Recent standardisation work on non-intrusive evaluation of voice quality in IP environments

Vincent Barriac[1]

[1] *France Télécom R&D, DIH/EQS, 22307 Lannion cedex, France, Email: vincent.barriac@francetelecom.com*

Introduction

There are plenty of ways to address the objective (i.e. by modelling or measuring) evaluation of voice quality of telecommunications services. Generally, these different methods are divided in two big families, depending on whether test signals are used or not: the intrusive ones (also known as end-to-end, e.g.: PESQ [1]), and the non-intrusive ones (also known as single-ended, e.g.: P.SEAM [2]).

But one can also make a difference between methods needing access to the voice signal itself (e.g.: PESQ, P.561 [3] or P.SEAM) and the ones based on modelling of quality from transmission parameters (e.g.: P.VTQ).

In the specific case of voice over the Internet Protocol (VoIP), as far as non-intrusive measurement is concerned, one is facing a specific difficulty: the signal is encapsulated into packets and must be reconstructed before being processed. This heavy and resource-consuming processing is most of the time incompatible with the capacities of network elements. This explains the work currently undertaken by several standardisation bodies to address modelling of voice quality in VoIP services based on protocol information.

Overview of standardisation bodies and activities

Before adopting standard measurement or modelling methods, one must define common metrics on which those methods will be based.

Several organisations are (or have been) working on that topic:

- IETF (IPPM): RFCs 3550 [4] (RTP, replaces 1889), 2678 (connectivity) [5], 2680 (packet loss) [6], etc.
- ITU-T
    - SG13 : performance of IP-based networks associated metrics and QoS classes, (Y.1540 [7] and Y.1541 [8])
    - SG 12 : subjective and objective metrics and thresholds for perceived QoS (G.1xx and P.8xx series),
- ETSI (TIPHON): TS 101-329-2 (definition of QoS classes, based on metrics and thresholds) [9].

Now, the definitions of metrics specific for VoIP are converging and are mostly based on IETF RFCs definitions. But for some degradations, like bursty packet loss or jitter, a common definition or modelling is still missing.

As far as measurement methods are concerned, some standards are emerging :

- P.SEAM: single-ended psycho-acoustical model,
- P.561 and P.562 : non-intrusive measurement and analysis of single parameters,
- P.VTQ, to be selected by the ITU,
- ETSI Guide EG 201.329-5, annex E, VQMon [10].

Focus on ITU-T Q.16/12

Question 16 of the Study Group 12 of the International Telecommunications Union Standardisation Sector (ITU-T) for the study period 2001-2004 , named "In-service non-intrusive assessment of voice transmission performance", is addressing both the update of existing ITU-T recommendations and the creation of new ones, in particular in the field of VoIP.

The measurement tools that can process basic non-intrusive measurements are known as INMDs (In service Non-intrusive Measurement Devices). The ITU-T Recommendations defining how an INMD should work and how to analyse the measurement results are respectively P.561 and P.562 [11].

P.561 has been revised in 2002, to take into account new type of connection for the measurement probes and new parameters related to transit on IP network (most of them being taken from IETF RFCs). Now, it defines 4 classes of INMDs (from A to D). Classes A to C concern INMDs for circuit switched (TDM) networks, whereas class D addresses the case of packet switched networks.

A class D INMD should be able to measure on both transmission directions, with a given accuracy and in a given value range the following parameters:

- active speech level (in dBm)
- psophometric noise level (in dBm0p)
- attenuation (in dB) and delay (in ms) on the echo path
- IP delay variation (in ms)
- IP packet loss ratio (in %)

As for psycho-acoustical models like P.SEAM, the four first parameters (similar for the ones defined for classes A to C) can only be measured if on can access the speech signal. In IP networks, it will be possible only if the content of RTP packets is not protected (e.g. no use of sRTP).

P.561 defines as well optional parameters, both for TDM and for VoIP (e.g.: connectivity, round trip delay). It refers as well to P.SEAM for the measurement of a global single-ended MOS score.

P.562 is currently under revision in order to be adapted to the new class D of measurement devices.

