VoIP Speech Quality Simulation and Evaluation

L. F. Sun[1], G. Wade[1], B. M. Lines[1], D. Le Foll[2], E. C. Ifeachor[1]

[1]School of Electronic, Communication & Electrical Engineering,
University of Plymouth, United Kingdom
[2]Wavetek Wandel Goltermann, Plymouth, United Kingdom

Abstract

The paper first presents briefly the current ITU P.861 PSQM objective speech quality measurement algorithm. Then the influence of packet loss and packet size on objective speech quality are simulated and analysed on a VoIP simulation platform. The limitations and possible improvements of the PSQM algorithm for use in VoIP applications are also given, together with future work.

Keywords

Voice over IP, QoS, Objective Speech Quality, Speech Quality Assessment

1. Introduction

Common VoIP network connections normally include the connection from phone to phone, phone to PC (VoIP Terminal or H.323 Terminal) or PC to PC. The end-to-end speech transmission quality will depend on the quality of the gateway (G/W) or VoIP/H.323 terminal and IP network performance.

Current research, worldwide, is concentrating on how to guarantee IP Network performance in order to achieve the required Quality of Service (QoS). Also, the impact of network parameters such as packet loss and jitter on speech quality have been broadly analysed (ETSI TR, 1999) (Yamamoto and Beerends, 1997). On the other hand, research is underway to improve the speech quality for "best effort" IP networks, and different compensation strategies for packet loss (Rosenberg, 1997) and jitter (Rosenberg and Qiu et al, 2000) have been proposed to improve speech quality even under poor network conditions.

Regardless of the strategy that is used to improve IP network performance or gateway/terminal performance, the purpose is to achieve a satisfactory speech transmission quality. The final judgement of speech quality still depends on the end user's perception. Subjective speech quality MOS (Mean Opinion Score) scores are considered the most powerful and recognised measure of speech quality, although the exact MOS value depends upon the measurement conditions. Since subjective measurement is time-consuming and expensive, objective speech quality measurement has been proposed to estimate the subjective quality of a network. Typical objective measurement methods include PSQM (Perceptual Speech Quality Measurement) (ITU, 1998) and PAMS (Perceptual Analysis/Measurement System) (ETSI EG, 1999). PSQM has been chosen as the ITU standard (P.861, 2/98) for objective speech quality measurement. Since these objective measures were originally

developed for the assessment of speech quality for low bit rate codecs, the impact of packet loss or variable delay (two important impairments in VoIP) were not considered in their first versions. Current work in ITU Study Group 12 therefore focuses on new objective speech quality assessment methods for VoIP, GSM and other networks. Modified PSQM or PAMS (e.g. PSQM+ (ITU, 1997), PSQM99, PAMS release 2.0 and 3.0) and other new algorithms have been proposed for the competition of the new ITU standard, which is expected to be available at the end of this year (ITU, 2000).

In this paper, we will first introduce briefly the ITU P.861 objective speech quality measurement algorithm in section 2. Then the structure, basic function and main parameters of a VoIP simulation platform are presented in section 3, preliminary test results about the influence of packet loss and packet size on objective speech quality are also given. In section 4, the limitations and possible improvements for PSQM while used in VoIP applications are presented, together with future work.

2. Objective Speech Quality Measurement

Objective perceptual speech quality measurement systems normally use two input signals, namely a reference signal and the degraded signal measured at the output of the network or system under test. Due to non-linearity arising from the codec, the signals should be speech recordings or artificial speech-like test signals. Typical measurement methods are PSQM and PAMS. Signal processing normally includes pre-processing, psycho-acoustic modelling, and a speech quality estimation model. The differences between these algorithms lie in differences between models. For example, the ITU P.861 PSQM algorithm consists of a perceptual model and a cognitive model (Figure 1), whilst PAMS includes an auditory transform (psychoacoustic model) and perceptual layer processing.

