Assessment of Effects of Packet Loss on Speech Quality in VoIP

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 on speech quality in Voice over Internet Protocol (VoIP) applications by using ITU-T G.107, the E-model, whose parameters currently only cover limited VoIP scenarios. Several packet loss rates, packet sizes and error concealment techniques for codec G.729 are examined. Mean Opinion Score (MOS) is used as an index for speech quality and is measured by Perceptual Evaluation of Speech Quality (PESQ) algorithm. These effects on speech quality are assessed in the equipment impairment factor domain and then formulated into the E-model. The validation test shows good accuracy of the proposed formula, the prediction errors range between ±0.10 MOS for most cases with an absolute maximum of 0.14 MOS.

1. Introduction

Voice over Internet Protocol (VoIP), the transmission of packetized voice over IP networks, has gained much attention in recent years. It is expected to carry more and more voice traffic for its cost-effective service. However, the current Internet, which was originally designed for data communications, provides *best-effort* service only, posing several technical challenges for real time VoIP applications. Speech quality is mainly impaired by packet loss, delay and delay jitter. Assessment of perceived speech quality in the IP networks becomes an imperative task to manufacturers as well as service providers.

Speech quality is judged by human listeners and hence it is inherently subjective. The Mean Opinion Score (MOS) test, defined by ITU-T P.800 [1], is widely accepted as a norm for speech quality assessment. However, such subjective test is expensive and time-consuming. It is impractical for frequent testing such as routine network monitoring.

Objective test methods have been developed in recent years. They can be classified into two categories: signal-based methods and parameter-based methods. Signal-based methods use two signals as the input to the measurement, namely, a reference signal and the degraded signal, which is the output of the system under test. They identify the audible distortions based on the perceptual domain representation of two signals incorporating human auditory models. These methods include Perceptual Speech Quality Measure (PSQM), Measuring Normalizing Blocks (MNB), Perceptual Analysis Measurement System (PAMS), and Perceptual Evaluation of Speech Quality (PESQ). Among them, PSQM and PESQ were standardized by ITU-T as P.861 and P.862 respectively. Parameter-based methods predict the speech quality through a computation model instead of using real measurement. A typical model is the E-model, as defined by ITU-T G.107 [2]. The E-model includes a set of parameters characterizing the end-to-end voice transmission as its input, and the output can be transformed into a MOS scale for prediction.

In the E-model, the delay impairment factor *Id*, and the equipment impairment factor *Ie*, are used to represent the degradation on speech quality due to delay and packet loss in VoIP scenarios, as their names imply. The recommended *Ie* values are tabulated in ITU-T G.113 [3], for limited testing conditions. These values are provisional only, as they were determined in single or a few tests.

Some works have been carried out on the effects of packet loss on speech quality. Particularly, [4][5] examined these effects in the MOS domain for certain packet loss rates and packet sizes. In [4], a formula was suggested based on the subjective MOS test, where linear PCM, and random packet loss were used, and the lost packets were replaced by silence. It modeled that MOS drops logarithmically with increasing packet loss rate or packet size. In [5], several common speech coders, and random packet loss were used without error concealment; the same formula as in [4] was used to fit MOS measured by PAMS.

However, MOS usually varies from speech to speech under the same testing conditions. The coefficients of the formula suggested above would be different if they were derived from another set of samples. A speech sample independent formula is much preferred. Such work can be done by transforming the MOS scale into the *Ie* scale used in the E-model, and assigning a stable *Ie* value to each impairment condition. In [6], the effects of packet loss,

0-7803-8108-4/03/$17.00 © 2003 IEEE

with one frame per packet and delay jitter, on VoIP speech quality were examined in the *Ie* domain.

In this paper, we focus on the effects of packet size on speech quality. ITU-T G.729 [7], one of the prevalent coders in VoIP, is used in simulation, as it has a smaller frame size of 10ms compared with that of ITU-T G.723.1. Several frames of G.729 can be encapsulated into one packet in real VoIP applications with tolerable overall delay. Random packet loss is assumed and several error concealment techniques are examined. The effects are formulated in the *Ie* domain, and finally incorporated into the E-model, extending its capability of speech quality prediction in VoIP scenarios.

The rest of the paper is organized as follows: Section 2 reviews the E-model. Section 3 describes the simulation system design and measurement methods. Section 4 presents the simulation results and the proposed formula. Section 5 gives the model validation test results. Finally, Section 6 concludes the paper and suggests some future studies.

