

Principles of Speech Quality Assessment using Objective Methods

1. Scope

This application guide describes the background to speech quality assessment for telephony. It goes on to explain the general principles of objective testing with the Digital Speech Level Analyser (DSLAI).

2. Mean Opinion Score (MOS)

Throughout the history of telephony the process of transmitting a speech signal from one place to another has resulted in distortion. The Malden brochure “Measuring Speech Channel Performance in Digital Networks” mentions some of the sources of these distortions. The nature and variety of these distortions has changed very significantly with the advance of technology. For some time now, the dominant method of classifying the quality of the received speech signal has been through the use of the Mean Opinion Score (MOS).

MOS is a result obtained from a *Subjective Testing* process in which *subjects* listen to recorded samples of telephony speech and write down their *opinion* using a pre-agreed numerical scale. The *mean* of their opinions is calculated, hence Mean Opinion Score. There are various types of MOS, but the most commonly used is the Absolute Category Rating (ACR) where the subjects are required to state their *absolute* opinions without reference to a “good quality” recording. Within the ACR method, two different MOS measures can be assessed – Listening Quality (by far the most commonly used) and Listening Effort. Listening Quality (LQ) is a measure of the intrinsic quality of the speech signal. Listening Effort (LE) is a measure of the degree of effort required by the listener to discern the speech signal from the various distortions to which it may have been subjected. The Degradation Category Rating (DCR) has been included in the table below as it provides another interpretation of the MOS scale.

Mean Opinion Score scales and their interpretation are summarised in Table 3, below.

Subjective Testing must be rigorously controlled by experts if the results are to be meaningful. Consequently, whilst subjective testing is the definitive method for speech quality evaluation, it is time-consuming and tends to be expensive.

3. Objective Methods

Objective “measurement” of speech quality consists of the use of specialised algorithms which attempt to predict the results of subjective testing. Such algorithms include Perceptual Evaluation of Speech Quality (PESQ, ITU-T Recommendations P.862 & P.862.1), Perceptual Analysis/Masurement System (PAMS) and Perceptual Speech Quality Measure (PSQM, which was ITU-T Recommendation P.861). The advantages of an algorithm which is successful at predicting the results of subjective testing are clear:

- The expertise is contained in the algorithm, so that in general specialist expertise is not needed to make use of it (although the quality of the interface between algorithm and network is crucial)
- Results are available on demand and at relatively low cost;
- Results are highly repeatable.

The PESQ algorithm has been subjected to rigorous evaluation in many applications (for example, different codecs, transmission methods and types of signal impairment) and is probably the best general purpose tool for the objective prediction of MOS. PESQ has been adopted across the telecommunications industry as a convenient and accurate method of predicting MOS. In common with other algorithms it essentially compares a captured “degraded” signal with a corresponding “reference” signal which was injected into the network. The comparison is complex but essentially it attempts to weight signal degradations according to how noticeable they are likely to be to a human listener, hence the use of the word “perceptual” in the titles of each of the algorithms mentioned above.

It is important to understand what a tool such as PESQ does and does not provide. PESQ predicts the users’ perception of the quality of a received speech telephony signal. In doing so, it offers an end-to-end assessment of all the factors which may influence listening quality. On the other hand, it does not make any assessment of conversational quality and specifically it does not take into account delay or echo, although delay can be accurately measured using a DSLA with PESQ, and talker echo can be measured using the DSLA. Finally, it is worth noting that listening quality is not a measure of intelligibility.

4. Test Signal Requirements

A test signal used with PESQ, for example, must be speech-like in order to make a meaningful measurement of the performance of a speech channel which includes digital coding. "Speech-like" means that the signal has adequately representative samples of real sequences of sounds. In addition, the spectral content must be appropriate. That is, the test signal must be filtered to the Intermediate Reference Send (or modified Intermediate Reference Send) characteristic (IRS, mIRS) as described in ITU-T Rec. P.830. This characteristic approximates the filtering which occurs between the microphone and junction with the network. Note that where there is acoustic coupling to the network under test (for example, when using a Head and Torso Simulator – HATS) this filtering occurs naturally and so the test signal does not require IRS or mIRS conditioning. Refer to Application Notes DSLA-301 and DSLA-302 for more information on testing with the DSLA through acoustic interfaces.

