New Methods for Voice Quality Evaluation for IP Networks

Lingfen Sun and Emmanuel Ifeachor

Department of Communication and Electronic Engineering
University of Plymouth
Plymouth PL4 8AA, United Kingdom

The need to evaluate voice quality in VoIP (Voice over IP) applications is an important requirement for technical and commercial reasons. This may involve subjective and/or objective voice quality measurements, but existing methods may not always be appropriate for VoIP applications. The aims of the study reported in the paper are to investigate new subjective and objective measurement methods for VoIP applications. The contributions of the paper are two-fold. First, we present a new subjective, Internet-based MOS (Mean Opinion Score) test methodology which allows rapid assessment of voice quality. We conducted MOS tests using the new method as well as traditional MOS tests under different VoIP network conditions and compared the results using objective measurement methods. Preliminary results show that the Internet-based MOS test compares well with traditional MOS test (correlation coefficients of 0.95). Second, we propose novel conversational intrusive and non-intrusive speech quality measurement methods, based on the ITU PESQ and E-model to extend the applicability of existing methods. We illustrate the application of the novel approach to the derivation of model parameters for a new codec for VoIP applications (the AMR codec).

1. INTRODUCTION

The convergence of communications and computer networks has led to a rapid growth in real-time applications such as Voice over IP (VoIP). However, IP networks are not designed to support real-time applications and factors such as network delay, jitter and packet loss lead to unpredictable deterioration in perceived voice quality. A major challenge that faces network and service providers is how to measure or predict voice quality accurately and efficiently for Quality of Service (QoS) monitoring and/or control purposes to meet technical and commercial requirements.

Voice quality measurement can be carried out using either subjective or objective methods. The Mean Opinion Score (MOS) is the most widely used subjective measure of voice quality and is recommended by the ITU [1]. A MOS value is normally obtained as an average opinion of quality based on asking people to grade the quality of speech signals on a five-point scale (Excellent, Good, Fair, Poor and Bad) under controlled conditions. In voice communication systems, MOS is the internationally accepted metric as it provides a direct link to voice quality as perceived by the end user. The inherent problem in subjective MOS measurement is that it is slow, time consuming, expensive and cannot be used for long-term or large scale

voice quality monitoring in an operational network infrastructure. This has made objective methods very attractive for meeting the demand for voice quality measurement in communication networks.

Objective measurement of voice quality in modern communication networks can be intrusive or non-intrusive. Intrusive methods are more accurate, but normally are unsuitable for monitoring live traffic because of the need for a reference data and to utilise the network. A typical intrusive method is based on the latest ITU standard, P.862 Perceived Evaluation of Speech Quality (PESQ) Measurement Algorithm [2]. This involves a comparison of the reference and the degraded speech signals to predict the listening-only one-way MOS score.

Non-intrusive methods do not need a reference signal and are appropriate for monitoring live traffic. ITU-T E-model [3] is the most widely used non-intrusive voice quality measurement method and may be used to predict conversational MOS score directly from IP network and/or terminal parameters [4,5].

Subjective methods are crucial for benchmarking objective methods. The need remains for an efficient method for subjective MOS tests. In the paper, we introduce a new subjective, Internet-based MOS test methodology, intended to simplify MOS tests for VoIP applications. We conducted MOS tests using the new method as well as traditional MOS tests for speech samples under different VoIP network conditions, and compared the performance with the latest ITU-T objective measurement methods (e.g. PESQ and E-model). Preliminary results show that the Internet-based MOS test compares well with traditional MOS test method.

The PESQ algorithm provides a more accurate measure of quality, but it is intrusive and can only predict one-way listening speech quality. In practice, there is a need for objective measure of conversational speech quality to account for interactivity in voice communication. In this paper, we present a novel conversational, intrusive speech quality measurement method, based on a combination of PESQ and E-model.

The current E-model [3] and extended E-models [4,5] rely on subjective tests for the derivation of model parameters when they are used for VoIP applications. This is obviously time consuming, impractical and hinders the use of the E-model in new and emerging applications. Non-subjective derivation of model parameters has recently been proposed [6], but this is limited to only codec impairments and is unsuitable for VoIP applications. In this paper, we introduce a new objective method for deriving model parameters for VoIP applications. This should extend the applicability of the E-model to meet the needs of new and emerging applications.

