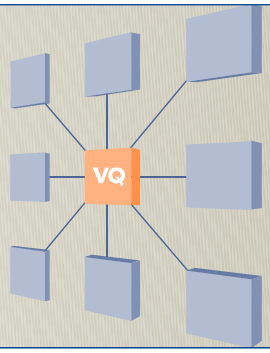
telchemy
Actively managing multimedia®



Application Note

Title **Voice Quality Measurement**

- **Series** Voice over IP Performance Management
- **Date** Nov 2014

Overview

This Application Note describes commonly-used call quality measurement methods, explains the metrics in practical terms and describes acceptable voice quality levels for Voice over IP services.

Contents

Introduction

Voice over IP systems can be affected by call quality and performance management problems. IT managers must understand basic call quality measurement techniques in order to successfully monitor and manage VoIP services and diagnose problems.

This Application Note describes commonly used call quality measurement methods, explains the metrics in practical terms, and describes acceptable voice quality levels for VoIP networks.

Definition of Call Quality

IP call quality can be affected by noise, distortion, too high or low signal volume, echo, gaps in speech, and a variety of other problems.

When measuring call quality, three basic categories are studied:

- **Listening Quality** -- Refers to how users rate the sound quality of what they hear during a call.
- **Conversational Quality** -- Refers to how users rate the overall quality of a call based on listening quality and their ability to converse during a call. This includes any echo or delay-related difficulties that may affect the conversation.
- **Transmission Quality** -- Refers to the quality of the network connection used to carry the voice signal. This is a measure of network service quality as opposed to the specific call quality.

The goal of call quality measurement is to obtain a reliable estimate of one or more of the above categories using either subjective or objective testing methods, that is, using human test subjects or computer-based measurement tools.

Listening Quality, Testing & MOS Scores

Subjective testing is the "time honored" method of measuring voice quality, but it is a costly and time consuming process. One subjective test methodology is the Absolute Category Rating (ACR) Test [1].

In an ACR Test, a pool of listeners rates the quality of a series of audio files using an opinion scale ranging from 1 to 5:

5	**Excellent**
4	**Good**
3	**Fair**
2	**Poor**
1	**Bad**

The average or Mean Opinion Score (MOS) for each audio file is then calculated from the group of individual scores. To ensure a reliable result for an ACR Test, a large pool of test subjects should be used (16 or more), and the test should be conducted under controlled conditions in a quiet environment. Generally, scores become more stable as the number of listeners increases. In order to reduce the variability in scores and to help with scaling of results, tests commonly include reference conditions using well-known impairments.

The chart below (Figure 1) shows the raw votes from an actual ACR Test that resulted in a MOS score of 2.4. The high number of votes for opinion scores "2" and "3" are consistent with the MOS score of 2.4; however, a significant number of listeners did vote scores of "1" and "4."

When analyzing the results of subjective tests, it is important to remember that the tests are truly "subjective," and that the results can vary considerably. Within the telephony industry, manufacturers often quote MOS scores associated with codecs; in reality, these scores are a value selected from a given subjective test.

Test labs typically use high quality audio recordings of phonetically balanced source text, such as the Harvard Sentences, for input to the VoIP system being tested. The Harvard Sentences are a set of English phrases chosen so that the spoken text will contain the range of sounds typically found in speech. Recordings are obtained in quiet conditions using high

