Voice Quality (VQ) in Converging Telephony and IP Networks

Definition

Voice quality (VQ) means different things, depending on one's perspective. On the one hand, it is a way of describing and evaluating speech fidelity, intelligibility, and the characteristics of the analog voice signal itself. On the other, it can describe the performance of the underlying transport mechanisms. However, VQ is defined as the qualitative and quantitative measure of the sound and conversation quality of a telephone call.

Overview

As the telephone industry changes—that is, as new technologies and services are added—existing technologies are applied in different ways, and new players become involved. Thus, maintaining the basic quality of a telephone call becomes increasingly complex. Although VQ has evolved over the years to be consistently high and predictable, it is now an important differentiating factor for new voiceover-packet (VoP) networks and equipment. Consequently, measuring VQ in a relatively inexpensive, reliable, and objective way becomes very important.

This tutorial discusses VQ–influencing factors as well as network impairments and their causes in a converged telephony and Internet protocol (IP) network, all from the perspective of the quality of the analog voice signal. Network performance issues will be discussed where appropriate, but the topic of VoP performance with regard to packet delivery is not covered in any real depth. VQ testing concepts, methods, and tools will als o be discussed.

Topics

1. Why Is VQ Again an Issue?
2. VQ Is Subjective
3. VQ Defined
4. Clarity
5. End-to-End Delay
6. Echo

1. Why Is VQ Again an Issue?

Traditional public switched telephone networks (PSTNs) have long since addressed the voice-quality problem by optimizing their circuits for the dynamic range of the human voice and the rhythms of human conversation. PSTNs have evolved to provide an optimal service for time-sensitive voice applications that require low delay, low jitter, and constant but low bandwidth. While these networks do not produce perfect quality, users have become accustomed to PSTN levels of VQ, and comparisons are often made in this context. That is, PSTN VQ is relatively standard and predictable.

IP networks, however, were built to support non–real-time applications, such as file transfers or e-mail. These applications are characterized by their bursty traffic and sometimes high bandwidth demand but are not sensitive to delay or delay variation.

If PSTNs and IP networks are to converge, IP networks (and the convergent points) must be enhanced with mechanisms that ensure the quality of service (QoS) required to carry voice. This point is especially important, considering that users of traditional telephone networks are used to quite high VQ standards. Providing comparable service quality in IP networks will drive the initial acceptance and success of VoP services, such as voice over IP (VoIP).

VoP technologies, particularly VoIP, have made maintaining VQ more complex by adding nonlinear compression and the need for timely packet delivery to networks not originally set up for these conditions. Transmission conditions that pose little threat to non–real-time data traffic can introduce severe problems to real-time packetized voice traffic.

Real-Time Bandwidth

Many data networks are not designed for the real-time bandwidth requirements of speech. Data networks typically have not needed to rely on streams of packets arriving at their destinations within narrow time windows (in other words, with relatively nonvarying delay). As voice signals are introduced into these networks, methods are employed to ensure this real-time transport, but VQ can still suffer if these methods do not work properly. Although real-time speech has a reasonably

low bandwidth requirement, it needs either a constant available bandwidth (for linear codecs) or direct available bandwidth (for low–bit-rate codecs). Another related condition has to do with bandwidth capacity in general. While many service providers have adequate capacity to handle the real-time voice traffic on their data networks without compromising other nonvoice traffic, linear and nonlinear voice compression techniques are still being used, particularly when voice is transmitted to the desktop. Nonlinear compression can be a major cause of reduced VQ.

Important Gateway Processes

VoP networks rely on network processes (often built into gateways) that help some voice-quality problems. For example, silence suppression is used to prevent packets from being created and transmitted during the quiet periods between spoken phrases. Also, echo cancellers are needed to eliminate echo that becomes perceptible when delay is introduced. If these kinds of processes do not work properly, VQ suffers.

Packet Loss

Packet network applications compensate for packet loss by retransmitting lost packets through the use of transmission control protocol (TCP). Data applications such as file transfers and e-mail are less sensitive to the time it takes for this to occur, but real-time voice traffic cannot tolerate this delay. In addition, VoIP networks use connectionless transfer protocols such as user datagram protocol (UDP) that do not guarantee delivery at all. Lost packets mean lost voice information.

Delay

The time it takes for a voice signal to be digitized, packetized, transmitted, routed, and buffered contributes to the delay experienced by a user. This delay can interfere with normal conversations and can exacerbate existing problems on the network such as echo.

Nonlinear Codecs

As alluded to above, an important reason to measure VQ is the continued development and use of nonlinear perceptual codecs. Nonlinear perceptual codecs compress voice such that the perceptually important information is preserved, but not necessarily the voice waveform. In other words, these codecs preserve how the voice sounds without preserving all of the frequency spectrum information. This nonlinear compression renders many traditional speech

measurements less useful; thus, the need for new measurement techniques emerges.

2. VQ Is Subjective

Generally speaking, VQ can be expressed (and therefore measured) primarily with respect to the talker and the listener who experience it. VQ should be approached from an end-to-end perspective; that is, regardless of the systems, devices, and transmission methods used, any voice-quality metric should be expressed in the context of the user's experience. But the end-to-end aspect of VQ is accompanied by the inherent subjective nature of this type of qualitative evaluation. What a listener considers high quality (or, for that matter, low quality) is influenced by expectations, context and environment, physiology, and mood.

These end-to-end and subjective characteristics of VQ make measuring it an interesting challenge. Testing methods and equipment must be able to address these issues directly, as well as provide data about the reasons for specific VQ–measurement results.

3. VQ Defined

At this point, VQ must be clearly defined before any discussion of its characteristics and components can proceed. Many factors can influence one's perception of the quality of a telephone call, ranging from the ease or difficulty in placing the call to the quality of the sound in the earpiece. At a very high level, basic telephone call quality is made up of three fundamental components:

- service quality
- sound quality
- conversation quality

Table 1 describes these three components in more detail.

