voipfuture

WHITEPAPER

VOIP QUALITY MONITORING BASICS

Mean Opinion Score (MOS)

Calculation & Aggregation

Voice quality is typically expressed in terms of the mean opinion score (MOS), which – from its origins – is a subjective rating of audio quality on a scale from 1 to 5. For many relevant use cases, such as customer experience management and SLA monitoring, it is not practical to ask test persons for their opinion. The original concept has therefore been extended to cover cases where MOS is not the result of an empirical study, but the output of a computer-based analysis.

Historically, MOS is a subjective measurement. Listeners in a "quiet room" score the quality of a call as they perceive it.

All monitoring tools that provide a MOS essentially try to estimate the outcome of an empirical study.

To avoid any confusion between the different types of MOS, ITU-T Recommendation P.800.1 specifies terminology to distinguish the area of application, i.e. the source of a specific MOS value. The recommendation defines the following identifiers:

- LQ refers to Listening Quality, i.e. the quality is determined by what the listening party perceives
- CQ refers to Conversational Quality, i.e. the quality is determined by a conversational situation
- TQ refers to Talking Quality, i.e. the quality as perceived by the talking party only (potentially influenced by echo, double talk, etc.)
- S refers to Subjective, i.e. MOS determined through an empirical study
- O refers to Objective, i.e. MOS determined via automated end-to-end quality measurements, e.g. using PESQ or POLQA
- E refers to Estimated, i.e. MOS calculated using a planning model, such as the E-Model defined in ITU-T G.107

This leads to the following set of acronyms:

	Listening-only	Conversational	Talking
Subjective	**MOS-LQSy**	**MOS-CQSy**	**MOS-TQSy**
Objective	**MOS-LQOy**	**MOS-CQOy**	**MOS-TQOy**
Estimated	**MOS-LQEy**	**MOS-CQEy**	**MOS-TQEy**

The "y" at the end of the above acronyms is a placeholder for a descriptor of the respective reference audio bandwidth. Reference bandwidths can be narrowband ("N", 300-3400 Hz) or wideband ("W", 50-7000 Hz).

In practice these indices are often omitted or shortened and one needs to derive the area of application or MOS source from the context.

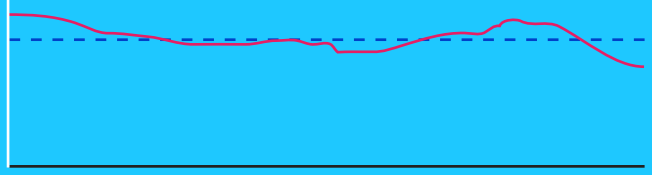
MOS estimates should reflect user experience. This whitepaper shows how to measure the users' quality of experience. In addition, it explains how Voipfuture Qrystal calculates and aggregates MOS values.

One MOS value can tell very different stories

Let's say the MOS value of a call is 3.8 - not good, not really disappointing

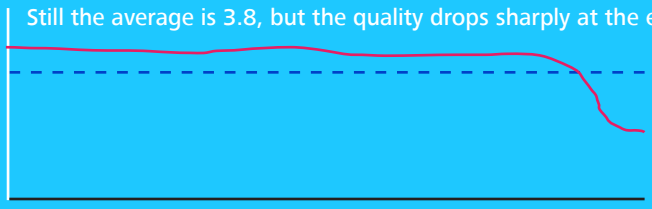
MOS 3.8

The simplest case: a call with small fluctuations around an average MOS of 3.8



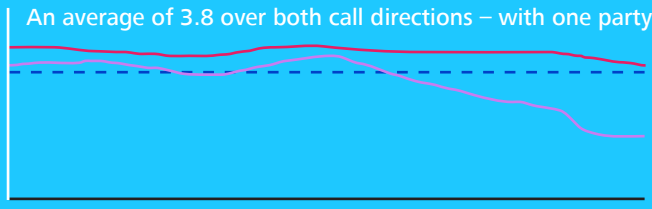
MOS 3.8

Still the average is 3.8, but the quality drops sharply at the end – the user terminated the call



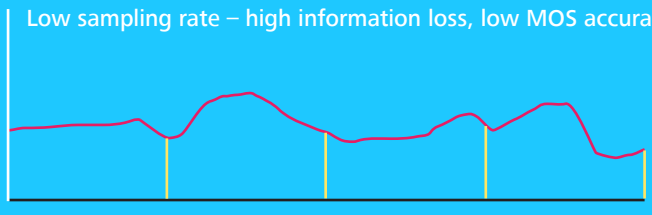
MOS 3.8

An average of 3.8 over both call directions – with one party experiencing worse quality