2. The E-model review

The E-model is a computational model for use in end-to-end transmission planning. It is defined in ITU-T G.107, and detailed guidelines and planning examples are given in ITU-T G.108. The E-model assesses the combined effects of transmission parameters that affect the conversation quality of narrow band telephony [2]. The parameters cover a wide range of impairment factors, such as handset acoustic characteristics, noise, delay, echo, quantization distortion, equipment impairment and so on. These factors are available at time of planning, either from the internationally accepted standards, network experience or from measurements. The primary output of the E-model is a transmission rating factor *R*, which can be transformed into other quality measures, such as MOS, Percentage Good or Better (%GoB) or Percentage Poor or Worse (%PoW) for prediction purposes. MOS is obtained from *R* by:

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

One important parameter of the E-model is the equipment impairment factor *Ie*, which represents the impairments caused by low bit rate codecs. Provisional *Ie* values for some codecs under conditions of packet loss are given in ITU-T G.113.

3. Simulation design

We investigate the effects of packet loss and packet size on speech quality. The simulation block diagram is shown in Figure 1.

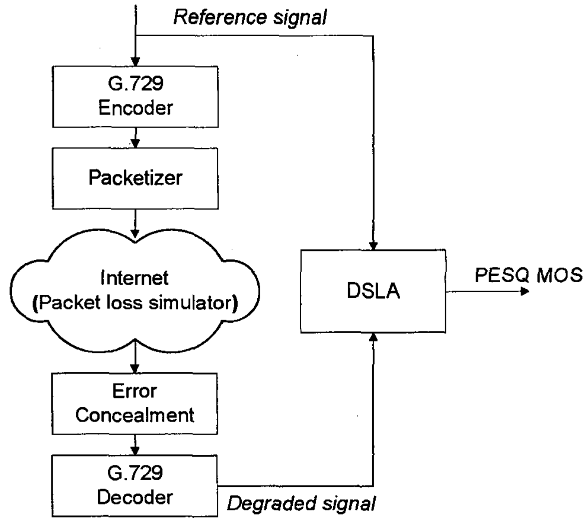

Figure 1. Simulation block diagram

The ANSI C source code for codec G.729 was obtained from ITU-T. The reference signal was encoded by G.729 and filled into packets. Then, random packet loss was simulated during transmission. At the receive side, an error concealment method was implemented to recover the missing packets before decoding. Finally, MOS was measured by PESQ.

In the simulation, one, two, three, four and five frames were encapsulated into a packet in turn, each corresponding to a packet size of 10, 20, 30, 40 and 50 ms. Packet loss was introduced successively from 0 to 20 percent. Totally, 19 non-equally spaced packet loss rates were tested, with more data points assigned to low packet loss scenarios. Moreover, three error concealment methods, namely, repetition, silence and built-in methods were considered. The former two replace the missing packet by previous correctly received packet or silence respectively; the latter regenerates a packet by triggering the internal erasure concealment algorithm in G.729.

3.1. Reference signal selection

The reference signal selection should follow the criteria given by ITU-T P.830 [8] and P.800 [1]. The reference speech sample should include bursts separated by silent periods, and bursts are normally 1–3 seconds in duration. Also, it should be active for 40-80% of the length.

In general, we selected two sets of reference signals for different purposes. Set 1 contained 20 samples and was used for deriving the formula we proposed in the next section; set 2 contained 5 samples, arbitrarily selected from other sources, and was used in validation tests. In each set, speech samples were chosen from two male and two female speakers, stored in 16-bit, 8000 Hz linear PCM format, and were roughly 8 seconds in duration with 50% of active speech intervals.

3.2. MOS measurement

MOS was measured by the PESQ metric, the most recent ITU-T standard for objective speech quality assessment. PESQ combines the merits of PAMS and PSQM99 (an updated version of PSQM), and adds new methods for transfer function equalization and averaging distortions over time. It can be used in a wider range of network conditions, and gives higher correlation with subjective tests than other objective algorithms [9][10]. In contrast to the conversational model, PESQ is a listening-only model; the degraded sample is time-aligned with the reference sample during preprocessing. The PESQ MOS values do not reflect the effects of delay on speech quality.

MOS measurement was conducted by a tool called Digital Speech Level Analyzer (DSLA) [11], which implements the PESQ algorithm. DSLA is manufactured by Malden Co. Ltd., and it includes a batch processor, which we used for automatically processing a large number of speech pairs without intervention.

4. Simulation results

The simulation was independently run 10 times under the same testing conditions. The MOS results were averaged out and the standard deviation was kept within 0.085 MOS.