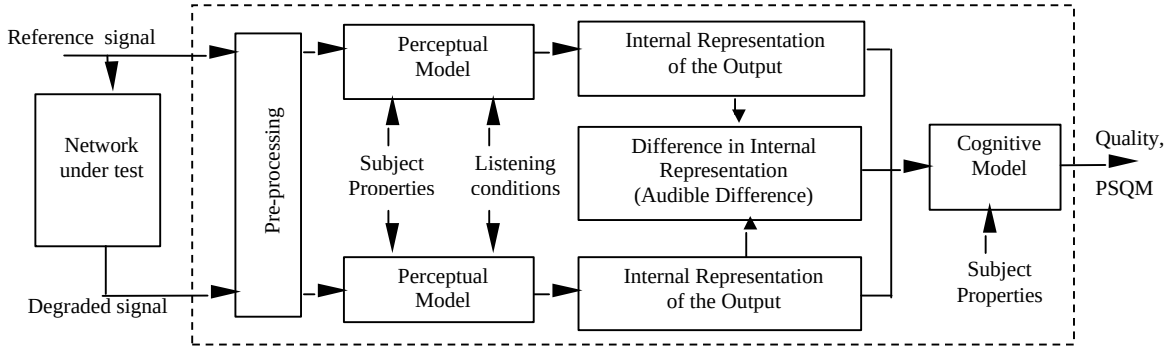

Figure 1. Structure of PSQM

As an example, we summarise the processing of PSQM as follows:

- Pre-processing

In order to compare the two signals, a pre-processing unit is used to perform delay adjustment (time alignment), loudness adjustment (equates loudness) and duration adjustment.

- Transformation

Each signal is passed through a “perceptual model”. This transforms the signal into a psychophysical representation that approximates human perception. These internal

representations make use of the psychophysical equivalents of frequency (critical band rates) and intensity (compressed sone).

- Calculation of perceptual difference distance, Noise Disturbance N or PSQM value

The perceptual difference distance is calculated between the two model output signals. This perceptual distance is expressed as a noise disturbance N_i for frame i (frame length 32ms), or N (PSQM value) by averaging for the whole speech segment. PSQM value indicated the degree of subjective quality degradation caused by the whole system under test. The PSQM value has a range from 0 to 6.5. 0 means no degradation (perfect quality), whereas 6.5 indicates the highest degradation.

- Mapping to objective MOS

The PSQM value is useful in itself for expressing speech quality degradation. In order to estimate subjective quality, mapping from the PSQM value to MOS score is necessary. The mapping part is not included in the ITU P.861 documents (ITU, 1998) and is also not taken into account in our current test.

3. VoIP Simulation Platform and Speech Quality Evaluation

3.1 VoIP Simulation Platform

An experimental VoIP speech quality evaluation system is set up as shown is Figure 2. Sender (A) and receiver (B) are two PC running a VoIP terminal simulation program under Linux. The third PC (C) works as a router running NIST NetDisturber (NIST, 2000), which can emulate various network problems by forwarding packets under specific parameters like packet loss, delay or jitter, between two network interface cards under a Linux system.

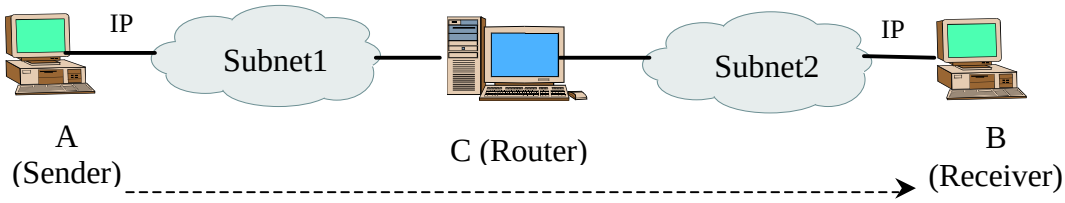

Figure 2. VoIP simulation platform

The sender process includes coder, packetizer and socket interface as shown in Figure 3. The receiver process covers socket interface, depacketizer, decoder, playout buffer, and sound driver interface as shown in Figure 4. It also includes an objective speech quality evaluation block, which completes ITU P.861 PSQM objective speech quality measurement algorithm.

