# Voice over IP

## Sounding good on the Internet

Princy Mehta and Sanjay Udani

© EXPERT GALLERY/J. NISHINAKA

The bulk of information conveyed over public telecommunication networks is voice. To do this, circuit-switched networks are employed. While circuit switching provides adequate voice quality, it can be highly inefficient. In contrast, the Internet's packet-switched networks are much more efficient but ill suited for voice without judicious implementation. Voice over Internet Protocol (VoIP) wants to provide the efficiency of a packet-switched network while rivaling the voice quality of a circuit-switched network.

Because voice applications are real time, they are intolerant of lengthy delays, packet losses, out-of-order packets and jitter. All these problems gravely degrade the quality of the voice transmitted to the recipient. Unfortunately, wireless networks exacerbate the problems that are intrinsically prevalent in their wire line counterparts: a higher frequency of dropped packets, larger latency and more jitter.

VoIP can be implemented in several ways. A Public Switched Telephone Network (PSTN)-based telephone can communicate with a VoIP application, and vice versa. These telephones can also communicate with each other where part of the call is routed over the Internet instead of solely over a dedicated circuit. Finally, two VoIP applications can communicate directly without accessing the PSTN.

## Components of VoIP

The Public Switched Telephone Network (PSTN) is the collection of all the switching and networking equipment that belongs to the carriers that are involved in providing telephone service. In this context, the PSTN is primarily the wired telephone network and its access points to wireless networks, such as cellular. The overall technology requirements of an Internet Protocol (IP) telephony solution can be split into four categories: signaling, encoding, transport and gateway control.

The purpose of the *signaling* protocol is to create and manage connections between endpoints, as well as the calls themselves. Next, when the conversation commences, the analog signal produced by the human voice needs to be *encoded* in a digital format suitable for transmission across an IP network. The IP network itself must then ensure that the real-time conversation is *transported* across the available media in a manner that produces acceptable voice quality. Finally, it may be necessary for the IP telephony system to be converted by a *gateway* to another format-either for interoperation with a different IP-based multimedia scheme or because the call is being placed onto the PSTN.

Figure 1 displays processing that must occur between the user's voice input and output. The diagram illustrates the necessary steps to achieve packetized voice, from data processing by the digital signal processor (DSP) to transmitting packets over IP. There are numerous packet-handling processes that must be encountered; hence, a nontrivial amount of latency (time delay) is present, which affects perceived voice quality.

## SS7

Once a user dials a telephone number (or clicks a name hyperlinked to a telephone number), signaling is required to determine the status of the called party—*available* or *busy*—and to establish the call. Signaling System 7 (SS7) is the set of protocols (standards for signaling) used for call setup, teardown, and maintenance in the Public Switched Telephone Network (PSTN). It is currently the one being used in North America to establish and terminate telephone calls. SS7 is implemented as a packet-switched network and typically uses dedicated links, nodes and facilities. In general, it is a non-associated, common channel out-of-band signaling network allowing switches to communicate during a call. SS7 signals may traverse real or virtual circuits on links that also carry voice traffic.

However, the industry is moving toward a converged network infrastructure to provide a more efficient and effective way of handling increased call volumes as well as deliver new and enhanced services. The integration of SS7 and IP will provide significant benefits. Figure 2 depicts a type of VoIP network utilizing an SS7-to-IP gateway. SS7 provides the call control on either side of the traditional PSTN, while H.323/Session Initiation Protocol (SIP) provides call control in the IP network. (Neither H.323 nor SIP alone has a complete set of IP telephony protocols.) The media gateway provides circuit-to-voice conversion.

## H.323

H.323, ratified by the International Telecommunication Union-Telecommunication (ITU-T), is a set of protocols for voice, video, and data conferencing over packet-based networks, such as the Internet. The H.323

0278-6648/01/$10.00 © 2001 IEEE

protocol stack is designed to operate above the transport layer of the underlying network. Therefore, H.323 can be used on top of any packet-based network transport, for instance TCP/IP, to provide real-time multimedia communication.

H.323 specifies protocols, including Q.931, H.225, H.245, and ASN.1, for real-time point-to-point audio communication between two terminals on a packet-based network that does not provide a guaranteed quality of service (QoS). The scope of H.323, however, is much broader and encompasses networking multipoint conferencing among terminals that support not only audio but also video and data communications.

