Assessing Voice Quality in Packet-Based Telephony

Delay, echo, codec type, and packet loss all come into play in achieving the quality of PSTN calls in packet-based telephony.

Jan Janssen, Danny De Vleeschauwer, Maarten Büchli, and Guido H. Petit
Alcatel Bell

The perceived quality of a telephone call largely depends on two factors: distortion, the difference between the received signal and the original one; and mouth-to-ear delay, the time between the moment the speaker makes an utterance and the moment the listener hears it. Different kinds of networks introduce different amounts of distortion and delay, so calls transported in different ways attain different quality levels.

To assess the quality of phone calls transported over a public switched telephone network (PSTN), which suffer very little from distortion, we can rely on the International Telecommunications Union's recommendations G.114 and G.131, which specify the tolerable mouth-to-ear delays for calls with and without echo control.[1,2] (The ITU-T series G recommendations on transmission systems and media, digital systems, and networks are available online at www.itu.int/rec/recommendation.asp?type+products&lang=e&parent=T-REC-G.) Nowadays, however, packet-based networks are becoming increasingly popular for telephony because they offer a cheap alternative to PSTN long distance. Packet telephony integrates both data and real-time voice traffic on the same infrastructure, and it allows easier introduction of new, multimedia services. Moreover, it is much more flexible in terms of codecs — the encoder/decoders used in telephony to reduce bit rate on the transmission paths. PSTNs are bound to a single codec, G.711, but packet-based networks can use any codec supported by both user terminals.

The price of this flexibility is complexity: Transporting voice over a packet-based infrastructure is likely to introduce more delay and distortion. Besides the propagation and switching delays encountered in a PSTN, delays from coding, packetization, queuing, and dejittering come into play. Furthermore, mouth-to-ear delays can differ considerably from one direction to the other, which almost never occurs in a PSTN. Distortion can

1089-7801/02/$17.00 ©2002 IEEE

stem from the use of a low-bit-rate codec or the loss of voice packets in the network or in the dejittering buffer. Controlling both the mouth-to-ear delay and the distortion is key to offering high-quality packet-based voice calls.

Our goal in this article is to extend recommendations ITU-T G.114[1] and G.131[2] to cover distorted phone calls transported (partly) over a packet-based network. We will assume the user terminals to be optimally tuned and focus on how network parameters — delay, packet loss, jitter, and so on — influence voice quality. We'll then discuss how those parameters can be quantified and incorporated into a model that lets us predict the quality of any packet-based phone call.

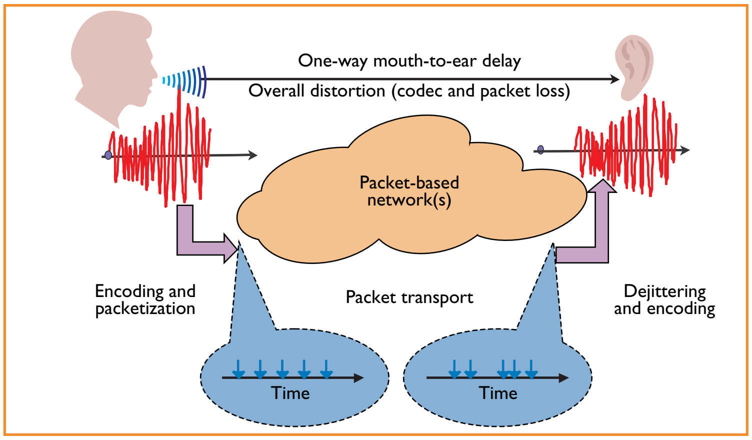

One-way mouth-to-ear delay
Overall distortion (codec and packet loss)
Packet-based network(s)
Encoding and packetization
Packet transport
Dejittering and encoding
Time
Time

Figure 1. Three essential stages in packetized phone call transport: encoding and packetization; packet transport; and dejittering and decoding.

Principles of Packetized Phone Call Transport

Figure 1 shows the three essential stages in the packetized transport of voice calls. In the first stage, the digitized voice signal (for example, G.711) is encoded and packetized. This packetization and encoding operation can take place either in the user terminal or in a gateway — for example, between a PSTN and a packet-based network. If the operation occurs in a gateway, we assume that the circuit-switched transport of the voice signal from the user terminal to the gateway introduces negligible delay and distortion.

The packetization delay T_{pack} is the time needed to collect all voice samples that end up in one packet: That is, for every T_{pack} a new voice packet is produced. T_{pack} scales linearly with the payload size, and its choice is a tradeoff between effective bit rate (codec plus overhead bit rate) and delay. That is, the larger the packets, the smaller the relative influence of overhead bytes on the effective bit rate — and the reverse.

In addition, the voice-encoding process, performed by a digital signal processor, takes some time, as do the other processes that run on the DSP. One such process is an algorithm that detects whether the incoming signal is a pure speech signal or consists of fax, modem, or dual-tone multifrequency (DTMF) tones that should bypass the voice encoder. Such an algorithm cannot make an instantaneous decision based on a single sample; it needs time to collect several. This process introduces a *look-ahead delay*. Some codecs introduce a similar look-ahead delay in the encoding process.

Packet Transport

In the second stage, the resulting flow of voice packets is transported over a packet-based network consisting of several access or core nodes. The network delay — that is, the delay incurred by transporting a voice packet over the network — breaks down into two parts: a deterministic part, the minimal network delay $T_{\mathrm{net,min}}$, and a stochastic part, the total queuing delay T_{que}.