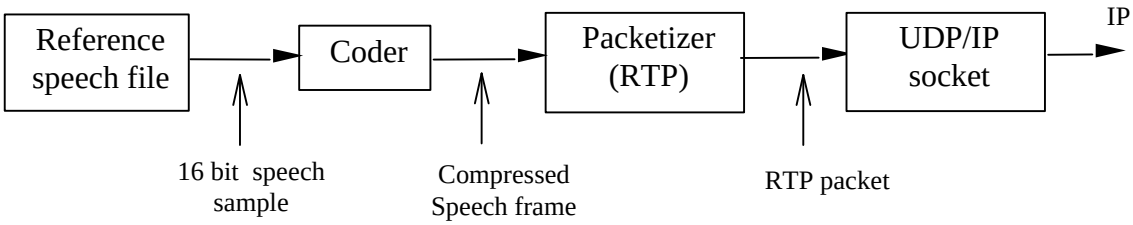

Figure 3. Sender process

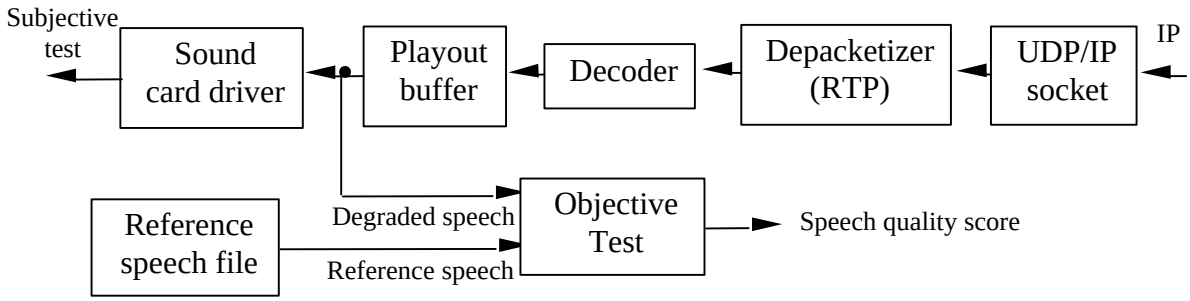

Figure 4. Receiver process

3.2 Test Conditions and Parameters

RTP/RTCP (Real-time Transport Protocol/ Real-time control Protocol) are chosen as the transport protocol for VoIP's real-time speech transmission. No signalling is considered in the experiment platform and only RTCP is used for setting up the connection between the sender and receiver.

ITU G.729A (8Kbps, low complexity version), G.723.1 Dual-rate (6.3/5.3Kbps) and ETSI GSM-FR (13Kbps) speech codecs are simulated in VoIP simulation platform. The type and frame size for each codec are shown in Table 1. VAD (Voice Activity Detection) is not included in the simulation. All the frames (for active or silent speech frame) are with the same length and have the same packet loss probability under simulation. For G.723.1, the high-pass filter and post-filter are enabled.

Codec	Type	Bit rate (Kb/s)	Frame Size (ms)	Frame length (bytes)	Lookahead (ms)	Encode Algorithmic Delay (ms)
GSM-FR	RPE-LTP	13	20	33	0	20
G.729	CS-ACELP	8	10	10	5	15
G.723.1	MP-MLQ	6.3	30	24	7.5	37.5
	ACELP	5.3	30	20	7.5	37.5

Table 1. Codec type and frame information

The RTP payload may include one to several speech frames according to the packet size. For example, if choosing three frames per packet, then the payload size is 24 bytes for G.729 as shown in Figure 5. The overhead of RTP/UDP/IP is 40bytes. Clearly, the more frames in one packet, the higher efficiency of transmission bandwidth and the longer delay for packetizing.

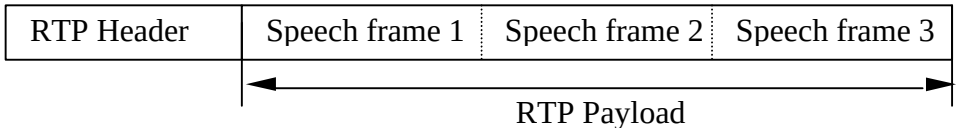

Figure 5. RTP packet structure (e.g. 3 speech frames per packet)

For P.861 PSQM, the frame length is 256 samples for 8000 Hz sampling rate (32ms) and adjacent frames overlap each other by 50%. The default global calibration factors (S_p and S_l) for long test sentence are used in the experiment, no other calibration work is done.

Two speech files in the ITU corpus are used in the experiment. Sentence 1 is female speaker and sentence 2 is male speaker. Both are about 5 to 6 seconds in duration and contain 16 bit signed linear PCM speech samples at 8kHz.