The remainder of the paper is structured as follows. In Section 2, a new Internet-based MOS test and results of its use to assess voice quality under different VoIP network conditions are presented. New intrusive and non-intrusive conversational speech quality measurement methods are presented in Sections 3 and 4, respectively. Section 5 concludes the paper.

2. INTERNET-BASED MOS TEST

The traditional MOS test methodology has been in existence for about 20 years [7] and today its use range from the assessment of codec's quality to the assessment of VoIP network quality. The stringent test requirements for traditional tests have not changed (e.g. the use of a sound-proof room) in that time and are essential for a proper assessment of voice quality in many cases, e.g. quality assessment of codecs, as the difference between codecs may be subtle and difficult to detect. However, for VoIP applications, new impairments, such as packet loss,

are much more perceptible than impairments from codecs. This has led us to investigate the possibility of conducting MOS tests under normal working/studying environments, as this is more realistic and subjects are more relaxed. In a sound-proof room, some subjects may find it uncomfortable, psychologically, to carry out tests in the confined environments. This has led to an Internet-based subjective test methodology, which has the following advantages:

- It is closer to reality than the traditional method. Subjects remain in familiar environments, e.g. an office or a laboratory, to carry out the test. This is clearly less stressful and the test can be done at the subject's own pace.
- It is possible to organise subjective tests at more locations around the world.
- It allows easier access to a larger number of subjects (e.g. 40 - 80 subjects can be tested at the same time in one or two large rooms, e.g. a laboratory).

Overall, it has the benefits of efficiency, realism, wide access and ease of organisation. It can save money and time compared to P.800. Of course, the main disadvantage of Internet-based MOS test is the lack of a controlled testing environment (e.g. very low background noise) compared to P.800.

In a previous Internet-based MOS study [8], we carried out the tests without control (subjects did their own tests on their own computer, in their own office and at their own preferred time slot). We have extended this by introducing a measure of control to reduce the impact of different working environments on the results. In the Internet-based MOS test method, all subjects sit in a large project room which they use regularly. It is not a sound-proof room, but it is quiet and has Internet access.

In the next section, we will present the set-up used to evaluate voice quality using both subjective and objective methods. The preliminary test results and analysis are then presented.

2.1. Voice Quality Evaluation

Figure 1 depicts the set-up used for the voice quality evaluation. It is a PC-based software system that allows the simulation of key processes in voice over IP and speech quality measurement. Objective voice quality measurements were made with the ITU PESQ and E-model to enable us to compare Internet-based MOS tests with traditional MOS tests. Reference speech files were first encoded using G.723.1 codec [9] and then processed in accordance with network parameter values in trace data files (see later) and then decoded to generate degraded speech signals (a fixed jitter buffer, for simplicity, was used to remove the effects of jitter. Packets that arrive too late are discarded).

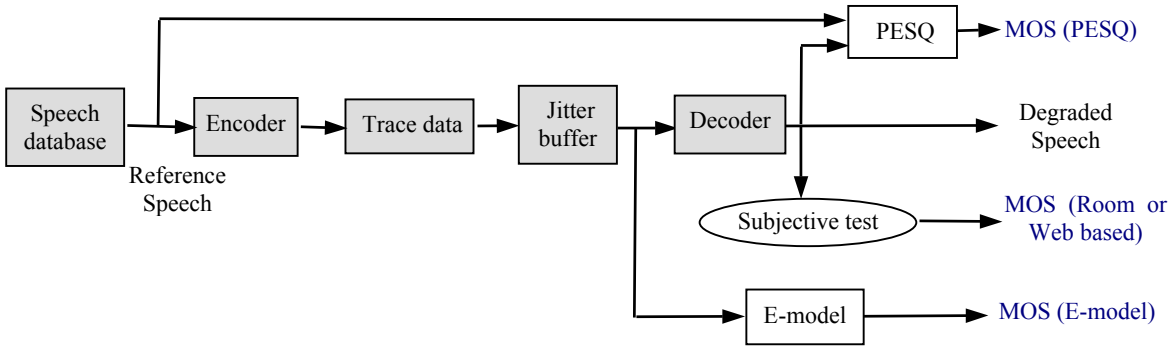

Figure 1. VoIP Speech Quality Evaluation Set-up

We collected Internet trace data between the UK and USA, UK and China and UK and Germany using a UDP/IP probe tool in the past year. A detailed description of the trace data collection and the effects of jitter buffer can be found in our paper [10]. Some trace data (e.g. those with a 30 ms packet interval, consistent with G.723.1) was selected for the quality evaluation.

