Making IP Networks Voice Enabled

A 1999 White Paper Updated

This is a white paper that was originally written in June of 1999 when IP was just beginning to be used for real-time voice communications. It is reprinted here as the concepts are critical to understanding how to manage and operate networks for real-time traffic. Note that other than few a slight edits in the Introductions and Conclusions section reflecting the current use of IP as a real-time transport, the rest of the paper is as written in 1999. The observations and calculations are equally valid today. Note that the paper does not deal with HD or wideband audio. There is discussion of that in the PKE Consulting white paper " CoS and QoS - Managing bandwidth and Complexity". As it was originally authored by Phil Edholm at Nortel, it is reprinted here.

Introduction

This paper is focused to what are the operational characteristics required to deliver quality voice/audio over a packetized IP network. Within this domain, there are two key characteristics of the actual delivered voice quality; the realism of the sampling (rate and compression algorithms (and the network operational factors that impact the human perception of the “realism” of the communication. While there can be great debate on the subjective value of PCM versus CD quality audio and the comparison of various compression technologies (723 vs 729, lossless versus perceptual coding, etc.), as above, these discussions will be left to other forums. In this paper we will focus on the underlying network infrastructure operational issues that ultimately will impact the quality of voice over the underlying IP network.

Human Time versus Network Time

Voice and data networks have fundamentally evolved in different domains. Voice networks have always operated in the human time domain (see Figure 1). Human time is characterized by the perceptual human reaction to delay in the network. Humans have three distinct thresholds of delay perception, two of them evolutionary (or created, depending on ones view), and a third that is “trainable”. Figure 2 shows these three “delay thresholds” and their relative value. The first is echo, the second the interactive latency of human “CSMA/CD, and the last the patience

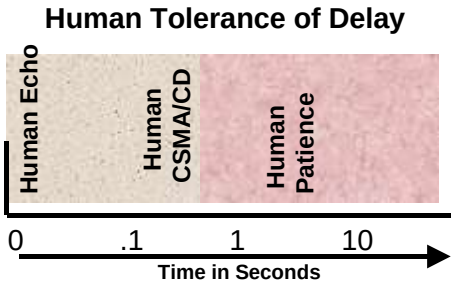


Figure 1 Relative Human Delay Factors

factor.

The echo latency model is actually based on your ear hearing your speech through the bone structure. The human brain naturally accepts and utilizes this "feedback" in the course of conversation. When an echo is generated from the telephony system (from a digital to analog conversion point or a cheap handset that couples from the speaker to the microphone), a potential problem is created. When the echo exceeds 15 msecs behind the original voice and arrives with sufficient volume, the human brain is confused and communicates the result as a garbled sound. .As framing rates for all IP based telephony service are a minimum of 10 msecs and the frames go both directions, the absolute minimum delay with infinitely fast compression and networks would be 20 msecs, guaranteeing that echo cancellation is required. Echo cancellation attempts, through complex algorithms implemented in DSPs to cancel the echo out of the voice path and spare the brain. Again, the value of echo cancellation and the relative merits of a variety of techniques and implementations will be left to others. As an aside, the typical Local Branch Exchange and PBX avoid requiring echo by maintaining an end to end delay of less than 5 msecs. This is achieved through the simple technique of sampling the voice every 125 microseconds and using synchronous TDM transmission technology. Echo is critical on operational VoIP solutions where the actors that create echo (talking off the net, talking to users with "cheap" phones) will occur. Due to the ubiquity of the PSTN, it is reasonable to assume that all VoIP implementations will require echo cancellation at some point(s) in the network.

The human CSMA/CD factor is a direct extension of how we communicate when we are face to face. We typically use a very Ethernet-like protocol (Carrier Sense Multiple Access) to share the talking bandwidth between us. When two people talk at the same time we use a back-off algorithm to resolve the collision (while the Ethernet uses a egalitarian binary back-off, the human equivalent appears to be hierarchical in nature – subordinate defers to boss, husband to wife, or devolves to the true aloha forms of unintelligible babble). Interestingly, the 1-persistant nature of Ethernet (where a node transmits immediately upon the availability of the channel, guaranteeing a collision when two nodes have packets queued) also is a direct emulation of the human environment, where most listeners are not actually listening, but waiting for the current speaker to stop speaking so they can begin talking. Above and beyond the CSMA/CD, humans use an acking protocol during the conversation to assure the communication is moving forward. During a communication, the speaker will pause occasionally, allowing the listeners to "ack" with a head nod or to "nack" by beginning to speak with their own views. When humans are distanced over a telephone system, this "acking protocol" becomes virtual., the speaker paused for a perceptible delay, waiting for the listener to reply. IN this case, the absence of response (silence) is taken as a virtual "ack" and the speaker proceeds. The challenge in this area is that human perception of a "delay" is between 200-300 msecs. A speaker will pause, but will begin talking again after that delay. If the overall latency (end to end – speakers mouth, back to the ear) is over the 200-300 msecs, then a collision will occur when the listener starts to talk and the original speaker talks also. This phenomenon is easily identifiable to anyone using a satellite long distance, where the overall delay exceeds a half a second. This "virtual collisions" make human communication very difficult, and

