How Does It Sound?

To check voice quality, designers of compressed and packetized networks are enlisting computer-based techniques

⬤ **PAUL DENISOWSKI**
Agilent Technologies Inc.

The likelihood that the plain old telephone system will not endure unchanged over the next decade seems pretty well accepted within the telecommunications industry. All of the major communications equipment manufacturers, including those whose primary business has been traditional telephony, have committed substantial resources to developing equipment for networks in which voice is carried as digital data, often compressed, along with nonvoice data over a common packet-switched infrastructure.

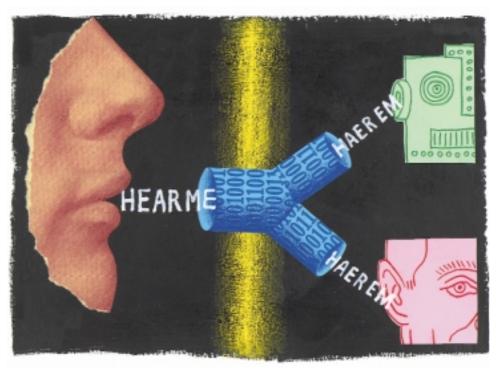These networks are of various kinds, both wireline and wireless. Central to the thinking behind them is the assumption that, in the future, voice will constitute only a minor fraction of the total traffic to be carried. It will therefore be wisest, the reasoning goes, to optimize the networks for data communication and to fit voice in as well as possible.

That is easier said than done. While data networks are designed to provide large amounts of capacity whenever they can, they may delay data packets until such capacity is available. Such behavior is a poor match for voice traffic, which needs only modest capacity but can tolerate only short delays.

Before long, therefore, most of these data-oriented

networks will be fitted with mechanisms for dedicating or reserving capacity for time- and loss-sensitive voice data. But, at present, few have such capability. As a result, voice conversations can be plagued by delays, echoes, and dropped fragments. These effects are aggravated by speech compression, which is implemented to conserve transmission capacity.

Natural speech is compressible because it has a lot of redundancy, so dropping a few packets of uncompressed speech may not affect the perceived quality very much. Compression, though, removes most of the redundancy, so every lost packet hurts. In any event, sending compressed speech

GENE GREIF

over a data-oriented network with no special provisions for handling it generally degrades voice quality.

That loss of quality is one of the main barriers to widespread acceptance of voice-over-packet networks. Consumers, long trained by the public switched telephone network (PSTN) to expect what is known as toll-quality voice, will inevitably compare the new voice networks with the old one for quality. Manufacturers and network operators must therefore be able to make such comparisons as well.

Subjective measures, like taking the mean of the opinions of a group of listeners, were an obvious starting point. Newer, more objective techniques, developed in Britain, the Netherlands and elsewhere, involve comparing the received sound with the transmitted sound and scoring the differences. But how? What exactly is voice quality, and how can it be quantified in an objective and reproducible manner? With some answers in hand, it is possible to pin down the factors that influence the perception of voice quality by the users of these networks, and the steps that can be taken to ensure that an acceptable level of quality is met.

Speech coding and compression

Both speech coding and compression have been used in the PSTN for over 20 years. With the exception of the local customer loop (normally analog, but increasingly digital), almost all voice is carried in digital format. In traditional phone networks, the standard method for converting analog voice signals to digital form is to sample them 8000 times per second and then encode each sample as an 8-bit binary word, as specified in detail in ITU-T standard G.711 (the ITU-T being the Telecommunication Standardization Sector of the Geneva-based International Telecommunication Union, a specialized agency of the United Nations). The result is the familiar 64-kb/s digital data stream known in telephony as a DS-0, the lowest rung in the digital signal hierarchy.

Speech compression is associated by most engineers with digital signal processing, but has actually been practiced in the telephone network since the all-analog era. In the early days it took the form of low-pass filtering of the analog signal. Although the human auditory range is generally regarded as extending up to about 20 kHz, telephone networks band-limit voice signals to approximately the bottom 4 kHz of the speech signal. Doing so greatly simplifies the design (and lowers the cost) of low-noise amplifiers and network equalization circuitry, but not without compromising speech quality. Although most of the energy in human speech falls below 4 kHz, the small fraction at higher frequencies does affect intelligibility: just try to distinguish between the words "fine" and "sign" over the telephone.

In addition to this low-pass filtering, voice-over-packet networks commonly compress voice signals further, employing complex algorithms to preserve or emphasize only those parts of the input speech signal of interest to a human listener. The idea is not to reproduce faithfully the waveforms and spectra of the original signal; rather, it is to transmit just the parts needed for reconstruction of the signal to the listener's satisfaction. The design of the algorithms owes as much to psychoacoustics as it does to engineering.