Figures 2, 3 and 4 show the simulation results for repetition, built-in and silence concealment methods respectively. In general, speech quality drops with increasing packet loss rate or packet size. In Figure 3, the speech quality rendered from the 40 ms packet is almost identical to that of the 50 ms packet. This is probably caused by the gradual attenuation of adaptive and fixed-codebook gains in the built-in error concealment routine; after 4 consecutive frame erasures, the power of the regenerated frame is so small that it has little effect on speech quality. Also, compared with Figures 2 and 4, the curves in Figure 3 disperse wider each from the other (except the curves for 40 ms and 50 ms cases), suggesting more impact of frame size on speech quality for the built-in method. Similarly, frame size has less impact when the silence method is used, as shown in Figure 4.

Based on the above measurement results, a formula was proposed to quantify these effects on speech quality in the *Ie* domain. The block diagram is shown in Figure 5 and explained below.

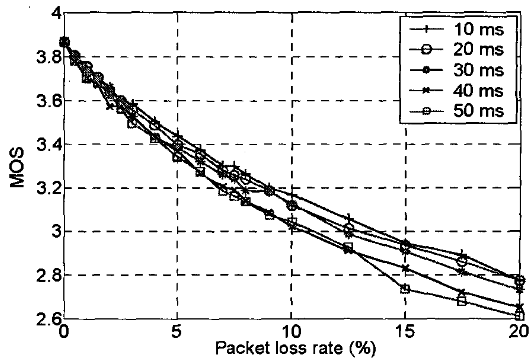

Figure 2. MOS for the repetition concealment

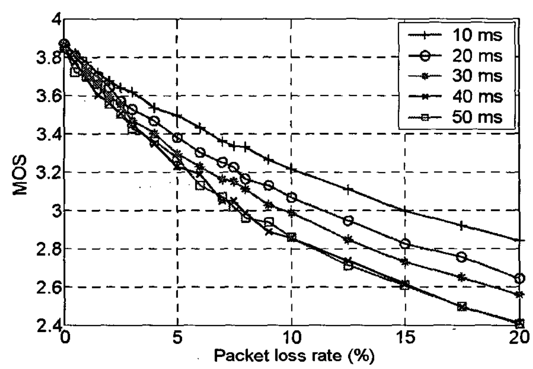

Figure 3. MOS for the built-in concealment

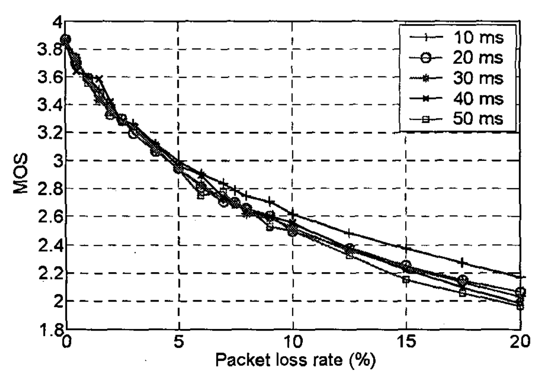

Figure 4. MOS for the silence concealment

The measured MOS value was transformed into the E-model rating factor R by taking the inverse of (1). Also, *Ie,mea*, denoting the *Ie* value from measurement, was derived from R with all the other parameters set to their default values. However, *Ie,mea* only reflects the impairment for this specific speech sample set. It is neither stable over different sample sets, nor always consistent with ITU-T recommended values in [3]. For example, *Ie* should be 10 for G.729 without packet loss, while our measured MOS value is 3.867 for this case, suggesting an *Ie* value of 17.128 instead.

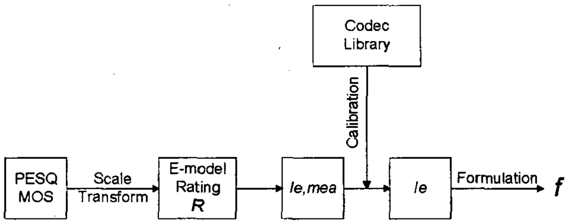

Figure 5. *Ie* formulation block diagram

A calibration stage was applied to correct this bias due to speech samples. The method is specified in ITU-T P.833 [12]. Speech samples are degraded by several common codecs, either independently or in tandem without packet loss, and their *Ie,mea* values are obtained. On the other hand, the expected *Ie* values, denoted by *Ie,exp*, for these reference conditions are available from [3]. A linear interpolation line can be made for pairs of *Ie,mea* and *Ie,exp*:

$$Ie,exp = a \cdot Ie,mea + b \qquad (2)$$

where coefficients a and b are found by the least square fitting. Then, (2) is applied to all the *Ie,mea* values. The resulting *Ie* values usually satisfy the framework of the E-model, and are considered to be stable, independent of speech samples.

