QUALITY ISSUES FOR PACKET-BASED VOICE TRANSPORT

D. De Vleeschauwer, A. Van Moffaert, M.J.C. Büchli, J. Janssen, G.H. Petit
Alcatel Bell, Network Strategy Group
Francis Wellesplein 1
B-2018 Antwerp, Belgium
E-mail: {danny.de_vleeschauwer, annelies.van_moffaert, maarten.buchli, jan.janssen, guido.h.petit}@alcatel.be
Phone: +32(0)32408196. Fax: +32(0)32404888

Abstract

This paper studies the influence of the mouth-to-ear delay and distortion (due to voice compression and packet loss) on the quality of a phone call, since these parameters are likely to be larger when the call is transported over a packet-based network instead of over a circuit-switched network. First, the need for echo controlling packetized phone calls is discussed. Second, it is shown that some codecs, in particular predictive codecs, do not attain high enough quality at low bit rates. In the same context also the potential danger of transcoding is recognized. Third, the merit of a packet loss concealment technique to considerably increase the robustness against packet loss is demonstrated. Next, the bounds on the mean one-way mouth-to-ear delay and packet loss that need to be respected in order to attain traditional PSTN quality, are derived for standard codecs (even the recently developed Adaptive MultiRate (AMR) codec) and various levels of echo control (i.e. perfect echo control and standard-compliant echo control). Finally, a gateway-to-gateway scenario in which the transport between the gateways is governed by a Service Level Specification (SLS), is discussed and a numerical example is given to show how the quality bounds can be met in this scenario by tuning the gateway parameters correctly.

1 Introduction

The quality of a telephone call depends on the parameter settings in the user terminal and on the parameters of the network over which the call is transported. In this paper, we assume that the user terminals are optimally tuned and study the influence of the network parameters. Offering quality to telephone calls transported over a Public Switched Telephone Network (PSTN) has been understood already for a long time. The main topic of this paper is to investigate how quality can be offered for calls (partly) transported over packet-based networks.

Although currently most of the core of the PSTN is digital, the access parts (e.g. the local loop) are in a lot of cases still analog. There are exceptions however, where even the access is digital, e.g., ISDN access and GSM access. In the 4-to-2-wire hybrids of those analog access parts hybrid echo may be introduced. Additionally, acoustic echo may also be introduced in the user terminals (even when the transport is digital end-to-end). In any case, the level of the echo can be controlled with an echo controller (see ITU-T Recommendation G.168 [3].

In the PSTN the one-way mouth-to-ear delay mainly consists of propagation delay and switching delay, and hence, it is practically completely determined by the physical distance between both calling parties. An exception is GSM access, where the transport over the air interface alone already introduces about 100 ms of delay [7].

The analog access part of a PSTN is nowadays so short that the distortion introduced in that part of the network is negligible. Over the core of a PSTN the voice signal is (mostly) transported in the G.711 codec format, a format that only introduces a negligible amount of distortion with respect to the analog format. Hence, for most telephone calls transported over a PSTN there is practically no (additional) distortion involved. There are exceptions however, where some distortion is introduced by signal compression: on some transoceanic links the voice is sometimes compressed and in GSM access the voice is transported in a compressed format over the air interface.

When there is little distortion of the voice signal (and when optimally tuned user terminals are utilized), the level of the echo and the one-way mouth-to-ear delay mainly determine the quality of telephone calls transported over a PSTN. It is known that some echo and some delay can be tolerated. ITU-T Recommendations G.114 [1] and G.131 [2] specify the mouth-to-ear delay that can be tolerated (for undistorted voice and) for the case with and without echo control.

The packet-based transport of telephone calls is more flexible than the transport over a PSTN. A packet-based network is not so tightly bound to one codec as the PSTN is to the G.711 codec (which only takes frequencies up to 3.1 kHz into account and has a bit rate of 64 kb/s). Any codec that both user terminals support can be utilized. Wide-band codecs (which take frequencies in the speech signal below 7 kHz into account) could be used to improve the intelligibility of the speech. Remark that the bit rate of such a codec is not necessarily higher than the 64 kb/s of the G.711 codec (as the G.711 codec is not very efficient). However, in this paper we only consider low-bit-rate narrow-band codecs, i.e., codecs that like the G.711 only take the frequencies up to 3.1 kHz

into account but compress the voice signal to a smaller bit rate than 64 kb/s, possibly at the expense of the introduction of some distortion. On top of this bit rate reduction Voice Activity Detection (VAD) can easily be exploited in packet-based networks, whilst in a PSTN this is impossible.

The price to pay for this additional flexibility is additional complexity: more delay and distortion are likely to be introduced. On top of the delays that also occur in the PSTN, packetization, codec, queuing and dejittering delay come into play [10]. Moreover, the mouth-to-ear delays may considerably differ from one direction to the other, a fact that (practically) never occurs in a PSTN. Distortion may stem from the use of a low-bit-rate codec or from the loss of voice packets in the network or the dejittering buffer. Fortunately, as will be shown in this paper, the one-way mouth-to-ear delay(s) and the distortion can be kept under control by tuning the devices in the network properly.

In the next section we first point out in what a packetized phone call differs from a phone call switched over a PSTN as far as quality is concerned. Section 3 quantifies how the echo level, the mouth-to-ear delay(s) and the distortion (through encoding and packet loss) influence the quality of a telephone call by means of the E-model. In Section 4 we present a method to tune the parameters such that adequate quality is attained, when the characteristics with which the voice packets are transported are known, e.g. through a Service Level Specification (SLS). Finally, in the last section we draw the main conclusions.

2 Principles of the packetized transport of phone calls

As illustrated in Figure 1 there are three essential stages in the packetized transport of phone calls.

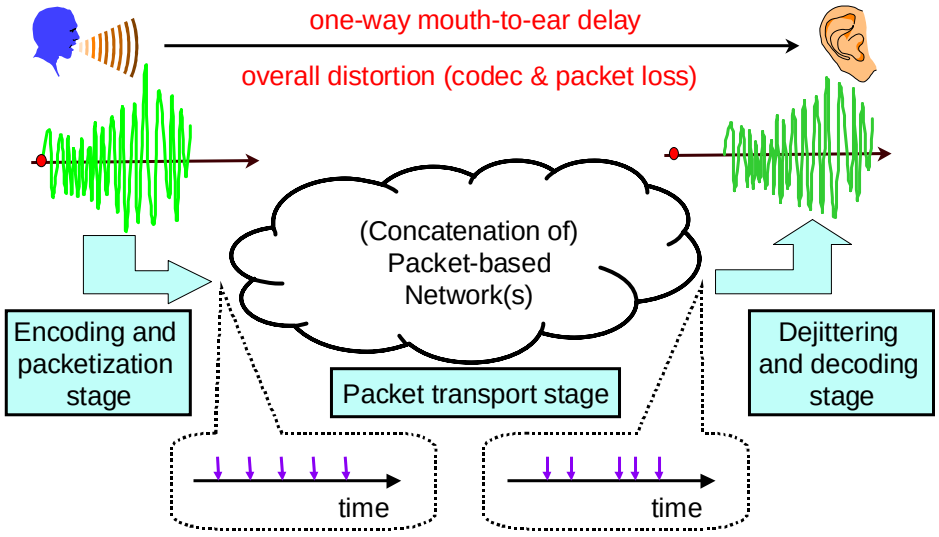


Figure 1: Three essential stages in the packetized transport of phone calls.

