VIAVI Solutions

White Paper

Voice Analysis for Mobile Networks

Audio Quality Scoring Principals for Voice

Quality of experience analysis for voice

Defining user quality of experience for voice calls over network infrastructure has been around for many decades, from its simplest origin in which people would sit in a room listening to the calls and scoring the quality of the network subjectively to today's compute intensive assessment algorithms. The key commonality over the years with all of the assessment techniques is the delivery of a quality metric or score commonly known as a Mean Opinion Score (MOS). Generally, a MOS rating of between 4 and 5 represents a good quality of experience for that voice call.

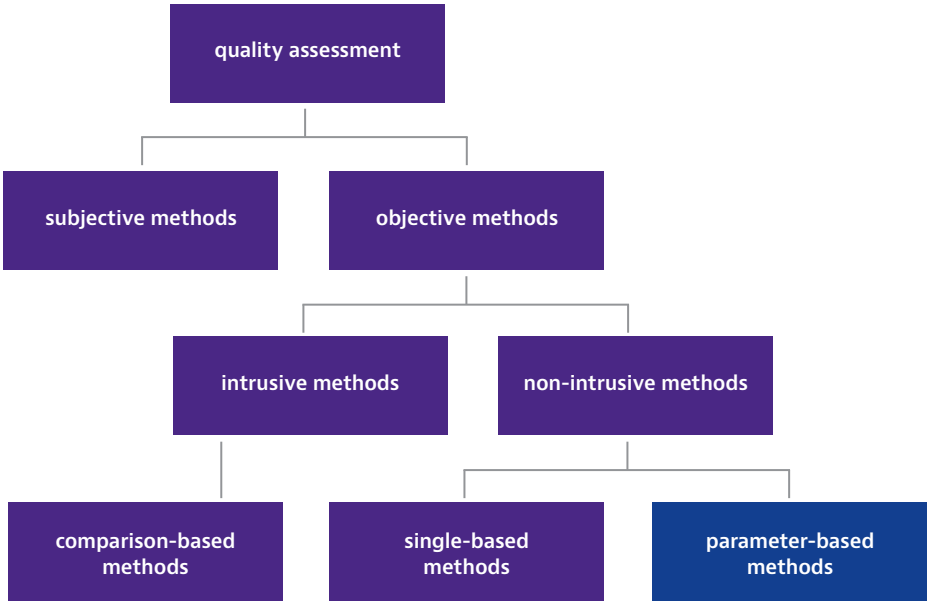

Figure 1: Voice quality assessment techniques

With the move from circuit to packet switched networking a correlation is needed between the voice quality scoring mechanism and the influence that the network packet transport has on voice quality performance. Quantitative measurements for network performance are collectively called Quality of Service (QoS) metrics, QoS enables user's express network performance objectively. As a result the subjective method is replaced with an objective method.

Figure 1 above outlines the techniques used to assess voice quality, with the focus of this paper on voice quality analysis for mobility, the aim is to select a suitable measurement technique from the objective methods.

Objective assessment can be categorized as Intrusive or Non-Intrusive. Intrusive methods are based on comparing a transmitted voice sample with the received voice sample (comparison based methods). The most widely recognized of these techniques is called Perceptual Evaluation of Speech Quality (PESQ) or its successor which has been adapted for mobility Perceptual Objective Listening Quality Assessment (POLQA). The comparison methodology means analysis is done offline or when the call has been completed. As the technique involves comparing file samples it is compute intensive. This makes intrusive assessment of voice somewhat limited in terms of the number of calls that can be assessed.

The alternative is to use Non-Intrusive techniques to determine voice quality, the methods are based on inspecting the audio signal in isolation, either by analyzing the audio signal (signal based methods) or by analyzing the IP network packets transport layer parameters (parameter based methods). The parameter based assessment methods enables live assessment of voice quality and indeed is far more scalable in terms of the number of calls that can be assessed for voice quality.