In a general H.323 implementation, three logical entities are required: gateways, gatekeepers and multipoint control units (MCUs). Terminals, gateways, and MCUs are collectively known as endpoints. It is possible to establish an H.323-enabled network with just terminals, which are H.323 clients. Yet for more than two endpoints, a MCU is required. It can be combined with a terminal, gateway or gatekeeper.

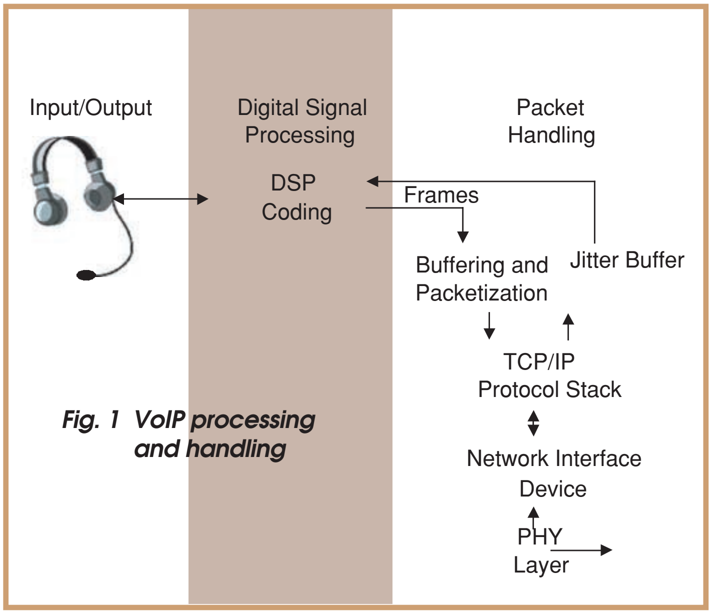

***Fig. 1 VoIP processing and handling***

## SIP

Session Initiation Protocol, SIP, defined by the Internet Engineering Task Force (IETF), is a signaling protocol for telephone calls over IP. Unlike H.323, however, SIP was designed specifically for the Internet. It exploits the manageability of IP and makes developing a telephony application relatively simple. SIP is an application-layer control (signaling) protocol for creating, modifying and terminating sessions with one or more participants.

SIP can be employed to initiate sessions and invite members to sessions that have been advertised by other means, such as via multicast protocols. The signaling protocol transparently supports name mapping and redirection services. This allows the implementation of intelligent network telephony subscriber services. These facilities also enable personal mobility-the ability of end users to originate and receive calls and access subscribed telecommunication services on any terminal in any location. This mobility can be augmented via wireless VoIP.

SIP supports five facets of establishing and terminating multimedia communications:

• User location: determination of the end system to be used for communication;

• User capabilities: determination of the media and media parameters to be used;

• User availability: determining the called party's willingness to engage in communications;

• Call setup: "ringing," establishing call parameters at both called and calling party;

• Call handling: including transfer and termination of calls.

SIP can also initiate multiparty calls using a multipoint control unitMCU or a fully-meshed interconnection instead of a multicast. Gateways that connect PSTN parties can also use SIP to set up calls between them. The protocol is designed as part of the overall Internet Engineering Task Force (IETF) multimedia data control architecture. It incorporates many protocols, for example Resource Reservation Protocol (RSVP) and Real-Time Transport Protocol (RTP), for proper functionality and operation.

## H.323 vs. SIP

H.323 and SIP are competing to obtain dominance of IP telephony signaling. Currently, there is no clear-cut winner. However, the standards appear to be evolving such that the best features of each are being implemented in the other's protocol.

For instance, the evolution of H.323 from versions 1 through 4 has focused on decreasing call setup delay from several round trips to be on par with SIP's 1.5 round trips. This reduces its signaling overhead. Obviously, this convergence is highly desirable. (Both support the majority of required end-user functions comparatively equally, such as call setup, teardown, call holding, call transfer, call forwarding, call waiting and conferencing.)