In the first stage, the digital voice signal (i.e., a voice signal lowpass-filtered with cut-off frequency at 3.1 kHz that is sampled at 8 kHz and quantized with a linear 13-bit quantizer) is encoded and packetized. This packetization and encoding operation can be performed either in the user terminal or in a gateway. In the latter case we assume that the transport of the voice signal from the user terminal to the gateway (possibly over an analog access network) merely introduces a negligible amount of delay and distortion.

The packetization delay T_{pack} is defined as the time needed to collect all voice samples that end up in one packet, and as such scales linearly with the payload size. The choice of the packetization delay is a trade-off between efficiency (the larger the packets, the smaller the relative influence of overhead bytes) and delay. In fact the effective bit rate R_{eff} that is needed to transport a voice flow over a packet-based network is defined as

$$R_{eff} = R_{cod} + \frac{S_{OH}}{T_{pack}} \quad , \tag{1}$$

where R_{cod} is the net codec bit rate and S_{OH} number of overhead bits per voice packet.

Also the encoding performed by a Digital Signal Processor (DSP) needs some time. Besides the voice encoding process other processes run on the DSP as well. An example is an algorithm that detects whether or not the incoming signal is a pure speech signal or consists of (fax, modem or DTMF) tones in order to bypass the voice encoder in the latter case. Such an algorithm needs to collect a few samples, as it cannot make an instantaneous

decision based on only one sample. This process introduces delay referred to as look-ahead delay. Some encoders themselves already introduce a similar look-ahead delay.

In the second stage, this flow of packets is transported over a packet-based network consisting of several access and backbone nodes. In the transport of the voice flow over this network some delay is incurred. The network delay can be split into two parts: a deterministic part, referred to as the minimal network delay $T_{net,\min}$, and a stochastic part, referred to as the total queuing delay. The minimal network delay mainly consists of the propagation delay (of 5 μs per km), the sum of all serialization delays, the route look-up delay, etc. If somewhere the packets are transported over an unreliable channel, e.g. an air interface, Forward Error Correction (FEC) techniques, like interleaving coupled with (Reed-Solomon) block or convolutional channel codes, also contribute an amount T_{FEC} to the minimal network delay.

The total queuing delay T_{que} is the sum of the queuing delay in each node. The queuing delay in one network node is due to the competition of several flows for the available resources in the queue of that node. The total queuing delay is responsible for the jitter introduced in the voice flow. The tail distribution function of the total queuing delay is defined as

$$F(T) = \text{Prob}\lfloor T_{que} > T \rfloor \quad . \tag{2}$$

Remark that the inverse of this function evaluated in P, i.e. $F^{-1}(P)$, gives the (1-P)-quantile of the total queuing delay.

In the transport over the network a fraction $P_{loss,net}$ of the packets may get lost. In the case where an unreliable medium (e.g. an air interface) is traversed, there exists a trade-off between packet loss in the network and FEC delay introduced in the network

$$P_{loss.net} = G(T_{FEC}) \quad . \tag{3}$$

The function $G(.)$ is non-increasing. For reliable channels $G(T_{FEC}) \equiv 0$, and there is no gain in choosing $T_{FEC} > 0$. In this paper we do not consider the transport over an unreliable medium, but refer the interested reader to [14] and [15].

In the last stage the jittered packet flow is dejittered and decoded. Since the decoder needs the packets at a constant rate, dejittering is absolutely necessary. Dejittering a voice flow consists of retaining the fastest packets in the dejittering buffer to allow the slowest ones to catch up. The fastest packets are the ones that do not have to queue in any of the nodes. So, in principle, the fastest packets have to be retained for a time equal to the maximal total queuing delay in the dejittering buffer. Because voice codecs can tolerate some packet loss and because waiting for the slowest packet frequently introduces too much delay, often the fastest packets are retained in the dejittering buffer for a time equal to the (1-P)-quantile of the total queuing delay. This means that a fraction P of the packets will be lost, because they arrive too late. This packet loss introduces distortion. Because it is usually not known if the first arriving packet is a slow or fast one, a static dejittering mechanism retains the first arriving packet a time T_{jit} in the buffer and then reads the buffer at a constant rate. Dynamic dejittering algorithms are able to gradually learn whether or not the first arriving packet was a fast or slow one and compensate in that way for the total queuing delay of the first packet.

The decoding and echo control processes finally also introduces some delay.

The dejittering, decoding and echo control can be performed either in the user terminal or in a gateway. In the latter case we assume that the transport of the voice signal from the gateway to the user terminal (possibly over an analog access network) again merely introduces a negligible amount of delay and distortion.

To conclude this section we bring together the impact of all stages on the one-way mouth-to-ear delay T_{M2E} and the overall packet loss P_{loss}.

First, we consider a packetized phone call that is statically dejittered. In that case the one-way mouth-to-ear delay (in one direction) can be split up in the following terms

$$T_{M2E} = T_{pack} + T_{DSP} + T_{net,\min} + T_{que,1} + T_{jit} \quad , \tag{4}$$

where 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 analog access parts if a gateway is involved and the delay T_{FEC} introduced by the scheme to protect the transport over an unreliable channel), $T_{que,1}$ is the total queuing delay of the first arriving packet and T_{jit} is the dejittering delay. The DSP delay T_{DSP} is lower bounded by the sum of all look-aheads, i.e., even if technology keeps evolving culminating in DSPs with a dazzling processing power, the look-aheads remain unaffected. The minimal network delay $T_{net,\min}$ is lower bounded by the total propagation delay. Since the total queuing delay $T_{que,1}$ of the first packet is

stochastic, the one-way mouth-to-ear delay of eq. (4) also is. For static dejittering mechanisms the dejittering delay T_{jit} is usually chosen on the safe side, i.e., such that in worst case (when the first arriving packet happens to be a fast one) at most a fraction $P_{loss,jit}$ of the packets get lost. Hence,

$$T_{jit} = F^{-1}\left(P_{loss,jit}\right) \quad . \tag{5}$$

Second, we consider a dynamically dejittered packetized phone call. When the dynamic dejittering mechanism is set to tolerate a packet loss of $P_{loss,jit}$, the dejittering delay is gradually adjusted to compensate for the total queuing delay $T_{que,1}$ of the first packet, so that after a transition period the one-way mouth-to-ear delay tends to

$$T_{M2E} = T_{pack} + T_{DSP} + T_{net,\min} + F^{-1}\left(P_{loss,jit}\right) \quad . \tag{6}$$

Comparing eq. (6) with eq. (4) combined with (5), we see that adaptive dejittering can (eventually) economize on the one-way mouth-to-ear delay by an amount equal to $T_{que,1}$.

