

Assessment of VoIP Quality over Access Networks

M. Amir Mehmood, Pakistan Internet Exchange, IT Infrastructure Division, PTCL, Lahore, Pakistan

Tariq M. Jadoon, Lahore University of Management Sciences, Lahore, 54792, Pakistan

Noor M. Sheikh, University of Engineering and Technology, Lahore, 54890, Pakistan

Abstract— This paper assesses VoIP quality over access networks in Pakistan using a delay jitter measurement methodology for evaluating the perceptual quality of voice calls using the ITU-T G.107 speech quality E-model. Passive measurements for voice calls in the presence of background Internet data traffic for G.723.1 and G.729a codecs are carried out using a non-intrusive parametric model. The R-factor and resultant Mean Opinion Scores (MOS) were calculated at different link loads and congestion hot spots were identified. The study highlights the inadequacy of access networks for handling VoIP traffic at current in Pakistan and suggests alleviating congestion by increasing capacity in access networks.

Index Terms—VoIP, Perceptual Quality Assessment, E-Model.

I. INTRODUCTION

The Internet is evolving into the ubiquitous packet switched infrastructure that aspires to provide an Integrated Broadband Network seamlessly integrating voice, video, data and multimedia traffic. Converging telephone and IP networks entails providing the same “toll-quality” service over best-effort IP networks. This is a significant engineering challenge bearing in mind that we now consider the high voice quality standards that are a hallmark of the public switched telephone network (PSTN) for granted.

Voice quality is ultimately adjudged by the listener and thus, speech quality is inherently *perceptual* or subjective in nature. Using a numeric scale ranging from 1 (unacceptable) to 5 (excellent), the Mean Opinion Score (MOS) test provides a widely accepted measure for subjective speech quality [1]. However, assessing speech quality through surveys is a time consuming and expensive process. A viable alternative is to develop *quality models* that simulate human rating behaviour by correlating perceptual QoS with quantifiable parameters. This is not a straightforward process as objective metrics do not necessarily correlate well with ‘perceptual quality’. A number of quality models and tests that provide objective MOS measures by correlating well with subjective scores have been developed. Rix classifies

these tests as either intrusive or non-intrusive test methods [2]. Intrusive testing methods involve comparing a reference acoustic speech signal with a degraded version of the signal received through the system under test. The ITU-T standardised the Perceptual Speech Quality Measure (PSQM) [3] in 1996. Problems in aligning the reference and degraded signal which are especially accentuated in VoIP networks necessitated improving PSQM and a new model called the Perceptual Evaluation of Speech Quality (PESQ) [4] was standardised by the ITU-T as P.862 in 2001. The E-model is a non-intrusive parametric model that is well-established as a transmission quality model. It is defined in ITU-T G. 107 [5] and is based on the principle that transmission impairments combine additively into a single psycho-acoustic transmission rating (R-factor) on a scale of 0 to 100. The R-factor can further be translated into a MOS through a simple transformation.

Sending speech as packets over the Internet entails sampling the original voice signal at a fixed rate and converting each sample in to a fixed number of bits. This constant bit rate stream is then either directly filled in packets of an appropriate size or is processed in frames of 10-30ms duration and compressed before packetization. Packets are subsequently prefixed with RTP/UDP/IP headers. Thus, a sample must wait for an algorithmic, processing and packetization delay before it can be placed on the wire. VoIP packets that traverse the Internet are subject to two principal impairments namely, packet loss and packet delay. Loss may either be due to congestion and may lead to packets being discarded at intermediate nodes or it may result from a failure of network components such as links and/or nodes. Packet delay has a fixed component as a consequence of the propagation and transmission delay as well as a variable component as a result of variable queueing delays packets encountered along buffers at intermediate nodes whilst traversing the Internet. Thus, packets received at the receiver do not have the same temporal relationship as they did at the sender resulting in delay jitter. An appropriately sized fixed or adaptive dejitter or playout buffer compensates for most of this at the expense of an added delay. Delay jitter manifests itself as packet loss for packets that arrive latter than a maximum threshold and degrades the quality perceived by the listener. The conversational quality of a call is primarily affected by the end-to-end delay in addition to the packet loss and delay jitter. These parameters constitute the network QoS

This work was supported in part by a research grant from the PTCL R&D Fund R&DF/Thematic-01/2004/06.