The minimal network delay $T_{\mathrm{net,min}}$ consists mainly of the propagation delay (5 microseconds per kilometer), the sum of all serialization delays, and the route lookup delay. If somewhere along the route the network transports the packets over an unreliable channel (an air interface, for example), forward error correction techniques — such as interleaving coupled with Reed-Solomon block or convolutional channel codes — also contribute an amount, T_{FEC}, to the minimal network delay. We assume that route updates are so infrequent that the probability of one occurring during a phone call is negligible, and, hence, the minimal network delay is constant. Yet, it is simple to extend the framework we describe in this article to the more general case.

In a single network node, the queuing delay comes from several flows competing for the available resources in that node's queue. The total queuing delay T_{que} is the sum of the queuing delays in all traversed nodes — this delay results in *jitter* in the voice flow. That is, a flow of voice packets that entered the network with constant interarrival times does not leave the network in that way because some voice packets are delayed more than others. *Jitter* is the delay difference between the slowest packet still to be considered on time and the fastest packet.

During transport over the network, a fraction $P_{\mathrm{loss,net}}$ of the packets can get lost either because of overflowing queues or erroneous transport over unreliable links. For losses caused by unreliable links, we can make a tradeoff between packet loss in the network and the FEC delay T_{FEC}.[3]

Dejittering and Decoding

In the last stage of packetized phone call transport, the jittered packet flow is dejittered and decoded. Because the decoder needs the packets at a constant rate, dejittering is absolutely necessary. Dejittering a voice flow consists of retaining the fast packets, which did not have to queue at any of the nodes, in the dejittering buffer to allow the slow ones to catch up. When a voice call's first packet arrives at the receiver, there must be some mechanism to decide its play-out time. Unfortunately, the receiver usually does not know whether the first arriving packet is slow or fast. *Static dejittering mechanisms* retain the first packet in the dejittering buffer under the assumption that it was a fast one and then read the buffer at a constant rate.

In principle, if no packets are to be lost, the dejittering buffer then should retain the first packet for a time equal to the maximal total queuing delay. However, because this might introduce too much delay, and because voice codecs can tolerate some packet loss, the practice is different: The first packet remains in the buffer for a time equal to a $(1 - P_{loss,jit})$-quantile of the total queuing delay — that is, the queuing delay value that is exceeded only by a fraction $P_{loss,jit}$ of the voice packets. This means that for the worst-case situation — when the first arriving packet happens to be a fast one — a fraction $P_{loss,jit}$ of the packets will be lost in the dejittering buffer because they arrive too late. When the first packet is not a fast one, fewer packets will be lost, but the mouth-to-ear delay will be greater.

We can avoid this increase in delay by using *dynamic dejittering mechanisms*. These algorithms gradually learn from the packets already received whether the first arriving packet was a fast or slow one and thereby compensate for the first packet's total queuing delay while still adhering to the tolerated packet loss value.[4] That is, these mechanisms tune dejittering delay T_{jit} so that in the long run, the sum of the first packet's queuing delay $T_{que,1}$ and the dejittering delay T_{jit} becomes equal to the $(1 - P_{loss,jit})$-quantile of the total queuing delay.

Finally, the decoding and echo-control processes also introduce delay.

Dejittering, decoding, and echo control can be performed either in the user terminal or in a gateway. For the gateway case, we again assume that the transport of the voice signal from the gateway to the user terminal introduces only negligible delay and distortion.

Total Impact

Now let's bring together the impact of all three stages on the one-way mouth-to-ear delay TM2E and the overall packet loss Ploss.

We can split a packetized phone call's mouth-to-ear delay (in one direction) into several terms:

$$T_{M2E} = T_{pack} + T_{DSP} + T_{net,min} + T_{que,1} + T_{jit}. \quad (1)$$

T_{pack} is the packetization delay; T_{DSP} is the sum of encoding, decoding, look-ahead, and echo-control delays; $T_{net,min}$ is the total minimal network delay (possibly including the delays over the access parts if gateways are involved and T_{FEC} introduced by the scheme to protect transport over an unreliable channel); $T_{que,1}$ is the queuing delay of the first arriving packet; and T_{jit} is the dejittering delay. T_{DSP} is lower-bounded by the sum of all look-aheads because no matter how dazzlingly fast DSPs become, the look-ahead delays will remain. $T_{net,min}$ is lower-bounded by the total propagation delay.

For packetized phone calls, the mouth-to-ear delay in one direction is not necessarily the same as that in the reverse direction; each of the terms in equation 1 can differ from one direction to the other.

Distortion stems from encoding the voice signal and from packet loss in the network $P_{loss,net}$ or in the dejittering buffer $P_{loss,jit}$. For values of interest — that is, small packet loss values of the order 10^{-2} or smaller — the total packet loss ratio nearly equals the sum of the packet losses in the network and the dejittering buffer. Packet loss (and even the codec format) can also differ from one direction to the other.

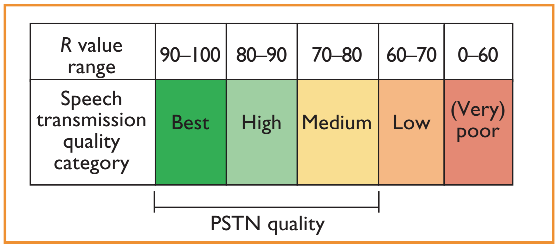
Parameters Determining Call Quality

Now let's look at how the quality of voice calls, mainly determined by mouth-to-ear delay and distortion, can be predicted using a standardized model originally developed by a European Telecommunications Standards Institute (ETSI) group — the E-model.

