Voice Over Packet Test Applications

Using GoldenVoice™ impairment measurements as a method of generating voice quality scoring with GoldenMOS™

A white paper describing a practical method of relating impairment measurements to MOS, PSQM, and PESQ utilizing GMOS technology

Ameritec

mailto:askzeke@ameritec.com

VoP Test Applications
Using GoldenVoice impairment measurements as a method of generating voice quality scoring with GoldenMOS, Golden PSQM (G-PSQM) and Golden PESQ (GPESQ)

Contents

Generating MOS Based on Measured Impairments

I. Introduction

Since 1998 Ameritec has been a provider of Voice Over Packet test solutions for companies whose business focus is the development of voice over packet networks and network components. The solution Ameritec has adopted for testing packet-switched networks is unique in the market segment. The market has seen methods of estimating voice quality scores using a host of subjective and quasi-objective methods. But, while these methods provide an estimation of voice quality, they do little to give any indication as to the sources of the impairments that adversely impact the scores. The great challenge for a developer has been to determine how to improve the performance of their product without metrics that highlight the potential sources of the problem. This is why Ameritec's focus for a VoP test product has been based on stimulating the network with a stimulus of a known characteristic, GoldenVoice, and then reporting results based on objective measurements. The results of these tests provide developers a broad range of measurements that allow them to evaluate the product's performance based on "*real*" numbers. In a white paper authored and published by Ameritec in August 2000, "*Filling the VoID in VoIP Testing*", we provided developers with insight into the correlation of impairments with specific elements of a packet-switched network.

While individual objective measurements are an excellent tool, there has been a continuing desire on the part of the development community to identify an internationally adopted standard based on the use of measured impairments to calculate a voice quality score. In May 2000 the International Telecommunications Union (ITU) released ITU-T Recommendation G.107 – The E-Model, a computational model for use in transmission planning. The standard was originally released in 1996 and then updated in 2000 to address transmission planning of networks comprised of packet-based components. The output of the E-Model is a R-Rating, or R-Factor, that is used to calculate a Mean Opinion Score (MOS), %GOB (percent good or better) and %POW (percent poor or worse). On the basis of market demand and the worldwide acceptance of MOS, Ameritec's GoldenMOS™ application has focused on using the algorithm from Annex B of ITU-T Recommendation G.107 to calculate MOS. The combination of Ameritec's recently released VoPtest solution and GoldenMOS gives the development and validation community a comprehensive set of tools that not only allows them to measure impairments, but to automatically produce a MOS based on the results of that test data.

The content of this document will focus on Ameritec's implementation of the E-Model equation into a FeatureCall ME™ and Conductor™ software tool named GoldenMOS™. FeatureCall ME™ is the Graphical User Interface (GUI) used to manage and control Ameritec's analog and digital models of AM2 Niagara, AM2S Squirt, and Crescendo Call Generators. Conductor is the Graphical User Interface (GUI) used to manage and control Ameritec's Allegro and Fortissimo Call Generators. FeatureCall ME™ and Conductor™ supports test environments built from Ameritec products for testing from single line IADs (Integrated Access Device) to multiple DS3 configurations from a single control tool – where scalability and functionality are the criteria: Ameritec is the solution.

II. Voice Scoring Standards

The market today is faced with a wide choice of voice scoring methods. Each one has its own benefits and liabilities. Some tests are designed to evaluate the quality of a conversation between a calling and a called party, while others are specifically designed to measure degradation caused by codecs, and not the overall characteristics of the call through the network.

A. ITU-T Recommendation P.800 – MOS (Mean Opinion Score)

This method of estimating mean opinion score (MOS) of a conversation has been the standard for grading the voice quality of networks (circuit-switched or packet-switched) since 1996. The P.800 is actually comprised of many different types of listen only tests and offers opinions on: (1) a Listening-quality scale; (2) a Listening-effort scale and (3) a Loudness-preference scale. Other scales are defined in P.800 but the above listed are the

most common. The tests are performed by a group of 40-60 *"screened"* human listeners who grade conversational clips based on a scale of 1-5 with 5 representing an excellent MOS. All other methods of assessing the speech quality reference the P.800 recommendation
The disadvantage of this method is that all measurements are subjective (human listeners) and are not reproducible. Therefore they are expensive and time consuming. A given MOS is affected by language, age and gender of the speaker. In addition to the subjective nature of the tests, no meaningful data is presented to assist in the isolation of network problems that impact the ability to comprehend the content of the conversation. Also this type of testing does not lend itself to automation because the nature of the tests requires people.

B. ITU-T Recommendation P.861 – PSQM/PSQM+ (Perceptual Speech Quality Measurement)

The PSQM algorithm was developed to overcome the subjective nature, time and expense of MOS testing and to provide an objective test algorithm. In addition to developing a method of objective testing the goal of P.861was to isolate the environment of the speaker and the listener and only measure the degraded speech that occurs through a speech codec. Therefore the principal application of PSQM is to characterize speech codecs and ***not*** to test an end-to-end call. The test consists of statistically comparing the output of the codec against the input of the codec and reporting degradation on a scale of 0 to 6.5 with 0 being a perfect comparison. The test supports the use of artificial voices that conform to P.50 or real voices recorded under the stringent requirements of P.830.