The reference speech database was taken from the TIMIT data set [11]. Each speech sample consists of four short sentences spoken by four different male and female speakers in order to keep a balanced design. Each speech sample was about 10 to 15 seconds long. A total of 10 speech samples were chosen for the VoIP quality evaluation. We also chose ten different network conditions from the Internet trace data set, covering packet loss rate (including late arrival loss due to jitter) from 0% to 30%. Ten degraded speech samples were generated and used for the quality evaluation.

Subjective Tests were carried out using two methods -- Internet-based (or web-based) and P.800-based tests (Room-based). A website for the MOS test was created at the following URL:
http://www.tech.plymouth.ac.uk/spmc/people/lfsun/mos/

The 10 degraded speech samples were put on the web, together with a brief instruction about the MOS test. 15 undergraduate students were invited to attend the controlled Internet-based MOS test. The tests were carried out in the project laboratory which they use regularly. The room was quiet and similar to a normal office environment. Brief instructions were given by a supervisor before the test. The students were asked to perform the test at their own pace. The tests took about 15 minutes to complete. When all the students had submitted their opinion scores, the MOS score were calculated and expressed as Web_MOS.

In order to compare the results of the controlled Internet-based MOS tests with similar P.800 tests, we carried out another round of MOS test in a small, quiet room (a sound-proof room was not available on-site and is another motivation for investigating the web-based approach). The room is about 8 square meters and 3.5 meters high with a desktop and a laptop PCs. A similar test procedure to the web-page was created locally. The same 10 degraded speech samples were chosen. The 15 students were invited again to carry out the tests, one by one. The test spanned over two days because of the numbers involved and their availability. The MOS score for the room-based MOS test was expressed as Room_MOS.

We also conducted MOS tests using the ITU PESQ algorithm [2] and the E-model [3] in order to compare the subjective test results with objective measurements. Listening-only speech quality measurements were considered in order to keep the same conditions. By comparing the reference speech and the degraded speech, an objective MOS score was obtained from the PESQ algorithm. This MOS score was referred to as MOS (PESQ) or PESQ_MOS. For the E-model, only the effects of the Equipment Impairment (I_e) were taken into account (the effects of delay, I_d, was not considered). This gives a listening-only MOS score which is referred to as MOS (E-model) or E-model_MOS.

Table 1. Objective and subjective MOS scores for different speech samples

Test samples	1	2	3	4	5	6	7	8	9	10
PESQ	3.18	2.65	2.85	3.74	2.02	1.95	2.42	2.59	2.93	2.54
E-model	2.90	2.50	2.56	3.92	1.00	1.04	1.63	2.07	2.71	2.41
Room-based	3.36	2.65	2.85	3.31	1.41	1.15	2.13	2.13	2.97	2.34
Web-based	3.37	2.32	2.52	4.00	1.22	1.11	2.19	2.04	2.90	2.35
Loss rate (%)	5.68	9.51	8.85	0.21	29.5	23.1	18.5	14.6	7.25	10.6

2.2. Test Results and Analysis

The results for the 10 degraded speech samples for each of the four methods of voice quality measurement (PESQ, E-model, Internet-based and Room-based MOS tests) are summarised in Table 1. The sequences of the test speech samples from 1 to 10 are the same with that on the MOS test website. The calculated packet loss rates are included to give an indication of the impact of the network impairment..

The relationships between the MOS scores and packet loss rates for the different MOS test methods are depicted in Figure 2. From the figure, it can be seen that the MOS scores for all four evaluation methods decrease with increasing packet loss rate. When the packet loss rate is low, the E-model, PESQ and Web-based MOS scores are quite close. When packet loss rate is high, PESQ seems to over predict the voice quality, whilst the E-model does the opposite. As the E-model predicts voice quality directly from network parameters (e.g. packet loss rate), it does not consider factors such as packet loss location. Also as VAD (voice activity detection) was not activated in the simulation, packet loss in the silence period will not be perceived by subjects, but it was still taken into account in the E-model calculation. This is partly why the E-model gives lower MOS scores compared to the other methods when the packet loss rate is high. Room-based and Web-based MOS scores are close, except in the case when there is almost no packet loss. This is probably because the background noise (e.g. from the fan) of the computer for Room-based tests is higher than those for Web-based test.

