Speech Quality Prediction in VoIP Using the Extended E-Model

Lijing Ding and Rafik A. Goubran
Department of Systems and Computer Engineering, Carleton University
1125 Colonel By Drive, Ottawa, ON, K1S 5B6, Canada
{lding, goubran}@sce.carleton.ca

***Abstract*-This paper investigates the effects of packet loss and delay jitter on speech quality in voice over Internet protocol (VoIP) scenarios. A new formula is proposed to quantify these effects and incorporated into ITU-T G.107, the E-model. In the simulation, codecs ITU-T G.723.1 and G.729 are used; random packet loss and Pareto distributed network delay are introduced. The prediction errors range between – 0.20 and + 0.12 MOS. The formula extends the coverage of the current E-model, and is very useful in MOS prediction as well as network planning.**

I. INTRODUCTION

Today, voice over Internet protocol (VoIP) has emerged as an important application and is expected to carry more and more voice traffic. However, the present Internet only offers *best-effort* service due to its nature; speech quality is mainly impaired by packet loss, delay and delay jitter.

Speech quality is determined by the listener's perception, and hence it is inherently subjective. The Mean Opinion Score (MOS) test is widely accepted as a norm for speech quality rating. However, the subjective MOS test is time-consuming and expensive. In recent years, several objective MOS measures were developed, such as Perceptual Analysis Measurement System (PAMS) and Perceptual Evaluation of Speech Quality (PESQ). They measure the audible distortions based on the perceptual domain representation of two signals, namely, a reference signal and a degraded signal which is the output of the system under test. On the other hand, ITU-T G.107 [1] defines the E-model, a computational model combining all the impairment parameters into a total value. The E-model is not a measurement tool, but an end-to-end transmission planning tool; the output can be transformed into a MOS scale for prediction.

In the current E-model, the impairment from packet loss is represented by *Ie*, the equipment impairment factor. The *Ie* values are tabulated in ITU-T G.113 appendix I [2], for limited testing conditions in terms of packet loss rates, error concealment methods and number of frames per packet. They are provisional only, as they were determined in single or a few tests. In addition, the E-model does not take into account impairment from delay jitter.

In this paper, we investigate the effects of packet loss and delay jitter on speech quality. Codecs G.723.1 and G.729 are used. Several error concealment methods, the Pareto distributed network delay, and a fixed buffer policy are simulated. A new parameter is added to represent the impairment from delay jitter. A new formula is proposed to quantify all these effects and finally incorporated into the E-model.

The rest of the paper is organized as follows: Section II reviews the E-model. Section III describes the simulation system design and measurement methods. Section IV presents the simulation results and our proposed formula. Finally, Section V concludes our work.

II. THE E-MODEL REVIEW

The E-model assesses the combined effects of varying transmission parameters that affect the conversation quality of narrow band telephony [1]. The principle of the E-model is based on the assumptions that transmission impairments can be transformed into psychological factors and psychological factors on the psychological scale are additive. The primary output of the E-model is a transmission rating factor *R*:

$$R = Ro - Is - Id - Ie + A \tag{1}$$

where *Ro* represents the basic signal-to-noise ratio, *Is* represents the impairments occurring simultaneously with the voice signal, *Id* represents the impairments caused by delay, and *Ie* represents the impairments caused by low bit rate codecs. The advantage factor *A* can be used for compensation when there are other advantages of access to the user. *R* can be transformed into a MOS scale by:

$$MOS = \begin{cases} 1 & R < 0 \\ 1 + 0.035R + R(R-60)(100-R)\cdot 7\cdot 10^{-6} & 0 < R < 100 \\ 4.5 & R > 100\ . \end{cases} \tag{2}$$

III. SIMULATION DESIGN

The simulation block diagram is shown in Fig. 1 and explained below.