The PSQM+ algorithm, as described in Appendix II of P.861, is an update to the original PSQM algorithm and is intended to overcome some of the deficiencies as defined by the ITU
with its release. While the changes broaden the application of PSQM+ by incorporating changes that compensate for channel errors and making the algorithm more applicable to a wider range of codecs, it still does not support delay as a variable.

The key to both PSQM and PSQM+ is time alignment. The algorithms have very limited tolerance for any amount of delay variation. Delay is a characteristic of the packet-switched network, and the wide ranges of delays are generally a result of the load conditions of access
nodes, or gateways, which make variable delays in *real-life* a network characteristic. Also, neither PSQM nor PSQM+ provide meaningful data to assist in the isolation of problems that
impact the score of the algorithms.

While both PSQM and PSQM+ are good resources for evaluating the performance, or relative
quality, of a speech codec, they do not completely replace the requirement for MOS to make
an end-to-end estimation of objective quality.

C. ITU-T Recommendation P.862 – PESQ (Perceptual Evaluation of Speech Quality)

A draft of PESQ was released in 2001 to the development community by the ITU as a "Method for end-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs." The intent of the algorithm is to broaden the application from testing only speech codecs but also for use in testing network applications. The algorithm takes into consideration filtering, distortion, channel impairments and some amount of variable delay. In the Scope (Section 4) of P.862 it states, "Although correlations between objective and subjective scores in the benchmark were around 0.935 for known and unknown data, the PESQ algorithm cannot be used to replace subjective testing."

PSQM, PSQM+ and PESQ all share a common problem, they have a very limited tolerance for any amount of delay variation. Delay is a characteristic of the packet-switched network, and the wide ranges of delays are generally a result of the load conditions of access nodes, or gateways, which make variable delays in *real-life* a network characteristic. Also, neither PSQM, PSQM+ nor PESQ provide meaningful data to assist in the isolation of problems that impact the score of the algorithms.

D. PAMS – Perceptual Analysis and Measurement System

The PAMS system provides detailed analysis of the performance of a speech codec accessed through an analog line. As with PSQM and PESQ, the PAMS algorithm compares a known reference model with a degraded test signal and reports the spectral anomalies found in the test sample. The PAMS algorithm uses time alignment, level alignment, equalization, auditory manipulation and other techniques of conditioning the sample to accurately report detected variances from the reference model. The output of the PAMS algorithm is a listening effort score (1-5), listening quality score (1-5) and distortion measurements. The listening quality score is generally associated with the subjective quality measurement (MOS) of P.800.

PAMS tends to be a resource intense testing tool and therefore is very expensive to incorporate into a multi-channel test tool.

E. ITU-T Recommendation G.107 – The E-Model

The E-Model is a tool designed for transmission planners and takes in account the characteristics of the transmission fabric, the handset and the environments of the speaker and the listener. The object of G.107 is to model the traits of an end-to-end conversation and, based on projected impairments, estimate the voice quality, or MOS, of the connection. The output of this equation gives network planners the projected performance of the networks that they intend to implement.

The E-Model has several characteristics that make it especially suited for the development and validation community, because it isolates impairments of delay, speech codec characteristics, packet loss and the listener environment -- all variables that affect the performance of today's networks.

While G.107 defines the equation, G.108 provides the application rules for the usage of the E-Model and in Section 9.6 discusses the use of computer programs for the specific equation parameter input as an acceptable utilization of the equation.

III. What is GoldenMOS?

GoldenMOS is a combination of the data collection ability of an Ameritec Call Generator and the power of FeatureCall ME™ and Conductor™ to process data to produce channelized MOS scores and related impairment data. The GoldenMOS application resident in FeatureCall ME™ and Conductor™ is Ameritec's implementation of the G.107 recommendation and calculates both the R-Factor and MOS. GMOS builds on the objective measurements produced by the VoP feature package. Figure 1 is a sample of the report generated by the GoldenMOS application.

GoldenMOS

QoS REPORT

Unit: DS3, Script: CAS-T1 Term VoP QoS 93Q0057.qsc, Codec: G.711(w/o PLC)

