Library: White Papers

Assessing VoIP Call Quality Using the E-model

Contents

Assessing VoIP Call Quality Using the E-model

Data networks haven't traditionally been reported on using a single metric, since there are many factors to consider. Yet, in the telephony world, a single number is typically given to rate call quality. Voice over IP (VoIP) is an example of a data network application that needs a single metric upon which to benchmark, trend, and tune. The E-model provides a powerful and repeatable way to assess whether a data network is ready to carry VoIP calls well. IxChariot employs the E-model to predict call quality in data networks. Its test results also show the underlying network attributes that influence the calculation.

1. Understanding Call Quality

Call quality testing has traditionally been subjective: picking up a telephone and listening to the quality of the voice. The leading subjective measurement of voice quality is the MOS (mean opinion score) as described in the ITU (International Telecommunications Union) recommendation P.800 [1]. However, asking people to listen to calls over and over can be difficult and expensive to set up and execute. Considerable progress has been made in establishing objective measurements of call quality.
Various standards have been developed:

PSQM (ITU P.861) / PSQM+ - Perceptual Speech Quality Measure

MNB (ITU P.861) - Measuring Normalized Blocks

PESQ (ITU P.862) - Perceptual Evaluation of Speech Quality

PAMS (British Telecom) - Perceptual Analysis Measurement System

The E-model (ITU G.107)

PSQM, PSQM+, MNB, and PESQ are part of a succession of algorithm modifications starting in ITU standard P.861 [2]. British Telecom developed PAMS, which is similar to PSQM. The PSQM and PAMS measurements send a reference signal through the telephony network and then compare the reference signal with the signal that's received on the other end of the network, by means of digital signal processing algorithms. Functional voice measurement tools such as Ixia's IxVoice have implemented PSQM, PAMS and PESQ measurements. These measurements are good in test labs for

analyzing the clarity of individual devices; for example, it makes sense to use PSQM to describe the quality of a PBX.

However, these approaches are not really well suited to assessing call quality on a data network, since they don't know about data networking. The models used are not based on data network issues, so they can't map back to the network issues of delay, jitter, and datagram loss.

MOS also comes from the telephony world. MOS is the widely accepted criterion for call quality, and the vendors that implement these scoring algorithms all map their scores to MOS. In using MOS with human listeners, a large number of people listen to audio and give their opinion of the call quality. This certainly works well, but you can guess it's pretty expensive to have a number of people standing around each time you make a tuning adjustment. The good news is that the human behavioral patterns have been heavily researched and captured. The ITU P.800 standard describes how humans react - what score they would give - as they hear audio with different aspects of delay or datagram loss. This mapping between network characteristics and a quality score makes MOS valuable for doing network assessments and tuning.

ITU recommendation G.107 [3,4] introduced the E-model. The output of an E-model calculation is a single scalar, called an "R factor," derived from delays and equipment impairment factors. Once an R factor is obtained, it can be mapped to an estimated MOS.

IxChariot [5] uses a modified form of the Emodel in its voice-readiness testing. It calculates an R factor and converts that to an estimated MOS. IxChariot tests work by generating real-time transport protocol (RTP) streams that mimic VoIP traffic. The RTP traffic flows between two endpoints in a data network. Each time a test is run, measurements are collected for the one-way delay time, the number of datagrams lost, the number of consecutive datagrams lost, and the amount of variability in the arrival time of the datagrams (known as jitter). These measurements capture what's important for voice quality: how the two people at the two telephones perceive the quality of their conversation.

2. Calculating an R Factor

The R factor calculated by the E-model ranges from 100 down to 0, where 100 is excellent and 0 is poor. The calculation of an R factor starts with the unadulterated signal. If there's no network and no equipment, quality is perfect. In equation form, we say:

$$R = R_0$$

But, the network and the equipment impair the signal, reducing its quality as it travels from end to end:

$$R = R_0 - I_d - I_e$$

where:

I_d: delays introduced from end-to-end

I_e: impairment introduced by the equipment

Three data network measurements influence these delay and equipment impairments: one-way delay, jitter, and lost data. They're also influenced by the codec, which has an implicit delay and impairment function.

