithm.

PESQ: Perceptual Evaluation of Speech Quality is a mechanism for automated assessment of the speech quality perceived by the user of a telephony system. It is standardized as ITU-T recommendation P.862. It is now one of the most broadly used objective quality measurement methods in telecommunications and IP networks.
PESQ is designed to predict the perceptual quality of a degraded audio signal by analyzing specific parameters such as noise, errors, coding distortions, delay, delay jitter, time wrapping, and transcoding. PESQ combines the excellent psycho-acoustic and cognitive

model from PSQM+ with a time alignment algorithm from PAMS which perfectly enables it to handle delay Jitter(OPTICOM, 2007). . The PESQ-algorithm is illustrated in Figure 1:

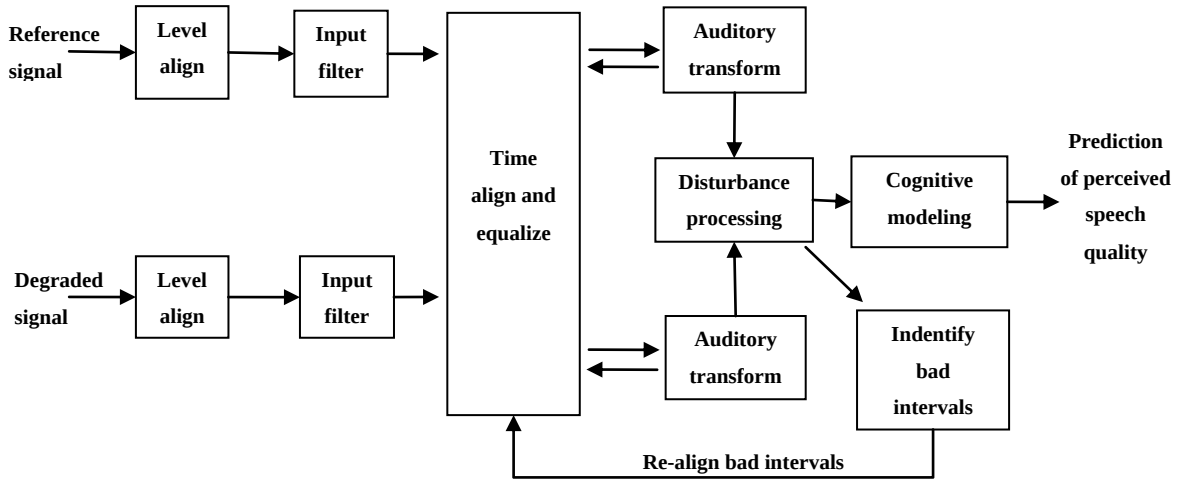


Figure 1-Block diagram of the PESQ algorithm(OPTICOM, 2007)

The PESQ algorithm consists of two main parts (Storm, 2007; QUALCOMM, 2008): *Conversion to the psychoacoustic domain* and *Cognitive modelling*

In the first part of the PESQ algorithm, the signals are processed and converted to psychoacoustic domain. As the gain of the system under test may vary depending on the interface used for measurement (e.g. ISDN), the first step of the processing is to align both original and degraded signals to the same constant power level in order to compensate for any gain or attenuation of the signal in the *level alignment* block. The level alignment block is the same as the normal listening level used in subjective tests.

Because the algorithm needs to model the signal that subjects would actually hear, the filtering that occurs in the handset/receiver in a listening test is modelled and compensated in the *Input filter* block. PESQ assumes that the handset's frequency response follows the characteristics of an IRS (Intermediate Reference System) receiver as used in subjective tests. Therefore PESQ will model the IRS-like receive filtered versions of the original speech signal and degraded speech signal. The IRS filtered signals will later be used in the time alignment procedure and the perceptual model block.

In order to allow for corresponding signal parts of the original and degraded files to be compared, PESQ computes the time delay values. The resulted time delay values will be used in the perceptual model.

In the *Auditory transform* block, a psychoacoustic model is applied to the signal, which will map them into internal representation in the time-frequency domain by a short-term *Fast Fourier Transform (FFT).* The purpose of this model is to mimic the properties of human hearing. During this mapping the filtering in the tested system is compensated and the influence of time varying gain is also removed.

In the cognitive model part of the PESQ algorithm, during the FFT the intensity of the spectrum is warped into a modified Bark scale, called the *pitch power density*, which mimics how the human ear transforms intensity into perceived loudness and reflects the human sensitivity at lower frequencies. The achieved representation is called the *Sensation Surface.* The sensation surface of the degraded signal is subtracted from the sensation surface of the reference signal in the *Disturbance Processing* block, taking into account how the brain perceives differences. The result is a disturbance density signal. Two different disturbance parameters are calculated; the *absolute (symmetric) disturbance* and the *additive (asymmetric) disturbance*. Next the two disturbance parameters are aggregated along the frequency axis resulting in two *frame disturbances.* Due to incorrect time alignment for an interval of speech, the disturbance signal may contain an interval of poor disturbance (above a threshold of 45). In this case, in the *Identify Bad Intervals* block, the time alignment and the subsequent PESQ processing is redone for the bad interval. If this resulting disturbance signal is better, the new result will be used instead.

The frame disturbance values and the asymmetrical frame disturbance values are aggregated over intervals of 20 frames. These summed values represent how distorted the speech is during very short periods of time (QUALCOMM, 2008). The values are then aggregated over the entire active interval of the speech signal. The final estimation of the perceived speech quality or the PESQ raw-score is a linear combination of the average disturbance value and the average asymmetrical disturbance value, which ranges from 0.5 to 4.5. The resulted PESQ raw-score has shown a poor correlation with MOS-LQS in some cases. P862.1 annexation introduces a transform function (see Equation 2-1) to achieve a better performance:

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945\,x + 4.6607}} \qquad (2\text{-}1)$$

Where $x$the PESQ is raw score and y is the MOS-LQO score.

This transform function maps the PESQ raw-score into the MOS-LQO (MOS-Listening Quality Objective) which improves on the original PESQ (P.862) by correlating better to subjective test results(GL Communications, 2007) .

Figure 2 shows the MOS-LQO mapping function, which gives a score from 1.02 to 4.55.

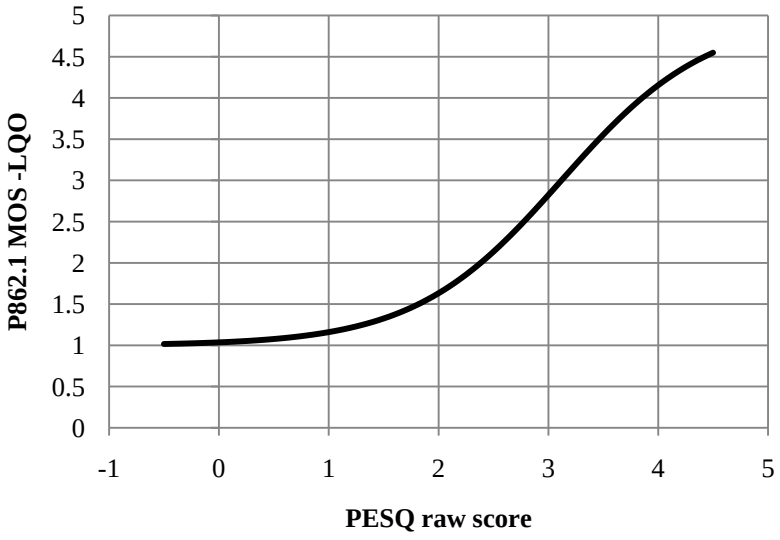


Figure 2- MOS-LQO transform function

It has to be noted that the mapping function proposed in P.862.1 predicts on a MOS scale. This mapping function converts the raw P.862 scores to MOS-LQO values and is a general calibration of PESQ score, derived as an average statistical function across a large number of subjective test results data in different contexts and languages and is not supposed to predict the MOS of a single experiment.

2.1.3.2. Non-Intrusive Methods

Intrusive quality measurement techniques require the reference signal to be inserted into network or the device under test. However, due to the extra traffic that intrusive methods would generate, a non-intrusive method that is only based on single sided monitoring may be sometimes more desirable especially in case of live networks. Objective test methods can be generally categorized to “Signal-based” and “Parameter-based” methods (Ding and Goubran, 2003a; Sun, 2005).

Signal-based: Also referred to as output-based techniques, these techniques work based on predicting the voice quality directly from the degraded speech signal obtained fro the system under test. Signal-based methods such as PAMS, MNB, PSQM and PESQ use the original signal and the degraded output signal to measure the perceptual speech quality.

3SQM: stands for "Single Sided Speech Quality Measure" developed for non-intrusive voice quality testing. It is based on ITU-T recommendation P.563.
Figure 3 illustrates the block diagram of 3SQM non-intrusive analysis algorithm (OPTICOM, 2004).

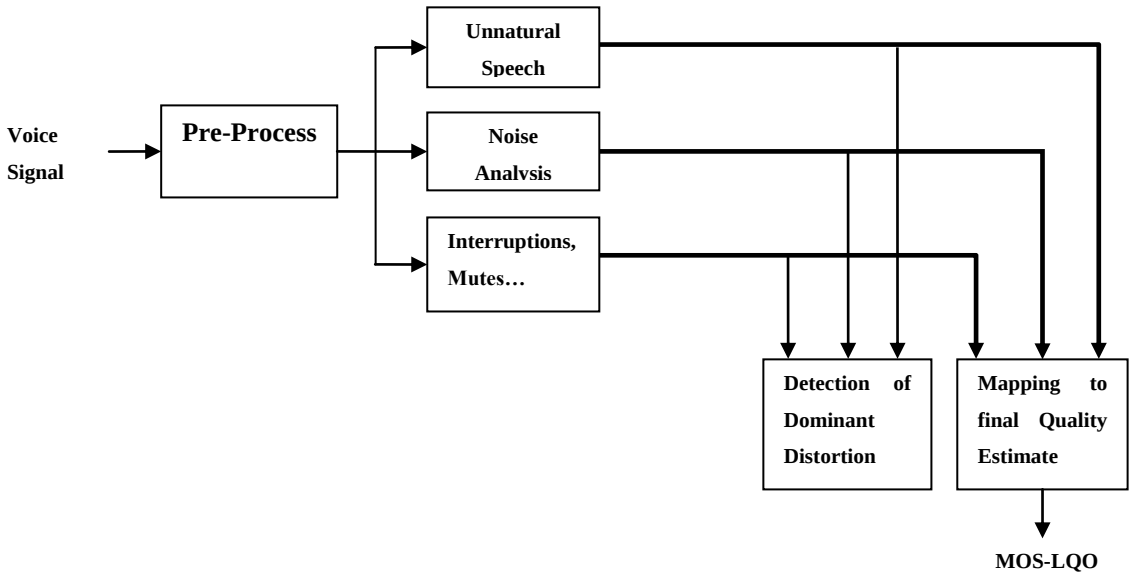


Figure 3-Block diagram of 3SQM algorithm

However, non-intrusive methods are commonly considered to be less reliable and accurate as intrusive measurement techniques due to the missing information of the source signal, and are used in the industry jointly with intrusive methods or for deriving a course quality indicator for the speech signal.
A number of other non-intrusive methods have been proposed in the literature with different approaches to voice quality measurement (Gray *et al.*, 2000; Clark, 2001; Conway, 2002). One proposed technique to measure the quality of a network degraded speech stream is to use vocal tract models to identify distorting parts of the signal. The aim of this method is to predict how it is plausible that the voice signal be generated by the "human vocal production

system"(Gray *et al.*, 2000). This method is said to offer accuracy approaching that of subjective and intrusive methods(Barrett and Rix, 2002).

Parameter-based: These techniques measure the voice quality using IP network impairment parameters such as packet loss rate, delay and jitter. Parameter-based methods such as the E-model use a computational model instead of using the real measurement.

E-Model: The E-Model, is based on the basic principle that: "Psychological factors on the psychological scale are additive" (ITU-T, 2003). The purpose of the model is to predict the subjective effect of combinations of impairments using stored information on the effects of individual impairments to help network planners design networks. In terms of VoIP, this means that the contribution of each impairment factor that affects a voice call can be computed separately even though such factors may be correlated. E-model includes the transmission statistics as well as the voice application characteristics like codec quality and impacts of packet loss and the late packet discard on the codec. Therefore, using E-model, the speech quality can be estimated by means of the R factor once the network and application statistics have been captured for a well-known codec(ITU-T, 2003; Carvalho *et al.*, 2005).
"VQmon" technique introduced by (Clark, 2001) is based on the E-model and uses Markov chain to model the packet loss characteristics of a VoIP call. It also considers the impact of time varying impairments such as bursty packet loss and recency. This technique is claimed to provide results with good correlation with subjective speech quality measurements.
Another non-intrusive method introduced by (Conway, 2002) is based on constructing a "pseudo-packet" by capturing the VoIP packet streams and their sequence and timestamps from the network and replacing their payload with a payload that they would have if they had been used to carry test voice signals. Existing objective quality evaluation algorithms can then be applied to the output without requiring any knowledge of the original transmitted voice signal. One advantage of this method over the other two mentioned methods is that it makes use of the sophisticated processing algorithms such as PSQM and PESQ. Hence, it exploits the results of considerable work that has been standardized, developed and accomplished by ITU-T and therefore inherits the accuracy provided by such methods. Improved objective speech quality measurement methods can also be incorporated to this method as they become available in the future.

2.2. Novel methods of objective quality measurement

Although there have been many advances in mobile non-voice services such as multimedia and data, speech quality is still one of the most significant factors in customer satisfaction in the mobile market. ”The ability to accurately measure customer experience is a vital tool in the fight to improve return on capital expenditure, stimulate usage, and reduce churn”(Barrett and Rix, 2002). An ideal speech quality measure must be capable of continuously monitoring all calls in the network, not biased for different channel conditions and with excellent correlation to actual speech quality(Karlsson *et al.*, 1999).

Speech quality has been traditionally measured in GSM networks using RxQual (Coded bit error rate) parameter. RxQual is a simple measure obtained by transforming the average bit error rate (BER) over a 0.5 second period, to a scale of 0 to 7. RxQual measure’s relation to true speech quality depends on the channel conditions and is not capable of capturing many factors such as the distribution of bit errors over time, frame erasures, handovers and different speech codecs when measuring the perceived speech quality. As a result, it is hard to use RxQual measure in order to optimise networks through speech quality (Karlsson *et al.*, 1999; Ericsson, 2006).

SQI measure is an objective, transmission based , integrity measure based on such parameters given by measurements that cellular systems perform on the radio-link as bit error levels, erased frames, stolen frames, hand-off situations, DTX-activity and statistics on distribution of each of mentioned parameters(Karlsson *et al.*, 1999). In order to estimate perceived speech quality, these parameters will be combined together with knowledge of speech codec capability. Therefore, In order to tune the SQI measure to the characteristics of each codec, parameters and transforms need to be obtained-using live recordings- and applied to the model for each codec. Performance comparisons presented has shown that SQI measure has better performance than other methods such as PSQM and RxQual measure in terms of correlation to the actual speech quality when compared to the results of subjective comparative listening tests. SQI measure can be easily used to continuously supervise all calls in the network when used in uplink.

In the last few years, voice over IP (VoIP) protocol has become an important application running over TCP/IP networks. More and more voice traffic is expected to be carried over IP networks due to its cost-effective service. Speech quality is the most visible and important aspect for QoS for VoIP applications. Since VoIP applications are real-time applications, impairments such as delay and packet loss will directly affect the end-to-end speech quality.

Many studies have been carried out to investigate the impact of different IP network parameters on the perceived speech quality. Two major models have been used to predict the speech quality directly from network parameters.

By combining E-model and PESQ intrusive algorithm, a new methodology for developing models for non-intrusive prediction of speech quality was presented by Lingfen and Ifeachor in 2006. Using this new methodology they have developed non-linear regression models to predict perceived speech quality for modern codecs such as G.729, G.723.1, AMR and iLBC. Another advantage of the presented method is that it is a general method and can be applied to other media such as audio, video - by including additional parameters - and can also be used in automated multimedia systems to control sender-bit-rate and adaptive codec type/mode in order to acquire better perceptual quality non-intrusively (Sun and Ifeachor, 2006).

Ding and Goubran proposed a new formula for speech quality evaluation by extending the E-model and using PAMS to measure MOS score for G.723.1 and G.729 codecs. Their new formula quantified parameters such as packet loss, delay jitter and buffer size and incorporated them into the E-model. The impact of each parameter was first examined separately and then the combined effect was examined as well by introducing them jointly. The new extended E-model formula showed good accuracy in the simulation for separated impairments as well as combined impairments when packet loss rate was lower than 10%(Ding and Goubran, 2003b).

Lingfen and Ifeachor proposed another new method for speech quality measurement based on PESQ algorithm and E-model. In this method, the degraded speech signal is generated by decoding the speech signal that has been first encoded and then processed in accordance with network impairment parameter values. The achieved degraded signal will be then processed along with the reference signal by PESQ to achieve conversational MOS score. Since this method is based on objective tests rather than subjective methods, it can be easily extended to other network conditions and codecs(Sun and Ifeachor, 2004).

Artificial neural network (ANN) models have been recently used to objectively predict speech quality of networks. Because of its ability to learn, ANN models have the advantage of adapting to the dynamic environment of IP network networks such as the Internet, compared to E-model - which is static.

In 2002, Lingfen and Ifeachor developed a new ANN based model for speech quality prediction -directly from IP network parameters- by investigating the impact of packet loss, Codec (G.729, G.723.1 and AMR) and gender of the talker. The results of their study showed

high prediction accuracy of the ANN model. They concluded that the loss pattern, loss burstiness and the gender of the talker meaningfully affect the perceived speech quality. Even though, deviation in speech quality is dependant on the packet size and codec, they found no significant correlation between packet size and the perceived quality for a given packet loss rate.

In voice over IP applications (VoIP), due to the nature of internet and TCP/IP-based networks, perceived speech quality is primarily impaired by delay, variation in delay (Jitter) and packet loss. (Sun and Ifeachor, 2003).

Many studies have been conducted on effects of packet loss on the perceived speech quality. Many showed that packet loss has a dramatic effect on the speech quality. Simulations using various codecs, random packet size and different error concealment methods have shown that MOS drops dramatically by increasing the packet loss (Yamamoto and Beerends, 1997; Duysburgh *et al.*, 2001).

Frame size is an important parameter that affects the speech quality. Using E-model and by simulating random packet loss, different packet sizes and various error concealment techniques for G.729 codec, (Ding and Goubran, 2003a) concluded that MOS drops dramatically when large frame size was used.

Other novel methods that do not need to use the original speech, training database or any perceptual quality measurement methods have also been proposed. An objective quality evaluation method using digital speech watermarking, as explained by (Cai *et al.*, 2007) is principally based on the techniques of discrete wavelet transform and quantization. Based on the fact that the embedded watermark will have the same distortion as the original speech, this method can predict speech quality by calculating the percentage of correctly extracted watermark bits. The experimental results of this method under MP3 compression, low-pass filtering, and Gaussian noise distortions showed accurate results with close correlation to PESQ results for both male and female speakers.

2.3. Applications of Speech quality measurement in 3G

Speech quality measurement has many applications in 3G and VoIP networks such as testing speech and channel codecs, signal processing algorithms and handsets through to entire network In 3G planning, procurement, optimisation, network monitoring, upgrades and network operation(Barrett and Rix, 2002). End-user’s perception of speech quality can be

efficiently used to improve the power control(Rohani *et al.*, 2006b) and link adaptation(Rohani and Zepernick, 2004) techniques in both GSM and UMTS wireless systems. Studies have shown that by applying speech quality measurement to the outer loop power control (OLPC) in 3G systems, significant capacity increase can be achieved compared to the use of conventional measures(Rohani *et al.*, 2006a).

Speech quality may be used as an effective measure in design of control systems in telecommunications networks. The effect of Jitter can be compensated by employing playout buffer algorithms at the receiver(Sun and Ifeachor, 2004). Formerly, parameters such as buffer delay and loss performance were used for choosing and designing such buffer algorithms. Perceived speech quality can be effectively used as a better measure to control the playout buffer to maximize MOS values in accordance to delay, loss, and rate(Fujimoto *et al.*, 2002; Boutremans and Le Boudec, 2003).

In addition, speech quality measurement models can be efficiently used for voice quality monitoring, perceptual buffer design and optimization and other QoS control purposes(Sun and Ifeachor, 2004).

However, current objective measurement techniques may not always accurately predict the true speech quality. In real-life situations, end-user's perception of the speech quality depends on many other conditions and impairment. It can be discussed that even the formal subjective quality measurement results may not correlate well with the real day-to-day end users' perception of the speech quality. Also the performance and the limitations of these objective methods in measuring the perceived quality with a good correlation with user's perception in real life live networks should be further analysed.

2.4. Limitations of existing objective quality measurement

Advanced algorithms have become available that can accurately measure the speech quality as perceived by a user. However, as many of these measures may have deficiencies and biased results in different network and speech conditions, more studies need to be conducted in order to generate more accurate prediction models and/or to include different parameters and conditions in existing models.

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 accurately predict perceived quality in live network as a result of improper time-alignment as reported by (Qiao *et al.*, 2008). Also 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. (Ditech, 2007) reports that there are significant known limitations to the PESQ algorithm in with regards to its time alignment and psychoacoustics model.

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).

It should also be noted that none the existing objective measurement techniques provides a comprehensive evaluation of a two-way transmission quality. It only measures the effects of one-way speech distortion and noise on speech quality. The effects of loudness loss, delay, sidetone, echo, and other impairments related to two-way interaction(ITU-T, 2001) are not reflected in the PESQ scores.

In most cases the assumptions about the behaviour of network losses do not reflect reality. Some methods introduced in the literature are based on the assumption of a linear relationship between MOS and packet loss or that the impacts of delay and packet loss on voice quality are linearly additive. Also some studies have suggested that same equation may be used to calculate packet loss effects for all codecs. Nonetheless, these assumptions may not be generalized and are uncertain for different codecs especially new codecs(Sun and Ifeachor, 2004).

Speech quality perceived by a listener is to some extent relative. In a research conducted by (Sun and Ifeachor, 2002) the impact of the gender of the talker (extracted from decoder) on perceived quality was investigated in addition to packet loss and codec. The results have shown that the perceived quality of the female talker tend to be lower than of the male talker.

The language of speech can also affect the perceived speech quality. Since, using linear prediction model for production of speech will result in different performance in different languages, talker dependency due to the codec algorithm, is generally inevitable for different CELP-based codecs such as G.729, G.723.1 and AMR(Sun and Ifeachor, 2002).

In addition, both E-Model and ANN based methods are dependant on databases obtained by subjective tests. Hence, the databases are limited and inadequate data exist to cover all network conditions and different scenarios. As a consequence, the effect of a number of network parameters has not been fully investigated yet.

The future investigations will focus on further analysis of other patterns such as the loss pattern with the aim of identifying additional perceptually relevant parameters and including more accurate features of speech content into objective models. For VoIP, real Internet VoIP data will be used for objective models like the ANN (Artificial neural network based model) (Sun and Ifeachor, 2002). Furthermore, models can be optimized by examining more speech data for analysis of parameters' dependency to the perceived quality. Optimizing such models can lead to efficient and non-intrusive QoS models and control mechanisms for VoIP and telecommunication networks.

Furthermore, future studies will focus on developing new joint-models that can effectively represent the speech quality of the call by incorporating different codecs, different metrics, different algorithms and the co-effect of different network parameters.

The efforts are being made to improve existing objective models and develop new models that have higher correlations with true speech quality. In order to achieve this, the new tools should be able to predict the quality in all the new mobile IP and VoIP environments. Future objective measurement models will have to combine quality and intelligibility measurements. This may be possible by extending the PESQ cognitive model to include intelligibility factor and give a common score for both quality and intelligibility.

Real-world Voice over IP scenarios are far more complicated. Voice activity detection and various transcodings might be used and voice quality may be distorted as a result of loss burstiness and different frame sizes. Future works will focus and the impact of such impairments on the speech quality for different codecs and scenarios.

2.5. Summary

In this chapter the theories and literature related to the research has been reviewed. The main factors affecting the quality of service in mobile networks were described in the context of speech quality measurement; and subjective and objective measurement tests were compared. Although subjective quality tests are the most accurate method for measuring the speech quality, they come at a high cost and are often very time-consuming. Therefore objective measurements have been developed to predict the perceptual opinion score of the system based on the data from several subjective tests. Next, intrusive and non-intrusive methods were introduced and two main intrusive (PESQ) and non-intrusive (3SQM) models were examined in detail.

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 surveys, it can be seen that there is still work to be done in the area of objective quality measurement. PESQ does not always accurately predict perceived quality. Further research can be done in the area to pinpoint the flaws and strengths of each objective model, which can help to further improve the accuracy of each model or may lead to the development of new, more accurate objective measurement techniques. 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.

3. Testbed installation and Enhancement

The aim of the project is implemented by means of a voice quality testing system using Asterisk software. The first main objective of the project is to setup and develop a testbed for speech quality measurement and evaluation. The testbed will then provide the platform for speech quality measurements and further experiments in the field such as other codecs and media like video and SIP calls.

3.1. Testbed Architecture

Figure 4 illustrates a schematic diagram of the testing system used for voice quality measurement.