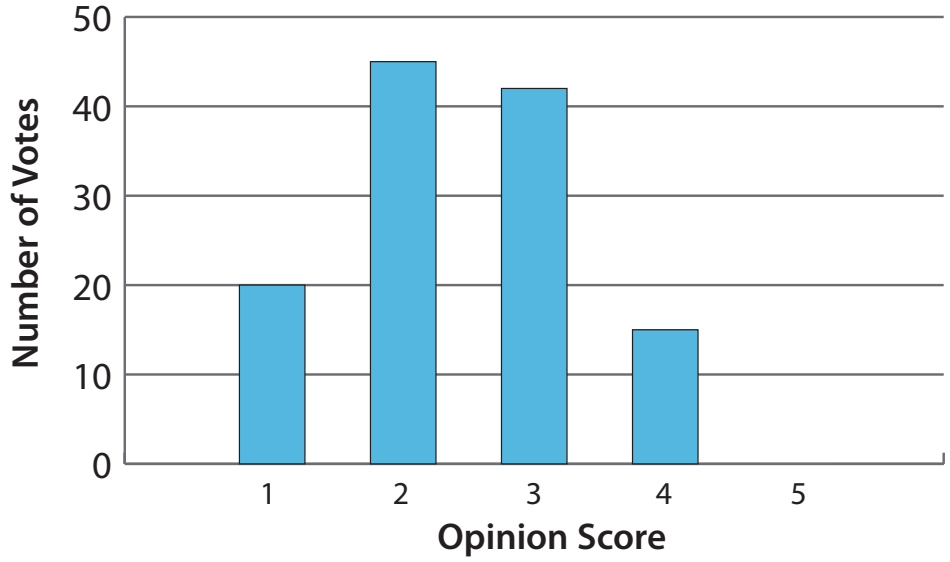


Figure 1: Chart Showing Listener Votes For an Actual ACR Test

resolution (16-bit) digital recording systems and are adjusted to standardized signal levels and spectral characteristics. The International Telecommunications Union (ITU) and the Open Speech Repository are sources of phonetically balanced speech material.

In order to differentiate between listening and conversational scores, the International Telecommunication Union (ITU) introduced the terms MOS-Listening Quality (MOS-LQ) and MOS-Conversational Quality (MOS-CQ) with the additional suffixes (S)ubjective, (O)bjective and (E)stimated [2]. Hence, a listening quality score from an ACR test is called a MOS-LQS.

In addition to ACR, other types of subjective tests include the Degradation Category Rating (DCR) and Comparison Category Rating (CCR) [1]. DCR methodology looks at the level of degradation for the impaired files and produces a DMOS score. The Comparison Category Rating (CCR) Test compares pairs of files and produces a CMOS score.

Conversational Quality Testing

Conversational quality testing is more complex, and hence, used much less frequently. In a conversational test, subjects are typically placed into interactive communication scenarios and asked to complete a task—such as booking a flight or ordering a pizza—over a telephone or VoIP system. Testers introduce effects such as delay and echo, and the test subjects are asked for their opinion on the quality of the connection.

The effect of delay on conversational quality is very task dependent. For non-interactive tasks, one-way delays of several hundred milliseconds can be tolerated; for highly interactive tasks, even short delays can introduce conversational difficulty.

The task dependency of delay introduces some question over the interpretation of conversational call quality metrics. For example, consider two identical VoIP system connections with 300 milliseconds of one-way delay. One supports a highly interactive business negotiation, while the other supports an informal chat between friends. In the first example, users may say that call quality was bad; in the second case, the users probably would not even notice the delay.

Sample-Based Objective Testing

P.862.x (PESQ) [3] and its successor P.863 (POLQA) [4] are so-called "full reference" models designed to measure objective listening quality (MOS-LQO). Full reference models compare an undistorted reference file with a test file that may be distorted through encoding or network induced artifacts such as packet drops. P.862.x and P.863 compare the input audio signals in the psychoacoustic domain, which requires the signals to be transformed using Fast Fourier Transforms, proper level scaling and temporal alignment. Because the process is fairly computationally intensive and requires access to the undistorted reference signal, full reference models are mainly employed in a laboratory environment.

VQmon® and the E Model

VQmon [5] is an advanced VoIP perceptual quality estimation algorithm that incorporates support for key international standards including ITU-T P.564, ITU-T G.107, ITU-T G.1020, ETSI TS 101 329-5 Annex E and IETF RFC 3611. VQmon is a "no-reference" algorithm that does not use the original reference signal, and therefore is able to derive call quality scores using typically less than one thousandth of the processing power needed by the P.862.x and P.863 approaches.