The author's e-mail addresses are (amir.mehmood@ptcl.net.pk jadoon@lums.edu.pk and adstec@mailcity.com)

parameters and can be mapped to a perceptual QoS measures such as the MOS through quality models such as the E-model. This paper examines the perceptual VoIP quality over access networks using a simple delay jitter methodology similar to that suggested by Cole and Rosenbluth [6]. The balance of the paper is organized as follows. Section II reviews the E-model. Section III describes the experimental set-up, measurements and results. Section IV concludes the paper by summarizing key results.

II. E-MODEL

The E-model was developed as a standard for measuring the transmission quality of narrowband telephony by ETSI [7] and latter standardized by the ITU-T as recommendation G.107. The E-model is a computational model that determines a transmission quality rating called the ‘R’ factor from the transmission parameters to predict the quality of the “mouth to ear” (M2E) speech path. The typical range for the R factor is 0-100 for the PSTN and corresponding values of the MOS are 1-5 as shown in Fig 1.

The basic principle of the E-model is that the various impairments contributing to the overall perception of voice quality are additive when converted to the appropriate psycho-acoustic scale (R). The basic formula for the E-model is:

$$R = R_0 - I_s - I_d - I_{e-eff} + A \quad (1)$$

where R_0 is the signal to noise ratio of the connection.

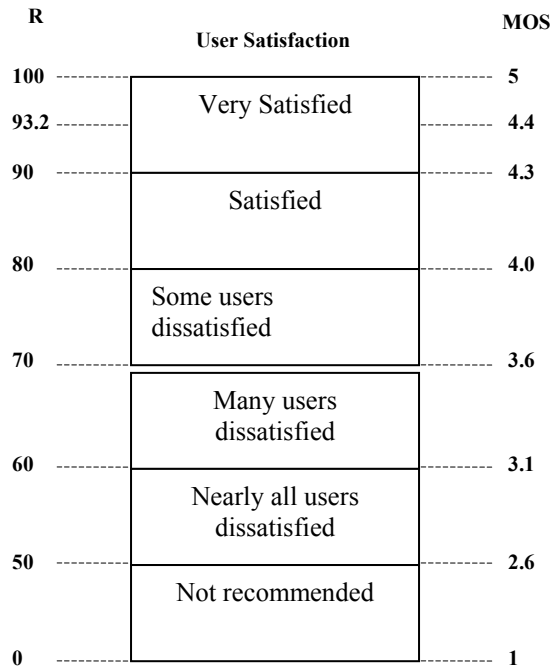


Fig. 1: Voice Quality classes

I_s are impairments simultaneous to voice signal transmission. It considers non-optimum sidetone, quantizing distortion, overall loudness and other impairments, which occur more or less simultaneously with the voice transmission.

I_d are impairments caused by delay after voice signal transmission. It is a mathematical summary of transmission delay, talker echo and sidetone.

I_{e-eff} are effective equipment impairments (e.g. due to codecs). This value depends on the quality of reproduced speech as rated by a user and is usually calculated from the MOS assigned to a particular compression algorithm.

Finally, A is the advantage factor given to a particular voice signal. The user expects some level of degradation in the speech signal and psychologically anticipates that the quality will not be as good as that of the PSTN for different links such as mobile connections, etc. The advantage factor compensates for this.

The default values of all the aforementioned parameters (using the year 2000 revision Annex A of [5]) result in $R = 93.2$. The E-model is a widely used computational model and several enhancements have been proposed to the basic model to cater for VoIP traffic [8]. Cole and Rosenbluth [6] outline a methodology for incorporating packet loss and delay variations as a reduction in base transport metrics through the equipment impairment factor. Basically, the end-to-end delay variations result in loss at the dejitter buffer due to the arriving packet stream underflowing or overflowing the decoder’s dejitter buffer and this is incorporated by adjusting I_{e-eff} appropriately. Other components that contribute to a further reduction in QoS are the end-to-end delay as well as the packet loss.

III. EXPERIMENTAL SET-UP

In order to perform an assessment of the perceptual quality of VoIP traffic on access links, an experimental set-up using a simple delay jitter measurement methodology was employed for a number of access links as depicted in Fig 2.