when combined with language barriers, any make it entirely unusable. This factor is why the open Internet VoIP, often with half second or higher delays, is known as push –to-talk. To successfully deploy a VoIP Telephony solution, the system must operate within the 250 msec parameters (300 msecs max.). This round trip time must include all latency and delay throughout the system end to end in either direction.

The final form of delay is the patience factor. While the previous delay criteria are physiological, the patience factor is trained. This is the time that a user is willing to wait for an event to happen after the enter key is struck.. I refer to this generally as "network time", because it often exceeds a second and is the domain where file transfer and web browsers operate. The typical PC operates in human time at the keyboard/monitor interface and in network time out the back Ethernet port, Personally, I find my patience is directly impacted by the delays common in windows (influencing me to be more patient and zen-like) and the incredible performance achieved to the Internet by my cable modem service at my home. My tolerance for delay has decreased by 90% due to the cable modem (versus 28.8 dial-up)

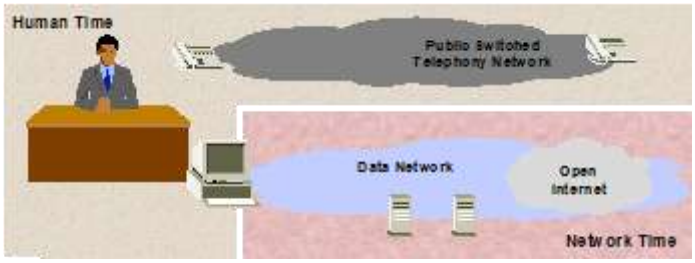


Figure 2 Human versus Network Time in Today's Networks

Figure 2 shows the issue of trying to support real-time traffic on networks that are not designed to these characteristics. The core goal of an IP network that can support voice is to build the network so it can support both network and human time applications (Figure 3).

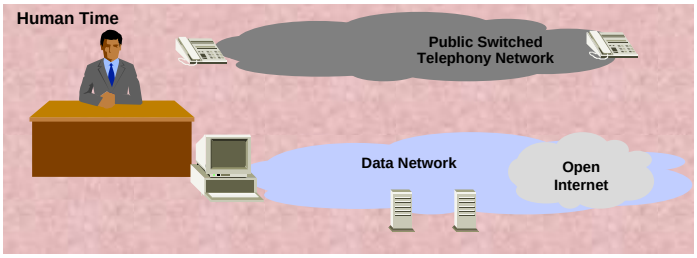
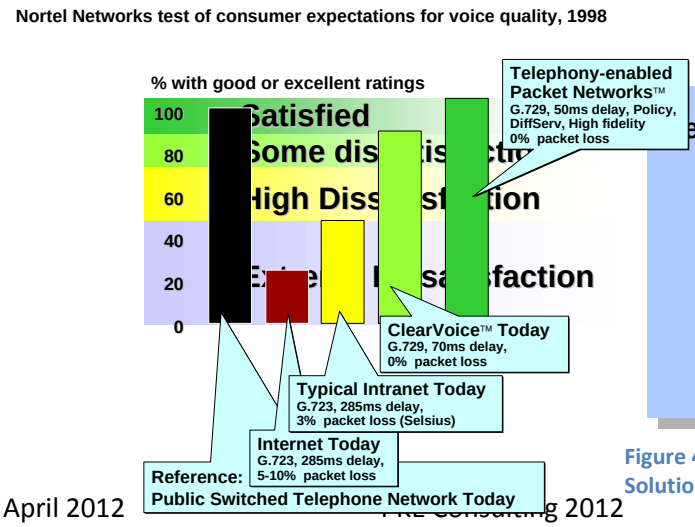


