

Implementing VoIP: A Voice Transmission Performance Progress Report

J. H. James, Bing Chen, and Laurie Garrison, AT&T

ABSTRACT

Aiming to introduce voice over IP networks and services in ways that satisfy the voice quality expectations of our customers, we have been conducting laboratory studies of how VoIP transmission affects voice quality while also carefully monitoring and managing several field implementations of VoIP. This article summarizes much of what we have learned in this work, and we hope it provides a useful progress report on the industry's evolution to VoIP. We review our data on the voice quality effects of packet loss, delay, speech coders, packet loss concealment algorithms, and the compression option of suppressing transmission during silence. Because the familiar problem of echo has emerged repeatedly in the VoIP environment, we review this issue in some detail. Packet loss and delay variation measurements made on private VoIP networks are reviewed, and the data here are encouraging. We finish by making our case that the network planning tool known as the E-model is currently an inexact predictor of VoIP network performance.

INTRODUCTION

Enough experience (both laboratory and field) with voice over IP (VoIP) has now accumulated for us to assemble a scorecard summary of what we understand about managing voice quality in the VoIP environment. This article reviews the lessons learned to date.

While such familiar topics as speech coding options, packet loss, and packet loss concealment strategies are addressed, our focus is on what we have learned about the voice quality effects of these variables. We also take an in-depth look at the issue of speech path delay and the challenge of controlling echo in the VoIP environment. The primary reason echo remains a significant performance factor is that most customers access VoIP networks via two-wire analog circuits.

The topic of the present usefulness of the E-model [1] for predicting the voice quality of VoIP networks is addressed. The view presented

here is that the supporting database for this model's VoIP network predictions is currently thin, and the model's rule for putting things all together is in need of work.

PACKET LOSS

Above some threshold rate, VoIP network packet loss introduces audio distortions that cause voice quality to decrease as the rate of packet loss increases. That said, on any particular connection this general effect can be modulated by:

- The distribution of the lost packets
- The packet loss concealment (PLC) algorithm in use

Early White Papers on VoIP generally repeated the assertion that packet loss did not become a significant problem until it reached a 5 percent rate. This assertion provided poor guidance; the more realistic situation is captured in Fig. 1 where we plot some mean opinion score (MOS) results from our Voice Quality Assessment (VQA) laboratory.

The data are collapsed over a variety of PLC algorithms and are based on the G.711 (64 kb/s) coder and a 20 ms packet size. The PLC algorithms used ranged from simply inserting silence for the missing audio to the use of the G.711 Appendix I algorithm that does a good job of masking the audio effects of up to 3 percent packet loss. Considering all the qualifying factors, we believe that VoIP networks must hold packet loss below 1 percent in order to deliver a level of voice quality that is public switched telephone network (PSTN) equivalent.

Intuitively, it would seem that the negative effects of packet loss would be exacerbated where packets are lost in clusters or bursts rather than in isolation. But where the overall rate of packet loss is held constant, the data are more complicated.

In Fig. 2 we plot the results of an MOS study in which we corrupted speech signals by introducing either 1, 2, or 5 percent packet loss and within each percentage distributed the lost packets in either a pseudorandom fashion where no two packets were lost in succession or in bursts

where up to eight consecutive 20 ms packets could be lost. What we observe is that at the lower 1 and 2 percent rates there is no significant effect of the distribution (random vs. bursty) variable, whereas with the high 5 percent rate opinion suffers more when packets are lost in bursts.

The coder used was G.711, and the PLC algorithm used was to replay a sample of the last valid speech information received for 30 ms into the packet loss episode and then insert silence for the duration of any longer episode. Since we have replicated the packet loss distribution effect shown in Fig. 2 with other PLC algorithms and coders, we believe it is robust. It is worth noting that where we used a high-quality PLC algorithm, such as that defined in G.711 Appendix I, the absolute MOS scores are significantly improved.

In our testing of field-deployed private VoIP networks (AT&T's and others) we have generally *not* seen troublesome rates of packet loss. Resource losses due to equipment failures and unprotected maintenance events have resulted in periods of high packet loss but thankfully these have been rare. Where other factors have been properly managed (sufficient network capacity provided, a high-quality coder in use, etc.) these VoIP networks have consistently supported PSTN equivalent voice quality.