Notice that for packetized phone calls the mouth-to-ear delay in one direction is not necessarily the same as that in the reverse direction as each of the terms in eq. (4) (or eq. (6)) may differ from one direction to the other.

Distortion stems from the encoding of the voice signal and from packet loss $P_{loss,net}$ in the transport over the network or from the packet loss $P_{loss,jit}$ in the dejittering buffer, i.e.,

$$P_{loss} = 1-(1-P_{loss,net})(1-P_{loss,jit}) \quad . \tag{7}$$

Notice that also the packet loss (and even the codec format) may differ from one direction to the other.

In the next section we determine how this one-way mouth-to-ear delay and this distortion impact the quality of the call.

3 Parameters determining the quality of a phone call

3.1 The E-model

The E-model is a tool to predict how an "average user" would rate a phone call of which the characterizing transmission parameters are known. Similar proprietary models exist (see the references in [16]), but the E-model has the advantage that it is standardized in ITU-T Recommendation G.107 [4]. Based on an extensive set of subjective experiments, a scale, referred to as the *R*-scale, was defined in [8] upon which impairments are approximately additive in the range of interest. Four types of impairments and an advantage factor were identified, that is,

$$R = R_0 - I_s - I_d - I_e + A \quad . \tag{8}$$

The first term R_0 groups the effects of noise and is amongst other a function of the level of the circuit noise and the (effective) level of the room noise (present at both sides). The second term I_s includes impairments that occur simultaneously with the voice signal, such as those caused by quantization, by too loud or too soft a connection and by a non-optimum side tone. The third term I_d comprises delayed impairments, including impairments caused by talker and listener echo or by a loss of interactivity. It is mainly a function of the level and the delay of the echo with respect to the original signal and the mouth-to-ear delays in both directions. The fourth term I_e covers impairments caused by what is referred to as "the use of special equipment" in ITU-T Recommendation G.107 and groups effects due to distortion. It is a function of the type of low-bit-rate codec used and the fraction of lost packets. The fifth term *A*, referred to as the expectation factor, expresses the decrease in rating a user is willing to tolerate because of the "access advantage" that certain systems have over traditional wire-bound telephony. As an example, the expectation factor *A* for mobile telephony (e.g. GSM) is 10.

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				

Table 1: Quality classes according to ITU-T Recommendation G.109.

Based on the rating *R* subjective user reactions can be predicted, such as the Mean Opinion Score (MOS) a judging panel would give to the call or the percentage of users finding the quality "Good or Better" (GoB). Moreover, as defined in ITU-T Recommendation G.109 [5] the rating *R* maps to certain quality classes: a rating

R in the ranges [90,100[, [80,90[, [70,80[, [60,70[, [50,60[corresponds to “best”, “high”, “medium”, “low” and “poor” quality, respectively. A rating below 50 indicates unacceptable quality. Throughout this paper, the classes are color coded according to Table 1.

In the next paragraphs we study the impact of the one-way mouth-to-ear delay(s) (via I_d) and the distortion (via I_e) on the quality of a packetized phone call. Other factors, like background noise and a connection that is too loud, also impair the quality (via R_0 and I_s) of a packetized phone call, but as these factors are not fundamentally different from a traditional PSTN call, they were not considered. Furthermore, as the objective was to make a fair comparison between the quality of packetized phone calls and traditional wire-bound PSTN calls, the expectation factor A was set to zero.

From Eq. (8) it can be seen that two calls with the same rating R can give a different subjective impression. One call might produce crystal clear, undistorted speech (e.g. $I_e = 0$) but suffer from a relatively large delay (e.g. $I_d = 10$). Another call might slightly distort the speech (e.g. $I_e = 10$), while its delay is not noticeable (e.g. $I_d = 0$). The E-model merely predicts that a judging panel will award the same MOS to both calls and the same percentage of users will find both calls GoB, albeit for different reasons.

Consider a packetized phone call between two parties, referred to as party 1 and party 2 (see Figure 2). Based on the E-model, we evaluate how party 1 will judge the call, that is, what rating R he will assign to it. The influence of delay is studied first, followed by the influence of distortion.

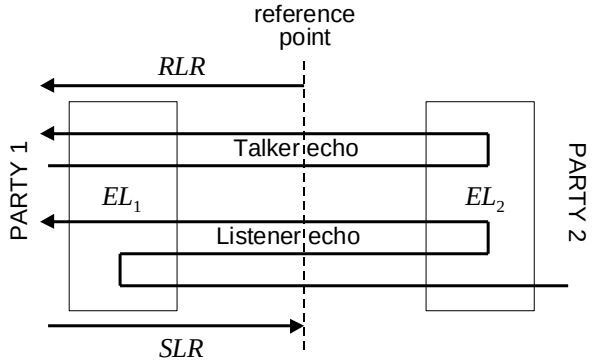


Figure 2: Talker and listener echo.

3.2 Influence of Mouth-to-Ear Delay

If there is some delay from party 1 to party 2 and vice versa, the rating R decreases by an amount equal to the impairment I_d. This impairment I_d is the sum of three contributing impairments: impairments due to talker echo, due to listener echo and due to the loss of interactivity. The impairment associated with talker and listener echo depends on the delay and the level of the respective echoes with respect to the original signal. We assume that the echoes (if any) are generated in devices (4-to-2-wire hybrids or user terminals) very close to the calling parties, i.e. that there are no echoes introduced somewhere in (hybrids in) the middle of the network. In that way only the mouth-to-ear delay $T_{M2E,12}$ from party 1 to party 2 and the one $T_{M2E,21}$ from party 2 to party 1 play a role. Remember that in a packet-based environment these two delays may differ.