Table 1. Details of Service Quality, Sound Quality, and Conversation Quality

Service Quality	Sound Quality	Conversation Quality
• offered services—such as calling card, 1-800/900 services, follow-me, and voice mail • availability of users in other countries or regions • network availability—down time, busy signals • reliability—such as dropped calls or wrong number • price	• loudness • distortion • noise • fading • crosstalk	• loudness distortion noise • fading • crosstalk • echo • end-to-end delay • silence suppression performance • echo canceller performance

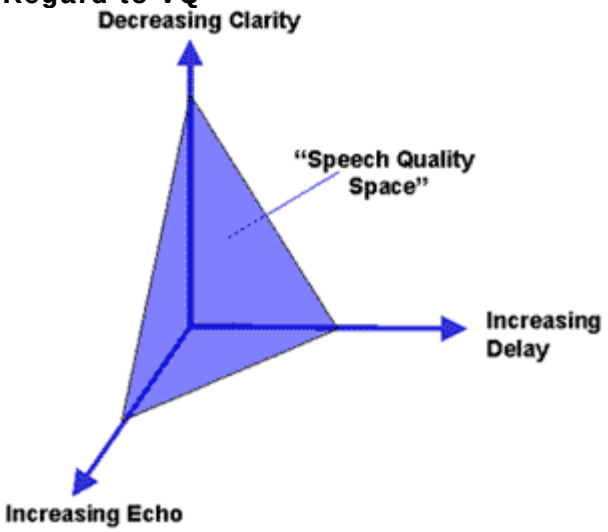
The components in *Table 1* impact perceived quality whether the telephone call occurs over traditional PSTN lines, emerging VoIP networks, or a hybrid of both, and they often depend on each other when it comes to a user's ultimate judgment of the quality of a given telephone call. For example, questionable sound quality is frequently tolerated, ignored, or unnoticed when service quality is very high. Users of cell phones or overseas satellite links tolerate or ignore sound-quality problems because of the usefulness of the call itself. Another example involves conversation quality. When a perceptible time lag between phrases spoken by talker and listener exists, many users perceive such a time lag as a sound-quality or service-quality problem.

Many aspects of service quality are closely tied to service provider business issues and network provisioning decisions and less closely connected to the technical aspects of network performance and network device operation. Yet, sound quality and conversation quality seem to be closely related and quite dependent on the details of network deployment and performance. For this reason, VQ is defined as the qualitative and quantitative measure of the sound and conversation quality of a telephone call.

Given the previous definition of VQ, three elements (shown in *Figure 1*) emerge as the primary factors affecting VQ, particularly in the case of networks using VoP or VoIP technologies.

- **clarity**—a voice signal's fidelity, clearness, lack of distortion, and intelligibility
- **end-to-end delay**—the time it takes a voice signal to travel from talker to listener
- **echo**—the sound of the talker's voice returning to the talker's ear

Figure 1. Relationship among Clarity, Delay, and Echo with Regard to VQ



The relationships among clarity, delay, and echo can be quite complex, as shown in *Figure 1*. If you think of VQ as a single plotted point in the graph, you can see that VQ improves as the point is plotted closer to the intersection of the three lines. In other words, as the distance between the voice-quality point and the intersection increases, VQ decreases.

Perception of One Aspect Affects Perception of Overall VQ

One of the main reasons clarity, delay, and echo are grouped together is that many users will report unacceptable VQ if only one aspect of VQ is unacceptable. For example, users rarely distinguish between distortion and annoying echo; they simply report unacceptable call quality. While network equipment manufacturers and service providers often need to distill voice-quality issues into distinct areas, users do not.

Clarity and Delay Are Orthogonal Aspects of VQ

Distortion and fidelity are independent of end-to-end delay in that a voice signal can experience significant delay yet sound very good. The converse is also true: A voice signal can sound very distorted but travel end-to-end too quickly to be perceived by the user. For VQ to be perceived as acceptable, however, clarity must be reasonably good, and delay must be reasonably short.

Echo Depends on Delay and Affects Clarity

As this tutorial later explains in detail, echo is perceptible only when network delay (defined in this case as the roundtrip delay from the talker to the point of echo) is above a certain threshold. In other words, echo coming from any remote point in the network will not be heard unless it is delayed long enough to be audibly separate from the original spoken phrase. Similarly, perceived clarity is often negatively impacted by audible echo, even though the level of distortion in the echo signal itself may be quite low.

Important Note—*Figure 1* provides a conceptual model only. It is true that VQ is influenced by clarity, delay, and echo, and that the relationships between them are generally shown by the graph. However, no known mathematical relationship exists that can be used to derive a single VQ number or a vector the length of which uniquely quantifies VQ. Any representation of VQ, whether it is for individual devices or VoP systems, must include at least a clarity and a delay component and optionally an echo component.

Breaking VQ into three distinct areas such as clarity, delay, and echo make evaluating VQ a manageable process. While there are many aspects of VoP telephony that can be measured—and some will be discussed later in this tutorial—clarity, delay, and, to a lesser degree, echo form the basis of most voice-quality concepts and test techniques.

4. Clarity

In the context of voice-quality testing, clarity describes the perceptual fidelity, the clearness, and the nondistorted nature of a particular voice signal. Clarity can also be described as speech intelligibility, indicating how much information can be extracted from a conversation. However, it is possible to understand what is said during a voice conversation but still experience poor clarity. For example, voice that is distorted and not easily heard can still be understood.

The subtle, yet important, distinction between clarity and intelligibility illustrates just one part of the complexity involved when attempting to quantify VQ. Clarity, and a person's evaluation of it, depends on numerous factors. For example,

certain frequency bands are more important for perceived clarity than others. Human listeners are more likely to find that distortion or attenuation in the 1,000-to-1,200-Hz band decreases clarity and intelligibility more than distortion or attenuation in the 250-to-800-Hz band. Another example is that complete sentences are usually much better understood as a result of the logical word flow in a sentence (and, therefore, perceived as having higher clarity) than a sequence of unrelated words, even if the random word sequence is less distorted.

What are the influencing factors for clarity in a combined IP/PSTN telephony network? *Figure 2* shows a typical implementation.