Sp Ch	GMOS	GMOS Median	RFactor	RFactor Median	G-PSQM	G-PSQM Median	G-PESQ	G-PESQ Median
2 1	4.4		92.1		0.7		4.3	
2 2	4.4		92.1		0.7		4.3	
2 3	4.4		92.1		0.7		4.3	
2 4	4.4		92.1		0.7		4.3	
2 5	4.4		92.1		0.7		4.3	
2 6	4.4		92.1		0.7		4.3	
2 7	4.4		92.1		0.7		4.3	
2 8	4.4		92.1		0.7		4.3	
2 9	4.4		92.1		0.7		4.3	
2 10	4.4		92.1		0.7		4.3	
2 11	4.4		92.1		0.7		4.3	
2 12	4.4		92.1		0.7		4.3	
2 13	4.4		92.1		0.7		4.3	
2 14	4.4		92.1		0.7		4.3	
2 15	4.4		92.1		0.7		4.3	
2 16	4.4		92.1		0.7		4.3	
2 17	4.4		92.1		0.7		4.3	
2 18	4.4		92.1		0.7		4.3	
2 19	4.4		92.1		0.7		4.3	
2 20	4.4		92.1		0.7		4.3	
Span 2	4.4	4.4	92.1	92.1	0.7	0.7	4.3	4.3

<u>Figure 1</u>
GoldenMOS Report

IV. Generating Data for the E-Model Equation

The key functionality of Ameritec's implementation is the ability to use the call generator to report all the necessary values of delay and voice packet loss to properly execute the E-Model equation and calculate the R-Rating. In order to generate the data it requires a stimulus that represents the spectral dynamics of speech and a detection mechanism that is capable of measuring impairments, and detecting other anomalies in the call. GoldenMOS is a listen-only test. Therefore the specialized script used to collect the data for the GoldenMOS application is based on all measurements being made on the Terminate side of the call. Figure 2 is a diagram of the call flow required to generate data for the GoldenMOS application.