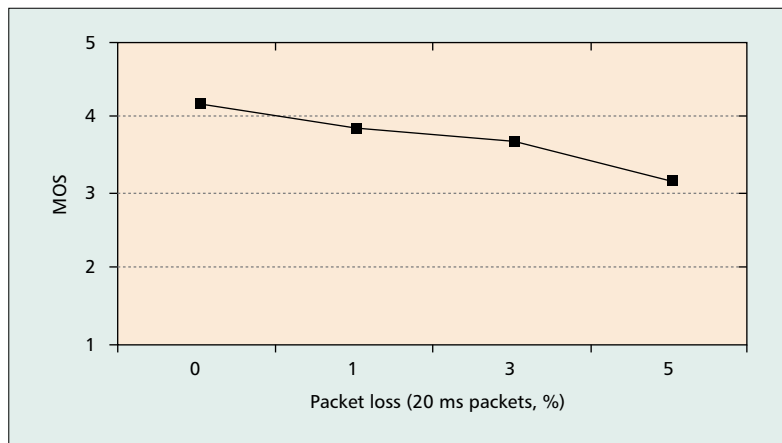
SPEECH CODING

In theory, the VoIP network/service architect has an array of speech coder choices. However, in practice, gateway vendors too often only support a limited number of coding options. Typically, a gateway will come with the G.711 coder and one, or at most two, low-bit-rate (LBR) coders. These LBR coders come with some speech quality penalty and, for non-waveform coders, some delay and complexity penalty.

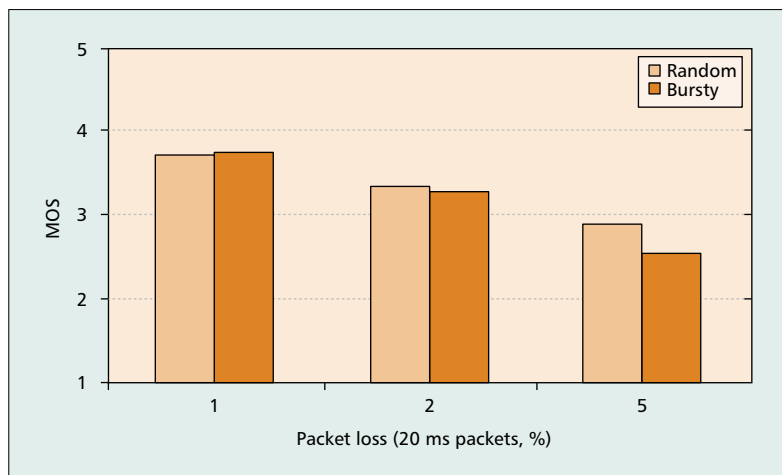
In isolation, coder performance generally follows bit rate (the higher the better). But in the VoIP environment more than one coder may be used on any end-to-end connection since the converging networks (digital cellular, broadband access, VoIP, PSTN, etc.) will frequently use different coders. Therefore, in considering coder performance it is important to consider the effects of processing speech through them multiple times. Such processing is normally referred to as *coder tandeming* although when the multiple encode-decode episodes involve different coders the term *transcoding* is often used.

In our studies of the effects of coder tandeming we have found that the process by which coder distortion accumulates over multiple encode-decode episodes is complex. Depending on the particular coder(s) involved, we see MOS performance penalties ranging from 0.06 to 0.56 for a single episode of coder tandeming. We further find that the size of this penalty is not a simple function of how well the particular coders perform in isolation (i.e., for a single coding).

This tandeming issue has motivated work on the development of strategies that obviate intermediate codings. One such tandem-free operation (TFO) strategy [2] transmits both an LBR coded stream and a back-up bit-robbed PCM stream. Another strategy to obviate transcoding



■ Figure 1. Packet loss effect averaged over PLC algorithm.



■ Figure 2. Random vs. bursty loss for G.711 with PLC.

has been proposed by Kang *et al.* [3] Theirs is a bitstream mapping strategy that takes advantage of the fact that many LBR coders use the same basic coding method. They studied the case where both the G.729 and IS-641 coders are in use on a connection, and showed both voice quality and delay advantages for mapping these two bitstreams as opposed to decoding one stream and then recoding it into the other.

Taking all the potential operating conditions into account, we have recommended that for VoIP services striving to maintain toll quality voice performance, under most operating conditions, one of the following coders should be used: G.711, G.726, G.728, or G.729E.