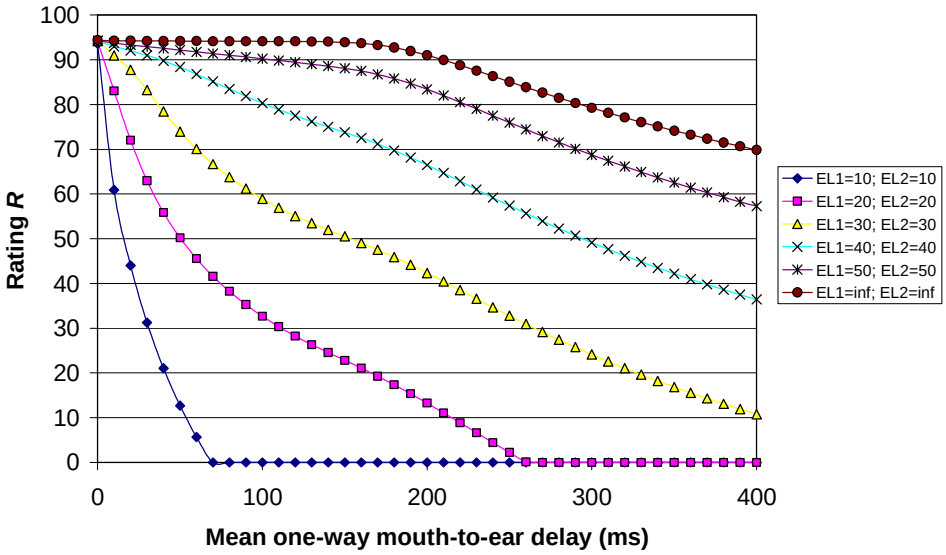
Talker echo disturbs party 1, who hears an attenuated and delayed echo of his own words $T_{M2E,12}+T_{M2E,21}$ after he uttered them. This echo is caused by a reflection close to party 2. This echo is attenuated by $SLR+RLR+EL_2$ (expressed in dB) with respect to the original signal. Here, EL_2 is the echo loss close to party 2 (measured with respect to a certain reference point) [8] and the Send Loudness Rating *SLR* and Receive Loudness Rating *RLR* are defined as the attenuation of the signal from party 1 to the reference point and vice versa respectively. The sum $SLR+RLR$ is usually (tuned to) about 10 dB, a value that we assume in the remainder of this paper.

Second, listener echo also disturbs party 1, who hears the original signal from party 2 followed by an attenuated echo of this signal $T_{M2E,12}+T_{M2E,21}$ 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. Hence, the attenuation of the listener echo with respect to the original signal heard by party 1 is EL_1+EL_2 (expressed in dB).

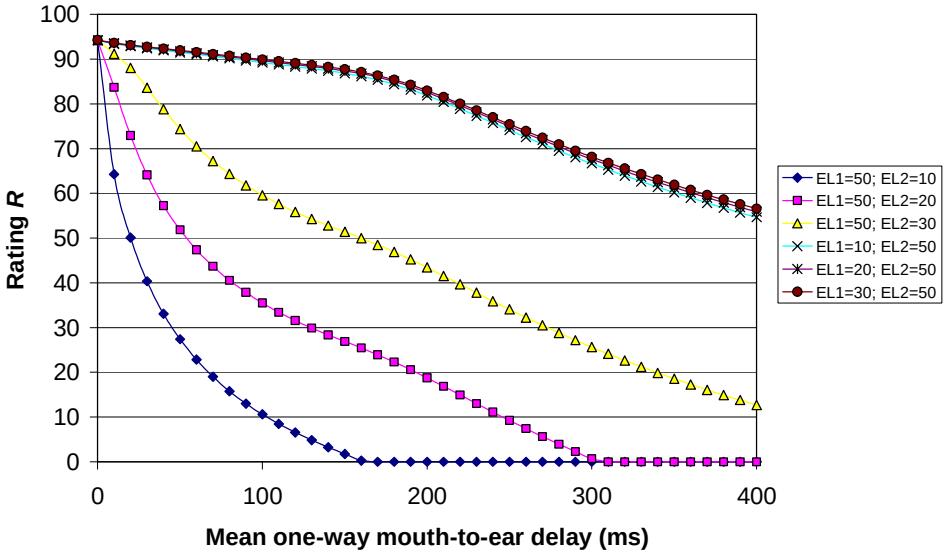
Echo may occur in the hybrid if the packetized phone call is terminated over a local PSTN or in the caller’s user terminal. For PSTN calls from traditional handsets, where echo is mainly caused by the 4-to-2-wire hybrids, a typical value for the echo loss is of the order of 20 dB [8]. The same value is valid for packetized phone calls where the call is terminated via a gateway over a local loop to a traditional handset. Acoustic echo is usually

small for traditional handsets. It is likely to be higher for other kinds of terminal, such as PCs and handsfree phones (resulting in an echo loss of e.g. 10 dB). An echo controller increases the echo losses EL_1 and EL_2. A standard-compliant echo controller [3] should increase the echo loss by 30 dB. Perfect echo control, which increases the echo losses EL_1 and EL_2 to infinity, can be achieved at moderate computational cost. Since 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, it is recommended that the echo controller in the gateway compensates for the echo generated in the hybrids of the PSTN over which the call is terminated and the echo controller in the terminal compensates for the acoustic echo this terminal generates itself.

The third delay-related factor that may 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 sum of both mouth-to-ear delays $T_{M2E,12}+T_{M2E,21}$.



(a)



(b)

Figure 3: The rating *R* as function of the mean one-way mouth-to-ear delay for undistorted voice (i.e. the G.711 codec without packet loss) and for various echo loss values (a) in case both echo loss values (expressed in dB) are the same and (b) in case both echo loss values (expressed in dB) are different.

Hence, under the above mentioned assumptions the impairment I_d is a function of $T_{M2E,m}$, EL_1 and EL_2, with

$$T_{M2E,m} = \frac{T_{M2E,12} + T_{M2E,21}}{2} \tag{9}$$

the mean one-way mouth-to-ear delay. Figure 3 illustrates the behavior of this function. Figure 3(a) shows how the rating R drops (due to an increase in I_d) as the mouth-to-ear delay increases for different values of the echo loss for the case when the echo losses at both end points are equal ($EL_1 = EL_2$) and when there is no distortion, i.e. $I_e = 0$. The impairment associated with delay is strongly influenced by this echo loss value. Notice that the rating R is a non-increasing function of the mouth-to-ear delay. The intrinsic quality of a phone call is defined as the rating R associated with a zero mouth-to-ear delay. The intrinsic quality of a packetized phone call transported without packet loss in the G.711 format and with all other parameters optimally tuned, corresponds to a rating R of about 94. This rating is referred to as $R_{\text{int},G.711}$. Figure 3(a) shows that if echo is perfectly controlled ($EL_1 = EL_2 = \infty$), the phone call retains its intrinsic quality up to a mean one-way mouth-to-ear delay of about 150 ms.

ITU-T Recommendations G.114 [1] and G.131 [2] specify the following tolerable mouth-to-ear delays for traditional PSTN calls:

- Under normal circumstances (i.e. if the echo loss is at least 20 dB), echo control is needed if the mouth-to-ear delay is larger than 25 ms;
- When the echo is adequately controlled:
  - a mouth-to-ear delay of up to 150 ms is acceptable for most user applications;
  - a mouth-to-ear delay between 150 ms and 400 ms is acceptable, provided that one is aware of the impact of delay on the quality of the user applications; and
  - a mouth-to-ear delay above 400 ms is unacceptable.