VQmon incorporates support for time varying IP impairments (typically caused by network congestion) and has been independently shown to provide significantly more accurate and stable metrics than other algorithms such as G.107 (E Model).

The E Model [6] was originally developed within the European Telecommunications Standardization Institute (ETSI) as a transmission planning tool for telecommunication networks, and was standardized by the ITU as Recommendation G.107 in 1998. Some extensions to the E Model that enable its use in VoIP service quality monitoring were developed by Telchemy and have been standardized in ETSI TS 101 329-5 Annex E. [7]

The objective of the E Model is to determine a transmission quality rating, i.e., the "R" factor, that incorporates the "mouth-to-ear" characteristics of a speech path. The R factor scale and typical R values vary somewhat depending on the codec type (see Table 1). The R factor is a conversational quality measurement that can be converted to estimated conversational and listening quality MOS scores (MOS-CQ and MOS-LQ).

The E Model is based on the incorrect premise that the effects of impairments are additive, i.e. linear. Non-linear approaches have been found to be better suited to describe the relationship between key impairment factors. In particular, VQmon's non-linear impairment factor combination model has been shown to improve the accuracy of estimated MOS scores when there is a high level of several dissimilar impairments (for example, packet loss and echo).

Impairments caused by network congestion tend to be highly time-varying, with high packet loss "bursts" occurring intermittently between "gaps" of low or no packet loss. VQmon outperforms the E Model by incorporating the effects of time-varying IP network impairments, providing a more accurate estimate of user opinion. In addition, VQmon incorporates extensions to support wideband and super-wideband/fullband codecs.

Unlike the E Model, VQmon was specifically developed for real-time service performance monitoring in live network environments. Since its development in early 2000, VQmon has achieved wide industry acceptance, with over 300 million agents currently deployed in a broad range of network, telecom, silicon solutions and test equipment.

Comparing Voice Quality Metrics

The chart in Figure 2 shows the relationship between the R factor generated by the E Model and MOS. The "official" mapping function provided in ITU G.107 gives a MOS score of 4.4 for an R factor of 93 (corresponding to a typical unimpaired G.711 connection, i.e., the equivalent of a regular telephone connection).

Recent ACR subjective test data suggests that a MOS score of 4.1 to 4.2 would be more appropriate for an unimpaired G.711 call. This

Table 1: R Factor Scale and Typical R Values by Codec Type

Codec Type	Audio BW Range	Sample Rate	R Factor Scale	Typical R Values
Narrowband	30 Hz - 3.4 kHz	8 kHz	0 - 100	50 - 93
Wideband	50 Hz - 7 kHz	16 kHz	0 - 129	50 - 108
Super-Wideband	50 Hz - 14 kHz	32 Khz	0 - 179*	50 - 177*
Fullband	20 Hz - 20 kHz	44.1 / 48 kHz	0 - 179*	50 - 177*

* Although fullband codecs use a wider audio bandwidth than super-wideband codecs—which can be beneficial for encoding music, noises and sounds—they generally do not provide a perceptual difference for speech signals. For this reason, the same R factor scale can be applied to both fullband and super-wideband codecs.