A key take-away about all of the objective assessment methods is that they are based on computer algorithms that produce the MOS score. During their development and validation they are all subjected to a calibration process during which the algorithm-generated MOS scores are compared to subjectively produced MOS scores (real people listening and scoring the same audio samples) to align the results. Under ideal conditions, this results in a POLQA MOS score of between 4 and 5 correlating to a MOS sampling score of 4 and 5 for non-intrusive voice quality assessment.

Ultimately, the decision to choose one technique over the other comes down to what it is that's under test and cost. For example, if a user is looking to determine how a new mobile phone behaves on a network they most likely use POLQA, or if its assessment of networking elements then a user would opt for the lower cost option and use non-intrusive analysis.

Correlating MOS ratings to network quality of service

The most relevant method of measuring voice quality in order to evaluate the experience that the subscriber is having is to use a listening-only like MOS score. Highlighted below are the different cases used to distinguish voice quality depending on the assessment method which is in use.

MOS-LQS	Subjective tests carried out according to ITU-T Recommendations P.830, P.835 and P.840 give results in terms of MOS-LQS.
MOS-LQO	The score is calculated by means of an objective model which aims at predicting the quality for a listening-only test situation.
MOS-LQO (electrical)	This kind of measurement is performed at electrical interfaces of the terminal only. In order to predict the listening quality as perceived by the user, assumptions for the terminals are made in terms of IRS or corrected IRS frequency response; this implicitly includes the assumption of a sealed condition between the handset receiver and the user's ear. ITU-T Recommendation P.862 falls into this category.
MOS-LQO (acoustic)	This kind of measurement is performed at acoustical interfaces of the terminal only. In order to predict the listening quality as perceived by the user, this measurement includes the actual telephone set products provided by the manufacturer or vendor. In combination with the choice of the acoustical receiver in the lab test ("artificial ear"), there will be a more or less leaky condition between the handset's receiver and the artificial ear. Consequently, for more realistic test scenarios, there may be a degradation of the measured MOS value, while for more artificial test scenarios there may be a negligible difference. ITU-T Recommendation P.863 partially falls into this category.
MOS-LQO (acoustic)	This kind of measurement is performed at transport interfaces. The model is based on a scalar transmission rating value, R, which is then converted to MOS-LQO value. ITU-T Recommendation G.107 falls into this category.

Detailed comparison of quality assessment methods

As highlighted in the previous chapter there is a number of assessment techniques used to derive voice quality which are dependent on the location of the listening point. The following figure outlines an architecture in which each of the unique measurement techniques is used to derive a reliable and repeatable voice quality score. This figure is also useful to outline the associated benefits and cost associated with each MOS score technique.