The E-model calculation in IxChariot considers the percentage of datagram loss, datagram loss burstiness (calculated from maximum consecutive datagram loss), the delay introduced by the jitter buffer, as well as data lost due to jitter buffer overruns, and the behavior of the codec.

A MOS can range from 5 down to 1. An estimate of the MOS can be directly calculated from the R factor, the quality rating of the E-model.

Figure 1. R factor values from the E-model are shown on the left, with MOS values on the right. The likely opinion of human listeners is shown in the middle.

The inherent degradation that occurs when converting an actual voice conversation to a network signal and back reduces the theoretical maximum R factor with no impairments to 93.2, and so the highest MOS is 4.41.

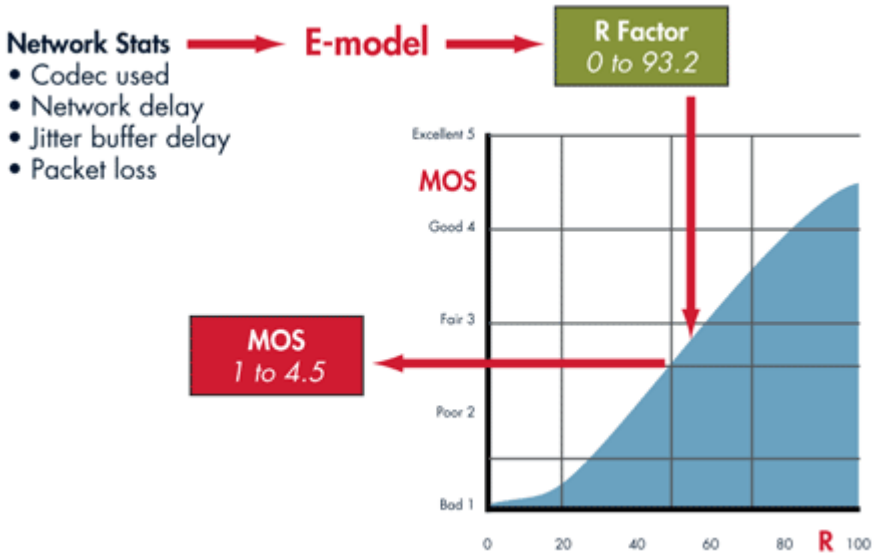

Figure 2. The E-model calculation takes as its input network statistics. Its output is an R factor, which is converted to a MOS estimate.

Let's look at each of the components in detail: codecs, delay, jitter, and loss.

3. Codec Selection

Codecs are the hardware or software used to convert an analog voice signal to digital and back. Five of the most common codecs are shown below.

Codec	Bit rate (kbps)	Frame time (ms)	Look ahead (ms)	Codec impairment
G.711	64.0	10	0	0
G.729	8.0	10	5	11
G.723.1-MPMLQ	6.3	30	7.5	15
G.723.1-ACELP	5.3	30	7.5	19
G.726	32.0	10	0.125	7