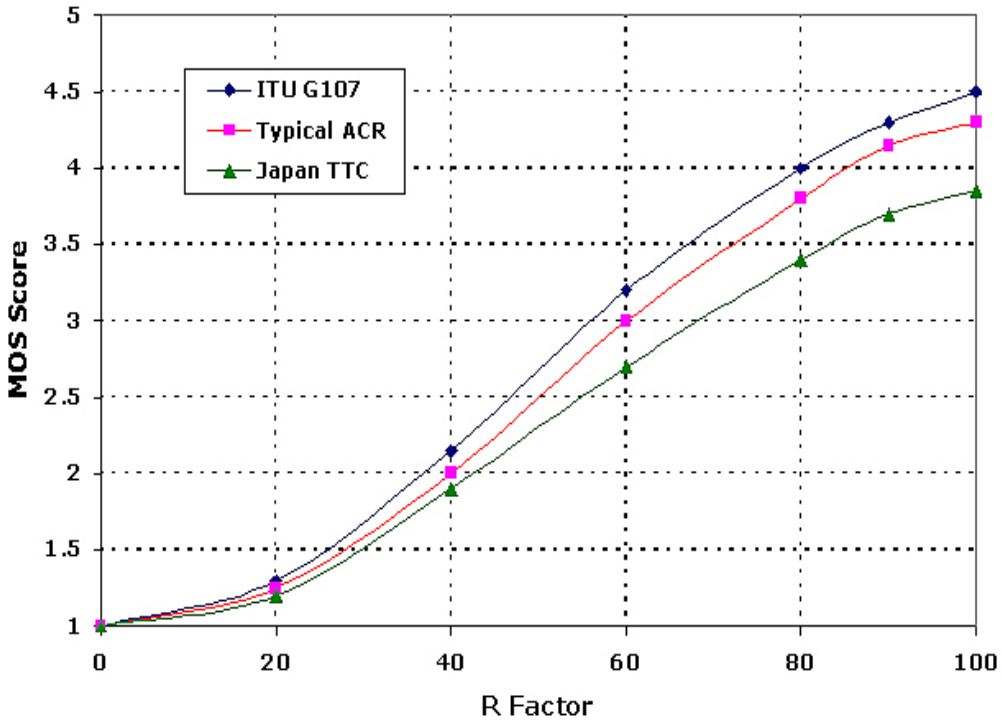


Figure 2: *Chart Showing the Relationship Between R factor and MOS Score (Narrowband Scale)*

would provide a slightly different mapping for "Typical ACR" than shown in the graph in Figure 2.

In Japan, the Telecommunication Technology Committee (TTC) developed an R factor to MOS mapping methodology that provides a closer match based on the results of subjective tests conducted in Japan. The TTC scores are traditionally lower than those in the US and Europe due in some part to cultural perceptions of quality and voice transmission.

Therefore, the chart above shows three viable mappings from R to MOS:

- ITU G.107 mapping
- ACR mapping
- Japanese TTC mapping

The use of wideband (or super-wideband/ fullband) codecs can introduce additional complications. An ACR test is on a fixed 1-5 scale, and is really a test that is relative to some reference conditions. Wideband tests use the same MOS scale as narrowband tests; therefore, a wideband codec may have (for example) a MOS score of 3.9 even though it sounds much better than a narrowband codec with a MOS of 4.1. This is not the case for R factors, which have a scale that encompasses both narrowband and wideband. Therefore a wideband codec may result in an R factor of 105 whereas a typical narrowband codec may result in an R factor of 93.

The following is an example of VQmon output for the same reference file evaluated in three different bandwidth contexts. This example used the AMR-NB codec at 12.2 kbps, with no loss impairments present.

Narrowband context:	4.16 MOS
Wideband context:	2.77 MOS
Super-Wideband context:	2.36 MOS