Figure 2. Example of a Combined PSTN/VoIP Network

Each of the following network components has an impact on voice clarity:

- **PSTN telephone**—influences clarity through the quality of its loudspeaker and microphone, the loudness of the transmitted and received signal, and the acoustic echo generated between the loudspeaker and microphone

- **PSTN network**—uses digital voice transmission for greater efficiency in the backbone; digitizing analog voice signals often affects voice clarity

- **VoIP gateway**—interconnects the PSTN with the IP network using voice and signaling schemes

Gateway components affecting clarity are the speech codec, silence suppression mechanism, and comfort noise generator. The IP network, even without active voice components, affects clarity through its tendency to lose packets and add extensive jitter to voice packet delivery.

The H.323 terminal (an application on a PC or an IP telephone) also affects the clarity through its speech codec, silence suppression mechanism, and microphone and loudspeaker quality.

Packet Loss

Packet loss is not uncommon in IP networks. As the network, or even some of its links, becomes congested, router buffers fill and start to drop packets. Another cause can be route changes as a result of inoperative network links. An effect

similar to packet loss occurs when a packet experiences a large delay in the network and arrives too late to be used in reconstructing the voice signal.

For non–real-time applications, such as file transfers, packet loss is not critical. Packet protocols provide retransmission to recover dropped packets. However, in the case of real-time voice information, packets must arrive within a relatively narrow time window to be useful to reconstruct the voice signal. Retransmissions in the voice case would add extensive delay to the reconstruction and would cause clipping or unintelligible speech.

To avoid packet loss for real-time applications, mechanisms are required in the IP network to assure minimum throughput for selected applications. These mechanisms minimize packet loss and delay for higher-priority traffic, such as voice. Various router mechanisms can be used to meet this objective. These include prioritization schemes, such as weighted fair queuing (WFQ), and router flow control mechanisms, such as the Internet Engineering Task Force's (IETF) multiprotocol label switching (MPLS) tagging scheme or the use of type-of-service (ToS) bits in the IP header. To use these mechanisms, a network administrator must decide what priority and resources to provide for each specific service class and configure the network accordingly. A more dynamic alternative for assigning resources is the resource reservation protocol (RSVP), which permits a terminal or voice gateway to request a specific IP QoS.

Regardless of which is used, a deeper problem remains. QoS is defined on an end-to-end basis and therefore requires that sufficient network resources be provided throughout the network path. This is not an overwhelming issue for an enterprise network or a single Internet service provider (ISP) environment, where all resources can be administered through one network manager. However, it is almost impossible to administer when multiple ISPs or service providers are involved, as is the case in virtually every national or international long-distance call. In addition, this fulfillment of QoS assumes that all routers in the network are equally capable of identifying voice traffic and providing the required network resources. This is still the exception rather than the rule in today's IP networks because standards for many of these mechanisms have not been finalized and implemented by equipment manufacturers.

Speech Codecs

A speech codec transforms analog voice into digital bit streams, and vice versa. In addition, some speech codecs also use compression techniques, removing redundant or less important information to reduce the amount of transmission bandwidth required. In other words, many codecs compress voice signals by preserving only those parts of the voice signal that are perceptually important. In the context of VQ testing, the phrase *perceptually important* refers to those parts of the audio signal that have the largest impact on a human's perception of the

signal, particularly if those parts are distorted or omitted. Perceptual importance is determined via an understanding of human physiology and cognitive psychology. Consequently, and depending on the type of codec used, the actual voice waveform may not be reproduced at the receiving end of a VoP conversation. Codecs such as G.711 can be thought of as linear because they come very close to reproducing the waveform. However, low–bit-rate codecs such as G.729 and G.723.1 try to reproduce the subjective sound of the signal rather than the shape of the speech waveform and are therefore generally thought of as nonlinear.

Essentially, compression is a balancing act between VQ, local computation power, and the delay and network bandwidth required. The greater the bandwidth reduction, the higher the computational cost of the codec for a given level of perceived clarity. In addition, greater bandwidth savings generally cause higher computational delay and therefore significantly increase the end-to-end delay. The network planner must make an informed trade-off between bandwidth, VQ, and delay.

A codec's effect on VQ is also influenced by packet size, packet loss, and any error-correction mechanisms used by the codec itself.

Other Factors Affecting Clarity

Other factors affect voice clarity. Some are the kinds of things one would expect in any audio or digital transmission channel, and others are specific to VoP networks.

Noise

All noise, regardless of its source, has the potential to reduce the clarity of a voice signal. Noise can originate from analog lines or from bit errors on data transmission lines. If it is introduced prior to the voice signal being digitized, it will be faithfully reproduced by the codec, if possible. Noise introduced after a voice signal has been converted back to analog will further distort the voice signal.

Voice Activity Detectors

Discussed in more detail later, voice activity detectors (VADs) can introduce clarity degradations by inadvertently removing (clipping) parts of speech utterances.

Echo

Speech that is echoed back to the speaker such that it is perceived during conversations can have a significant (albeit indirect) effect on perceived clarity. For example, if you can hear your own voice echoed back to you as you are talking, this can be annoying and perhaps disruptive.

External Environmental Factors

It is possible to have excellent audio quality on a telephone speaker, but, as a result of room noise, end-user mood, end-user expectations, and other intangible factors, the audio quality could still be perceived as unacceptable. This affects testing methods and makes true subjective testing with human subjects more difficult.

5. End-to-End Delay

Delay is the time required for a signal to traverse the network. In a telephony context, end-to-end delay is the time required for a signal generated at the talker's mouth to reach the listener's ear. End-to-end delay is the sum of the delays at the different network devices and across the network links through which voice traffic passes. Many factors contribute to end-to-end delay.

PSTN Delay

PSTN delay is most often the result of transmission delay on long-distance trunks. The delay is especially high when satellite links are involved (a geostationary satellite link has a transmission delay of about 250 milliseconds). In addition, switching delay in network nodes is relatively small when compared to transmission delay. In the vast majority of cases, PSTNs exhibit relatively low delay, which is primarily a function of transmission distance.

IP Network Delay

IP network delay is primarily determined by the buffering, queuing, and switching or routing delay of IP routers.