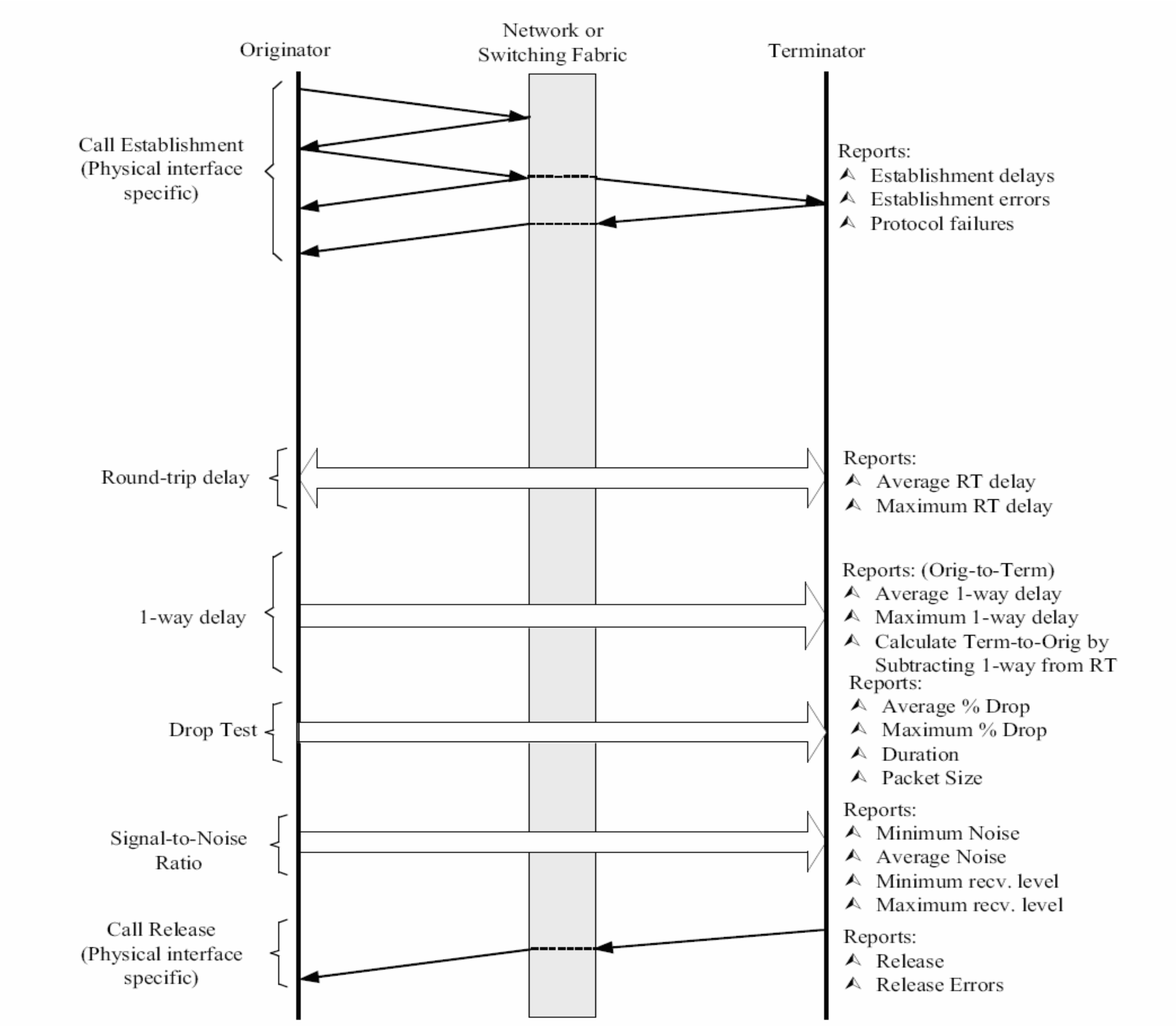


Figure 2
GoldenMOS Call Flow Diagram

A. The characteristics of GoldenVoice

GoldenVoice is a complex tone sequence developed by Ameritec in response to the testing demands of packet-switched communications systems and simulates the spectral range and dynamics of human speech. A key component of packet-network gateways are Vocoders that are optimized for human speech and typically do not pass the simple test tones traditionally used in circuit-switched voice communications testing. GoldenVoice was developed from earlier work in analog voice transmission impairment testing (Ameritec is a world leader in TIMS testing products) and is a test stimulus that will pass reliably through all modern Codecs (G.711, G.723.1, G.726, G.728, G.729A/B, etc.) using various packet loss concealment (PLC) methods and voice activation detection (VAD) schemes. Because of the dynamics of GoldenVoice it provides a stimulus that aggressively exercises Codecs across the voice spectrum, thus providing the user with a tool that is not affected by age, gender or language biases associated with recorded voice clips.

B. VoP^3 Measurements Required for GoldenMOS

The heart of the GoldenMOS application is a specialized script written by Ameritec to collect and report the measurements necessary to data fill the E-Model equation. The E-Model equation requires directional delay measurements, the % packet loss, Signal-to-Noise Ratio, circuit noise and receive signal level. Other data input is required to identify the Codec Type, used by the GoldenMOS application, and voice packet size. Identification of the voice packet size is a mandatory script parameter used by the application to calculate the % packet loss.

1. Directional Delays

The E-Model requires both directional delay and round trip delay measurements. Using a synchronized timing mark distributed between different call generators, or different physical interfaces, measurements are made on a local connection or connections spread across a wide area. Synchronization is provided locally or globally by an internal clock or Ameritec's CRS-AMSG (available GPS timing option). The 1-way delay measurement is made by launching a 100-millisecond wide GoldenVoice packet on the timing mark, receiving it on the far end and measuring the deviation between the receipt of the packet and the timing mark (see Figure 3).

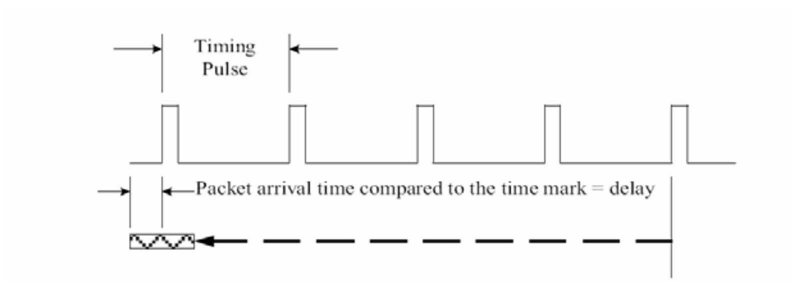


Figure 3
1-way delay measurement

Round trip delay is measured by launching a 100-millisecond GoldenVoice pulse from the terminator (listener) towards the originator (speaker). The pulse is recognized by the originator, held for a period of time and sent back to the terminator. The delay is calculated by measuring the total trip time, less the hold time at the originator, which is a constant in the script.

2. Percentage of Voice Packet Loss

Packet loss is measured by sending a fixed duration GoldenVoice pulse from the originator to the terminator and listening for a loss of signal greater than a specified packet size threshold (see Figure 4). Since the measurement is a fixed period of time, and the size of the voice packet is defined in the script, the total number of packets is counted and, using the number of lost packets measured, the percentage of voice packet loss is calculated.