SILENCE SUPPRESSION

Individual connections have spare capacity due to the silence periods that occur as a customer assumes the role of listener. This fact has led to the strategy of having multiple connections share the bandwidth of large transmission pipes. This digital speech interpolation strategy, now more frequently referred to as either silence suppression or discontinuous transmission, requires a voice activity detector (VAD). Early on, VAD designs were prone to clip the speech signal. This clipping problem has been successfully

Those familiar with the laboratory data from which G.114 was formulated know that this depiction of delay's effect on voice quality is a simplified one. Very different results were seen by different contributing laboratories, and the resulting recommendation was a compromise.

addressed by the silence suppression features associated with the G.729 (Annex B) and G.723.1 (Annex A) coders, although we still see clipping with some “outboard” features used with the G.711 coder.

This speech clipping issue aside, silence suppression carries with it a quality penalty when used on connections that are characterized by background noise that gets coupled into the connection. With silence suppression in use, the transmission of this noise is blocked during “silence” periods and in an attempt to mask this process some artificial (comfort) noise is introduced. Attempts to have this artificial noise match the suppressed connection noise are seldom achieved; what the customer hears is an often annoying “pumping” of different sounding noise types.

Because of this noise-pumping problem, we have recommended against the use of silence suppression in VoIP services where a PSTN equivalence goal is set and high environmental noise is a likely operating condition for customers, or where music-on-hold signals are common. We also support the option of using one of the LBR coders that perform relatively well in noise, such as G.726 or G.729E, rather than silence suppression, when the value proposition of the VoIP service requires compression but noisy customer environments must be planned for.

CONNECTION DELAY

The transmission performance effects of connection delay are normally separated into two areas: the problems associated with the breakdown of the normal flow of conversations and the problem of echo. We follow that convention here. We will also touch on the subject of VoIP network delay variation, or jitter, since this variable can be the source of packet loss and can have the negative effect of driving adaptive dejitter buffers to their maximum length.

THE CONVERSATION EFFECTS OF DELAY

Discussions of VoIP typically flag the issue of increased end-to-end delay and discuss the effects of this delay in terms of its potential for interfering with the normal cadence of voice conversations. A real-time expectation guides our conversation behavior and, where this expectation is violated, the back-and-forth nature of the conversation begins to break down as we start to talk over each other (double talk) and, consequently, become more hesitant about switching between the role of listener and talker. It is this type of problem that motivated the development of the International Telecommunication Union — Telecommunication Standardization Sector (ITU-T) Recommendation G.114 on delay [4]. Basically, G.114 advises that one-way delay can accumulate up to 150 ms without effect, but beyond that point negative consequences begin to gradually accrue.

Those familiar with the laboratory data from which G.114 was formulated know that this depiction of delay's effect on voice quality is a simplified one. Very different results were seen by different contributing laboratories, and the resulting recommendation was a compromise.

Given that the introduction of VoIP will increase connection delays, and sometimes increase them beyond the 150 ms recommendation, we have taken another look at this issue in a study designed to be somewhat more “naturalistic” than the traditional laboratory study of delay. Our procedure was to initiate voice calls to members of the subject pool the AT&T VQA Lab maintains in support of our MOS studies. In this study 24 of these subjects were called at home and engaged in a conversation that lasted approximately 5 min. Such calls are common since they are used to confirm dates/times when the subject will participate in an MOS study. We set the one-way delay on these connections to 200 ms. We concluded the call by asking, “Is there anything about this connection that you find different from your typical call?”

In response to this question *not one person* mentioned a delay-related issue. While most of these people indicated that the connection sounded the same as their normal connections, others offered up such observations as “tinny” sounding voice, and “crackle” noise on the connection. But again, no one commented on the time dimension or reported problems of interrupting or double-talk.