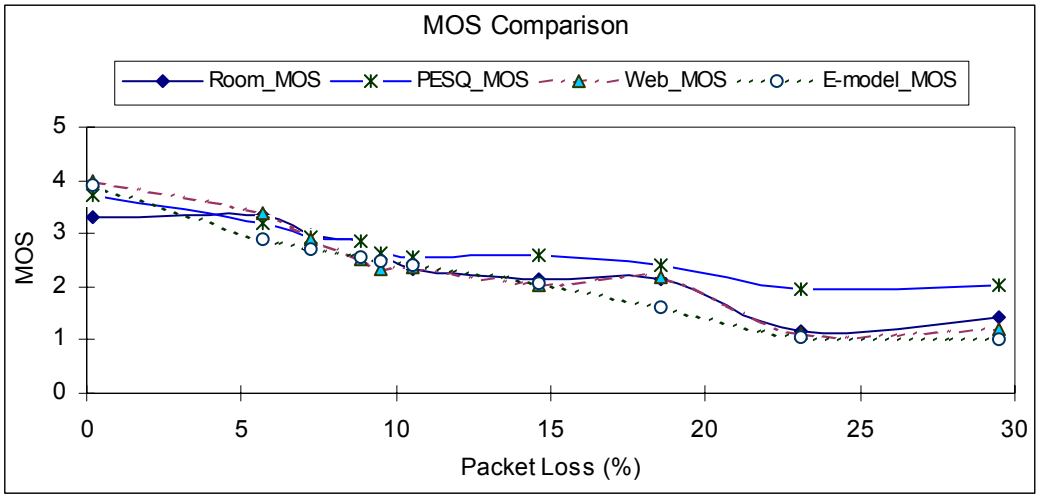

Figure 2. MOS comparison for objective and subjective test methods

Table 2. Correlation coefficients for MOS comparison

Name	PESQ vs Room_MOS	PESQ vs Web_MOS	E-model vs Room_MOS	E-model vs Web_MOS	Web_MOS vs Room_MoS	E-model vs PESQ
Correlation Coefficients	0.933	0.984	0.935	0.964	0.952	0.975

The Pearson correlation coefficient between the results of subjective and objective methods were calculated and the results are shown in Table 2. From the table, it can be seen that the Internet-based MOS test (Web_MOS) compares well with the traditional MOS test (Room_MOS) (correlation coefficients of 0.95). This suggests that with the Internet-based MOS test it is possible to obtain similar results to those of traditional MOS tests for VoIP applications. For objective measurement methods (E-model and PESQ) and subjective methods (Room-based and Web-based), the correlation coefficients are between 0.93 to 0.98. This shows that the two objective methods can both predict subject MOS score well, although both seem to predict Web-based MOS better than Room-based MOS.

3. INTRUSIVE CONVERSATIONAL SPEECH QUALITY MEASUREMENT

PESQ is an intrusive method and can only predict one-way listening speech quality. It does not consider the impact of end-to-end delay which is important for interactivity in communications. We propose a conversational quality measurement method which exploits the accuracy of the PESQ algorithm and the delay model of the E-model, see Figure 3.

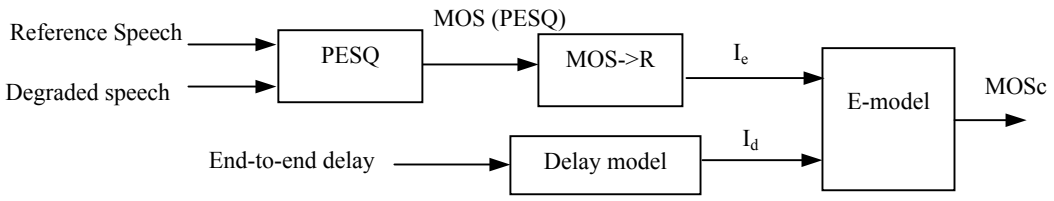

Figure 3. An intrusive conversational speech quality measurement

As shown in Figure 3, the listening *MOS* score is first obtained using PESQ (referred to as *MOS (PESQ)*). The *MOS* score is then converted to a rating factor (the *R* factor) and then to an equipment impairment value (the I_e value). The conversational MOS score, *MOSc*, is obtained by combining the I_e value and the effects of end-to-end delay (I_d values). The detailed procedures are given in the following steps.