Packet Capture Delay

Packet capture delay is the time required to receive the entire packet before processing and forwarding it through the router. This delay is determined by the packet length and transmission speed. Using short packets over high-speed trunks can easily shorten the delay but potentially decrease network efficiency.

Switching/Routing Delay

Switching/routing delay is the time the router takes to switch the packet. This time is needed to analyze the packet header, check the routing table, and route the packet to the output port. This delay depends on the architecture of the route engine and the size of the routing table. New IP switches can significantly speed up the routing process by making routing decisions and forwarding the traffic via hardware as opposed to software processing.

Queuing Time

Due to the statistical multiplexing nature of IP networks and to the asynchronous nature of packet arrivals, some queuing (thus, delay) is required at the input and output ports of a packet switch. This delay is a function of the traffic load on a packet switch, the length of the packets, and the statistical distribution over the ports. Designing very large router and link capacities can reduce but not completely eliminate this delay.

VoIP Device Delay

VoIP gateways and VoIP terminals also contribute significantly to end-to-end delay as a result of signal processing at both the sending and the receiving sides of the link. This processing includes the time codecs required to encode the analog voice signal into a digital signal and to decode the digital voice signal back to analog. Some codecs also compress the voice signal, thereby extracting redundancy, which further increases delay due to the necessary computation. The higher the compression, the more voice bits must be buffered. The more complex the processing, the longer this delay component.

At the transmit side, packetization delay is another factor. Packetization delay is the time needed to fill a packet with voice data. The longer the packet size, the more time is required. Using shorter packet sizes can shorten this delay, but this will decrease network efficiency because more packets have to be sent, each with nearly redundant header information.

On the receive side, voice packets must be delayed to compensate for variation in packet interarrival times (also known as jitter). Even packets generated with constant spacing in time will generally arrive at the receiver with a randomly spaced distribution as a result of the different buffering and queuing times packets experience and the varying transmission routes in the IP network. Jitter smoothing using jitter buffers is required because speech codecs need a constant flow of data without gaps. Delay caused by jitter buffering can be reduced by designing a network with less jitter at each node, with as few nodes as possible. The size of the jitter buffer itself can also be optimized, and many modern jitter buffers will adapt to existing jitter in order to keep their size as small as possible.

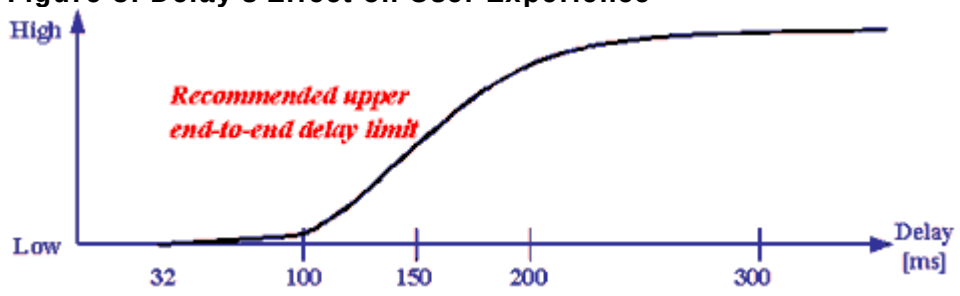
Using mechanisms that prioritize voice traffic over other traffic in the network can significantly reduce jitter.

No matter how well VoIP devices and networks are designed, a fundamental delay exists that simply cannot be eliminated. That is, some delay will always be introduced as a result of the physical limits of packetization, processing time, and propagation time. Consider an example in which IP packets each contain 20 ms of voice data. It takes 20 ms to fill (packetize) the very first packet. Assume that the codec imposes a further delay of 10 ms for framing and computation. A jitter buffer size of at least one frame (20 ms) can be expected at the receive end of the link. Add transmission times, router processing times, and other miscellaneous sources of delay, and one arrives at 60 ms. It is clear that 30 ms (packetization plus codec computation/framing) is a fundamental lower limit on end-to-end delay in this example. The delay cannot be made any smaller.

How Much Delay Is Too Much?

How much delay is too much? Delay does not affect VQ directly but instead affects the character of a conversation. Below 100 ms, most users will not notice the delay. Between 100 ms and 300 ms, users will notice a slight hesitation in their partner's response. This hesitation can affect how each listener perceives the mood of the conversation. In this situation, conversations can seem cold. Interruptions are more frequent, and the conversation gets out of beat. Beyond 300 ms, the delay is obvious to the users, and they start to back off to prevent interruptions. At some point, conversation is virtually impossible. Obviously, shorted delay results in better conversation quality and in better perceived overall VQ (see *Figure 3*).

Figure 3. Delay's Effect on User Experience

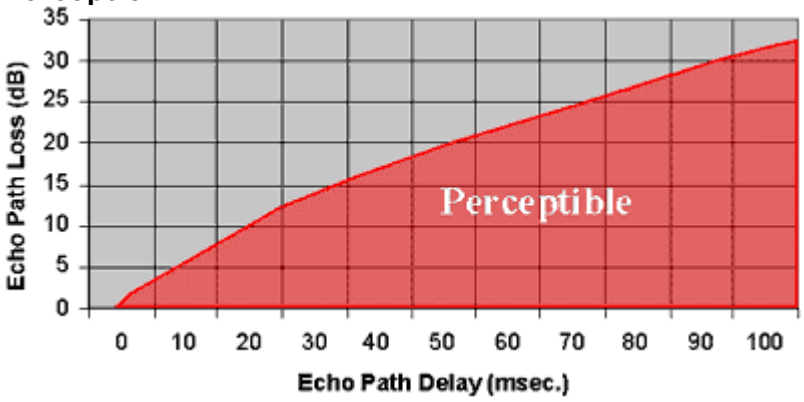


An interesting phenomenon related to delay has to do with echo. Generally speaking, echo exists in many PSTNs, but because of where echo originates and the relatively short end-to-end delays in these networks, echo is often not noticed. However, when VoIP levels of delay are introduced, echo often becomes noticeable.