## Voice coders

An efficient voice encoding and decoding mechanism is vital for using the packet-switched technology. The purpose of a voice coder (vocoder)-also referred to as a codec (coding/decoding)-is to use the analog signal (human speech) and transform and compress it into digital data. A number of factors must be taken into account including bandwidth usage, silence compression, intellectual property, look-ahead and frame size, resilience to loss, layered coding, and fixed-point vs. floating-point digital signal processor (DSPs).

The bit-rate of available narrowband vocoders ranges from 1.2 to 64 kbps, with an inevitable effect on the quality of the restituted voice. There is ordinarily, but not always, a trade-off between voice quality and bandwidth used. Using today's most efficient vocoder allows quasi-toll quality bandwidth usage to be as low as 5 kbps. Toll quality is recognized as the standard of a long-distance PSTN call. As newer and more sophisticated algorithms are developed, this bit-rate will decrease. This will permit more samples to be squeezed more efficiently while minimally sacrificing quality, if at all.

The algorithmic delay introduced by a coding/decoding sequence is the frame length plus the look-ahead size. A vocoder with a small frame length has a shorter delay than one with a longer frame length, but it introduces a larger overhead. Most implementations choose to send multiple frames per packet. Thus, the real frame length to take into account is the sum of all frames stacked in a single IP packet. The smaller the

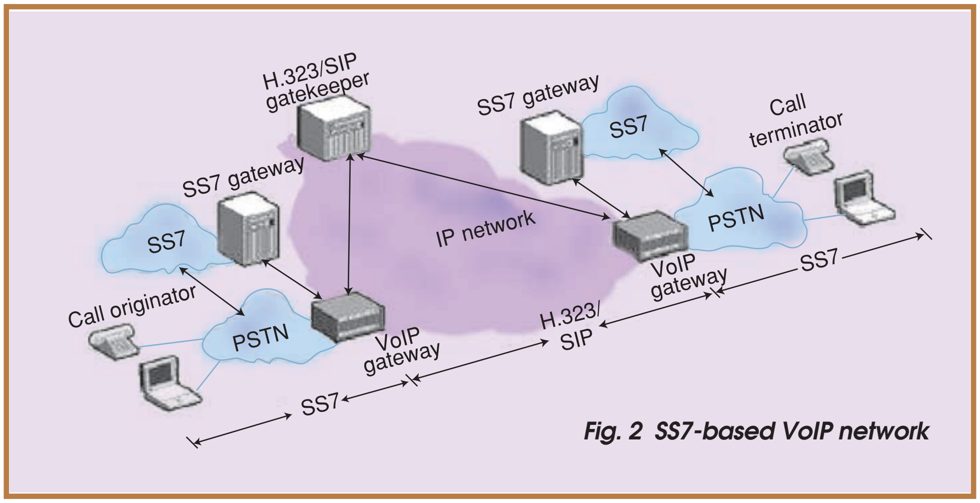

*Fig. 2 SS7-based VoIP network*

frame size, the more frames in an IP packet; thereby, there is minimal influence on latency.

## ITU-T specs

The International Telecommunication Union-Telecommunication (ITU-T) has a rigorous process in approving vocoders. Before a vocoder is chosen, the ITU evaluates its mean opinion score (MOS) and often requires toll quality or better. To determine the MOS, trained evaluators rate the overall quality of speech samples and assign a subjective score. Three popular ITU-approved vocoders are summarized in Table 1; the expected MOS can range from a scale of 1 (bad) to 5 (excellent).

## Future voice coders

One recently established vocoder is the Mixed Excitation Linear Predictive (MELP) vocoder, which utilizes a miniscule 2.4 kbps. Another high quality speech vocoder is being developed based on the Multi-Band Excitation (MBE) model operating at both 2.4 kbps and 1.2 kbps. The trend in industry appears to be developing vocoders that utilize less bandwidth than their predecessors do.

Since the early 1990s, the ITU has forged ahead from the 64 kbps G.711 to the more recent G.723.1 specification that consumes merely one-twelfth of that bandwidth. This bandwidth savings commonly comes at the cost of lower quality and robustness to hostile network environments. Given the inevitable increase in the average user's bandwidth over time, perhaps this effort would be better directed at improving quality first, then addressing bandwidth.