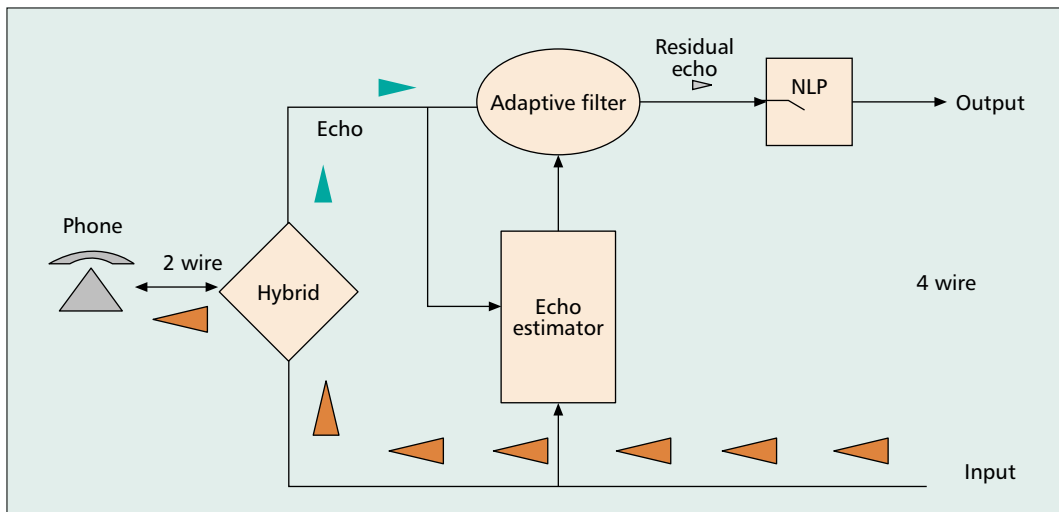
Given the naturalistic flavor of this exercise, and the fact that the normal expectation for these test participants was that they were conversing on a local call, the fact that no one pointed to the delay dimension suggests to us that one-way delay of up to 200 ms does not introduce an obvious transmission impairment. This and other data have led us to relax our concerns about introducing VoIP networks and services that push end-to-end delays into the 150–200 ms region.

ECHO

It was understood, especially after the effort to solve the echo problems introduced with digital cellular technology, that VoIP networks would require the use of echo cancellers. And since echo canceller technology was mature, and extensive industry performance requirements in place, the avoidance of echo problems seemed to be a fairly simple network design matter. This hope was not realized.

Early problems were found with the canceller designs embedded in the DSP fabric of the VoIP gateways. “Bugs,” insufficient code allowances, and untested designs were found, and our product testing period was extended by the cycle of modification and retest. While many designs performed reasonably well against the set of tests and requirements defined in ITU-T G.168 [5], certain vendors relied too heavily on this testing since we found obvious and often severe problems when actual speech signals were processed through these cancellers rather than the specific “speech-like” test signal required by the G.168 standard. In fact, this problem led us to develop a set of tests that probe the same performance dimensions as G.168 (i.e., convergence speed, double talk recognition, etc.) but with samples of speech that capture echo canceller performance when analyzed within the standard MOS test process.

Early introduction of a VoIP link into the core network exposed a problem we refer to as



■ Figure 3. A simplified echo canceller model.

Due in large part to careful network planning, the VoIP field networks we have tested in have been essentially free of the more critical problem where echo persists on a call.

initial echo. Initial echo is the experience of echo at the beginning of a call, typically on the first word or couple of words out of one's mouth. Figure 3 provides a simplified echo canceller model to aid discussion of this problem.

To understand the initial echo problem one needs to understand:

- That the front-end process (the adaptive filter process) of an echo canceller needs some time to work
- That while the adaptive filter is converging on the echo, some significant level of residual echo energy is likely to escape the echo canceller and be reflected back at the talker
- That the duration of this initial echo period is a complex function of the echo return loss (ERL) for the particular connection, the design of the adaptive filter, and the threshold used by the echo canceller to operate its second-stage process (aka its nonlinear processor, NLP) that blocks the residual echo energy by functionally opening the return path
- That customer sensitivity to this initial echo increases with increasing connection delay

The addition of a VoIP link does *not* make the network echo canceller work more slowly; the same residual echo is present before and after the introduction of the VoIP link. It is just that without the additional VoIP delay, the initial echo is below the customer's perception threshold; with this additional delay, the echo is now heard.