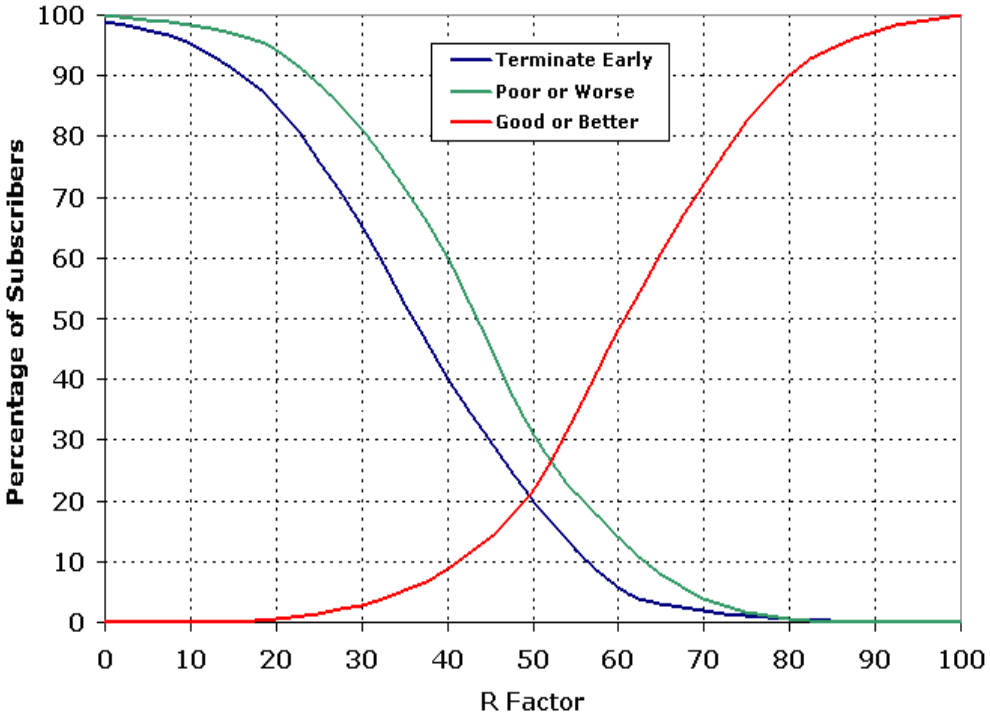


Figure 3: *Chart Showing the Relationship Between R Factor and Subscriber Opinion (Narrowband Scale)*

User perception of call quality can also be affected by variable bitrate codecs that switch between narrowband and wideband sampling rates during a call. In particular, transitions from wideband to narrowband are associated with a drop in quality. Calls with frequent switching between narrowband and wideband may be perceived by users as more annoying than calls that start as narrowband and remain consistent throughout the call.

Acceptable Voice Quality Levels

The chart in Figure 3 shows the relationship between R factor and the percentage of subscribers, i.e., users, that would typically regard the call as being Good or Better (GoB), Poor or Worse (PoW) or Terminate the Call Early (TME). For example at an R factor of 60, over 40% of subscribers would regard the call quality as "good;" Nearly 20% of subscribers would regard the call quality "poor." And, almost 10% would terminate the call early.

Table 2 on the following page shows a typical representation of call quality levels. Generally, an R factor of 80 or above (narrowband) or 100 or above (wideband) represents a good objective; however, there are some key things to note:

- Since R factors are conversational metrics, the statement that R factors should be 80 or more for narrowband or 100 or more for wideband implies both a good listening quality and low delay. Stating that (ITU scaled) MOS should be 4.0 or better is not the same as assuming that this is MOS-LQ and does not incorporate delay. Saying that R should be 80 or higher and MOS should be 4.0 or higher is not consistent. Telchemy introduced the notation R-LQ and R-CQ to deal with this; hence, an R-LQ of 80 for narrowband *would* be comparable with a MOS of 4.0.
- The typically manufacturer-quoted MOS for G.729A is 3.9, implying that G.729A could not meet the ITU scaled MOS for “Satisfied.” However, G.729A is widely used and appears to be quite acceptable. This problem is

due to the scaling of MOS and not the codec. Typical ACR scores for codecs should be compared to an ACR scaled range. For example, "Satisfied," would range 3.7 to 4.1 and hence the G.729A MOS of 3.9 would be within the Satisfied range.

Summary

Voice quality measurement plays an essential role in managing the performance of Voice over IP systems. IT managers should be familiar with the various methods available for measuring voice quality and understand the advantages and limitations of each method. In particular, objective, no reference measurement tools such as Telchemy's VQmon can provide real-time feedback on the perceptual quality of live calls in a production environment.

When specifying call quality objectives, it is important to be clear about terminology—either specify R factor (R-CQ) or MOS-CQ, or the combination of MOS-LQ and delay. If you use wideband and narrowband codecs, then be aware that you need to interpret MOS scores as "narrowband MOS" or "wideband MOS" in order to avoid confusion.