As already noted, the whole point in using compression is to conserve system capacity: algorithms have been developed for encoding speech into data streams with rates below 8 kb/s, while maintaining adequate perceptual quality for many applications. This is a noteworthy reduction from the 64 kb/s produced by straightforward digitizing in accordance with G.711. Still, this saving has a price. In general, the greater the savings in data rate, the worse the perceived quality. In addition, the higher the level of compression, the longer it takes to encode the speech samples (which adds delay), and the more computational power is required on both ends (higher cost, greater complexity) in order to maintain reasonable voice quality.

What is voice quality?

The ability to quantify voice quality in an objective manner is important for several reasons. First, it is useful to be able to compare the quality of a voice-over-packet network to the PSTN, in part because the latter has become the standard for acceptable voice quality. Additionally, engineers need some means for determining the effects on voice quality of changes in their designs or variations in network conditions, for example, the amount of packet jitter (delay variation). Lastly, objective measurements of voice quality are valuable from a business perspective. They allow service providers to compare their own and their competitors' offerings, while also serving as the basis of voice-quality service level agreements.

Three main factors determine how an individual perceives voice quality: delay, echo, and clarity. All three must fall within certain bounds for the quality of the received voice signal to be judged acceptable.

The long and the short of delay

Delay (also known as latency) is the easiest of the three factors to understand and to quantify; it is the time the sounds take to travel from speaker to listener. The amount of delay in the circuit-switched public phone network does not vary much for terrestrial calls; it is usually on the order of a few tens of milliseconds, and is largely a function of distance.
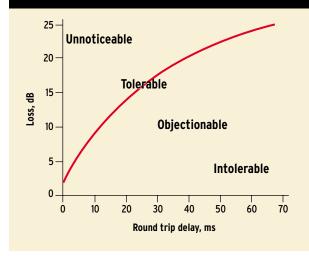
In certain circumstances, however, delay can easily go into the hundreds of milliseconds, even in the traditional voice network. Most people have encountered this amount of delay during an overseas satellite-carried call because of the distances involved. After all, the path up to a geostationary satellite and back is something in excess of 70 000 km, so that traversing it at the speed of light takes almost 250 ms.
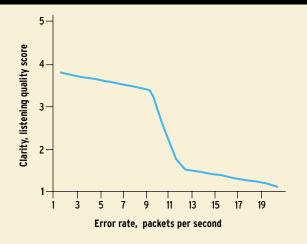
In situations where end-to-end delay exceeds 150 ms, the ability to hold a conversation is impaired. The parties involved begin to interrupt or "talk over" each other because of the time it takes to realize the other party is speaking. When delay becomes high enough, conversations degrade into a half-duplex mode, with conversation taking place strictly in one direction at a time.

● Quantifying Speech Quality

How annoying is an echo? It depends on the echo's delay and on how its strength compares with that of the original signal. The red line represents tolerable combinations. Below the line lie objectionable echoes.

Speech clarity exhibits a threshold effect when plotted as a function of signal quality. For the system studied here, transmission errors have little effect on clarity so long as they occur at a rate below about 9 packets per second. Any higher, and clarity rapidly becomes unacceptable.

Voice-over-packet networks typically run very close to, if not over, the 150 ms threshold. Latency in these networks stems from many factors, such as propagation delay, link speed, buffering, and coding (how long it takes to encode the speech). Note, too, that unlike delay in the circuit-switched PSTN, delay in packetized voice networks often varies with time of day and day of week or with network conditions.

Digital cellular systems are also subject to significant delays. Although most of them today run over circuit- and not packet-switched networks, they do make extensive use of compression. Thus, although they are not subject to the transmission delay uncertainties of packet-switched networks, they are susceptible to coding delays.

Echo, sometimes desirable

At a basic level, echo is simply hearing your own voice reflected back to you. More technically, echo occurs when some of the transmitted signal appears on the receive path. One common (and desirable) form of echo is sidetone—hearing your own voice in the earpiece of the telephone with essentially zero delay. In fact, most people find the absence of sidetone disturbing, believing that if you can't hear yourself, the other person can't hear you either.

Echo can be electrical or acoustic in origin. Electrical echoes are often caused by poor impedance matching or crosstalk. Acoustic echo can arise from coupling between the speaker and microphone at the far end, for example, in a speakerphone.

The level of annoyance caused by echo increases with both its intensity and its delay [see figure above left]. That is why sidetone is rarely a problem: its delay is so short that it would have to be very loud indeed to cause annoyance. Voice-over-packet networks, where delay is typically an order of magnitude higher than in traditional voice networks, are another story.