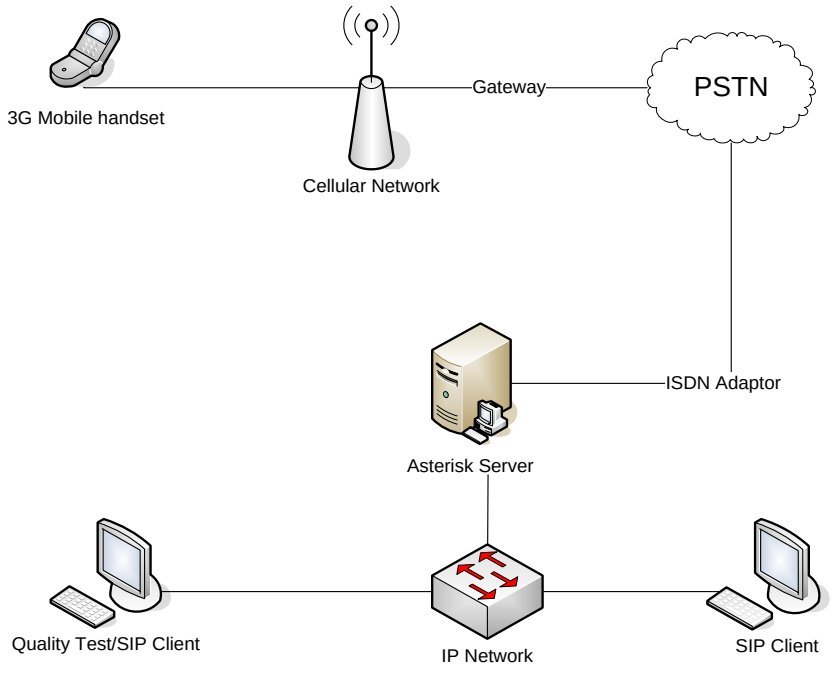


Figure 4- Testbed platform for speech quality evaluations

3G handsets: Calls are initiated from a 3G mobile handset to the landlines setup to pass the traffic through Asterisk server in order to route the call and also capture the required network parameters and record the calls.

Local PCs: The local PCs are used mainly for analysis purposes and as SIP clients. They are also used for storing the samples. Opticom Opera software and Audio Score software that are used for record/play and speech quality measurement are Windows-based softwares which are installed on local PCs.

3.2. Asterisk server

Asterisk is an open source, full featured PBX system. It supports almost all standard call features on station interfaces, such as Caller ID, Call Waiting, various types of Call Forward, Stutter Dial tone, Three-way Calling, Call Transfer, Directory Service (ADSI) enhancements, Voicemail, Conferencing, Least Cost Routing, VoIP gateway, Call Detail Records and full IVR capability. Also, Asterisk is programmable at many layers, from the lowest-level C code, to high level AGI scripting (similar to CGI) and extension logic interfaces (Spencer, 2008). The user of the Asterisk can easily manipulate, customize and develop the operations and the logic of handling the calls and need not know anything about the physical interface, protocol, or codec of the call they are working with due its design architecture.

Asterisk utilizes a modular design that makes it easy to operate and to manipulate and develop various components smoothly and independently, without having to deal with technical difficulties concerning the other components(Meggelen *et al.*, 2005). Figure 5 illustrates the modular architecture of Asterisk server design (Sacchi *et al.*, 2007). When the server starts, Dynamic module loader initializes all the parameters required for connection such as channels, dial plans, installed codecs. These links will be then linked with the appropriate internal Application Program Interfaces (APIs). When the server is ready, the Switching core is responsible for accepting the calls coming from the interfaces. Handling the calls will be done according to the dial plan. The Application launcher is handles physical operations such as ringing the handsets telephones. Asterisk applications allow it to connect any Interface, phone, or packet voice connection, to any other interface or service. The Codec translator module and its components provide transparent connection between different channels using different codecs. Hardware components such as ISDN cards or FXO cards

physically manage call initialisation and call reception issued by the Asterisk server. Asterisk's codec, file format, channel, CDR, application, switch and other API's separates developer/user from the complexities of the entire system(Spencer, 2008).

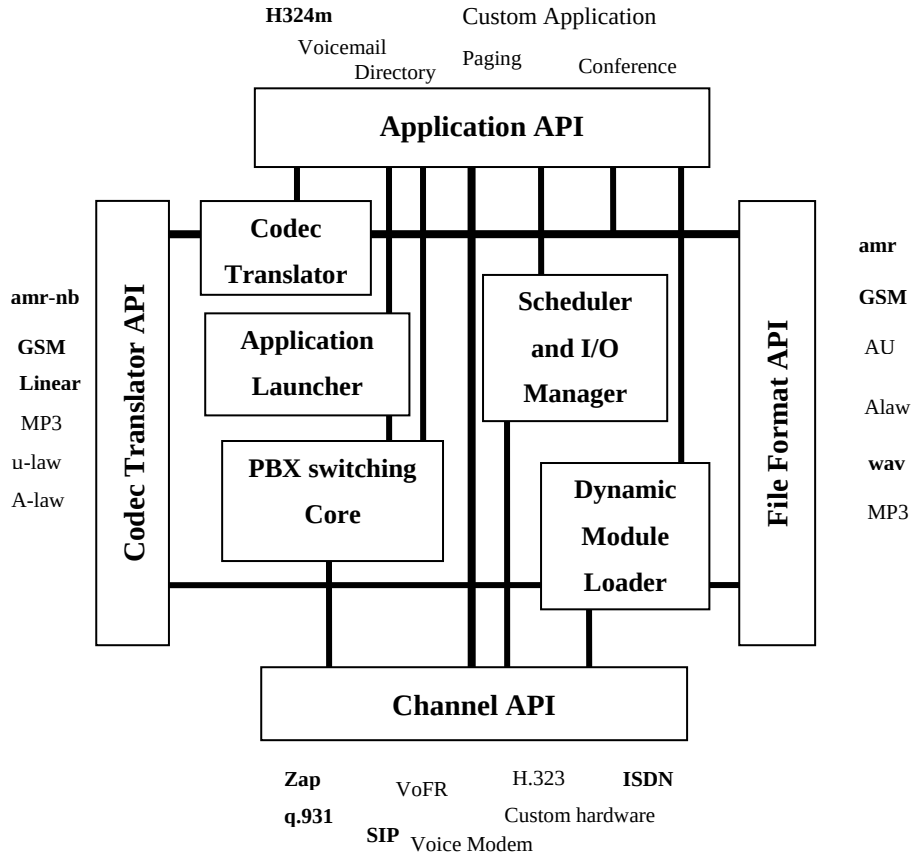


Figure 5- Asterisk modular server architecture diagram

Asterisk supports most of the codecs used in VOIP and telecommunication systems such as GSM, G.711, G.726, G729, iLBC and Speex. Other codecs required for this project such as AMR can be added through adding the required codes (patching the source code).

Asterisk is also capable of transcoding between various codecs. This means that Asterisk automatically translates between two codecs when required. This feature can be used when Asterisk acts as a mediator between two media or when mixing different speech codecs. (e.g. Calling from SIP phone to mobile phone).

3.3. Codecs and file formats

Asterisk provides transparent translation between all of the following codecs (Spencer, 2008):

- 16-bit Linear 128 kbps
- G.711u (μ-law) 64 kbps
- G.711a (A-law) 64 kbps
- IMA-ADPCM 32 kbps
- GSM 6.10 13 kbps
- MP3 (variable, decode only)
- LPC-10 2.4 kbps

Other codecs such as G.723.1 and G.729 can be passed through transparently.
In terms of file formats supported, Asterisk supports a variety of audio file formats:
Supported formats include:

- **raw:** 16-bit linear raw data
- **pcm:** 8-bit mu-law raw data
- **vox:** 4-bit IMA-ADPCM raw data
- **mp3:** MPEG2 Layer 3
- **wav:** 16-bit linear WAV file (8000 Hz)
- **WAV:** GSM compressed WAV file (8000 Hz)
- **gsm:** Raw GSM compressed data
- **g723:** Simple G723 format with timestamp

When a file is played back on the channel, Asterisk automatically chooses the least expensive format for that device.
Asterisk can play any file types if the format and codec is available for that file type. If provided with different file types, Asterisk will select the file type with the lowest impact on the systems performance. This selection is based on the translation weight values that Asterisk calculates when starting up. The translation results can be seen using the following command:

```
CLI> core show translations
```

3.4. Operating system

Asterisk is an open source/free software implementation of a PBX system. Asterisk is designed for the GNU/Linux operating system and can be installed on almost all existing distributions. However, there are minor differenced between distribution due to different kernel versions and the changes made to the original Linux kernel (Vanilla kernel) by the organizations for each distribution. In our project, because it was necessary to use bristuff (discussed in later sections), we were bound to find the best distribution to install all the patches and kernel modules required for supporting our BRI ISDN card. We installed Asterisk on three different Linux distributions. It was finally decided that Fedora core 8 is a convenient distribution, allowing that all our required packages could be easily compiled and installed.

3.5. Asterisk Installation

Libpri and Zaptel packages are necessary for the Asterisk installation because we need to have ISDN functionality. PRI and BRI cards will need the Zaptel module in order to work correctly. If using PRI cards, Libpri modules will have to be installed as well. Because we use BRI with zaphfc module, which depends on Zaptel module for loading and also uses PRI functionalities in Asterisk (maps all the BRI functionalities to BRI in Asterisk). It is necessary that we have both Libpri and Zaptel packages installed before installing the Asterisk.
It is important to install the packages in order. The order has to be: Libpri, Zaptel, and asterisk.

Installing Libpri:

```
# cd libpri
# make
# make install
```

Installing Zaptel:

```
# cd zaptel
# make
# make install
```

Note that if you are using older versions of Asterisk (like 1.2.x), you may need to pass your kernel version to make command. For example if you have a 2.6 kernel you should type:

```
# make linux26
```

Installing Asterisk:

cd asterisk

For Asterisk version 1.4 and above we should start configure script?

```
# ./configure
```

It is possible to customize the installation by using the menu provided by using the following command (Optional):

```
# make menuselect
# make
# make install
# make samples
```

To verify the asterisk installation, we can start asterisk daemon by typing `safe_asterisk` and connect to its console by typing `asterisk -vvvvvr`. Or we can start and connect to Asterisk in console mode by typing 'asterisk –vvvvvgc'. We should check for any errors or warnings when asterisk is loading to check whether all the applications and modules are being loaded properly.

In order to play Music-On-Hold (MOH), before making Asterisk you have to install mpg123 package. By installing mpg123 Linux will be able to decode and handle the MP3 file format. We can install the mpg123 packages that come with Asterisk.

```
# make mpg123
```

Linux Kernel Sources package (or Kernel Headers) must be installed on the system prior to installing Asterisk because we need to install a kernel module. In order to compile the zaptel and zaphfc packages on the system, it is necessary to have the kernel source version matching the kernel version that is already running on the system. To check whether the Linux kernel source is installed on the system, first we must find out the current kernel version running on the system by typing:

```
# uname –r
```

Or view the contents of the version file in /proc directory:

```
# cat /proc/version
```

The results of these commands will show the version of the kernel currently running on the system. It is required that we have the kernel source and headers version matching this version. Depending on the distribution of the Linux, the source and headers should be obtained and installed so the asterisk can be compiled without any problems.
Also, by default, during Zaptel and Zaphfc installation, Linux will look for the kernel source directory in `/usr/src` directory. Therefore, two symbolic links need to be created before compiling these packages.

```
# ln –s /usr/src/'uname –r' /usr/src/linux-2.6
# ln –s /usr/src/'uname –r' /usr/src/linux
```

3.5.1. Installation on Suse

The following packages need to be installed prior to before compiling the Asterisk:

Subversion
kernel-source - <*for current kernel version*>
gcc
make
ncurses
ncurses-devel
openssl

openssl-devel

zlib

zlib-devel

We need to make sure all of these packages are already installed before installing the Asterisk. To install any required packages in Suse using YaST, type:

yast

3.5.2. Installation on Debian

To install Asterisk on a server running Debian with kernel 2.6, some additional packages are required. Make sure all the following packages are installed on the system:

- Cvs
- zlib1g-dev
- newt header
- bison
- ncurses-dev
- libssl-dev
- libnewt-dev
- initrd-tools
- procps

If any of these packages are not installed. We can use aptitude application to install it:

```
# apt-get install <PACKAGE_NAME>
```

3.5.3. Installation on Fedora Core

The following packages need to be installed prior to installing Asterisk:

- Bison
- bison-devel
- ncurses
- ncurses-devel

- zlib
- zlib-devel
- openssl
- openssl-devel
- gnutls-devel
- gcc
- gcc-c++

In order to check for the availability of the packages, type:

```
# rpm -q <PACKAGE_NAME>
```

If any of those packages are not installed install them by using yum

```
# yum install <PACKAGE_NAME>
```

Important Note: Fedora does not install the kernel sources into the `/usr/src/<kernel version>` like other Linux distributions. The default place for kernel's sources is `/usr/src/kernels/<kernel-version>`.

3.6. AMR Support in Asterisk

In order for Asterisk to support AMR codec, the source code needs to be patched and recompiled. The patch adds AMR-NB support to Asterisk. For Installing AMR Patch, follow these instructions(García Murillo, 2007):

1. Create the asterisk directory

```
# mkdir asterisk
# cd asterisk
```

2. Checkout fontventa repository. This repository contains the patch and the Makefile required for incorporating the 3GPP AMR C-codes into the asterisk codes.

```
# svn checkout http://sip.fontventa.com/svn/asterisk/fontventa
```

3. Checkout Asterisk. In case you are using Bristuff, you can skip this step and use the asterisk provided in Bristuff package.

```
# svn checkout
http://svn.digium.com/svn/asterisk/branches/1.4/asterisk
# cd asterisk/
```

4. Add AMR to Asterisk

```
# patch --dry-run -p0 < ../fontventa/amr/amr-asterisk-patch.txt
# patch -p0 < ../fontventa/amr/amr-asterisk-patch.txt
# cd codecs
# ln -s ../../fontventa/amr/amr_slin_ex.h
# ln -s ../../fontventa/amr/slin_amr_ex.h
# ln -s ../../fontventa/amr/codec_amr.c
# mkdir amr
# cd amr
```

5. Download AMR code from 3GPP website

```
# wget http://www.3gpp.org/ftp/Specs/archive/26_series/26.104/26104-
700.zip
# unzip -j 26104-700.zip
# unzip -j 26104-700_ANSI_C_source_code.zip
# ln -s ../../../fontventa/amr/Makefile
```

6. Build Asterisk

```
# cd ../..
# ./configure
# make
```

6. Configure AMR: app_h324m encodes AMR inside the ast_frame in RTP octed aligned mode. (RFC 4867 section 4.4).

To configure the AMR codec to use octed aligned mode, add this to codecs.conf:

```
[amr]
octet-aligned=1
```

In order to verify that the AMR codec is properly installed type:

```
*CLI> core show codecs
```

Figure 6 shows the results showing that AMR codec is properly loaded and working.

```
*CLI> core show codecs
Disclaimer: this command is for informational purposes only.
        It does not indicate anything about your configuration.
        INT    BINARY       HEX   TYPE       NAME    DESC
-------------------------------------------------------------------
          1 (1 <<  0)      (0x1)  audio       g723   (G.723.1)
          2 (1 <<  1)      (0x2)  audio        gsm   (GSM)
          4 (1 <<  2)      (0x4)  audio       ulaw   (G.711 u-law)
          8 (1 <<  3)      (0x8)  audio       alaw   (G.711 A-law)
         16 (1 <<  4)     (0x10)  audio   g726aal2   (G.726 AAL2)
         32 (1 <<  5)     (0x20)  audio      adpcm   (ADPCM)
         64 (1 <<  6)     (0x40)  audio       slin   (16 bit PCM)
        128 (1 <<  7)     (0x80)  audio      lpc10   (LPC10)
        256 (1 <<  8)    (0x100)  audio       g729   (G.729A)
        512 (1 <<  9)    (0x200)  audio      speex   (SpeeX)
       1024 (1 << 10)    (0x400)  audio       ilbc   (iLBC)
       2048 (1 << 11)    (0x800)  audio       g726   (RFC3551)
       4096 (1 << 12)   (0x1000)  audio       g722   (G722)
       8192 (1 << 13)   (0x2000)  audio        amr   (AMR NB)
      65536 (1 << 16)  (0x10000)  image       jpeg   (JPEG image)
     131072 (1 << 17)  (0x20000)  image        png   (PNG image)
     262144 (1 << 18)  (0x40000)  video       h261   (H.261 Video)
     524288 (1 << 19)  (0x80000)  video       h263   (H.263 Video)
    1048576 (1 << 20) (0x100000)  video      h263p   (H.263 Video)
    2097152 (1 << 21) (0x200000)  video       h264   (H.264 Video)
```

Figure 6- Loaded Codecs in Asterisk (notice AMR Codec)

During our installations, it appeared the code in codecs/amr did not build in the process of building asterisk. To solve this issue:

In codecs/Makefile change this section:

```
$(LIBAMR): @$(MAKE) -C amr
```

To this:

```
$(LIBAMR): @$(MAKE) -C amr all
```

3.7. Bristuff

BRIstuff is a package provided by www.junghanns.net for enabling BRI functionality in Asterisk using ISDN BRI cards. It is set of patches and BRI drivers, along with download and patching scripts for Asterisk, Zaptel and Libpri. After the patches have been applied, Asterisk can use BRI telephony interface cards from (such as HFC chipsets that we use in our testbed platform) through the Zaptel channel driver interface (chan_zap). Some features of Bristuff include:

Support for ISDN/BRI in Asterisk ZAP channels
Asterisk zaphfc: module driver for supporting many simple ISDN cards that use the Cologne HFC-s chipset.

Bristuff is essentially a distribution of Asterisk with many modifications (Cohen, 2007). It has an install script that downloads some specific versions of Zaptel, libpri and Asterisk, patches them and installs them. It is necessary that we do not mix the versions of these packages. Otherwise it is possible that we break the compatibility and not get the excepted results.

3.8. Asterisk Configuration

The configuration files for Asterisk are stored in /etc/asterisk. Except for zaptel.conf file which is the configuration file for zaptel module; all the configuration files we refer to are located in this directory. They can be edited using any text editor in Linux. Each configuration file has a specific syntax that we have to get familiar with and follow when configuring various settings in Asterisk (such as dial plans and SIP phones.

3.8.1. Configuring ISDN line

ISDN Devices can be configured in either NT or TE mode:

NT Mode: NT stands for network terminator. Network terminator acts as the interface between an ISDN user and the ISDN provider. It can be a small hardware box to which the

user has to connect the ISDN devices via an interface called *S0*, or it can be integrated into the ISDN card.

When connecting multiple devices to the ISDN connection, the network terminator (NT) behaves as master and synchronizes the communication on the S0 bus. All other device will behave as slaves. This functionality of the network terminator is called NT mode. Not all ISDN devices are normally capable of running in NT mode. Some special ISDN cards with HFC chipsets can run in NT mode, and can directly communicate with other ISDN user devices via a crossed cable.

There are various channel driver modules for ISDN devices that can be used in Asterisk. However, NT mode is only supported by the mISDN, zaphfc, vISDN and the sirrix channel drivers(Digium, 2007). You will need your device to be in NT mode if you want to connect your asterisk server to a PBX and the PBX cannot be put into NT mode.

TE mode: TE stands for Terminal Equipment; an ISDN telephone is an example of this, although a PBX could also be configured to be in TE mode. As a rule of thumb, when connecting to ISDN phones, the PBX will need to use NT mode. When connecting PBX'es together, one of them will need to be in TE mode and the other one in NT mode.

3.8.2. Channel Configuration

In order to configure the channel, the Zap channel module needs to be configured and loaded to work with Asterisk. It allows Asterisk to communicate with the Zaptel device driver, used to access the telephony interface cards (in this case the ISDN bri interface .The interface parameters are configured in zaptel.conf and Asterisk's Zap channel module is configured via the zapata.conf file.

3.8.2.1. Configuration File /etc/zaptel.conf

The zaptel.conf file is where the required interface parameters are configured for the Zaptel card.Within the zaptel.conf file, first the type of signaling that the channel will use is defined as well as the number and the type of channels to load. Theses settings in the configuration file will then be used to configure the channels with the ztcfg command as seen in Figure 7.

Ztcfg program parses the zaptel.conf file and configures the hardware elements in the system. Three main elements are configured in the zaptel.conf file(Tims, 2008):

- An identifier for the interfaces on the card within the dialplan (this is an
- The type of signaling that will be used for the interface
- The tone language associated with a particular interface, as found in zonedata.conf. By specifying this parameter, channels used by the system can be set to give familiar tones. For example by setting the `loadzone` to `UK`, British users can hear familiar UK tones.

The `loadzone` and `defaultzone` options need to change from:

```
loadzone=nl
defaultzone=nl
```

To:

```
loadzone=uk
defaultzone=uk
```

Also to configure the signaling for the HFC based ISDN BRI card the following options need to be set within zapte.conf file;

```
span=1,1,3,ccs,ami
bchan=1-2
dchan=3
```

3.8.2.2. Zap Channel Module Configuration

Before running ztcfg program, we need to load the Zaptel driver module in the system. We can do it by running the following commands:

```
# modprobe zaptel
# insmod zaphfc.ko modes=1
```

In order to make the driver modules load at the system start-up, edit the file `/etc/rc.d/rc.local` and add these lines:

```
# modprobe zaptel
# insmod /home/Mohammad/bristuff-0.4.0-test6/zaphfc/zaphfc.ko
# sleep 10
# ztcfg –vvv
```

```
# ztcfg -vvv

Zaptel Version: 1.4.7.1
Echo Canceller: MG2
Configuration
=====================

SPAN 1: CCS/ AMI Build-out: 133-266 feet (DSX-1)

Channel map:

Channel 01: Clear channel (Default) (Slaves: 01)
Channel 02: Clear channel (Default) (Slaves: 02)
Channel 03: D-channel (Default) (Slaves: 03)

3 channels to configure.

Changing signalling on channel 1 from Unused to Clear channel
Changing signalling on channel 2 from Unused to Clear channel
Changing signalling on channel 3 from Unused to HDLC with FCS check
```

Figure 7- Zaptel configuration results

3.8.2.3. Configuration File /etc/asterisk/zapata.conf

For one hfc card, the signalling setting should be changed from:

```
signalling = bri_cpe_ptmp
```

to

```
signalling = bri_net_ptmp
```

Because BT circuits are ISDN2e we must set the pridialplan to unknown in order to set up the D-channel properly:

```
pridialplan=unknown
```

In order to check whether the ISDN line is up and running, we can check the status of the span associated with the ISDN BRI card. The status must show *Up* and *Active* to show that the line is properly working:

```
*CLI> pri show span 1
Primary D-channel: 3
Status: Provisioned, Up, Active
Switchtype: EuroISDN
Type: CPE (PtMP)
Window Length: 0/7
Sentrej: 0
SolicitFbit: 0
Retrans: 0
Busy: 0
Overlap Dial: 0
T200 Timer: 1000
T203 Timer: 10000
T305 Timer: 30000
T308 Timer: 4000
T309 Timer: -1
T313 Timer: 4000
N200 Counter: 3
```

3.8.3. Dial Plan Configuration

The dial plan in Asterisk is where the behaviour of all connections through the PBX is configured. By defining various dial plans we control how the incoming and outgoing calls are handled and routed.
The format of the extensions lines is fairly simple:

```
exten => extension number, command priority, command
```

Every line of the dial plan must start with an **exten =>**, which indicates to the asterisk that there is a command to be followed on this particular line. The **extension number** can be a digit or a character and is the number that the caller is trying to reach. This is the number of the ISDN landline in case of our Testbed platform. The **command priority** is the order in which the commands have to be followed by asterisk. **Command** is the issued instructions to Asterisk, telling it what to do. There are several options and configurations for this and usually a limited set of the options will be used for a certain type of operation, depending on the complexity of the task. The configuration file "extensions.conf" contains the "dial plan" of Asterisk. The commands used in our testbed platform can be found in Appendix-C.

3.8.3.1. Configuration File /etc/asterisk/Extensions.conf

The content of 'extensions.conf' is divided into 4 main sections. There are two categories of content which are: static settings and definitions, and executable dial plan components. The executables are also referred to as *contexts*. The settings are grouped into [general] and [globals]. Here is where the system administrator can define the names of contexts. After the [general] and [globals] categories, the rest of the extensions.conf is used for defining the dial plan. The dial plan consists of a number of contexts, each of which includes a collection of extension lines. It is also possible to use macros, which are reusable execution patterns, like procedures in any programming language. Important parts the extensions.conf used for playing back speech samples for the testbed platform can be found in Appendix-C.

3.8.4. Configuring SIP

```
# cd /etc/asterisk
```

The files we are particularly interested in are sip.conf and extensions.conf. The first thing we will need to do is editing sip.conf configuration file as shown below:

```
# vi sip.conf

[phone1]
type=friend
host=dynamic
username=User1
secret=password
dtmfmode=rfc2833 ;
context=from-sip
callerid="phone1" <1000>
```

Type: Choices are *friend*, *peer* or *user*. *Peer* is usually used when Asterisk is contacting a proxy and user is used for SIP clients that only make calls. Type f*riend* is used when the client acts as both a peer and a user.

Context: Setting the context is extremely important. In most cases this should be different from the context used in zap channel configuration. It should also be the same for all the sip clients that need to make calls to each other. If a phone is not in a valid context you will not be able to use it. A dial plan entry with the same context needs to be created in extensions.conf file in order to handle the calls in asterisk.