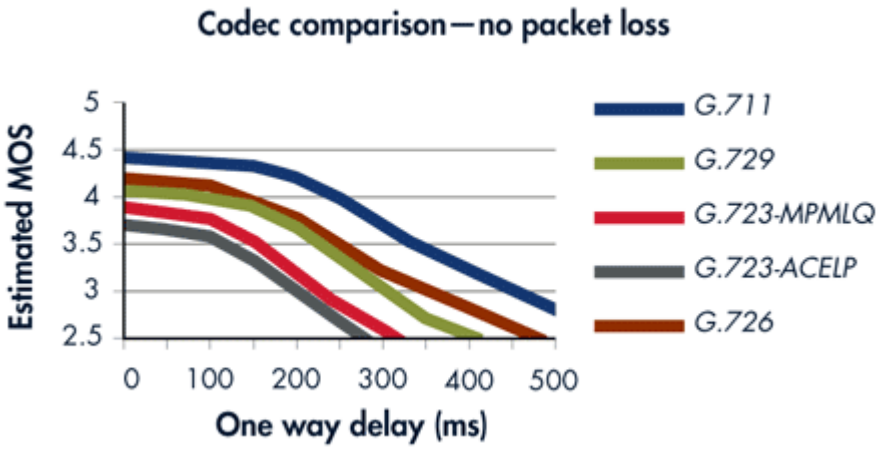

Figure 3. Five representative codecs are shown, with their nominal payload rate.

The G.711 codec gives the best voice quality, since it does no compression, introduces the least delay, and is less sensitive than other codecs to datagram loss. Other codecs, like G.729 and the G.723 family, consume less bandwidth by doing compression. The fact that they use less bandwidth is good, since you can get more concurrent calls, but the compression they do reduces the clarity, introduces delay, and makes the voice quality very sensitive to lost datagrams.

The codec impairments, as shown above, can reduce the R factor significantly. They are added directly into the "Ie" portion of the R-factor equation. For example, using the G.723.1-ACELP codec causes 19 points to be subtracted directly from the 93.2 points available in the R factor.

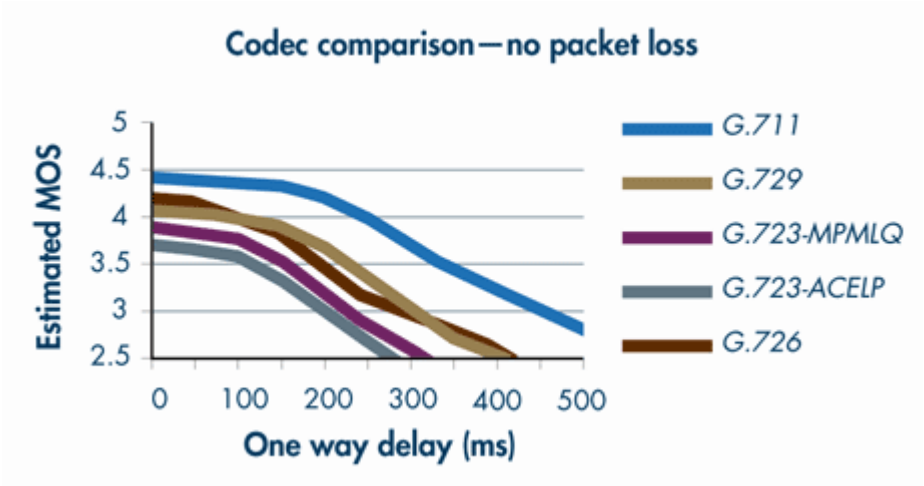

Figure 4. If there were no jitter and no datagram loss, the MOS would be influenced only by the network's one-way delay and choice of VoIP codec. This graph shows the effect of one-way delay for five codecs.

4. One-way Delay

One-way delay, the time it takes to get data across the network, is the primary indicator of the "walkie-talkie" effect. Humans are used to having conversations where they both talk at the same time. Most listeners notice when the delay is more that than about 150 ms; when it exceeds 200ms, they find it disturbing and describe the voice quality as poor.

The one-way delay that's measured from end to end is actually made up of four components:

Propagation delay: the time to travel end-to-end across the network. The propagation delay between Singapore and Boston is much longer than the delay between New York and Boston.

Transport delay: the time to get through the network devices along the path. Networks with many firewalls, many routers, and slow WANS introduce more delay than a LAN on one floor of a building.

Packetization delay: the time for the codec to digitize the analog signal and build frames - and undo it at the other end. The G.729 codec has a higher packetization delay than the G.711 codec, because it takes longer to do its compression.

Jitter buffer delay: the delay introduced by the receiver to hold one or more datagrams, to damp variations in arrival times.