A. Reference Signal and Codec Selections

Two sets of clean speech samples were used as the reference signals. In each set, speech samples were chosen from two male and two female English speakers, and stored in 16-bit, 8000 Hz linear PCM format, roughly 8 seconds in duration with 50% of active speech intervals. Specifically, set 1 contained 20 samples and was used for deriving the formula

0-7803-7974-8/03/$17.00 © 2003 IEEE

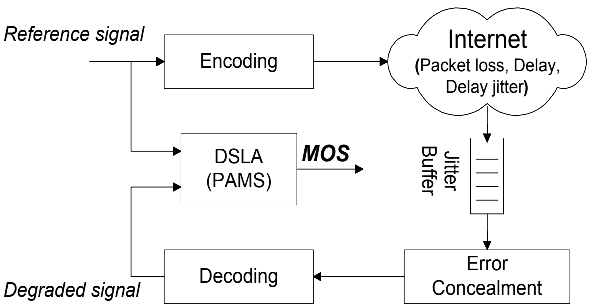

Fig. 1. Simulation block diagram

we proposed in the next section; set 2 contained another 5 new samples, arbitrarily selected from other sources, for use in validation tests.

Codecs G.723.1.B and G.729 were used. The ANSI C codes were obtained from ITU-T. The former is a floating-point dual-rate algorithm with the high rate 6.3 kb/s and low rate 5.3 kb/s; each frame is 30 ms. The latter is a fixed-point algorithm with rate 8.0 kb/s; each frame is 10 ms.

B. MOS Measurement

The MOS value was measured by PAMS due to the availability. PAMS is an objective measurement algorithm designed for robust end-to-end speech quality assessment [3]. A tool called Digital Speech Level Analyzer (DSLA) [4] was used in MOS measurement. DSLA is manufactured by Malden Electronics Ltd., UK, and includes the PAMS algorithm.

In the paper, each reported MOS value was the average of 10 repeated tests under the same conditions; the standard deviation was kept within 0.13 MOS.

C. Impairments Simulation

The simulation codes were written in MATLAB. In simulating packet loss, one frame per packet and random packet loss were assumed. In simulating delay jitter, Pareto distributed network delay was assumed.

It is now widely accepted that Internet traffic is self-similar [5][6]. The degree of self-similarity can be expressed by the Hurst parameter H ($0.5<H<1$); this degree increases when H increases. The Pareto distribution is a suitable model for such traffic [7] with the probability distribution function:

$$f(t)=\frac{\beta a^{\beta}}{t^{\beta+1}} \quad with\ \alpha>0,\beta>0\ and\ t\geq\alpha \tag{3}$$

where α is the location parameter and β is the shape parameter. β is associated with H by: $\beta = 3-2H$ [7].

The delay jitter is calculated as the difference between the inter-departure and interarrival times of two consecutive arrival packets. To obtain a steady output stream, a jitter buffer is used at the receive side. The received packet is held for a while before being played out. This amount of holding time is the measure of jitter buffer size. The buffer size can be fixed or adaptively adjusted during a call. To examine the relationships between MOS and buffer size, we considered a fixed buffer policy in [8], E-policy, which expands the playout time in order to preserve information by inserting pauses into the output stream.

IV. RESULTS

Packet loss and delay jitter were introduced separately at first, and then they were introduced together in the simulation. The simulation and formulation results were given in this section. The proposed formula was validated and showed good accuracy.

A. Effects of Packet Loss on Speech Quality

The effects of replacing the lost packet by nothing (splicing), silence and previous packet (repetition) were examined. The results for G.723.1.B 6.3kb/s with packet loss up to 10% are shown in Fig. 2. Speech quality drops with increasing packet losses. Compared with splicing and silence substitution, the repetition method performs best. We further introduced the packet loss up to 20% for the repetition method. The results for G.729 and G.723.1.B are shown in Fig. 3. Codec G.729 gives the highest MOS, G.723.1.B 5.3 kb/s gives the lowest MOS and G.723.1.B 6.3 kb/s performs in between.

Some former researches [9][10] examined the relationships between packet loss rate and MOS, a typical formula was suggested in [10]:

$$predicted\ MOS = MOS\_opt - Cln(1+loss) \tag{4}$$

where *MOS_opt* is the optimal MOS value without impairment, and C is a constant factor which differs with codecs or error concealment methods. However, for a given testing condition, C generally depends on the speech samples used and is not fixed.