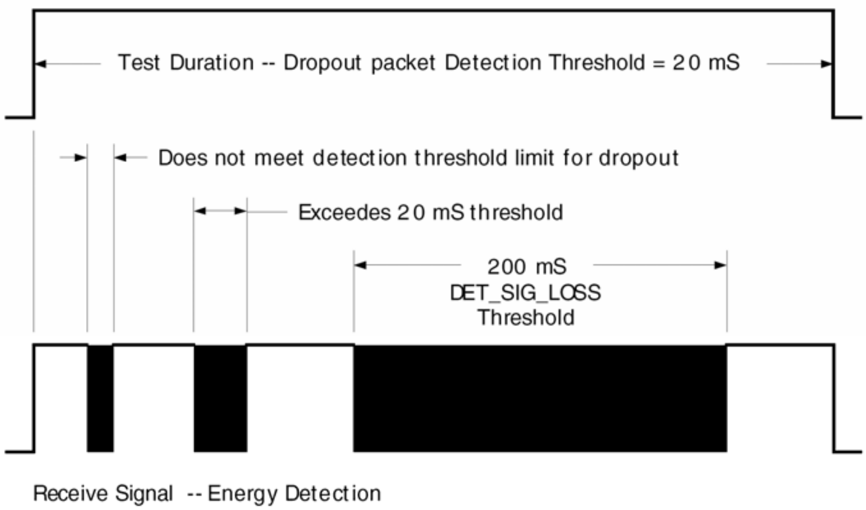


Figure 4
Packet Dropout Measurement

3. Signal-to-Noise Ratio (SNR)

The SNR value is a key measurement used to assess the performance of analog voice transmission and until recently was not considered an appropriate measurement for packet-switched networks. This opinion has recently changed as developers now realize that the SNR value is a true indicator of the listenability at the receiver's ear – which is an analog signal. Noise and distortion are injected into the circuit as the signal passes through Codecs, vocoders, transcoders, A-D and D-A converters, 2-wire to 4-wire converters and other network elements. Noise and distortion only become problematic when they exceed a given level, for instance, when they exceed the typical background noise level of –45 dBm.

The primary function of a signal to noise ratio measurement is to determine the ratio, measured in dB, of the signal energy and the non-signal energy. In a traditional circuit-switched environment the test was done with a single tone. Since gateways are designed to eliminate single frequency tones a new pair of GoldenVoice patterns, GVSNRmeasure1 (tone 29) and GVSNRmeasure2 (tone 30), were developed and are exclusively used for the SNR measurements. The GVSNRmeasure1 has a spectral range centered around 1,000 Hz and the GVSNRmeasure2 has a spectral range centered around 2,000 Hz. One pattern is used to do the forward (terminator-to-originator) measurement and the other is used for the backward (originator-to-terminator) measurement. The distinct GoldenVoice patterns used for directional testing eliminate any potential of triggering the echo-cancellers that would suppress a similar return pattern.

In an 8,000 Hz PCM network – standard analog circuit – the IEEE defines the SNR value to have a range of 0-35 dB with an optimum value of 35 dB. As the relationship between signal energy and noise energy changes, the SNR value changes. The object of the test is to determine the SNR value plus other non-signal characteristics on the channel. The channel receiving the GoldenVoice SNR signal is equipped with a specialized DSP application that can detect and evaluate not only the GoldenVoice pattern but also the characteristics of other energy present on the channel that exists within the voice band of 0 to 4,000 Hz.

4. Receive Level

The receive level is a by-product of the SNR measurements. The GoldenVoice pattern is sent from the originator at –9 dBm. The GoldenMOS script compares the receive level to the defaulted transmit level and calculates an overall loudness rating (OLR).

V. Generating GoldenMOS using Measured Impairments

The challenge before packet-switched network developers is to constantly improve the performance and quality of their products. These products are typically assessed by the subjective observations of the end user. Information that a developer uses to assess quality has little meaning to the consumer. Voice packet loss and buffer sizing mean nothing to the end user, but everything to the engineer attempting to isolate and correct various component flaws, or weaknesses in a system.

The combination of VoP and GoldenMOS attacks this issue.

Recommendation G.107 is defined as a transmission planning tool, and based on projected impairments, forecasts a MOS. While G.107 describes the basic E-Model algorithm, there are other documents that add definition and clarification to the requirements of G.107.

- ITU-T Recommendation G.108 adopted 9/1999 – Application of the E-model: a planning guide
- ITU-T Recommendation G.113 adopted 2/1996 with Appendix I adopted 2/2001 – Transmission impairments
- ITU-T Recommendation G.114 adopted 5/2000 – One-way transmission time

- ETS ETR 250 adopted 7/1996 – Speech communication quality in a 3.1 kHz communication network for handsets

A. Implementing the E-Model

The basic desire of the development community is to perform the voice quality measurement as close as possible to the network's connect point of the test target, thus eliminating the extraneous characteristics of an attached network and the speaker and the listener environment. The E-Model provides a powerful and repeatable method of assessing whether a data network is capable and ready to carry VoIP calls as well as performing voice-readiness testing.

B. Calculating the R-Factor

The E-model is a complex formula taking into account the various impairment conditions that exist in a network to compute a single score, called an "R-factor". Once an R-factor is obtained, it can be correlated to an estimated Mean Opinion Score (MOS). R-factor values range from 100 (excellent) down to 0 (poor); a MOS can range from 5 down to 1. An estimated MOS can be directly calculated from the E-model's R-factor.