Figure 3 Extending Data Networks to Human Time

Figure 4 shows testing that Nortel Networks did in the late 90s in our their laboratories that measures the user acceptability of various communication values related to IP. Note



elay Impact on Perceived Voice Quality

Round Trip Delay	%Poor or Worse
150 msec	3
175 msec	4
200 msec	5.5
225 msec	6.5
250 msec	8
300 msec	10.5

Figure 4 Subjective Perceived Quality of different VoIP Solutions

that both the open Internet, with dramatic latency and packet loss and the Corporate Internet have performance that is not acceptable for either business or general personal use. The goal of VoIP Telephony is to achieve the right capability, where in addition to meeting the criteria for loss and delay, higher quality voice sampling (CD quality) can be used with compression to generate an even more realistic voice call than PCM can today. The chart on the right is a representation of perceived quality in subjective tests as delay is introduced. The measure her is the percentage of poor or worse ratings as delay is increased. Typical Telcos use a maximum of 5% tolerance for this test. Therefore, a maximum latency of 200 – 250 msecs is the maximum.

Achieving Human Time in IP Networks

As was discussed early, a key criteria to achieving a voice network that can operate to our human perceptions is the echo delay. As the typical frame size will dictate that echo cancellation is required for many calls, echo cancellers are a way of life in the IP voice world. The other key criteria is meeting the 250 msec latency goal for user acceptable telephony. To accomplish this, we must take into account all of the latency in the network, from the speakers mouth, through the system back to the speakers ear. The round trip is crucial as this is the delay the speaker senses and uses to determine his communications patterns. If we assume that the two paths are basically the same, we can analyze a uni-directional path with a maximum latency of 125 msecs. While a goal of 125 msecs will provide in most cases an acceptable result, a goal of 95 msecs assures that there will be no discernible difference to the PSTN. This then is the range of total one way delay that must be achieved to deliver quality voice.

The latency between the speaker's voice and the receiver ear (the one way path) is the sum of a number of components: Subjective Perceived Quality of different VoIP Solutions

- Quantitization time (10-30 msecs) - the frame time to generate a discretely sampled element
- Compression time (2-10 msecs) - the time it takes to compress the voice into packet form
- Packetization time (1msec) - processing time to assemble and transmit the packet
- Network transit time (variable) - the time to transition the network
- Reverse packet and decompression (2-3 msecs) - the time it takes on the receiving end to output the samples
- Latency is for a round trip

Figure 5 shows a timing diagram of the transmission of voice over an IP network and the additive effects of the delays. The two lines on the right represent 95 and 125msec (180 or 250 msecs round trip) respectively from the beginning of the latency at the left.. All packetized voice systems must have a minimum of two frames (packets) worth of voice sample delay. On the chart, frame 1 is in the input buffer of the transmitting node and frame 3 is in the output buffer. As these two frames are "synchronized" (the next buffered frame arrives just in time), they are essentially the same delay. Therefore, the beginning of frame 3 is shown as the end of the "acceptable" window of latency. The net result is one frame of delay for the sampling. Another frame is in the receive buffer as a jitter packet

(frame 2 in the drawing). The net result is that at least two frames of delay are required. On the chart the frames are shown as 20 msecs, where actual frames will vary from 10-30 msecs. On the chart, the additive effects of compression, packetization, and the reverse process are shown.

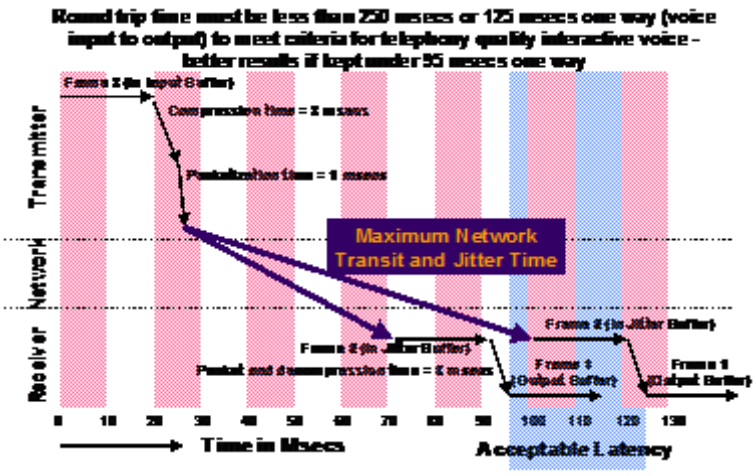


Figure 5 Time Model for VoIP Transmissions

For purposes of this analysis, the compression is assumed to take about 3 msecs. This is a reasonably low estimate. It is critical to note that different compression techniques have different times to complete. Based on implementation data, a typical G723 compression will take twice as long as G729. This is purely based on the number of DSP cycles required for 723 versus 729. However the 3 msecs is a reasonable minimum. Using these factors, both the Non-transport latency and the available latency for transport can be defined. As can be seen on the drawing the total non-transport and available transport latency is:

