MOS Technology Brief

Mean Opinion Score Algorithms for Speech Quality Evaluation

Mean Opinion Score (MOS) is commonly used to rate phone service speech quality, expressed on a scale from 1 to 5, where 5 is best. A score of 4 is normally considered "Toll Quality" for calls placed over PSTN/TDM networks. MOS is a function of many factors, including the type of network and codecs used, wiring and premises equipment, and even the handset used to place the call.

MOS was originally determined using subjective listening tests, where a panel of trained experts judged recorded speech samples to assign an averaged score. Test equipment calculates MOS using sophisticated algorithms that are designed to closely approximate the results of subjective listening tests.

MOS is an overall indication of speech quality, aggregating dozens of speech impairments into a single score. Because it is a general metric reflecting many factors, the MOS score should not be used exclusively to characterize speech quality. The effects of echo, delay, call volume and clipping may not significantly influence MOS scores calculated by some algorithms, while actual callers may be quite irritated by their presence, especially if they affect a significant portion of the call.

A thorough speech quality assessment is best performed by reviewing primary quality measurements such as noise, distortion, echo, and delay, along with MOS from a number of algorithms, if possible. Knowing the strengths and weaknesses of the MOS algorithms used is critical to accurately interpret the resulting scores.

Tektronix®
Communications

MOS Algorithms

There are a number of industry-standard MOS calculation algorithms, each originally designed for a specific application. Some algorithms consider only packet-based (IP) statistics, whereas others include analog measurements such as noise, volume, echo, and distortion to enhance accuracy and repeatability.

Since the human ear is an analog device, and sound is an analog signal, it is important to include analog signal analysis when evaluating speech quality - MOS scores calculated using both analog & IP measurements most closely approximate a real caller's perception of speech quality.

PESQ ITU T P.862

The PESQ algorithm is based on a psychoacoustic, empirically-refined model designed to evaluate speech quality as perceived by actual callers.

The algorithm compares a reference speech file with a recording of the same file after being transmitted through a network or device under test, measuring the effects of one-way distortion and noise on speech quality.

Because the original reference and degraded recording are time-aligned and amplitude normalized for comparitive analysis, PESQ does not take into account the effects of delay, echo, and attenuation.


Inputs: Analog Measurements
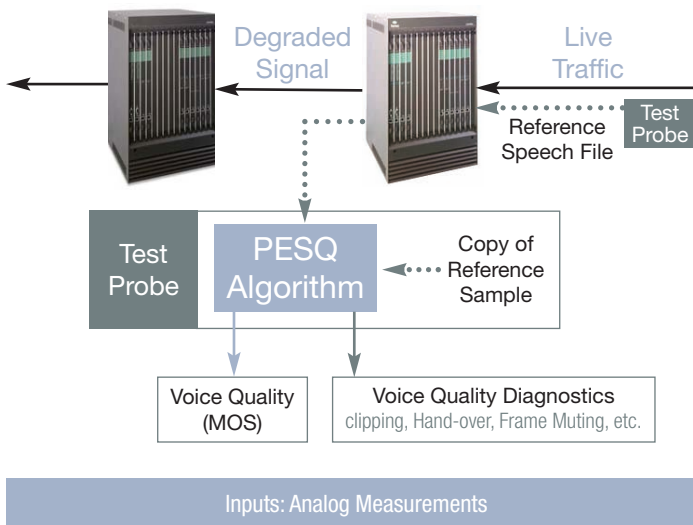
P.563 Listening MOS

P.563 is an extension to the PESQ P.562 algorithm that allows it to calculate MOS non-intrusively from actual speech samples or call recordings, suitable for single-ended testing.

The P.563/PESQ combination includes the effects of noise, echo, delay, clipping, frame mutes, packet-loss, codec and network type.