In its simplest form, the R-factor formula is:

Rfactor = 0 < (Ro – Is – Id – Ie + A) < 100

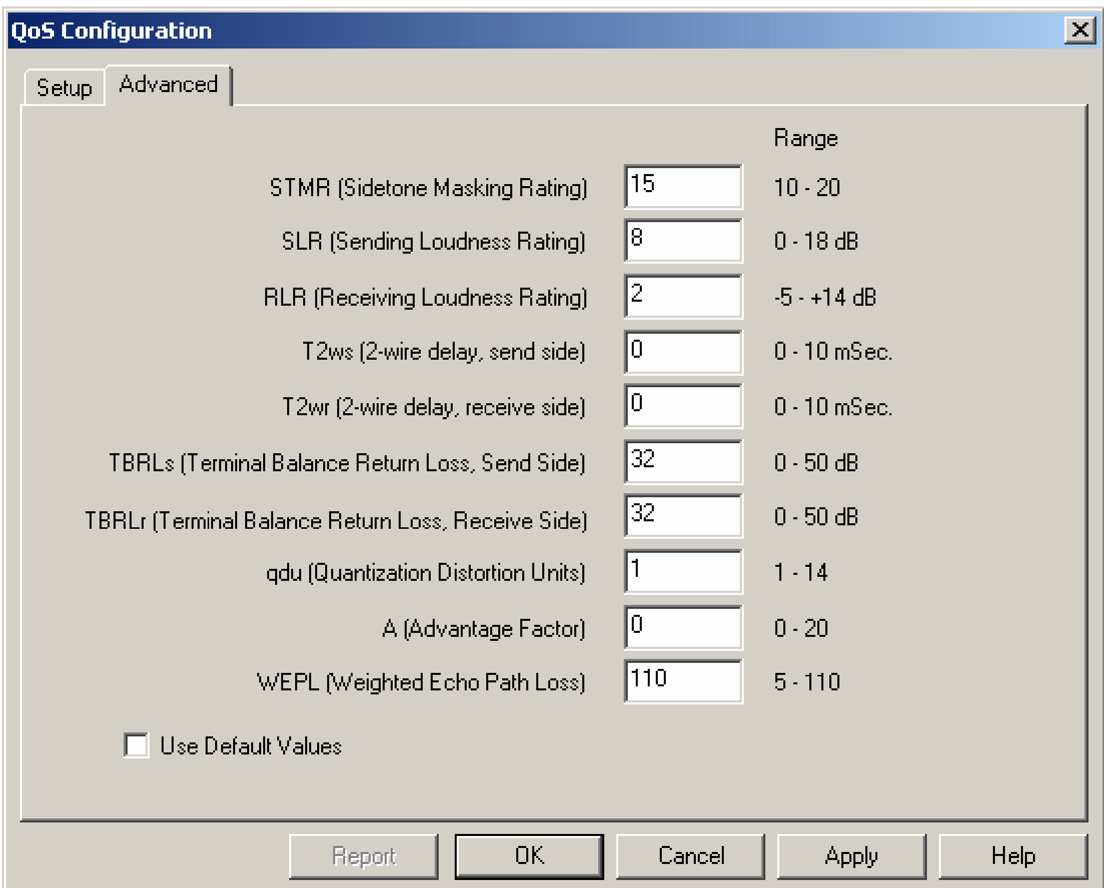
Where,

Ro = Signal to Noise Ratio
Is = Simultaneous Impairment Factor
Id = Delay Impairment Factor
Ie = Equipment Impairment Factor
A = Advantage Factor

Each of these above mentioned factors are determined by computing extensive formulas found within the G.107 document. However, shown in this manner, this equation highlights that the critical measurements needed to calculate the R-factor include signal-to-noise ratio, delays within the network, types of equipment (codecs, vocoders, etc.) used in the network and the type of network utilized such as wireless, land-line, etc. (Advantage factor).

While Ameritec's call generators are actually measuring the necessary values required to compute the R-factor value, the user is required to only enter those values into the equation that relate to the type of network they are testing. These values and their associated definitions are:

Codec type	G.711, G.729, G.723.1, etc.
Advantage Factor	Conventional wire (0), cellular network in a building (5), cellular network in a moving vehicle (10), satellite (20). Default is 0.
TBRLs	Terminal Balance Return Loss, send side. Default is 32.
TBRLr	Terminal Balance Return Loss, receive side. Default is 32.
T2ws	The 2-wire delay on the send side. Default is 0.
T2wr	The 2-wire delay on the receive side. Default is 0.
STMR	Sidetone Masking Rating. Default is 15.
qdu	Quantization Distortion units. Default is 1.



A. Codec Type

The type of equipment used in the network has a significant impact on the measured R-factor and equivalent GMOS value. These codecs/vocoding schemes for example have differing delay characteristics and tolerances for dropouts. Additionally, there are network characteristics that the user must take into account when selecting a codec type. For example, the user may select a G.711 codec or a G.711 with bursts codec that not only describe the vocoding scheme used but also describes the nature of the packet errors encountered on the circuit. Packet errors such as dropouts that occur in a burst fashion can have a more pronounced effect on voice quality than errors that are randomly distributed over time.

B. Advantage Factor

The value associated with the Advantage factor differs depending on the type of network the user is testing. This factor can best be described as a compensation value associated with the user's tolerance of poor circuit quality for convenience of placing a call. For example, a user may expect that a call placed on a wireless network will suffer from dropouts and poor voice quality but he is willing to accept that for the convenience of using a mobile phone.