It can be seen from Figure 3(a) that for an echo loss of 20 dB, the rating R drops below 70 at a mouth-to-ear delay of 25 ms and for calls with perfect echo control, the rating R drops below 70 at a mouth-to-ear delay of 400 ms. Hence, ITU-T Recommendations G.114 and G.131 ensure that traditional PSTN calls have a rating R of at least 70. Also, the interactivity bound of 150 ms can be observed in Figure 3(a) for infinite echo loss.

Figure 3(b) shows how party 1 rates the call in case the echo losses at both end points are different. It can be seen that party 1 experiences a low quality if the echo loss EL_2 close to party 2 is not high enough, even if the echo controller close to party 1 (i.e., his "own" echo controller) is standard-compliant. Alternatively, if the echo controller close to part 2 is good enough, the echo controller close to party 1 does not impact the quality experienced by party 1 a great deal. Hence, the party with the best echo control will experience the worst quality (if all other factors are equal for both parties).

3.3 Influence of Distortion

If the voice signal party 1 hears is distorted, the rating 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. Remark that it is common practice, but not strictly mandatory, to transport the voice in the same format in both directions.

We first consider the influence of compressing the voice signal. As the G.711 codec just samples the (low-pass filtered) voice signal at 8 kHz and quantizes the samples with a non-uniform logarithm-like 8-bit quantizer, it introduces hardly any distortion. The packetization delay can be any multiple of 0.125 ms.

Predictive codecs (e.g. the G.726 codec) predict the sample to be encoded based on the previous ones (already encoded) and quantize the prediction error in 2, 3, 4 or 5 bits, resulting in a net codec bit rate R_{cod} of 16, 24, 32 and 40 kb/s respectively. 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 (referred to as 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 net codec bit rate R_{cod} is the code word size (in bits) divided by the frame length. Some of these codecs require a look-ahead in order to estimate the vocal track model parameters more accurately. Since the packetization delay is an integer multiple of the voice frame, and hence, is at least one voice frame, the larger the voice frame is, the larger the minimal delay the codec introduces, is. 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.

Recently a new codec the Adaptive MultiRate (AMR) codec [9] was developed in the framework of the third generation mobile network. It has a voice frame length of 20 ms (as all GSM codecs) and the particularity that the vocal track parameters can be quantized in a different number of bits, resulting in code words of variable size, from voice frame to voice frame, and hence, in a variable bit rate.

Figure 4 summarizes the distortion impairment associated with some standardized codecs. The points on this figure are rate-distortion pairs determined by experiments reported in [6]. Also three lines connecting similar pairs are drawn on this figure. This is a straight line when there are just two pairs or a quadratic best fitting curve in case there are more pairs. One line is associated with the G.726 codec and gives the rate-distortion trade-off for predictive codecs. It can be seen that at low bit rates predictive codecs introduce a lot of distortion. Another line is associated with the G.728 codec. This codec has a better rate-distortion trade-off than predictive codecs but does not reach the full potential of codecs of the vocoder type, as its voice frame size is too small. Also the older GSM-FR and GSM-HR codecs do not reach the full potential of vocoder codecs. A third line is drawn through the state-of-the-art codecs of the vocoder type (i.e. the G.729, G.723.1 and GSM-EFR codec) and as such gives the rate-distortion trade-off for vocoder codecs. It can be seen that vocoder codecs have the best rate-distortion trade-off. Although the AMR codec has not been characterized yet in terms of how much distortion it introduces at what bit rate, the latter curve on Figure 4 (labeled "AMR") forms a very good initial estimate.

A VAD scheme, which detects if the signal contains active speech or background noise, can be used to further reduce the overall bit rate to be sent. Good VAD schemes hardly introduce any additional distortion.

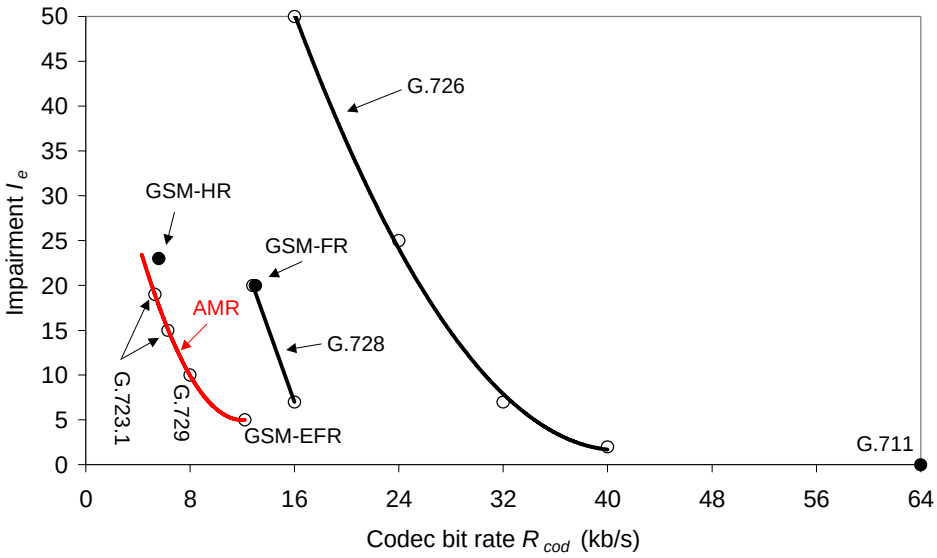


Figure 4: Impairment I_e for several standardized codecs. The points are experimental rate-distortion pairs. Based on this experimental input the rate-distortion (codec bit rate R_{cod} vs. Impairment I_e) curve is interpolated for the G.726 codec, the G.728 codec and the AMR codec.

The distortion impairment I_e associated with a codec increases as the packet loss ratio increases. Only a few results are known and are summarized in Figure 5. In that figure we draw the quadratic curves that best fit the experimental data (i.e. the points in that figure) reported in [6], which gives experimental data for four codecs under the assumption that voice packets are lost at random. Although other results are not yet known some trends can be observed.

The sensitivity to packet loss depends on the Packet Loss Concealment (PLC) technique used by the codec. In contrast to the G.711 codec, most state-of-the-art low-bit-rate codecs (e.g. G.729, G.723.1 and GSM-EFR) have a built-in PLC scheme. However, a (proprietary) PLC scheme can be implemented on top of the G.711 codec. From Figure 5 it can be seen that for the codecs that use PLC, the impairment increases by about 4 units on the *R*-scale per percent packet loss (for low loss values). If no PLC scheme is implemented on top of the G.711 codec, the distortion impairment increases by 25 units on the *R*-scale for each percent packet loss (for low loss values).