The calculation of Id involves taking the sum of these delay components and processing them through the E-model equation. Measuring response time (round-trip delay) and dividing the resulting time measurement by two isn't always a good approximation of one-way delay. Response time hides assumptions about the symmetry of the paths between two

endpoints. The two RTP streams in a VoIP call can take different paths through an IP network.

Performance endpoints calculate one-way delay explicitly, rather than just taking the round-trip time and dividing it in half. The endpoints start with flows similar to those used by the Network Time Protocol (NTP) [6]. They maintain local copies of their clocks, because there can be many simultaneous connections. Also, the internal clocks in every different operating system and computer platform seem to be a little different, and the clocks drift apart over time.

The endpoints maintain virtual (software) clocks for each partner involved in a VoIP test. The virtual clocks consist of the offset between the microsecond clocks maintained by the two endpoints.

The microsecond clock is a high-resolution clock that's maintained independently of the operating system's system clock. The endpoints compare their respective views of the clocks prior to the start of each test and periodically during a test run. They also measure clock synchronization and drift between test runs, to establish a track record for the expected delay. Our one-way delay algorithms have proven robust in measurements with thousands of endpoint pairs. We've also verified their effectiveness in testing with stratum 1 GPS timeservers.

5. Jitter

A jitter value captures the amount of variability in the arrival times of the datagrams at the receiver.

The sending side sends datagrams at a regular periodic rate, say every 20 or 30 ms. Ideally, the receiving side would receive the datagrams at the same rate, in which case there's no jitter. However, all kinds of things can happen in data networks, and some datagrams arrive quickly while others arrive more slowly. If slow datagrams arrive too late, they are discarded to make way for the datagram that follows them. One method of damping the variability of arrival rates is to put a "jitter buffer" between the network layer and the VoIP application. A jitter buffer holds datagrams at the receiving side. It can compensate for variability of arrival rates and also deal with datagrams that arrive out of order. It hands the arriving datagrams to the processing application in order, at a more consistent rate. However, since the jitter buffer needs to hold the datagrams for some time to do this damping, it further increases the delay. And, compounding the problems somewhat, datagrams can be lost when a jitter buffer is overrun.

IxChariot tests simulate both the delay and data loss effects of a jitter buffer, showing the effect of various size buffers on the estimated MOS.

6. Lost Data

Datagrams that are lost generally can't be recovered, so they appear as momentary gaps in the conversation. Some tiny gaps are okay, but a consistently high rate of lost datagrams or periods where lots of datagrams are lost are disturbing to human listeners.

It is "bursts of loss" that degrade quality most significantly. A burst is generally considered a loss of five consecutive datagrams or more. Human listeners don't readily notice lower quality if loss of datagrams is randomly distributed, just a few at a time. There's some effect, as shown in the following two figures, but the quality decline is mostly because of the combination with delay. Bursts of loss, however, have devastating effect, and are weighted heavily in the E-model calculation.

