Assessment of Speech Quality for VoIP Applications using PESQ and E -Model

H.A.Khan and L.Sun

Signal Processing and Multimedia Communications,
University of Plymouth, Plymouth, United Kingdom
e-mail: L.Sun@plymouth.ac.uk

Abstract

The aim of the paper is to investigate and assess speech quality for VoIP applications using the latest ITU-T standards (i.e. ITU-T P.862 PESQ and G.107 E-Model) and to compare the results with subjective tests results. . The speech quality metrics used in this experiment are MOS-LQS, MOS-LQO and MOS-CQE. The impact of packet loss rate (including packet burstness) on speech quality was investigated based on a VoIP testbed (including NistNET network emulator and XLite/ Skype VoIP terminals). The preliminary results show that PESQ achieves a higher correlation rate (81%) with subjective test results than that of E-model (74%). Some issues related with how to test speech quality properly in the experimental testbed are also discussed.

Keywords

QOS, MOS-LQO, MOS-LQS, MOS-CQE, VoIP, Emulator and NistNET

1 Introduction

Voice over IP is getting popular day by day, due to its obvious benefits. However, with the popularity and growth in the field of VoIP, standardization also became an important part of the industry. VoIP technology is used to transfer voice data over the traditional data networks using a set of protocols specialized for voice communication. Voice over Internet Protocol is growing at a very fast pace. The advantages of this technology are low cost and transfer of other data then just voice (images, video etc) over long distances using computer networks. However, the quality of this service is often compromised with cost and bandwidth. Also, the quality of this technology is affected by network issues like jitter, delay, distortion, packet drop etc. However, to regulate and maintain quality of the service, there are certain Quality Measuring Methods that are used today in VoIP field.

There are many speech quality measurement methods, for example, the ITU-T PESQ and ITU-T E-model, which have been widely used in industry for speech quality assessment for VoIP products and systems. However, it is unclear how well the PESQ and E-model is when compared with subjective test results.

The main aims of this paper are (1) To compare and correlate Objective speech quality (PESQ) and Estimated Speech Quality (E-model) with the subjective method of speech quality (2) To investigate and analyze the effect of packet loss ratio over the voice quality in an IP network (3) To get a better knowledge of Speech quality measuring methods. Also analyzing and evaluating the reasons behind the unusual behavior will be a part of the project.

For this purpose, we have set up a VoIP speech quality testbed and investigated speech quality using PESQ and E-model and compared the objective test results with subjective test results. The preliminary results show that PESQ correlates better then E-model. The Pearson Correlation that was calculated between PESQ and Subjective tests came out to be 81% while the correlation between E-model results and Subjective results came out to be 74%. During the experiment, we also find that the NistNET network emulator software should be given some time in-order to achieve drop rate that has been specified. We find out that time equivalent to 100 ICMP echo packets take (200 seconds) should be given to the network and NistNET for drop rates up to 20% be achieved.

The remainder of the paper is structured as follows. In Section 2, an overview of VoIP and speech quality measurement methods is introduced. In Section 3, the testbed used in VoIP quality assessment and methodology of testing is presented. In Section 4, test results and analysis are shown for drop ratios of 4%, 8%, 12%, 16% and 20% with loss correlation of 0.3, 0.5 and 0.8. Section 5 contains the conclusions that are obtained after analyzing the results. All the references are mentioned in the end of this paper.

2 Voice over Internet Protocol and Quality Measurement Methods

To define it ,VoIP is set of technologies that is used to transfer Voice data over computer networks rather than traditional PSTN systems(Dudman 2006). If we look into the architecture of Voice over IP, it essentially consists of end points that are capable of converting analogue voice into digital data, a network, a receiving end that would convert digital voice data into audible sound, and a server to register and control all the activities. There are many factors affecting the speech quality in VoIP network (Opticom 2007). There are certain methods that are used to measure the speech quality over an IP network. These methods can be divided into intrusive and non intrusive methods.

ITU-T defines MOS as the values on a predefined scale that subjects assign to their opinion of the performance of the telephone transmission system used either for conversation or for listening to spoken material (ITU.T.Recommendation.P.800.1 2006). MOS is arithmetic mean of all the scores collected in a test. We will take into account the main three types of MOS tests, that are Subjective based testing, Objective based testing PESQ ITU-T P.800 and Estimation based testing, E-model ITU-T G.107.