6. Echo

From a telephony perspective, echo is the sound of the talker's voice returning to the talker's ear via the telephone's speaker. In other words, echo occurs when the talker's voice signal leaks from the transmit path back into the receive path. If the time between the original spoken phrase and the returning echo is short (25 to 30 ms), or if the echo's level is very low (approximately -25 dB), it probably will not cause any annoyance or disruption to voice conversations. In many PSTN environments, echo exists but occurs so close in time to the source speech that it is very rarely an issue (exceptions can include the echo you might hear while participating in an overseas satellite call). In fact, a special type of echo with a delay of about 28 ms (often called side tone) is desired because it is reassuring for a talker to hear his or her own voice in the earpiece while speaking. It is when the echo that is loud enough to be heard passing through networks with enough delay to be perceptible to the speaker (usually around 30 ms and above) that the quality of a voice call becomes problematic.

Figure 4. Relationship between Echo Levels, Delay, and Perception



What Causes Echo?

In the vast majority of cases, echo is caused by an electrical mismatch between analog telephony devices and transmission media in a portion of the network called the tail circuit. A tail circuit is everything connected to the PSTN side of a packet voice gateway: all the switches, multiplexers, cabling, private branch exchanges (PBXs), or everything between the voice gateway and the telephone. Specifically, this electrical mismatch occurs between a four-wire ear-and-mouth (E&M) trunk line or digital transmission channel and a two-wire foreign exchange office (FXO) line. This local-loop, four-wire–to–two-wire conversion happens in a device known as a hybrid that separates send-path and receive-path signals in order to carry them on separate pairs of wires or transmission channels. Because the methods used to separate send signals from receive signals

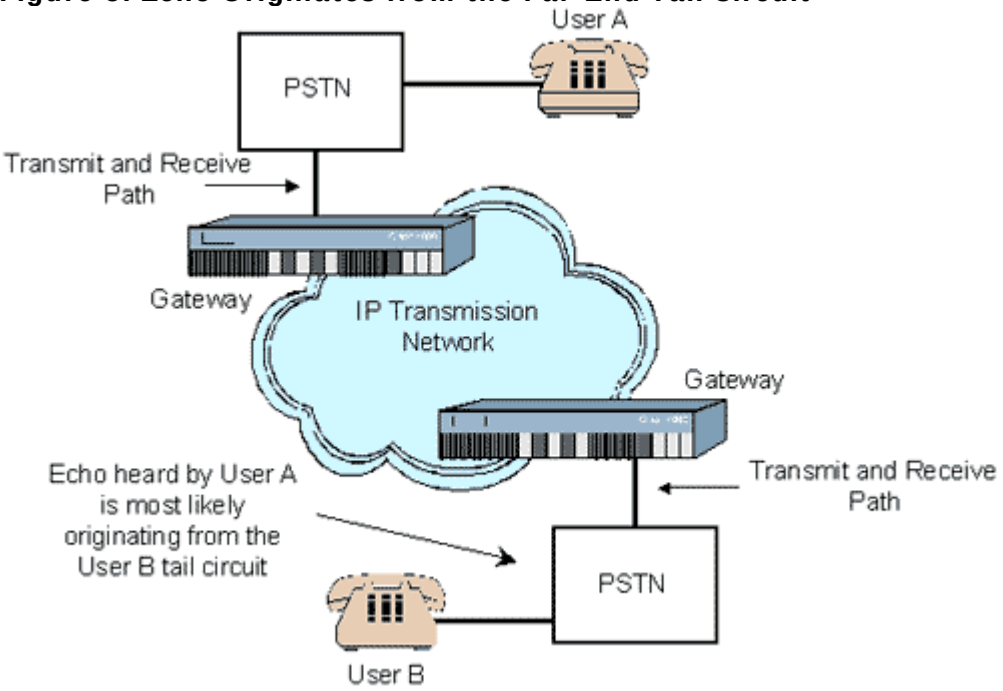
are often not ideal, some of the received signal leaks onto the send path and is perceived as echo.

Another cause of echo can be acoustic coupling problems between a telephone's speaker and microphone. For example, consider the handset of a traditional telephone or the hands-free set of a speaker telephone or personal computer (PC) terminal. This is called acoustic echo.

What Makes Echo Perceptible?

As mentioned, roundtrip delay introduced into the voice path by VoP networks such as VoIP can often cause existing echo originating from an analog tail circuit to become perceptible and even annoying. Echo that originates between an individual's telephone and the PSTN central office (CO) is not perceptible because it returns to one's ear too quickly. Even echo from the far-end tail circuit usually returns quickly enough or is attenuated enough to not be heard. VoP network components, however, introduce into the voice path a fundamental and unavoidable end-to-end delay that often exceeds the 32-ms threshold mentioned earlier. If echo is produced in the far-end PSTN analog tail circuit, at least twice this delay (known as roundtrip delay) will pass before the echo reaches the near-end talker's ear. Thus, even attenuated echo can become perceptible. As near-end echo will not be heard, one can often correctly conclude that any perceptible echo originates from the far-end tail circuit. *Figure 5* illustrates this point.

Figure 5. Echo Originates from the Far-End Tail Circuit



How Do Networks Deal with Echo?

To deal with unwanted echo, functional components known as echo cancellers are deployed in the local exchange, the VoIP gateway, or the VoIP PC terminal, usually as close as possible to the tail circuit that causes the echo. Referring back to *Figure 5*, an echo canceller next to the hybrid on User B's side of the network faces out at User B and cancels the echo of User A's voice that would otherwise be heard by User A.

To eliminate unwanted echo, echo cancellers form a mathematical model of the tail circuit they monitor. They then use this model (along with representations of the signal likely to be echoed, such as User A's voice) to estimate the expected echo. This estimated echo is then subtracted from the speech originating on the tail circuit side of the echo canceller (User B's voice). Thus, normal speech is allowed to pass through the echo canceller, but echoes of received speech are removed.