In the E-model, the impairment from packet loss is represented by a fixed *Ie* for a given condition. So the speech quality can be predicted by direct calculation rather than doing real measurement. *Ie* is independent of the speech samples. Currently, only limited, inflexible and provisional *Ie* values are available.

To extend the E-model to cover more packet loss rates, error concealment methods and packet sizes, the *Ie* values are derived by the following two steps [11]:

1. Scale transformation. The measured MOS is transformed into the equipment impairment factor scale by (2) and (1) in turn, with all the other parameters set to their default values.
2. Linear interpolation. The *Ie* value from step 1, denoted by *Ie,mea*, does not necessarily equal that in ITU-T standard, denoted by *Ie,std*. According to the reference conditions in [11], a linear interpolation line:

$$Ie,std = a \cdot Ie,mea + b \tag{5}$$

is used to calibrate *Ie,mea* to *Ie,std*. In our case, for

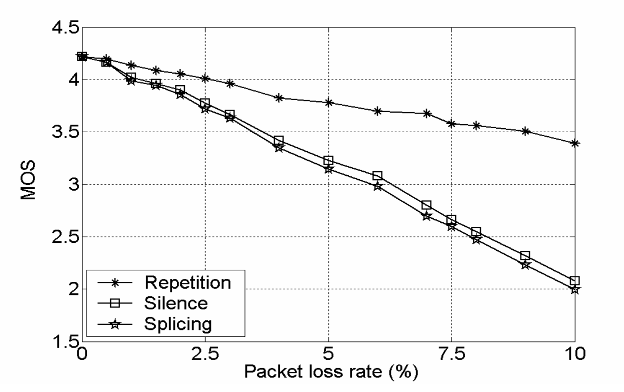

Fig. 2. MOS vs. packet loss for G.723.1.B 6.3 kb/s under different error concealment methods

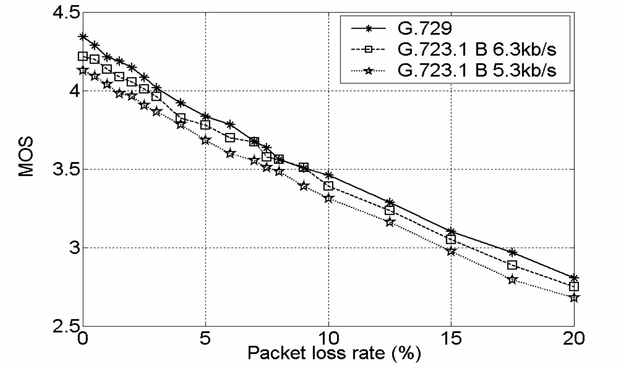

Fig. 3. MOS vs. packet loss for G.729 and G.723.1.B with replacing the lost packet by repetition method

speech set 1, *a* and *b* are found to be 0.8532 and 8.2640 respectively. After calibration, a stable *Ie* value, consistent with the ITU-T standard, is obtained.

We modeled *Ie* increased logarithmically with packet loss rate:

$$Ie = Ie\_opt + C1 \cdot ln(1 + C2 \cdot loss\_rate) \qquad (6)$$

where *Ie_opt* is the optimum (without packet loss) *Ie* from [2], *loss_rate* is the amount of packet loss in percent, and factors *C1, C2* are constants used to adjust the shape of the curve.

The fitting of our proposed model was examined by the correlation coefficient ρ and root mean square error σ. The results for the repetition method with up to 20% packet loss are summarized in Table I.

The model can also be applied to other cases by using different sets of *C1* and *C2* in (6). For the silence substitution concealment and ITU-T suggested *Ie* values in [2], the results are summarized in Table II.