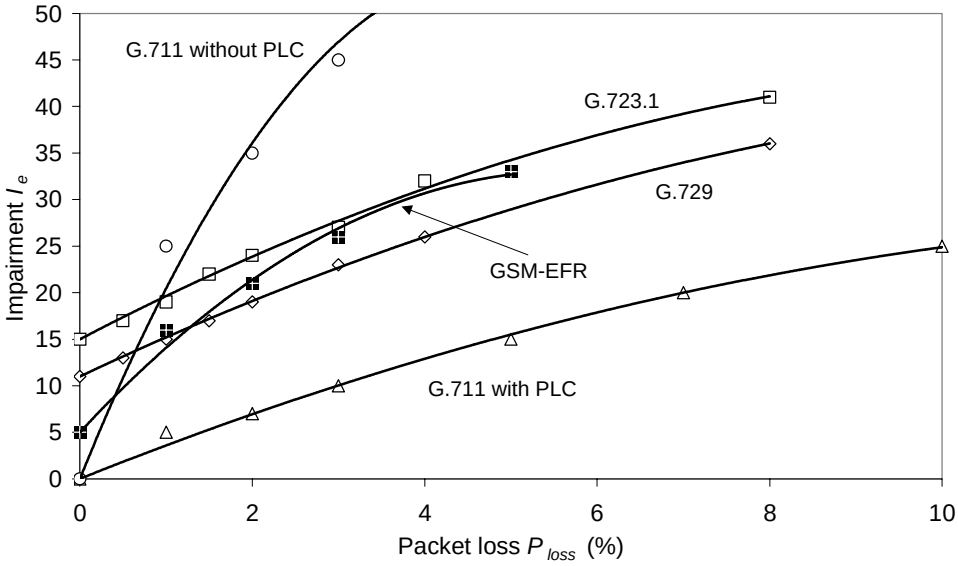


Figure 5: Distortion impairment I_e as a function of the packet loss P_{loss}.

Figure 5 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). The G.723.1 codec was only used at 6.3 kb/s. Comparing the slopes of the curves on Figure 5, we see that the G.711 codec with PLC, is slightly less sensitive to packet loss than the G.729 codec, which in turn is a bit less sensitive than the G.723.1 codec. From the results it cannot be concluded if this is due to the bit rate of the codecs (high-bit-rate codec formats contain more redundant information, and hence, are probably less sensitive to loss) or to a smaller packetization interval. Also from Figure 5, it can be seen that the PLC technique of the GSM-EFR codec does not perform so well as the PLC techniques of the other considered codecs. The conclusion from this paragraph is that in lossy environments a PLC is highly recommended.

CODEC	G.711 (64kb/s)	G.726 (40kb/s)	G.726 (32kb/s)	G.726 (24kb/s)	G.726 (16kb/s)	G.728 (16kb/s)	GSM-FR (13kb/s)	G.728 (12.8kb/s)	GSM-EFR (12.2kb/s)	G.729 (8kb/s)	G.723.1 (6.3kb/s)	GSM-HR (5.6kb/s)	G.723.1 (5.3kb/s)
G.711 (64kb/s)	94	92	87	69	44	87	74	74	89	84	79	71	75
G.726 (40kb/s)	92	90	85	67	42	85	72	72	87	82	77	69	73
G.726 (32kb/s)	87	85	80	62	37	80	67	67	82	77	72	64	68
G.726 (24kb/s)	69	67	62	44	19	62	49	49	64	59	54	46	50
G.726 (16kb/s)	44	42	37	19	0	37	24	24	39	34	29	21	25
G.728 (16kb/s)	87	85	80	62	37	80	67	67	82	77	72	64	68
GSM-FR (13kb/s)	74	72	67	49	24	67	54	54	69	64	59	51	55
G.728 (12.8kb/s)	74	72	67	49	24	67	54	54	69	64	59	51	55
GSM-EFR (12.2kb/s)	89	87	82	64	39	82	69	69	84	79	74	66	70
G.729 (8kb/s)	84	82	77	59	34	77	64	64	79	74	69	61	65
G.723.1 (6.3kb/s)	79	77	72	54	29	72	59	59	74	69	64	56	60
GSM-HR (5.6kb/s)	71	69	64	46	21	64	51	51	66	61	56	48	52
G.723.1 (5.3kb/s)	75	73	68	50	25	68	55	55	70	65	60	52	56

Table 2: Transcoding matrix.

The voice signal does not need to 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. Since all (considered) standard codecs need an 8 kHz stream of uniformly quantized voice samples at the input, the code words of the first codec need to 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. The intrinsic quality associated with all combinations of two codecs can be found in Table 2 (using the color code of Table 1). The diagonal entries in this table correspond to tandeming two codecs of the same type. Table 2 readily shows that transcoding can be very harmful to the quality of a call. In practice, the order in which the codecs are tandemed has a small influence, which cannot be seen in (the symmetric) Table 2 because, as impairments are considered to be

additive in the E-model, asymmetries cannot occur. The conclusion from Table 2 is that transcoding should be avoided.

4 Controlling voice quality

The conclusion from Section 3 is that for our purposes the rating R can be written as

$$R = R_{\text{int},G.711} - I_d\left(T_{M2E,m}, EL_1, EL_2\right) - I_e\left(\text{codec}, P_{loss}\right) \quad . \tag{10}$$

The combined effect of the first and second term is illustrated in Figure 3. The third term is displayed in Figure 5.

4.1 Quality bounds

Since the echo control bound of 25 ms is almost always exceeded when the phone calls are transported over a packet-based network, echo control is strongly recommended for packetized phone calls. A good echo controller, i.e., an echo canceller compliant with ITU-T recommendation G.168 [3], can increase the echo loss (from 20 dB usually occuring in the PSTN) to 50 dB. With an echo controller with a non-linear element perfect echo control, in which case the echo loss is increased to infinity, can be achieved.

From Figure 3 it is clear that in the case of perfect echo control at both sides, the intrinsic quality of the call is attained if the mean one-way mouth-to-ear delay is kept below 150 ms. From eq. (10) we notice that this intrinsic quality is solely determined by the distortion impairment I_e, which in turn is determined by the codec(s) used and the overall packet loss experienced. Since the intrinsic quality $R_{\text{int},G.711}$ of an undistorted call is about 94 and the bound for traditional quality is 70, there is an impairment budget of 24, part of which is consumed by the codec(s) (see Figure 4). Once the codec has been chosen, the remainder of the margin can be consumed either by allowing the mean one-way mouth-to-ear delay to exceed 150 ms or by tolerating some packet loss. The bound on the mean one-way mouth-to-ear delay for a certain codec is derived by subtracting the impairment associated with that codec (displayed in Figure 4) from the curves of Figure 3 and determining where the curve associated with perfect echo control drops below 70. The bound on packet loss for a certain codec is derived by determining in Figure 5 for which packet loss value the impairment budget of 24 is just not consumed. The bounds for the AMR codec are derived under the assumption that the interpolation (i.e. the curve in Figure 4 labeled "AMR") is valid and that per percent packet loss 4 units are added to the impairment I_e. The fourth column of Table 3 and Table 4 give the codec-dependent bounds on the mean one-way mouth-to-ear delay and packet loss, respectively, when the echo is perfectly controlled [10].

standard body	short name	codec bit rate (kb/s)	T_{M2E} (ms) EL = infinite	T_{M2E} (ms) EL = 50 dB
ETSI	GSM-HR	5.6	177	29
ITU-T	GSM-FR	13	210	106
ITU-T	G.711	64	400	291
ITU-T	G.728	12.8	210	106
		16	322	243
ITU-T	G.726 G.727	16	NA	NA
		24	NA	NA
		32	322	243
		40	375	276
ITU-T	G.723.1	5.3	219	131
		6.3	251	187
ITU-T	G.729	8	294	223
ETSI	GSM-EFR	12.2	342	256
3GPP	AMR	4.75	197	72
		5.15	214	117
		5.9	239	172
		6.7	262	198
		7.4	280	213
		7.95	293	223
		10.2	332	250
		12.2	342	256

Table 3: Tolerable mean one-way mouth-to-ear delay T_{M2E} bounds when there is no packet loss in the case of perfect echo control (EL=inf) and an echo loss EL = 50 dB.
(NA = Traditional PSTN quality (R = 70) is Not Attainable.)