A test signal must also have known properties in terms of sounds, noise and levels. High noise levels will mask impairments or cause incorrect codec operation. In some specialised applications such as codec testing it may be desirable to identify and isolate particular sounds in order to study the effects of those sounds on the system under test. The ASTS consists of a number of independent files which can be used alone or in combination to achieve a test signal having the required properties in terms of speaker gender, duration and sound (phoneme) content.

The Artificial Speech Test Stimulus (ASTS) supplied by Malden Electronics satisfies all of the above requirements and is very suitable for use with PESQ and the other algorithms supplied for use with the DSLA. The ASTS was compiled from an analysis of conversational speech, with greatly reduced redundancy of sounds. This means that tests can be made much shorter than is possible when using real speech.

The ASTS is supplied in both 8k and 16k sample rate versions, suitable for, respectively, standard narrowband telephony channels and narrowband/wideband (7.5kHz) channels.

Natural speech can also be used as a test stimulus, subject to the same conditions mentioned above. The recording must be free as far as possible of background noise and the speech quality must be good. This implies that a recording made, for example, using a standard-quality microphone in an office and a PC sound card is unlikely to be suitable. Recordings should be in 16 bit, 8k or 16k sample rate. The requirements for appropriate filtering of the recorded source material are identified above.

5. Suggested measures and influence on MOS

It is difficult to generalise about what constitutes a "standard" set of measures for speech quality. Specific test environments will dictate specific measurements of particular interest. In general, the following measures together give a reasonable "snapshot" characterisation of network performance. Repeated testing is required to determine the stability of performance, as mentioned below.

Suggested measures, for each direction:

- a) Listening Quality MOS
- b) Mean delay
- c) Echo level
- d) Delay variation
- e) Received speech level
- f) Received noise level
- g) Lost speech (due to lost packets and voice activity detection)

Note that items e) to g) are taken account of in the MOS prediction; whether it is necessary to measure them depends on whether the objective is to fully understand the underlying reasons for network performance, or simply to know how users will perceive the performance.

6. E-Model – ITU-T Rec. G.107

The E-Model is "... a computational model for use in transmission planning" (ITU-T G.107). It is an attempt to take into account all the factors which might influence perceived conversational quality, but it is not a measurement tool. The R Factor (R), on a scale of 0-100, is a "goodness" factor calculated as shown:

$$R = R_o - I_s - I_d - I_{e-eff} + A$$

where:

- R_o = Signal-to-noise ratio
 I_s = Simultaneous impairment factor (loudness, sidetone, quantisation distortion)
 I_d = Delay impairment factor (absolute delay, talker echo, listener echo)
 I_{e-eff} = Effective equipment impairment factor (may be derived from MOS-LQ; – see ITU-T Rec. P.834)
 A = Advantage factor

The G.107 document itself offers a number of cautions about the method:

- "Some experimental investigations suggest that the general tendency of the equipment impairment factors

is too pessimistic, so that a hidden security margin may be incorporated.”

- “The E-model supposes that different kinds of impairments are additive on the scale of the transmission rating factor R. This feature has not been checked to a satisfying extent.”
- “Some experiments show that the E-model disregards some masking effects occurring for talker sidetone, namely in conjunction with circuit noise, room noise at receive side, and low delay talker echo (<10 ms).”
- “Up to now it has not been clarified under which conditions the given values for the advantage factor should be applied. It is expected that these values may depend, e.g., on the user group, and that the absolute values will change in long term.”

In general, it may be more useful – particularly for troubleshooting - to measure the parameters listed in paragraph 5 above.

7. Test Guidelines: Speech Quality Assessment using the DSLA

Planning the test - 4 steps:

Step 1: Select the measurement end points – see Table 1

Step 2: Define the test category – see Table 2

Step 3: Select the test signal:

- Test signals must be speech or speech-like. The Phonytalk™ test process in DSLA uses carefully prepared material (ASTS). When real speech has to be used – for example when testing must be done with speech of a particular language – the signals must be of very high quality with low noise content, must cover a number of speakers of both genders and must be arranged into segments of 8-15 seconds of active speech (i.e. the duration excluding silence periods). The use of professionally prepared material such as the NTT (*) database is recommended.

Step 4: Select or write the test script

- Many test requirements can be met by using the Quick Start Examples supplied with DSLA.
- The naming of Quick Start Example schedules reflects the end point types. For example, Telephone-Handset.sch uses the channel A telephone line port and the channel B Handset port.
- When variations are required it is often beneficial to edit one of the standard scripts rather than start from nothing.