Non-Transport Latency = 2 x (frame length) + compreession + packetization + de-packetization + de-compression

Available Transport Latency = Total Latency (95 or 125 msces) – Non-Transport Latency

Obviously the total delay is very dependent on the frame size. In the drawing a reference frame of 20 msecs has been used. In actual implementations, the frame length is defined in the standard. The frame length is a trade-off of latency versus efficiency. Fro example, if the frame was 1450 bytes long (or 11600 samples at 125 microsecs or over one second of actual conversation) the efficiency would be very high (over 95%). This means the actual transmitted data stream would be about 8.5 kbps for a voice stream of 8 Kbps. However, that one second frame when doubled in each direction would yield a delay of over 4 seconds. This is more akin to talking to the moon than to the next city. Alternatively, is we are to attempt to duplicate the latency in the PSTN where voice samples are sent each 125 microsecs, in essence, placing a compressed byte in a separate packet every 1 msec, the overhead balloons. As the header is over 44 bytes, the overhead is 44 times the voice data rate. For a compressed rate of 8 Kbps, this means the transmitted stream on the network would be 352 Kbps! This is obviously a poor answer if VoIP is to save money, especially when it would take two T1 lines to send one voice call! This is one of the reasons that AAL2 in ATM was designed to support micro-cells within a larger cell to optimize the factors of latency and overhead.

The answer is to pick a specific frame size, typically from within the range of 10-30 msecs per frame. G729 and G723 have picked respectively 10 and 30 msecs as their frame size. The means that 729 favors latency at the expense of lower efficeiny and higher network data rate, while 723 chooses better efficiency at the cost of higher latency and potentially reduced voice quality. Table 1 shows the relative values for non-transport latency, Available transport latency to meet the 95 and 125 msec total criteria, and bandwidth to transmit an 8 Kbps compressed stream. For this comparison, the 729 stream has a 3 msec compression time, and the 723 stream has a 6 msec compression time and both have the same 1 msec for packetization and 3 msecs for de-packetization and de-compression. The bandwidth is shown for continuous voice and does not include the value of silence suppression, which can reduce the actual transmission by up to 50%.

	G729	G723
Total Non-Transport Latency	27 msecs	67 msecs
Available transport latency at 95 msecs	68 msecs	28 msecs
Available transport latency at 125 msecs	98 msecs	58 msecs
On network transmit bandwidth	43 Kbps	20 Kbps

Actual transmission rates are typically about 14 Kbps for 723 and around 25 Kbps for 729.

Having established the parameters for operation, what are the factors in transmission that add latency? The first delay factor is the actual clocking of the data packet onto the media. As the packets are typically of minimum size (64 bytes), even on a 10 Mbps Ethernet, the actual packet transmission time is only 5 microseconds. As often network run at higher speed, packet latency is generally not critical. The one exception to this is when low speed lines are used, either as modem connections from a remote user or when fractional T1 lines are uses from a branch office. It takes 15 msecs to transmit a 54 byte packet over a 28.8 modem and 7 msecs to send over a 64K trunk line. This is further impacted by the potential of the voice packet being immediately behind a maximum length data packet (1500 bytes). In this case the voice packet would be delayed a total of 420 msecs over the 28.8 modem and 190 msecs on the 64K line. Obviously, reasonable voice latency and low speed lines are not compatible without other factors to operate the line. A second major delay factor is the propagation delay of the packet through a WAN environment. Excluding active components, the transmission fiber, typically at 80% of the speed of light, is 38 msecs for 5,000 miles (8,000 kilometers). I the traditional PSTN network, where end delays in PBXs and COs are less than 5-6 msecs, this is the dominant delay. In our IP network it becomes a significant part of the latency. Another factor is the forwarding time in packet processing nodes (routers) through the network. Each router hop adds a measurable time to the packet, from less

than 1 msec up to 3 msecs. The final additional latency factor is added packets in the transmission path (jitter packets). To prevent under-runs at the receiving node, additional packets are buffered in the receiving node. In the event of an delay through the network transmission (due to lack of or poor QoS/CoS), the packets in the buffer assure that the voice is not interrupted. However, each additional frame that is held in the buffer adds delay to the overall path. Therefore, if there are two addition jitter packets, the latency is increased by 20 msecs for 729 and 60 msecs for 723.