Subjective scoring is done through the scoring of many subjects. These tests are carried out in a very controlled environment, so that external disturbance elements are not involved. The subjects are presented with a number of degraded samples and are asked to number mark them from 1 to 5, depending on the perceived speech quality. In objective types of test, the score is calculated from an objective models that predicts the scores as would have done by subjective testing(PESQ 2001). PESQ is an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs(ITU-T 2007a). PESQ software is modeled and designed in such a way that it can it can analyze certain degradation factors inside an audio file as compared to its reference file.

Estimation based methods of speech quality measurements are usually non-intrusive that are estimated using the parameters of the networks. The method we have used in our project is E-Model , the principal behind the working of E-Model is that the Psychological factors on the psychological scale are additive.(ITU-T 2005). Prediction models like these are useful in network planning and to replace the need for the sample comparison(Horrocks). , E-Model estimates the value from 1 to 100. The equation (Ding and Goubran 2003)by which E-Model is calculated is given below

$$R = R_o - Is - Ie - Id + A$$

R_o is the over all signal to noise ratio including the over all circuit and signal noise effects. A is the advantage factor. Id stands for any delay that can be caused by the network while Ie counts for the packet loss. In this section, we will restrict the theory explanation only for IeI.e. as it is being used in our project and a thorough explanation is required for any reader to understand the concept behind the packet loss calculations. Usually, packet loss is not purely random; in fact it occurs with a conditional probability. Hence Ie, when packet loss is not random, we write it as $Ie - eff$ and defined by

$$Ie - eff = Ie + (95 - Ie).\left(\frac{Ppl}{(Ppl/BurstR) + Bpl}\right)$$

Where Ppl = percentage of packet loss
$BurstR$ = Correlation of packet loss
Bpl = Packet robustness value

The Ie and Bpl values for different codecs are taken from ITU publication(ITU-T 2007b).The value of BurstR can be calculated using

$$BurstR = \frac{1 - \frac{Ppl}{100}}{q}$$

the equation can be rewritten(Sun and Ifeachor 2004) as

$$R = 93.2 - I_{e-eff}$$

Hence, we will calculate the R values from above equations and we can use (ITU-T 2005)the equations defined in ITU-T Recommendation G.107 to convert R values into MOS values.

3 Testbed Setup and Methodology:

The architecture deployed for our testing is a WAN (Wide area network), that consists of many routes that are lossy and possess delays. However the main concentration in this analysis would be on packet loss in the network. The end points which were discussed are two computers that contain SIP Phones. Below is the general picture of the architecture that we are looking to experiment with**.**

In this testbed, we are using a system that is a network emulator, and will be used to emulate network scenarios with different packet loss. The network emulator that we will use is called NistNET(NistNET 2001). NistNET is used here to introduce delay and packet loss for experimenting different network scenarios i.e. heavy packet loss with high correlation like 0.8 or 80% etc. The NistNET runs on a Linux base Operating System, and here in our testbed, we have used SuSE 9. The main values for packet loss used in our project were 0%, 4%, 8%, 16% and 20% with the correlation probability of 0.3, 0.5 and 0.8.