Inputs: Analog Measurements
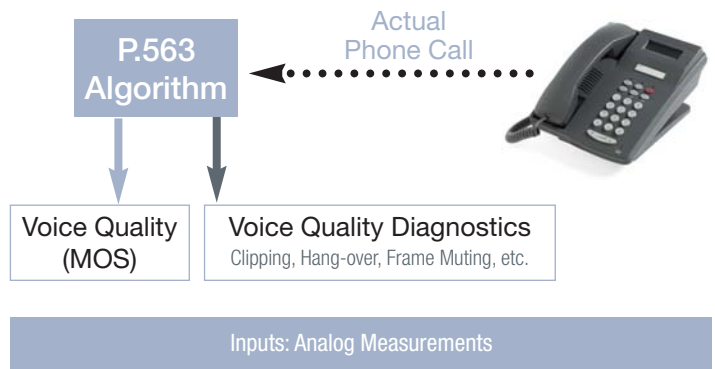
Typical MOS Scores

4.0   3.5   ~3
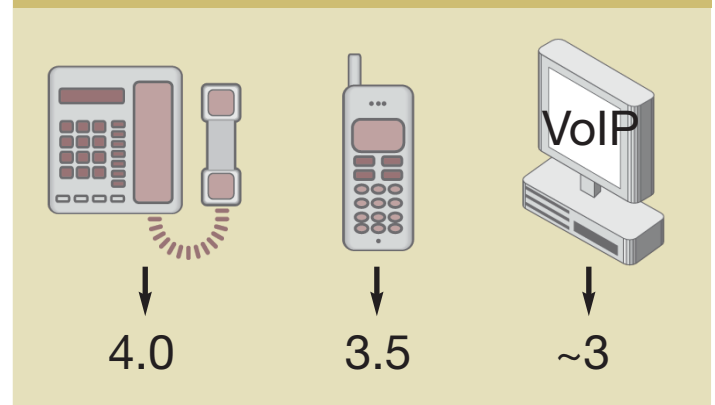
VQES Algorithm

A sophisticated, statistics-based algorithm that estimates end-user satisfaction and perceived voice quality of a call; VQES includes the effects of low volume (speech power), noise, distortion, echo and delay. The VQES algorithm also calculates P(UDI), the Probability of Unusable, Difficult or Irritating calls (0-100%), a metric reflecting a caller's level of frustration with the overall quality of the call. VQES MOS is affected by echo, but not delay, whereas the P(UDI) accounts for the effects of both.
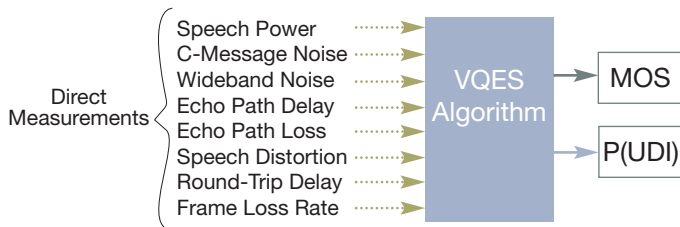
The graph below shows how MOS calculated by the VQES algorithm is affected by echo. Echo is characterized by echo path loss (EPL) and echo path delay (EPD), two interrelated metrics. The combination of these measurements affects the perceived effect of echo on speech quality. For example, both an increase in echo path delay or low echo path loss make echo more pronounced and distinct from the original signal.
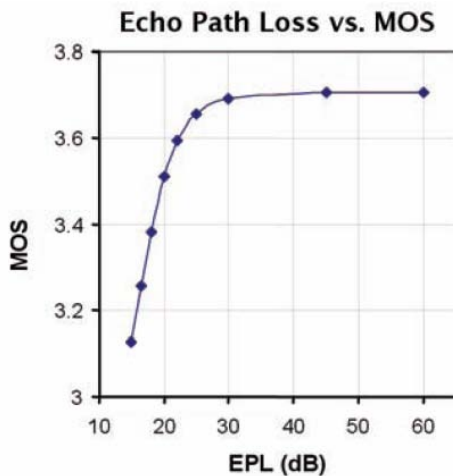
Advanced MOS Applications

Accelerated Call Quality Testing

Some active test systems use specially processed natural speech reference files that have silence and syllabic repetition removed. The most advanced files compress hours of natural speech into test calls lasting only a few minutes – providing greatly accelerated MOS measurements with the degree of accuracy and precision normally associated with longer tests.

