D. De Vleeschauwer, J. Janssen, G. H. Petit, F. Poppe

Quality bounds for packetized voice transport

The transport of high quality packetized voice requires bounds on the mouth-to-ear delay and distortion to be respected.

Introduction

For traditional (wire-bound) Switched Telephone Network (STN) calls, which do not suffer from distortion, the key factor that determines the quality is the mouth-to-ear delay, defined as the delay incurred from the moment the talker utters the words until the instant the listener hears them. ITU-T Recommendations G.114 [1] and G.131 [2] report the mouth-to-ear delays that can be tolerated for undistorted voice. The bounds on these delays depend on the level of echo disturbing the voice call.

Voice calls can also tolerate some distortion, that is, the voice signal heard by the listener does not need to be an exact copy of the voice signal produced by the talker. In the case of packetized voice calls, distortion may be introduced by the codec that compresses the voice signal or by the loss of voice packets.

Controlling both the mouth-to-ear delay and distortion is the key to offering high quality packetized voice calls.

E-model

The E-model [3,4,5,6] predicts the subjective quality of a telephone call based on its characterizing transmission parameters. It combines the impairments caused by these transmission parameters into a rating *R*, which can be used to predict subjective user reactions, such as the Mean Opinion Score (MOS) or the percentage of users finding the quality Good or Better (GoB). The *R* scale was defined so that impairments are approximately additive in the *R* range of interest. The rating *R* is given by:

$$R = R_0 - I_s - I_d - I_e + A$$

The first term R_0 groups the effects of noise, such as background noise and circuit noise. The second term I_s includes impairments that occur simultaneously with the voice signal, such as those caused by quantization, by too loud a connection and by too loud a side tone. The third term I_d encompasses delayed impairments, including impairments caused by talker and listener echo or by a loss of interactivity. The fourth term I_e covers impairments caused by the use of special equipment; for example, each low bit rate codec has an associated impairment value. This impairment term can also be used to take into account the influence of packet loss. The fifth term *A* is the expectation factor, which expresses the decrease in the rating *R* that 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.

ITU-T draft Recommendation G.109 [7] states that 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 article, the classes are color coded according to *Table 1*.

As far as quality is concerned, a packetized voice call introduces more delay and distortion than a traditional STN call.

First, the delay for packetized voice calls, where the most important contributions are encoding, packetization, propagation, queuing, service, dejittering and decoding delay, is larger than for a traditional circuit-switched voice call, where the mouth-to-ear delay is mainly made up of the

R-value Range	100 - 90	90 - 80	80 - 70	70 - 60	60 - 0
Speech Transmission Quality Category	best	high	medium	low	(very) poor

PSTN Quality

Table 1 – Quality classes

Origin	Standard	Type	Codec bit rate (kbit/s)	Voice Frame(ms)	Look ahead(ms)	Algorithmic Delay(ms)	Ie	Instrinsic Quality
ITU-T	G.711	PCM	64	0.125	0	0.125	0	94.3
	G.726 G.727	ADPCM	16				50	44.3
			24				25	69.3
			32				7	87.3
			40	0.125	0	0.125	2	92.3
	G.728	LD-CELP	12.8	0.625	0	0.625	20	74.3
			16				7	87.3
	G.729(A)	CS-ACELP	8	10	5	15	10	84.3
	G.723.1	ACELP	5.3	30	7.5	37.5	19	75.3
		MP-MLQ	6.3				15	79.3
ETSI	GSM-FR	RPE-LTP	13	20	0	20	20	74.3
	GSM-HR	VSELP	5.6	20	0	20	23	71.3
	GSM-EFR	ACELP	12.2	20	0	20	5	89.3

Table 2 – Major parameters of standard codecs

propagation delay and switching delay. Most low bit rate codecs are frame-based, that is, they encode a voice interval of a certain duration, referred to as the voice frame, in a single encoding operation. Some codecs even need to collect the voice signal of an interval (referred to as the look-ahead) after the voice frame that is being encoded. The lengths of these intervals are given in *Table 2* for standard codecs. Since a packet must transport at least one voice frame, the lower bound on the packetization delay is set by the voice frame length. Similarly, as the encoder has to wait until the look-ahead has been collected, the lower bound on the encoding delay is determined by the look-ahead length. Hence, the mouth-to-ear delay is lower bounded by the algorithmic delay of a codec [8], which is the sum of the voice frame and the look-ahead length. The algorithmic delays of various standard codecs are given in *Table 2*.