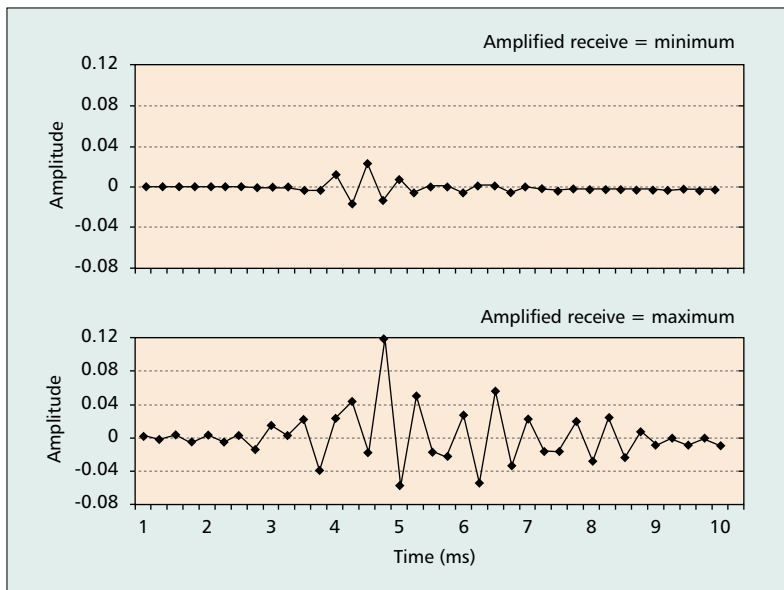
The problem of initial echo was associated with a small number of customers whose lines were poorly balanced to the network and therefore characterized by low ERL levels (often in the 8–12 dB range). This limiting factor was offset by the fact that it was often the same customers, those calling the customers having the low ERLs, who were repeatedly exposed to the problem.

This problem motivated attempts to improve the convergence speed of the echo canceller, either through the design of faster operating adaptive filters or, in some instances, by lowering the canceller's threshold for the operation of its NLP. Interestingly, we are beginning to see

another approach to this problem: having the echo canceller act like an echo suppressor at the very beginning of a call, where initial echo might be a problem, then to switch to a conventional echo canceller design once enough time has elapsed for the adaptive filter to have built up sufficient echo control. Our experience with this approach is limited to the laboratory, where it looks promising but, as with most wrinkles in well understood designs, this needs to be followed carefully into the field environment.

The customer impact of the initial echo problem was lessened by the fact that it was confined to the very beginning of a call. Due in large part to careful network planning, the VoIP field networks we have tested in have been essentially free of the more critical problem where echo persists on a call. The notable exception here is in the consumer access (cable/digital subscriber line, DSL) VoIP environment where a media terminal adaptor (MTA) interfaces the customer's phone with their broadband network. This MTA contains both a requisite hybrid circuit and an echo canceller. Given the physical closeness of the MTA's echo canceller to the source of the electrical echo, the delay associated with the echo path is relatively short, and the industry has elected to use cancellers with an 8 ms operating range in the MTA. While on paper 8 ms of operating range appears sufficient to control echo in this environment, we have found less than adequate echo control under certain conditions.

Troubleshooting consumer access VoIP echo problems, we found the common denominator to be the use of a telephone having an amplified receive feature that is set to its high level. An effect of this amplification is that the far-end party's speech signal emerges from the set's receiver with enough amplification that it can couple back into the set's transmitter and be heard as echo at the far end. In this case the MTA echo canceller must deal with a composite echo signal that is the product of the reflection coming from the hybrid circuit and the acoustic echo coming from the handset. As we show in Fig. 4, the addition of the acoustic path echo can



■ **Figure 4.** Effects of the handset amplifier on the echo impulse response.

be associated with enough extra delay that the composite echo seen by the MTA canceller can have at least some energy delayed beyond 8 ms.

The top panel of Fig. 4 shows the echo's impulse response (i.e., time-energy characteristic) when the handset in use behind the MTA has an amplified receive capability but is set to its *minimum* setting, whereas the bottom panel displays the expanded (i.e., significant energy now appearing later in time) and amplified impulse response that results when that same handset's amplified receive feature is set to *maximum* gain.

It is instructive to understand that, as shown in Fig. 4, an echo signal is distributed over some period of time (roughly 5 ms). When all, or at least the bulk, of the energy falls outside the operating range of the canceller, what is heard is what is generally defined as echo (i.e., a second and delayed version of what just came out of one's mouth). But where only some energy, normally a function of frequency, falls beyond the operating range of the canceller, the impairment heard can range from that of intermittent echo to that of voice-correlated noise.

We are still wrestling with the echo problems associated with these 8 ms MTA echo cancellers. What we know strongly suggests that the adoption of a 16 ms canceller design would solve these problems. We appreciate the fact that increasing the operating range of an echo canceller requires additional processing power and thus comes at a cost. But from what we have seen so far in this consumer VoIP environment, it looks like CableLabs' suggestion [6] that vendors might want to use a canceller with greater than 8 ms of range as a product differentiator is a good one.