VoIP Stress Testing

Test calls up to 24 hours in duration that report MOS at frequent intervals can be used as a "BERT" test for speech, helping operators to identify periodic issues and long-term call degradation. The quality of VoIP calls often worsen over time, as jitter buffers reach capacity and drop packets.

Listening MOS-based Tests

Algorithms that use actual speech to calculate MOS are ideal for single-ended testing where a human caller, IVR or Voicemail system can be used as a test source. Common applications include test solutions that can be used by VoIP customers to check the quality of their line, and tests that validate call quality to off-net locations where probes cannot be installed.

MOS calculated by the VQES algorithm

RTCP & RTCP-XR - IETF RFC-3611

RTCP (Real Time Control Protocol) is the control portion of the Real-time Transport Protocol (RTP) specification defined by RFC-3611, the standard commonly used to transport voice and video packets over IP networks. Standard RTCP packets contain basic quality information about the media transmission, including packet loss, duplication, jitter, packet time-to-live (TTL) and hop limit values. RTCP Extended Report (RTCP-XR) packets carry additional service quality details, including burst statistics, round-trip delay (RTD), signal and noise level, echo return loss, R-factor, listening and conversational MOS, and jitter buffer configuration. RTCP does not define a MOS algorithm per-se; it specifies a method to report media transmission quality.

Session end-points and network elements record and calculate the metrics exchanged in RTCP sender and receiver (RTCP-SR/RR) report blocks; the information can be used by compliant devices to dynamically adjust their configuration to optimize transmission quality. RTCP and RTCP-XR packets are also used by test systems to measure and report service quality. The additional quality metrics contained in XR packets are normally calculated by end-devices or test equipment using the E-Model algorithm, using standard RTCP packet statistics and context-appropriate default values as inputs (following page).
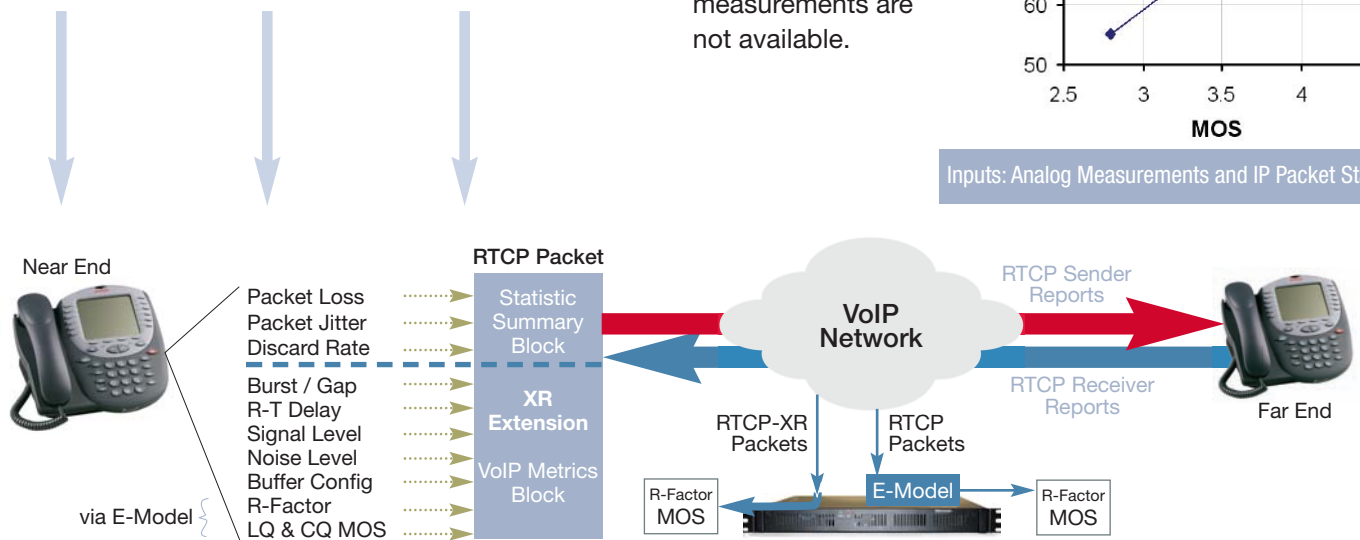
E-Model G.107, 108, 109