Three sets of experiments were performed:

1. Cable-Modem to Distribution Router-1 LER through ISP-1
2. Dial-up Modem to Distribution Router-1 LER through ISP-2
3. Cable Modem to Dial-up Modem between ISP-1 and ISP-2.

Voice sessions were established between soft phone clients using the Session Initiation Protocol (SIP). Microsoft MSN Messenger ver 7.0. was selected as the soft phone client due to its low average mouth-to-ear M2E delay in comparison with other IP phones or soft client phones [9].

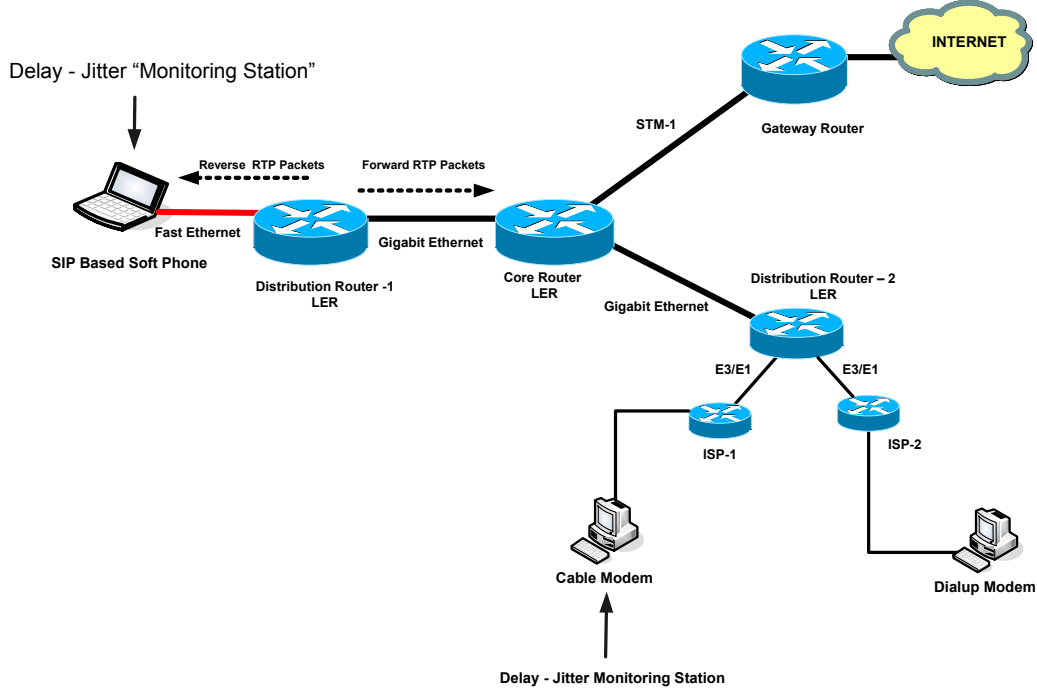


Fig. 2: VoIP delay jitter measurement set-up

The first two sets of experiments entailed establishing calls from an Intel 1.5 GHz Centrino processor based laptop connected through a fast Ethernet port of a Cisco 7500 series Distribution Router-1 LER placed at the Pakistan Internet Exchange node at Lahore to other machines with soft phone clients connected to access networks of ISPs via 64 kbps cable-modems as well as 56 kbps dial-up modems as shown in Fig. 2. While, the third set of experiments was between soft phone clients connected to two different ISPs. It is important to note here that the traffic characteristics presented to the VoIP soft phone applications are a snapshot of the instantaneous traffic conditions presented to the VoIP streams. Nevertheless, the readings shown are representative of the large degree of variation in the loads and jitter present in the access network. The voice streams were captured using an open source packet capturing application (Ethereal Ver. 0.10.10) installed on a delay-jitter monitoring station.