As PSQM is not suitable for variable delay happened during silence period or talkspurt, we do not consider the jitter adjustment in the experiment. Only a fixed size of jitter buffer (playout buffer) is considered in order to compensate some late packets at the cost of a buffer delay.

Time-alignment at the beginning of test sequence is considered. Except the codec's internal loss concealment, no other external concealment algorithms are taken into account in the simulation. The combination impact of packed loss, packet size and codec type on objective speech quality is the main purpose of the experiment in the paper.

3.3 Preliminary test results and analysis

We first test the PSQM values (both N_i and N) for G.723.1, G.729 and GSM-FR without any frame/packet loss. Then for G.729, we choose the 5% random frame loss for one frame per packet (independent or single frame loss) and 5% frame loss for 5 frames per packet (burst frame loss). The PSQM N_i vs frame *i* and corresponding speech waveform for the first 1.3 seconds of sentence 1 are shown in Figure 6 and 7. The PSQM value (N) for sentences 1 and 2 are shown in Table 1.

It is clear that all N_i values for codecs without packet loss are relatively stable within a limited range (no obvious peak). The reference and degraded speech waveform are similar (no gap) as shown in Figure 7 (A) and (B) for G.729. The PSQM values reflect well the subjective test results (MOS) for 3 codecs (The MOS scores from (Rudkin and Grace et al, 1997) are also listed in Table 2, MOS score for G.723.1 (5.3Kbps) is not available). However when a packet loss occurs, especially burst frame loss, there is an obvious peak in N_i curve for the lost period. It is also clear from comparing two waveforms of Figure 7 (A) for reference speech and (C) for degraded speech with burst loss that one frame is concealed by G.729's built-in one frame loss concealment algorithm. Four silent frames in the case of 5 consecutive frame loss follow this concealment frame. As the G.729 (G.723.1 is similar) decoder is highly dependent on the past state, the burst loss packets cause a divergence of encoder and decoder state. Even if subsequent packets after the burst loss sequence are received, they will not be decoded correctly. We can see this phenomena from the waveform after the burst frame loss in Figure 7 (C) and from the wider peak of N_i curve for G.729 (C) in Figure 6. If a random single frame is lost as shown in Figure 7 (D), the lost frame is concealed by the codec's concealment scheme. There is no gap in the waveform of the degraded signal while compared with the reference signal and also a narrower N_i peak for G.729 (D) in Figure 6.

From the test results and our analysis, we classify the influence of a packet loss on an active speech frame (the frame for a talkspurt, not for a silence period) into the following 3 categories.

- concealment frame (for a lost frame concealed by codec's built-in concealment algorithm or by external concealment algorithms)

- real-lost frame (for a lost frame filled by silence or comfort noise)
- inferred frame (for a normal received frame, which can not be decoded correctly due to lack of the parameters of the previous frames.)

Except the above 3 frames related with packet loss, the others are normal frames, which only suffer the normal codec impairment.

The PSQM N_i value and its variation (e.g. peak and its width) can also be used for analysis and estimate the end-to-end packet loss and the influence of packet loss on objective speech quality.

In addition, we deliberately erase 1/2/3/4/5/6 consecutive frames for every 100 frames and get the PSQM values as shown in Table 3. It is clear that the longer the burst frame loss size, the more serious the objective speech quality degradation. Burst frame loss (corresponding to the larger packet size) has much more influence on perceived speech quality.