C. Terminal Balance Return Loss (TBRL)

This value has both a send side and a receive side component. The TBRL can be measured using a Transmission Impairment Test Set (TIMS) and is the return loss value measured from one end of the circuit to the other end. Return loss indicates how well the input and output impedances are matched throughout the circuit. If the circuit is properly matched then very little of the transmitted energy will be returned on the received pair and the Return Loss reading (which is a measure of the transmitted energy versus received energy) will be a large reading.

D. Two-wire Delay

This value represents the delay of the 2-wire portion of the network loop. Although Ameritec's call generators used in a testing application would not have a 2-wire delay element associated with it, it may be applicable to place a value in this setting that would simulate an actual network loop with subscribers at the end.

E. Sidetone Masking Rating

This parameter addresses the optimum amount of tone that that is returned in a typical hybrid from the mouthpiece to the earpiece. As networks have evolved users may now be utilizing digital phone sets, computers, etc. to perform the functions that previously were performed with a generic 2600 series handset. These devices have to simulate or compensate for this tone, otherwise, the quality of service could be affected. For example, if the SideTone is too quiet, a user may tend to speak louder, and hence, a higher value of STMR may be required. Conversely, a SideTone that is too loud may result in a user speaking too softly and thereby preventing the party on the other end from adequately hearing the conversation. In either case, both of these conditions affect GMOS scoring.

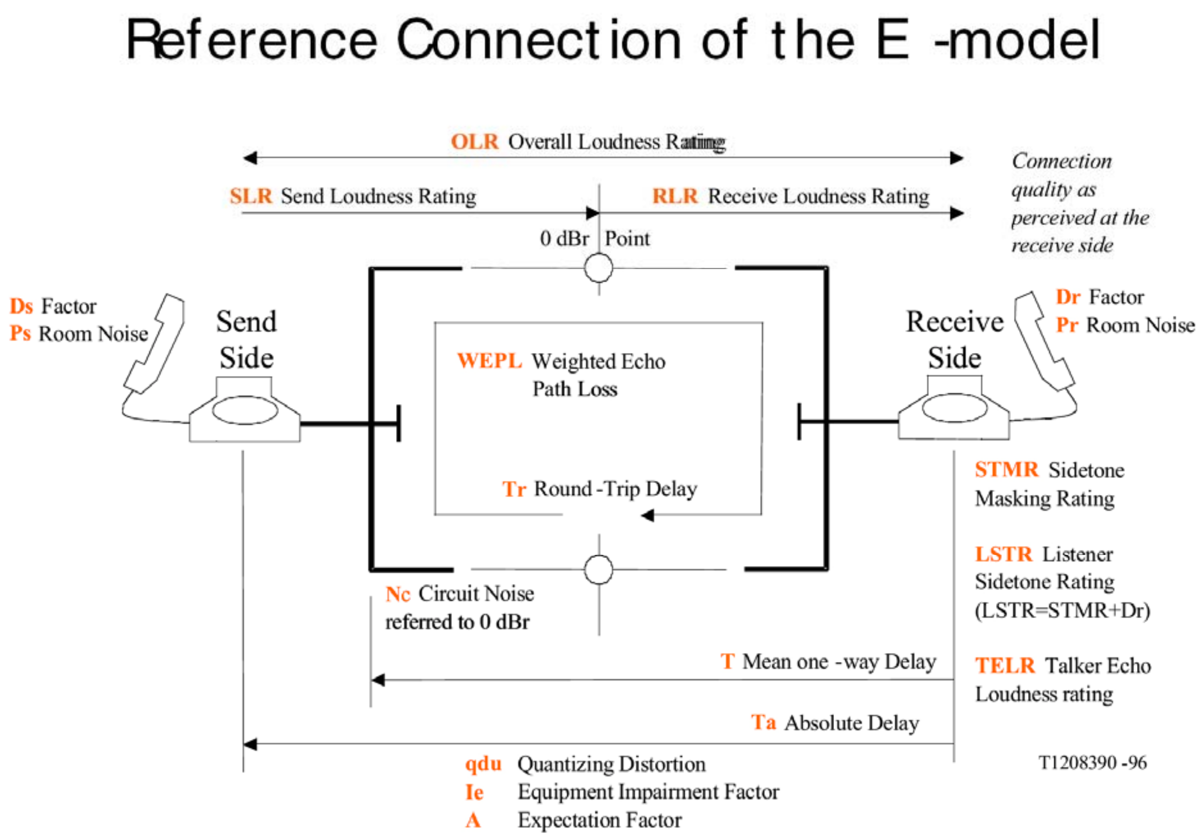


Figure 1/G.107 – Reference connection of the E - Model

F. Quantization Distortion Unit:

This parameter has a default value of 1 and represents an analog to digital conversion and digital to analog conversion within a network. Every additional D/A and A/D conversion adds an additional qdu unit to this default value. Also, should equipment elements contain digital gain or loss, additional qdu's should be added to this value as well.