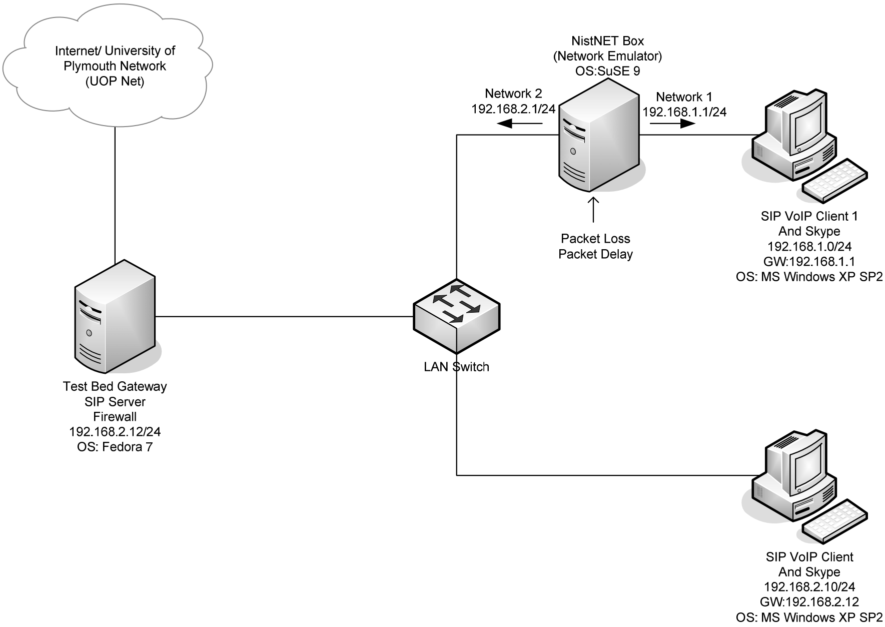


Figure 1: Test bed architecture

The whole network has been divided into two subnets. The NistNET system has two network cards connected to it and each is connected to a separate network. So any data that is moving from network 1 to network 2, as shown in Fig 1, will suffer all those degradations (Loss, jitter, delay etc) as defined in the NistNET box. However, the testbed is also connected to external network through a gateway as shown in the figure above. The reason behind giving an external network access to the network is to do the testing on the Propriety based P2P VoIP clients e.g. Skype and Google talk, which need their own servers to communicate. However, for the SIP VoIP terminal used in this project is X-Lite which gives us the flexibility of choosing the codecs for VoIP communication. One important factor is to transfer voice sample from the sender to receiving end. For this purpose, a cable who's both ends are 3.5" audio jacks, is used. One end is connected to microphone/audio in port while other is connected to speakers/audio out port of the same system. Thus any audio file played in the system will be directly transferred to the microphone port and thus will be used for input voice sample for VoIP end terminal. Another method is to use virtual cable, that is, software based sound driver, works same like physical cable connected between audio out and in ports. Virtual audio cable can process sound output into the audio input /Line in of the soundcard. We used physical cable for our experimentation.

NistNET is a software based network emulation package that runs on Linux(Khasnabish 2003). NistNET allow a single Linux base system as a router and performs Fire-wall like functions to emulate a wide variety of network base functions(Panwar 2004). The Linux Platform on which NistNET runs in our case is SuSE 9 Enterprise edition. The end clients are the two end systems that are operating on Windows XP. These end systems act as the VoIP end points. The VoIP end terminal software used here were Skype and X-Lite SIP. The gateway is also a Linux base platform that is fedora 7. It also contains two network interfaces; one is connected to the University of Plymouth Network while other is connected to our project network. The IP address for the one interface that is connected to the University network is on DHCP while other is on a static IP. The firewall settings' are the default in the system.

The main aim of the project is to assess and correlate PESQ and E-Model with Subjective MOS results, thus the testing has to be carried out with ITU-T recommended voice sample and observe the results. The below mentioned are the three different methods used to assess the quality of speech or device network.

The samples reach the receiving end after passing through an emulated lossy network. We compare the reference signal with the degraded signal in Opticom Opera software to get the Objective PESQ MOS-LQO results, while in subjective testing; Human beings are used as subjects for grading the quality of the sample.

We took 18 samples, which are in different combination of loss percentage (0%, 4%, 8%, 12%, 16% and 20%) and Loss correlation (0.3, 0.5 and 0.8).The reference sound file (b_f1_eng.wav) was placed in the beginning while the remaining 18 samples were placed randomly in a playlist. Subjects are asked to listen to the files in given order and were asked to mark them in the range of 1 to 5.Later; all scores for a file were averaged to obtain MOS-LQS score.

The methodology we use here is that one system transmits the original sample (reference audio file), while it passes through the NistNET system that introduces the present network parameters (delay loss etc), and then it reaches the other system where it is recorded and analyzed. So, the transmitting system plays the audio file in any of the ordinary audio player. The Audio out is inserted into Audio In (Microphone) port of the sound card through an audio cable that has both ends 3.5" audio jacks. On the other end of the network, the voice sample is received by the VoIP client, which is recorded either by the built in recorder or by any other software. In this experiment, Audacity software was used for recording with Skype while Built-in recording function was used for X-Lite recording. This file is then used in Opera software with reference to its original audio file to compute PESQ (MOS-LQO) score.

The third part of the testing was to calculate the R value from E model. As we have discussed before that the E-model is basically an estimation of the voice quality that the voice sample will have depending on the network conditions. As the R values are dependent on the network values, we will just consider packet loss and no delay is taken into the consideration. The shortened general E-model equation (Sun and Ifeachor 2004)is given by

$$R = R_o - I_{e-eff}$$

Where I_{e-eff} is calculated using equations as described in literature review, where R_o value is taken as 93.2, in this way, we get the values of R for the corresponding packet loss ratio. Also, we know the mapping function of MOS scores to R values. Hence we can use that as well for mapping the R values to MOS values.