The fifth column of the same tables show how these bounds reduce when the echo control is not perfect, but still very good EL = 50 dB (i.e., compliant with ITU-T Recommendation G.168 [3]) at both sides. For the packet loss

bounds it was assumed that the mean one-way mouth-to-ear delay was exactly 150 ms. The bounds for the AMR codec are derived under the same assumption specified above. It can be seen that if the performance of the echo controller drops from perfect to slightly less than perfect, this can have a drastic effect, especially on the mean one-way mouth-to-ear delay bound.

standard body	short name	codec bit rate (kb/s)	P_{loss} EL = infinite T_{M2E} <150ms	P_{loss} EL = 50 dB T_{M2E} =150ms
ITU-T	G.711no PLC	64	1.2	0.9
ITU-T	G.711 PLC	64	9.6	6.1
ITU-T	G.723.1	6.3	2.1	0.6
ITU-T	G.729	8	3.5	1.7
ETSI	GSM-EFR	12.2	2.5	1.5
3GPP	AMR	4.75	0.7	NA
		5.15	1.2	NA
		5.9	1.9	0.3
		6.7	2.6	1.1
		7.4	3.2	1.6
		7.95	3.5	2.0
		10.2	4.6	3.0
		12.2	4.8	3.2

Table 4: Tolerable packet loss P_{loss} bounds for a mean one-way mouth-to-ear delay below 150 ms in the case of perfect echo control and for a mean one-way mouth-to-ear delay of 150 ms for an echo loss EL = 50 dB. (NA = Traditional PSTN quality (R = 70) is Not Attainable.)

4.2 Controlling the delay and distortion in a gateway-to-gateway scenario

In this section we consider a gateway-to-gateway scenario illustrated in Figure 6. Phone calls originate from and terminate at traditional phone sets and are switched over a local PSTN to gateways, between which the voice signals are transported over a QoS-enabled IP backbone administered by one network manager [12]. Between each pair of gateways a traffic pipe is defined. We assume symmetric pipes. The transport of the voice packets over this pipe is governed by a Service Level Specification (SLS) [13]. The SLS is completely defined by specifying values for $P_{loss,net}$, $T_{net,\min}$ and enough information to describe the function $F(.)$ of eq. (2) as accurately as possible. As described above the latter requirement boils down to specifying as much quantiles of the total queuing delay as necessary.

The gateway parameters that can be tuned are the packetization delay T_{pack} and the dejittering delay T_{jit} (or equivalently the dejittering loss $P_{loss,jit}$ (see eq. (5)). We assume that adaptive dejittering is used and is converged to its optimal value, and hence, eq. (6) determines the one-way mouth-to-ear delay. We furthermore assume that there is no packet loss in the backbone $P_{loss,net} = 0$ and that the minimum network delay $T_{net,\min}$ is primarily determined by propagation (of 5 µs per km). Hence, this delay $T_{net,\min}$ is determined once the physical distance between the gateways is known.

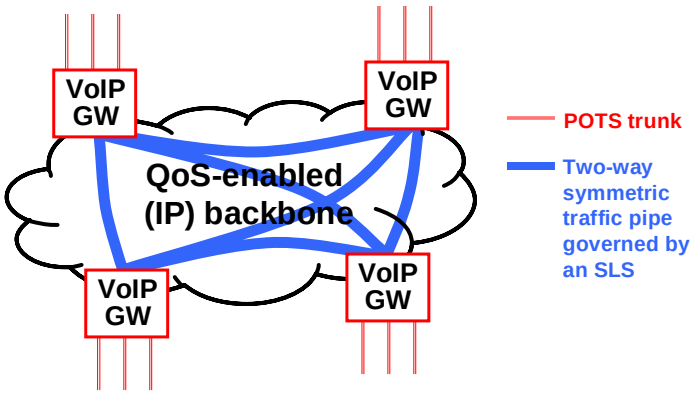


Figure 6: A gateway-to-gateway scenario to transport phone calls.

The choice in packetization delay T_{pack} is a trade-off between efficiency (see eq. (1)) and delay (see eq. (6)). We know from the previous section that under perfect echo control at both sides a total budget of 150 ms can be consumed without hampering the quality. However, from Table 3 it can be seen that if the performance of the echo control reduces to “nearly perfect”, but still is standard-compliant, the bound on the mean one-way mouth-

to-ear delay can be below 150 ms in some cases. The packetization delay is typically chosen between 10 and 80 ms. From eq. (1) it follows that since the overhead S_{OH} is 320 bits (consisting of 20 IP, 8 UDP and 12 RTP bytes) for Voice over IP (VoIP), an overhead bit rate between 32 kb/s and 4 kb/s, respectively, is introduced.

The flexibility in the choice in dejittering loss $P_{loss,jit}$ (or equivalently the dejittering delay T_{jit}) is governed by the number of quantiles that are specified in the SLS, i.e. how many points of the function $F(.)$ are given. If only the maximum total queuing delay (i.e. the (1-P)-quantile with $P = 0$) is given, only this total queuing delay can be used as dejittering delay. The more quantiles are specified, the more flexible the choice can be.

To conclude this section we give an example. Consider a phone call from Europe to the US. We assume $T_{net,\min}$ = 50 ms, $P_{loss,net} = 0$ and that the SLS in both directions is as described in Table 5(a). Furthermore, we assume a DSP delay T_{DSP} = 15 ms, an echo loss $EL_1 = EL_2$ = 50 dB, and that the G.729 codec (at 8 kb/s) is used. Table 5(b) gives the effective bit rate R_{eff} calculated with eq. (1) and Table 5(c) (using the color code of Table 1) gives the rating R calculated with eq. (10) for various values of the packetization delay and dejittering loss. From these tables it can be concluded that a packetization delay of 30 ms and a dejittering loss of 10^{-3} leads to a good compromise between effective bit rate (R_{eff} = 18.7 kb/s) and quality (R = 79).