Host: This can be either set to *dynamic* or the IP address of the SIP client. Use Dynamic if the IP addresses are allocated by a DHCP server in your network. This will tell asterisk to negotiate the IP address with the SIP client. If a DNS server exists in the network, then we can enter the client name in the **Host** field.

Dtmfmode: This field specifies how the client handles DTMF signalling. This entry is not installation specific and depends on the type of SIP phone used to connect to the Asterisk. For X-lite softphone which we have used for our project, the default setting (rfc2833) should be set. Other options are *inband*, *rfc2833*, or *info*.

```
exten => 1000,1,Dial(SIP/phone1,20,tr)
exten => 2000,1,Dial(SIP/phone2,20,tr)
exten => 3000,1,Dial(SIP/phone1&SIP/phone2,20,tr)
```

3.9. Summary

This chapter focused on the testbed platform that was built in order to transport the calls from 3G mobile network into the quality test equipment for studying the quality of the speech over live network. Asterisk open-source PBX is used as the main component of the platform and the system is connected to the mobile network using an ISDN bri interface card. Customizations has been made to the standard installation of the Asterisk such as AMR support, Q.931 channel configuration and Zaphfc and Zaptel modules, in order to meet the requirements of the designed testbed platform. Detailed setup considerations, all the required configurations for Asterisk and step-by-step instructions for building the testbed is provided in this chapter. The next chapter will cover the methods and considerations of objective and subjective measurement over live network using the quality test platform that has been set up during this chapter.

4. Methodology and Experiment Design

The second main objective of the project is to collect live speech samples and evaluate the quality of the recorded samples using objective quality measurement tools such as PESQ(ITU-T P.862) and 3SQM(ITU-T P.563) .

4.1. Selection of speech samples

(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 at most 12*s* long should be used for the experiments. However, because PESQ can be applied to speech up to 30 s long, each speech sample can be 8-30*s* long including any silence before, after and between sentences.
Speech samples from (ITU-T P.50) database were used for all the subjective and objective measurements. P.50 consists of several speech samples from different languages. For each language, there are 8 female and 8 male voices. The names of the files in the database are self explanatory. For example *B_eng_m1*.wav is the first male voice in the British English section of the database. The language selected for the experiments is British English and all the 8 male and the 8 female voices were used. The samples were first saved as 16bit binary raw format and converted to wav files to before encoding with GSM and AMR codecs. More details about the conversion process are provided in section 4.2 of this chapter.

4.1.1. Record and Play Software

The softwares used for playing and recording the speech samples need to be reliable and tested to ensure that it does not introduce unwanted distortions to the speech samples. The software that is used for this purpose is a Motorola speech quality test tool called *Audioscore*. The software needs to be installed on a windows platform with the customized soundcard driver for VX 440 soundcard.

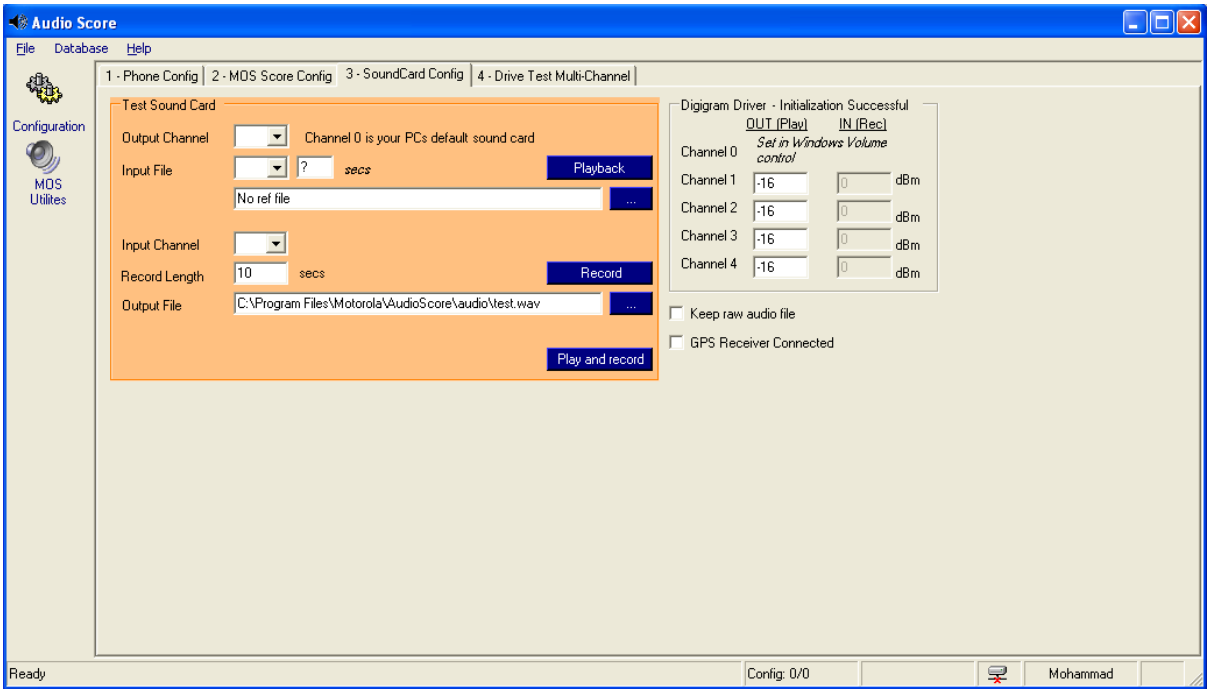


Figure 8- Audio Score Soundcard Config tab used for playing and recording speech samples

4.1.2. Sound card

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. The soundcard that is used for this project is a Digigram VX pocket 440 PCMCIA card. This sound card was suggested to us by Motorola experts that have been using this for their experiments. We were also provided with a windows driver for this soundcard specifically tailored for using with Audio Score software to record and play WAV files. Using the driver and Audio Score together we can make sure that no distortions are introduced by the operating system or the soundcard when playing or recording.

4.1.3. Cable

In order to perform play and record operation from/to the mobile handset, the handset needs to be connected to the local PCs to record the speech signals. To connect the mobile handset, 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 cable that is used for this purpose is made by modifying a Mono hands-free headset and connecting the microphone and speakers directly to XLR connectors which will then be connected to the soundcard EMU. This cable has been designed and suggested to us by the Motorola team.

Using two cables together, it is possible to experiment with two mobile phones, without having to connect through the asterisk server. This is useful especially if we are interested in analyzing delay.

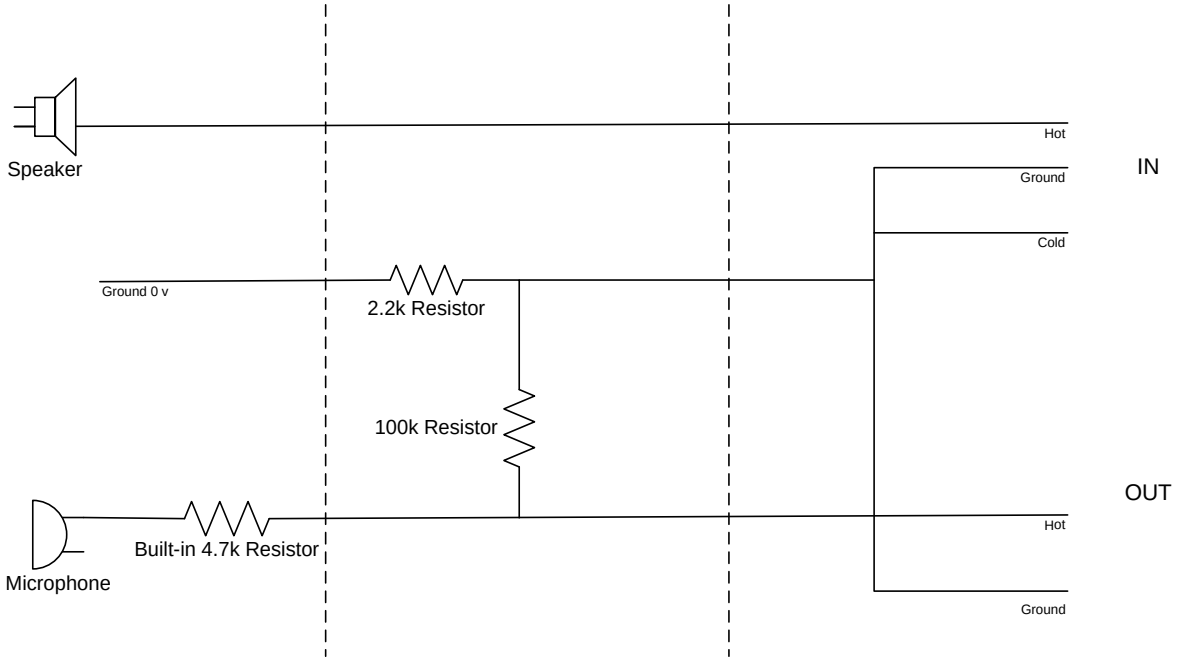


Figure 9- resistors added to the cable to match the voltage level

Figure 9 illustrates the resistor network added to the microphone circuit in order to match the voltage on the both sides. Using this setup, the cable can connect the mobile network directly to the EMU socket of the soundcard.

4.2. Encoding of the selected Sample Speech files

Codecs are mathematical models used to encode and compress analogue audio information from analogue voice signals to a digitally encoded version. Voice compression algorithms take into account the human brain's ability to interpret what we believe we should hear and form an impression from incomplete information rather than from what is actually heard.

(Meggelen *et al.*, 2005). This human brain's ability allows many of these models to gain compression by using lossy compression algorithms. However, this may come at the cost of losing the quality. Thus, depending on the algorithm used, codecs vary in the sound quality, the bandwidth required, and the computational requirements and should be selected for each system individually according to the quality requirements of that system. The purpose of the various encoding algorithms is to achieve a balance between efficiency and the quality required for the system.

Each service, program, handset, gateway or mobile operator, supports several different codecs, and may allow different codecs to pass through or talk to each other or negotiate which codec they prefer to use.

One of the purposes of this research is to evaluate the quality of the speech using various Audio codecs. The main two codecs that were studied in the experiments are AMR and GSM.

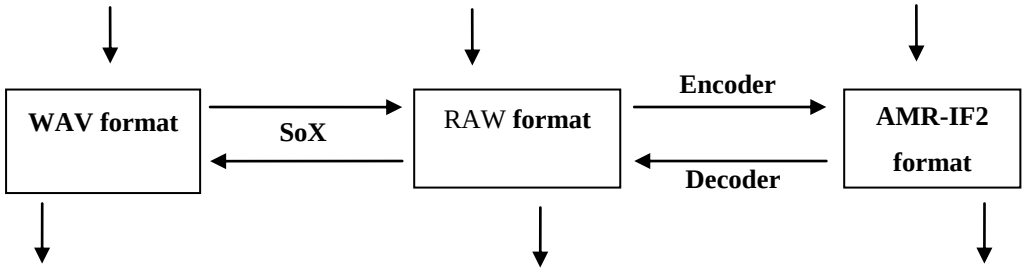


Figure 10-Input/output diagram for encoding and decoding sample audio files

4.2.1. Experiments with GSM Codec

Using the testbed platform, we can evaluate the quality of speech using different codecs. One codec that is mainly used in all the mobile networks is GSM. In order to do the experiment with GSM codecs, our reference speech samples need to be converted to GSM format. The resulted GSM files will then be play back over the mobile network and recorded on the local PCs. The reference and degraded files will then be fed to PESQ tools and the results can be analyzed to evaluate the speech quality. Figure 11 shows a schematic diagram of the GSM experiment process.

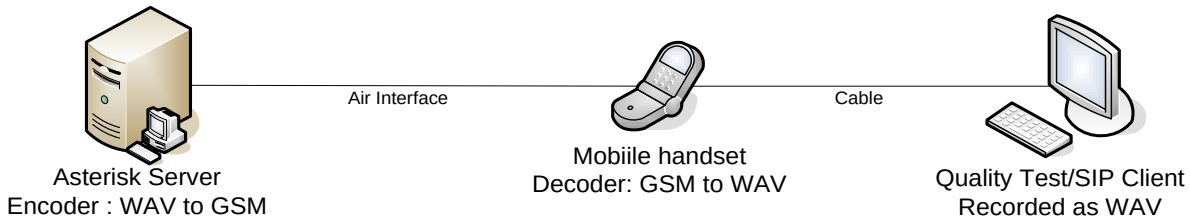


Figure 11- GSM experiment- GSM Encoding and Decoding process

To convert the reference speech files from binary files provided in ITU-T database, first SOX application was used. To convert .16P binary files to GSM format the file needs to be converted to .wav format first:

```
sox -t  sw -r 16000  <input>.16p -t wav -r 8000 <output>.wav
```

Asterisk can only play wav files with sampling rate of 8000. And our cable is a Mono cable. Therefore we use 8000 sampling rate and one channel (mono) for converting our files.
The WAV file can then be converted to GSM format:

```
sox <wavfile>.wav -r 8000 -c 1  <output>.gsm
```

The first experiments using GSM files resulted in very poor quality. One of the important points to consider when recording the files is that both reference and degraded signals MUST be recorded at the same sampling rate. If the sampling rate of the signals is not identical, The MOS scores calculated by PESQ will be very poor and not useful.
In addition to comparing the reference signals with the recorded signals, we also investigated the quality degradation resulted by only encoding and decoding the speech samples (Encoder loss). The results showed that the quality of the speech signals decreased significantly by only encoding to GSM format and then Decoding to WAV format again. Digging further into this, it appeared that this also depends on the rate of the GSM encoding. Full rate GSM and Enhanced Full Rate (EFR) GSM formats will give better results rather than normal GSM formats.

Audio files described by the following four characteristics(Bagwell, 2005):

Rate: The sample rate is in samples per second. For example, 8000 or 16000.
Data size: The precision the data is stored in. Most popular are 8-bit bytes or 16-bit words.
Data Encoding: The type of encoding algorithm used. Examples are u-law, ADPCM, or signed linear data.
Channels: The number of channels in the audio data. Mono and Stereo are the two most common.
Header-less data, or commonly referred to as raw data does not provide any information about the file. In such case, enough information must be passed to *SoX* on the command line so that it knows what type of data the file contains.

4.2.2. Experiments with AMR Codec

Experiments with AMR codec basically follow the same concept as GSM codecs. The AMR encoded speech samples with be played by the Asterisk server and recorded on the mobile side. AMR codec is a patented codec and requires license for installing in commercial systems. However, the encoder/Decoder for AMR codec is freely provided by 3GPP. There are two possible AMR frame type settings in the encoder/decoder. Specification (3GPP TS 26.101) describes these two possible frame types: *interface format 1* and *2* (abbreviated as IF1 and IF2). Theses formats describe the generic frame structure of the speech codec. IF2 frame type is a byte-aligned frame types that is used for making 3G calls in the Motorola handsets. When wrapping AMR encoded voice in .3gp files for playback in asterisk, AMR mime type with IF1 can also be used in the decoder setting. AMR files need to be encoded with IF2 format. To do this, we need to compile the AMR encoder and decoder with –IF2 option by changing the compiler options from ETSI (the default setting) to IF2:

In ‘Makefile.gcc’ file, change:

```
CFLAGS_NORM  = -O4 -DETSI
CFLAGS_DEBUG = -g -DDEBUG –DETSI
```

To

```
CFLAGS_NORM  = -O4 -DIF2
CFLAGS_DEBUG = -g -DDEBUG -IF2
```

4.2.2.1. Asterisk H324m Library

The H.324M protocol is the standard used in UMTS 3G video calls. In order to play back AMR encoded speech samples, it is necessary that this library is properly installed and the modules are loaded when starting Asterisk. This library allows Asterisk to bridge calls between a 3G mobile handset and an IP phone (SIP or H323), place/receive or record 3G calls on the Asterisk. The library deals with deals with the H223, H245, WNSRP, AMR IF2 format. It supports both MPEG-4 and h263 video formats. For playing back AMR encoded samples, each AMR encoded speech file will be wrapped in a .3gp file format and will be played back on the channel using H324 mp4_play() command. The app_mp4.c application must also be installed from http://sip.fontventa.com/. The only hardware required is one ISDN interface card (bri or pri). In order to put AMR encoded speech samples into the .3gp files the following commands can be used. The required dial plan configurations are provided in Appendix-C.

```
# mp4creator -create=<file.amr> <file.3gp>
# mp4creator -hint=1 <file.3gp>
# mp4info <file.3gp>
```

.3gp file extension is technically the same as .mp4. But the .mp4 does not support AMR IF2 format whereas .3gp does. Thus, .3gp is the correct suffix, but Asterisk and mpeg4ip do make make a distinction between the two file formats.

4.3. Objective measurements

4.3.1. Quality tests based PESQ

PESQ is a method to objectively measure end-to-end user perceived speech quality by comparing the original degraded signal. PESQ-algorithm requires a reference speech file and a degraded speech file, which is a copy of the reference processed by the system under test. These two samples will be compared in the algorithm resulting in a PESQ Raw-score. The speech samples should be natural recorded speech and artificial voice is recommended to be avoided. Also using a live network to record the degraded speech samples is preferable, since using synthetic network impairments such as noise, packet loss or jitter may not properly

model the signal characteristics during the normal operation of a mobile network. Also we should avoid recording the degraded samples at high levels where amplitude clipping may occur. The reference speech should be as distortion-free as possible. PESQ accepts both 8 kHz and 16kHz sample rates for both the reference and the degraded sample. However, Asterisk only accepts 8 kHz bit rate files. Therefore all speech samples are first converted from binary raw files to mono 8 kHz before encoding. Figure 12 demonstrates the block diagram of the objective measurements setup using PESQ.

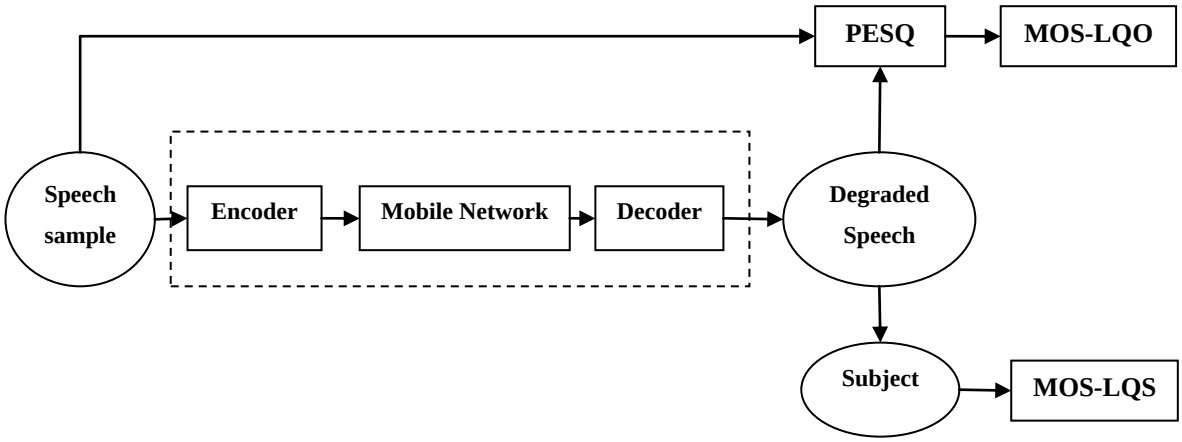


Figure 12-PESQ speech quality evaluation set up

Reference samples of 7-8 seconds length are sent through the network under test, and the received listening quality is analyzed in comparison to the original by PESQ.

4.3.1.1. ITU-T PESQ source code

The latest version of the PESQ code is version 1.2. It can be obtained from ITU-T website for free and needs to be compiled using a C compiler to work.

In order to compile PESQ C-code:

Download the compressed package from ITU-T website:

http://www.itu.int/rec/T-REC-P.862-200102-I/en

Uncompress the package:

```
# unzip T-REC-P.862-200102-I!!SOFT-ZST-E.zip
```

In the source folder, compile the C-code using gcc with –lm option. Type:

```
#  cd p862/software/Sourcecode
#  gcc -o pesq *.c –lm
```

This will give the working PESQ binary file. Alternatively, Make application can be used to compile the source code. To use make, a Makefile needs to be created. Appendix-A shows a sample Makefile created for compiling PESQ C-code.

4.3.2. Quality tests based on 3SQM

3SQM is a non-intrusive, single sided measurement, which means it is not based on a comparison with a reference signal like PESQ. Figure 13 shows the block diagram of 3SQM measurements using the testbed platform.

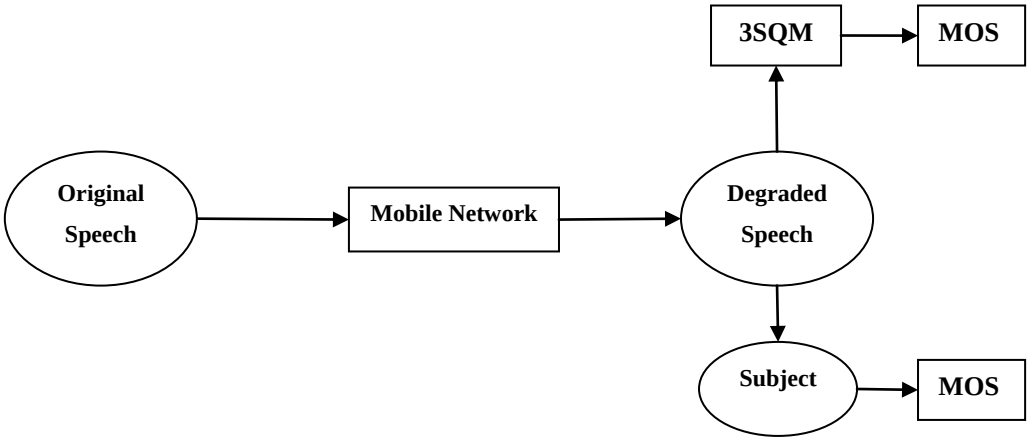


Figure 13- 3SQM speech quality evaluation set up

The source code for the 3SQM measurement tool can be obtained from 3GPP website under P.563 specification, Download the compressed package from ITU-T website:

http://www.itu.int/rec/T-REC-P.563/en

Uncompress the package:

```
# unzip T-REC-P.563-200405-I!!SOFT-ZST-E.zip
```

Compile the C-code by running the `./configure` script and then `make install` command. This will give the working P563 binary file.

4.3.3. Analysis tools for Quality measurement

Although both PESQ and Opera can compensate for small delays between playing and recording and speech samples (PESQ algorithm time aligns the original and degraded files before measuring the quality), it is sometimes required to edit the recorded samples in order to lessen the time difference as much as possible.

4.3.3.1. Audacity Audio Editing Software

Audacity is an open source digital audio editor application(Sourceforge.Net, 2008). It is a well-known tool capable of importing and exporting audio formats such as wav and MP3, recording and playing sounds and it also has a very well-designed graphical user interface for showing wave forms. It is possible to view and edit both original and degraded signal together on a same timeline. In this project, Audacity is used in conjunction with PESQ quality measurement tools for analyzing and editing .wav files. Figure 14 shows the basic interface of Audacity.

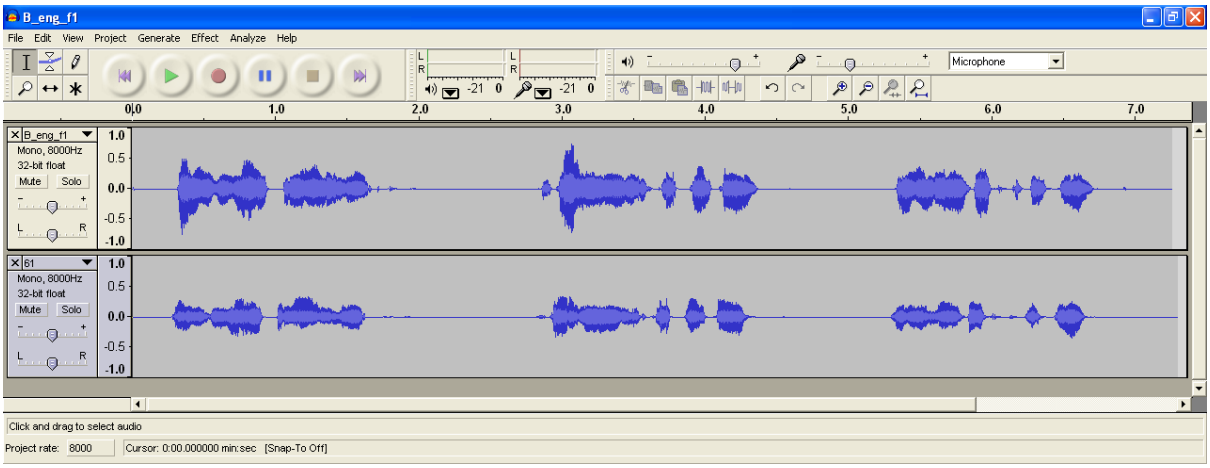


Figure 14-Audacity, Recorded and degraded signal waveform

4.3.3.2. Opera –Digital Ear

The Software provides perceptual evaluation of speech quality according to ITU-T recommendation with a graphical user interface along with other useful analysis information such as original and degraded signal wave forms, Jitter, Attenuation measured in dB, etc. Additional features that are useful for analyzing quality of speech samples include (Opticom, 2008) :

- Delay Jitter vs. Time min/max scores and graph (PESQ)
- Indicating call clarity
- Delay histogram
- Time-Signal Graphs
- Time-Quality measurement Graphs

Figure 15 shows the basic Opera interface.

