A Technical White Paper on Sage's PSQM Test

Renshou Dai

August 7, 2000

1 Introduction to PSQM

1.1 What is PSQM test?

PSQM stands for Perceptual Speech Quality Measure. It is an ITU-T P.861 [1] recommended objective method of estimating the subjective quality of voice-band speech codecs. The recommendation is based on the work of Beerends *et al* [2] [3]. The essence of the PSQM algorithm is to measure the distortion experienced by a speech signal when transmitting through various codecs and transmission media. It differs from the signal-to-noise type of measurement in that the distortion is not measured in the normal physical domain (time or frequency domain, for example). Instead, the distortion is measured in an 'internal psychoacoustic domain' to mimic the sound perception of human subjects (phone users) in real-life situations so that the measured distortion can be easily correlated with human perceptions. This is done by converting the physical-domain signals into the perceptually meaningful psychoacoustic domain through a series of nonlinear processings such as time-frequency mapping, frequency warping, intensity warping, loudness scaling, asymmetric masking and cognitive modeling etc.

1.2 What signal is used for PSQM test?

The testing source signal used along with PSQM is an ITU-T P.50 recommended artificial voice [4] with an active speech level of -20dBm. The artificial voice, which includes both genders (male and female), is aimed to reproduce the essential characteristics of human speech for the purposes of characterizing linear and nonlinear telecommunication systems and devices that are intended for the transduction or transmission of speech. The essential characteristics of a speech signal include the long-term average spectrum, short-term spectrum, instantaneous amplitude distribution, voiced and unvoiced structure of speech waveform and syllabic envelope.

2 PSQM Measurement results

2.1 PSQM value:

The output of the PSQM algorithm is called the PSQM value. It indicates the degree of subjective quality degradation caused by the whole communication system under test. The PSQM value ranges from 0 to 6.5. 0 means no degradation (perfect quality), whereas 6.5 indicates the highest degradation.

2.2 Mean-Opinion-Score (MOS):

As reported in Beerends work [2] [3], the PSQM value can be used to accurately predict an objective MOS score that has a high statistical correlation with the subjective MOS score obtained through human listening test. This objective MOS has a range from 1 to 5 and is inversely proportional to the PSQM value. MOS 5 means excellent speech quality, whereas MOS 1 indicates the worst. The exact conversion formula from PSQM to MOS has not been published. In Sage's implementation, the MOS value is obtained from the PSQM value through the following logistic function [5]:

$$MOS = \frac{4}{1 + e^{0.66PSQM - 2.2}} + 1$$

2.3 Gains:

Besides the PSQM and MOS values, the PSQM test also measures the overall system gains along each test direction. The gain value is not measured on any single frequency point. It is measured within the whole voice band (300Hz to 3400Hz) which practically indicates how much gain or loss a speech signal will experience when traveling from the transmitting end to the receiving end. The PSQM algorithm does not consider a flat gain change as speech quality degradation. But in a real life situation, the phone users may object to excessive gain or loss. Therefor, these gain values are important additional information to testers. The measurement precision is ±0.1dB.

2.4 Round-trip delay:

In real phone call, long delay certainly degrades the voice quality. But the delay effect is not reflected in PSQM test since it is a synchronized measurement and the synchronization process already factors out the delay effect. For this reason, Sage's implementation of PSQM also includes a round-trip delay measurement, which will provide the testers additional information that is not reflected in the PSQM test *per ce*. The algorithm is different from the absolute delay test specified in IEEE 743 (and patented by Sage) where the delay is estimated by measuring the phase delays of the AM modulated tones. For certain compressive codecs (CELP Vocoders, for example), the phase information of the tone signal may not be well preserved, therefor the measurement may not be accurate. The round trip delay measured during PSQM test is based on the cross-correlation of a frequency-hopped spread spectrum signal. It is more suitable for applications through compressive codecs and digital channels. The measurement precision is ±0.125ms.

3 Application Scope

3.1 Digital Distortions:

PSQM is perfectly applicable to the situation where digital distortions are the dominant cause of speech quality degradation. These digital distortions include voice compression, quantization (digitization) noise, codec transcoding errors (or noise), packet-cell-datagram loss, random or bursty bit errors etc. Table 1 lists the PSQM and MOS values for some codecs, in which the causes of degradation are mostly due to quantization and compression.

3.2 Analog type distortions:

If the main causes of quality degradation are additive noise and band-limiting attenuation and envelop-delay distortions which are typical of analog transmission media, then certain precaution

Coding Technique	Standard	Bit Rate (kbps)	PSQM	MOS
A-law, μ-law PCM	G.711	64	< 0.3	4.5
ADPCM	G.726	32	1.5	4.1
LD-CELP	G.728	16	1.6	4.0
MP-MLQ/ACELP	G.723.1	6.4(5.3)	1.8	3.8
CS-ACELP	G.729	8	1.6	4.0
LPC	USFS-1015	2.4	4.4	2.3
CELP	USFS-1016	4.8	3.0	3.0
VSELP	IS-54/IS-136	7.95	2.4	3.5
ACELP	IS-641	7.4	1.6	4.0
RPE-LTP	GSM Full-rate	13	2.4	3.5
VSELP	GSM Half-rate	5.6	2.4	3.5
ACELP	GSM Enhanced Full-rate	12.2	1.6	4.0

Table 1: Characteristics and PSQM/MOS numbers of standard speech coders. All the tabulated MOS and PSQM numbers are statistical mean values associated with each coder. During real test, the numbers may vary around, depending on the test signal. When using male voice, the measured MOS numbers wll be higher than those in the table. When using female voice, the measured MOS numbers will be lower. This agrees with the real performance of most speech coders.

should be taken when using PSQM test. PSQM test can certainly indicate the degradation caused by these distortions. But whether or not the degradation measured by PSQM algorithm in this case truly corresponds to human perception remains to be studied. Table 2 and Table 3 list the PSQM and MOS values when noise and band-limiting factors are applied to the speech signal. Although the PSQM value correctly indicates the increase of distortion, yet whether or not the predicted MOS score is the same as how human perceives the distortion is unknown.