Second, in contrast to circuit-switched voice calls, as a result of voice compression and packet loss during transport or in the dejittering buffer, the distortion of packetized voice calls is not negligible.

Alcatel has studied the impact of the one-way mouth-to-ear delay (via I_d) and the distortion (via I_e) on the quality of a packetized voice 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 voice call, but as these factors are not fundamentally different from a traditional STN call they were not considered. Furthermore, as the objective was to make a fair comparison between the quality of packetized voice calls and traditional wire-bound STN calls, the expectation factor *A* was set to zero.

From Equation 1 it follows that two calls with the same rating *R* can give a totally 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. $I_d = 0$). However, the E-model 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 voice call between two parties, referred to as party 1 and party 2 (see Figure 1). Using the E-model, Alcatel calculated how party 1 will judge the call, that is, what rating *R* will be assigned to it. The influence of delay was studied first, followed by the influence of distortion.

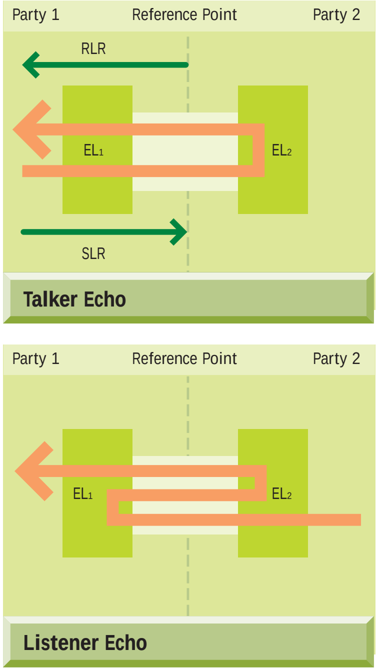
Influence of Mouth-to-Ear Delay

If the voice signal party 1 hears is delayed, the rating *R* decreases by an amount equal to the impairment I_d associated with the mouth-to-ear delay. This impairment is the sum of three contributing impairments: talker echo, listener echo and loss of interactivity.

First, talker echo disturbs party 1, who hears an attenuated and delayed echo of his or her own voice. This echo is caused by a reflection close to party 2. The level of this echo is strongly influenced by the echo loss EL2 close to party 2 (measured with respect to a certain reference point) [5].

Second, listener echo also disturbs party 1, who hears the original signal from party 2 followed by an attenuated echo of the signal. This echo is determined by a reflection close to party 1 with attenuation EL1, followed by a reflection close to party 2 with attenuation EL2.

Echo may occur in the hybrid if the packetized voice call is terminated over a local STN or in the callers' terminal equipment. For STN calls from traditional handsets, where echo is mainly caused by the



Talker Echo

Listener Echo

Figure 1 – Talker and listener echo

EL : Echo Loss
SLR : Send Loudness Rating
RLR : Receive Loudness Rating

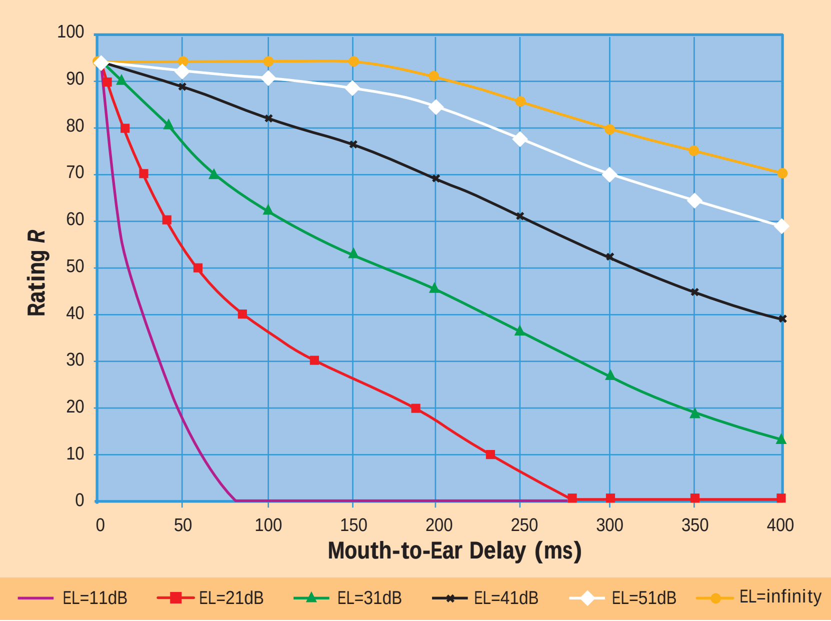


Figure 2 – The rating *R* as function of the mouth-to-ear delay for undistorted voice and for various echo loss values