The E-Model rates conversational quality (CQ) and listening quality (LQ) using R-factor, which can be converted into respective MOS_{CQ} and MOS_{LQ} ratings, as well as two other aggregate metrics, the Good-or-Better (GoB) and Poor-or-Worse (PoW) indices. MOS_{CQ} incorportaes the effects of delay and echo on conversations, while MOS_{LQ} does not.

The E-Model was originally designed for network planning & codec testing - to measure the delay, echo, and distortion that result from digital speech compression/decompression and transcoding. Vendor-specific extensions have increased the versatility and accuracy of the E-Model by including VoIP-specific IP & analog measurements in R factor calculations.
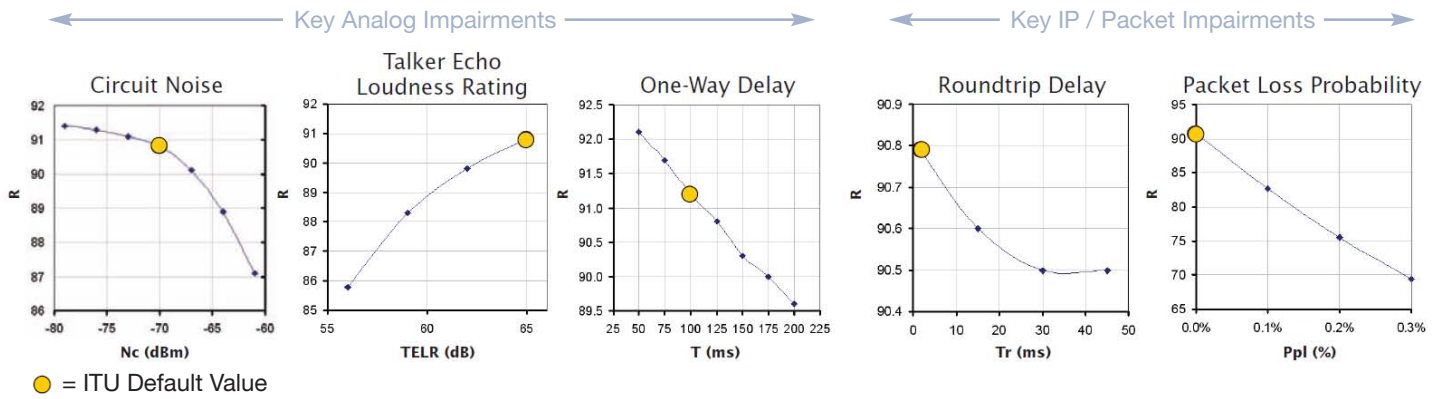
The E-Model algorithm is commonly used with IP only statistics, with nominal "default" values substituted for the missing analog measurements. The graphs below show how IP/Packet-based MOS can exhibit substantial error when actual analog measurements are not available.