- Quick Start scripts beginning with the word “Network” are designed for making tests using two DSLA’s, one at one location (“local”) and the other at another location (“remote”).

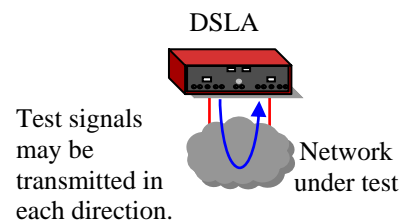
8. Understanding the Results

Very wide variations in MOS (± 1) can indicate inappropriate measurement techniques or an intermittent connection. If these sources of degradation can be eliminated then the network itself might be at fault. In a cellular network there could be cell loading, handover or rf reception problems. In a VoIP network the settings for QoS might be inconsistent or some equipment might not support QoS so high data loads cause significant packet loss or jitter.

Small variations in MOS (± 0.1) can be expected and are unlikely to cause end users any difficulty. Very few subjects can discern 0.1 MOS variations in a subjective test.

Variations in MOS of ± 0.5 are audible and excessive. The comprehensive range of techniques employed in DSLA enable users to understand the factors (codec distortions, jitter, packet loss, speech level) that might contribute to these poor scores.

Variations in speech level across networks frequently cause poor scores as the noise floor is raised in gain stages to compensate for low levels elsewhere. Inadvertent inclusion of network segments employing low bit rate codecs will also degrade the score. Voice Activity Detector settings are critical in ensuring speech is transmitted in its entirety without front end clipping. Inappropriate jitter buffer design and deployment is easily identified from the lost speech statistics and the graphical views in DSLA. The trade-off between increasing jitter buffer size to improve speech quality and the increase in delay is readily determined.



* See http://www.ntt-at.com/products_e/speech/ for details.

End Point	Advantages	Disadvantages
Analogue - Handset	Allows consistent connection approach to all network and phone types (VoIP, PSTN, cellular...)	Need to determine and standardise levels with each new handset; calls must usually be set up manually*
Analogue - Telephone Line	Easy and reliable connection type for PSTN, so offered on most equipment	Includes analogue loop which can be a source of atypical performance
Digital - ISDN	A direct connection into a network	Not always available and does not include terminal performance
Digital - IP	Convenient test point for isolating VoIP degradations and evaluation of VoIP terminal device (phone, gateway)	Good for IP to PSTN performance measurement, but provides an estimate only for the PSTN to IP path since an actual IP phone is not used, nor its characteristics taken into account.

* DSLA offers automatic call set up for some IP phone types

Table 1: Measurement End points

Test Category	Typical Purpose	Method
Connection check	Verify call set-up and speech path between end points	Measure speech level in each direction, for say 5 seconds. Check received signal and noise levels.
Quick quality check	Rapid assessment of instantaneous performance	Measure speech quality in each direction using say 6-8 seconds of active speech.
Multi-level quality check	Verify loss plan	Perform one speech quality test in each direction for a range of levels in (say) 3dB steps; plot speech quality score and output speech level against input speech level
Thorough quality check	Measure average speech performance for a wide range of speech sounds; identify incidence of intermittent problems	As for quick quality check, but repeated so as to use all ASTS speech files in both male and female voices.
Long duration quality test	Show performance variations over time, e.g. in relation to network load	Generally preferable to run repeated short duration tests over an extended period.
Engineering evaluation	Characterise performance as fully as possible - applications include regression testing, DSP and codec evaluation.	Use four different speech levels; analysis and statistical inference on results, including as appropriate speech quality, delay, delay variation, speech and noise levels, talker echo level.

Table 2: Test Categories

Mean Opinion Score, MOS	Listening Quality (Absolute Category Rating, ACR)	Degradation Category Rating, DCR	Listening Effort (Absolute Category Rating, ACR)
5	Excellent	Imperceptible	Complete relaxation possible; no effort required
4	Good	Just perceptible but not annoying	Attention necessary; no appreciable effort required
3	Fair	Perceptible and slightly annoying	Moderate effort required
2	Poor	Annoying but not objectionable	Considerable effort required
1	Bad	Very annoying and objectionable	No meaning understood with any feasible effort

Table 3: Mean Opinion Score