E-Model

The E-model (http://portal.etsi.org/stq/presentations/emodel.pdf) is a tool for predicting how an "average user" would rate the voice quality of a phone call with known characterizing transmission parameters. Similar proprietary models exist,[5] but the E-model has the advantage of being standardized in *Recommendation ITU-T G.107*.[6]

Based on an extensive set of subjective experiments, the E-model's developers defined an addi-



R value range	90–100	80–90	70–80	60–70	0–60
Speech transmission quality category	🟩 Best	🟩 High	🟨 Medium	🟧 Low	🟥 (Very) poor

PSTN quality

Figure 2. Phone call quality classes. Recommendation ITU-T G.109 maps call quality ratings to five classes.

tive rating scale R that assesses the quality of a phone call by quantifying the various transmission impairments. The scale identifies four types of impairments and an expectation factor:

$$R = R_0 - I_s - I_d - I_e + A. \qquad (2)$$

R_0 groups the effects of noise; it is a function of, among other things, the circuit noise level and the effective room noise level on both ends of the call. I_s includes impairments that occur simultaneously with the voice signal, such as those caused by quantization, by a too-loud or a too-soft connection, and by a nonoptimal side tone. I_d comprises delay impairments, including impairments caused by talker and listener echo and by a loss of interactivity. It is mainly a function of the echo's level and delay with respect to the original signal and the mouth-to-ear delays in both directions. I_e covers impairments caused by what G.107 calls "the use of special equipment" and groups effects that arise from distortion. It is a function of the type of low-bit-rate codec used and the fraction of lost packets. A is the expectation factor, which expresses the decrease in rating R that a user is willing to tolerate because of the system's access advantage over traditional wire-bound telephony. For example, for mobile telephony (such as GSM), A ranges between 5 and 10.[6]

Based on the R rating, we can predict subjective user reactions, such as what mean opinion score (MOS) a judging panel would award the call. Moreover, as defined in recommendation G.109,[7] the R rating maps to certain quality classes (see Figure 2): An R from 90 to 100 corresponds to best quality; 80 to 90 is high quality; 70 to 80 is medium; 60 to 70 is low; and 50 to 60 is poor. A rating below 50 indicates unacceptable quality. Throughout this article, we have color-coded these classes as in Figure 2.

From equation 2, we can see that two calls with the same R rating can give totally different subjective impressions. One call might produce crystal-clear, undistorted speech ($I_e = 0$) but suffer from a relatively large delay ($I_d = 10$). Another call might slightly distort the speech ($I_e = 10$), but its delay might not be noticeable ($I_d = 0$). The E-model predicts that a judging panel will award the same MOS to both calls, albeit for different reasons.

Now let's use the E-model's I_d and I_e factors to study the impact of one-way mouth-to-ear delay and distortion on the quality of a packetized phone call. Other factors, such as background noise and a connection that is too loud (which show up in R_0 and I_s), also impair call quality, but we won't consider these because they are not fundamentally different from what they would be in a traditional PSTN call. Furthermore, because our objective is to make a fair comparison between the quality of packetized and traditional wire-bound PSTN calls, we will set expectation factor A to zero.

Consider a packetized phone call between party 1 and party 2. Based on the E-model, let's evaluate how party 1 will judge the call in terms of R.

Mouth-to-Ear Delay

If there is some delay from party 1 to party 2 and vice versa, R decreases by an amount equal to impairment I_d. Recall that I_d is the sum of three contributing impairments arising from talker echo, listener echo, and loss of interactivity. The impairments associated with talker and listener echo depend on the delay and the level of those echoes with respect to the original signal. We assume that no echoes are introduced in the middle of the network — that is, any echoes originate in devices very close to the calling parties. In particular, hybrid or electrical echoes find their origin at the 4-to-2-wire hybrids (devices that convert 4-wire transport to 2-wire transport) of an analog PSTN (due to an impedance mismatch); acoustic echoes are due to the acoustic coupling between the speakers and the microphone at the user terminals. Thus, only the mouth-to-ear delays $T_{M2E,12}$ (from party 1 to party 2) and $T_{M2E,21}$ (from party 2 to party 1) play a role. Remember that in a packet-based environment these two delays can differ.

Talker echo disturbs party 1, who hears an attenuated and delayed echo of his/her own words $T_{M2E,12} + T_{M2E,21}$ after uttering them. This echo is caused by a reflection close to party 2. This echo is attenuated by $SLR_1 + RLR_2 + EL_2$ (expressed in dB) with respect to the original signal (see Figure 3). Here, EL_i is the echo loss close to party i — that is, how much quieter, in decibels, the echo is than the original signal (measured with respect to a certain reference point); SLR_i, the send loudness rating, is the attenuation of the signal from party i to the reference point; RLR_i, the receive loudness rat-