B. Effects of Delay Jitter on Speech Quality

The effects of delay jitter on speech quality were examined by using different delay distribution parameters and jitter buffer sizes. In the simulation, β of the Pareto distribution was selected from 1.2 to 1.9 at intervals of 0.1, which was equivalent to *H* ranging from 0.55 to 0.90; the fixed buffer size *T* was selected from 30 to 100 ms at intervals of 10 ms. The results for G.723.1.B 5.3 kb/s are shown in Fig. 4 as an example. MOS drops when *H* increases or *T* decreases.

We added a new parameter, *Ij*, jitter impairment factor, into the E-model to quantify its effects on speech quality:

$$R = Ro - Is - Id - Ie - Ij + A \qquad (7)$$

Ij is a function of *H* and *T*. To derive stable *Ij* values, the same steps as in subsection A were applied here as well. We modeled *Ij* increased parabolically with *H*, and decayed exponentially with *T*:

$$Ij = C1 \cdot H^2 + C2 \cdot H + C3 + C4 \cdot e^{-T/K} \qquad (8)$$

where *C1, C2, C3, C4* are coefficients and *K* is a time constant. The results are summarized in Table III.

C. Combined Effects on Speech Quality

The combined effects of packet loss and delay jitter on speech quality were examined by introducing them jointly. In the simulation, *H* was selected to be 0.60, 0.75 and 0.90, and *T* was selected to be 50, 70 and 90 ms respectively. The packet loss rate was up to 20%, and the lost packets were concealed by using the repetition method. The results for G.723.1.B 5.3kb/s with *H* = 0.75 are shown in Fig. 5. The results for the rest are similar. MOS drops with increasing packet losses or increasing *H* or decreasing *T*.

TABLE I
FACTORS *C1* AND *C2* - REPETITION

Codec	*Ie_opt*	*C1*	*C2*	ρ	σ
G.723.1.B-5.3[1]	19	37.40	0.05	0.9989	1.3080
G.723.1.B-6.3[1]	15	36.59	0.06	0.9986	0.9286
G.729[1]	10	25.05	0.13	0.9987	0.9548

1. One frame/packet, repetition method, up to 20% packet loss.

TABLE II
FACTORS *C1* AND *C2*

Codec	*Ie_opt*	*C1*	*C2*	ρ	σ
G.723.1.B-5.3[1]	19	71.38	0.06	0.9988	1.6581
G.723.1.B-6.3[1]	15	90.00	0.05	0.9983	0.9164
G.729[1]	10	47.82	0.18	0.9997	0.3682
G.723.1.A+VAD-6.3[2]	15	30.50	0.17	0.9991	0.5396
G.729A+VAD[3]	11	30.00	0.16	0.9998	0.2945

1. One frame/packet, silence insertion, up to 10% packet loss.
2. One frame/packet, no concealment, up to 16% packet loss.
3. Two frames/packet, no concealment, up to 16% packet loss.

TABLE III
COEFFICIENTS AND TIME CONSTANT

Codec	*C1*	*C2*	*C3*	*C4*	*K*	ρ	σ
G.723.1.B-5.3	-8.3	22.3	-1.1	9.0	40	0.9350	0.6937
G.723.1.B-6.3	-23.7	45.4	-6.8	9.7	36	0.9478	0.7799
G.729	-15.5	33.5	4.4	13.6	30	0.9499	0.7836