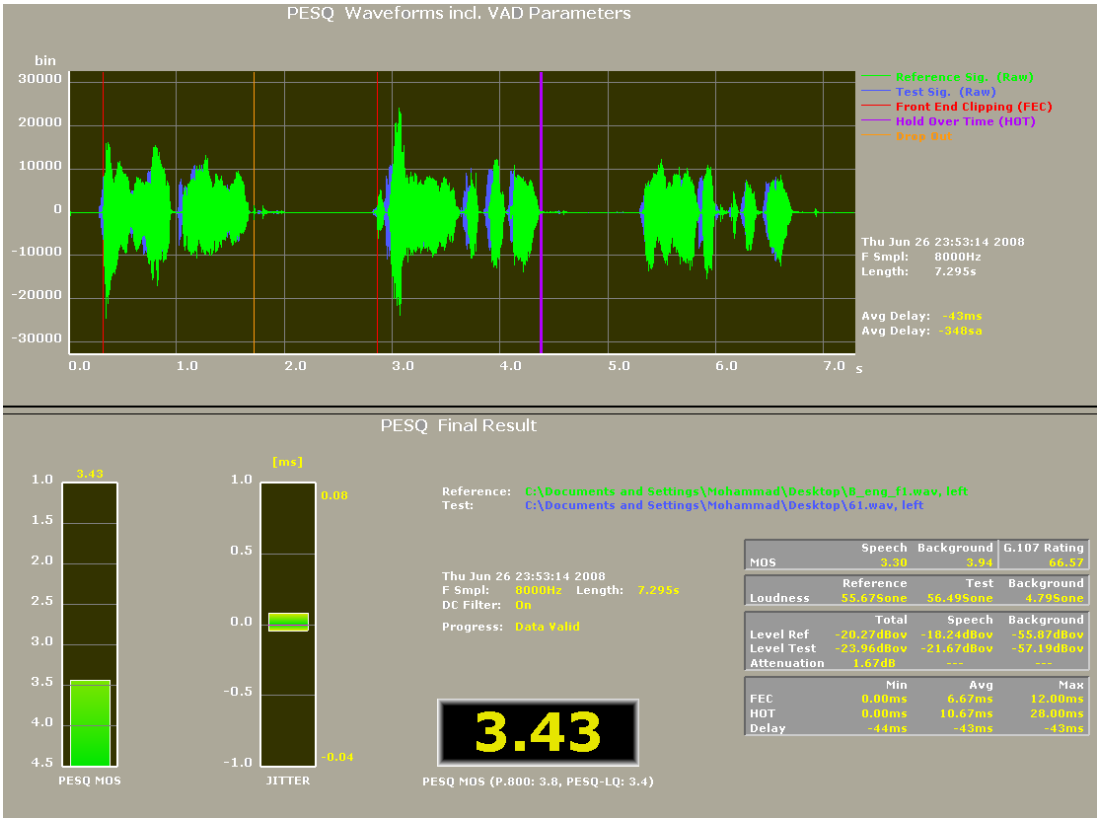


Figure 15- Opera Interface showing waveform and PESQ Final Result

Generally the final measurement results (MOS scores) given by Opera are identical to the results of ITU-T PESQ tool. However the additional information Opera provides with the results can be quite useful for analyzing the results in next stages of the project.

4.4. Subjective measurement design and considerations

The ITU recommendation P.800(ITU-T, 1996) explains how to perform a subjective speech quality measurement.

4.4.1. ITU.T P.800 subjective measurement specification

Conducting an informal speech quality assessment based on ITU.T P.800 specification requires specialist equipment such as soundproof rooms and acoustic equipment as well as voting machines and communication equipment for communicating with the subjects during the test.

According to (ITU-T P.800) specification, eligible subjects should have been selected at random from the normal service users, and should not have been directly involved in work connected with assessment of the quality of service, or related work such as speech coding; and they should not have been participated in any subjective test for at least the previous six months.

In terms of the opinion scales for the test, various five-point category-judgement scales may be used for different purposes.

4.4.2. Informal Subjective quality test procedure

The subjective test that was carried out in this research project is an informal subjective test. It has not been conducted in sound-proof facilities or in a tightly controlled environment. But all the efforts was made to comply with the ITU.T P.800 in terms of the selection of the source material, random selection of the subjects, minimum number of the subjects and the eligibility of the subjects.

The purpose of this experiment was to investigate the accuracy of the PESQ and 3SQM objective measurement models. 30 samples with the highest difference in their PESQ and 3SQM results were selected from 192 samples, 17 of which were GSM encoded samples and 13 were AMR-encoded samples. Table 2 shows the list of files used in the subjective

measurement experiment. The names of the speech files were changed as seen in table and the files were mixed together to avoid any particular ordering in the samples. The score sheets with instructions and voice files were sent out to colleagues and friends. The instructions to subjects and the score sheet are found in Appendix-D.

Table 2- Files used in the informal subjective test

Filename	DEGRADED	Operator	Gender	PESQ	3SQM
A1	../GSM-V3/B_eng_m6.wav	Vodafone	M	3.367	2.320
A2	../GSM-V3/B_eng_m5.wav	Vodafone	M	3.271	1.863
A3	../GSM-V3/B_eng_f1.wav	Vodafone	F	2.498	3.444
A4	../GSM-V2/B_eng_m8.wav	Vodafone	M	3.302	1.432
A5	../GSM-V2/B_eng_m6.wav	Vodafone	M	3.336	2.313
A6	../GSM-V2/B_eng_m5.wav	Vodafone	M	3.228	1.847
A7	../GSM-V1/B_eng_m8.wav	Vodafone	M	3.078	1.280
A8	../GSM-V1/B_eng_m5.wav	Vodafone	M	3.153	1.923
A9	../GSM-T3/b_eng_f8.wav	Three	F	2.825	3.629
A10	../GSM-T3/b_eng_f7.wav	Three	F	2.916	4.089
A11	../GSM-T3/B_eng_f2.wav	Three	F	2.579	3.349
A12	../GSM-T3/B_eng_f1.wav	Three	F	2.806	3.730
A13	../GSM-T2/B_eng_m8.wav	Three	M	2.99	1.428
A14	../GSM-T2/B_eng_m2.wav	Three	M	3.072	1.602
A15	../GSM-T1/B_eng_m8.wav	Three	M	2.955	1.485
A16	../GSM-T1/B_eng_m6.wav	Three	M	3.18	1.814
B1	../AMR-V6/b_eng_f8.wav	Vodafone	F	3.091	4.050
B2	../AMR-V6/b_eng_f7.wav	Vodafone	F	3.107	4.693
B3	../AMR-V6/b_eng_f5.wav	Vodafone	F	2.863	3.842
B4	../AMR-V6/B_eng_f1.wav	Vodafone	F	3.163	4.199
B5	../AMR-V5/b_eng_f7.wav	Vodafone	F	3.401	4.683
B6	../AMR-V4/b_eng_f8.wav	Vodafone	F	3.092	4.039
B7	../AMR-V4/b_eng_f7.wav	Vodafone	F	3.108	4.498
B8	../AMR-V4/B_eng_f1.wav	Vodafone	F	3.163	4.265
B9	../AMR-V3/B_eng_m8.wav	Vodafone	M	3.335	2.107
B10	../AMR-V3/b_eng_f7.wav	Vodafone	F	3.426	4.688
B11	../AMR-V6/b_eng_f5.wav	Vodafone	F	2.863	3.842
B12	../AMR-V2/b_eng_f7.wav	Vodafone	F	3.392	4.662
B13	../GSM-T3/B_eng_f2.wav	Three	F	2.579	3.349
B14	../AMR-V1/b_eng_f7.wav	Vodafone	F	3.109	4.563

4.5. Comparison between objective and subjective results

It is known that subjective votes vary from experiment to experiment, depending on the context of the experiment. When comparing subjective and objective scores, we should take

into account the fact that, even in similar conditions, the scores given by subjects will not be generally the same in two different subjective tests. Subjective scores are affected by many factors such as the balance of the other conditions in the test, individual preferences of each subject (Malden, 2004), the user's expectation and culture, used equipment and the quality range included in the experiment (ITU-T P.862.3). So one subjective test can not be directly compared with another subjective test. Objective quality scores do not show such behaviour because an objective model is not supposed to predict the absolute MOS of a single subjective experiment. It is therefore necessary to "compensate for these systematic differences" (Malden, 2004) before comparing subjective and objective scores. Therefore it is unrealistic to except the objective measurement results given by a model such as PESQ or 3SQM to give exactly the same scores as every subjective test.

However, one set of scores can be mapped on to another set of scores for the same condition. The difference between two sets of scores is usually a curve, plus small errors (ideally). This curve is the 'mapping function' function that approximately maps one set of scores on to the other. For the order to be preserved, the mapping function should be monotonic (one-to-one). The biggest risk according to (ITU-T, 2007) in this case, "is that the regression line is not monotonic, which in most cases can be checked visually." This means that the mapping should be constrained to be monotonic across the range of the data in order to preserve the order of the objective scores.

Equation (4-1) shows the general form of the mapping function for this operation. Using the cubic polynomial regression method, the unknown β parameters will be estimated. The completed function can then be used to scale the objective scores onto the same scale as the subjective votes.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \qquad (4\text{-}1)$$

This relationship between PESQ and 3SQM scores and MOS-LQS is modelled using a monotonic 3rd order polynomial. The polynomial function can be applied to map the objective scores for the objective model onto the same scale as MOS-LQS in this experiment. R statistical package was used for the calculations in this project(R Development Core Team, 2007).The technique applies equally to PESQ, PESQ-LQO and 3SQM scores. PESQ-LQO is generally closer to listening quality than PESQ raw score, but the comparison between either PESQ raw score or PESQ-LQO and subjective MOS is still affected by the difference in

subjective MOS scales from experiment to experiment, depending on the context of the experiment.

The ITU-T has accepted that a monotonic cubic polynomial, optimised for minimum mean squared error, which minimizes the root mean square error (RMSE) or maximizes the correlation between the two data sets, should be used for comparing subjective and objective measurements. Also the correlation coefficients for this mapping function have to be calculated separately for each experiment.

Although this is the most effective method to map between subjective and objective scores, doing this in the exact and correct way, complex mathematics and special numerical tools are required which are not easily available. For the general case using PESQ and 3SQM, it is recommended to draw a scatter-plot and add a cubic polynomial regression line and read the correlation given for the regression line(ITU-T, 2001; ITU-T, 2004).

By calculating the correlation coefficient of the data series, the closeness of the fit between the objective and the subjective scores are measured. This is calculated with the Pearson's formula (see Equation 4-2), after mapping the objective to the subjective scores:

$$r = \frac{\sum(x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (4\text{-}2)$$

In this formula, *xi* is the subjective MOS score for each sample, and $\bar{x}$ is the average over the subjective MOS votes, y_i is the mapped objective score for the samples and $\bar{y}$ is the average over the mapped objective MOS scores (Malden, 2004).

4.6. Summary

This chapter outlines the main methods implemented in this study for collecting samples and conducting quality measurements based on the established quality measurement platform. 16 British English samples from the ITU-T P.50 are selected and encoded by GSM and AMR encoder. Recording were made during weekdays and at different times of the day. 96 of the samples that had the list clipping were selected from the recorded samples and objective measurements based on PESQ and 3SQM models. *Motorola Audioscore* and Opera digital ear softwares were used for editing and analyzing the recorded speech samples together with

3SQM and PESQ ITU-T original c-code. Besides, as one of the objectives of the project was to evaluate different objective quality measurement methods and investigate the accuracy of each model, carrying out an informal subjective quality test that conforms to the ITU-T P.800 would be a good approach. Therefore 30 samples with the highest differences between their PESQ and 3SQM results were selected, and the results of the informal quality test were compared with the objective measurement results. The closeness of fit between the subjective and objective results was also calculated in order to examine the correlation of subjective and objective techniques.

5. Objective and Subjective measurement Results

5.1. Encoder/decoder effect on the speech quality

As described in previous sections, a major source of quality degradation in telecommunications networks can be the speech codec used to encode and compress the speech signals in the network. Codecs are designed to gain bandwidth by compressing some parts of the signal when encoding it to the digital version. Though depending on the type of algorithm used, this costs some degree of quality loss.

The objective quality scores of the speech samples after encoding and decoding are summarized in Figure 16. The last column in the histogram shows the percentage of the samples that achieved a score of 3.5 and above, which is an acceptable quality score known as *communication quality*. All the PESQ scores are between 3 and 4 and 50% of them are above between 3.5 and 4. PESQ-LQ results are a transformed form of PESQ raw scores and tend to not differ significantly from them except in very low quality levels. But 3SQM results are consistently lower than that of PESQ scores. 50% of the samples are in the poor (between 2 and 3) and from the other half only less than 10% of them achieved communication quality.

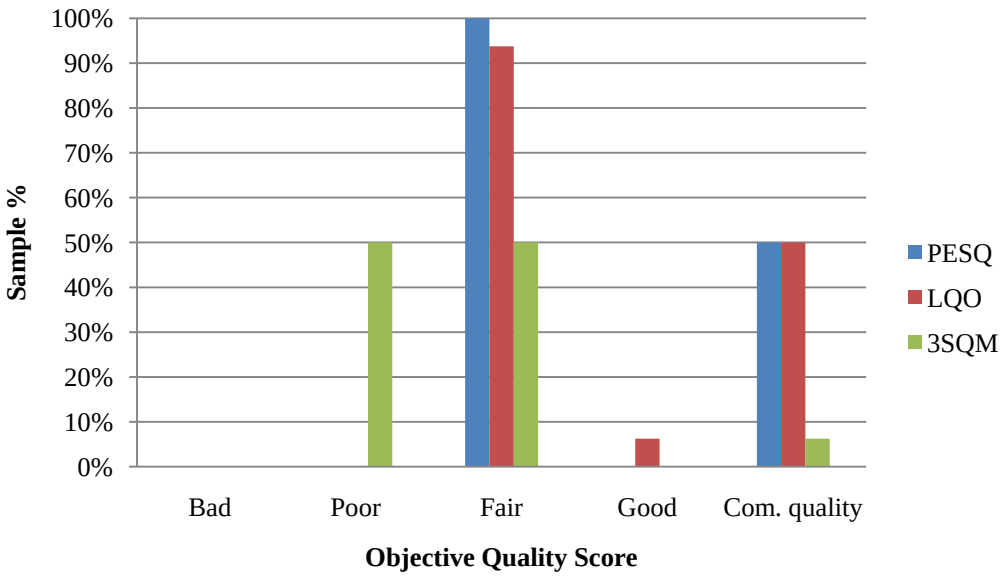


Figure 16- objective measurement results (GSM) after encoding/decoding

Table 3 shows more detailed statistical results about the objective scores after GSM encoding. The average objective MOS for 3SQM model is significantly lower than the average score for PESQ and PESQ-LQO. Average score for PESQ-LQO is slightly higher than PESQ raw scores, which is the effect of the transform function used to calculate the PESQ-LQO scores. Also the standard deviation of 3SQM results is higher than the standard deviation for both PESQ and PESQ-LQO, which shows a higher variation in the 3SQM results.

In both PESQ and 3SQM samples no sample scored 'good' or 'excellent' expect for 6.25% in PESQ-LQO that are in good category. Therefore it can be inferred that the scores will be at the most in the fair and poor category after being sent through the live network, where they will be affected by more impairments and the quality is highly expected to decrease more.

Table 3-Statistical summary of objective scores after GSM encoding/decoding

	PESQ	LQO	3SQM
Bad	0.00%	0.00%	0.00%
Poor	0.00%	0.00%	50.00%
Fair	100.00%	93.75%	50.00%
Good	0.00%	6.25%	0.00%
Com. quality	50.00%	50.00%	6.25%
Average	3.555313	3.613303	2.949592
STDEV	0.234971	0.312567	0.460223

The AMR codec (12.2 kbit/s bit rate) showed a significantly higher quality scores when compared to GSM codec results. Table 4 presents a statistical summary of the quality scores given by PESQ and 3SQM to the samples after AMR encoding.

Table 4- ITU-T samples after AMR encoding/decoding

	PESQ	LQO	3SQM
Bad	0.00%	0.00%	0.00%
Poor	0.00%	0.00%	37.50%
Fair	50.00%	31.25%	43.75%
Good	50.00%	68.75%	18.75%
Com. quality	100.00%	100.00%	37.50%
Average	3.95075	4.099259	3.31149
STDEV	0.106888	0.112818	0.615608

The difference between 3SQM and PESQ scoring is more visible in the experiments with AMR encoder. As Figure 17 shows, PESQ results are all in the fair and good quality category, and all of them have scores more than 3.5, which indicates a quite good quality. However 3SQM results are scattered between poor, fair and good and only 37.50% of them are in the communication quality category. The standard deviation for 3SQM is much higher than PESQ scores' standard deviation which again shows a higher variation in 3SQM results. Also the average score for PESQ is significantly higher than the average score for 3SQM.

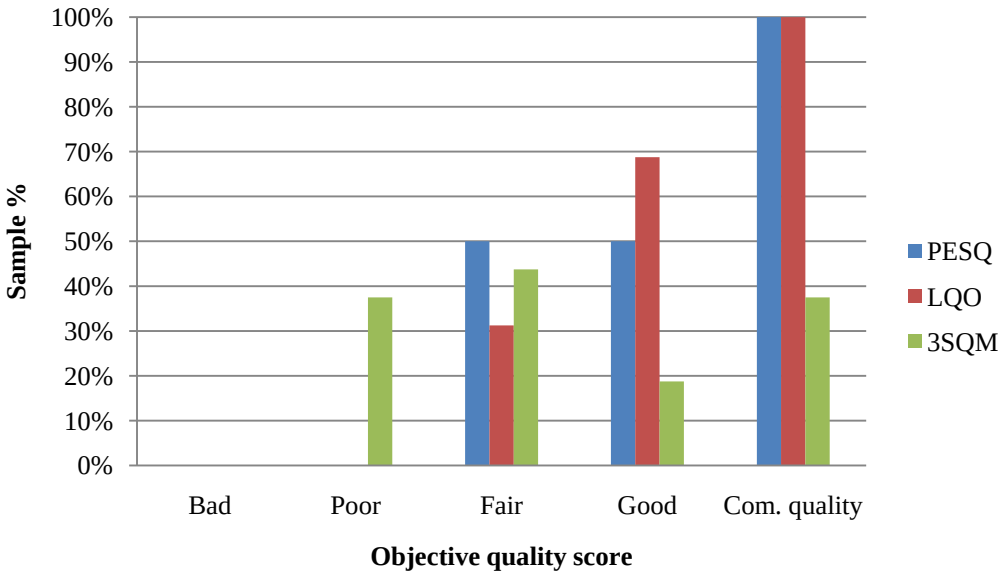


Figure 17- ITU-T samples objective measurement results (AMR)

5.2. Objective measurements on live network calls

5.2.1. Comparison between PESQ and 3SQM results

The objective measurements conducted on the 96 GSM encoded samples recorded over live mobile network resulted as excepted. From the results from the codec-only experiments, it was expected that the quality scores would drop even more when the samples were played back through the network. The quality scores for the GSM samples over live network were mostly in the poor and fair categories, except for a negligible 1.04% of 3SQM results.(see Table 5)

Also in the experiments with AMR codec (See Table 6), the majority of objective scores were in the fair category (between 3 and 4) and the rest were in poor category. No samples scored between 1 and 2 (Bad category). 15.63% of the 3SQM scores were in good category whereas no PESQ score was between 3 and 4.

Table 5- Statistical summary of objective measurments for GSM live recordings

	PESQ	PESQ-LQO	3SQM
Bad	0.00%	1.04%	11.46%
Poor	39.58%	52.08%	58.33%
Fair	60.42%	46.88%	40.63%
Good	0.00%	0.00%	1.04%
Com. quality	1.04%	3.13%	8.33%
Average	3.034167	2.880833	2.730831
STDEV	0.262414	0.371346	0.583026

Table 6-Statistical summary of objective measurements for AMR live recordings

	PESQ	PESQ-LQO	3SQM
Bad	0.00%	0.00%	0.00%
Poor	16.67%	30.21%	20.83%
Fair	83.33%	69.79%	63.54%
Good	0.00%	0.00%	15.63%
Com. quality	15.63%	21.88%	43.75%
Average	3.258583	3.201083	3.42036
STDEV	0.256977	0.371277	0.552167

3SQM results in all cases showed a lower average and a higher variation compared with PESQ and PESQ-LQO results. However, by looking at the trend lines in Figure 18 and Figure 19, we can see that both 3SQM and PESQ results follow patterns in categorizing the samples in bad, poor, fair and good quality that are similar to a degree. It can be concluded that overall there are similarities between the results of the two algorithms. But inconsistencies and differences between the quality scores for individual samples need to be investigated more in order to find out which algorithm is more accurate.

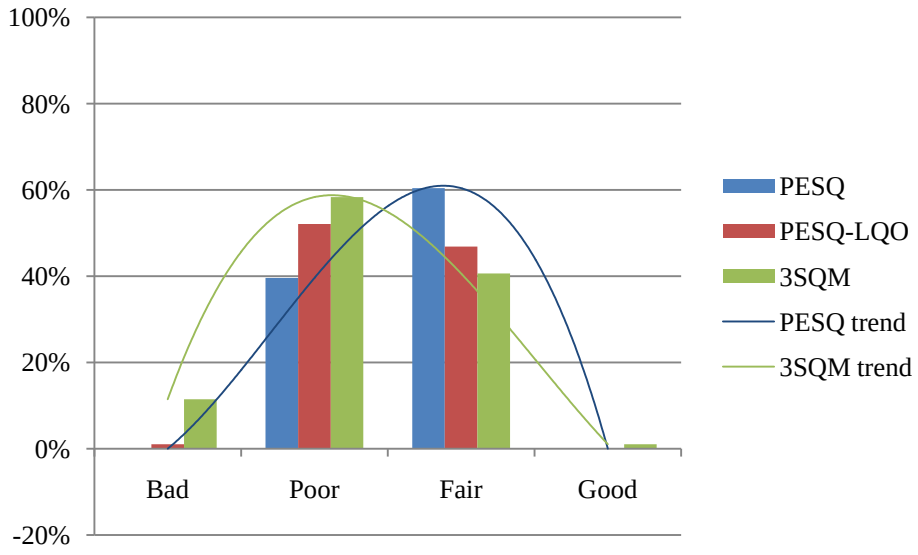


Figure 18-Objective Measurement results for GSM live recordings

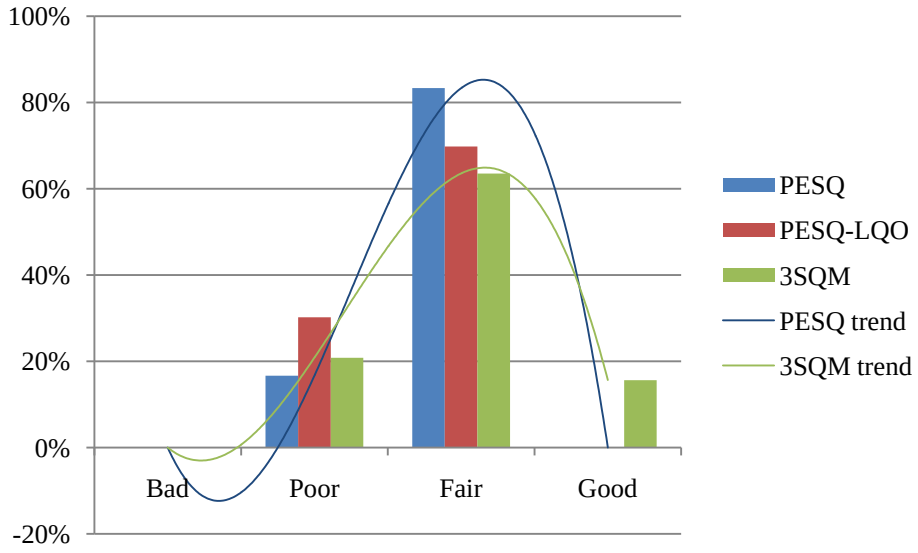


Figure 19- Objective Measurement results for AMR live recordings

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 that had the highest difference between 3SQM and PESQ MOS is shown in Figure 20 (B-eng-m8 using GSM codec) and Figure 21 (B-eng-m6 using GSM codec). As indicated in the figure, many parts of the signal, particularly at the beginning of the recording are lost, which result in a low MOS, since PESQ compares the two signals for estimating the objective MOS .Also when listening to the speech, some parts of the speech, especially the beginning is not intelligible. Therefore Regarding it as a "bad" sample with a score of 1-1.5 is reasonable.

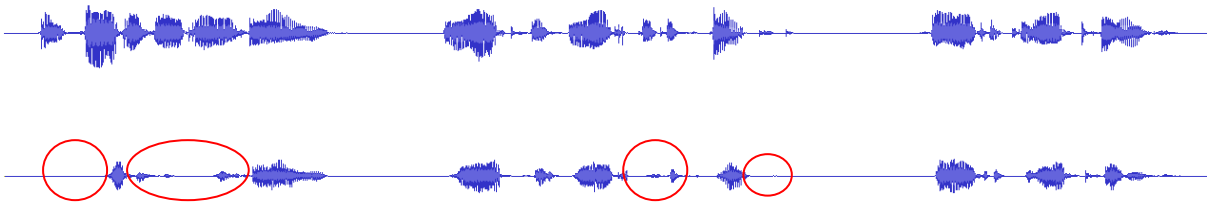


Figure 20- B-eng-m8.wav, original and degraded speech samples(Vodafone set 2)

Figure 21- B_eng_m6.wav, original and degraded speech samples (Vodafone set 3)

On the other hand, for some other samples the reason for such low scores is not so obvious. Figure 22 shows the same sample recorded in another time. The waveform is fairly well and does not seem to have a very low MOS score. However, 3SQM score is 1.43.