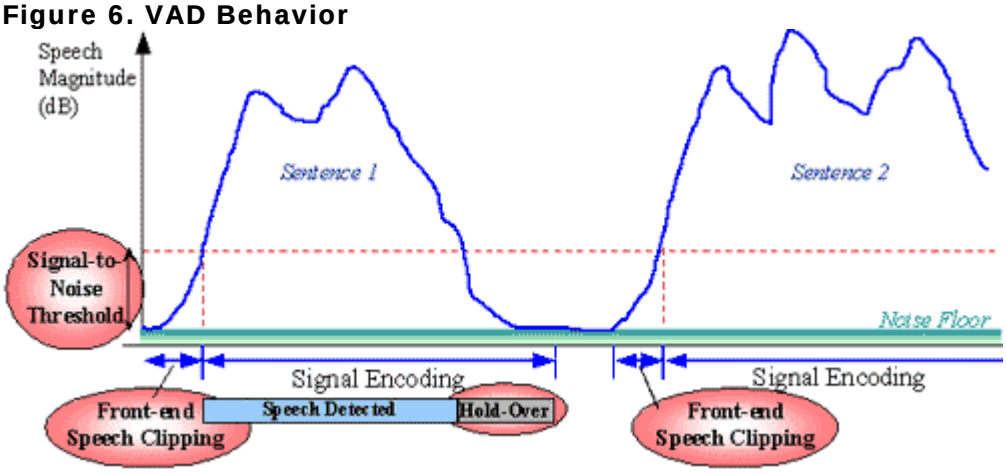
An interesting characteristic of most modern echo cancellers is their ability to adapt to signal and tail circuit conditions. In other words, at the start of a voice call, echo cancellers take some finite time to converge on the echo estimate that will be subtracted from far-end speech signals. For example, at the beginning of a VoIP telephone call that terminates through an analog tail circuit, echo may be perceptible but quickly diminishes as the echo canceller converges. A point of failure (or poor performance) for many echo cancellers is when the talker at the far end interrupts the near-end talker (a condition known as doubletalk). Echo cancellers work with the assumption of a linear and time-invariant tail circuit. Doubletalk, however, causes the tail circuit to appear to be nonlinear, resulting in echo canceller divergence (in other words, its echo estimate becomes more inaccurate). In this case, the interrupting speech can become distorted.

7. Silence Suppression and Comfort Noise Generation

To use bandwidth more efficiently, VoIP networks employ functionality known as silence suppression or voice activity detection. A VAD is a component of a voice gateway or terminal that suppresses the packetization of voice signals between individual speech utterances, such as during the silent periods in a voice conversation. VADs generally operate on the send side of a gateway and can often adapt to varying levels of noise versus voice. Thus, similar to adaptive jitter buffers and echo cancellers, VADs can converge on appropriate thresholds to optimize their performance for a given voice conversation. As human conversations are essentially half-duplex in the long term, the use of a VAD can realize approximately 50 percent reduction in bandwidth requirements over an

aggregation of channels. *Figure 6* depicts the behavior of a VAD and its parameters.

Figure 6. VAD Behavior



While a VAD's performance does not affect clarity directly, if it is not operating correctly, it can certainly decrease the intelligibility of voice signals and overall conversation quality. Excessive front-end clipping (FEC), for example, can make it difficult to understand what is said. Excessive holdover time (HOT) can reduce network efficiency, and too little holdover time can cause speech utterances to feel choppy and unconnected when cutting in—even in short speech pauses.

Complementary to the transmit-side VAD, a comfort noise generator (CNG) is a receive-side device. During periods of transmit silence, when no packets are sent, the receiver has a choice of what to present to the listener. Muting the channel (playing absolutely nothing) gives the listener the unpleasant impression that the line has gone dead. A receive-side CNG generates a local noise signal that it presents to the listener during silent periods. The match between the generated noise and the true background noise determines the quality of the CNG.

8. Testing VQ

Traditionally, VQ testing techniques involved comparing waveforms on a screen and measuring signal-to-noise ratio (SNR) and total harmonic distortion (THD), among others. These and other linear measurements are useful only in certain cases because they assume that changes to the voice waveform represent unwanted signal distortion. These testing methods also assume that telephony circuits are essentially linear. However, in VoIP and other VoP networks, particularly when low–bit-rate speech codecs such as G.729 and G.723.1 are used, neither waveform preservation nor circuit linearity can be assumed. These codecs try to reproduce the subjective sound of the signal rather than the shape of the speech waveform, rendering traditional testing methods more or less ineffective. And, as discussed before, the bursty and time-insensitive nature of packet

networks exposes the need for other testing methods as well. Finally, because of their heightened importance, the performance of echo cancellers, voice activity detectors, and other processes must be tested directly.

Measuring Clarity

Because of the inherent subjective nature of voice-quality testing, one obvious method to quantify quality is to have relatively large numbers of human listeners rate VQ as part of a controlled and well-defined test process. The advantage of this method is that clarity evaluations are derived directly from the individuals who will experience the voice call. Another advantage is the statistical validity provided by numerous evaluators. This, in fact, has been the method used for many years and is defined as mean opinion score (MOS) in International Telecommunications Union–Telecommunications Standardization Sector (ITU–T) specification P.800.

In spite of its obvious advantages, MOS has one distinct and significant disadvantage: it is expensive both in terms of time and effort. Parading tens or even hundreds of human listeners through a VQ test lab to evaluate the performance of a single set of telephony devices or software products would seem not to be the most efficient method. Experimental conditions must be tightly controlled, test results must be carefully analyzed, and the whole process must be repeated when new equipment or voice-encoding methods are developed. So how can clarity be measured in a repeatable, objective, and reasonably inexpensive manner?

One method is perceptual speech-quality measurement (PSQM), defined by ITU–T Recommendation P.861. Originally created to evaluate speech codecs, the PSQM algorithm provides a method by which speech within the voice bandwidth of 300 to 3400 Hz can be objectively measured for distortion, the effects of noise, and overall perceptual fidelity. Simply put, PSQM is an automated human listener.