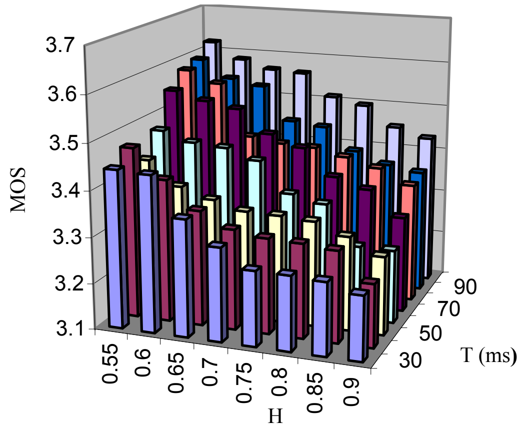

Fig. 4. MOS vs. *H* and *T* for G.723.1.B 5.3kb/s

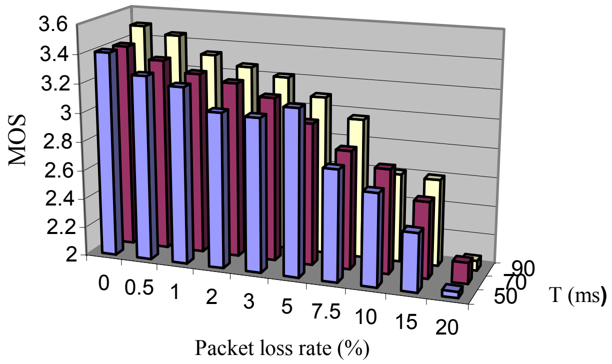

Fig. 5. Combined effects for G.723.1.B 5.3kb/s with *H* = 0.75

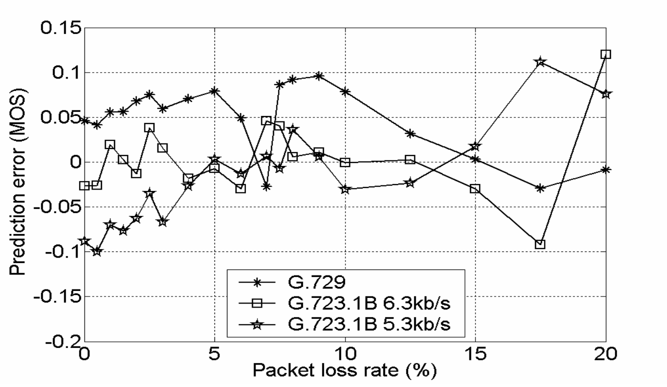

Fig. 6. MOS prediction errors for packet loss

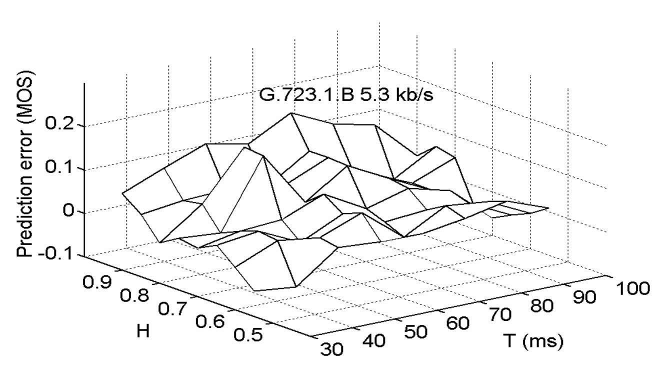

Fig. 7. MOS prediction error surface for G.723.1.B 5.3kb/s

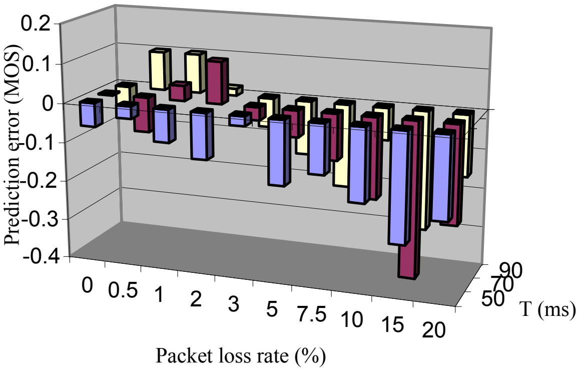

Fig. 8. MOS prediction errors for G.723.1.B 5.3kb/s, *H* = 0.75

We assumed the derived *Ie* and *Ij* were additive in the extended E-model formula. Thus, MOS can be predicted by (7) and (2) in turn, where *Ie* and *Ij* were given by (6) and (8) respectively.

D. Validation Tests

Validation tests were run to determine the accuracy of our proposed formula in MOS prediction. Speech set 2 was used in validation tests under the same testing conditions.