The ITU-T G.107 [3] defines the relationships between the *R* to *MOS* as in (1).

$$\begin{aligned} MOS &= 1 && for \quad R \le 0 \\ MOS &= 1 + 0.035R + R(R-60)(100-R)7\times10^{-6} && for \quad 0 < R < 100 \\ MOS &= 4.5 && for \quad R \ge 100 \end{aligned} \tag{1}$$

However, Equation (1) cannot be inverted directly to obtain the R values because it covers the R-values between 0 and 6.5, which maps to MOS scores below 1. Thus, the R-values are normally restricted to the range [6.5, 100] , with R-values below 6.5 assigned a $MOS = 1$

before inversion. Candono's Formula [15] can be used to obtain the R-values from the MOS, but the equations are very complicated. Thus, we propose to use a simplified 3rd order polynomial fitting (Equation 2) to obtain the equation for mapping from MOS to R values. The fitting curve and original curve from G.107 are shown in Figure 4.

$$R = 3.026MOS^3 - 25.314MOS^2 + 87.060MOS - 57.336 \quad (2)$$

If we consider only the equipment impairment, R values can be converted to I_e using Equation (3) (a default R value of 93.2 is used [3]).

$$I_e = 93.2 - R \quad (3)$$

The delay impairment factor, I_d, represents all impairments due to delay of voice signals, and includes impairments due to Listener Echo, Talker Echo and Absolute delay. I_d can be calculated by a series of complex equations [3]. The relationships between I_d and one-way delay can be expressed by a simplified equation (4) according to [5] . The corresponding fitting curve and the curve from G.107 [3] are shown in Figure 5.

$$I_d = 0.024T_a + 0.11(T_a - 177.3)H(T_a - 177.3)$$
$$where \begin{cases} H(x) = 0 & \text{if } x < 0 \\ H(x) = 1 & \text{if } x \geq 0 \end{cases} \quad (4)$$

Considering I_d and I_e, E-model R factor can be simplified as in (5).

$$R = 93.2 - I_d - I_e \quad (5)$$

From R, the conversational MOS score ($MOSc$) can be calculated using Equation (1). Overall, the conversational MOS score can be obtained from comparing the reference and degraded speech samples and taking into account the effects of end-to-end delay.

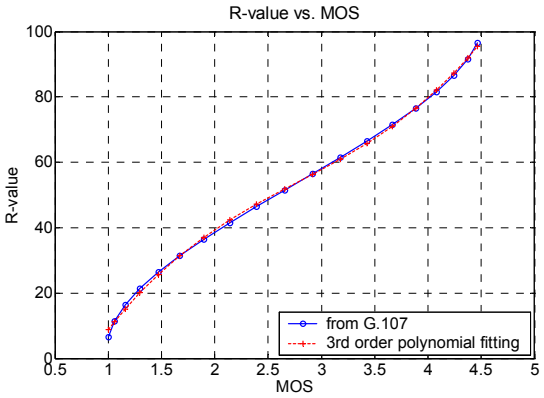

Figure 4. R-value vs. MOS score

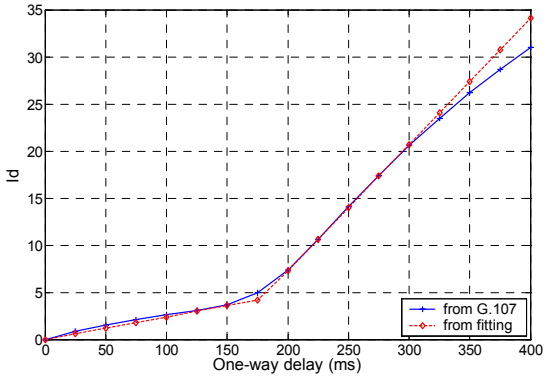

Figure 5. I_d vs. one-way delay

4. NON-INTRUSIVE OBJECTIVE MEASUREMENT

The basic E-model has been extended in several directions (e.g. extended or simplified models) and used for non-intrusive voice quality monitoring for VoIP applications. However, the equations for equipment impairment (I_e) are still based on subjective tests. Such tests are required in order to derive the model parameters for I_e for each codec (e.g. new emerging codecs), application (e.g. with/without *VAD*) and network condition (e.g. packet size, network packet loss such as random or burst packet loss). This is obviously time consuming and impractical and hinders the development of the E-model for future applications.