## Transport

Once signaling and encoding occur, Real-time Transport Protocol (RTP) and Real-Time Control Protocol (RTCP) are utilized to move the voice packets. Media streams are packetized according to a predefined format. RTP provides delivery monitoring of its payload types through sequencing and time stamping. RTCP offers insight on the performance and behavior of the media stream, such as voice stream jitter. RTP and RTCP are intended to be independent of the signaling protocol, encoding schemes, and network layers implemented.

**Table 1 Summary of ITU vocoders**

| Voice coder | Bit-rate | Frame length | Expected MOS |
|---|---|---|---|
| G.711 (PCM) | 64 kbps | 1 ms | 4.1 |
| G.723.1 (MP-MLQ) | 6.3 kbps | 30 ms | 3.9 |
| G.723.1 (ACELP) | 5.3 kbps | 30 ms | 3.65 |
| G.726 (ADPCM) | 32 kbps | 0.125 ms | 3.85 |
| G.729A (CS-ACELP) | 8 kbps | 10 ms | 3.7 |

## RTP

Real-time Transport Protocol (RTP) provides end-to-end delivery services for data with real-time characteristics. Those services include payload type identification, sequence numbering, time stamping and delivery monitoring. Applications typically run RTP on top of User Datagram Protocol (UDP) to make use of its multiplexing and checksum services. In fact, both protocols contribute parts of the transport protocol functionality; however, RTP may be used with other apposite network-layer or transport-layer protocols.

RTP does not intrinsically provide any mechanism to ensure timely delivery or provide other Quality of Service guarantees. Instead, RTP relies on lower-layer services to provide them. A signaling protocol also must set up the connection and negotiate the media format that will be used. RTP does not guarantee delivery or prevent out-of order delivery, nor does it assume that the network can reliably deliver packets in sequence.

## RTCP

Real-Time Control Protocol (RTCP) is based on the periodic transmission of control packets to all participants in the session. It uses the same distribution mechanism as the data packets. The underlying protocol must provide multiplexing of the data and control packets. RTCP performs the following functions:

- Provide feedback on the quality of the data distribution (primary function);
- Carry a persistent transport-layer identifier for a Real-time Transport Protocol (RTP) source, canonical name;
- Controls the rate in order for RTP to scale up to a large number of participants; and
- Conveys minimal session control information.

## Gateway control

Gateways are responsible for converting packet-based audio formats into protocols understandable by PSTN systems. The aforementioned signaling protocols provide more services than are necessary, such as service creation and user authentication, which are irrelevant for gateways. Vendors have gravitated towards simplified Device Control Protocols rather than all-encompassing signaling protocols.

The IETF standard Media Gateway Control Protocol (MGCP) is a merger between the Internet Protocol Device Control and the Simple Gateway Control Protocol. The Megaco protocol (H.248), which is still evolving, is MGCP's progeny. It contains all of MGCP's functionality, plus superior controls over analog telephone lines and the ability to transport multiple commands in a single packet.

Media gateways will be the junctions that provide paths between switched and packet networks for voice. When media gateways are initially set up for communication, a vocoder approach normally is used. Megaco-related standards will enable support of existing and new applications of telephone service over hybrid telephone networks containing an assortment of technologies.

## Wireless networks

An emerging trend for implementing VoIP-and, in general, connecting computing devices-is in wireless networks. A wireless local area network (WLAN) is a data transmission system designed to provide location-independent network access between computing devices by using radio waves rather than a cable infrastructure. WLANs give users wireless access to the full resources and services of the LAN across a building or campus environment.

For voice applications, wireless networks aggravate the problems already prevalent in wireline networks: a higher frequency of dropped packets, larger latency and more jitter. Furthermore, there are additional security issues: it is relatively easier for an unauthorized device to surreptitiously eavesdrop on a conversation. Finally, interference between different wireless technologies must be considered when they are both operating on the same frequency band.

## QoS

The basic routing philosophy on the Internet is "best-effort." This attitude serves most users acceptably but it is not adequate for the time-sensitive, continuous stream transmission required for VoIP.

Quality of Service (QoS) refers to the ability of a network to provide better, more predictable service to selected network traffic over various underlying technologies, including IP-routed networks. QoS features are implemented in network routers by:

- Supporting dedicated bandwidth;
- Improving loss characteristics;
- Avoiding and managing network congestion;
- Shaping network traffic; and
- Setting traffic priorities across the network.