Signal-to-Noise-Ratio	PSQM	MOS
40dB	0.07	4.9
30dB	0.5	4.7
20dB	1.93	3.8
10dB	4.5	2.2
5dB	5.9	1.4
0dB	6.5	1.0

Table 2: Simulated PSQM Results with Additive White Gaussian Noise using male voice

4 Limitations

The fact that we have to embed the telemetry and synchronization signals inside the test signal implies that this test will have a reliability (or robustness) limit.

Based on in-house tests, for analog-type of impairments, it's been found that the PSQM test can tolerate one-way attenuation of 30dB, 0dB signal-to-noise ratio (synchronization and telemetry signal to added white noise), -3dB echos, and other envelop and attenutation distortions etc.

Pass-Band Bandwidth	PSQM	MOS
200-3600Hz	0.03	4.9
300-3200Hz	0.35	4.8
400-3000Hz	1.22	4.3
500-2900Hz	1.69	4.0
600-2800Hz	2.7	3.4
700-2700Hz	4.0	2.6
800-2500Hz	6.5	1.0

Table 3: Simulated PSQM Results with Band-Limiting Distortions using male voice

When testing through a Wavetek (with VSELP Vocoder loopback and induced bit-errors) and a TDMA cellular phone, it's been found that the PSQM test can run through the testing path with as high as 7% bit-error-rate (induced in the compressed vocoder data-packets). Considering the fact that a typical Vocoder can only handle less than 3% bit-error-rate, this 7% bit-error-rate tolerance is quite 'bullet-proof'.

No matter how reliable the test is, there will always be the case when it fails. When this happens, the results will be displayed as 'FAIL', and the test will time-out after about 8 seconds.

5 Understand the test options

As currently implemented on 93x, the user-selectable options are:

Gender: One can choose between MALE and FEMALE. This controls whether to use male artificial voice or female artificial voice for testing. In almost all cases, the female test signal will indicate more quality degradation (higher PSQM value and lower MOS value), because female voice has higher pitch and has higher frequency components. This is also consistent with the results from subjective listening tests. Default is male.

Test duration: This dictates how long the test will be performed along each direction. One can choose between 1s and 64s. An ideal test duration is between 10 and 15s. Default is 10s.

Seed/Voice Pattern: The artificial voice is essentially a controlled random sequence. The initial seed determines the whole sequence. So when selecting different seed, a different pattern of artificial voice signal will be generated. Theoretically, different pattern of artificial voice signals should all give the same PSQM result, because they are all statistically identical (ergodicity). But in practice, when the test duration is short, the test results may vary a little when choosing different seed. This is not a bug. It is a normal statistical phenomenon. Right now, the customer can choose between 1 and 128, which means, the user can select one out of 128 different voice patterns for testing. This is an advantage over the use of real speech samples, where the number of choices is inevitably limited by storage space. But with artificial voice, one has almost an infinite number of choices. If not selected, the seed will be randomly chosen for you each time you start the test.

Speed: This controls how fast the artificial voice signal alternates between voiced segment (vowels) and unvoiced segment (fricatives). When choosing SPEED 'fast', the speech signal will sound like a fast talker. When choosing SPEED 'slow', the speech signal will sound like a slower talker. And SPEED 'medium' means a normally-paced artificial voice signal.

6 Compatibility Issues

6.1 Is Sage's PSQM Test Compatible with Others Implementation?

The answer is no. ITU-T P.861 only recommends the essential PSQM algorithm. It does not specify how synchronization and telemetry should be done. Different implementer will have different proprietary implementation schemes. Plus, the test source signal is different. Sage's implementation uses artificial voice that can be repeatedly generated in real time. Others may use stored real speech samples.

6.2 Is Sage's PSQM Result Comparable with Others PSQM Test?

The answer is yes. However, the PSQM result does depend on the test signal. If different implementation uses different test signal, the results will not be exactly same. But they should agree within certain range (±0.5, for example).

References

[1] "Objective quality measurement of telephone-band (300-3400Hz) speech codecs," *ITU-T Recommendation P.861*, Aug., 1996.

[2] J. G. Beerends and J. A. Stemerdink, "A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio. Eng. Soc.*, Vol. 42, pp. 115-123, March 1994.

[3] J. G. Beerends and J. A. Stemerdink, "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio. Eng. Soc.*, Vol. 40, pp. pp. 963-978, Dec. 1992.

[4] "Artificial Voices," *ITU-T Recommendation P.50*, March 1993.

[5] S. Voran, "Objective Estimation of Perceived Speech Quality, Part I: Development of the Measuring Normalizing Block Technique," *IEEE Transactions on Speech and Audio Processing*, July 1999.