4-to-2-wire hybrids, a typical value for the echo loss is 21 dB [5]. The same value is valid for packetized voice calls terminated over a local STN to traditional handsets. Echo loss is likely to be lower for other kinds of terminal, such as personal computers and handsfree phones. The echo losses EL1 and EL2 can be increased by using an echo controller, which should be deployed as close to the source of echo as possible, that is, in the gateways between the PSTN and the packet network, or in the terminals. A simple echo controller can increase the echo loss by 30 dB. Perfect echo control, in which the echo losses EL1 and EL2 increase to infinity, can be achieved at moderate computational cost.

The third delay-related factor that may disturb party 1 is the loss of interactivity. If the mouth-to-ear delay is too large, an interactive conversation becomes impossible.

Alcatel has used the E-model, which takes all these impairments into account, to calculate the rating *R* given by party 1 in the case of undistorted voice (see Figure 2). In the case of packetized voice calls, undistorted calls are calls transported without packet loss in the G.711 format. Figure 1 shows the influence of the mouth-to-ear delay on the rating *R* for different values of echo loss when the echo losses at both end points are equal (EL1 = EL2). The impairment associated with delay is strongly influenced by this echo loss value.

Observe that the rating *R* is a non-increasing function of the mouth-to-ear delay. The intrinsic quality of a voice call is defined as the rating *R* associated with a zero mouth-to-ear delay. The intrinsic quality of apacketized voice call transported without packet loss in the G.711 format corresponds to R = 94.3. Figure 2 shows that if echo is perfectly controlled (EL1 = EL2 = ∞), this voice call retains its intrinsic quality up to a mouth-to-ear delay of 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 21 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 2 that for an echo loss of 21 dB, the rating R drops below 70 at a mouth-to-ear delay of 25 ms. 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 2 for infinite echo loss.

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 has two sources: encoding of the voice signal from party 2 and packet loss during the transport of voice packets from party 2 to party 1.

Table 2 summarizes the distortion impairment and intrinsic quality (using the color code of *Table 1*) associated with each standard codec [9]. The distortion impairment I_e associated with a codec increases as the packet loss ratio increases. Figure 3, based on [9], shows this effect for four codecs, assuming that voice packets are lost at random. This figure 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). Results are not yet known for other packetization intervals.

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 low bit rate codecs (i.e. G.729, G.723.1 and GSM-EFR) have a built-in PLC scheme. However, a PLC scheme can be implemented on top of the G.711 codec. For the codecs that use PLC, the impairment increases by about four 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).

The voice signal does not need to be transported in the same format end-to-end. Somewhere along the route, voice 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 additive on the R scale. The intrinsic quality associated with all combinations of two codecs can be found in *Table 3* (again using the color code of *Table 1*). The diagonal entries in this table correspond to tandeming two codecs of the same type. Hence, 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 3* because, as impairments are considered to be additive in the E-model, asymmetries cannot occur.

Quality Bounds

If the mouth-to-ear delay, echo loss and distortion impairment are known, the quality of a packetized voice call (i.e. its rating *R*) can be derived from Figure 2, as follows. First, identify the curve on Figure 2 that corresponds to the given echo loss. Then, using this curve, read the rating *R* corresponding to the given mouth-to-ear delay. Finally, subtract the distortion impairment I_e from this rating *R*.

As stated before, if there is no echo control, the echo loss is likely to be (smaller than) 21 dB for packetized voice transport. For this value of the echo loss, the rating *R* drops rapidly as the mouth-to-ear delay increases. Hence, if there is no echo control, there is only a very small delay budget for which traditional PSTN quality ($R \geq 70$) can be guaranteed. As mentioned previously, the lower bound for the mouth-to-ear delay for packetized voice is the algorithmic delay. Since, in the case of low bit rate codecs this algorithmic delay is larger than the delay budget corresponding to 21 dB, calls transported using this codec format require echo control [8].