Voice applications have different characteristics and requirements from those of traditional data applications. Because they are innately real-time, voice applications tolerate minimal delay in delivery of their packets. Additionally, they are intolerant of packet loss, out-of-order packets, and jitter. To effectively transport voice traffic over IP, mechanisms are required that ensure reliable conveyance of packets with low and controlled latency.

Another approach utilizes Resource Reservation Protocol (RSVP) which is a relatively new protocol developed to enable the Internet to support QoS. Using RSVP, a VoIP application can reserve resources along a route from source to destination. RSVP-enabled routers will then schedule and prioritize packets to fulfill the QoS. RSVP is part of the Internet Integrated Service (IIS) model, ensuring best-effort service, real-time service, and controlled link sharing.

While QoS is an extension to IPv4-the current version of IP-IPv6 (the successor of IPv4) will inherently support QoS. However, IPv6 also has a much larger packet header, so it is possible that while QoS will alleviate much of the jitter and congestion voice packets presently suffer, it could come at the cost of increased latency. IPv6 headers necessitate 40 bytes, compared with 20-byte IPv4 headers, thus doubling the overhead. This may pose trouble for vocoders that only succeed with diminutive packets. Nevertheless, this larger packet overhead can be partially offset if IPv6 provides for efficient compression schemes for the header.

## Packet loss

UDP cannot provide a guarantee that packets will be delivered at all, much less in order. Packets will be dropped under peak loads and during periods of congestion. Due to time sensitivity of voice transmissions, the normal TCP-based retransmission schemes are not appropriate. Approaches used to compensate for packet loss include interpolation of speech by replaying the last packet and sending redundant information. Packet losses greater than 10 percent are generally intolerable, unless the encoding scheme provides extraordinary robustness.

## Jitter

Inasmuch as IP networks cannot guarantee the delivery time of data packets (or their order), the data will arrive at very inconsistent rates. The variation in inter-packet arrival rate is jitter, which is introduced by variable transmission delays over the network. Removing jitter to allow an equable stream requires collecting packets and storing them long enough to permit the slowest packets to arrive in time to be played in the correct sequence. The jitter buffer is used to remove the packet delay variation that each packet encounters transiting the network. Each jitter buffer adds to the overall delay.

## Latency

Latency is the time delay incurred in speech by the Internet Protocol (IP) telephony system. One-way latency is the amount of time measured from the moment the speaker utters a sound until the listener hears it. Round trip latency is the sum of the two one-way latency figures that compose the user's call. The lower the latency, the more natural interactive conversation becomes; accordingly, the additional delay incurred by the VoIP system is less noticeable. In PSTN calls, the round trip latency of calls originating and terminating within the continental United States is under 150 ms.

In a VoIP implementation used to reduce costs, studies suggest that users will tolerate one-way latency of up to 200 ms. The 1996 ITU Recommendation G.114 for one-way end-to-end transmission time limit is:

- Under 150 ms: acceptable for most user applications;
- 150 to 400 ms: acceptable provided administrators know of the transmission time impact on the quality of user applications; and
- Over 400 ms: unacceptable for general network planning purposes.

Two difficulties are echo and talker overlap that result from a high end-to-end delay in a voice network. Echo-wherein the speaker's voice is reflected back-becomes a problem when the round-trip delay is more than 50 ms. Since echo is perceived as a significant quality obstacle, the VoIP system must address the need for echo control by implementing echo cancellation. Talker overlap-the problem of one caller stepping on the other talker's speech-is made worse when the one-way delay is greater than 250 ms. The end-to-end delay budget, therefore, is the major constraint and driving requirement for reducing latency through a packet network.

## Bit-rate vs. voice quality

As previously mentioned, many developers have focused on designing vocoders that consume progressively lower bandwidth. Moreover, many algorithms were created for using voice over a reliable circuit-switched connection rather than the packet-based network the Internet utilizes. This effort might be misdirected. Most applications of VoIP rely on connectivity to the Internet, where the vast majority of its users have a 28.8 kbps or higher connection. Nonetheless, developers are still pursuing ultra-low bandwidth vocoders instead of improving the quality of low bandwidth vocoders already in existence. Perhaps this effort is intended to allow users to concurrently enjoy other bandwidth-consuming applications, such as browsing the World Wide Web.