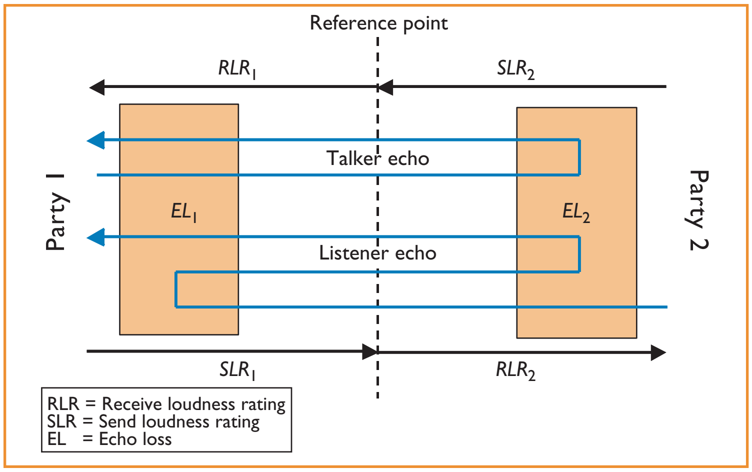


Figure 3. Talker and listener echo. This graph illustrates how and where echoes are produced.

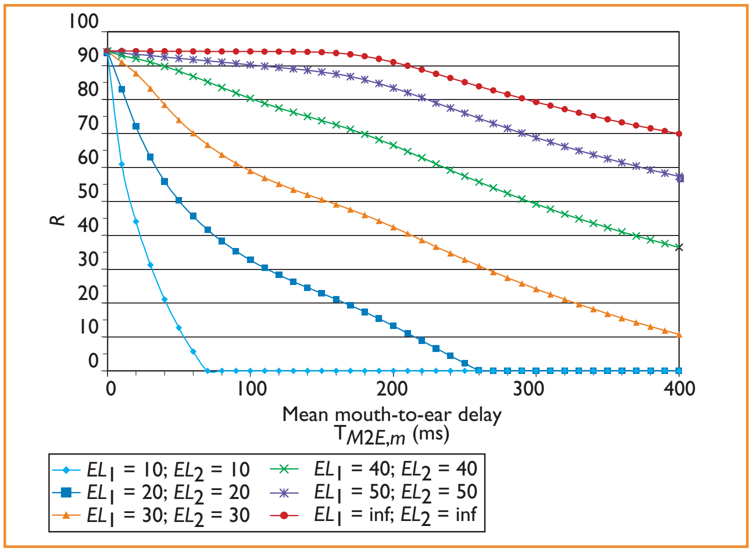


Figure 4. Call-quality rating. This graph shows rating R as a function of mean mouth-to-ear delay $T_{M2E,m}$ for undistorted voice and echo loss values between 10 dB and infinity.

ing, is the attenuation from the reference point to party *i*.[8] In this example, we take $SLR_1 = SLR_2 = 7$ dB and $RLR_1 = RLR_2 = 3$ dB, which are common values for these attenuations.

Listener echo also disturbs party 1 who hears the original signal from party 2 followed by an attenuated echo of this signal $T_{\mathrm{M2E,12}} + T_{\mathrm{M2E,21}}$ after the original signal. The level of this echo is determined by a reflection close to party 1 with attenuation EL_1, followed by a reflection close to party 2 with attenuation EL_2. As the original signal is attenuated by $SLR_2 + RLR_1$, the listener echo is attenuated $EL_1 + EL_2$ dB more.

As we mentioned, echo can occur in a hybrid (if the packetized phone call is terminated over a local PSTN) or in the caller's user terminal. For PSTN calls from traditional handsets, where the chief cause of echo is the 4-to-2-wire hybrids, a typical value for the echo loss is of the order of 20 dB.[8] The same value is valid for packetized phone calls that are terminated via a gateway over a local loop to a traditional handset. Usually, there is only a limited amount of acoustic echo introduced in traditional handsets and IP phones. In other kinds of terminals, such as PCs and hands-free phones, we can expect much more acoustic echo (resulting in an echo loss of, for example, 10 dB).

An echo controller increases echo losses EL_1 and EL_2. A standards-compliant echo controller[9] should increase the echo loss by at least 30 dB. At moderate computational cost, we can achieve perfect echo control, increasing echo losses EL_1 and EL_2 to infinity. Because it gradually gets more difficult to control the echo as it is more delayed with respect to its original signal, the echo controller should be deployed as close to the source of echo as possible. Hence, we recommend that the echo controller in the gateway compensate for the echo generated in the hybrids of the PSTN over which the call is terminated, and that the echo controller in the terminal compensate for the acoustic echo this terminal generates.

The third delay-related factor that might disturb party 1 is the loss of interactivity. If the mouth-to-ear delays are too large, an interactive conversation becomes impossible. The impairment associated with the loss of interactivity is a function of the mean mouth-to-ear delay,

$$T_{\mathrm{M2E,m}} = (T_{\mathrm{M2E,12}} + T_{\mathrm{M2E,21}})/2. \tag{3}$$

Figure 4 depicts the behavior of I_d as a function of $T_{\mathrm{M2E,m}}$, EL_1, and EL_2 for undistorted voice ($I_e = 0$). For simplicity, we consider the case where the echo loss values at both end points are equal. Rating R drops as the mouth-to-ear delay increases. In particular, the impairment associated with delay is strongly influenced by the echo loss value: the lower the echo loss value, the lower the quality for a specific delay value.

We define a phone call's intrinsic quality as the R rating associated with a zero mouth-to-ear delay. For undistorted phone calls (for example, those transported without packet loss in the G.711 format) with all other parameters optimally tuned, this intrinsic quality corresponds to an R rating (referred to as $R_{\mathrm{int,G.711}}$) of about 94.