It is assumed here that perfect echo control is achieved, in which case the intrinsic quality of the call is attained if the mouth-to-ear delay is kept below 150 ms. 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 of an undistorted call is 94.3 and the bound for traditional quality is 70, there is an impairment budget of 24.3, part of which is consumed by the codec (see *Table 2*). Once the codec has been chosen, the remainder of the margin can be consumed either by allowing the mouth-to-ear delay to exceed 150 ms or by allowing some packet loss. *Tables 4* and *5* give the codec-dependent bounds on the packet loss and mouth-to-ear delay, respectively, assuming only one of these phenomena is allowed to occur. Note that packet loss could be traded off against mouth-to-ear delay (e.g. by varying the dejittering delay), as long as the impairment budget is not exceeded.

Conclusions

The E-model has been used to study the quality of packetized voice calls. With regard to quality, more delay and distortion are introduced for packetized voice calls than for traditional STN calls.

Since the tolerable mouth-to-ear delay budget is smaller than the minimal packetization delay if voice

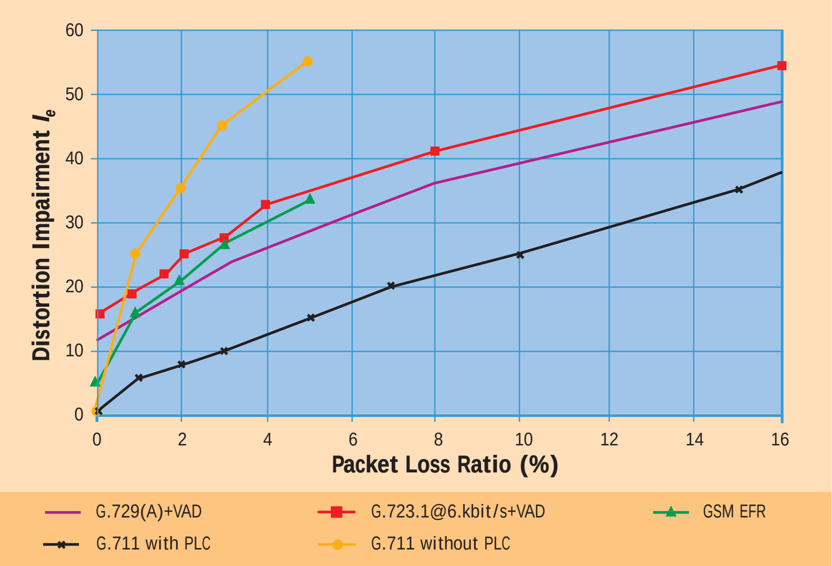


Figure 3 – Distortion impairment as a function of the packet loss

VAD : Voice Activity Detection
PLC : Packet Loss Concealment
EFR : Enhanced Full Rate

CODEC	G.711 (64kbit/s)	G.726 (40kbit/s)	G.726 (32kbit/s)	G.726 (24kbit/s)	G.726 (16kbit/s)	G.728 (16kbit/s)	GSM-FR (13kbit/s)	G.728 (12.8kbit/s)	GSM-EFR (12.2kbit/s)	G.729 (8kbit/s)	G.723.1 (6.3kbit/s)	GSM-HR (5.6kbit/s)	G.723.1 (5.3kbit/s)
G.711 (64kbit/s)	94.3	92.3	87.3	69.3	44.3	87.3	74.3	74.3	89.3	84.3	79.3	71.3	75.3
G.726 (40kbit/s)	92.3	90.3	85.3	67.3	42.3	85.3	72.3	72.3	87.3	82.3	75.3	67.3	71.3
G.726 (32kbit/s)	87.3	85.3	80.3	62.3	37.3	80.3	67.3	67.3	82.3	77.3	72.3	64.3	68.3
G.726 (24kbit/s)	69.3	67.3	62.3	44.3	19.3	62.3	49.3	49.3	64.3	59.3	54.3	46.3	50.3
G.726 (16kbit/s)	44.3	42.3	37.3	19.3	0	37.3	24.3	24.3	39.3	34.3	29.3	21.3	25.3
G.728 (16kbit/s)	87.3	85.3	80.3	62.3	37.3	80.3	67.3	67.3	82.3	77.3	72.3	64.3	68.3
GSM-FR (13kbit/s)	74.3	72.3	67.3	49.3	24.3	67.3	54.3	54.3	69.3	69.3	59.3	51.3	55.3
G.728 (12.8kbit/s)	74.3	72.3	67.3	49.3	24.3	67.3	54.3	54.3	69.3	64.3	59.3	51.3	55.3
GSM-EFR (12.2kbit/s)	89.3	87.3	82.3	64.3	39.3	82.3	69.3	69.3	84.3	79.3	74.3	66.3	70.3
G.729 (8kbit/s)	84.3	82.3	77.3	59.3	34.3	77.3	69.3	64.3	79.3	74.3	69.3	61.3	65.3
G.723.1 (6.3kbit/s)	79.3	75.3	72.3	54.3	29.3	72.3	59.3	59.3	74.3	69.3	64.3	56.3	60.3
GSM-HR (5.6kbit/s)	71.3	67.3	64.3	46.3	21.3	64.3	51.3	51.3	66.3	61.3	56.3	48.3	52.3
G.723.1 (5.3kbit/s)	75.3	71.3	68.3	50.3	25.3	68.3	55.3	55.3	70.3	65.3	60.3	52.3	56.3