The VoIP stream from the soft phone is encoded as RTP/UDP/IP packets and is transmitted as the forward RTP stream while the captured voice streams from the 'far' side is the reverse RTP stream. Jitter calculations are carried out in accordance with RFC 3550 [10] for both directions. If S_i and S_j are the sender's and R_i and R_j the receiver's RTP timestamps for packet i and j , respectively. The Inter packet spacing D may be expressed as:

$$D(i,j) = (R_j - R_i) - (S_j - S_i) = (R_j - S_j) - (R_i - S_i) \quad (2)$$

This inter-packet spacing gives a clear illustration of the end-system view of packet delay variation. The inter arrival

jitter can be calculated successively for each packet received using D according to the formula [10]:

$$J(i) = J(i-1) + (|D(i-1,i)| - J(i-1))/16 \quad (3)$$

The codec used in the 1st set of experiments was G.729a. Each VoIP frame had a link-layer frame size of 94 bytes. 54 of these bytes make-up the RTP/UDP/IP/Ethernet headers and the remaining 40 bytes were the payload containing compressed speech. The 2nd and 3rd sets of experiments were performed using the G.723.1 codec. Each VoIP frame was 78 bytes in length and once again 54 bytes were header overhead, while the remaining 24 bytes were compressed speech.

The jitter was calculated using (3) in the 1st set of experiments (cable modem to distribution router-1 LER) for two link loads, i.e. 47.8% and 97.6%. An increase in the link load clearly manifests itself in greater jitter as shown in Fig. 3. The corresponding empirical probability density functions (pdf) of the inter-packet spacing for the cable modem to distribution router-1 LER through ISP-1 are shown in Fig. 4. The mean inter-packet spacing is approximately 20ms while the standard-deviation of the inter-packet spacing increases from 1.455ms to 6.161ms when ISP-1's link utilization increases from 47.8% to 97.6%. For comparison the pdf of the 2nd set of experiments (dial-up modem to distribution router-1 LER through ISP-2) is also shown on the same graph. The pdf of the dial-up modem connection has a variance of 12.33ms and a mean of approximately 30ms. Fig. 5 shows the one-way inter-packet spacing D calculated using (2) between a dial-up and cable modem connection for experiment 3.

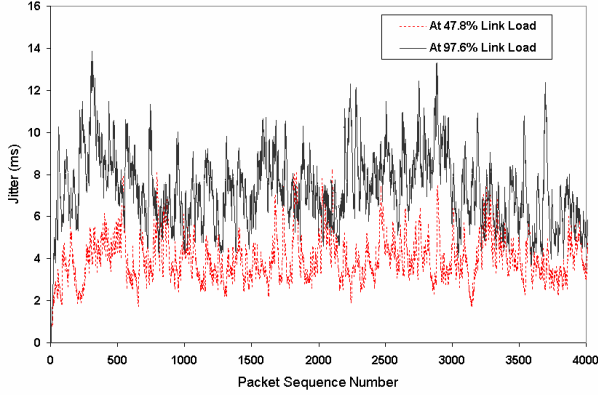


Fig. 3: Jitter between a cable modem and distribution router

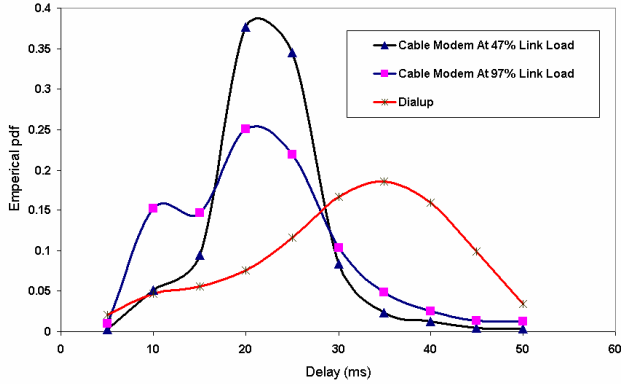


Fig. 4: Empirical probability density function of the inter-packet spacing

Fig. 6 shows the corresponding empirical pdf of inter-packet spacing. This is a bimodal distribution and clearly indicates that a significant number of packets get severely delayed as evident from the second hump while a delayed packet followed by a packet on time results in clumping of packets (apparent from the first hump). The mean inter-packet spacing is 30ms. However, the standard deviation of the inter-packet spacing is 49.31ms. Further, 2.57% packets are lost as a consequence of network congestion.