But the most innovating topic of ITU-T Q.16/12 is the selection of a new standard for the evaluation of voice quality in VoIP services based on IP protocol information, known as P.VTQ.

The P.VTQ selection process

P.VTQ is supposed to address a large variety of network and terminal conditions. The concept is very simple: based on the collection of information provided by the analysis of various IP protocols (signalling + headers of the RTP packets), assess the global quality of the transmission, as closely correlated to the users' perception as possible.

P.VTQ, as a global model, is supposed to issue a single quality score (either a mean opinion score or a transmission rating score). To help analysing the measurement results, it is also mandatory that intermediate metrics are issued as well.

Launched in 2002, the competition for the selection of the future standard is still going on. Two candidate models are competing :

- PsyVoIP, from Psytechnics
- VQMon, from Telchemy

The competition is mostly based on a comparison of global scores obtained with candidates models with unknown references:

- Subjective scores
- Objective LQ (P.862) MOS scores

The new databases recorded for the P.VTQ competition cover a wide scope in terms of types of IP terminals (gatewayss, IP phones, softphones) and of network conditions (delay distribution, packet loss, based on simulation or on replay of real network conditions). For half of the databases, the characteristics of the IP terminals will remain unknown. For the other half, the models will have the possibility to adjust their scores based on a calibration performed a priori. In fact, only PsyVoIP is using calibration.

The databases have been recorded and processed following the reference methods by several independent labs. The processing by the two candidates and the final comparison still needs to be done but will be finished by the end of 2004.

Advantages and limitations of P.VTQ

As mentioned in the introduction, P.VTQ, as a parametric model, requires much less processing power and memory than a psycho-acoustical modem like P.SEAM. It can thus be easily embedded in network elements (including end-points, like IP-phones), whereas P.SEAM is rather destined to external expert probes.

The expected counterpart is a reduced reliability compared to reference scores (it is expected to have correlations of about 80%, when P.SEAM is close to 90%). But this drawback is compensated by the availability of intermediate metrics, which are very useful to analyse and understand the causes of quality impairments, and also by the fact that P.VTQ is designed for a specific application (VoIP), what requires less model training to be rather accurate.

Another problem, common to all single-ended measurement methods, is the fact that P.VQT is obliged to assume a given behaviour of the end devices (extra delay or packet loss, bad loudness ratings of terminals, etc.). This is partially compensated by the process of calibration present in PsyVoIP.

Last, but not least : P.VTQ can be seen, because of its low complexity and its possibility to issue transmission rating values, as a simple adaptation of the E-model to the context of VoIP. Some can be then tempted by a proprietary implementation of the E-model, not compliant with P.VTQ (and free !). As long as the results of the competition are not known, and even after (if they show no big difference compared to an implementation of the E-model), this risk is present. This is why P.VTQ must have been successfully validated against the widest range of conditions before becoming a standard. The limited number of independent parties involved in the selection process is endangering this objective.

References

[1] ITU-T P.862 Recommendation, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, 2001

[2] ITU-T SG12, Contribution D 183, Draft new Recommendation P.SEAM (for consent), 2004

[3] ITU-T P.561 Recommendation, In-service non-intrusive measurementdevice - Voice service measurements, 2002

[4] IETF RFC 3550, RTP: A Transport Protocol for Real-Time Applications, 2003

[5] IETF RFC 2678, IPPM Metrics for Measuring Connectivity, 1999

[6] IETF RFC 2680, A One-Way Packet Loss Metric for IPPM, 1999

[7] ITU-T Y.1540 Recommendation, Internet protocol data communication service - IP packet transfer and availability performance parameters, 2002

[8] ITU-T Y.1541 Recommendation, Network performance objectives for IP-based services, 2002

[9] ETSI TS 101-329-2, End-to-end Quality of Service in TIPHON Systems;Part 2: Definition of Speech Quality of Service (QoS) Classes, 2001

[10] ETSI TS 101-329-5, End-to-end Quality of Service in TIPHON systems;Part 5: Quality of Service (QoS) measurement methodologies, 2002

[11] ITU-T P.562 Recommendation, Analysis and interpretation of INMD voice-services measurements, 2000