DELAY VARIATION (JITTER)

The processing of speech packets in the media gateways and their network transmission between gateways is not a time-invariant process. Rather, there is normally delay variation associated with

this whole process. To absorb this delay variation, the egress gateway temporarily holds (buffers) incoming packets, thus reducing the need to discard packets whose arrival time is offset by this variation while also allowing the original fixed-rate audio stream to be recovered and delivered to the customer.

Our experience with private VoIP networks (AT&T's and others) as well as with broadband access networks (cable and DSL) is that jitter tends to be small. We typically see fairly tight distributions on packet delay variation where a single packet arriving more than 10 ms from its expected arrival time is a rare event. Such results have biased us toward recommending smaller fixed-size dejitter buffers, and lower limits on the size to which adaptive buffers can grow.

VOIP NETWORK PERFORMANCE PLANNING AND THE E-MODEL

As we hope the preceding indicates, successful VoIP networks are, and continue to be, deployed. In part this success is due to the fundamental laboratory work that has explored how voice quality is affected by such VoIP variables as packet loss. It is also due in part to the extensive predeployment testing of products and architectures that has allowed problems to be found and solved before they can affect customers.

Traditionally, our own network performance planning has made use of performance models that capture existing knowledge and provide a means of predicting from it. The model in use in the industry today is the E-model, and there is some disagreement regarding its utility for helping us plan in the evolving VoIP network environment. Some believe that the E-model is developed to a point where it can be used to make strong predictions about the customer effects of certain VoIP network designs and performance states. We believe the E-model's VoIP-related database is still thin and somewhat "noisy." On this second point, it is our observation that gaining valid and reliable subjective testing data about the voice quality-VoIP network relationship is made more difficult by the fact that we are now dealing with time-varying impairments such as packet loss that challenge traditional subjective testing methods.

This issue of the strength of the E-model's supporting database will fade as more data are applied, but even then a more sophisticated rule for combining the effects of multiple sources of impairment on a connection will have to be developed. The rule used in the model currently is one of simple addition. Problems with this assumption are discussed in the companion article by Takahashi *et al.* [7]. We have pursued this matter extensively and offer some data of our own.

We conducted an MOS study in which we varied two impairment sources: speech coder and one-way delay. We chose three different combinations of these two variables that yield the same predicted MOS level according to the E-model. Specifically, we achieved an R-value of 72.6, which translates into a predicted MOS of

3.72. The specific coder/delay combinations were:

- The G.711 coder that has no amount of equipment impairment (I_e) associated with it, and 358 ms of one-way connection delay
- The G.729 coder, that has 10 units of I_e , with 264 ms of delay
- A tandem episode of encodings involving the G.729 and G.728 coders that yields 17 units of I_e , with 205 ms of delay

In deriving the R-value results we assume that the best possible result is $R_o = 93.2$, and that only I_e and I_d reduce that level. The specific I_e values and the R-to-MOS conversion we obtain from G.107 and G.113 [8].

Our means of acquiring actual (i.e., measured) MOS results was to have pairs (25) of test subjects converse over test connections for which we had control of both the speech coder in use and the amount of one-way connection delay.

The test subjects were instructed to use their experience with normal telephone connections as their reference in judging the quality of the test connections, and to speak back and forth so that both could listen to the connection quality. Each trial was 3 min long.

Our results are shown in Fig. 5. Clearly, we see a discrepancy between the predicted and measured MOS results. Just as clearly we can conclude that the predicted vs. measured difference is not some simple offset effect that can be remedied by some constant correction factor. We see three results: the E-model prediction is significantly pessimistic (the G.711/358 ms result), the E-model prediction is just about right (the G.729/264 ms result), and the E-model prediction is way too positive (the tandem encoding/205 ms result). The importance of these results is magnified, we believe, by the fact that the impairment sources used in the study, speech coder and connection delay, are not new, but rather have received considerable MOS attention.

From our point of view, then, the database used by the E-model to predict customer opinion of VoIP network quality needs to be built up and a more sophisticated rule developed to model how a customer perceives the overall quality of a connection when it is impaired in multiple ways. Until such progress is made we recommend that E-model predictions of VoIP network performance be interpreted with caution.