In our case, a and b were found to be 1.4374 and 14.8239 respectively. Calibrated *Ie* was obtained by applying (2) to *Ie,mea*, and the results for the built-in method are shown in Figure 6, as an example. All the curves start from the same point at which packet loss rate is zero, and increase disproportionally with packet loss rate or frame size.

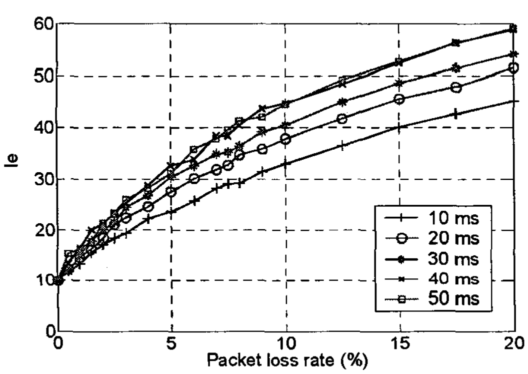

Figure 6. Calibrated *Ie* for the built-in concealment

In [6], when one frame per packet was studied, *Ie* was modeled to increase logarithmically with packet loss rate:

$$Ie = Ie\_opt + C1 \cdot ln(1 + C2 \cdot loss\_rate) \qquad (3)$$

where *Ie_opt* is the optimum (without packet loss) *Ie* from [3]; for codec G.729, this value is 10. *loss_rate* is the amount of packet loss in percent. Factors *C1* and *C2* are constants used to adjust the shape of the curve, and are found in the least square sense.

We adopted the same model here. For all the fifteen *Ie* curves (three error concealment methods, each with five packet sizes), factor *C1* was found to be: 21.54-28.66, and factor *C2* was found to be 0.15-0.56. The results show that *C1* is relatively stable over curves, while *C2* changes dynamically.

Mathematically, *C1* and *C2* can be respectively modeled by a function of packet size:

$$\begin{aligned} C1 &= f(no\_frame) \\ C2 &= g(no\_frame) \end{aligned} \qquad (4)$$

where *no_frame* is the number of frames per packet. However, this will make the overall expression of (3) complicated. To simplify, we fixed *C1* for each concealment method, as it was relatively unchanged, and found *C2* by the least square fitting as well. The simplification only increased the standard deviation of the prediction error a bit: an increase of 7, 6 and 25 percent for repetition, built-in and silence methods respectively. The results for factors *C1* and *C2* are given in Table 1.

Table 1. Factors *C1* and *C2*

Concealment	*no_frame*	*C1*	*C2*
	1	22.69	0.200
	2	22.69	0.211
Repetition	3	22.69	0.223
	4	22.69	0.257
	5	22.69	0.266
	1	25.21	0.150
	2	25.21	0.202
Built-in	3	25.21	0.238
	4	25.21	0.291
	5	25.21	0.291
	1	25.71	0.423
	2	25.71	0.491
Silence	3	25.71	0.493
	4	25.71	0.484
	5	25.71	0.517

Then, a third order polynomial function $g(no\_frame)$, given by:

$$g(x) = D_1x^3 + D_2x^2 + D_3x + D_4 \tag{5}$$

was used to fit the factor $C2$ in Table 1. The coefficients D_i (i = 1, 2, 3 and 4) are given in Table 2. Equation (5) is valid for up to 4 frames per packet for the built-in method, as $C2$s are identical for 40 ms and 50 ms cases. For the other two methods, (5) is valid for up to 5 frames per packet.

Table 2. Coefficients D_i of the polynomial $g(x)$

Concealment	D_1	D_2	D_3	D_4
Repetition	-0.0022	0.0208	-0.0410	0.2234
Built-in*	0.0055	-0.0410	0.1365	0.0490
Silence	0.0090	-0.0868	0.2652	0.2356

* Up to 4 frames per packet.

In summary, to accommodate the impacts of packet size on Ie, the overall formula (3) is modified to:

$$Ie = 10 + C1 \cdot \ln[1 + g(no\_frame) \cdot loss\_rate] \tag{6}$$

where $C1$ is given in Table 1, $g(no\_frame)$ is given by evaluating the polynomial, whose coefficients are specified in Table 2, at the point of no_frame. Equation (6) is valid for up to 20 percent of packet loss. The fitness of the proposed formula was examined by the standard deviation of the prediction error σ and the correlation coefficient ρ. The results are summarized in Table 3. An example of the curve fitting is shown in Figure 7, for the built-in method.