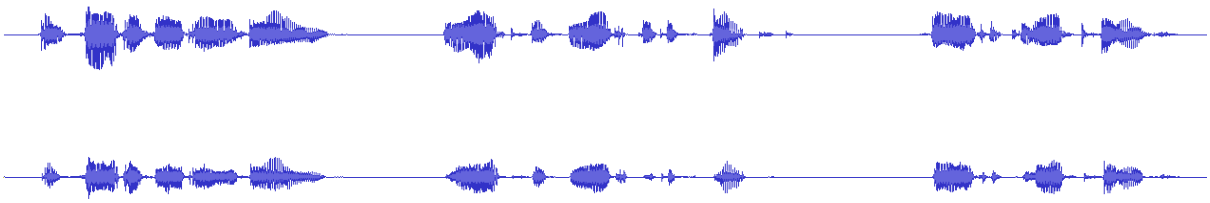


Figure 22- B-eng-m8.wav, original and degraded speech samples (Vodafone set 1)

By drawing the variations of Objective MOS over time, the results of PESQ scores can be further analyzed. Figure 23 and Figure 24 illustrate the PESQ quality score over time. In the second figure, there is no significant loss in the sample and quality scores are quite consistent despite few spikes.

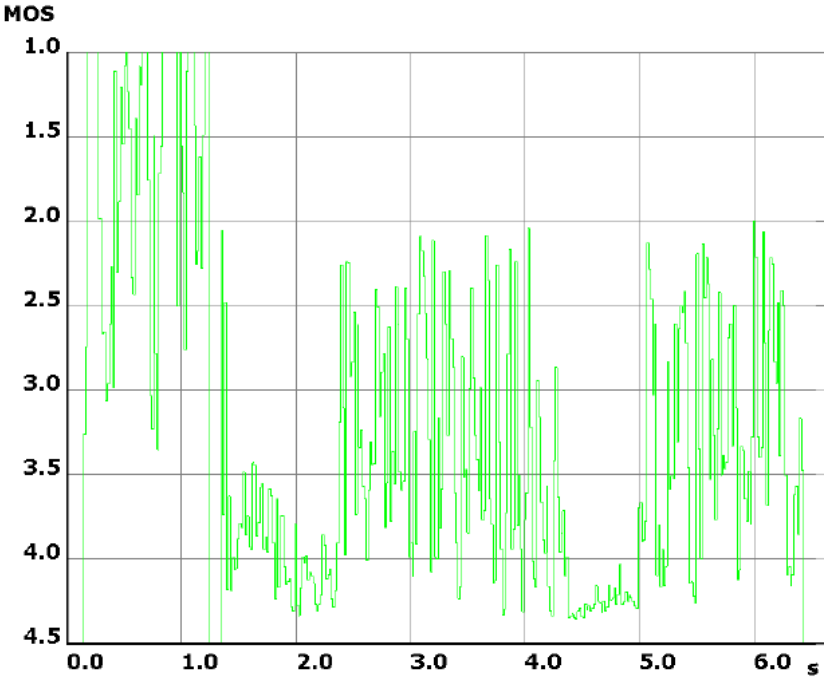


Figure 23- MOS vs. Time for B-eng-m8.wav (Vodafone set 2)

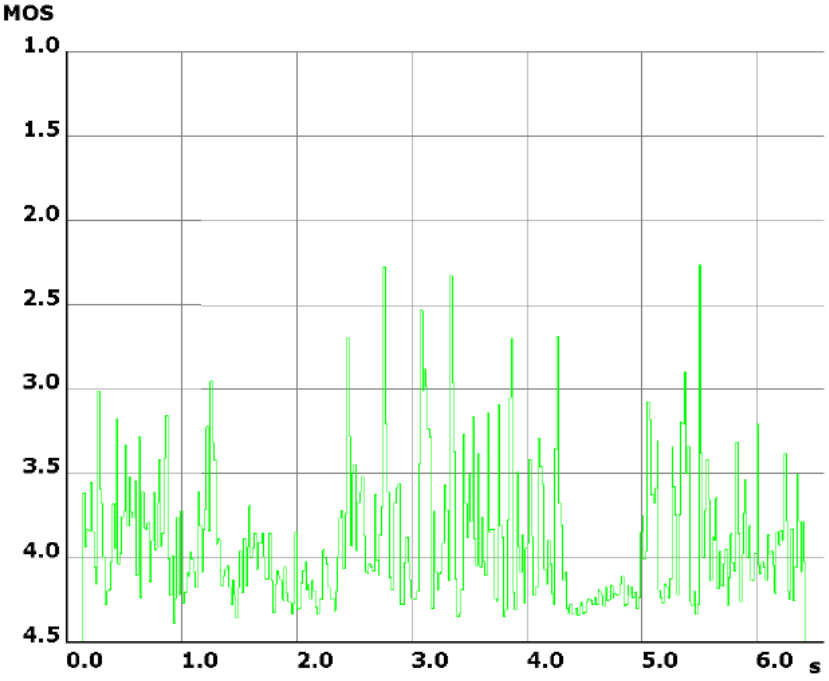


Figure 24-MOS vs. Time for B-eng-m8.wav (Vodafone set 1)

As can be seen in Figure 23, the lowest quality levels are at the beginning of the sample where the quality is in range of 1 to 1.5. The speech samples all consist of 3 short sentences separated by a silence gap. In this case the first sentence is almost completely distorted.

However the second and the third sentences have a relatively better quality especially at the end of each sentence and the quality fluctuates in range of 2 to 3.5. This may cause subjective scores to give a lower quality to the sample since most of the distortion can be heard by the subject at the beginning of the speech in the listening test and particularly in this case the loss is located at the beginning of the speech. But PESQ's score is the average of all the scores seen in the figure and will be lower. It can be concluded that position of loss in the samples can affect the score given by subjects. Also depending on the perceptual model of objective method, objective measurements may differ in the results.

5.2.2. Impact of the talker's gender on the objective quality scores

In order to investigate the impact of the talker's gender on the speech quality, we first measured the quality of the 8 male and 8 female samples after encoding and compared the results of the objective measurement with the intention of finding any meaningful differences between the quality score of the male and female talkers. The results for GSM codec is shown in Table 7 and Figure 25.

Table 7-PESQ results for GSM encoding/decoding, divided by gender

	Reference	Degraded	PESQMOS	Length
Female	B_eng_f1.wav	../gsm/B_eng_f1.wav	3.422	8 sec
	B_eng_f2.wav	../gsm/B_eng_f2.wav	3.281	8 sec
	b_eng_f3.wav	../gsm/b_eng_f3.wav	3.294	8 sec
	b_eng_f4.wav	../gsm/b_eng_f4.wav	3.408	8 sec
	b_eng_f5.wav	../gsm/b_eng_f5.wav	3.289	8 sec
	b_eng_f6.wav	../gsm/b_eng_f6.wav	3.293	8 sec
	b_eng_f7.wav	../gsm/b_eng_f7.wav	3.453	8 sec
	b_eng_f8.wav	../gsm/b_eng_f8.wav	3.288	8 sec
Male	B_eng_m1.wav	../gsm/B_eng_m1.wav	3.737	8 sec
	B_eng_m2.wav	../gsm/B_eng_m2.wav	3.63	8 sec
	B_eng_m3.wav	../gsm/B_eng_m3.wav	3.822	8 sec
	B_eng_m4.wav	../gsm/B_eng_m4.wav	3.752	8 sec
	B_eng_m5.wav	../gsm/B_eng_m5.wav	3.843	8 sec
	B_eng_m6.wav	../gsm/B_eng_m6.wav	3.664	8 sec
	B_eng_m7.wav	../gsm/B_eng_m7.wav	3.882	8 sec
	B_eng_m8.wav	../gsm/B_eng_m8.wav	3.827	8 sec

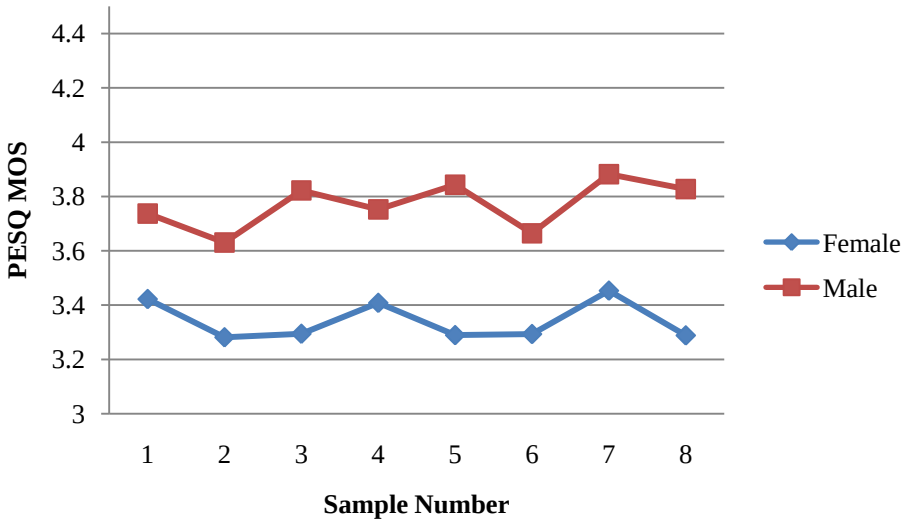


Figure 25-GSM codec encoding/decoding results for British English Samples

As it can be seen in Table 7 and Figure 25, in GSM encoded samples, male voices have a higher PESQ raw score and therefore higher PESQ-LQO results. The box plots in Figure 26 show a visible higher average PESQ score for male samples approximately close to the maximum score observed for female samples.

In order to show whether these results are statistically significant, T-test hypothesis test was used. Null hypothesis is rejected with p value of almost zero. For PESQ and PESQ-LQO the confidence interval was 0.21 to 0.39 and 0.32 to 0.56 respectively, which confirms that the PESQ scores for male samples are significantly higher than of female samples.

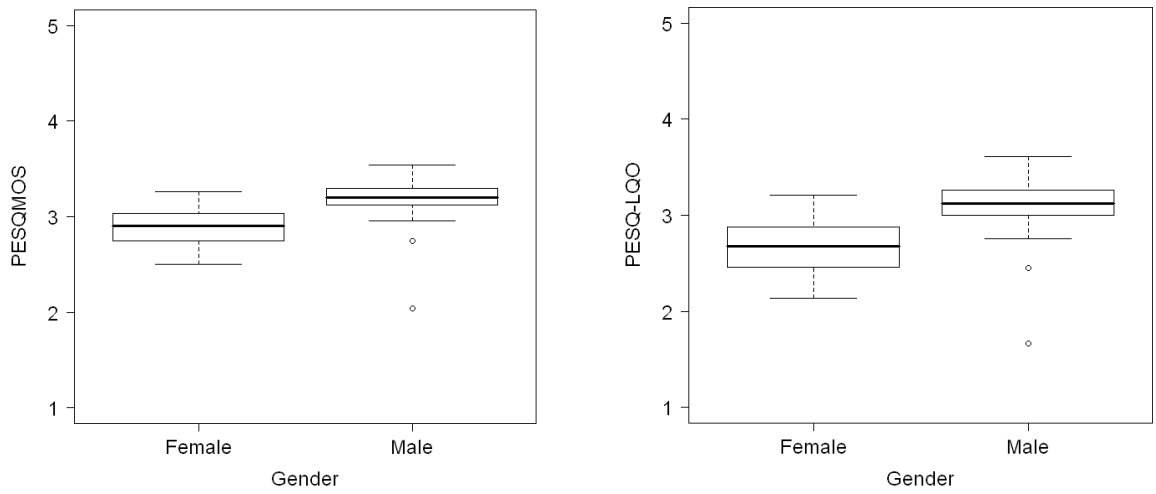


Figure 26-PESQ and PESQ-LQO quality score for male and female talkers in GSM experiments

In contrast, 3SQM results show higher PESQ scores for female samples as shown in Figure 27. This is rather an interesting result since the higher PESQ score for male samples was considered to be due to the way that GSM codec functions. However this results shows that this is dependent on the measurement algorithm rather than the codec in this case.

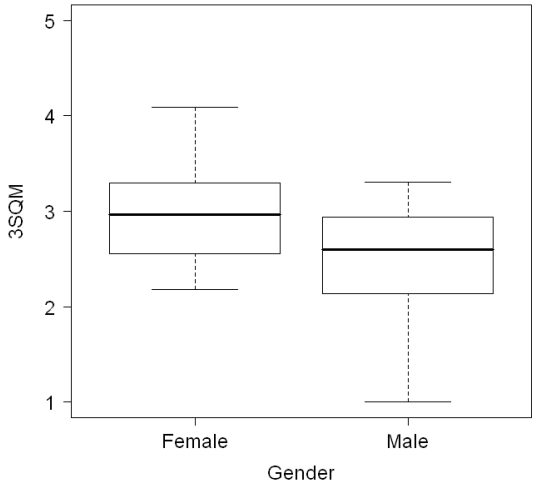


Figure 27- 3SQM quality score for male and female talkers in GSM experiments

Like the results from GSM experiment, PESQ results for samples encoded with AMR also show higher scores for male samples.

Table 8- PESQ results for AMR encoding/decoding, divided by gender

	Reference	Degraded	PESQMOS	Length
Female	B_eng_f1.wav	../amr/B_eng_f1.wav	4.109	8 sec
	B_eng_f2.wav	../amr/B_eng_f2.wav	3.81	8 sec
	b_eng_f3.wav	../amr/b_eng_f3.wav	3.805	8 sec
	b_eng_f4.wav	../amr/b_eng_f4.wav	4.035	8 sec
	b_eng_f5.wav	../amr/b_eng_f5.wav	3.844	8 sec
	b_eng_f6.wav	../amr/b_eng_f6.wav	3.812	8 sec
	b_eng_f7.wav	../amr/b_eng_f7.wav	4.005	8 sec
	b_eng_f8.wav	../amr/b_eng_f8.wav	3.783	8 sec
Male	B_eng_m1.wav	../amr/B_eng_m1.wav	3.909	8 sec
	B_eng_m2.wav	../amr/B_eng_m2.wav	3.995	8 sec
	B_eng_m3.wav	../amr/B_eng_m3.wav	4	8 sec
	B_eng_m4.wav	../amr/B_eng_m4.wav	4.024	8 sec
	B_eng_m5.wav	../amr/B_eng_m5.wav	3.965	8 sec
	B_eng_m6.wav	../amr/B_eng_m6.wav	4.023	8 sec
	B_eng_m7.wav	../amr/B_eng_m7.wav	4.071	8 sec
	B_eng_m8.wav	../amr/B_eng_m8.wav	4.022	8 sec

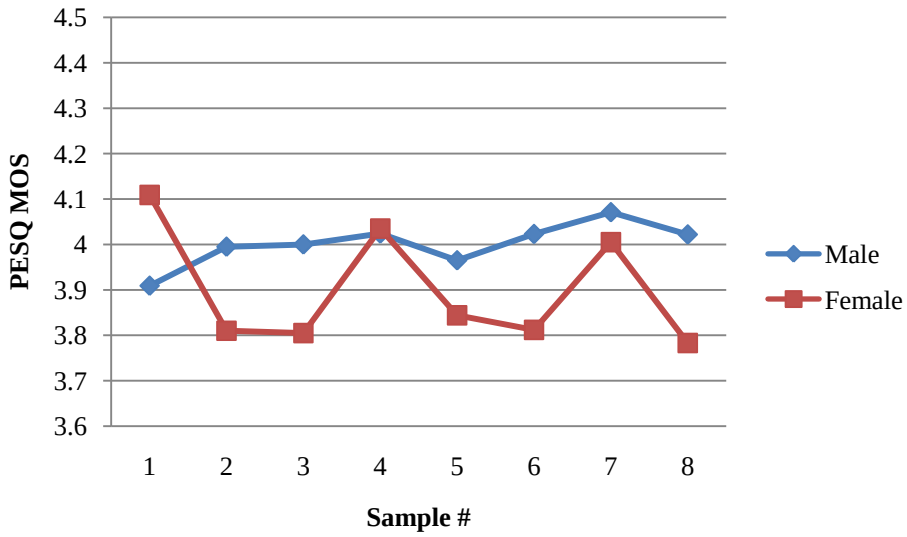


Figure 28-AMR Codec encoding/decoding results for British English Samples

By looking at the PESQ results presented in Table 8 and Figure 28, the gender of the caller has a significant effect on the quality of the voice. Male samples have a higher average quality score and a significant higher minimum point as can be seen in Figure 29. This hypothesis can be investigated using t-test statistical hypothesis test. T-test results also showed that the gender of the caller has a significant effect on the voice quality in both PESQ and PESQ-LQO results. However, 3SQM results yield the opposite. The right box plot in Figure 29 show that male samples have a lower average quality score (by 0.3), and a lower minimum point, which is contrary to the results of PESQ experiments.

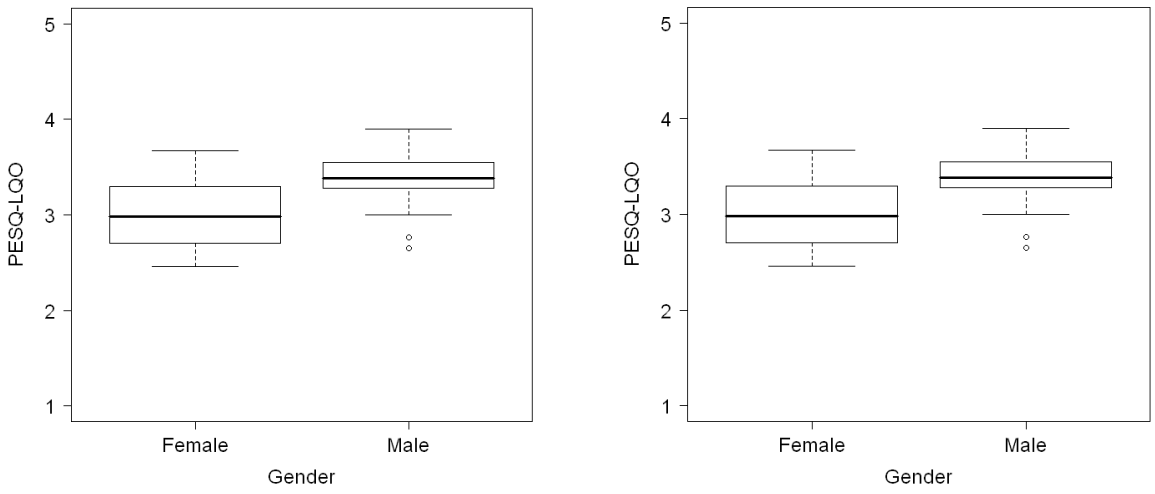


Figure 29-PESQ-LQO and 3SQM scores for AMR samples divided by gender

5.2.3. Impact of the Time of call on the objective quality scores

In order to investigate the impact effect of the time of call on the quality of the speech samples, the samples were recorded at 3 different times of the day during *weekdays.* The results were then categorised into the 3 recording times and the averages were compared with each other, as Figure 30 and Figure 31 show.

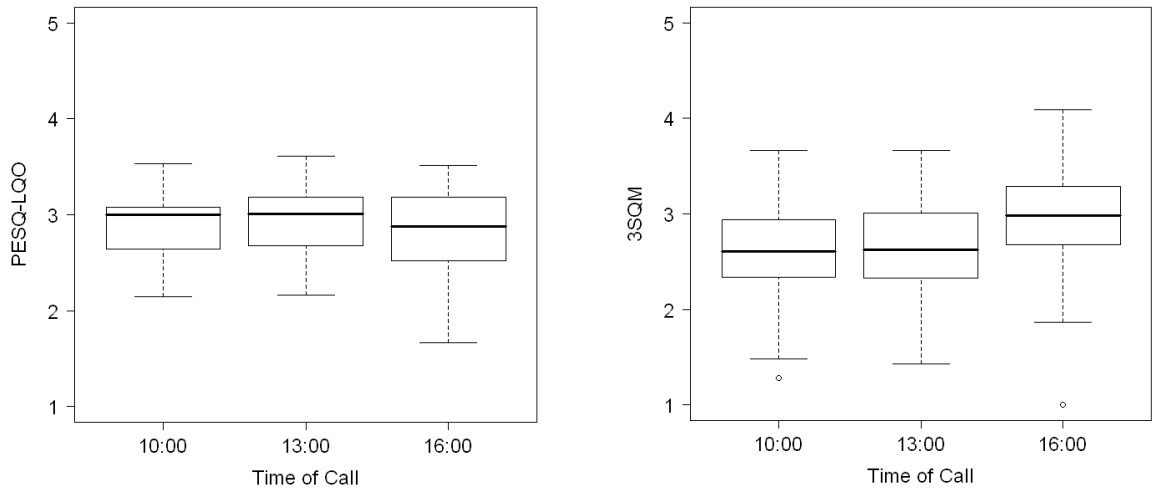


Figure 30-PESQ-LQO and 3SQM scores for GSM encoded samples grouped by the time of call

Samples were recorded at 10:00am, 13:00 and 16:00 and two sets were recorded for each time. The average PESQ score for the samples recorded at 13:00 and 10:00 are almost equal but the average score for the samples recorded at 16:00 are lower than other samples. Interestingly, 3SQM samples show the exact opposite result for the samples recorded at 16:00.

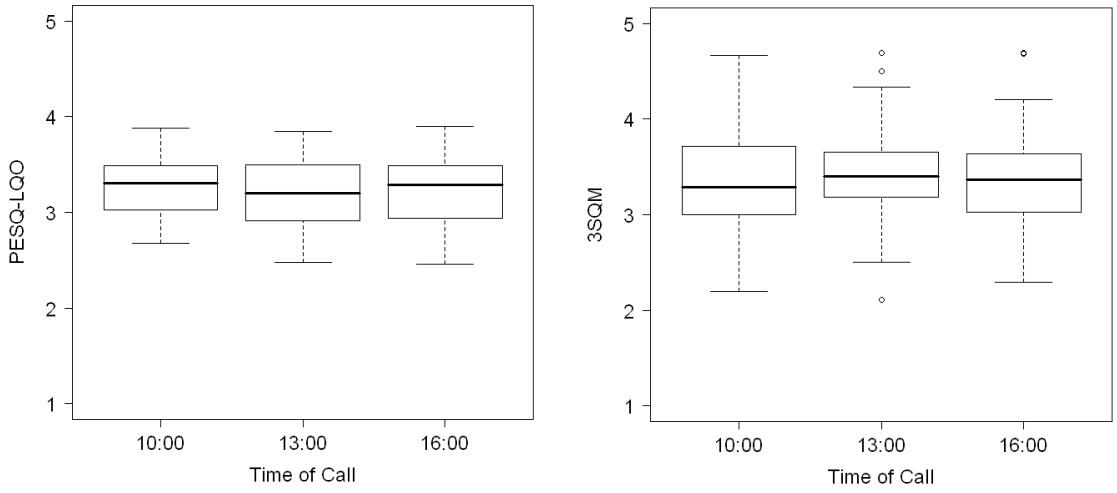


Figure 31- PESQ-LQO scores for AMR encoded samples grouped by the time of call

Despite the small natural differences between the samples, the time of call did not have any meaningful impact on the quality of the calls in the experiments in general. However, as the quality of service in mobile networks is dependant on many factors such as network load, number of customers, the infrastructure of the network, distance of the caller from the base station, and radio conditions, these results can not be generalized and are only used for this experiment to narrow down the parameters that has had an impact on the speech quality of the recorded calls. More graphs showing the results of this experiment for PESQ can be found in Appendix-F.

5.2.4. Does the Mobile Operator affect the objective quality scores?

Figure 32 compares the PESQ scores between the two operators used for recording the GSM samples. The comparison between the quality scores of two operators were only possible for the experiments with GSM codec, since video calls from mobile to landlines are blocked in 'Three' operator and the AMR samples were recorded only over 'Vodafone' network. It can be seen that the average score for recordings over Vodafone is slightly higher than the 'Three' samples. Despite this small difference in the average scores, results do not present evidence for a meaningful difference in the quality levels of the networks. Therefore it can be concluded that in similar recording conditions, the network over which the sample has been recorded has not affected the quality of the call. This conclusion is only limited to this case study and can not be generalized unless more controlled experiments are carried out to investigate this effect.