CONCLUSIONS

The major conclusions we have come to regarding voice quality and VoIP are:

- Well designed and managed VoIP networks are being deployed that provide toll-quality voice performance.
- For this high quality goal to be achieved, packet loss needs to be held below 1 percent.
- PLC algorithms are useful.
- The benefits to the VoIP service provider of using compression can come with significant quality penalties, especially where multiple codings are likely end-to-end and/or where high noise levels (or music on hold) might be a common operating condition.

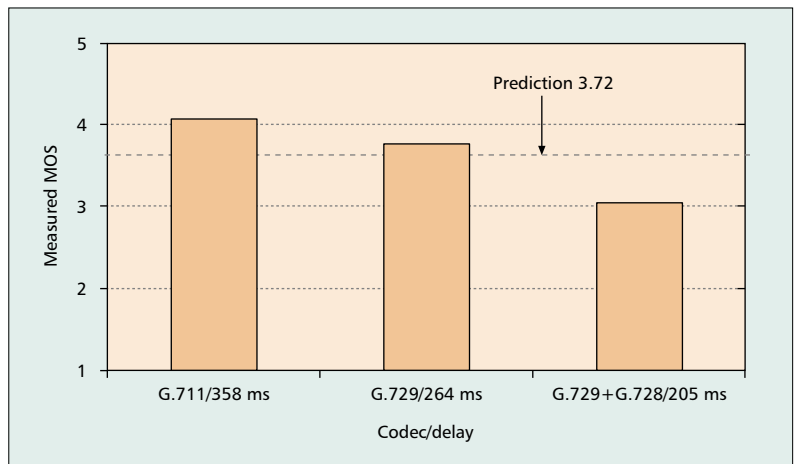


Figure 5. Measured MOS vs. E-model prediction.

- The fact that VoIP networks will sometimes be associated with one-way delays approaching 200 ms will likely not have significant customer effects.
- Echo control remains a challenge for VoIP.
- Current efforts to model VoIP network performance are probably premature.

REFERENCES

- [1] ITU-T Rec. G.107, "The E-Model, A Computational Model for Use in Transmission Planning," July 2002.
- [2] ETSI TS 128 062 V5.4.0 (2003-09), "Digital Cellular Telecommunications System (Phase 2+); Universal Mobile Telecommunications System (UMTS); Inband Tandem Free Operation (TFO) of Speech Coders; Service Description."
- [3] H.-G. Kang, H. K. Kim, and R. V. Cox, "Improving the Transcoding Capability of Speech Coders," *IEEE Trans. Multimedia*, vol. 5, no. 1, Mar. 2003, pp. 24–33.
- [4] ITU-T Rec. G.114, "One-Way Transmission Time," May 2003.
- [5] ITU-T Rec. G.168, "Digital Network Echo Cancellers," June 2002.
- [6] PacketCable™ Audio/Video Coders Specification, PKT-SP-CODEC-I05-040113, Jan. 13, 2004.
- [7] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS Assessment Technologies for VoIP," *IEEE Commun. Mag.*, this issue.
- [8] ITU-T Rec. G.113, "Transmission Impairments Due to Speech Processing; Appendix I: Provisional Planning Values for the Equipment Factor I_e and Packet-Loss Robustness Factor B_{pl} ," May 2002.

BIOGRAPHIES

J. (JIM) H. JAMES (jhjames@att.com) is currently technical manager of AT&T's Voice Quality Assessment Laboratory. He received a Ph.D. in the area of experimental psychology from Northern Illinois University in 1977. From 1977 to 1979 he was a postdoctoral research fellow at Yale University. He joined AT&T Bell Laboratories in 1979 and throughout his career has studied the voice quality effects of evolving technology.

BING CHEN (bingchen@att.com) received her B.S. in mechanical engineering from TaiYuan Institute of Technology, China, and her M.S. and Ph.D. in psychology from the University of California, San Diego. She joined Bell Laboratories in 1990 and has been engaged in the quality assessment of audio, multimedia, and network equipment.

LAURIE GARRISON (lfgarrison@att.com) received her Ph.D. in cognitive psychology from the State University of New York at Buffalo in 1991. Prior to joining AT&T, she worked on neural network character recognition for Eastman Kodak. Since becoming a member of AT&T's Voice Quality Assessment Laboratory in 1993, she has been involved in the performance evaluation of speech coders, echo cancellers, and end-to-end network quality.