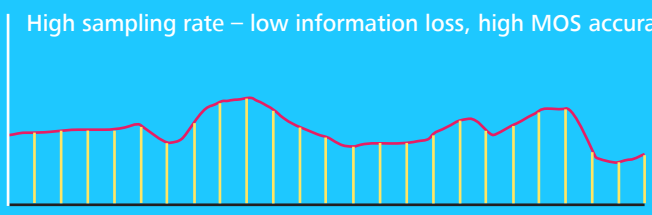


Obviously averages are misleading. A meaningful MOS requires time slicing – the slicing interval determines the quality of information.

Low sampling rate – high information loss, low MOS accuracy



High sampling rate – low information loss, high MOS accuracy



Voipfuture Qrystal is a passive monitoring system that corresponds to the Mode B method described in ITU-T Recommendation P.564. It performs passive measurements of the RTP flows and calculates an MOS_{LQE} and MOS_{CQE} using assumptions about the end points. While this describes the general approach – which is shared by many passive monitoring tools – there is something very special about the Voipfuture approach that makes a big difference.

MOS_{LQE} and MOS_{CQE} Calculation

Voipfuture's unique technology determines highly accurate listening quality estimates for fixed 5-second time slices of every RTP stream. The MOS_{LQE} estimates are calculated based on the E-Model defined in ITU-T G.107 using a number of input parameters, such as

- The codec
- Information about burst loss, i.e. consecutive packet loss (CPL)
- Information about the critical loss density (CLD), i.e. the number of packets received in between loss events
- Information about individual packet interarrival times (PIT)

Voipfuture developed the fixed time slicing technology to account for the varying nature of voice quality in IP networks. Conventional averaging per call leads to a significant loss of information. For example, the average quality of a perfect 5-minute call that suffers from severe impairments in the last 10 seconds will be close to perfect. Nevertheless, the parties will hang up and report a bad user experience to the help desk. Averaged metric data does not help to troubleshoot or even just confirm such problems.

Fixed time slice analysis of call quality enables impairments to be pinpointed in time, to identify time-correlated events in the network and to provide exact information on the user experience. This approach is also recommended by the TM Forum[1].

The quality data for each 5-second segment is summarized in a quality data record (QDR). The following figure illustrates the accuracy with which the QDRs reflect the media stream quality.

The basis for the calculation of the MOS_{LQE} is the R-factor formula defined in ITU-T Recommendation G.107. For passive monitoring systems such as Voipfuture Qrystal, the only relevant parameter for this formula is the equipment impairment factor ($I_{e\text{-}eff}$):

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

Details about the formula and the meaning of its parameters can be found in the aforementioned ITU-T standard.

[1] TM Forum GB934, 'Best Practice: Voice over IP SLA Management'

Voipfuture Qrystal derives an accurate $I_{e\text{-}eff}$ value that reflects all impairments for a fixed 5-second time slice. The $I_{e\text{-}eff}$ is calculated using input from the Voipfuture algorithm that analyzes the CPL, CLD and PIT histograms available for every 5-second interval of an RTP stream.

Section 9.1 of ITU-T P.564 requires that speech quality be estimated using audio samples of 8-30 seconds in length. Therefore, a sliding window approach is used to combine two 5-second into one 10-second time window so that sufficient information is available to calculate an R-factor value. The R-factor value is then converted to an MOS score via standardized conversion tables.

Note that Qrystal provides MOS_{LQE} not only on the conventional narrowband but also on the super wideband scale which accounts for increased user expectations when wideband codecs are introduced. Furthermore, if delay is available from RTCP timing information, the MOS_{CQE}, i.e. the conversational quality estimate is provided as well.

MOS Aggregation

Obviously, quality per fixed unit is useful for troubleshooting, e.g. when it comes to identifying time-correlated events on the media plane. However, quality data for fixed time units also enable service quality information to be compared and aggregated in a meaningful way. Using conventional average MOS per call comparison and aggregation is difficult to say the least. For example, an average MOS per call does not enable you to tell whether a 5-minute call with MOS 3.9 is better or worse than a 1-minute call with MOS 4.0 as there is no common reference for comparison.

Comparison and aggregation becomes simple if data is available for fixed time units. The MOS values of a stream's time slices can be aggregated into distributions of MOS-qualified call minutes, or MOS Minutes for short. To this end, the 5-second MOS values are entered into a MOS Minute histogram with five bins as defined in ITU-T Recommendation G.107.