G. Weighted Echo Path Loss:

Weighted Echo Path Loss (WEPL) is defined as the sum of the Echo Return Loss (ERL) from the send and receive sides of the circuit, plus the sum of the node losses. WEPL along with the delay in the circuit contributes to the delay impairment due to the listener echo. This generally has a minor affect on the overall quality.

C. Calculating a GoldenMOS value

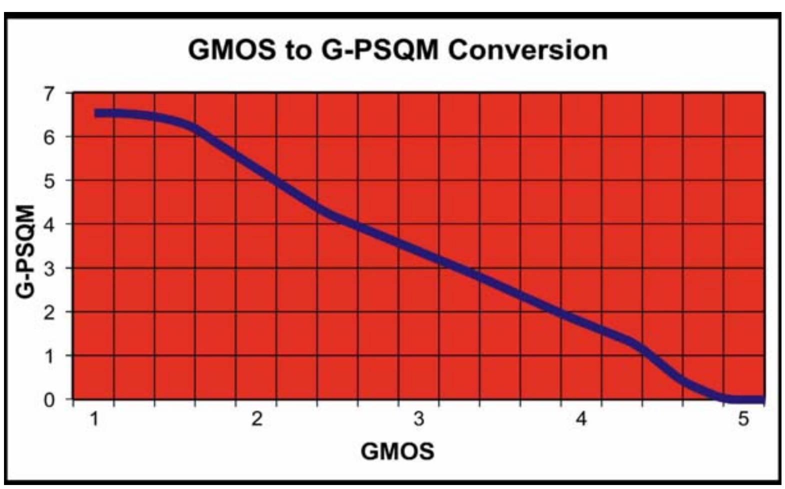
The relationship between the R-factor and the MOS value is shown in the figure below. The GoldenMOS application calculates the GMOS value automatically from the R-factor.

The MOS score is related to the R-factor value by the equation

$$MOS = 1 < 1 + (0.035*Rfactor) + (Rfactor(Rfactor - 60)*(100 - Rfactor)*7.0e\text{-}06) < 4.5$$

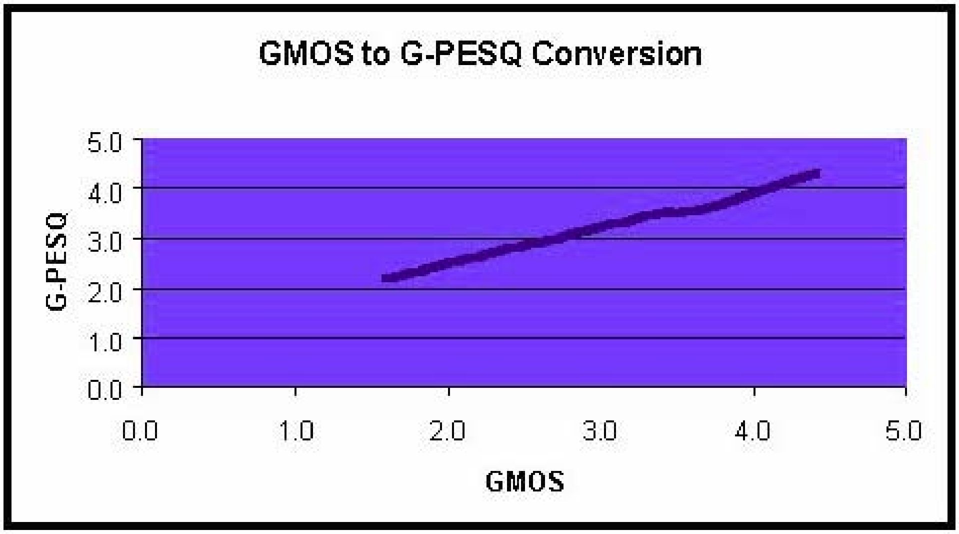
D. Calculating a G-PSQM value

Once a GMOS value has been determined by the method described above, it is then possible to correlate this value to a G-PSQM value using the figure below.



E. Calculating a G-PESQ value

Once a GMOS value has been determined by the method described above, it is also possible to correlate this value to a G-PESQ value using the figure below.



VI. Summary

What Ameritec Call Generators offer are performance, accuracy, capability and scalability. Ameritec Call Generators are designed to make precise impairment measurements, and FeatureCall ME™ and Conductor™ and GoldenMOS calculates the GMOS, G-PSQM and GPESQ values. All the functions are done simultaneously on every channel in the call generator without impacting the performance of the test equipment – a single tool that is capable of satisfying the needs of every department in your organization.

Ameritec Call Generators are available in a wide range of physical interfaces -- Analog, T1/E1 CAS, ISDN-PRI, SS7 and DS3 – and provide interworking between the different interfaces. No longer is your testing confined to one piece of equipment in one location but with a GPS clocking source testing can also be done over a wide area duplicating the characteristics that are representative of actual deployed products.

You can count on Ameritec to provide the tools, the resources and the support to make you a success. If you have questions on testing your application: askzeke@ameritec.com.

Ameritec