Acronyms

ACR	Absolute Category Rating (Test)
CCR	Comparison Category Rating (Test)
CMOS	Comparision Mean Opinion Score
DCR	Degradation Category Rating (Test)
DMOS	Degradation Mean Opinion Score
ETSI	European Telecommunications Standardization Institute
GoB	Good or Better (Score)
IETF	Internet Engineering Task Force
IP	Internet Protocol
ITU	International Telecommunications Union
MOS	Mean Opinion Score
MOS-CQ	Mean Opinion Score - Conversational Quality
MOS-LQ	Mean Opinion Score - Listening Quality
PESQ	Perceptual Evaluation of Speech Quality
POLQA	Perceptual Objective Listening Quality Assessment
PoW	Poor or Worse (Score)
R Factor	Transmission Quality Rating
TME	Terminate the Call Early (Score)
TTC	Telecommunications Technology Committee
VoIP	Voice over Internet Protocol

Table 2: *Typical Representation of Call Quality Levels*

User Opinion	R Factor (Narrowband)	R Factor (Wideband)	MOS (ITU Scaled)	MOS (ACR Scaled)
Very Satisfied	90 - 100	115 - 129	4.3 - 5.0	4.1 - 5.0
Satisfied	80 - 90	100 - 115	4.0 - 4.3	3.7 - 4.1
Some Users Satisfied	70 - 80	90 - 100	3.6 - 4.0	3.4 - 3.7
Many Users Dissatisfied	60 - 70	80 - 90	3.1 - 3.6	2.9 - 3.4
Nearly All Users Dissatisfied	50 - 60	65 - 80	2.6 - 3.1	2.4 - 2.9
Not Recommended	0 - 50	0 - 65	1.0 - 2.6	1.0 - 2.4

Frequently Asked Questions (FAQ)

What is the objective of call quality measurement?

The goal of call quality measurement for Voice over IP is to assess the listening quality, conversational quality or network transmission quality (or a combination) of the service using either subjective or objective testing methods.

What's the difference between subjective and objective testing for measuring call quality?

Subjective testing uses human subjects to listen to audio samples and rate their quality on an opinion scale. Examples of subjective test methodologies include ACR, DCR and CCR tests.

Objective testing takes objective measurements from audio samples and applies an algorithm to the data to obtain an estimate of voice quality as perceived by users. Examples of objective test methodologies include ITU-T P.862.x (PESQ), P.863 (POLQA), G.107 (The E Model), and Telchemy's VQmon.

What is a Mean Opinion Score (MOS)?

In subjective testing, a Mean Opinion Score is the arithmetic mean or average of all of the individual opinion scores resulting from a single test. In IP telephony, MOS is commonly used to measure the listening and conversational quality of a VoIP call on a scale from 1 to 5, with 5 being best.

Objective test methods such as VQmon, PESQ and POLQA produce objective MOS scores that are designed to reflect, as accurately as possible, subjective speech quality as perceived by human subjects.

What are "full reference" and "no reference" testing models?

A "full reference" test algorithm compares an original reference signal to the impaired signal and analyzes the difference between the two. Full reference tests are generally used in dedicated testing environments and are not suitable for real-time performance monitoring in VoIP networks. PESQ and POLQA are examples of full reference test models for measuring voice quality.

"No reference" tests analyze only the impaired signal and do not need the original reference signal. Examples of no reference test models include VQmon and the E Model.

What is the E Model?

The E Model (ITU-T G.107) is a transmission planning tool that predicts voice quality as it would be perceived by a typical telephone user by calculating the impact of various types of impairments—including noise, echo, delay and packet loss—on the quality of a call. The E Model equation produces a numeric "R factor" value that can be mapped to a MOS value.

What is VQmon?

VQmon is a perceptual quality estimation algorithm developed by Telchemy, which is widely used to monitor and measure the quality of Voice over IP. VQmon incorporates support for key international standards including ITU-T P.564, ITU-T G.107, ITU-T G.1020, ETSI TS 101 329-5 Annex E and IETF RFC 3611.