We can also see from Figure 4 that for an echo loss of 20 dB, R drops below 70 at a mouth-to-ear delay of 25 ms, while for calls with perfect echo control, R drops below 70 at a mouth-to-ear delay of 400 ms. Recommendations ITU-T G.114 and G.131 mention the same bounds for undistorted (PSTN) calls;[1,2] thus, we conclude that the notion of "PSTN quality" relates to R ratings of 70 or higher. Moreover, a phone call retains its intrinsic quality up to a mean mouth-to-ear delay of about 150 ms if the echo is perfectly controlled (EL_1 = EL_2 = infinity), which also corresponds to the G.114 and G.131 quality statements.[1,2]

Compression and Packet Loss

If the voice signal that reaches party 1 (in Figure 3) is distorted, R decreases by an amount equal to the distortion impairment I_e. This impairment is a function of at least two parameters: the codec used by party 2 to encode the voice signal and packet loss P_{loss} during the transport of voice packets from party 2 to party 1. It is common practice, but not strictly mandatory, to transport the voice in the same format in both directions.

First let's consider the influence of compressing the voice signal. As we mentioned, the G.711 codec introduces almost no distortion by sampling and quantizing the voice signal. The packetization delay can be any multiple of 0.125 ms.

Predictive codecs (the G.726 codec, for example) predict the sample to be encoded based on the previous, already-encoded samples and quantize the prediction error in 2, 3, 4, or 5 bits, resulting in a codec bit rate of 16, 24, 32, or 40 Kbps. Again, the packetization delay can be any multiple of 0.125 ms.

Codecs of the vocoder type are based on a model for the human vocal track. These codecs first segment the speech signal in intervals of constant duration called voice frames. Then, for each consecutive voice frame, they estimate and quantize the parameters of the vocal track model and collect all quantized parameters in a code word. The codec bit rate equals the code word size divided by the frame length. Some of these codecs require a look-ahead to estimate the vocal track model parameters more accurately. The packetization delay is an integer multiple of the voice frame. Most vocoder codecs have a frame length between 10 and 30 ms — the G.729 codec has 10 ms; the G.723.1 codec, 30 ms; and all GSM codecs, 20 ms. An exception is the G.728 codec, which has a voice frame length of 0.625 ms.

Figure 5 summarizes the distortion impairment associated with some standardized codecs. The points on this figure are rate-distortion pairs determined by experiments reported in Appendix I of recommendation G.113.[10] The three lines connect similar pairs — a straight line for two pairs and quadratic best-fitting curves for more than two pairs. The solid black line, associated with the G.726 codec, gives the rate-distortion tradeoff for predictive codecs. We see that at low bit rates, predictive codecs introduce a lot of distortion. The dashed black line corresponds to the G.728 codec, and the solid gray line is drawn through the state-of-the-art vocoder-type codecs — G.729, G.723.1, and global system for mobile communication/enhanced full rate (GSM-EFR). The G.728 codec has a better rate-distortion tradeoff than predictive codecs but does not reach the full potential of vocoder-type codecs, as its voice frame size is too small. The older GSM-FR and GSM-HR codecs do not reach the full potential of vocoder codecs, either.

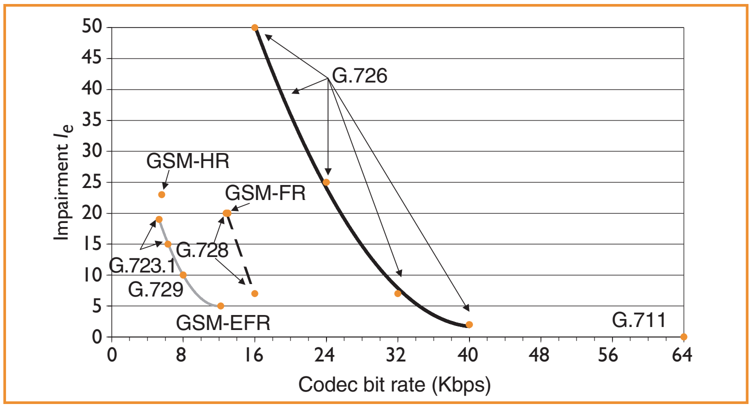


Figure 5. Distortion impairment I_e for several standardized codecs. This figure shows for several codec families how the distortion decreases for increasing codec rates.

A voice activity detection (VAD) scheme, which detects whether the signal contains active speech or background noise, can further reduce the overall bit rate to be sent. Good VAD schemes introduce almost no additional distortion.

The distortion impairment I_e associated with a codec increases as the packet loss ratio increases. Figure 6 shows the quadratic curves that best fit the experimental data[10] (shown as points) for four codecs, assuming that voice packets are lost at random. Figure 6 deals only with one specific packetization interval per codec: 10 ms for G.711, 20 ms for G.729, and GSM-EFR 30 ms for G.723.1 at 6.3 Kbps. Although we do not yet have results for other codecs and packetization intervals, we can identify some trends.