PSQM evaluates the quality of voice signals in much the same way that nonlinear codecs encode and decode voice signals. It evaluates whether a particular voice signal is distorted according to what a human listener would find annoying and distracting. To do this, PSQM takes a clean voice sample and compares it to a more or less distorted version using a complex weighting method that takes into account what is perceptually important—for example, the physiology of the human ear and cognitive factors related to what human listeners are likely to notice. PSQM provides a relative score that indicates just how different the distorted signal is with respect to the original from the perspective of the human listener via the algorithm. PSQM shows whether the distorted voice signal is better or worse than the original. Because of the way PSQM works, this distortion

score corresponds very closely to how a statistically large number of human listeners would react in the same test situation (for example, MOS).

PSQM was originally and specifically designed to measure the perceived quality of voice as impacted by voice compression codecs. However, certain impairments, such as packet loss, introduced by data network transmission, are not adequately reflected in PSQM scores. Therefore, an enhanced version of PSQM, known as PSQM+, was developed to correlate more to MOS scores in the presence of network impairments.

Another important model for measuring perceived clarity that has recently been developed is the perceptual analysis measurement system (PAMS). PAMS uses a similar perceptual model as PSQM and shares the purpose of providing a repeatable, objective means for measuring perceived VQ. PAMS uses a different but effective signal-processing model than PSQM and produces different types of scores. It provides a listening quality score and a listening effort score, both of which correlate to MOS scores and are on the same 1 to 5 scale.

Measuring Delay

As mentioned previously, end-to-end delay can have a significant effect on the quality of a voice conversation. Remember that delay does not affect the sound of a voice conversation but rather the rhythm and feel of the conversation. There are two primary ways to measure delay in a VoP environment: acoustic packet Internet groper (PING) and maximum length sequence (MLS) normalized cross-correlation. Ideally, both methods should be used to ensure that delay measurements are accurate and consistent, because delay can change in a dynamic VoIP environment.

Acoustic PING

Acoustic PING is just what might be expected from the name. A narrow audio spike is transmitted from one end of the audio channel to the other, and the time it takes to travel end-to-end is measured. This simple method, however, is susceptible to noise and attenuation because the actual spike can be masked by other noise spikes on the channel or strongly attenuated such that it will not be detected. In addition, the relative narrowness of the spike makes it vulnerable to packet loss (that is, the spike itself may only be one or two packets long). Acoustic PING should be augmented with other methods to ensure accuracy and consistency.

MLS Normalized Cross-Correlation

It is possible to use digital signal processing (DSP) techniques in which a special test signal is transmitted onto the system under test, and the received signal and original test signal are then analyzed together to determine end-to-end delay. This method, called MLS normalized cross-correlation, uses a test signal that sounds very much like white noise and, in fact, exhibits many of the same characteristics. Unlike white noise, MLS noise is a repeatable and predictable noise pattern that enhances analysis calculations.

Using this method, the delay value calculated is actually a subset of the information obtained. Delay calculated in this way is much more accurate, provides higher resolution results, and is more noise resistant than acoustic PING methods.

Measuring Echo

Several aspects exist in measuring echo. Initially, one may need to characterize echo levels and echo delay. In addition, one may need to measure how well echo cancellers deal with echo. Finally, one may find it useful to evaluate just how annoying echo is to users of the telephony system.

Echo Characterization

Characterizing echo almost always involves measuring echo levels and the length of time it takes for an echo to return to the talker. The amount that echo is attenuated before it reaches the talker's ear is often referred to as echo return loss (ERL). ERL is an important parameter because many echo cancellers are unable to deal with echo that has not been attenuated by some amount. In addition, the time that passes before echo is heard (known as echo delay) must be within a certain window to allow echo cancellers of reasonable processing power to be effective. ERL and echo delay could be considered tail circuit design parameters and certainly have an impact on the type and configuration of the echo canceller used. It is also useful to know these echo characteristics so that decisions can be made as to whether an echo canceller is the right solution or whether tail circuit redesign is needed to solve a given echo problem.

Echo Cancellers

Measuring the actual echo that may or may not exist on the network should also be accompanied by a direct evaluation of echo canceller performance. To do this, test personnel must often simulate tail circuit behavior (echo delay, ERL, and frequency response) and be able to control various aspects of that behavior.

Important parameters to measure when evaluating an echo canceller are as follows:

- **convergence time**—the time required for an echo canceller to adapt to the local tail circuit and provide adequate echo reduction
- **cancellation depth**—the reduction in echo strength achieved (measured in decibels)
- **doubletalk robustness**—a measure of whether the echo canceller loses its cancellation ability under conditions of simultaneous talking from both ends of the connection.

Perceived Annoyance Caused by Echo (PACE)

One very useful measure of the overall quality of a voice connection is to what extent echo is perceived as a problem by the users of that connection. Similar to voice clarity, this point is essentially a subjective judgment and requires very special measurement algorithms to achieve an objective, reliable, and repeatable result. The ITU–T has defined methods by which echo characteristics can be measured. G.165 is an algorithm that uses white noise, while G.168 uses speech frequency test signals. However, these methods seem best suited for laboratory testing and are not suitable for low–bit-rate codecs in which the waveform of the voice signal is not always preserved. However, by using an objective, perception-based algorithm such as PSQM or PAMS, it is possible to evaluate the effect echo has on a user's perception of quality in both a test lab environment and in deployed VoP networks.

Measuring VADs

Another VoP device component with a performance level that can be measured directly is the voice activity detector. The goal in this case would be to measure FEC and HOT and perhaps CNG match. Ideally, it is necessary to produce test signals that simulate the conditions with which the VAD will be presented—or, rather, predominant voice levels accompanied by low-level noise. One successful method is to produce a hybrid test signal comprised of a finite voice band noise burst accompanied by a very–low-level and distinct tracer dye tone. The noise burst (which simulates a speech utterance) and tracer dye tone are sent through a network containing a VAD and received at the other end. The received noise burst width is compared to the original to determine the FEC, and the tracer dye tone is used to detect when the VAD closes (HOT).