The methods of dealing with echo are straightforward, at least in principle: suppress it or cancel it. Echo suppression is

● Defining Terms

Bark scale: a nonlinear scale of frequency bins corresponding to the first 24 critical bands of hearing. The bin center frequencies are only 100 Hz apart at the low end of the scale (50 Hz, 150 Hz, 250 Hz ...) but get progressively further apart at the upper end (4000 Hz, 4800 Hz, 5800 Hz, 7000 Hz, 8500 Hz...).

Circuit-switching: the traditional technique for establishing calls in the public switched telephone network. A circuit is established ("nailed up") between the communicating parties before useful communication begins. When the conversation ends, the circuit is "torn down," making the resources it used available for others.

Codec: coder-decoder, a device for encoding analog speech signals into digital form and vice versa. As the term is currently used, codecs differ from analog-to-digital and digital-to-analog converters in that they not only digitize speech signals, they also process (compress and decompress) them.

Packet-switching: a communications technique that, like the postal system, does not require a permanent end-to-end connection between the sender and the recipient. Instead, packets are dumped into the system and then routed from point to point based on information in their headers—just as letters are routed from post office to post office based on information written on their envelopes.

PAMS: British Telecom's perceptual analysis/measurement system.

PSQM: KPN Research's perceptual speech quality measurement.

the simpler of the two; it turns off the receive path while the transmit path is active. The problem is that the echo suppressor circuitry needs a perceptible length of time to determine that a speaker has finished talking. An annoying half-duplex (one way at a time) communication can result, not terribly different from that caused by long transmission delays except that it is imposed by the equipment instead of by human nature.

A superior method is echo cancellation, in which an echo canceler so to speak remembers the transmitted sound and then subtracts any weakened and/or delayed version of that sound that appears on the receive path. Thanks to the availability of powerful yet reasonably-priced digital signal processors for implementing echo cancellation, echo suppression has been largely replaced by echo cancellation.

Echo cancellation works best when echo delay times are short. It is perhaps best regarded as a "clean-up" tool for eliminating residual effects after other techniques for reducing delay have done all they can. It is also worth remembering that echo-canceler characteristics, such as how long they remember the transmitted signal, can vary, making them more or less suitable for use in a given voice-over-packet network.

Clarity hard to quantify

Of the three factors defining voice quality, clarity has been the hardest to quantify or measure objectively. In a general sense, clarity corresponds to how intelligible the received speech is—for example, how well (or whether) the listener can make out the words, identify the speaker, or notice nuances such as the speaker's emotional state.

One important thing to bear in mind about clarity is the highly nonlinear relationship between cause and effect. As in many digital technologies that involve compression (notably MPEG-2), clarity is subject to a so-called cliff effect [see figure on upper right of opposite page]. With increasing signal impairments, clarity degrades only slightly up to a given point, after which the received speech rapidly becomes incomprehensible. The exact location of the "cliff" is usually determined experimentally.

Historically, clarity has been measured using a technique known as the mean opinion score (MOS). Such a score is obtained by having a group of listeners rank a speech sample on a scale of 1–5, where 1 is very bad, 5 is excellent and 4 is normally considered toll quality (what is heard on the PSTN). Since it is logistically difficult to set up and execute a well-constructed MOS test, MOS testing is particularly ill-suited for long-term measurement, such as making measurements every 5 minutes for an entire week.

To address these shortcomings, various computer-based methods have been developed to create objective and reproducible measurements of perceived voice quality. Most define an acoustic and perceptual (or cognitive) process for human speech in order to determine how well a received speech sample compares with the original signal from the point of view of a human listener.

Two clarity measurements are currently in wide use. One is Perceptual Speech Quality Measurement (PSQM), originally developed by KPN Research in the Netherlands and now specified in ITU-T P.861. The other is Perceptual Analysis/Measurement System (PAMS), which was developed by British Telecom in the United Kingdom.

Both PSQM and PAMS use natural speech or speech-like samples as their inputs. Generally, a given speech sample is played over the speech path, where it presumably undergoes various types of impairments during encoding, packetization, transmission, and decoding. The received speech sample and the original are then used as the inputs to the clarity algorithms. Typically tests are made using speech samples from both male and female speakers, with the utterances carefully chosen to exercise the system with a comprehensive range of phonological (speech sound) patterns.

The four steps

For both methods, the first step is to time-align the reference (original) and received (degraded) signals [see figure, next page]. The job is not as simple as it sounds. First, the algorithm has no a priori knowledge of how much delay has occurred in the system. In packet-based networks, moreover, there is the problem of time warping, in which different packets are subjected to different amounts of delay, so the received signal may differ in length from the original. One way to align the signals in time is to adjust the time difference between them until the cross-correlation (a mathematical measure of similarity) between them is maximized.