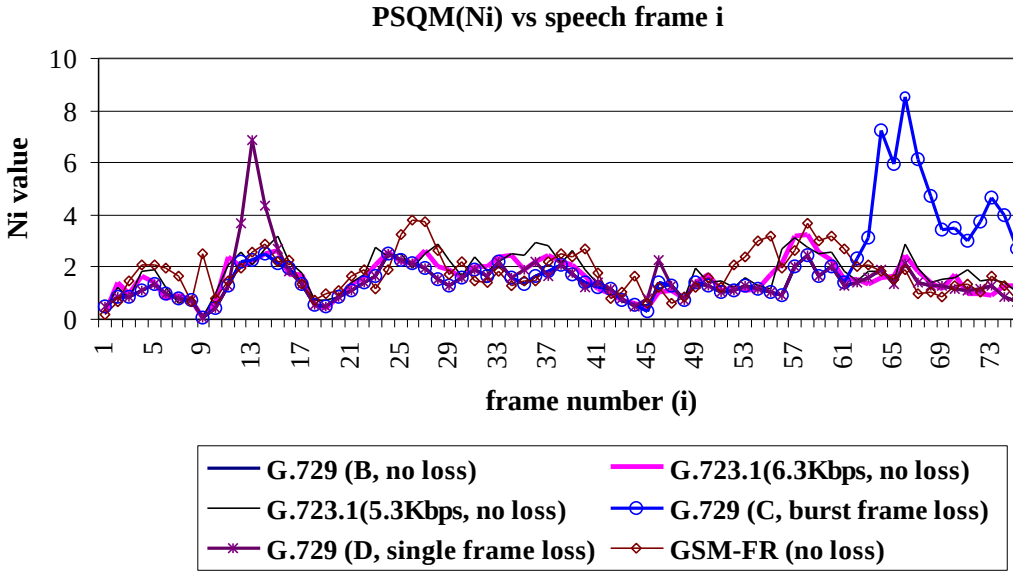

Figure 6. PSQM (N_i) value versus speech frame

Codec type packet loss	G.723.1 (5.3Kbps) no loss	G.723.1 (6.3Kbps) no loss	GSM-FR (13Kbps) no loss	G.729 (B) (8Kbps) no loss	G.729 (C) 5% random frame loss (for 1 frame per packet)	G.729 (D) 5% burst frame loss (for 5 frames per packet)
PSQM (N) Sentence 1	1.74	1.51	1.64	1.35	1.64	2.17
PSQM (N) Sentence 2	1.81	1.71	1.77	1.60	1.85	2.04
MOS	-	3.8	3.7	4.0	-	-