Self-Test

1. Transmission conditions that pose little threat to non–real-time data traffic can introduce severe problems to real-time packetized voice traffic, such as important gateway processes, packet loss, delay, nonlinear codecs, and _____.

    a. real-time bandwidth

    b. low bandwidth

2. VQ is defined as the qualitative and quantitative measure of the sound and conversation quality of a telephone call.

    a. true

    b. false

3. VQ is comprised of three factors: clarity, delay, and _____.

    a. jitter

    b. signal-to-noise ratio

    c. echo

    d. silence suppression

4. Echo coming from any remote point in the network will not be heard unless it is delayed long enough to be audibly separate from the original spoken phrase.

    a. true

    b. false

5. Clarity describes the perceptual fidelity, nondistorted nature, and _____ of a particular voice signal.

    a. loudness

    b. delay

    c. clearness

    d. randomness

6. Packet loss can be caused by congestion, router changes as a result of inoperative network links, and occasions on which a packet experiences a large delay in the network and arrives too late to be used in reconstructing the signal.

    a. true

    b. false

7. IP network delay is comprised of the following: switching/routing delay, queuing time, and _____.

    a. packet capture delay

    b. noise

8. Delay affects the character of a conversation, and between _____ ms, users will notice a slight hesitation in their partner's response.

    a. 25 and 50

    b. 50 and 100

    c. 100 and 300

9. A special type of echo with a delay of about _____ ms (often called side tone) is desired because it is reassuring for a talker to hear his or her own voice in the earpiece while speaking.

    a. 28

    b. 50

    c. 100

10. Echo is caused by an electrical mismatch between analog telephony devices and transmission media in a portion of the network called the tail circuit.

    a. true

    b. false

11. When the echo is loud enough to be heard passing through networks with enough delay to be perceptible to the speaker (usually around 30 ms and above), the quality of a voice call becomes problematic.

    a. true

b. false

12. VADs generally operate on the receive side of a gateway and can often adapt to varying levels of noise versus voice.

    a. true

    b. false

13. Traditionally, voice-quality testing techniques involved comparing waveforms on a screen and measuring _____, among others. These and other linear measurements are useful only in certain cases because they assume that changes to the voice waveform represent unwanted signal distortion.

    a. jitter

    b. SNR and THD

    c. clarity

14. Important parameters to measure when evaluating an echo canceller are convergence time, cancellation depth, and _____.

    a. frequency of response

    b. doubletalk robustness

    c. echo delay

15. The goal of measuring a VAD would be to measure FEC and HOT and perhaps CNG match.

    a. true

    b. false

Correct Answers

1. Transmission conditions that pose little threat to non–real-time data traffic can introduce severe problems to real-time packetized voice traffic, such as important gateway processes, packet loss, delay, nonlinear codecs, and _____.

    a. real-time bandwidth

b. low bandwidth

See Topic 1.

2. VQ is defined as the qualitative and quantitative measure of the sound and conversation quality of a telephone call.

a. true

b. false

See Topic 3.

3. VQ is comprised of three factors: clarity, delay, and _____.

a. jitter

b. signal-to-noise ratio

c. echo

d. silence suppression

See Topic 3.

4. Echo coming from any remote point in the network will not be heard unless it is delayed long enough to be audibly separate from the original spoken phrase.

a. true

b. false

See Topic 3.

5. Clarity describes the perceptual fidelity, nondistorted nature, and _____ of a particular voice signal.

a. loudness

b. delay

c. clearness

d. randomness

See Topic 4.

6. Packet loss can be caused by congestion, router changes as a result of inoperative network links, and occasions on which a packet experiences a large delay in the network and arrives too late to be used in reconstructing the signal.

    a. true

    b. false

    See Topic 4.

7. IP network delay is comprised of the following: switching/routing delay, queuing time, and _____.

    a. packet capture delay

    b. noise

    See Topic 5.

8. Delay affects the character of a conversation, and between _____ ms, users will notice a slight hesitation in their partner's response.

    a. 25 and 50

    b. 50 and 100

    c. 100 and 300

    See Topic 5.

9. A special type of echo with a delay of about _____ ms (often called side tone) is desired because it is reassuring for a talker to hear his or her own voice in the earpiece while speaking.

    a. 28

    b. 50

    c. 100

    See Topic 6.

10. Echo is caused by an electrical mismatch between analog telephony devices and transmission media in a portion of the network called the tail circuit.

    a. true

b. false

See Topic 6.

11. When the echo is loud enough to be heard passing through networks with enough delay to be perceptible to the speaker (usually around 30 ms and above), the quality of a voice call becomes problematic.

    a. true

    b. false

    See Topic 6.

12. VADs generally operate on the receive side of a gateway and can often adapt to varying levels of noise versus voice.

    a. true

    b. false

    See Topic 7.

13. Traditionally, voice-quality testing techniques involved comparing waveforms on a screen and measuring _____, among others. These and other linear measurements are useful only in certain cases because they assume that changes to the voice waveform represent unwanted signal distortion.

    a. jitter

    b. SNR and THD

    c. clarity

    See Topic 8.

14. Important parameters to measure when evaluating an echo canceller are convergence time, cancellation depth, and _____.

    a. frequency of response

    b. doubletalk robustness

    d. echo delay

See Topic 8.

15. The goal of measuring a VAD would be to measure FEC and HOT and perhaps CNG match.

    a. true

    b. false

    See Topic 8.

Glossary

CNG
comfort noise generator

CO
central office

DSP
digital signal processing

E&M
ear and mouth

ERL
echo return loss

FEC
front-end clipping

FXO
foreign exchange office

HOT
holdover time

IETF
Internet Engineering Task Force

IP
Internet protocol

ISP
Internet service provider

ITU–T

International Telecommunications Union–Telecommunications Standardization Sector

MLS
maximum length sequence

MOS
mean opinion score

MPLS
multiprotocol label switching

PACE
perceived annoyance caused by echo

PAMS
perceptual analysis measurement system

PBX
private branch exchange

PC
personal computer

PING
packet Internet groper

PSQM
perceptual speech quality measurement

PSTN
public switched telephone network

QoS
quality of service

RSVP
resource reservation protocol

SNR
signal-to-noise ratio

TCP
transmission control protocol

THD
total harmonic distortion

TOS
type of service

UDP
user datagram protocol

VAD
voice activity detector

VoIP
voice over Internet protocol

VoP
voice over packet

VQ
voice quality

WFQ
weighted fair queuing