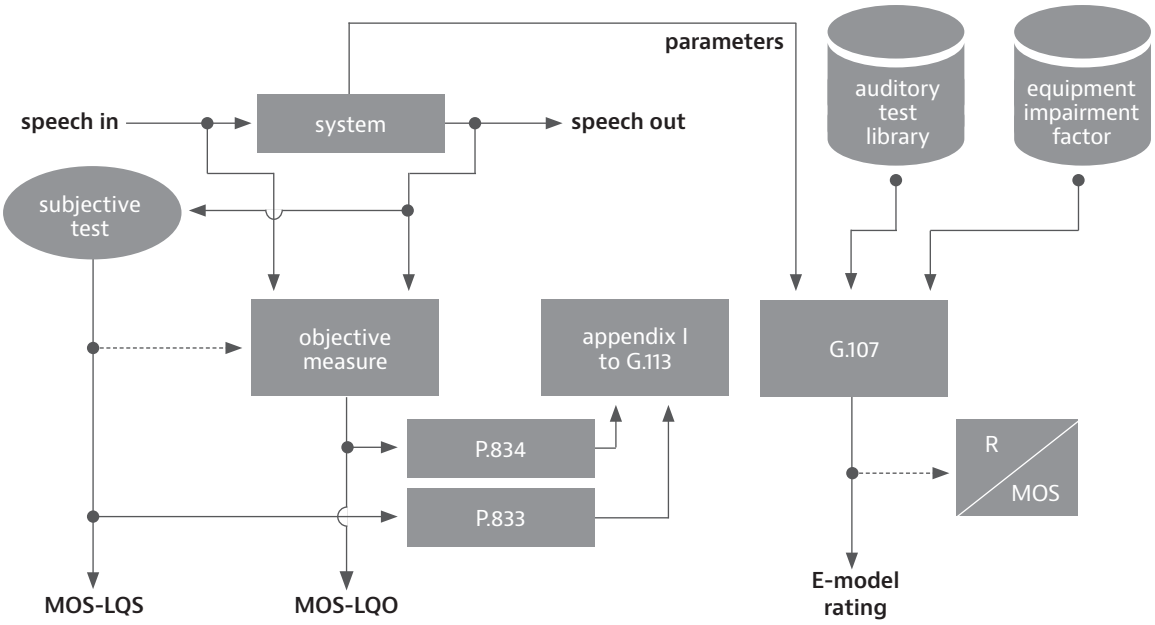

Figure 2: Voice quality analysis algorithm architecture

Subjective method

The limitation of subjectively based methods is that they are time-consuming, and so evaluating various different scenarios — especially at the scale required for evaluating wireless telephony — is prohibitively expensive.

These include the ITU-T Recommendations P.800, P.830, P.831, P.832, P.835 and P.840.

Comparison based methods

Comparison based methods give the best accuracy, as they analyze the difference between the audio signals transmitted from one caller and then received by the other caller. They require that the transmitted signal is known, so they are the most computing-intensive method.

This makes them expensive for assessing the quality of thousands of simultaneous calls. Beside the effect of the impairments of the transmission network, comparison based methods also capture the effects of transcoding, echo cancelation and all other type of audio signal impairments. Audio signal impairments are not relevant when testing and evaluating the effect of the transport layer network.

For Example ITU-T Recommendation P.863 (POLQA)

Signal based methods

Signal based methods give good accuracy, but are better suited to laboratory-testing scenarios as there is an inherent need to extract the actual audio stream in some form and then subject this to the analysis algorithm. With these methods, it is also possible to scale the covered number of scenarios significantly; but they require specialized solutions and are computing intensive.

Another benefit is that these methods do not require comparison samples as do comparison based methods, so they scale better for the quality assessment of thousands of simultaneous calls.

For Example ITU-T Recommendations P.862.1, P.862.2 (PESQ) and P.863 (POLQA)

Parameter based methods

Parameter based methods give good accuracy as well as scaling efficiently for assessing the quality of thousands of simultaneous calls. They do not require the comparison sample and do not need the full extraction of the audio stream, as they are based on analysis of the transport-layer parameters.

For example ITU-T Recommendation G.107 and G.107.1 (ITU E-Model)

TeraVM's MOS scoring is based on the algorithms outlined here. Unlike POLQA, the TeraVM voice MOS score is lightweight, meaning it does not require any other dedicated hardware for analysis purposes. As already indicated under ideal conditions, when there are no impairments, both MOS scores produce toll quality values between 4-5!

TeraVM quality assessment implementation for mobile

In testing eNB capacity, the voice-quality impairments are derived purely from the delay, jitter, out-of-sequence and lost-packet phenomena. The TeraVM audio- and video-quality assessment algorithm therefore expands the basic parameter based method by taking into consideration not only the transport-layer metrics, but also the metrics of the higher-layer audio transport protocol, and enhances these results by modeling the behavior of the UE (user equipment) by buffering the audio/video stream before starting the analysis.