4 Test Results and Analysis

Now in this part of the report, we will look into the results of the tests we conducted and will discuss about them. The three types of test we did were with a audio sample file over three different Quality measuring standards in the metrics of MOS-LQS, MOS-LQO and MOS-CQE.

4.1 Objective Tests based on PESQ (MOS-LQO)

We received the following results when objective testing was done

	Table 1				**Table 2**				**Table 3**			
Loss	**Test 1**	**Test 2**	**Test 3**	**Average**	**Test 1**	**Test 2**	**Test 3**	**Average**	**Test 1**	**Test 2**	**Test 3**	**Average**
0%	3.32	3.39	3.44	3.38	3.23	3.35	3.26	3.28	3.31	3.03	3.33	3.22
4%	3.01	3.01	2.76	2.93	3.02	3.1	3.1	3.07	2.9	3.15	2.89	2.98
8%	2.48	2.46	2.69	2.54	2.72	2.65	2.55	2.64	2.67	3.02	2.54	2.74
12%	2.35	2.41	2.28	2.35	2.6	2.5	2.52	2.54	2.71	2.73	2.37	2.60
16%	1.98	1.78	1.78	1.85	2.33	2.15	2.59	2.36	2.38	2.4	2.47	2.42
20%	2.19	2.36	2.03	2.19	2.43	2.17	2.47	2.36	1.35	2.25	2.01	1.87

Table 1: Results of objective testing

4.2 Subjective Tests

The results obtained from subjective testing are given as below

Table 1					Table 2					Table 3				
Loss Corr	Loss	Av .MOS-LQS	MOS-LQO	Av. MOS-LQO	Loss Corr	Loss	Av .MOS-LQS	MOS-LQO	Av. MOS-LQO	Loss Corr	Loss	Av .MOS-LQS	MOS-LQO	Av. MOS-LQO
0.3	0%	2.91	3.32	3.38	0.5	0%	2.83	3.23	3.28	0.8	0%	2.92	3.31	3.22
0.3	4%	3.23	3.01	2.92	0.5	4%	2.73	3.10	3.07	0.8	4%	2.80	2.90	2.98
0.3	8%	3.14	2.69	2.54	0.5	8%	2.29	2.72	2.64	0.8	8%	2.97	2.67	2.74
0.3	12%	2.25	2.41	2.34	0.5	12%	2.31	2.60	2.54	0.8	12%	2.40	2.37	2.6
0.3	16%	1.79	1.98	1.84	0.5	16%	2.14	2.15	2.35	0.8	16%	1.61	2.38	2.41
0.3	20%	1.81	2.19	2.19	0.5	20%	1.93	2.43	2.35	0.8	20%	1.39	2.25	1.87

Table 2: Results of subjective testing

4.3 Objective Tests based on E-model (MOS-CQE)

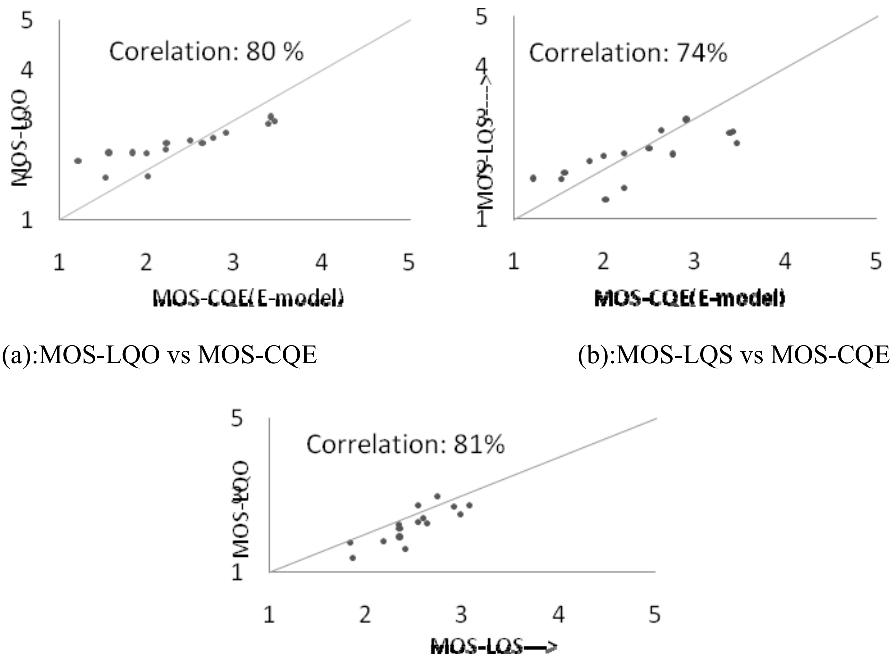
The results obtained from E-Model calculation are as follow