MOS vs. R factor
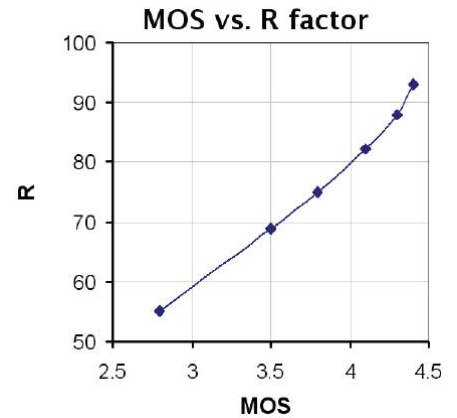
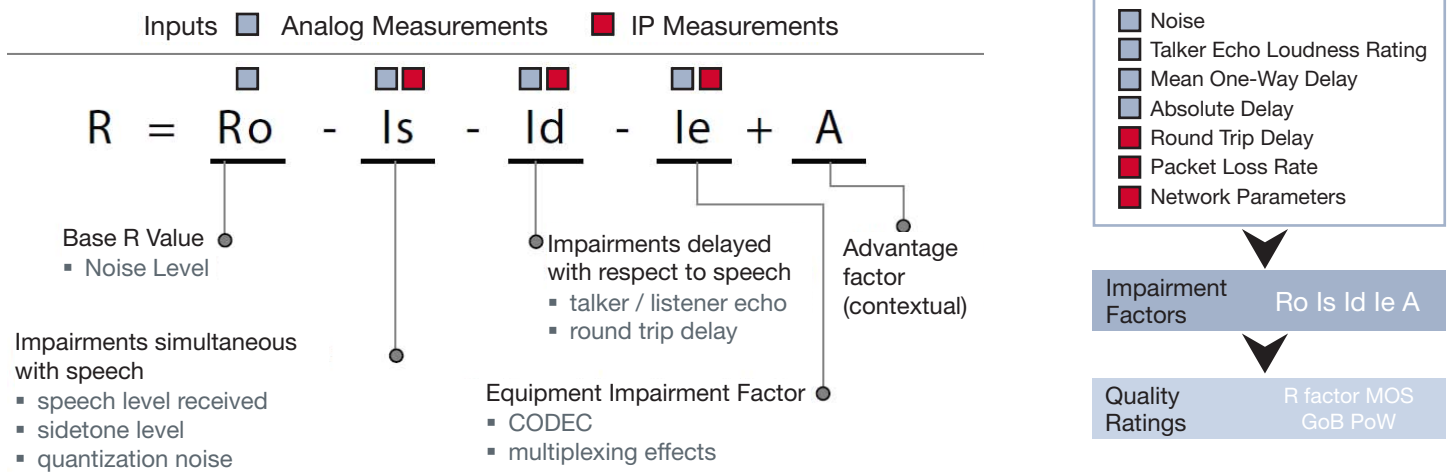Inputs: Analog Measurements and IP Packet Stats

Considers: Analog Measurements and IP Packet Stats

Key Analog Impairments ⟷ Key IP / Packet Impairments

Circuit Noise — R vs Nc (dBm)

Talker Echo Loudness Rating — R vs TELR (dB)

One-Way Delay — R vs T (ms)

Roundtrip Delay — R vs Tr (ms)

Packet Loss Probability — R vs Ppl (%)

○ = ITU Default Value

Values calculated using the Minacom E-Model simulator, using ITU recommended default values for all other variables

The E-Model Formula

Inputs ☐ Analog Measurements ☐ IP Measurements

$$R = Ro - Is - Id - Ie + A$$

Base R Value
- Noise Level

Impairments simultaneous with speech
- speech level received
- sidetone level
- quantization noise

Impairments delayed with respect to speech
- talker / listener echo
- round trip delay

Equipment Impairment Factor
- CODEC
- multiplexing effects

Advantage factor (contextual)

☐ Noise
☐ Talker Echo Loudness Rating
☐ Mean One-Way Delay
☐ Absolute Delay
☐ Round Trip Delay
☐ Packet Loss Rate
☐ Network Parameters

Impairment Factors — Ro Is Id Ie A

Quality Ratings — R factor MOS GoB PoW

About Tektronix Communications:

Tektronix Communications provides network operators and equipment manufacturers around the world an unparalleled suite of network diagnostics and management solutions for fixed, mobile, IP and converged multi-service networks.

This comprehensive set of solutions support a range of architectures and applications such as LTE, fixed mobile convergence, IMS, broadband wireless access, WiMAX, VoIP and triple play, including IPTV.

For Further Information:

Tektronix maintains a comprehensive, constantly expanding collection of application notes, technical briefs and other resources to help engineers working on the cutting edge of technology.

Please visit www.tektronix.com/communications

Contact Tektronix:

Please visit www.tektronix.com/communications

Phone:
1-800-833-9200 option 1
+1-469-330-4000

Locate your nearest Tektronix representative at:
www.tektronix.com/contactus

Tektronix®
Communications