The R -factor and resultant MOS of voice calls for each set of experiments can be calculated using the E-model. The R factor can be estimated by calculating individual terms in the RHS of (1). The main terms are the effective equipment impairment factor I_{e-eff} and I_d delay impairment factor while the other parameters can be assumed to be a fixed for a given experimental set-up:

$$I_{e-eff} = I_e + (95 - I_e) \cdot \frac{P_{pl}}{P_{pl} + B_{pl}} \quad (4)$$

where I_e is the equipment impairment factor. P_{pl} is the packet-loss percentage and B_{pl} is the packet-loss robustness factor.

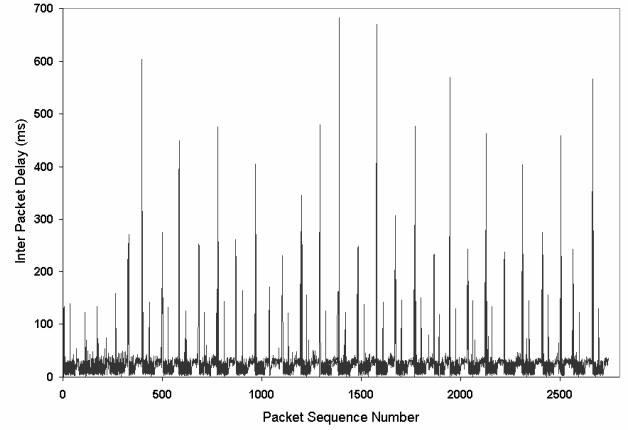


Fig. 5: One-way inter-packet delay between a dial-up and cable modem user.

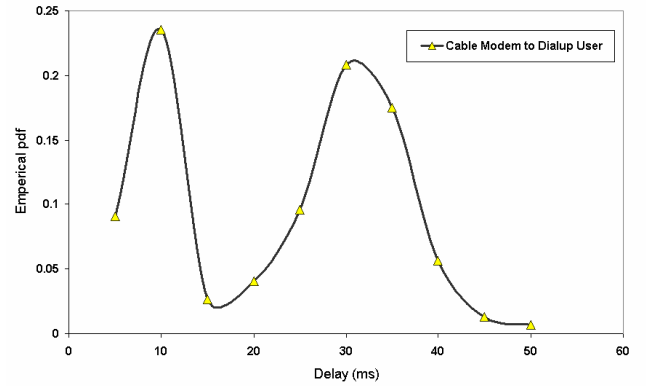


Fig. 6: Empirical probability density function of the inter-packet spacing

These values are codec dependent and have been suggested in ITU-T G.107 [5] and G.113 Appendix I [10].

Using a line of argument similar to the one presented by Cole and Rosenbluth [6], one can calculate a conservative bound on the probability of packet loss at the dejitter buffer as a result of jitter $P_{dejitter_buffer}$ by using Chebyshev's inequality. The dejitter buffer size is assumed to have a capacity of three packets. The total loss P is then given by:

$$P = P_{pl} + (1 - P_{pl}) \cdot P_{dejitter_buffer} \quad (5)$$

P would replace the packet loss P_{pl} in (4) to incorporate the effect loss due to jitter. This would result in an effective equipment factor I_{e-eff} incorporating the effects of VoIP packet loss as a consequence of jitter. The other main parameter affecting R is I_d . However, I_d is not a significant component of (1) as long as the end-to-end delay is within the 175ms bound [12]. Beyond this voice quality degrades rapidly. Factors contributing to I_d are the algorithmic delay as well as network latency (sum of propagation and transmission delays) and the decoding delay. The R values can now be calculated using (1) and the corresponding MOS scores can be obtained through expression (6).

TABLE I
Perceptual Quality Assessment for VoIP for Access Networks

Experiment	Codec	Link	I_e	P_{pl}	B_{pl}	I_{e-eff}	μ ms	σ ms	P %	I_{e-eff}^*	R	MOS
1	G.729a	48% (ISP-1)	11	0	19	11	20	1.46	0.133	11.58	81.62	4.1
		97% (ISP-1)	11	0	19	11	20	6.16	2.338	20.32	72.88	3.75
2	G.723.1	dial up (ISP-2)	15	0	16.1	15	30	12.33	4.223	31.62	61.58	3.18
3	G.723.1	cable to dial-up (ISP-1 to ISP-2)	15	2.57	16.1	26.01	30	49.31	68.375	79.75	13.45	1.06

* includes the effects of loss due to jitter

Table I shows the corresponding values for the different scenarios described in the experiments.