SLS specification	
P	(1-P)-quantile (ms)
1.E-01	1
1.E-02	3
1.E-03	10
1.E-04	30
1.E-05	130

(a)

T_{pack} (ms)	10	20	30	40
R_{eff} (kb/s)	40.0	24.0	18.7	16.0

(b)

$P_{loss,jit}$ \ T_{pack} (ms)	10	20	30	40
1.E-01	52	52	51	51
1.E-02	76	75	75	75
1.E-03	79	79	79	78
1.E-04	79	79	78	78
1.E-05	72	70	69	67

(c)

Table 5 : (a) SLS specification, (b) effective codec rate R_{eff} (kb/s) and (c) rating R for various values of the packetization delay and dejittering loss.

The question how to provision the SLSs, i.e., how to configure the routers in the network such that the quantiles specified in the SLS are attained is beyond the scope of this paper. We refer the interested reader to [11] and [17].

5 Conclusions

In this paper the quality issues associated with the packetized transport of phone calls were considered. Since for packetized phone calls more delay and distortion is introduced than for traditional PSTN calls, the impact of delay and distortion on the quality of the phone call was studied quantitatively with the E-model. The trade-offs involved in the choice of the packetization delay and dejittering loss were discussed. From this quality study the following conclusions were drawn.

For packetized phone calls echo control is highly recommended, if not required, since otherwise the tolerable mouth-to-ear delay budget risks to be too small. If the echo is perfectly controlled, the quality remains equal to the intrinsic quality up to a mouth-to-ear delay of about 150 ms. The intrinsic quality depends on the amount of distortion that is introduced. If the echo control is slightly less than perfect, but still standard-compliant, the quality decreases even for delays smaller than 150 ms.

The intrinsic quality associated with predictive codecs at low bit rates is lower than the traditional PSTN quality. Therefore, these codecs should not be used at a bit rate below 32 kb/s. For the same reason, transcoding should be avoided.

Under perfect echo control the margin between the intrinsic quality of a codec and the bound for traditional quality can either be consumed by allowing a mouth-to-ear delay above 150 ms or by allowing some packet loss. The maximum tolerable bounds on the mean one-way mouth-to-ear delay and packet loss are reported in this paper for the most common codecs and even the recently developed Adaptive MultiRate (AMR) codec. It is also shown how these bounds decrease if the echo control is slightly less than perfect, but still standard-compliant.

These tolerable bounds should be respected by any packetized phone call (gateway-to-gateway, IP-phone-to-IP-phone, mobile-phone-to-mobile-phone, gateway-to-IP-phone, etc.) if traditional quality is to be maintained.

Finally, to illustrate how these bounds can be used this paper considered a gateway-to-gateway scenario where the transport of the voice packets is governed by a Service Level Specification (SLS). The trade-offs involved were shown on a numerical example.

Acknowledgement

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

References

[1] "One-Way Transmission Time", ITU-T Recommendation G.114, February 1996.

[2] "Control of Talker Echo", ITU-T Recommendation G.131, August 1996.

[3] "Digital Network Echo Cancellers", ITU-T Recommendation G.168, April 1997.

[4] "The E-model, a Computational Model for Use in Transmission Planning", ITU-T Recommendation G.107, December 1998.

[5] "Definition of Categories of Speech Transmission Quality", ITU-T Recommendation G.109, September 1999.

[6] "Provisional Planning Values for the Equipment Impairment Factor I_e", Appendix to ITU-T Recommendation G.113 (Draft), September 1999.

[7] "Digital Cellular Telecommunications System; Technical Performance Objectives", ETSI Technical Report ETR 315, November 1996.

[8] "Speech Processing, Transmission and Quality Aspects (STQ); Overall Transmission Plan Aspects for Telephony in a Private Network", ETSI Guide 201 050, November 1998.

[9] "Mandatory speech codec; AMR speech codec; Interface to Iu and Uu", ETSI 3G TS 26.102 v 3.1.0 (2000-03), Release 1999.

[10] D. De Vleeschauwer, J. Janssen, G. H. Petit, F. Poppe, "Quality Bounds for Packetized Voice Transport", Alcatel Telecom Review, First quarter 2000, pp. 19-23, January 2000.

[11] D. De Vleeschauwer, A. Van Moffaert, J. Janssen, M.J.C. Büchli, G.H. Petit, B. Steyaert, H. Bruneel, "Determining the number of packet-based phones that can be supported by one access node", Accepted for the 14th ITC Specialists Seminar on Access Networks and Systems, Barcelona (Spain), 25-27 April 2001.

[12] D. Goderis, "Functional Architecture Definition and Top Level Design", IST project Tequila deliverable D1.1, http://www.ist-tequila.org/deliverables/d11_final.pdf, 11 September 2000.

[13] D. Goderis, Y. T'joens, C. Zaccone, C. Jacquenet, G. Memenios, G. Pavlou, R. Egan, D. Griffin, P. Georgatsos, L. Georgiadis, "Service Level Specification Semantics, Parameters and Negotiation Requirements", IETF Internet Draft draft-tequila-diffserv-sls-00.txt, November 2000.

[14] F. Poppe, D. De Vleeschauwer and G.H. Petit, "Guaranteeing Quality of Service to Packetised Voice over the UMTS air Interface", Proceedings of the Eighth International Workshop on Quality-of-Service (IWQoS 2000), Pittsburgh (PA), 5-7 June 2000.

[15] F. Poppe, D. De Vleeschauwer, 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", Accepted for IEEE Infocom 2001, Anchorage (AL), 22-26 April 2001.

[16] N.O. Johannesson, "The ETSI Computation Model: A Tool for Transmission Planning of Telephone Networks", IEEE Communications Magazine, pp. 70-79, January 1997.

[17] G. Van Hoey, S. Van den Bosch, P. de La Vallée-Poussin, H. De Neve, and G.H. Petit, "Capacity planning strategies for voice-over-IP traffic in the core network", IEEE Workshop on High Performance Switching and Routing (HPSR2001), Dallas (Texas), 29-31 May 2001.

Biography

Danny De Vleeschauwer is a research engineer participating in the QoS, Traffic and Routing Technology Project within the Network Architecture team of the Alcatel Network Strategy Group in Antwerp, Belgium.

Annelies Van Moffaert is a research engineer participating in the QoS, Traffic and Routing Technology Project within the Network Architecture team of the Alcatel Network Strategy Group in Antwerp, Belgium.

Maarten J.C. Büchli is a research engineer participating in the QoS, Traffic and Routing Technology Project within the Network Architecture team of the Alcatel Network Strategy Group in Antwerp, Belgium.

Jan Janssen is a research engineer participating in the QoS, Traffic and Routing Technology Project within the Network Architecture team of the Alcatel Network Strategy Group in Antwerp, Belgium.

Guido H. Petit is Director of the Network Architecture Team of the Alcatel Network Strategy Group in Antwerp, Belgium.