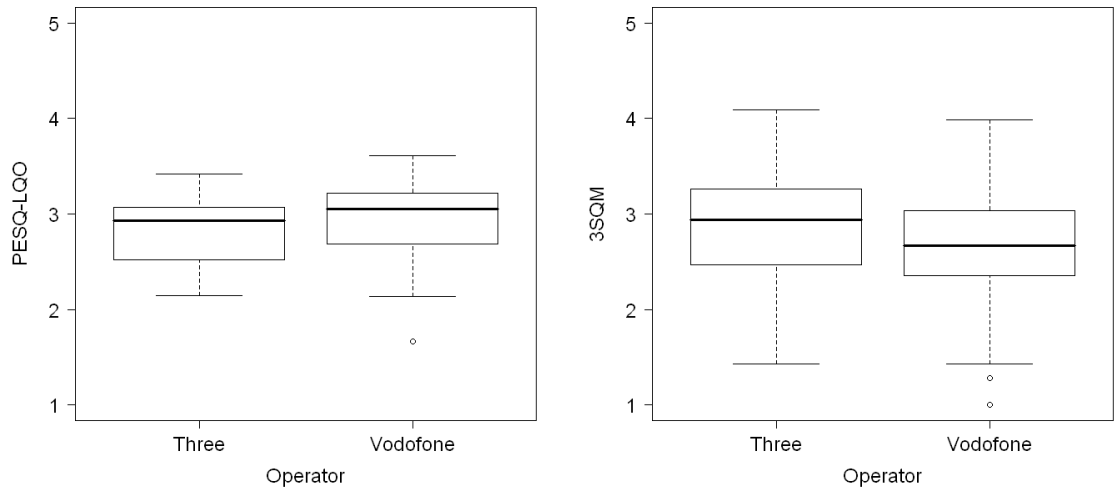


Figure 32- PESQ-LQO and 3SQM results grouped by network operator

5.2.5. Effect of the volume setting of the handset on the quality

Figure 33 compares the 6 sets of recordings made with AMR-encoded samples. The recordings were made with different volume settings on the handset. It can be seen that the 4th and 6th set of samples have relevantly lower average results. Knowing that PESQ algorithm's score does not take the loudness into account, the results could be because of the clipping on higher volume parts that can change the degraded speech samples and affect the results. On the other hand, it is interesting that 3SQM results of the same samples do not yield any significant difference between 6 sets. (Qiao *et al.*, 2008) also reports in their case study that different volume settings have different effect on the PESQ test result. When volume increases too much, clipped speech causes to a lower PESQ.

Table 9- Time and volume setting of the AMR recorded samples

Set 1: 16 samples, Weekday, 10:00 AM, Volume setting=6
Set 2: 16 samples, Weekday, 10:00 AM, Volume setting=7
Set 3: 16 samples, Weekday, 13:00 PM, Volume setting=6
Set 4: 16 samples, Weekday, 13:00 PM, Volume setting=7
Set 5: 16 samples, Weekday, 16:00 PM, Volume setting=6
Set 6: 16 samples, Weekday, 16:00 PM, Volume setting=7

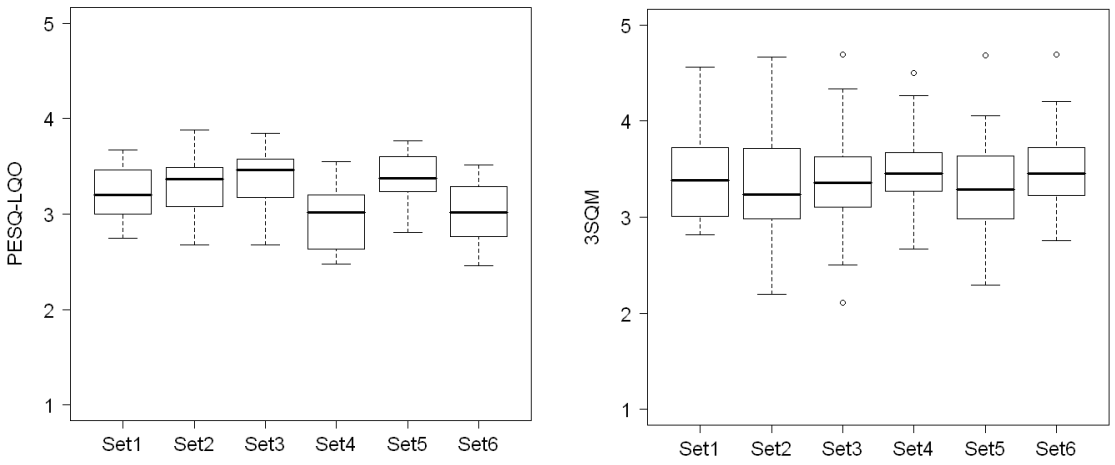


Figure 33-PESQ and 3SQM results of AMR samples, grouped by time and volume level

5.3. Informal Subjective Test

The subjective test conducted in this research project was an informal subjective test meaning that the test was not carried out in sound proof facilities and the participants were not invited to complete the test in a controlled environment. However efforts have been made for the test to conform to the ITU-T standards for subjective evaluation of voice quality in telephone networks in this study(ITU-T P.8001996).

5.3.1. Participants

50 subjects were initially asked to complete the test, from which 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 year's old. Also 3 out of 33 used speakers to listen the samples and the rest used earphones to complete the experiment.

5.3.2. Selection of Test Material

As the purpose of the subjective test was to investigate the accuracy of PESQ and 3SQM results, the main criteria in selecting the 30 speech samples used in this informal subjective test was the difference between the PESQ and 3SQM of the results. 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 issue discussed in 5.2.2, some female and some male samples were selected (12 male, 18 female). Since the subjects of were voluntarily participating in this test and there were no incentives to offer to the subjects for completing the experiments, only 30 samples were used to keep the experiment length between 10-15 minutes.

5.3.3. Test procedure

Instruction sheets and score sheets used were compiled based on the guidelines provided by ITU-T P.800. The subjects were asked to first adjust the volume setting of their computer using the reference signal provided with the samples listening quality. Once the subjects were confident with the volume level of their computer, they were instructed to listen to the speech samples only once and write down their opinion score in the score sheet, based on the Listening-quality scale provided in the instruction sheet. The speech samples, along with the instructions, were sent by email to the subjects and the completed score sheets were gathered via email as well.

5.3.4. Subjective Test 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 for most of the cases. This indicates that the individual results differ quite significantly from subject to subject. Nevertheless, it is known that people have quite different opinions and expectations when it comes to the quality. Overall, when comparing the results of subjective and objective tests, in most cases objective results correspond to the subjective results with random errors.
Table 10 shows the results of the informal subjective test. The *MOS-LQO and PESQMOS* columns show the results produced by the PESQ software, the column with *MOS* shows the average opinion score of the 33 subjects who completed the experiment and the last column shows the standard deviation of the subjective results. The complete results of the subjective test can be found in Appendix-F.

Table 10- Results of the informal subjective test

Speech File	PESQMOS	MOS-LQO	3SQM	MOS	STDEV
GSM-V3/B_eng_m6.wav	3.367	3.366	2.320	2.485	0.939
GSM-V3/B_eng_m5.wav	3.271	3.226	1.863	2.364	1.025
GSM-V3/B_eng_f1.wav	2.498	2.133	3.444	2.606	0.899
GSM-V2/B_eng_m8.wav	3.302	3.271	1.432	1.455	0.617
GSM-V2/B_eng_m6.wav	3.336	3.321	2.313	3.121	1.053
GSM-V2/B_eng_m5.wav	3.228	3.162	1.847	2.939	1.059
GSM-V1/B_eng_m8.wav	3.078	2.939	1.280	3.455	0.869
GSM-V1/B_eng_m5.wav	3.153	3.051	1.923	2.758	1.001
GSM-T3/b_eng_f8.wav	2.825	2.567	3.629	3.333	0.957
GSM-T3/b_eng_f7.wav	2.916	2.699	4.089	2.844	0.723
GSM-T3/B_eng_f2.wav	2.579	2.234	3.349	2.219	0.792
GSM-T3/B_eng_f1.wav	2.806	2.54	3.730	2.844	0.574
GSM-T2/B_eng_m8.wav	2.99	2.808	1.428	3.515	0.870
GSM-T2/B_eng_m2.wav	3.072	2.929	1.602	3.818	0.727
GSM-T1/B_eng_m8.wav	2.955	2.756	1.485	3.788	0.820
GSM-T1/B_eng_m6.wav	3.18	3.091	1.814	3.212	0.960
AMR-V6/b_eng_f8.wav	3.091	2.958	4.050	3.939	0.899
AMR-V6/b_eng_f7.wav	3.107	2.982	4.693	3.848	0.906
AMR-V6/b_eng_f5.wav	2.863	2.621	3.842	3.438	1.076
AMR-V6/B_eng_f1.wav	3.163	3.065	4.199	3.758	0.867
AMR-V5/b_eng_f7.wav	3.401	3.415	4.683	3.969	0.782
AMR-V4/b_eng_f8.wav	3.092	2.959	4.039	4.094	0.734
AMR-V4/b_eng_f7.wav	3.108	2.983	4.498	3.879	0.927
AMR-V4/B_eng_f1.wav	3.163	3.065	4.265	3.844	0.884
AMR-V3/B_eng_m8.wav	3.335	3.32	2.107	4.030	0.847
AMR-V3/b_eng_f7.wav	3.426	3.451	4.688	3.909	0.843
AMR-V6/b_eng_f5.wav	2.863	2.621	3.842	3.594	0.911
AMR-V2/b_eng_f7.wav	3.392	3.403	4.662	4.063	0.801
GSM-T3/B_eng_f2.wav	2.579	2.234	3.349	2.344	1.096
AMR-V1/b_eng_f7.wav	3.109	2.985	4.563	4.094	0.856

5.3.5. Comparison between Subjective and objective tests

Table 11 compares the average scores of objective models and the subjective MOS from the informal subjective test. In GSM samples, average 3SQM score is lower than the average subjective scores and average PESQ-LQO results seem to be closer to the average subjective votes. On the other hand, in AMR samples, average 3SQM is closer to the average subjective scores. The average of quality scores over all the 30 samples shows that objective results taken as a whole are fairly linked to the subjective results.

Table 11- Comparison between Objective and subjective average quality score results

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 12 presents a more detailed statistical breakdown of the subjective and objective results. It appears that 3SQM has more accurate results in higher quality levels and PESQ was more accurate in the fair category. 50% of the subjective results reached communication quality (are over 3.5) which is quite a different result compared with PESQ results whereas the 3SQM results show the exact percentage of the samples in the communication quality group.

Table 12- statistical summary of the subjective test results

	PESQ	LQO	3SQM	MOS-LQS
Bad	0.00%	0.00%	30.00%	3.33%
Poor	33.33%	56.67%	10.00%	33.33%
Fair	66.67%	66.67%	23.33%	50.00%
Good	0.00%	0.00%	36.67%	13.33%
Com. quality	0.00%	0.00%	50.00%	50.00%
Average	3.074933	2.9385	3.168085	3.30303
STDEV	0.248953	0.35939	1.22352	0.673707

Table 13 and Table 14 also present a partial comparison between the objective and subjective results for GSM and AMR samples.

Table 13- Partial statistical summary of subjetive test results for GSM codec

	PESQ	PESQ-LQO	3SQM	MOS-LQS
Bad	0.00%	0.00%	0.00%	0.00%
Poor	0.00%	0.00%	52.94%	5.88%
Fair	47.06%	58.82%	11.76%	52.94%
Good	52.94%	41.18%	35.29%	41.18%
Com. quality	0.00%	0.00%	17.65%	17.65%
Average	3.104944	2.980333	3.54545	3.707071
STDEV	0.212726	0.309646	1.244794	0.397828

Table 14- Partial statistical summary of subjetive test results for AMR codec

	PESQ	PESQ-LQO	3SQM	MOS-LQS
Bad	0.00%	0.00%	0.00%	0.00%
Poor	6.67%	23.33%	3.33%	0.00%
Fair	36.67%	20.00%	6.67%	30.00%
Good	0.00%	0.00%	33.33%	13.33%
Com. quality	0.00%	0.00%	40.00%	40.00%
Average	3.121875	3.00775	3.997274	3.693182
STDEV	0.286521	0.414533	0.892859	0.540972

In terms of the differences between the quality scores of male and female samples, the average subjective MOS for the 12 male samples was 3.078 and for the 18 female samples the average was 3.543, which shows that the female samples have a higher quality scores than the male samples. By comparing these results with the results of the objective measurements for the previous 192 objective measurements discussed in 5.2.2, it appears that the subjective results are more consistent with the 3SQM results in terms of categorizing the quality scores by gender (3SQM results also showed a higher average score for the female samples. However, we should consider that the subjective scores taken from only 30 samples from the 192 previously recorded samples, with a certain criteria and does not reflect the results of the entire objective measurements. Figure 34 compares the objective and subjective scores of the 30 samples used in the subjective measurements.

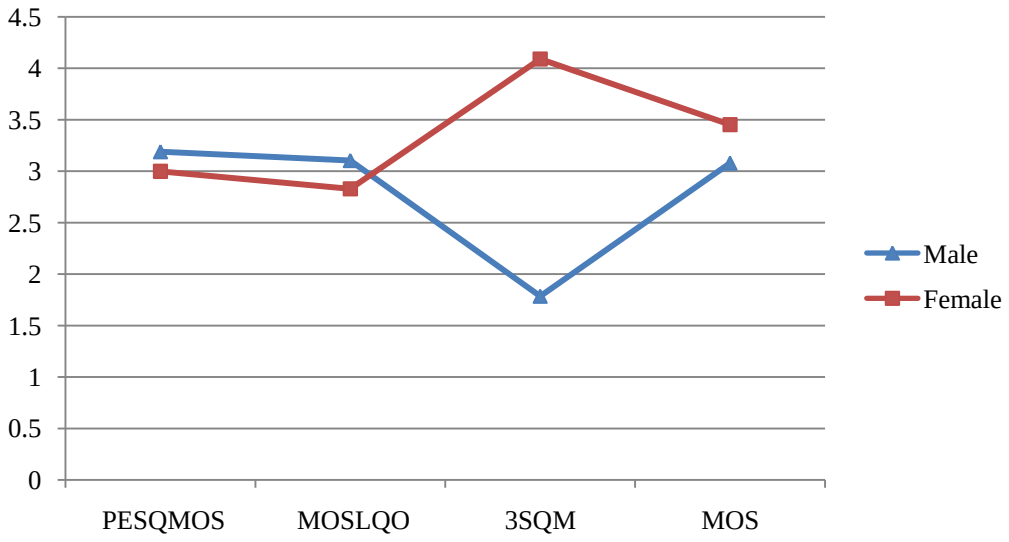


Figure 34 - Comparison of the taker's gender effect on objective and subjective score

Although both subjective MOS and 3SQM results show that the female talkers have a relatively higher average quality score than that of the male talker, the average scores for subjective results are closer to the average PESQ scores, in which the male talkers have a higher average quality score. The figure also shows a big gap between the results of male and female samples for 3SQM measurements. It sounds as if the 3SQM results are overestimated for female talkers and underestimated for male talkers. Considering that the main criteria for the selection of the samples for the subjective test has been the difference between the PESQ and 3SQM results, the gender issue can possibly be deduced as a main reason for the differences between the quality scores of the two objective methods. In order to further analyze the accuracy of the objective measurements, the correlation of each method with the subjective results needs to be investigated.

5.3.6. Correlation of Subjective and Objective measurements

As explained in the methodologies section, the results of the subjective quality test need to be treated carefully. The objective results have to be mapped using a 3rd order polynomial regression function before the correlation coefficient is calculated. Figure 35 illustrates the scatter plots for PESQ and 3SQM results before applying the mapping function.

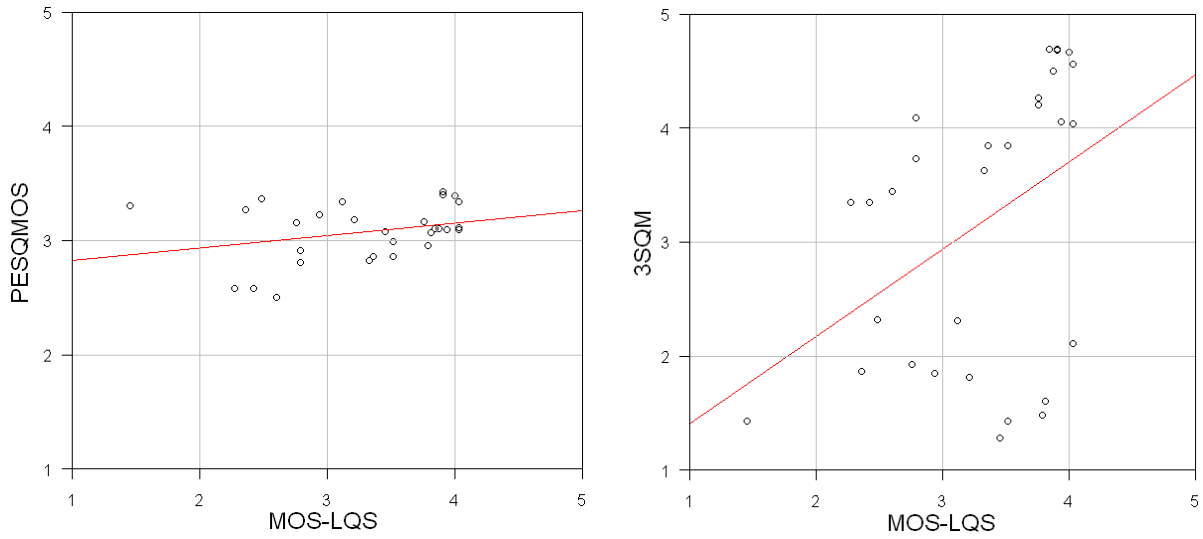


Figure 35-Objective vs. subjective measurement results before mapping

In general, the direct comparison between the results of objective and subjective tests implies that PESQ algorithm is more accurate in predicting the quality scores in medium to high quality levels and 3SQM is more accurate in higher quality levels. However 3SQM seems to

be more pessimistic in lower quality levels, which could be an explanation for the high standard deviation in 3SQM results that has been consistently seen in the experiments. By applying the mapping function, the objective scores are scaled into the subjective MOS and the correlation coefficient can be calculated for the mapped scores. Figure 36 and Figure 37 show the scatter plots of the objective versus subjective results before and after mapping. The correlation coefficient for PESQ results after mapping is 0.9433, which shows a high degree of correlation between PESQ and the 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 coefficient for 3SQM.

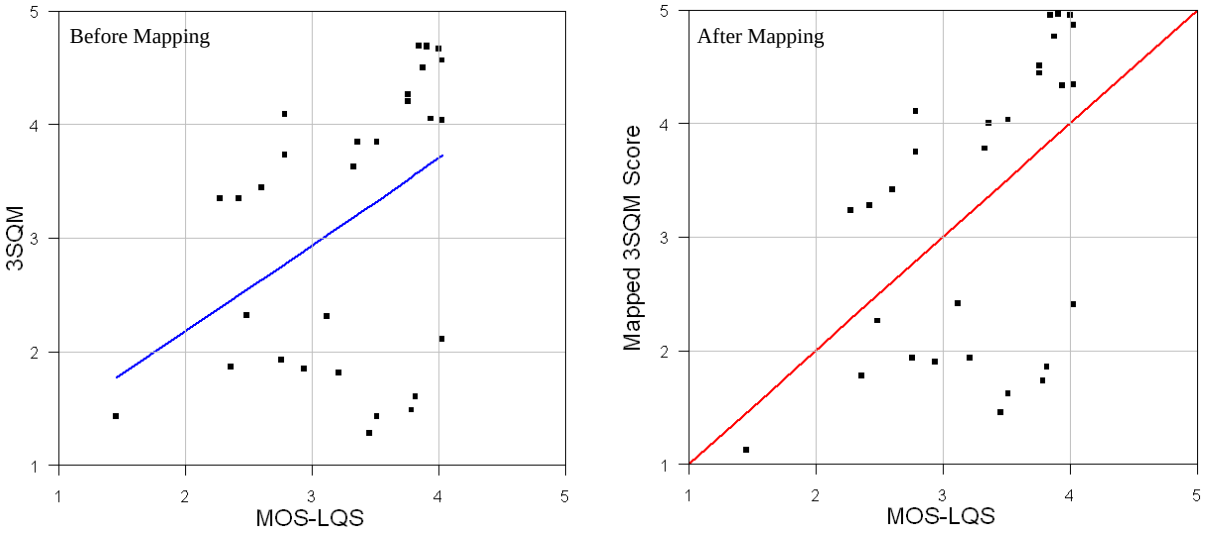


Figure 36- Mapping between 3SQM score and subjective MOS

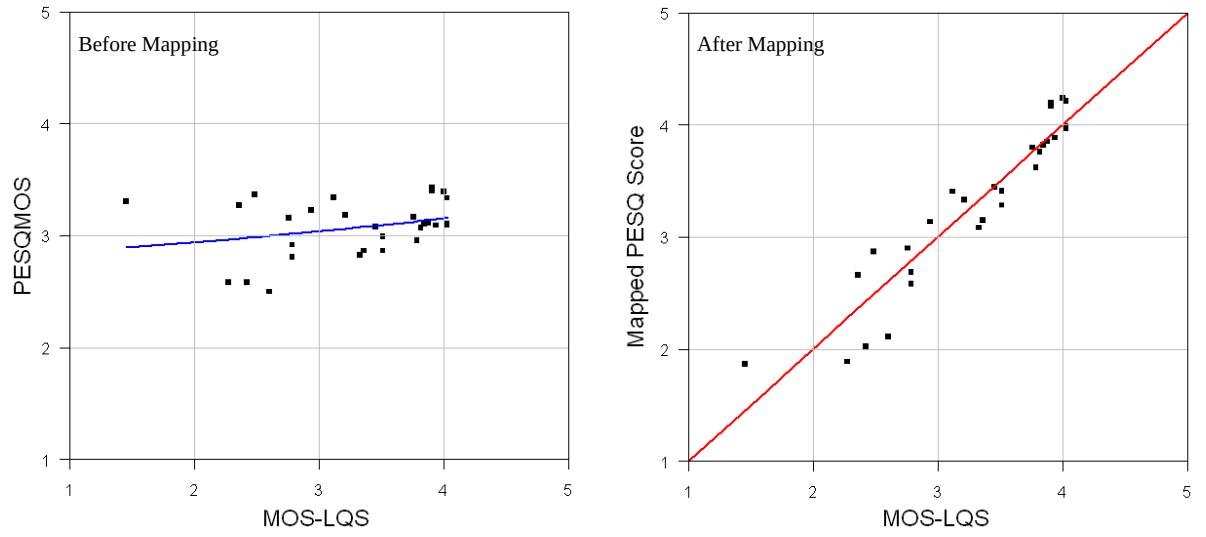


Figure 37-Mapping between PESQ score and subjective MOS

5.4. Concluding discussion

According to the correlation results, overall, PESQ and PESQ-LQO measures have a significantly better correlation with the subjective MOS and therefore are more reliable measurement techniques for predicting quality of speech in live 3G networks.

The interesting point when comparing the correlation results is that from the comparison done on the individual results in earlier section, there were cases that 3SQM predicted a better score for the samples and PESQ's prediction seemed to be less accurate. So a higher correlation result could be expected for 3SQM by doing a direct comparison between PESQ and 3SQM results. The question may raise that why such a mapping function is used that increases the correlation of the PESQ?

It should be taken into account that, objective quality measurement techniques and generally any kind of quality measurement are designed to predict an overall quality measure of the system under test and are not supposed to exactly predict every individual case. Furthermore, as explained in the methodologies section earlier, even the MOS scores of the same samples are known to vary between two different subjective tests. That is the reason why such methods of mapping between the results need to be undertaken to compensate for errors and uncontrolled variables, and analyzing the correlation between subjective and objective measurements may not be done directly or based on individual cases.

However by studying the individual cases in which one algorithm predicts the quality with a significant higher accuracy, the flaws and strengths in the algorithms can be identified and may result in improving the models and hopefully creating better prediction models in the future.

6. Conclusions and Future Work

Measurement of speech quality perceived by the customer has many constructive applications in 3G such as testing speech and channel codecs, signal processing algorithms and handsets through to entire network In 3G planning, procurement, optimization, network monitoring, upgrades and network operation.

The area of how to objectively measure speech quality is expanding fast and demands are growing for a comprehensive objective quality measurement technique. Subjective tests, however, are still more accurate but objective measurement has proved to have many constructive benefits and will continue to expand in various areas.

The main objective of this work was to investigate and evaluate the accuracy of PESQ and 3SQM objective measurement models in a live wireless 3G mobile network. A testbed platform was set up and voice quality tests were carried out for the speech signals recorded from mobile network to PSTN line through ISDN interface. The effect of a number of factors, namely voice codec, gender of the talker, time of call, and the network operator was studied. To investigate the accuracy of objective measurement results, an informal subjective test was carried out with 33 participants and the closeness of fit between the objective and subjective opinion scores was studied.

6.1. Conclusions

Below is the list of findings from the voice quality evaluations carried out during this project:

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. However, experiments with 3SQM algorithm showed relatively better MOS scores for female samples. Such inconsistencies in the results of PESQ and 3SQM show the

differences in the perceptual models and gain/attenuation compensation methods of the two models.

In the experiments, many inconsistencies between PESQ and 3SQM predicted opinion scores were observed. The results of the informal subjective test conducted to evaluate the accuracy of these two objective measurements, showed that PESQ and PESQ-LQO scores, when scaled into subjective MOS using a polynomial mapping function, have a high correlation with subjective results votes (0.943 and 0.891 respectively).