Adding together these latency factors will give an idea of the latency over the network. If we assume 2 jitter packets, a 5,000 mile optical path ad 10 router hops at 1 msec each, the total latency, including the Non-Transport Latency is:

= Non-Transport + Jitter + transmission + routers = Total

G729 = 27 + 20 + 38 + 10 = 95 msecs

G723 = 67 + 60 + 38 + 10 = 155 msecs

Obviously, the challenges of putting voice onto an IP network and meeting the key latency factors is a challenge. Under reasonable circumstances, G729 will offer voice that is close to equivalent to PSTN, while G723 is difficult at all. For the IT professional considering bringing IP Telephony into their organization, the challenge is to pick the right technologies and to manage the operational factors in the network to assure meeting the QoS/CoS goals.

There are three paths to building a WAN to support voice. A totally private WAN can be built, using circuits or public ATM/Frame Relay. Within the confines of this network, the QoS and system latency can be managed. A second option is to use an ISP that manages their network to assure latency. Today, certain ISPs will “guarantee” round trip latency to be less than 120 msecs anywhere in the US. This is equivalent to 60 msecs in either direction. If we incorporate this into the above analysis as the transmission/router path, the total latency one way with 2 jitter packets will be 107 msecs for 729 and 167 msecs for 723. Careful negotiation and operation in this space is critical to success.. The third option is to use the open Internet. Here the latency is not guaranteed in any way and typical latency in the net exceeds 250 msecs one way, making this impracticable for business applications.

Gone Forever

Typical data networks, in addition to operating in network time, use retransmission to assure reliable delivery of the information. In the voice world, the latency as discussed above makes retransmission of lost packets impracticable. Therefore, lost packets have a significant impact on the quality of the voice. When a packet is lost, the voice in it cannot be recovered. The better designed codecs will repeat the previous frame buffer in the event of a lost frame, minimizing the aural impact of a single lost frame, the problem is compounded either by multiple lost frames I sequence or continuous losses. Lost frames

come from two factors; packets lost through bit errors and packets lost through network congestion. Network congestion is managed through the QoS and latency as discussed above.

The loss of one or multiple voice packets may be noticeable. Obviously, the loss of a frame in 723 is moe noticeable than in 729 due to the length of time in each frame. As the loss will result in a stutter on wel designed codecs, keeping the losses to a relatively low level will be sufficient to guarantee customer acceptance.

The typical BER rate of Ethernet is 10^{-9} (1 billion bits without an error) and is typically 10^{-14} in optical backbones in carriers. Based on this, it is possible to estimate the loss of voice based on using a conservative BER of 10^{-8}. If we further assume 10 hops (ten segments, each with an additive BER), it is possible to calculate the timing of lost frames.

$$\frac{10^8 \text{ bits}}{(10 \text{ hops})(15000 \text{ bps} * 60\text{s/m})} = 11 \text{ minutes between occurrences}$$

Obviously, with today's networks, the typical BER will be low enough. The exception to this is the use of modem lines. A noisy modem line may have a BER of as low as 10^{-5}, yielding a lost packet every 7 seconds, which may be more noticeable to the user.

Conclusions

While it is well proven that IP networks are capable of carrying voice traffic and that IP networks will carry a significant percentage, if not the majority of the voice traffic, the challenges of delivering user acceptable voice over IP are large. Managing latency to reasonable levels so that easy human to human conversation can occur with the ease we are accustomed to is still a challenge in some situations. While great voice is achievable, the choice of coding techniques and the management of the network has a significant impact. The choice of a large frame encoding technique (such as 723), when combined with a slow transport (including additional jitter frames) will deliver performance that will be identifiable as inferior by the end user.

To achieve quality of voice, the network must support some reasonable form of QoS, delivering both managed latency as well as minimal packet losses. The lower the guarantees of latency through the net, the greater the requirement for jitter packets, increasing total latency. In addition to QoS/CoS, the management of BER is require on modem lines.

It is possible to operate voice over the IP network, but it takes significant analysis and management of the network to achieve acceptable performance. In addition to the latency and other factors analyzed in this paper, the other key factors of system reliability and telephony features must be considered in any decision. As Telecom professionals look to integrate their voice services into their IP infrastructures,

they need to make sure that the result will meet the business needs of their users and will not be compared to the two tin cans and a string.