The second step is gain-scaling of the reference and received signals, to bring them to the same power level. Again, this is more complicated than it sounds since, in most cases, the

> **Experience suggests that people will accept less than toll-quality voice in exchange for other benefits**
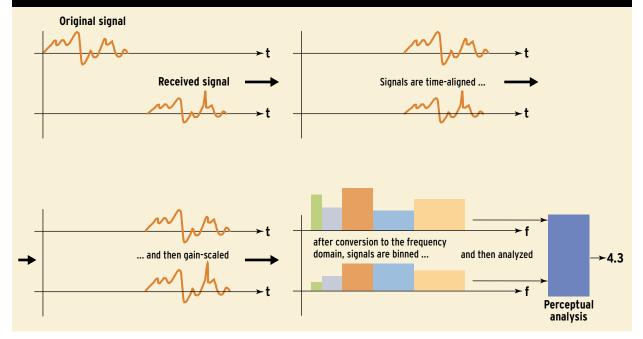
received signal will have suffered some impairment and cannot be regarded as merely an attenuated copy of the reference. Nevertheless, it is possible to bring them to the same power level in some average sense.

After that, the time-domain signals are transformed to the frequency domain and the resultant spectra assigned to bands, or bins, that reflect the known nonlinearity of human hearing with respect to frequency. The binning, which follows a scale known as a Bark scale, reflects the fact that humans resolve low frequencies better than high ones—that is, the bins are narrower at the low-frequency end of the scale and wider at the high-frequency end.

Then comes the critical part of the analysis. The contents of the bins are compared and processed using a perceptual model to determine how significant the differences seem to the human ear. The result of this processing provides clarity scores for each part of the utterance.

● Four Steps Prepare for Voice Clarity Analysis

Before a received signal and its original [upper left] can be compared, the two must be time-aligned [upper right], gain-scaled [lower left], converted to the frequency domain, and then binned [lower right]. In the binning process, values of signal amplitude are determined for each of a set of frequency ranges, or bins, whose width increases with increasing frequency—emulating the behavior of the human auditory system. It is the values in the bins that are compared using a perceptual model to generate a clarity score.

The output of the PSQM algorithm is a numeric score ranging from 0 to 6.5, with lower numbers indicating relatively better voice quality. PSQM was originally designed to assess and compare various speech codecs, not end-to-end networks. Various enhancements (referred to collectively as PSQM+) have therefore been added to it to allow for network testing. The mapping from PSQM to the traditional MOS listening quality scale is nonlinear.

PAMS produces two scores: a listening quality score (Ylq) and a listening effort score (Yle), both on a scale from 1–5, with higher scores meaning better quality. Like the PSQM clarity score, the listening quality score measures how closely the received speech resembles the original (again, in the judgment of a human listener). The listening effort score is different. Most useful when dealing with badly degraded signals, it does not assess sound quality, but rather measures the amount of mental effort a listener must exert to understand the meanings of the sentences.

The measure of usefulness of an objective speech quality assessment algorithm like PSQM or PAMS is how well its results correlate with, or predict, scores from a well-executed subjective speech quality (MOS) test using human listeners. Here the news is good. These perceptual algorithms typically have a very high correlation ($\rho > 0.9$) with subjective measurements. More traditional objective evaluation methods, such as signal-to-noise ratio, on the other hand, correlate rather poorly with subjective scores. Simply having a high signal-to-noise ratio does not guarantee high perceived speech quality.

KPN Research and British Telecom, the developers of PSQM and PAMS, respectively, are currently collaborating on a proposed new ITU-T standard for objective speech quality assessment model. Called Perceptual Evaluation of Speech Quality (PESQ), it will combine the strongest parts of its two predecessors—the perceptual model of PSQM and the time-alignment routine of PAMS. As a result, PESQ will presumably produce scores that correlate even more strongly with subjective mean opinion scores.

Evidently, despite substantial progress in this area over the last 10 years, much work still remains to be done. In technology, better techniques are needed for, among other things, minimizing delay and reducing transmission errors. In the field of psycho-acoustics, our understanding of how speech is heard and processed must improve.

Voice quality alone will not determine the success or failure of voice-over-packet networks. Their fate will be decided in the marketplace, where other factors will come into play. Experience suggests that people will accept less than toll quality voice in exchange for other benefits, like mobility, reduced cost, and advanced services. All the same, speech quality remains one of the great hurdles affecting market acceptance of such networks, and the ability to identify and quantify the factors influencing voice quality has a pivotal role to play in the development and deployment of this new technology.

SPECTRUM EDITOR: MICHAEL J. RIEZENMAN ●