The sensitivity to packet loss depends on the packet loss concealment (PLC) technique that the codec uses. In contrast to the G.711 codec, most state-of-the-art, low-bit-rate codecs (such as G.729, G.723.1, and GSM-EFR) have built-in PLC schemes. However, a standardized PLC scheme can

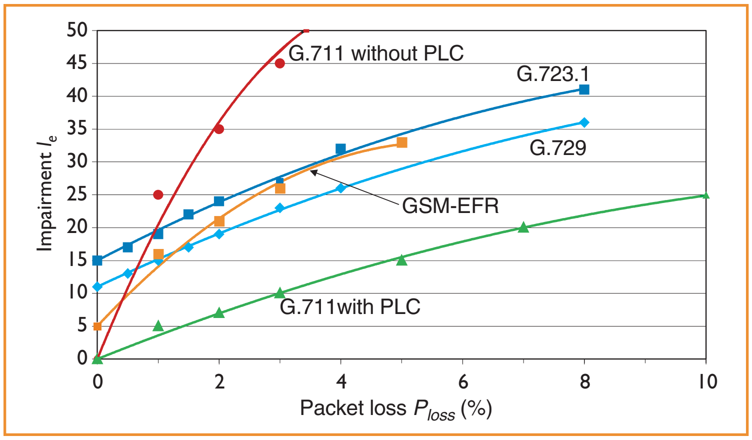


Figure 6. Distortion impairment I_e as a function of packet loss P_{loss}. This figure shows for several codecs how the distortion increases for increasing packet loss ratios.

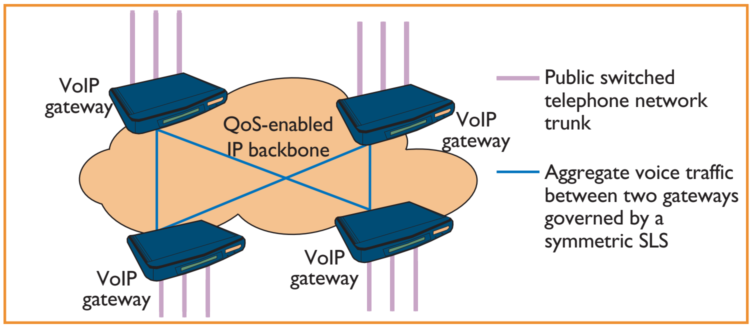

VoIP gateway
QoS-enabled IP backbone
VoIP gateway
VoIP gateway
VoIP gateway
Public switched telephone network trunk
Aggregate voice traffic between two gateways governed by a symmetric SLS

Figure 7. Gateway-to-gateway scenario for phone call transport. Voice packets travel over an IP backbone between gateways that switch to local PSTNs and terminate at traditional telephone sets.

be implemented on top of the G.711 codec, introducing 3.75 ms of additional delay.[11]

From Figure 6 we see that for the codecs that use PLC, impairment increases by about 4 units on the *R* scale per percent packet loss (for low loss values). Observe that the GSM-EFR codec's PLC technique performs worse than the PLC techniques of the other codecs considered. If no PLC scheme is implemented atop the G.711 codec, the distortion impairment increases by 25 units on the *R* scale for each percent packet loss (for low loss values).

In this respect, it is interesting to note that an average phoneme (the smallest unit of speech in any particular language) in human speech is 80 ms.[12] If it is transported in one packet, a single packet loss can eliminate the whole phoneme, and PLC techniques lack the information to compute an approximation of the lost information. Therefore, it is advisable to use packetization delays shorter than 80 ms. When the packetization delay is much smaller than 80 ms, PLC techniques might conceal even consecutive (bursty) packet losses.

Other techniques to repair losses do exist. They all boil down to transmitting additional information in future packets — FEC or the speech coded at a lower bit rate, and hence, lower quality.[12] As such, these techniques introduce additional delay and overhead bit rate.

The voice signal need not be transported in the same format end to end. Somewhere along the route, the voice signal might be transcoded from one codec format into another. Because all the standard codecs we have considered need an 8-kHz stream of uniformly quantized voice samples at the input, the code words of the first codec must be decoded before the signals can be encoded into another codec format. Consequently, the impairment terms associated with the two codecs should be added to obtain the overall distortion impairment I_e, because, in the E-model, impairments are approximately additive on the *R* scale. For example, using both the GSM-EFR and the G.729 codec at 8 Kbps for one phone call leads to an intrinsic quality of $R = 94 - 5 - 10 = 79$. We have published the intrinsic *R* values associated with all combinations of two codecs elsewhere.[13]

Case Study

From our discussion so far, we conclude that we can write the *R* rating as

$$R = R_{\text{int,G.711}} - I_d(T_{\text{M2E,m}}, EL_1, EL_2) - I_e(\text{codec}, P_{\text{loss}}). \tag{4}$$

That is, given the mean mouth-to-ear delay, the echo loss values, the codec used, and the packet loss encountered, we can predict the quality of the corresponding phone call. Figure 4 shows the combined effect of the first and second terms; Figure 6 displays the third term.

From Figure 4, we can also derive the tolerable bounds on the mouth-to-ear delay and distortion when we are aiming for PSTN quality (*R* greater than 70). Indeed, because the intrinsic quality $R_{\text{int,G.711}}$ of an undistorted call equals 94 and the bound for traditional quality corresponds to 70, we have an impairment budget of 24 on the *R* scale, part of which is consumed by the codec (see Figure 5). Once we have chosen the codec, the remainder of this impairment budget can serve to allow some packet loss or mouth-to-ear delay. Obviously, the echo loss value has a great influence on the tolerable mouth-to-ear delay bound. Elsewhere, we report the tolerable mouth-to-ear delay and packet loss bounds for perfect echo control and different codecs, under the additional assumption that either delay or packet loss is occurring, but not both simultaneously.[13]

Because practical network configurations will introduce both mouth-to-ear delay and packet loss, we'll now look at a case study for a gateway-to-gateway scenario. As depicted in Figure 7, phone calls originate from and terminate at traditional telephone sets and are switched over a local PSTN to voice-over-IP gateways, between which the voice signals are transported over a QoS-enabled IP backbone administered by one network manager.