The comparison between subjective and objective results shows that with such a high correlation, PESQ can be used reliably for objective speech quality measurements in live 3G networks. However, even with such high levels of correlation in the mapped results, we cannot expect exactly repeatable results and the ability to predict the score of an alternate model. Two individual cases were found in which 3SQM predicted the quality more accurately. Also in 5 cases 3SQM results showed to be defective and PESQ was clearly more accurate. Some of the PESQ's less accurate scores can possibly be due to the loss position in the degraded speech signal; but the exact reasoning will require more deep inspections.

3SQM could not supersede PESQ's intrusive analysis as expected; since it lacks the information from the reference signal. However, non-intrusive measurements have advantages in telecommunication networks and are also useful in identifying quality in Individual tests. Therefore, we recommend a co-existence of both measures when investigating speech quality in 3G mobile networks.

The objective tests on the speech samples encoded using GSM full rate encoder shows that the impact of the encoder on the speech samples is quite significant. All of the samples scored between 3 and 4 (fair quality) after being encoded and 35.29% of them did not achieve communication quality score (>3.5). With such quality loss in the encoder, it would be perceivable that the quality scores will be much less when sent through a live wireless network, where the signal will be distorted by many more impairments. The measurement results of the recordings made through live mobile network has also confirmed this. Almost 40% of the recorded samples were of poor quality (between 2 and 3) and only a negligible 1.04% scored over 3.5. Thus, the effect of the encoder should be carefully considered in the design process of the network, as it has a major impact on the perceived quality of the speech in a mobile environment.

AMR Codec had more appreciable quality scores. All of the voice signals had a MOS score accepted as communication quality when encoded using AMR encoder with 12.2 Kb/s bit rate. After being sent through the network, 83.3% of the samples scored between 3 and 4,

which is a fair quality. 16.67% of the samples scored between 2 and 3 and 21.88% scored above 3.5.

In terms of the differences in the speech quality between the two mobile operators used in the experiments, the results for both operators showed fairly the same range of quality scores in the experiments with GSM codec, and no significant differences between the networks was observed. For AMR codec, it was found in the investigations that the first operator (THREE) has blocked 3G video call signals from mobile to landline and comparison between the two operators was therefore not possible.

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. Nevertheless, it is known that people have quite different opinions and expectations when it comes to the quality. Overall, when comparing the results of subjective and objective tests, in most cases objective results correspond to the subjective results with random errors.

6.2. Limitations of the work

One of the limitations of this project was the lack of enough subjective information for comparing with objective measurements. As discussed in the methodologies section, the subjective test that was carried out for this project is an informal subjective test; and is not as reliable as a standard subjective measurement according to ITU-T specifications. A standard subjective measurement involves large numbers of individual listening tests by different subjects in order to statistically achieve a good subjective MOS score. The small amount of subjective votes may cause a less accurate mapping of objective measurements in the statistical calculations and yield overoptimistic correlation values.

It should also be taken into account that none of the objective measurement methods used in this study provides a comprehensive evaluation of the transmission quality in the mobile environment. Other impairments related to two-way speech distortion such as loudness loss, delay, sidetone, echo, and other such parameters have also a significant impact on the true speech quality perceived by the end-user.

The other limitation of our approach is that the effect of the end-to-end delay on the speech quality has not been investigated. All the recordings have been edited and the time delay between the reference and speech samples was eliminated before running any objective or

subjective tests. The effect of delays on the speech quality was out of the scope of this research and can be a good subject for further studies.
Finally, all the recordings in this project were made on the network downlink and no measurements were done in the uplink (recording from headset's microphone to the Asterisk server).

6.3. Suggestions for future work

As described in the literature review section, there are many factors in mobile and VoIP networks that can have an effect on the perceived end-to-end quality. Future work is therefore highly advisable to investigate more into the effect of other parameters such as the behaviour of other voice codecs, language of the talker, loss pattern and loss location in live end-to-end calls over 3G networks.
Also, discussed in the methodologies section, is the use of media and channel dumps that can be highly useful in pointing out the exact reasoning behind the effect of packet loss and loss location on the quality scores. More detailed information from voice calls can be gathered with this method, which can be incorporated with the information from the speech signals and obtain more accurate, perceptually-relevant quality information.
Further improvements to the testbed platform can be made by developing AMR capabilities in the Asterisk package. Also for eliminating the effect of the ISDN line or the any interface's effect on the quality, we recommend that the influence of the ISDN line on the quality of the call should be investigated.
More statistical techniques may be used to investigate the influence of factors and analyze the distribution of subjective votes and the influence of factors such as talker or subject. Methods such *confidence interval*, *T-tests* and *ANOVA* may be useful to estimate the distribution of the observations, significance of differences between the observations, and eliminate the factors that cannot be fully controlled.

References

3GPP (2005) Mandatory speech codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec frame structure. *3GPP Specification TS 26.101*

Bagwell, C. (2005) *SoX man page*. Available online at: http://linux.die.net/man/1/sox. (Accessed: 1/5/2008)

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

Boutremans, C. and Le Boudec, J. Y. (2003) *Adaptive joint playout buffer and FEC adjustment for Internet telephony*. Le Boudec, J.Y. (Ed). INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE. pp 652-662 vol.1.

Cai, L., Tu, R., Zhao, J. and Mao, Y. (2007) Speech Quality Evaluation: A New Application of Digital Watermarking. *Instrumentation and Measurement, IEEE Transactions on*, 56, (1) 45-55.

Carvalho, L., Mota, E., Aguiar, R., Lima, A. F. and de Souza, J. N. (2005) *An E-model implementation for speech quality evaluation in VoIP systems*. Mota, E. (Ed). Computers and Communications, 2005. ISCC 2005. Proceedings. 10th IEEE Symposium on. pp 933-938.

Clark, A. D. (2001) *Modeling the effects of burst packet loss and recency on subjective voice quality*. IP Telephony Workshop.

Cohen, T. (2007) *BRIstuff*. Available online at: http://www.voip-info.org/wiki/view/Bristuff. (Accessed: 18/5/2008)

Conway, A. E. (2002) *A passive method for monitoring voice-over-IP call quality with ITU-T objective speech quality measurement methods*. Communications, 2002. ICC 2002. IEEE International Conference on. pp 2583-2586 vol.4.

Corbun, O., Almgren, M. and Svanbro, K. (1998) *Capacity and speech quality aspects using adaptive multi-rate (AMR)*. Almgren, M. (Ed). Personal, Indoor and Mobile Radio Communications, 1998. The Ninth IEEE International Symposium on. pp 1535-1539 vol.3.

Digium. (2007) *Asterisk and the BRI Cards - types and useful information* Available online at: http://www.asteriskguru.com/tutorials/bri.html. (Accessed: 15/5/2008)

Ding, L. and Goubran, R. A. (2003a) *Assessment of effects of packet loss on speech quality in VoIP*. Goubran, R.A. (Ed). Haptic, Audio and Visual Environments and Their Applications, 2003. HAVE 2003. Proceedings. The 2nd IEEE Internatioal Workshop on. pp 49-54.

Ding, L. and Goubran, R. A. (2003b) *Speech quality prediction in VoIP using the extended E-model*. Goubran, R.A. (Ed). Global Telecommunications Conference, 2003. GLOBECOM '03. IEEE. pp 3974-3978 vol.7.

Ditech. (2007) *Limitations of PESQ for Measuring Voice Quality in Mobile and VoIP Networks*. Available online at: http://www.ditechnetworks.com/learningcenter/whitepapers/WP_PESQ_Limitations.pdf. (Accessed: 1/7/2008)

Duysburgh, B., Vanhastel, S., De Vreese, B., Petrisor, C. and Demeester, P. (2001) *On the influence of best-effort network conditions on the perceived speech quality of VoIP connections*. Vanhastel, S. (Ed). Computer Communications and Networks, 2001. Proceedings. Tenth International Conference on. pp 334-339.

Ericsson. (2006) *Speech Quality Measurement with SQI*. Available online at: www.ericsson.com/solutions/tems/library/tech_papers/tech_related/Speech_Quality_Measurement_with_SQI

Fujimoto, K., Ata, S. and Murata, M. (2002) *Adaptive playout buffer algorithm for enhancing perceived quality of streaming applications*. Ata, S. (Ed). Global Telecommunications Conference, 2002. GLOBECOM '02. IEEE. pp 2451-2457 vol.3.

García Murillo, S. (2007) *AMR patch* Available online at: http://sip.fontventa.com/trac/asterisk/browser/amr/README. (Accessed: 18/5/2008)

GL Communications. (2007) *Voice Quality testing*. Available online at: http://www.gl.com/ITUalgorithms.html. (Accessed: 1/82008)

Gray, P., Hollier, M. P. and Massara, R. E. (2000) Non-intrusive speech-quality assessment using vocal-tract models. *Vision, Image and Signal Processing, IEE Proceedings* -, 147, (6) 493-501.

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

ITU-T (1996) Methods for subjective determination of transmission quality *ITU-T Recommendation P.800,August 1996.*,

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 (2003) Rec. G.107. The E-Model, A Computational Model For Use in Transmission Planning.2003.

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.*,

Karlsson, A., Heikkila, G., Minde, T. B., Nordlund, M. A. and Timus, B. A. (1999) *Radio link parameter based speech quality index-SQI*. Heikkila, G. (Ed). Speech Coding Proceedings, 1999 IEEE Workshop on. pp 147-149.

Korhonen, J. (2001) *Introduction to 3G mobile communications.* Artech House, Norwood,MA.

Mahdi, A. E. (2007) *Voice Quality Measurement in Modern Telecommunication Networks*. 14th International Workshop on Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. . pp 25-32.

Malden. (2004) *Speech Quality Assessment Background Information For DSLA and MultiDSLA Users*. Available online at: http://files.teraquant.com/malden/Speech%20Quality%20Assessment.pdf. (Accessed: 1/8/2008)

Meggelen, V., Madsen, L. and Smith, J. (2005) *Asterisk™: The Future of Telephony.* (Second ed) Oreilly.

Möller, S. (2000) *Assessment and Prediction of Speech Quality in Telecommunications.* Kluwer Academic Publishers, Dordrecht, The Netherlands,. 272 pp.

Mullner, R., Ball, C. F., Ivanov, K., Winkler, H. A., Perl, R. A. and Kremnitzer, K. A. (2007) *Exploiting AMR-WB Audio Bandwidth Extension for Quality and Capacity Increase*. Ball, C.F. (Ed). Mobile and Wireless Communications Summit, 2007. 16th IST. pp 1-7.

OPTICOM. (2004) *3SQM- ADVANCED NON-INTRUSIVE VOICE QUALITY TESTING*. Available online at: http://www.opticom.de/download/3SQM-WP-290604.pdf. (Accessed: 15/7/2008)

OPTICOM. (2007) *Voice Quality Testing | PESQ*. Available online at: http://www.opticom.de/technology/pesq.html. (Accessed: 5/1/2008)

Opticom. (2008) *Technical Specificationfor the OPERA™ Software Suite*. Available online at: http://www.opticom.de/download/TechSpec_OPERA_SW_Suite.pdf. (Accessed: 23/6/2008)

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*. Available online at: http://www.qualcomm.com/common/documents/white_papers/PESQ_Limitations_Rev_C_Jan_08.pdf. (Accessed: 1/7/2008)

R Development Core Team (2007) *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. .

Rix, A. W., Bourret, A. and Hollier, M. P. (1999) Models of Human Perception. *BT Technology Journal*, 17, (1) 24-34.

Rix, A. W. and Hollier, M. P. (2000) *The perceptual analysis measurement system for robust end-to-end speech quality assessment*. Hollier, M.P. (Ed). Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on. pp 1515-1518 vol.3.

Rohani, B., Caldera, M. and Zepernick, H. J. (2006a) Benefits of perceptual speech quality metrics in modern cellular systems. *Electronics Letters*, 42, (21) 1250-1251.

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

Rohani, B. and Zepernick, H. J. (2004) *Application of a perceptual speech quality metric for link adaptation in wireless systems*. Zepernick, H.J. (Ed). Wireless Communication Systems, 2004. 1st International Symposium on. pp 260-264.

Sacchi, C., Piazza, M. and Natale, F. G. B. D. 2007. *Cost-Effective VoIP Services for Reducing Digital Divide in Developing Countries: Case of Study and Practical Implementation.* UNIVERSITY OF TRENTO, Trento. pp.

Sourceforge.Net. (2008) *The Free, Cross-Platform Sound Editor*. Available online at: http://audacity.sourceforge.net/. (Accessed: 1/9/2008)

Spencer, M. (2008) *Introduction to the Asterisk open source PBX*. Available online at: http://www.smbconsulting.com.do/docs/asterisk-whitepaper.pdf. (Accessed: 12/6/2008)

Storm, A. 2007. *Speech quality investigation using PESQ in a simulated climax system for ATM*. MASTER'S THESIS, Luleå University of Technology.

Sun, L. (2005) *Voice over IP and Voice Quality Measurement* Available online at: http://www.tech.plym.ac.uk/spmc/people/lfsun/publications/VoIP-Acterna-2005.ppt. (Accessed: 1/9/2008)

Sun, L. and Ifeachor, E. (2004) *New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks*. Ifeachor, E. (Ed). Communications, 2004 IEEE International Conference on. pp 1478-1483 Vol.3.

Sun, L. and Ifeachor, E. C. (2002) *Perceived speech quality prediction for voice over IP-based networks*. Ifeachor, E.C. (Ed). Communications, 2002. ICC 2002. IEEE International Conference on. pp 2573-2577 vol.4.

Sun, L. and Ifeachor, E. C. (2003) *Prediction of perceived conversational speech quality and effects of playout buffer algorithms*. Ifeachor, E.C. (Ed). Communications, 2003. ICC '03. IEEE International Conference on. pp 1-6 vol.1.

Sun, L. and Ifeachor, E. C. (2006) Voice quality prediction models and their application in VoIP networks. *Multimedia, IEEE Transactions on*, 8, (4) 809-820.

Tims. (2008) *Asterisk config zaptel.conf*. Available online at: http://www.voip-info.org/wiki-Asterisk+config+zaptel.conf. (Accessed: 10/7/2008)

Uvliden, A., Bruhn, S. and Hagen, R. (1998) *Adaptive multi-rate. A speech service adapted to cellular radio network quality*. Bruhn, S. (Ed). Signals, Systems & Computers, 1998. Conference Record of the Thirty-Second Asilomar Conference on. pp 343-347 vol.1.

Varga, I., De Lacovo, R. D. and Usai, P. (2006) Standardization of the AMR wideband speech codec in 3GPP and ITU-T. *Communications Magazine, IEEE*, 44, (5) 66-73.

Yamamoto, L. and Beerends, J. G. (1997) *Impact of Network Performance Parameters on the End-to-End Perceived Speech Quality* Expert ATM Traffic Symposium. Greece.

Appendix A – Makefile for PESQ

```
# Makefile for PESQ
# by Dr. Lingfen Sun, 15/06/2001
# For Linux

CC=gcc
CFLAGS= -O2 -D__unix__

PROGRAMS = pesq

all: $(PROGRAMS)
PESQOBJS = dsp.o pesqdsp.o pesqio.o pesqmain.o pesqmod.o

pesq: $(PESQOBJS)
    $(CC) -o pesq $(PESQOBJS) -lm

dsp.o : dsp.c
    $(CC) $(CFLAGS) -c dsp.c

pesqdsp.o : pesqdsp.c dsp.h pesq.h pesqpar.h
    $(CC) $(CFLAGS) -c pesqdsp.c

pesqio.o: pesqio.c dsp.h pesq.h pesqpar.h
    $(CC) $(CFLAGS) -c pesqio.c

pesqmain.o: pesqmain.c dsp.h pesq.h pesqpar.h
    $(CC) $(CFLAGS) -c pesqmain.c

pesqmod.o: pesqmod.c dsp.h pesq.h pesqpar.h
    $(CC) $(CFLAGS) -c pesqmod.c
```

Appendix B – Asterisk Zapata.conf

```
# Zapata.conf
# For Asterisk & ISDN2e
# Created by Mohammad Goudarzi
# 1/6/2008

[channels]

nocid=Unavailable
withheldcid=Withheld
Language=en
usecallerid=yes
pridialplan=unknown
prilocaldialplan=unknown
nationalprefix=0
internationalprefix=00
switchtype = euroisdn
signalling = bri_cpe_ptmp
echocancel=no
echocancelwhenbridged=no
immediate=no
overlapdial=yes
group = 1
context=isdn-in
callgroup=1
channel => 1-2
```

Appendix C – Asterisk extensions.conf configurations

```
# Extensions.conf
# For Asterisk & h324m playback
# Created by Mohammad Goudarzi
# 1/6/2008

[isdn-in]

; Extensions playing back for GSM-encoded samples

exten => 670526,1,Answer
exten => 670526,2,wait(2)
exten => 670526,4,DigitTimeout,5
exten => 670526,5,ResponseTimeout,10

; plays a beep on the channel when ready

exten => 670526,3,Background(beep)

; waits for the user to press number 1

exten =>
  1,1,Playback(/home/Mohammad/BRITISH_ENGLISH/gsm/B_eng_f1)
exten => 1,2,wait(3)
exten => 1,3,Hangup

; Extensions playing back for .3gp files via H324m gateway

exten => 670526,1,Answer

; takes the call into the h324m context to be handled via gateway

exten => 670526,1,h324m_gw(ISDN2@isdn-in-exts)

[isdn-in-exts]

exten => test2,1,Answer
exten => test2,2,Wait(5)
exten =>
  test2,3,mp4play(/home/Mohammad/newsamples/amr/B_eng_f1.3gp)
exten => test2,3,HangUp
```

Appendix D – Score sheet and instructions For the Subjective test

Subjective test of speech quality

Dear Participant

This subjective quality test is designed to collect your opinion score about the qualities of a series of recorded speech samples.

The test consists of 30 short audio files of 7-9 seconds. As the subject, your task is to listen to each file and write down your opinion score for it in the tables provided on page 2 of this form.

The duration of the test will be 10-15 minutes. The entire test should be taken at one time in a silent, undisturbed environment and without any other person's interference. It is recommended to use headphones for completing the test. However, if you want to use your speakers, please feel free to do so.

There are 31 files provided together with this form. The file ***Reference.wav*** is provided for you to help you adjust the volume setting of your headphones/speakers. Before starting the test, please set the volume on your headphones/speakers and listen to the ***Reference.wav file.*** Make sure the level is pleasant and what is said in the file can be heard clearly and without effort. Feel free to listen to the reference file as many times until you are confident about the volume level.

Also, please state the listening device (headphones or speakers), your age range, and your gender in questions 1 to 3. The options are provided in **bold with green background**. Please delete the options as appropriate. All the information will be kept confidential.

The rest of the files (A1-A16, B1-B14) are recorded samples that you are to evaluate. You are allowed to listen to them **ONLY ONCE** and write down your opinion score according to the table-1 in the tables provided in page 2. **Please only give whole marks as your opinion**. (Do not give decimal scores).

The ratings to be given are between 1 and 5 as follows (only whole must be stated):

Table 15- Opinion scores

Score	Quality of speech
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Thank you for your participation.

Mohammad Goudarzi

Subjective Test

1- Gender : [**M/F]**

2- AGE**: [0-20] [21-30] [31-40] [41-50] [50 and above]**

3- You are using **[SPEAKERS|HEADPHONES]** to listen to the speech samples.

4- Opinion scores: Please write down your opinion score in the following tables:

File Name	Score
A1	
A2	
A3	
A4	
A5	
A6	
A7	
A8	
A9	
A10	
A11	
A12	
A13	
A14	
A15	
A16	

Filename	Score
B1	
B2	
B3	
B4	
B5	
B6	
B7	
B8	
B9	
B10	
B11	
B12	
B13	
B14	

After you have completed the test, please Email your completed form to mohammad.goudarzi@postgrad.plymouth.ac.uk or Mgudarzi@Gmail.com

Thank you very much for you time.

Mohammad Goudarzi

August 2008

Appendix E – Results of objective measurements

Table 16- PESQ and 3SQM MOS - Set 1 (Three Operator)

REFERENCE	Gender	Time	Pure Codec	PESQMOS	MOSLQO	3SQM
B_eng_f1.wav	F	10:00	3.422	3.134	3.022	3.245181
B_eng_f2.wav	F	10:00	3.281	2.601	2.262	2.60452
b_eng_f3.wav	F	10:00	3.294	2.508	2.145	2.398949
b_eng_f4.wav	F	10:00	3.408	2.961	2.765	2.926527
b_eng_f5.wav	F	10:00	3.289	2.760	2.475	2.963213
b_eng_f6.wav	F	10:00	3.293	2.574	2.227	2.473013
b_eng_f7.wav	F	10:00	3.453	3.139	3.030	2.493427
b_eng_f8.wav	F	10:00	3.288	2.833	2.579	3.504643
B_eng_m1.wav	M	10:00	3.737	3.149	3.044	2.784209
B_eng_m2.wav	M	10:00	3.630	3.072	2.929	2.316673
B_eng_m3.wav	M	10:00	3.822	3.292	3.256	2.935487
B_eng_m4.wav	M	10:00	3.752	3.166	3.070	2.834415
B_eng_m5.wav	M	10:00	3.843	3.105	2.979	2.053734
B_eng_m6.wav	M	10:00	3.664	3.180	3.091	1.814774
B_eng_m7.wav	M	10:00	3.882	3.317	3.293	2.939605
B_eng_m8.wav	M	10:00	3.827	2.955	2.756	1.485266

Table 17-PESQ and 3SQM MOS - Set 2 (Three Operator)

REFERENCE	Gender	Time	Pure Codec	PESQMOS	MOSLQO	3SQM
B_eng_f1.wav	F	13:00	3.422	3.159	3.059	3.560824
B_eng_f2.wav	F	13:00	3.281	2.640	2.313	2.769376
b_eng_f3.wav	F	13:00	3.294	2.592	2.250	2.377095
b_eng_f4.wav	F	13:00	3.408	2.929	2.718	2.968553
b_eng_f5.wav	F	13:00	3.289	2.781	2.505	3.025886
b_eng_f6.wav	F	13:00	3.293	2.525	2.166	2.47707
b_eng_f7.wav	F	13:00	3.453	3.091	2.957	2.458547
b_eng_f8.wav	F	13:00	3.288	2.865	2.624	3.344531
B_eng_m1.wav	M	13:00	3.737	3.209	3.134	2.468915
B_eng_m2.wav	M	13:00	3.630	3.072	2.929	1.602145
B_eng_m3.wav	M	13:00	3.822	3.404	3.420	2.978743
B_eng_m4.wav	M	13:00	3.752	3.143	3.036	2.721782
B_eng_m5.wav	M	13:00	3.843	3.066	2.921	1.990037
B_eng_m6.wav	M	13:00	3.664	3.166	3.070	2.264232
B_eng_m7.wav	M	13:00	3.882	3.365	3.364	3.006207
B_eng_m8.wav	M	13:00	3.827	2.990	2.808	1.428077

Table 18-PESQ and 3SQM MOS - Set 3 (Three Operator)

REFERENCE	Gender	Time	Pure Codec	PESQMOS	MOSLQO	3SQM
B_eng_f1.wav	F	16:00	3.422	2.806	2.540	3.730866
B_eng_f2.wav	F	16:00	3.281	2.579	2.234	3.349128
b_eng_f3.wav	F	16:00	3.294	2.689	2.378	2.931982
b_eng_f4.wav	F	16:00	3.408	2.911	2.691	3.270076
b_eng_f5.wav	F	16:00	3.289	2.734	2.439	3.313303
b_eng_f6.wav	F	16:00	3.293	2.711	2.408	3.111601
b_eng_f7.wav	F	16:00	3.453	2.916	2.699	4.089891
b_eng_f8.wav	F	16:00	3.288	2.825	2.567	3.629477
B_eng_m1.wav	M	16:00	3.737	3.030	2.867	2.991126
B_eng_m2.wav	M	16:00	3.630	3.197	3.116	2.979338
B_eng_m3.wav	M	16:00	3.822	3.197	3.117	3.27206
B_eng_m4.wav	M	16:00	3.752	3.245	3.188	3.301248
B_eng_m5.wav	M	16:00	3.843	3.173	3.081	2.581107
B_eng_m6.wav	M	16:00	3.664	3.094	2.962	2.634784
B_eng_m7.wav	M	16:00	3.882	3.228	3.162	3.269579
B_eng_m8.wav	M	16:00	3.827	3.078	2.939	2.978634

Table 19- PESQ and 3SQM MOS - Set 4 (Vodafone Operator)

REFERENCE	Gender	Time	Pure Codec	PESQMOS	MOSLQO	3SQM
B_eng_f1.wav	F	10:00	3.422	3.214	3.142	3.662346
B_eng_f2.wav	F	10:00	3.281	2.836	2.583	2.713971
b_eng_f3.wav	F	10:00	3.294	2.848	2.600	2.514777
b_eng_f4.wav	F	10:00	3.408	3.158	3.059	3.165251
b_eng_f5.wav	F	10:00	3.289	2.914	2.696	2.926559
b_eng_f6.wav	F	10:00	3.293	2.708	2.404	2.534144
b_eng_f7.wav	F	10:00	3.453	2.904	2.682	2.180395
b_eng_f8.wav	F	10:00	3.288	3.025	2.860	3.264511
B_eng_m1.wav	M	10:00	3.737	3.247	3.190	2.613983
B_eng_m2.wav	M	10:00	3.630	3.199	3.119	2.361916
B_eng_m3.wav	M	10:00	3.822	3.130	3.016	2.950459
B_eng_m4.wav	M	10:00	3.752	3.277	3.234	2.517492
B_eng_m5.wav	M	10:00	3.843	3.153	3.051	1.923257
B_eng_m6.wav	M	10:00	3.664	3.152	3.049	2.225603
B_eng_m7.wav	M	10:00	3.882	3.485	3.533	2.868727
B_eng_m8.wav	M	10:00	3.827	3.078	2.939	1.280307