To improve the applicability of the E-model, we have extended it in two directions:

First, we have proposed the use of the Internet-based MOS test methodology to increase the efficiency of subjective MOS tests. As described above, the use of the Internet-based MOS test method to derive model parameters is more efficient. Second, we have proposed the use of an objective method, such as PESQ, to replace the subjective tests which are currently required for deriving the model parameters. As in Figure 3, the Equipment Impairment factor (I_e) can be derived directly from an objective measurement method (e.g. PESQ). Of course, the accuracy of the resulting model parameters will depend on the accuracy of the objective measurement method used, but this will improve as new objective measurement algorithms (e.g. the next generation of PESQ) become available.

As an example, we derived the I_e value for a new codec for VoIP applications using PESQ (the AMR [12] at the highest mode of 12.2 Kb/s. I_e model does not exists for AMR codecs at present in public domain). The procedures are as follows:

Step 1: Obtain MOS (PESQ) vs. packet loss rate for the AMR codec (Figure 6). Following the approach in Figure 1, we have used a random packet loss generator instead of real trace data. The packet size is set to 1 frame/packet (20ms). We obtained MOS (PESQ) at different packet loss rates (from 0 to 30% in steps of 3%) for the AMR Codec. Each MOS value was calculated by averaging over 25 different random seeds for both male and female speech samples from ITU-T dataset [13] to avoid the influence from packet loss location and gender.

Step 2: Convert the MOS vs. packet loss rate to I_e vs. packet loss rate as shown in Figure 7 (the curve from PESQ) using Equations (2) and (3). A logarithm fitting function, similar as in [5], can be derived as Equation (6). The fitting curve is also shown in Figure 7 (from fitting).

$$I_e = 13.2 + 15.84 * \ln(1 + 0.38 * loss) \qquad (6)$$

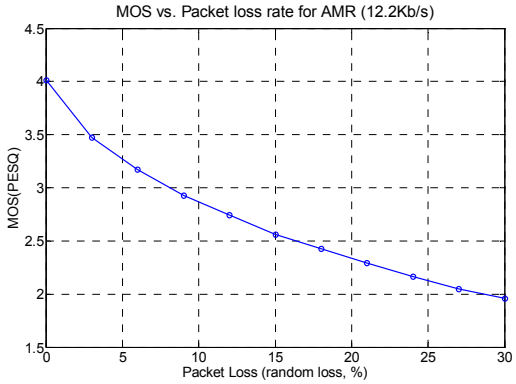

Figure 6. MOS vs. Packet loss for AMR

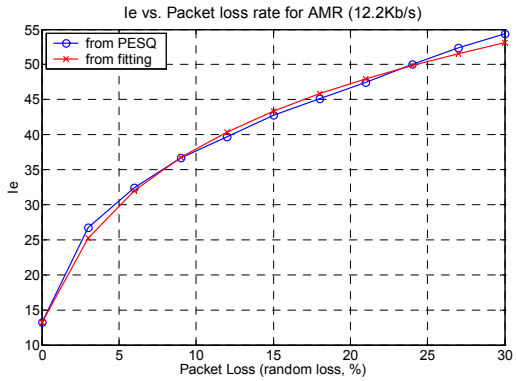

Figure 7. I_e vs. Packet loss for AMR

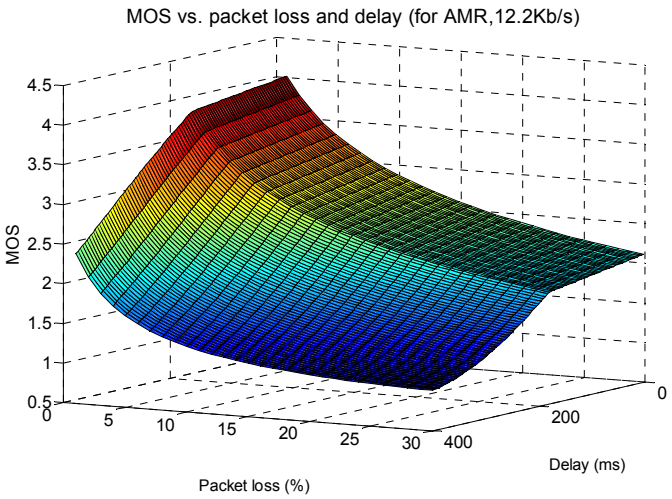

Figure 8. MOS vs. Packet loss rate and delay for AMR (12.2.Kb/s)

Step 3: Calculate the MOS for AMR codec (12.2 Kb/s mode) using Equations (6), (4), (5), and (1) for a given random packet loss rate and end-to-end delay. The MOS vs. packet loss rate and delay is shown in Figure 8.