For the effects of packet loss, the prediction errors for G.729 and G.723.1 are shown in Fig. 6. The errors range between ± 0.10 MOS for most cases; the maximum error is 0.12 MOS. For the effects of delay jitter, the prediction errors for G.723.1.B 5.3kb/s are shown in Fig. 7. For all three codecs, the errors mostly range between ± 0.10 MOS as well; the maximum error is 0.18 MOS. For the combined effects, the prediction errors for G.723.1.B 5.3kb/s with *H* = 0.75 are shown in Fig. 8. For all three codecs, the errors range between – 0.20 and + 0.10 MOS when packet loss rate is below 10%; they become pronounced (up to – 0.40 MOS) when packet loss rate is above 10%.

For the combined effects, when a jitter buffer was used, the impairment from delay jitter converted into the impairment from packet loss. However, the effects of packet loss did not satisfy additivity. In fact, MOS dropped at a decreasing rate with increasing packet loss, which can be substantiated by Fig. 3 or *Ie* values in [2]. Assuming the additivity of derived *Ie* and *Ij* in the extended E-model would underestimate the real speech quality to some extent. This was the reason why the prediction errors were negative in general if the packet loss rate was high.

V. Conclusions

This paper has extended the E-model in speech quality prediction in VoIP scenarios. The impairment from packet loss is modeled by a logarithmic function; the impairment from delay jitter is modeled as the sum of a parabolic function to reflect the impact of Internet traffic self-similarity, and an exponential function to reflect the impact of buffer size. Good accuracy is achieved by our extended E-model formula, especially for the separated impairments; the prediction errors lie in between ± 0.10 MOS for most cases. For the combined impairments, the formula still gives good prediction when the packet loss rate is below 10%, the errors range between – 0.20 and + 0.10 MOS.

Future work will focus on developing a joint model to effectively represent the combined effects of packet loss and delay jitter by examining the cross correlations between these two factors. Also, more testing conditions, such as using other error concealment methods (e.g. built-in and interpolative methods), other codecs (e.g. G.711, G.722 and G.728), different packet sizes, adaptive buffer algorithms, and using the PESQ metric, will be investigated.

ACKNOWLEDGMENT

The authors would like to thank CITO, OGSST and NSERC for their financial supports.

REFERENCES

[1] ITU-T G.107, *The E-model, a computational model for use in transmission planning*, 2000.

[2] ITU-T G.113, *Transmission impairments due to speech processing*, 2001.

[3] A.W. Rix and M.P. Hollier, **“**The perceptual analysis measurement system for robust end-to-end speech quality assessment,” *Proc. of ICASSP’00*, vol. 3, pp. 1515-1518, 2000.

[4] *Digital Speech Level Analyzer, User Guide*, Revision 3.5, Malden Electronics Ltd., http://www.malden.co.uk/downloads/medslahlp.pdf, 2001.

[5] V. Paxson and S. Floyd, “Wide area traffic: the failure of Poisson modeling,” *IEEE/ACM Trans. on Networking*, vol. 3, no. 3, pp. 226-244, 1995.

[6] M. Borella, S. Uludag, G. Brewster and I. Sidhu, “Self-similarity of Internet packet delay,” *ICC’97*, vol. 1, pp. 513-517, 1997.

[7] J. Gordon, “Pareto process as a model of self-similar packet traffic,” *GLOBECOM’95*, vol. 3, pp. 2232 –2236, 1995.

[8] W. Naylor and L. Kleinrock, “Stream traffic communication in packet switched networks: destination buffering considerations,” *IEEE Trans. on Comm.*, vol. 30, no.12, pp. 2527-2534, 1982.

[9] L. Yamamoto and J.G. Beerends, “Impact of network performance parameters on the end-to-end perceived speech quality,” *Expert ATM Traffic Symposium*, Greece, 1997.

[10] B. Duysburgh, S. Vanhastel, B. Vreese, C. Petrisor and P. Demeester, “On the influence of best-effort network conditions on the perceived speech quality of VoIP connections,” *Proc. of ICCCN’01*, pp. 334-339, 2001.

[11] ITU-T P.833, *Methodology for derivation of equipment impairment factors from subjective listening-only tests*, 2001.