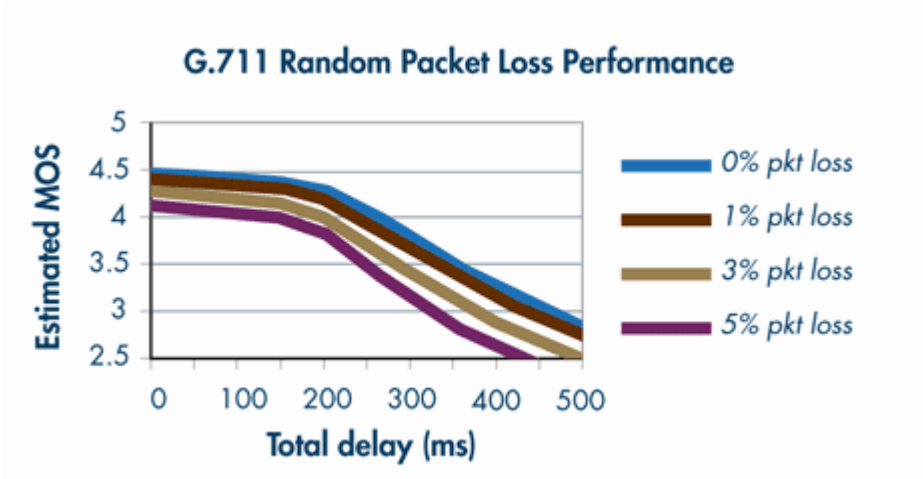

Figure 5. This chart shows the effect of randomized datagram loss when using the G.711 codec. The MOS estimate declines as the percentage of lost datagrams increases and as the delay increases.

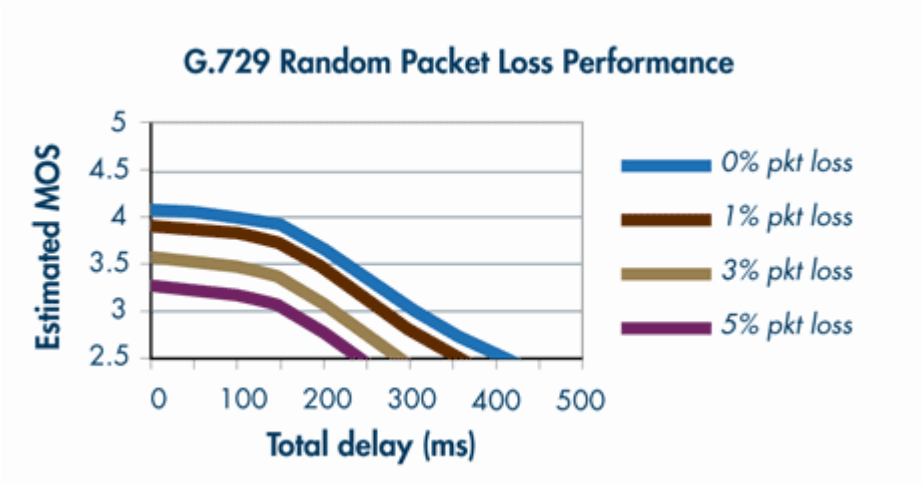

Figure 6. This chart shows the effect of randomized datagram loss when using the G.729 codec. With this codec, the MOS estimate starts lower than with G.711, and declines from there as the percentage of lost datagrams increases and as the delay increases.

7. Summary

IxChariot's granular measurements for one-way delay, jitter, and lost data can be overwhelming for someone not extensively trained. Our goal was to make the evaluation simple, so a single numerical score is used to estimate the quality of a voice conversation. We used the E-model, which takes these network attributes into account when calculating its R factor. The R factor is readily mapped to an estimated MOS.

Like all scores, it's strongest at the extremes, which results in a simple set of rules for those doing an assessment:

If the score is clearly high, the network passes the assessment.

If the score is clearly low, the network fails the assessment.

If the score is in the middle, the network is probably not in great shape, and more examination of the underlying data is called for.

Tools like IxChariot help you assess whether a data network is ready to carry VoIP traffic well. When doing a voice-readiness assessment using mean opinion scores, you'd like to see MOS values of **4.0 or higher** - otherwise users won't like how their phone conversations sound. When you test one call, do you get a score in that range? As you add more calls, how do the scores degrade? These are the types of questions IxChariot can answer when assessing VoIP call quality.

8. Additional Information

1. ITU-T Recommendation P.800, "*Methods for subjective determination of transmission quality.*"
2. ITU-T Recommendation P.861, "*Objective quality measurement of telephone-band (300-3400 Hz) speech codecs.*"
3. ITU-T Recommendation G.107, "*The E-model, a computational model for use in transmission planning.*"
4. ITU-T Recommendation G.108, "*Application of the E-model: A planning guide.*"
5. Network Time Protocol version 3, RFC 1305,**www.ietf.org/rfc/rfc1305.txt**

[**back** | **top of page** | **back to white papers & guides**]