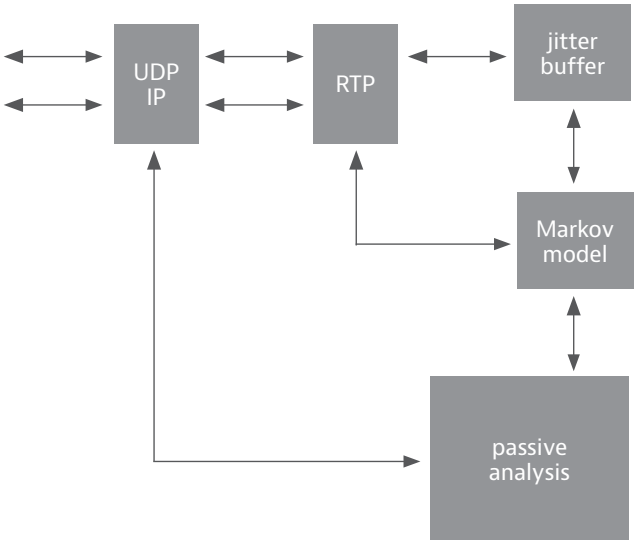

Figure 3: TeraVM MOS analysis algorithm architecture

While the analysis is being conducted the UE model also takes into account the Jitter buffer that all UEs use to mitigate the effects of the wireless environment and the delays and jitters inherent in it. The Markov Model is designed to measure the distribution of lost and discarded packets, it is designed to detect two primary states within calls: the burst state, in which the rate of packet loss and discard is high enough to cause noticeable degradation in quality, and the gap state.

Recommendations for meaningful VoLTE capacity testing

For evaluating the eNB capacity performance over the RF or CPRI (common public radio interface) interface, the recommendation is to always establish a baseline MOS score against which the performance is compared. This score can then be used to evaluate the eNB performance when:

- The number of simultaneous telephony calls is increased
- The calls are subjected to varying fading and mobility environments
- Different air-interface technologies are used to alleviate the effect of the above.

The MOS evaluation which is run with codecs operating at different audio bandwidths e.g. an ITU-T G.711 voice codec, condition will often yield a score above 4.0 in a narrow-band (300-3700 Hz) test; whereas it is more likely to yield a score in the range of 3.5-3.7 in a wideband (50-7000 Hz) test, due to the presence of the higher quality wideband samples.

These will quantify the performance of the eNB when evaluated using the MOS score and the transport layer parameters, which give an insight into the reasons for performance deviations on the air interface.

Conclusion

Understanding the performance of mobile networks can easily be assessed using voice quality measurement techniques. The key is to select the correct voice quality measurement technique as a wrong decision can be costly. The first step is to reflect on what it is that needs to be assessed i.e. is it the network transport layer or higher up at a device level?

The two most common technologies for voice quality assessment are non-Intrusive MOS scoring (as delivered in TeraVM) and POLQA, both of which can be used to characterize a voice transmission system. The major differences between the two is price, scalability and complexity of the solutions. A non-intrusive MOS score technique uses network packet parameters, making it ideal for network transport layer assessment. Plus non-intrusive MOS is less costly, scales well and has least complexity to implement.

On the opposite side is POLQA, which compares the received audio signal with the expected signal making it ideal for device level assessment. POLQA MOS scoring is highly accurate as it takes into account transmission network impairments, the effects of transcoding, echo cancelation, and any other type of audio signal alteration. As the POLQA algorithm is computationally intensive, it is not practical to use it for testing the speech quality with more than a handful of handsets.

Further comparison of the two voice quality measurement techniques reveals that under ideal conditions i.e. no impairments presence, that both techniques will deliver a MOS score between 4 and 5. Therefore, when it comes characterizing the performance of the mobile backhaul network, non-intrusive MOS scoring as delivered in TeraVM offers greatest savings, with maximum number of voice call quality measurement points with the least complexity.

VIAVI Solutions

To reach the VIAVI office nearest you,
visit viavisolutions.com/contact

voiceanalysis-wp-wir-nse-ae
30187527 900 1018