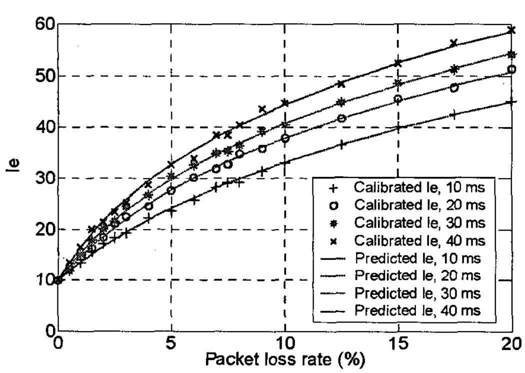

Figure 7. Curve fitting for the built-in concealment

Thus, for a given network condition, specifically, when packet loss rate, frame size and error concealment method are available, the Ie value for codec G.729 can be calculated by (6); R and MOS can be predicted in turn by the E-model.

Table 3. Fitness analysis

Concealment	ρ	σ
Repetition	0.9988	0.5600
Built-in*	0.9994	0.4553
Silence	0.9983	1.0177

* Up to 4 frames per packet.

5. Validation tests

To determine the accuracy of the proposed formula in MOS prediction, speech set 2 was used for validation purposes. Simulations were carried out under the same testing conditions and the prediction errors were calculated. It shows that, for all the three concealment methods, the prediction errors range between ±0.10 MOS for most cases; the absolute maximum error is 0.14 MOS. Figure 8 shows the prediction error distributions for the repetition method as an example.

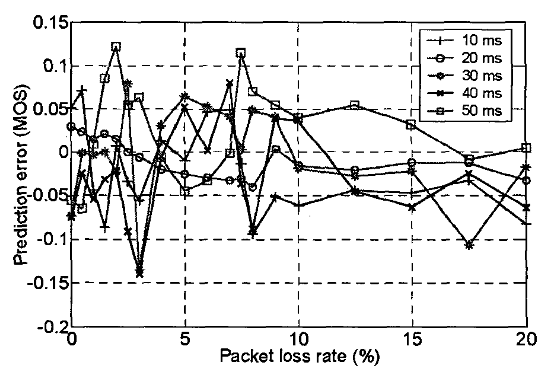

Figure 8. Prediction error for the repetition concealment

6. Concusions

This paper has investigated the effects of packet loss for codec G.729 on speech quality in VoIP applications. Random packet losses, different packet sizes, and several error concealment techniques are simulated. The results show that frame size has more impact on speech quality for the built-in method; MOS drops more quickly if a larger frame size is used. Less impact has been observed for the silence method. A formula is then proposed to quantify these effects in the Ie domain, and finally incorporated into the current E-model. MOS can be directly predicted from the formula for the given network conditions without doing real measurements. Very good prediction accuracy is achieved; the errors lie between ±0.10 MOS for most cases.

The real VoIP scenarios are much more complicated. To name a few, packet loss may be bursty, some techniques such as Voice Activity Detection (VAD) may be used, and transcoding may happen in the call path. Future work will focus on evaluating the impairments from other scenarios, such as those mentioned above. Also, more speech codecs (e.g. ITU-T G.711, G.722 and G.723.1) will be examined.

Acknowledgement

The authors wish to thank Communications and Information Technology Ontario (CITO), National Sciences and Engineering Research Council of Canada (NSERC), Ontario Graduate Scholarships in Science and Technology (OGSST) and Nortel Networks for their financial supports.

References

[1] ITU-T P.800, "Methods for subjective determination of transmission quality," 1996.

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

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

[4] L. Yamamoto and J.G. Beerends, "Impact of network performance parameters on the end-to-end perceived speech quality," in *Proc. Expert ATM Traffic Symposium*, Mykonos, Greece, 1997.

[5] 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," in *Proc. IEEE International Conference on Computer Communications and Networks*, Phoenix, USA, 2001, pp. 334-339.

[6] Lijing Ding and Rafik Goubran, "Speech quality prediction in VoIP using the extended E-model," to appear in *IEEE GLOBECOM*, San Francisco, USA, 2003.

[7] ITU-T G.729, "Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)," 1996.

[8] ITU-T P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," 1996.

[9] A.W. Rix, J.G. Beerends, M.P. Hollier and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Salt Lake City, USA, 2001, pp. 749-752.

[10] ITU-T 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," 2001.

[11] *Digital Speech Level Analyzer, User Guide*, Revision 4.0, Malden Electronics Ltd., Surrey, England, 2001, http://www.malden.co.uk/downloads/medslahlp.pdf.

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