Assume that between each pair of gateways, voice packets can be sent over this IP backbone. A symmetric traffic contract or service-level specification (SLS) governs the transport of aggregate voice traffic between any specific pair of gateways (in both directions). More precisely, we assume that the SLS specifies values for $P_{\text{loss,net}}$ and $T_{\text{net,min}}$, as well as a maximum value or some $(1 - P)$-quantiles of the total queuing delay.

Our concrete example is a phone call from Europe to the U.S. We assume that the SLS specifies that there is no packet loss in the backbone ($P_{\text{loss,net}} = 0$) and that the minimal network delay depends primarily on the propagation delay. Hence, we can determine delay $T_{\text{net,min}}$ once we know the physical distance between the gateways. Assuming that 10,000 km of cables must be propagated at a speed of 200 km/ms, we end up with a minimal network delay of 50 ms.

Table 1 lists the queuing delay quantiles specified in an example of such an SLS. These are typical values for a high-speed network where very occasionally — with a probability lower than 10^{-3} — signaling traffic of higher priority can block a voice queue.

Let's assume that we're using the G.729 codec at 8 Kbps and consider packetization delays of 10, 20, 30, and 40 ms. Because complete human phonemes should be transported in multiple packets for the PLC technique to have an optimal effect, we must avoid longer packetization delays. Another gateway parameter that we can tune is the tolerated dejittering loss, $P_{\text{loss,jit}}$. We assume that our call uses a dynamic dejittering algorithm, which, as we described earlier, tunes the dejittering delay so that eventually, the sum of this dejittering delay and the first packet's queuing delay $T_{\text{jit}} + T_{\text{que},1}$ becomes the $(1 - P_{\text{loss,jit}})$-quantile of the total queuing delay. Finally, for the last unidentified delay component of equation 1, the DSP delay, we'll suppose $T_{\text{DSP}} = 15$ ms.

For echo control, we'll consider two cases: $EL_1 = EL_2 = 40$ dB, which represents 20 dB of a 4-to-2 wire hybrid and an additional 20 dB of a poor echo controller; and $EL_1 = EL_2 =$ infinity, which represents perfect echo control.

Table 1. Queuing delay quantiles in SLS specification.

Packet Loss	(1 – *P*)-quantile (ms)
10^{-1}	1
10^{-2}	3
10^{-3}	10
10^{-4}	30
10^{-5}	130

Table 2. Effective codec bit rate for G.729 codec.

Packetization delay T_{pack} (ms)	Effective codec bit rate (Kbps)
10	40
20	24
30	18.7
40	16

The overhead S_{OH} per voice packet equals 320 bits (20 IP, 8 UDP, and 12 RTP bytes), resulting in an overhead bit rate between 32 Kbps and 8 Kbps for the considered range of packetization delays. Table 2 gives the effective bit rate. Figure 8 presents rating R calculated with equation 4 for various values of the packetization delay and dejittering loss, using the color code from Figure 2.

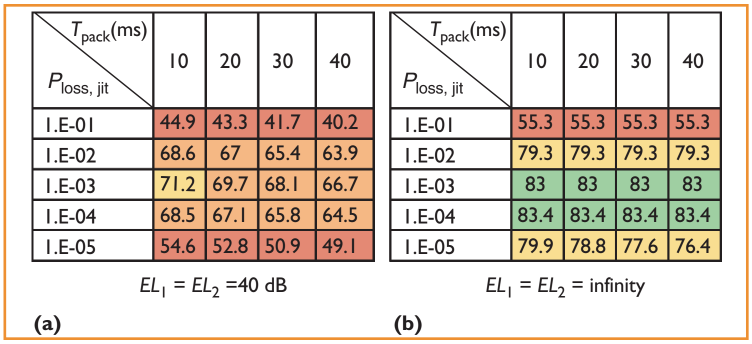
For $EL_1 = EL_2 = 40$ dB, we observe that the call can achieve medium PSTN quality only for a dejittering loss value $P_{\text{loss,jit}} = 10^{-3}$ and a packetization delay $T_{\text{pack}} = 10$ ms, resulting in a rather large effective bit rate of 40 Kbps. To achieve more efficient voice transport by choosing a larger packetization delay, we must settle for lower quality, dropping below the PSTN level.

If our call has perfect echo control, however, high quality is guaranteed for $P_{\text{loss,jit}} = 10^{-3}$ or 10^{-4} and all the packetization delays considered. Larger values of T_{pack} obviously result in the highest transport efficiency.