Table 3 – Matrice de transcodage

is transported in a low bit rate codec format, calls transported in this format need to be echo controlled. If the echo is perfectly controlled, the quality remains equal to the intrinsic quality up to a mouth-to-ear delay of 150 ms. The intrinsic quality depends on the amount of distortion that is introduced.

The intrinsic quality associated with some low bit rate codecs is lower than the traditional STN quality. Therefore these codecs should be avoided. For the same reason, transcoding should be avoided at all cost. 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 mouth-to-ear delay and packet loss bounds are reported here for the most common codecs. These bounds should be respected by any packetized voice call (phone-to-phone, PC-to-PC, mobile-phone-to-mobile-phone, phone-to-PC, etc) if traditional quality is to be maintained.

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 N.O. Johannesson: “The ETSI Computation Model: A Tool for Transmission Planning of Telephone Networks”, IEEE Communications Magazine, pp 70–79, January 1997.

4 P. Meschkat: “TPE: Transmission Planning (End-to-End) using the E-model (Supporting ETSI Guide 201 050)”, Windows Software Tool, Alcatel Telecom, December 1997.

5 “Speech Processing, Transmission and Quality Aspects (STQ); Overall Transmission Plan Aspects for Telephony in a Private Network”, ETSI Guide 201 050 (Draft), November 1998.

6 “The E-model, a Computational Model for Use in Transmission Planning”, ITU-T Recommendation G.107, December 1998.

7 “Definition of Categories of Speech Transmission Quality”, ITU-T Recommendation G.109, September 1998.

8 D. De Vleeschauwer, J. Janssen, G.H. Petit: “Delay Bounds for Low Bit Rate Voice Transport over IP Networks”, Proceedings of the SPIE Conference on Performance and Control of Network Systems III, volume 3841, pp 40–48, Boston (MA), 20-21 September 1999.

9 “Provisional Planning Values for the Equipment Impairment Factor I_e”, Appendix to ITU-T Recommendation G.113 (Draft), September 1999. ■

Origin	Standard	Codec Bit Rate (kbit/s)	PL Bound (%)
ITU-T	G.711 without PLC	64	1
	G.711 with PLC	64	10
	G.729(A)+VAD	8	3.4
	G.723.1@ 6.3 kbit/s+VAD	6.3	2.1
ETSI	GSM-EFR	12.2	2.7

Table 4 – Tolerable packet loss bounds for a mouth-to-ear delay below 150ms

PLC : Packet Loss Concealment
VAD : Voice Activity Detection

Origin	Standard	Codec Bit Rate (kbit/s)	M2E Delay Bound (ms)
ITU-T	G.711	64	400
	G.726 G.727	16	NA
		24	NA
		32	324
		40	379
	G.728	12,8	212
		16	324
	G.729(A)	8	296
	G.723.1	5,3	221
		6,3	253
ETSI	GSM-FR	13	212
	GSM-HR	5,6	180
	GSM-EFR	12,2	345

Table 5 – Tolerable mouth-to-ear (M2E) delay bounds when there is no packet loss

NA : Traditional PSTN quality is Not Attainable

Danny De Vleeschauwer is a research engineer participating in the Traffic and Routing Technology project within the Network Architecture department of the Alcatel Corporate Research Center in Antwerp, Belgium.

Jan Janssen is a research engineer participating in the Traffic and Routing Technology project within the Network Architecture department of the Alcatel Corporate Research Center in Antwerp, Belgium.

Guido H. Petit is Manager of the Traffic and Routing Technology project within the Network Architecture department of the Alcatel Corporate Research Center in Antwerp, Belgium.

Fabrice Poppe is a research engineer participating in the Traffic and Routing Technology project within the Network Architecture department of the Alcatel Corporate Research Center in Antwerp, Belgium.