For $R < 0$: MOS = 1

For $R > 100$: MOS = 4.5

For $0 < R < 100$:

$$\text{MOS} = 1 + 0.035R + 7.5 \times 10^{-6} R (R-60)(100-R) \quad (6)$$

For the first set of experiments:

$I_e = 11$, $B_{pl} = 19$ for G.729a and consequently $I_{e-eff} \approx 11$

as there is no packet loss in the network. Incorporating the effects of jitter require that the probability of packet loss due to jitter is calculated. A bound for this loss can be calculated using Chebyshev's inequality assuming a buffer that can hold three packets. This results in a modified effective equipment impairment factor I_{e-eff} which is then used to calculate the R-factor and consequently the MOS values. The MOS values calculated of the first set of experiments were 4.1 and 3.75 for a load of 48% and 97% respectively. By and large this represents a satisfactory perceptual quality.

For the second and third set of experiments, the G.723.1 codec was employed. The corresponding values of $I_e = 15$, $B_{pl} = 16.1$ and $I_{e-eff} = 15$ when there is no packet loss in experiment 2 while $I_{e-eff} = 26$ when $P_{pl} = 2.57\%$ in experiment 3. The algorithmic delay for G.723.1 is 37.5ms but this still results in the end-to-end delay remaining within the 175ms bound. This ultimately results in a transmission quality factor $R \approx 62$ for experiment 2 and consequently an MOS of approximately 3.2 suggesting that most, to nearly all users are dissatisfied. For experiment 3, the resultant value of MOS implies that all users are dissatisfied as can be seen from Table 1.

Experiments 1 and 2 have been conducted from access networks to the core and typically do not represent a realistic scenario for VoIP calls. In reality, most calls would originate from end-user equipment on one ISP and terminate at end user equipment in the same or other ISPs.

IV. CONCLUSIONS

In conclusion, the quality of access networks is not adequate to support VoIP services based on our measurements of the delay and jitter prevalent in access networks in Pakistan. The method employed for carrying out the perceptual quality assessment was a modified E-model method catering for effective equipment impairments as a consequence delay, loss and jitter over packet switched networks for G.729a and G.723.1 codecs. In order to provide a reasonable QoS, ISPs will need to increase bandwidth in their access networks or deploy QoS architectures such as Differentiated services to provide preferential treatment and thus a reasonable QoS for voice traffic.

REFERENCES

- [1] "Methods for subjective determination of transmission quality". ITU-T Rec. P. 800, Aug. 1996.
- [2] Rix, A. W., "Perceptual speech quality assessment - a review." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, (3), pp.1056-1059, May 2004.
- [3] "Objective quality measurement of telephone-band (300-3400 Hz) speech codecs." ITU-T Rec. P. 861, Feb. 1998.
- [4] "Perceptual Evaluation of Speech Quality (PESQ): An Objective end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." ITU-T Rec. P. 862, Feb. 2001.
- [5] "The E-Model, a computational model for use in transmission planning." ITU-T Rec. G.107, March 2003.
- [6] R. G. Cole, J. H. Rosenbluth, "Voice over IP performance monitoring." *ACM SIGCOMM Computer Communication Review*, vol.31 no.2, April 2001
- [7] "Speech Communication Quality from mouth to ear for 3.1 kHz Handset Telephony across Networks." ETSI ETR 250, July 1996
- [8] L. Ding, R. A. Goubran, "Speech quality prediction in VoIP using the extended E-model", *GLOBECOM 2003*, vol. 22, No. 1, Dec 2003
- [9] W. Jiang, K. Koguchi and H. Schulzrinne, "QoS Evaluation of VoIP End-points." *IEEE International Conference on Communications (ICC)*, Anchorage, Alaska, May 2003.
- [10] H. Schulzrinne, et. al., "RTP: A Transport Protocol for Real-Time Applications", RFC 3550, July 2003.
- [11] "Transmission impairments due to speech processing: Appendix I: Provisional planning values for the equipment impairment factor I_e and packet-loss robustness factor B_{pl} ", ITU-T Rec. G.113 Appendix I, May 2002.
- [12] "One-way transmission time", ITU-T Rec. G.114, May 2003.