This method can be extended to other speech codecs, current or new ones (above 4.8Kb/s [2]) and packet loss patterns (e.g. burst packet loss). It can be easily used to monitor/predict conversational speech quality from network impairments (e.g. packet loss rate and end-to-end delay) non-intrusively [5, 14].

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have investigated novel subjective and objective speech quality evaluation methods for VoIP applications. We have proposed a new Internet-based subjective MOS test and carried out informal subjective MOS tests. Preliminary results show that the Internet–based MOS test compares well to the Room-based MOS test (correlation coefficients of 0.95). In general, the two ITU objective test methods (PESQ and E-model) can predict subjective MOS scores well. We have introduced improved intrusive and non-intrusive conversational speech quality measurement methods which exploit the capabilities of PESQ and E-model.

Future work will focus in two directions. First, we will investigate further the Internet based MOS test methodology by undertaking a more extensive MOS tests. We wish to establish, for example, how the test environment affects the results, what the differences between controlled Internet-based test and uncontrolled Internet-based tests and between Internet-based MOS test and formal P.800-based MOS tests (in sound-proof room) really are. Secondly, we will investigate further the new intrusive and non-intrusive measurement methods and how to use them in new applications (e.g. in perceived quality driven QoS control systems).

ACKNOWLEDGEMENT

We would like to thank Mr. Chunshui Liu for his contribution during his MSc study at the university.

We are grateful to Acterna for part sponsorship of our work.

REFERENCES

1. ITU-T Rec. P. 800, Methods for subjective determination of transmission quality, August 1996.
2. ITU-T Rec. P. 862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, February 2001.
3. ITU-T Recommendation G.107, The E-model, a computational model for use in transmission planning, 2000.
4. A. D. Clark, Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality, IPTEL'2001, pp. 123-127, April, New York, 2001.
5. R. G. Cole and J.H. Rosenbluth, Voice over IP Performance Monitoring, Journal on Computer Communications Review, vol. 31, no.2, April 2001.
6. S Möller and J Berger, Describing Telephone Speech Codec Quality Degradations by Means of Impairment Factors, J. Audio Eng. Soc., Vol. 50, No. 9, September 2002, pp. 667-680.
7. ITU-T P.830, Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs, February 1996.
8. L Sun and E Ifeachor, Subjective and Objective Speech Quality Evaluation under Bursty Losses, Proceedings of On-line Workshop Measurement of Speech and Audio Quality in Networks (MESAQIN 2002), Prague, Czech Republic, Jan. 2002, pp.25 - 29.
9. ITU-T Recommendation G.723.1, Dual Rate Speech Coder for Multimedia Communication Transmitting at 5.3 and 6.3 kbit/s, March 1996.
10. L Sun and E Ifeachor, Prediction of Perceived Conversational Speech Quality and Effects of Playout Buffer Algorithms, to appear in the Proceedings of IEEE International Conference on Communications (ICC), Anchorage, USA, May 2003.
11. TIMIT data set, J. S. Garofolo, L. F. Lamel, W. M. Fisher, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus.
12. ETSI EN 301 704 V7.2.1 (2000-04), Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR) speech transcoding.
13. ITU-T Recommendation P.50, Appendix 1, Test signals, 1999.
14. A. P. Markopoulou, F. A. Tobagi, M.J. Karam, Assessment of VoIP Quality over Internet Backbones, Proc. of IEEE Infocom, June 2002.
15. C. Hoene, B. Rathke and A Wolisz, On the Importance of a VoIP Packet, In Proc. of ISCA Tutorial and Research Workshop on the Auditory Quality of Systems, Germany, April 2003 (http://www.tkn.tu-berlin.de/publications/papers/paper10.pdf).