Table 2. PSQM value for different codec and packet loss

Burst size	1	2	3	4	5	6
PSQM	1.38	1.53	1.96	2.05	2.17	2.30

Table 3. PSQM value for different burst frame loss size

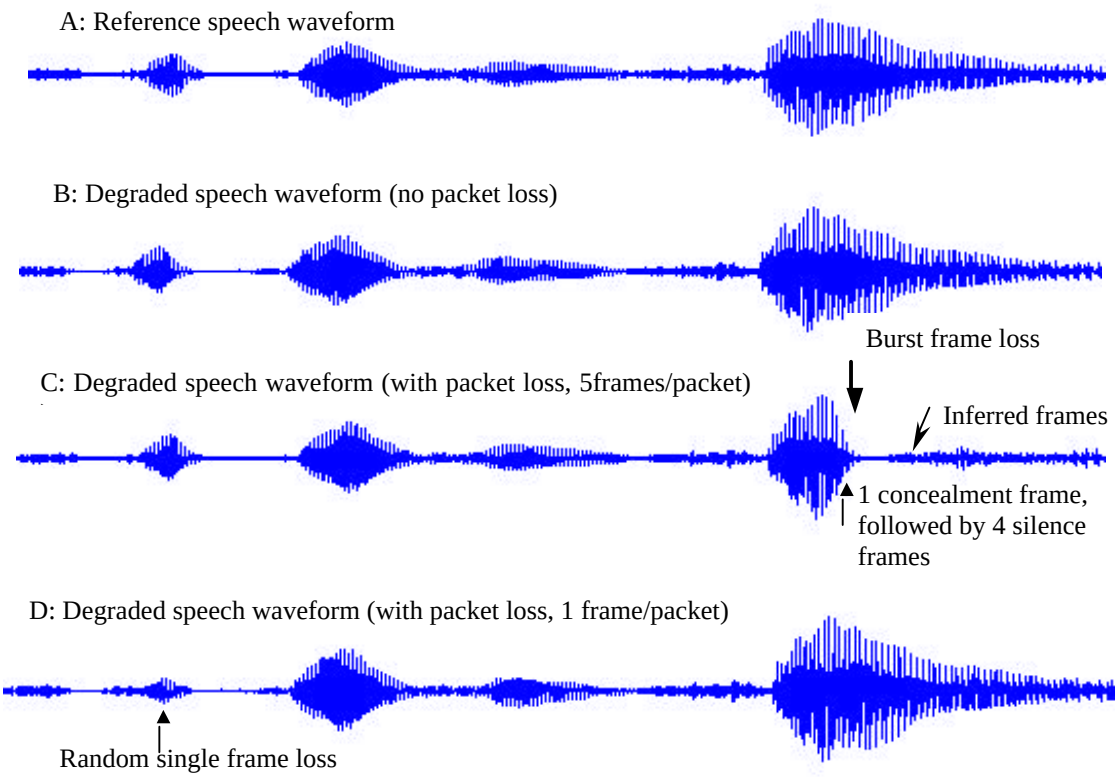

Figure 7. Reference and degraded speech waveform (G.729)

4. Future work

Simulation and analysis of VoIP speech quality under different codecs, packet size and network performances are ongoing.

As PSQM treats all frames (three types of loss influencing frames and normal frames) with the same processing method or the same weighting factors. PSQM+ was proposed by introducing an additional scaling factor especially for the lost (silence) frames, thereby compensating for the lost frame effect. While for the concealment and implicated frames, their impact on objective speech quality should also be taken into account with other additional scaling factors. The whole impact of packet loss on objective speech quality needs further research.

Another major problem for PSQM in VoIP applications, is the end-to-end jitter (not network jitter) (Sun and Wade et al, 2000), which is caused by the adjustment of jitter buffer. As we have mentioned in the paper, the PSQM algorithm only works well under strict time-alignment for two comparing signals and will give a very high PSQM value if two signals are not time-aligned (the PSQM value is almost meaningless if severe end-to-end jitter exists).

For an adjustment happening in a silent period, it is only necessary to perform realignment before the next talkspurt. For an adjustment happening in mid-talkspurt, time-alignment strategies e.g. cross-correlation could be used to find the matching signal frames for the calculation of the Noise Disturbance for each frame after the adjustment.

Obviously if buffer adjustment is very small, the effect could be imperceptible by the end user. In this situation, only delay jitter needs to be removed to keep the two signals aligned and the corresponding subjective impairment does not need to be considered. If the end-to-end delay jitter is greater than a subjective threshold, the playout impairment itself should be considered as one factor to weight the speech quality measurement algorithm.

5. References

ETSI, TR 101 329 V 2.2.2 (1999), "Telecommunication and Internet Protocol Harmonization Over Networks (TIPHON); General aspects of Quality of Service (QoS)", *European Telecommunications Standards Institute.*

ETSI, EG 201 377-1 V1.1.1 (1999), "Specification and measurement of speech transmission quality; Part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks", *European Telecommunications Standards Institute.*

ITU-T COM 12-20-E (1997), "Improvement of the P.861 Perceptual Speech Quality Measure", KPN Research, Netherlands, *International Telecommunication Union.*

ITU-T, P.861 (1998), "Objective quality measurement of telephone-band (300-3400 Hz) speech codecs", *International Telecommunication Union.*

ITU-T COM 12-117-E (2000), "Report of the question 13/12 rapporteur's meeting", Germany, *International Telecommunication Union.*

NIST Net home page (2000), *http://snad.ncsl.nist.gov/itg/nistnet/index.html.*

Rosenberg, J. (1997), "G.729 Error Recovery for Internet Telephony", Project Report, Columbia University, *http://www.cs.columbia.edu/~jdrosen/e6880/index.html.*

Rosenberg, J., Qiu, L. and Schulzrinne, H., (2000), "Integrating Packet FEC into Adaptive Voice Playout Buffer Algorithms on the Internet", *Proceedings of IEEE Infocom 2000,* Tel Aviv, Israel.

Rudkin, S., Grace, A. and Whybray, M.W., (1997), "Real-time applications on the Internet", *BT Technology J,* Vol. 15, No.2, pp. 209 – 224.

Sun, L., Wade, G., Lines, B., Ifeachor, E. and Foll, D.Le, (2000) "End-to-end speech quality analysis for VoIP", *IEE 16th UK Teletraffic Symposium,* Harlow, U.K.

Yamamoto, L.A.R. and Beerends, J.G. (1997), "Impact of network performance parameters on the end-to-end perceived speech quality", *Expert ATM Traffic Symposium,* Greece.