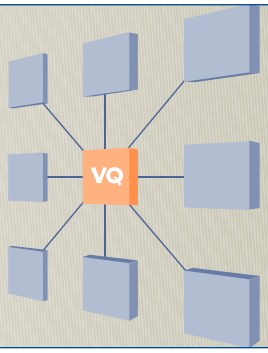
VQmon calculates the impact of time-varying IP impairments and has been independently shown to provide significantly more accurate and stable metrics than other algorithms such as the E Model. When measuring call quality, VQmon produces both listening and conversational quality MOS scores (MOS-LQ, MOS-CQ) and R factors (R-LQ, R-CQ).

Why is VQmon used for real-time performance monitoring?

VQmon was specifically developed as a non-intrusive tool for monitoring the quality of VoIP calls in real time. VQmon code is compact and resource-efficient, typically requiring less than one thousandth of the processing power needed for P.862.x and P.863 calculations. VQmon supports a wide set of industry standard and proprietary codecs, and can be embedded into a broad range of platforms including IP phones and gateways, mobile handsets, routers and switches, soft clients and test equipment.

What impact does the codec type (narrowband vs. wideband) have on call quality measurement?

Subjective tests (such as ACR) use the same opinion scale for narrowband and wideband codecs. Therefore, a test using a narrowband codec might receive a higher MOS score than a test using a wideband codec, even though the quality of the wideband sample is actually higher. To avoid confusion when testing both narrowband and wideband codecs, the MOS context should be specified (for example, "Narrowband MOS" or "Wideband MOS") or the quality level should be expressed as an R factor, which uses a wider scale for wideband codecs.



References

[1] ITU-T Recommendation P.800: Methods for subjective determination of transmission quality, August 1996

[2] http://www.cs.columbia.edu/~hgs/audio/harvard.html

[3] ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology, July 2006

[4] ITU-T Recommendation P.862 Perceptual Estimation of Speech Quality (PESQ)

[5] ITU-T Recommendation P.863 Perceptual Objective Listening Quality Assessment (POLQA)

[6] ITU-T SG12 Contribution D105, VQmon Description, January 2003

[7] ITU-T, "Conformance testing for voice over IP transmission quality assessment models," Recommendation P.564

[8] ITU-T G.107 The E Model: A computational model for use in planning

[9] ITU-T, "Performance parameter definitions for quality of speech and other voiceband applications utilizing IP networks," Recommendation G.1020

[10] ETSI TS 101 329-5 QoS Measurement for VoIP

[11] ITU-T RFC 3611: RTP Control Protocol Extended Reports (RTCP XR), November 2003

[12] ITU-T RFC 6035: Session Initiation Protocol Event Package for Voice Quality Reporting, November 2010

About Telchemy, Incorporated

Telchemy® is the global leader in Voice and Video over IP performance management technology with its VQmon®, DVQattest®, SQprobe® and SQmediator® families of service quality monitoring and analysis products. Telchemy has led the use of embedded software probe technology and the application of big data and analytics for VoIP performance management, and is positioned to be a leading provider of voice and video performance monitoring technology for the emerging SDN. Founded in 1999, the company has products deployed worldwide and markets its technology through over 140 leading networking, test and management product companies. Visit **www.telchemy.com**.

Telchemy Application Notes

Series **Voice over IP Performance Management**

Title	Voice Quality Measurement
Title	Impact of Delay on VoIP Services
Title	Impact of Echo on VoIP Services
Title	Data and Fax Modem Performance on VoIP Services
Title	Voice Quality Estimation in Wireless & TDM Environments
Title	Managing Enterprise IP Telephony
Title	Managing IP Centrex & Hosted PBX Services
Title	Managing Cable Telephony Services
Title	Managing Wireless LANs and Wi-Fi Services
Title	Six Steps to Getting Your Network Ready for Voice over IP

Download application notes from:
www.telchemy.com/application

Information

Email
sales@telchemy.com
info@telchemy.com
Phone
+1-866-TELCHEMY
+1-678-387-3000
Fax
+1-678-387-3008

Telchemy Incorporated
2905 Premiere Parkway
Suite 280
Duluth, GA 30097
USA

03.Nov.2014