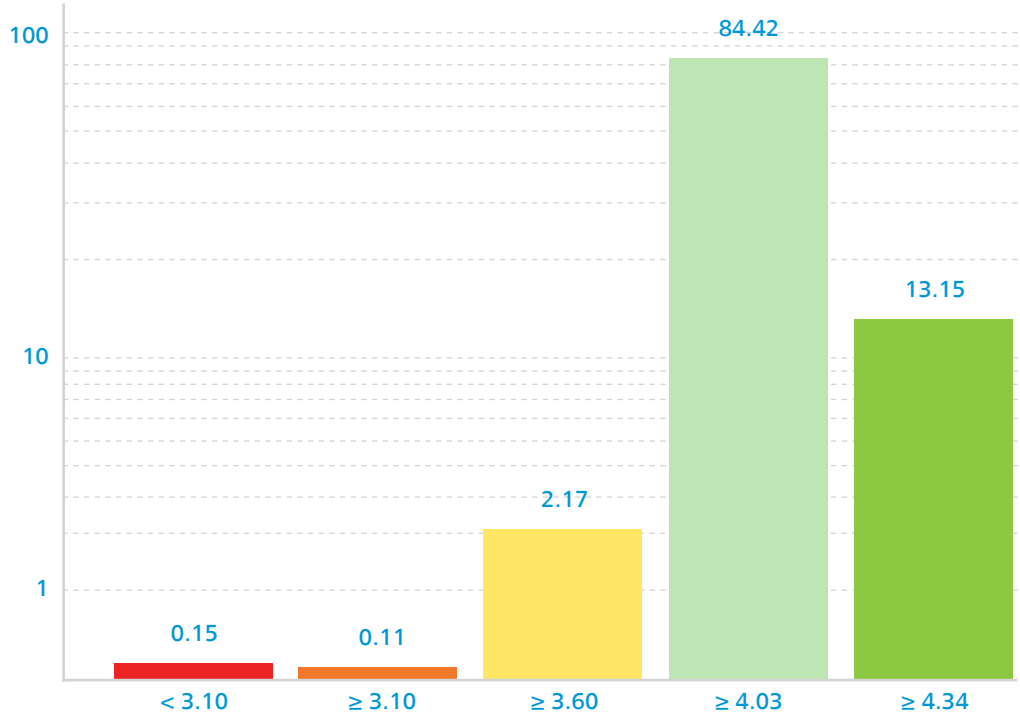
MOS CLASS	MOS	USER SATISFACTION
5	4.34	Very satisfied
4	4.03	Satisfied
3	3.60	Some users dissatisfied
2	3.10	Many users dissatisfied
1	2.58	Nearly all users dissatisfied

Using MOS Minutes it is easily possible to create advanced stream-level metrics, e.g.

- Good Stream: A "Good Stream" is a stream where all QDRs with a valid MOS have a score above 4.03, i.e. the user experience is good from beginning to end

- Good Minute Ratio (GMR): The Good Minute Ratio is calculated for every stream. It is the number of 'good' QDRs divided by the number of all QDRs. A 'good' QDR is a 5-second segment with an MOS > 4.03.

This type of aggregation is obviously not only possible on the level of a single stream or call, but also on entire groups of streams.



Aggregate statistics can be created on grouping criteria such as SIP trunks, media trunk, probes and numbering plan entries. For example, it is possible to create MOS Minute statistics for all streams originating from a specific provider in Mexico. These MOS Minute statistics can be found throughout the Voipfuture Qrystal Manager, providing concise information about the service quality of the calls and streams in focus.

MOS Minutes are also used for voice quality SLAs, where they enable service quality to be expressed by billing unit.

About

Voipfuture develops technology for monitoring and analyzing media quality in IP networks - to control and evaluate network performance and the quality of Voice over IP services.

Our monitoring solution covers all standard VoIP services including NGN, IMS and LTE networks.

Its integration into existing IT infrastructures is facilitated through the use of off-the-shelf server hardware from major vendors.

Our unique software technology is exceptionally scalable. It makes the products applicable to a wide range of network environments, from enterprise branch sites to major interconnection points and complete carrier networks.

Flexible deployment options provide for various applications

- Troubleshooting
- Network performance optimization
- Customer care applications
- Interconnection peering
- Customer SLA monitoring

Our products' open interfaces allow tight integration with existing network management solutions and customer experience management systems.

Wendenstr. 4
20097 Hamburg
Germany

+49 40 688 90 01 0
info@voipfuture.com
www.voipfuture.com