Conclusions

To bring PSTN-level voice quality to packetized telephony, controlling both delay and distortion is crucial. Our quantitative study of these impairments, through the standard E-model, led us to several conclusions:

- For packetized phone calls, echo control is highly recommended if not required. Other-



T_{pack}(ms) / $P_{loss, jit}$	10	20	30	40
1.E-01	44.9	43.3	41.7	40.2
1.E-02	68.6	67	65.4	63.9
1.E-03	71.2	69.7	68.1	66.7
1.E-04	68.5	67.1	65.8	64.5
1.E-05	54.6	52.8	50.9	49.1

$EL_1 = EL_2$ =40 dB

(a)

T_{pack}(ms) / $P_{loss, jit}$	10	20	30	40
1.E-01	55.3	55.3	55.3	55.3
1.E-02	79.3	79.3	79.3	79.3
1.E-03	83	83	83	83
1.E-04	83.4	83.4	83.4	83.4
1.E-05	79.9	78.8	77.6	76.4

$EL_1 = EL_2$ = infinity

(b)

Figure 8. Phone call quality rating R for the G.729 codec as a function of the packetization delay T_{pack} and dejittering loss $P_{loss,jit}$, according to the SLS specified in Table 1. The quality ratings are color coded according to Figure 2 and calculated for (a) 40-dB echo loss and (b) perfect echo control.

wise, the tolerable mouth-to-ear delay budget risks being too small to attain PSTN quality.

- The intrinsic quality associated with predictive codecs at low bit rates (below 32 Kbps) is lower than traditional PSTN quality. Thus, such codecs should not be used. Transcoding, also, should be avoided.
- The packet loss ratio (in the network and in the dejittering buffer) should be kept under control — say, below 10^{-2}. Packet loss concealment techniques definitely increase the robustness against packet loss considerably.

Our case study of a gateway-to-gateway scenario, where a service-level specification governs transport of the voice packets, illustrates how the E-model can help guide the choice of transmission parameters.

Acknowledgments

This work was carried out within the framework of the CoDiNet project sponsored by the Flemish Institute for the Promotion of Scientific and Technological Research in the Industry (IWT).

References

1. *Recommendation ITU-T G.114, One-Way Transmission Time,* Int'l Telecommunication Union, Geneva, 1996.
2. *Recommendation ITU-T G.131, Control of Talker Echo,* Int'l Telecommunication Union, Geneva, 1996.
3. F. Poppe, D. De Vleeschauwer, and G.H. Petit, "Choosing the UMTS Air Interface Parameters, the Voice Packet Size, and the Dejittering Delay for a Voice-Over-IP Call between a UMTS and a PSTN Party," *Proc. IEEE Infocom,* IEEE CS Press, Los Alamitos, Calif., 2001, pp. 805-814.
4. S.B. Moon, J. Kurose, and D. Towsley, "Packet Audio Play-Out Adjustment: Performance Bounds and Algorithms," *Multimedia Systems,* vol. 6, no. 1, Jan. 1998, pp. 17-28.
5. N.O. Johannesson, "The ETSI Computation Model: A Tool for Transmission Planning of Telephone Networks," *IEEE Comm. Magazine,* vol. 35, no. 1, Jan. 1997, pp. 70-79.
6. *Recommendation ITU-T G.107, The E-model, A Computational Model for Use in Transmission Planning,* Int'l Telecommunication Union, Geneva, 1998.
7. *Recommendation ITU-T G.109, Definition of Categories of Speech Transmission Quality,* Int'l Telecommunications Union, Geneva, 1999.
8. *ETSI Guide 201 050, Speech Processing, Transmission, and Quality Aspects (STQ); Overall Transmission Plan Aspects for Telephony in a Private Network,* European Telecommunication Stds. Inst., Sophia Antipolis, France, 1998.
9. *Recommendation ITU-T G.168, Digital Network Echo Cancellers,* Int'l Telecommunication Union, Geneva, 1997.
10. *Recommendation ITU-T G.113, Appendix I, Provisional Planning Values for the Equipment Impairment Factor I_e,* Int'l Telecommunication Union, Geneva, 1999.
11. *Recommendation ITU-T G.711, Appendix I, A High Quality Low-Complexity Algorithm for Packet Loss Concealment with G.711,* Int'l Telecommunication Union, Geneva, 1999.
12. V. Hardman et al., "Reliable Audio for Use over the Internet," *Proc. Int'l Networking Conf.,* Internet Soc., Reston, Va., 1995; available at www.isoc.org/HMP/PAPER/070/html/paper.html.
13. J. Janssen, D. De Vleeschauwer, and G.H. Petit, "Delay and Distortion Bounds for Packetized Voice Calls of Traditional PSTN Quality," *Proc. First IP-Telephony Workshop,* GMD Report 95, GMD, Berlin, 2000, pp. 105-110.

Jan Janssen is a research engineer in Alcatel's Network Strategy Group, Antwerp, Belgium. His main research interests are performance issues for packet-based voice/multimedia transport. He received an MSc and PhD in sciences (mathematics) from the Katholieke Universiteit Leuven, Belgium.

Danny De Vleeschauwer is a research engineer at Alcatel. His main research interests are signal processing, queuing theory, and multimedia over packet-based networks. He received the engineering degree and a PhD in applied sciences from the University of Ghent, Belgium.

Maarten Büchli is a research engineer at Alcatel. His main research interests are in the area of quality-of-service architectures and performance issues for packet-based voice. He holds an MSc in electrical engineering from the University of Twente, the Netherlands.

Guido H. Petit is director of the Network Architecture Department at Alcatel, and a part-time visiting professor at the University of Ghent, Belgium. He holds an MSc and PhD in chemistry from the University of Antwerp, Belgium. He is an IEEE associate.

Readers can contact the authors at {jan.janssen,danny.de_vleeschauwer,maarten.buchli,guido.h.petit}@alcatel.be.