Some developers are alternatively constructing higher quality vocoders that consume more bandwidth. They are amenable to trading-off bandwidth to achieve this quality. It is also critical for the vocoder to tolerate mishandled, dropped and out-of-order packets intrinsic in the User Data Protocol (UDP). Of equal importance, one-way latency should be confined to one-quarter of one second. Finally, the vocoder should maintain an optimally sized buffer to restrain jitter, echo, and talker overlap.

Since users may not endure inferior performance, the focus should be on high quality instead of ultra-low bit-rate. Manifestly, 64 kbps is too high for users dialing up via analog modems to connect to the Internet; nevertheless, a higher quality vocoder could be preferable to a low quality vocoder. In a corporate or broadband environment, even 64 kbps is just noise in the line when the average user is allotted hundreds, if not thousands, of kilobits per second.

Another possibility is developing higher bandwidth vocoders to allow something that the traditional telephone system can never do: transport high fidelity stereo audio. A potential application would be allowing users to call another VoIP application to listen to high quality, compressed music, for instance in MP3 format, consuming a mere 128 kbps. Of course, there are other issues involved, such as the server's ability to provide music at this fidelity while being able to scale.

## Summary

It remains to be seen when VoIP can emerge from a specialized application to mainstream voice communication. While VoIP technology may have progressed admirably, as gauged by protocol and vocoder maturity, it still has plenty of room for improvement as indicated by the following drawbacks:

- Erratic quality of voice transmissions;
- Unreliability of IP networks;
- Standards battles;
- Encroaching/competing wireless technologies; and
- Confusing human usability factors.

Reliability cannot be overemphasized. The PSTN operates with at least 99.999 percent specified availability and is available even during power outages. This cannot be said of modern VoIP applications; consequently, VoIP's reliability must improve in the near future for it to gain wide acceptance and let users sound good on the Internet.

## Read more about it

- Philip Carden, "Building Voice over IP," *Netw. Comput.*, May 2000.
- Linden deCarmo, "The media gateway control protocol: A simpler and more reliable voice over the internet," *Dr. Dobb's Journal*, May 2000.
- Bill Douskalis, *IP Telephony: The Integration of Robust VoIP Services.* Upper Saddle River, NJ: Prentice-Hall, 2000.
- M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg, "SIP: Session Initiation Protocol," RFC 2543, The Internet Society, Mar. 1999.
- Oliver Hersent, David Gurle, and Jean-Pierre Petit, *IP Telephony: Packet-Based Multimedia Communications Systems.* Harlow, England: Addison-Wesley, 2000.
- "Leveraging the intelligence of SS7 to improve IP-based remote access and other IP services," 3Com Corp., May 19, 1999.
- Alan Percy, "Understanding latency in IP telephony," Brooktrout Technology, Feb. 1999.
- H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," RFC 1889, Jan. 1996.

## About the authors

Princy Mehta earned his MS degree in Telecommunications and Networking Engineering from the University of Pennsylvania and his BS degree in Computer Engineering from Rutgers University. He is currently employed with Lockheed Martin Naval Electronics & Surveillance Systems-Surface Systems as a Member of Engineering Staff. A graduate of the company's Engineering Leadership Development Program, his professional endeavors include Voice over IP and network systems security, in which he is SANS GIAC certified. Prior to his recent responsibilities, Princy programmed in C and shell script, administered a variety of Unix systems, and developed embedded DSP applications for multiprocessors. E-mail: <prmehta@seas.upenn.edu>.

Sanjay Udani received his PhD, MSE and BSE/BS Economics degrees from the University of Pennsylvania. He is currently a Distinguished Member of Technical Staff in Verizon's Technology organization in Arlington, VA. He is an adjunct faculty member at the University of Pennsylvania, teaching a graduate telecommunications course. Prior to joining Verizon he was involved in VLSI and ASIC design at Intel, and dabbled with large-scale virtual environment network design as well as power management for mobile computing while in graduate school. E-mail: <udani@seas.upenn.edu>.

Note: The full (unabridged) versionof this paper is posted at http://www. cis.upenn.edu/~techreports/and is listed as Technical Report MS-CIS-01-31.