Table 20- PESQ and 3SQM MOS - Set 5 (Vodafone Operator)

REFERENCE	Gender	Time	Pure Codec	PESQMOS	MOSLQO	3SQM
B_eng_f1.wav	F	13:00	3.422	3.112	2.989	3.664518
B_eng_f2.wav	F	13:00	3.281	2.807	2.542	2.645496
b_eng_f3.wav	F	13:00	3.294	2.906	2.684	2.496519
b_eng_f4.wav	F	13:00	3.408	3.198	3.118	3.060165
b_eng_f5.wav	F	13:00	3.289	2.991	2.809	3.07773
b_eng_f6.wav	F	13:00	3.293	2.773	2.493	2.69378
b_eng_f7.wav	F	13:00	3.453	3.174	3.082	2.258284
b_eng_f8.wav	F	13:00	3.288	2.900	2.676	3.245063
B_eng_m1.wav	M	13:00	3.737	3.312	3.286	2.517618
B_eng_m2.wav	M	13:00	3.630	3.253	3.200	2.345622
B_eng_m3.wav	M	13:00	3.822	3.389	3.398	3.010846
B_eng_m4.wav	M	13:00	3.752	3.141	3.032	2.60685
B_eng_m5.wav	M	13:00	3.843	3.228	3.162	1.847584
B_eng_m6.wav	M	13:00	3.664	3.336	3.321	2.313401
B_eng_m7.wav	M	13:00	3.882	3.539	3.608	2.905891
B_eng_m8.wav	M	13:00	3.827	3.302	3.271	1.432557

Table 21- PESQ and 3SQM MOS - Set 6 (Vodafone Operator)

REFERENCE	Gender	Time	Pure Codec	PESQMOS	MOSLQO	3SQM
B_eng_f1.wav	F	16:00	3.422	2.498	2.133	3.444097
B_eng_f2.wav	F	16:00	3.281	2.903	2.679	2.721584
b_eng_f3.wav	F	16:00	3.294	2.933	2.724	2.45726
b_eng_f4.wav	F	16:00	3.408	3.236	3.174	3.135398
b_eng_f5.wav	F	16:00	3.289	3.049	2.895	2.947056
b_eng_f6.wav	F	16:00	3.293	2.779	2.502	2.576003
b_eng_f7.wav	F	16:00	3.453	3.261	3.211	3.985929
b_eng_f8.wav	F	16:00	3.288	3.023	2.857	3.252273
B_eng_m1.wav	M	16:00	3.737	3.331	3.314	2.806242
B_eng_m2.wav	M	16:00	3.630	3.243	3.184	2.366257
B_eng_m3.wav	M	16:00	3.822	3.468	3.510	3.125562
B_eng_m4.wav	M	16:00	3.752	3.292	3.257	2.925913
B_eng_m5.wav	M	16:00	3.843	3.271	3.226	1.863817
B_eng_m6.wav	M	16:00	3.664	3.367	3.366	2.320076
B_eng_m7.wav	M	16:00	3.882	2.744	2.453	2.756827
B_eng_m8.wav	M	16:00	3.827	2.037	1.662	1

Table 22- AMR Samples set-1

REFERENCE	Gender	Time	Pure Codec	PESQ MOS	MOSLQO	3SQM
B_eng_f1.wav	F	10:00	4.109	3.588	3.673	4.352682
B_eng_f2.wav	F	10:00	3.810	3.102	2.974	3.005803
b_eng_f3.wav	F	10:00	3.805	3.129	3.014	2.99125
b_eng_f4.wav	F	10:00	4.035	3.527	3.591	3.27514
b_eng_f5.wav	F	10:00	3.844	3.324	3.303	3.529126
b_eng_f6.wav	F	10:00	3.812	2.946	2.743	3.003455
b_eng_f7.wav	F	10:00	4.005	3.109	2.985	4.563834
b_eng_f8.wav	F	10:00	3.783	3.061	2.914	3.896432
B_eng_m1.wav	M	10:00	3.909	3.150	3.046	3.235255
B_eng_m2.wav	M	10:00	3.995	3.494	3.546	3.483409
B_eng_m3.wav	M	10:00	4.000	3.407	3.423	3.569574
B_eng_m4.wav	M	10:00	4.024	3.458	3.496	3.876206
B_eng_m5.wav	M	10:00	3.965	3.206	3.129	2.818061
B_eng_m6.wav	M	10:00	4.023	3.304	3.275	2.841736
B_eng_m7.wav	M	10:00	4.071	3.326	3.306	3.466073
B_eng_m8.wav	M	10:00	4.022	3.148	3.043	3.294389

Table 23-AMR Samples set-2

REFERENCE	Gender	Time	Pure Codec	PESQ MOS	MOSLQO	3SQM
B_eng_f1.wav	F	10:00	4.109	3.429	3.455	4.041712
B_eng_f2.wav	F	10:00	3.81	2.901	2.677	3.260056
b_eng_f3.wav	F	10:00	3.805	3.191	3.107	2.96642
b_eng_f4.wav	F	10:00	4.035	3.453	3.489	3.714962
b_eng_f5.wav	F	10:00	3.844	3.157	3.056	3.623552
b_eng_f6.wav	F	10:00	3.812	2.922	2.707	3.087771
b_eng_f7.wav	F	10:00	4.005	3.392	3.403	4.662499
b_eng_f8.wav	F	10:00	3.783	3.203	3.125	4.020241
B_eng_m1.wav	M	10:00	3.909	3.395	3.406	2.989347
B_eng_m2.wav	M	10:00	3.995	3.446	3.479	3.12675
B_eng_m3.wav	M	10:00	4	3.754	3.883	3.547726
B_eng_m4.wav	M	10:00	4.024	3.341	3.328	3.213354
B_eng_m5.wav	M	10:00	3.965	3.342	3.33	2.489841
B_eng_m6.wav	M	10:00	4.023	3.516	3.576	2.798691
B_eng_m7.wav	M	10:00	4.071	3.708	3.828	3.706044
B_eng_m8.wav	M	10:00	4.022	2.958	2.76	2.201109

Table 24- AMR Samples set- 3

REFERENCE	Gender	Time	Pure Codec	PESQ MOS	MOSLQO	3SQM
B_eng_f1.wav	F	13:00	4.109	3.514	3.573	4.331092
B_eng_f2.wav	F	13:00	3.810	3.025	2.860	3.376601
b_eng_f3.wav	F	13:00	3.805	3.048	2.893	3.090358
b_eng_f4.wav	F	13:00	4.035	3.445	3.478	3.505602
b_eng_f5.wav	F	13:00	3.844	3.157	3.057	3.588932
b_eng_f6.wav	F	13:00	3.812	2.899	2.673	3.115948
b_eng_f7.wav	F	13:00	4.005	3.426	3.451	4.688734
b_eng_f8.wav	F	13:00	3.783	3.315	3.291	3.970253
B_eng_m1.wav	M	13:00	3.909	3.439	3.469	3.157778
B_eng_m2.wav	M	13:00	3.995	3.615	3.709	3.221192
B_eng_m3.wav	M	13:00	4.000	3.725	3.848	3.542777
B_eng_m4.wav	M	13:00	4.024	3.650	3.753	3.67005
B_eng_m5.wav	M	13:00	3.965	3.343	3.331	2.498603
B_eng_m6.wav	M	13:00	4.023	3.514	3.574	2.770815
B_eng_m7.wav	M	13:00	4.071	3.489	3.540	3.341287
B_eng_m8.wav	M	13:00	4.022	3.335	3.320	2.107414

Table 25-AMR Samples set-4

REFERENCE	Gender	Time	Pure Codec	PESQ MOS	MOSLQO	3SQM
B_eng_f1.wav	F	13:00	4.109	3.163	3.065	4.265944
B_eng_f2.wav	F	13:00	3.810	2.828	2.571	3.497203
b_eng_f3.wav	F	13:00	3.805	2.760	2.476	3.233973
b_eng_f4.wav	F	13:00	4.035	3.072	2.929	3.502901
b_eng_f5.wav	F	13:00	3.844	2.859	2.617	3.584295
b_eng_f6.wav	F	13:00	3.812	2.764	2.481	3.209542
b_eng_f7.wav	F	13:00	4.005	3.108	2.983	4.498204
b_eng_f8.wav	F	13:00	3.783	3.092	2.959	4.039794
B_eng_m1.wav	M	13:00	3.909	3.150	3.045	3.368166
B_eng_m2.wav	M	13:00	3.995	3.494	3.546	3.415496
B_eng_m3.wav	M	13:00	4.000	3.365	3.362	3.6302
B_eng_m4.wav	M	13:00	4.024	3.474	3.518	3.716046
B_eng_m5.wav	M	13:00	3.965	3.204	3.127	2.66476
B_eng_m6.wav	M	13:00	4.023	3.305	3.276	2.938357
B_eng_m7.wav	M	13:00	4.071	2.880	2.646	3.359643
B_eng_m8.wav	M	13:00	4.022	3.149	3.044	3.30191

Table 26-AMR Samples set-5

REFERENCE	Gender	Time	Pure Codec	PESQ MOS	MOSLQO	3SQM
B_eng_f1.wav	F	16:00	4.109	3.547	3.618	4.05279
B_eng_f2.wav	F	16:00	3.810	3.025	2.860	3.253941
b_eng_f3.wav	F	16:00	3.805	2.804	2.537	2.999172
b_eng_f4.wav	F	16:00	4.035	3.458	3.496	3.598211
b_eng_f5.wav	F	16:00	3.844	3.156	3.055	3.651986
b_eng_f6.wav	F	16:00	3.812	2.916	2.698	2.967081
b_eng_f7.wav	F	16:00	4.005	3.401	3.415	4.683742
b_eng_f8.wav	F	16:00	3.783	3.315	3.291	3.980691
B_eng_m1.wav	M	16:00	3.909	3.440	3.471	3.189204
B_eng_m2.wav	M	16:00	3.995	3.676	3.787	3.009858
B_eng_m3.wav	M	16:00	4.000	3.769	3.901	3.628122
B_eng_m4.wav	M	16:00	4.024	3.653	3.758	3.511061
B_eng_m5.wav	M	16:00	3.965	3.344	3.332	2.458433
B_eng_m6.wav	M	16:00	4.023	3.339	3.325	2.688978
B_eng_m7.wav	M	16:00	4.071	3.656	3.762	3.327937
B_eng_m8.wav	M	16:00	4.022	3.313	3.288	2.291213

Table 27- AMR Samples set-6

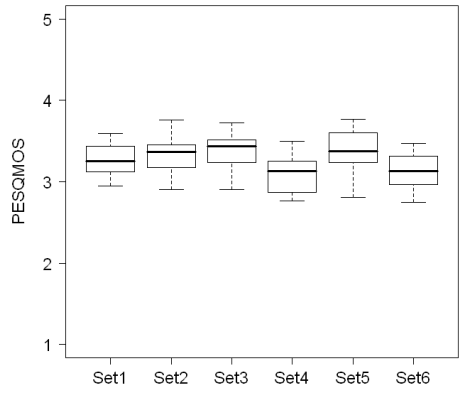
REFERENCE	Gender	Time	Pure Codec	PESQ MOS	MOSLQO	3SQM
B_eng_f1.wav	F	16:00	4.109	3.163	3.065	4.199852
B_eng_f2.wav	F	16:00	3.810	2.748	2.459	3.213067
b_eng_f3.wav	F	16:00	3.805	2.787	2.514	3.232659
b_eng_f4.wav	F	16:00	4.035	3.061	2.913	3.598777
b_eng_f5.wav	F	16:00	3.844	2.863	2.621	3.842438
b_eng_f6.wav	F	16:00	3.812	2.762	2.478	3.260485
b_eng_f7.wav	F	16:00	4.005	3.107	2.982	4.693055
b_eng_f8.wav	F	16:00	3.783	3.091	2.958	4.050754
B_eng_m1.wav	M	16:00	3.909	3.149	3.045	2.975586
B_eng_m2.wav	M	16:00	3.995	3.472	3.516	3.356924
B_eng_m3.wav	M	16:00	4.000	3.405	3.421	3.533625
B_eng_m4.wav	M	16:00	4.024	3.474	3.518	3.550071
B_eng_m5.wav	M	16:00	3.965	3.202	3.124	2.756811
B_eng_m6.wav	M	16:00	4.023	3.306	3.278	3.032593
B_eng_m7.wav	M	16:00	4.071	3.313	3.288	3.543667
B_eng_m8.wav	M	16:00	4.022	3.116	2.995	3.365416

Appendix F – Subjective measurement results

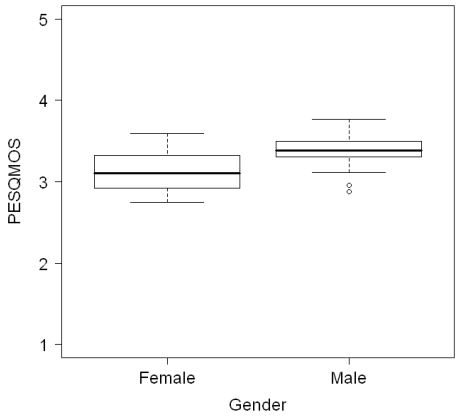
	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14
S1	2	1	2	1	2	2	3	1	3	4	2	3	2	2	2	1	5	4	4	3	3	3	4	4	2	4	4	4	3	5
S2	3	2	3	2	3	2	2	2	3	3	2	2	3	3	3	3	3	3	2	4	3	4	3	3	3	3	2	3	1	3
S3	4	3	4	2	3	3	4	3	3	4	3	3	4	4	4	4	4	5	4	5	4	4	4	4	5	4	2	4	3	5
S4	1	2	2	1	3	2	3	2	4	3	2	3	3	4	4	4	5	4	3	3	4	5	4	5	4	4	4	5	2	5
S5	1	2	2	1	2	2	3	3	3	2	2	3	4	4	4	4	4	4	4	4	4	4	4	5	5	5	4	5	4	5
S6	2	3	2	1	3	4	4	3	5	3	3	2	4	4	3	2	5	4	2	4	5	5	5	3	4	4	4	5	2	5
S7	2	2	2	2	3	3	3	3	4	3	3	3	4	4	4	3	4	4	4	4	4	5	4	4	4	4	4	5	3	4
S8	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	4	5	5	4	5	5	5	4	5
S9	1	2	1	2	2	2	2	3	1	2	2	4	2	3	3	3	3	3	5	4	3	4	4	5	4	4	4	3	5	5
S10	1	4	2	1	3	5	3	4	3	2	1	3	4	5	4	5	3	2	1	2	3	4	4	2	4	3	1	4	1	3
S11	4	5	4	2	5	4	3	4	2	2	1	2	4	4	5	5	3	3	2	4	3	4	3	3	4	2	4	3	1	3
S12	4	5	4	2	5	4	3	4	2	2	1	2	4	4	5	5	3	3	2	4	3	4	3	3	4	2	4	3	1	3
S13	3	1	3	1	2	2	4	1	4	3	2	3	4	5	4	3	5	5	4	4	4	4	4	3	5	4	3	4	3	4
S14	2	3	3	1	4	4	5	3	3	3	2	3	4	5	5	3	4	4	4	4	4	5	5	5	4	5	4	4	3	5
S15	3	4	4	2	5	5	5	4	4	4	3	3	5	4	5	3	5	5	5	5	5	5	5	5	5	5	5	5	4	5
S16	2	1	2	1	1	1	2	1	4	3	3	3	2	2	2	1	4	3	2	4	3	2	2	1	2	3	3	2	1	3
S17	3	2	3	1	4	4	4	4	1	2	1	2	5	4	5	4	2	3	4	3	4	5	4	4	5	4	3	4	1	2
S18	3	2	2	1	3	3	2	3	4	4	3	3	4	4	4	3	4	4	3	4	5	4	5	4	5	5	3	5	3	4
S19	3	2	3	1	4	3	4	2	2	2	1	2	3	3	3	3	3	4	2	2	4	3	2	3	4	3	2	3	1	3
S20	2	2	3	1	1	2	3	3	4	2	3	3	4	4	4	3	4	3	4	3	4	4	5	4	3	5	4	5	4	3
S21	3	3	3	2	3	4	4	4	4	3	3	3	3	4	4	4	5	5	5	5	5	5	5	5	4	5	5	5	3	5
S22	1	2	4	1	2	4	4	2	4	3	3	2	3	4	3	2	4	3	4	2	3	3	3	3	4	3	4	4	2	5
S23	3	3	3	2	4	4	4	4	4	2	2	2	2	3	2	3	5	5	5	5	5	5	5	5	5	5	5	5	4	5
S24	3	2	3	1	3	3	4	3	4	4	3	3	3	3	4	3	5	5	4	3	5	3	3	3	3	4	3	3	2	3
S25	2	2	3	1	4	2	4	3	5	3	3	3	3	4	4	3	5	4	4	4	4	5	4	4	4	4	4	4	3	4
S26	1	1	3	1	2	2	3	3	4	3	3	4	3	4	4	3	4	5	3	4	4	3	3	4	3	3	3	4	2	4
S27	3	2	1	1	4	2	2	1	3	3	2	1	2	3	3	4	2	4	3	3	3	3	3	3	5	4	2	3	1	3
S28	3	1	1	1	2	3	3	2	3	2	1	3	4	4	4	2	4	3	2	4	4	4	3	4	4	4	3	3	1	3
S29	3	2	1	3	3	1	4	2	4	1	1	3	5	5	4	3	4	3	2	3	2	4	3	3	5	4	4	3	1	4
S30	2	2	3	1	3	2	3	1	3	3	2	3	4	4	4	4	3	5	4	4	5	5	5	4	4	4	5	5	3	5
S31	3	3	2	2	4	3	5	3	3	3	2	3	3	4	4	3	4	4	3	5	5	4	5	5	5	4	3	4	3	4
S32	4	2	3	2	4	3	4	3	3	3	3	3	4	4	4	3	4	4	4	3	4	4	4	3	4	4	3	4	2	4
S33	2	2	2	1	3	3	4	3	3	2	3	3	4	4	4	3	3	2	2	4	3	3	3	3	3	3	3	4	3	4
MOS	2.485	2.364	2.606	1.455	3.121	2.939	3.455	2.758	3.333	2.788	2.273	2.788	3.515	3.818	3.788	3.212	3.939	3.848	3.364	3.758	3.909	4.030	3.879	3.758	4.030	3.909	3.515	4.000	2.424	4.030
STDEV	0.939	1.025	0.899	0.617	1.053	1.059	0.869	1.001	0.957	0.781	0.839	0.650	0.870	0.727	0.820	0.960	0.899	0.906	1.141	0.867	0.843	0.810	0.927	1.001	0.847	0.843	1.004	0.866	1.173	0.918

Appendix G – Statistical Results for PESQMOS

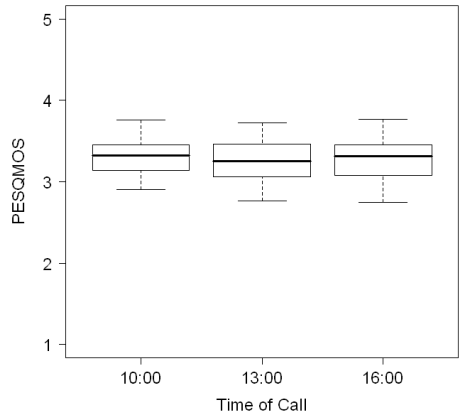
The graphs presented in Chapter 5 show the box-plots for the objective measurement results using 3SQM and PESQ-LQO, the same plots for raw PESQ scores are shown below:



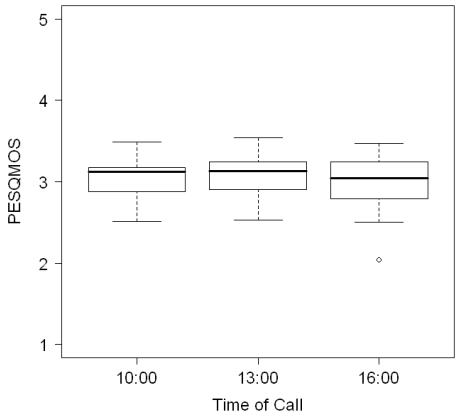
Effect of the volume setting on PESQ raw scores in AMR experiments



Talker's gender effect on the PESQ scores in AMR experiments



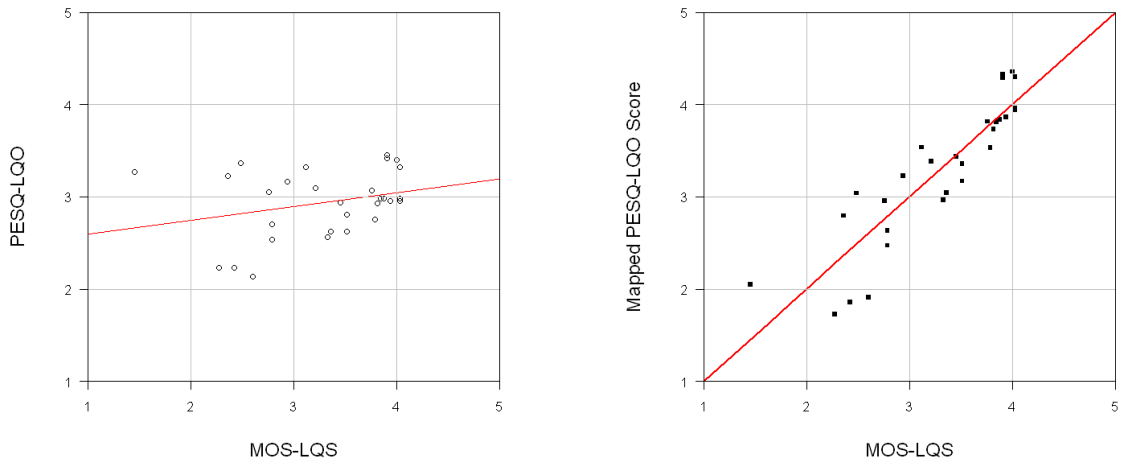
PESQ scores for AMR samples vs. Time of call



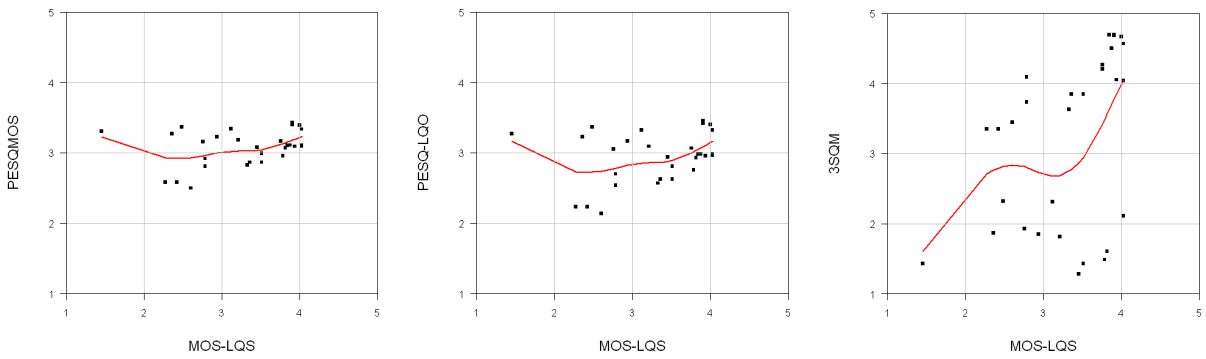
PESQ scores for GSM samples vs. Time of call

PESQ scores for GSM samples divided by Operator

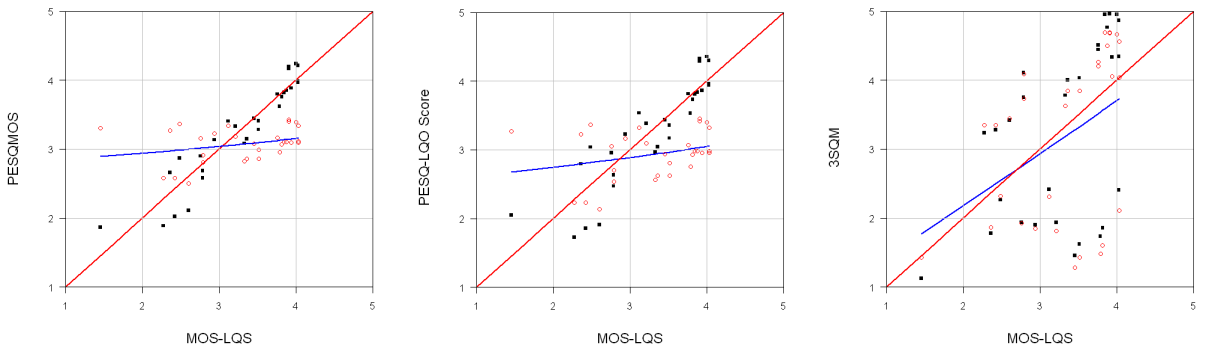
Appendix H – Graphs for mapping function and polynomial Calculations



Mapping between PESQ-LQO score and subjective MOS



Objective vs. Subjective measurements - 3rd order polynomial regression function



Objective vs. Subjective measurements before and after mapping