Table 1				Table 2				Table 3			
Loss Corr	Packet Loss	R value	MOS-CQE	Loss Corr	Packet Loss	R value	MOS-CQE	Loss Corr	Packet Loss	R value	MOS-CQE
0.3	4%	65.51	3.38	0.5	4%	66.11	3.41	0.8	4%	67.13	3.46
0.3	8%	51.04	2.63	0.5	8%	53.32	2.75	0.8	8%	56.17	2.9
0.3	12%	38.47	1.99	0.5	12%	43.13	2.22	0.8	12%	48.37	2.49
0.3	16%	27.78	1.52	0.5	16%	35.06	1.83	0.8	16%	42.93	2.21
0.3	20%	18.51	1.21	0.5	20%	28.79	1.56	0.8	20%	38.89	2.01

Table 3: Results of Estimation based testing

4.4 Analysis

The first analysis which we will do is the correlation between the three types of Voice quality measures. We have already collected the necessary data in the Testing portion of the project.



(a):MOS-LQO vs MOS-CQE

(b):MOS-LQS vs MOS-CQE

(c):MOS-LQO vs MOS-LQS

Figure 2: Pearson Correlation plots between (a) PESQ Objective Values MOS-LQO and E-Model values MOS-CQE (b) Subjective values MOS-LQS and E-Model values MOS-CQE (c) PESQ Objective values MOS-LQO and Subjective values MOS-LQS

The first correlation plot is between PESQ Objective values of MOS-LQO and E-model value of MOS-CQE, in which correlation comes out to be 80%. The correlation between Subjective results MOS-LQS and estimation based E-Model MOS-CQE results came out to be 74% and that is shown in Fig 2(b). Fig 2 (c) is the correlation between MOS-LQS and MOS-LQO. Thus we can see that the objective results have a better correlation as compared with Subjective and Estimate based results. Hence, PESQ objective MOS correlates much better with subjective MOS than E-model MOS.

If we look into the objective test results with correlation of 0.3, there is an interesting point, MOS values increased when packet drop is increased from 16% to 20%. During the testing phase, readings from 0% to 16% were taken continuously, but

NistNET was turned off and on again for 20% loss reading. Hence, when NistNET was started again, the MOS values were not in a continuous fashion but changed. To investigate that, we observed NistNET box with loss correlation of 0.8 with loss percentages of 4%, 8% and 20%. We checked by sending 10, 50 and 100 ping packets through the network and observed that the more the time or packets sent, better the packet drop percentage achieved. For 20% drop rate in NISTnet, the average loss rate in ping packets were 0%, 8% and 21% for 10, 50 and 100 packets sent. Also loss percentages of 4% and 8% were achieved when average packet loss in 100 Ping packets were observed. However, the average loss was not satisfactorily close to the set rate if observed with 10 or 50 ping packets. Therefore, we observed that when time equivalent to 100 ping packets (200 seconds) is given to NistNET before carrying out any test, the average loss rate is almost achieved and results are more accurate.

Another observation that we have was the analysis for the waveform of 16% drop and correlation of 0.3 in the objective testing. We will consider the waveform of the original file first.

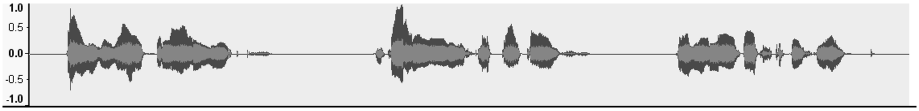


Figure 3: Original waveform

Now we will see the waveform of the file that suffers a loss of 16% with correlation of 0.3.The plot is shown below, which indicates that due to high amplitude at some points, the waveform was clipped.

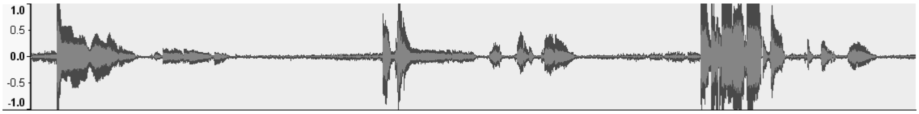


Figure 4: Waveform received from X-Lite

While the original file doesn’t have the elements of so high amplitude, some gain was added to the wave file that resulted in the high amplitude and eventually clipping for high amplitude values. Tests were carried out with same parameters, however, clippings or high gain was not achieved. Literature shows that X-Lite has an auto gain function, however, this was checked with Skype and compared with the waveform received from X-Lite, but results show no sign of high gain added by X-lite. Hence, proving that there was some gain added to the incoming speech data, but not from the X-Lite. The waveform for the same parameters as received from Skype is shown below, which resembles the real waveform and no increase in gain is seen.

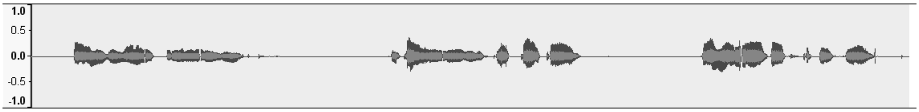


Figure 5: Waveform received by Skype

The reason for the high amplitude of the sound file received was that the system was not calibrated and was sent with high volume or it was the disturbance in the system for example audio cable was not connected properly etc. The standard method for objective and subjective MOS testings require certain calibrations, however, due to time and space constrains, these tests were carried out in a normal Lab environment. Hence we can say that before starting the experiment, the system should be calibrated in order to minimize the error or to localize any type of disturbance that can affect the speech quality. We also used virtual cable instead of physical connection between the input and output of sound card, and find out that the MOS results obtained from Virtual cable are better than physical cable. Hence, in short, system calibration and more use of controlled equipment (software based instead of hardware) should be used in order to minimize the external disturbances that can affect the test results.

5 Conclusion

In this experiment, we correlated PESQ and E-Model with subjective MOS results. A testbed was setup, which resembles a WAN environment. NistNET was used to introduce controlled packet loss in the network. Tests were carried out with two different VoIP end terminals, Counterpath X-Lite and Skype. Standard ITU-T speech sample was sent from one end client to other, and was recorded using audacity software or built-in recorder in X-Lite. The correlation between the Subjective and Objective scores, MOS-LQS and MOS-LQO, came out to be 81 percent. The correlations were measured using the Pearson correlation equations. Similarly the correlation between the Subjective scores (MOS-LQS) and Estimation based scores (MOS-CQE) came out to be 74%. While the correlation between the Objective and Estimation based results came out to be 80%. If we compare the three correlations that we have determined, it’s obvious that the Objective scores correlate better (80% and 81%) then subjective or estimation based scores. Thus over all, we can say that PESQ correlate much better than the E-Model. The other conclusions that we obtained from this experiment is about the NistNET (Network emulator) software. We came to an important observation that NistNET should be given enough time (the time of 100 ICMP ping packets echo relies) so that the packet drop rate is averaged at the set rate. This conclusion holds good for packet drop ratios of 20% or less. The final important observation which we obtained from the experiment is that the testbed should be calibrated carefully before carrying out any type of test. For example, there was a high gain input or any cable of the system not connected properly, which introduced this noise and high gain. Due to time constrains, some issues, especially of system calibration were left for future research and analysis. Virtual cable was also used instead of Physical cable that connects the audio out and audio in of the sound card through a driver software. The MOS scores obtained by using virtual cable were better than with physical cable. Hence, we can conclude that the system should be carefully calibrated for any of the test especially in terms of senders voice gain, and in controlled environment, where external disturbances can be minimized (using Virtual software cable instead of physical one) so that the results achieved are more accurate, or any issue can be localized and analyzed.

6 References

Ding, L. and Goubran, R. 2003. Speech Quality Prediction in VoIP Using the Extended E-Model. IEEE, Ottawa,ON. Available online at: http://portal.acm.org/citation.cfm ?id=1268177.

Dudman, J. 2006. voice over IP:what it is, why people want it, and where it is going. JISC. Available online at: www.jisc.ac.uk/media/documents/techwatch/tsw0604openoffice.odt.

Khasnabish, B. 2003. *Implementing Voice Over IP*. Published by Wiley-IEEE.

NistNET. 2001. NistNET Homepage. Available online at: http://snad.ncsl.nist.gov/nistnet/. (Accessed: 03/03 2008)

OPTICOM. 2007. Voice Quality Testing. Available online at: www.opticom.de. (Accessed: 8/08 2008)

Sun, L. and Ifeachor, E. 2004. New Models for Perceived Voice Quality prediction and their Applications in Playout Buffer Optimization for VoIP Networks. University of Plymouth, Plymouth. Available online at: http://www.tech.plym.ac.uk/spmc/people/lfsun/.