psytechnics
behind every word

Using PESQ to Test a VoIP Network

Application Note

Prepared by:
Psytechnics Limited

23 Museum Street
Ipswich, Suffolk
United Kingdom
IP1 1HN

t: +44 (0) 1473 261 800
f: +44 (0) 1473 261 880
e: info@psytechnics.com

Issue: 1.0

Date: January 2004

Table of Contents

Introduction

Voice over Internet Protocol (VoIP) networks are now here and proving to deliver acceptable service quality to a variety of customers. In order to roll out these solutions, good engineering discipline is required especially when rolling these networks into buildings with existing infrastructure. To ensure that these installations run smoothly a well-defined test and validation procedure is required. The basis of this procedure should be the ITU-T standard P.862 or PESQ™.

PESQ was developed and extensively tested by the world's experts in objective voice quality measurement to provide an engineering tool that can precisely measure the quality of a voice connection in terms that accurately reflect user perception. PESQ is used by all of the leading VoIP equipment vendors to validate designs prior to and after production and is the right choice to test an installed VoIP network.

This application note describes a methodology of how to test a VoIP connection using PESQ. It assumes that the reader has a basic understanding of PESQ, of voice quality issues and an understanding of VoIP. The application note ends with a summary of how to interpret the PESQ results.

Methodology

Connecting to a network

PESQ provides an end-to-end measure of voice quality by injecting a test signal at one end of a test connection and capturing a degraded version of that signal for comparative analysis at the far end of the test connection.

In most cases testing looks to establish the voice quality performance between components of a VoIP network and between the VoIP network and the PSTN. For this there are three basic types of network connection:

- Acoustic
- Analog electrical
- Digital electrical

From a practical field-testing perspective the later two are preferred over the first, as connecting acoustically requires a sound isolated environment. As such, this Application Note will concentrate on electrical connections.

Both the analog and digital connections can be divided into a further two types of connection. These are summarized in the table below along a main advantage and disadvantage of each technique.

Type	Sub-type	Description	Advantage	Disadvantage
Analog	Handset	Connects to handset port of analog or digital phones in place of the handset	Allows a consistent connection approach to all types of network, whether VoIP or PSTN	This is an undefined reference point in a phone and so level determination can be difficult
Analog	2-wire	Connects at the 2-wire line at the point where an analog phone would normally be located	Easiest way to connect to the PSTN and as such is offered by many test equipment products	Includes analog local loop which can itself be a source of atypical performance degradation
Digital	ISDN/PCM	Connects to an ISDN BR/PRI port or possibly direct into PCM network at E1/T1 or higher rate (ATM)	A perfect reference point into a network since no degradation prior to elements under test occur	Only suitable for PSTN connections and suffers from possible signaling incompatibility
Digital	IP	Connects to IP network as a VoIP media end-point	Provides a convenient test-point in a VoIP network for providing ad-hoc testing	Should only be used to test an IP to PSTN connection, as it is unable to emulate an IP phone so cannot measure PSTN to IP connections

When testing a VoIP connection a digital connection on the PSTN side and a handset connection onto an IP-phone are favored. This provides the most accurate view of a VoIP network's performance by including all the important components and excluding unnecessary analog degradations.

A Digital IP tester provides an invaluable test-point, especially for testing from the IP network to a PSTN network, but as noted above, for testing the reverse (PSTN to IP) direction it is unable to accurately validate actual end-to-end performance.

Choosing a Test Signal

PESQ testing requires a speech or speech-like test signal. Real speech recordings are easy to generate, although care over noise levels and clarity in their production is required, while speech-like test signals are more complex to manufacture but offer a more universal basis by which to compare results.

Psytechnics have an established artificial speech-like test signal (ASTS) for such testing; the ASTS signal has been used for voice quality testing in the industry since 1996. Whilst in BT, the founders of Psytechnics with the help of linguistics expert Professor John Local developed the ASTS signal. This signal takes a 4-

hour conversational corpus to determine the frequency of phonemes in everyday telephone conversations. The signal combines the most popular phoneme from each of the important signal processing groups with all linguistically legal combinations of the other popular phonemes to create a set of speech utterances that represent a full spread of speech sounds in a short duration. This signal has proven to be ideal for network testing.

A telephone system transmits speech between approximately 300Hz and 3.4 kHz. When testing a telephone system there is a requirement to make sure that the test signal does not contain significant signal components outside this range. With most connection types described above this requires that the test signal is pre-filtered. With acoustic testing there is no need to use a pre-filtered test signal while for handset testing this is not so clear as the filter characteristic may be distributed between the handset and the phone. However, since additional filtering will not generally affect the voice quality measurement, whereas a lack of it produces degradation, it is easiest if a pre-filtered signal is used for all but acoustic testing.

Earlier in this section possible issues with using recorded natural speech were mentioned. In addition to making sure there is a reasonably low noise floor (at least -75 dBov[1]) it is important to check the active speech level and the transient profile of the recording. The active speech level should be around -26dBov as this provides a good head room to avoid amplitude clipping as well as ample dynamic range to avoid quantization distortions in any test results. Transient profile describes how the speech recording varies with time, specifically the durations of continuous speech (a short sentence or utterance) and the amplitude of the utterance from beginning to end. For naturalness it is recommended that utterances should be between 1 and 3 seconds. As for amplitude profile, the talker should be at a consistent volume throughout the whole of an utterance. People have a natural tendency to start out reading loudly but tend to finish quietly when speaking into a microphone.

Knowing How Long to Test

Measurement stability is affected by test signal length, any signal containing less than 5 seconds of active speech is not recommended. The recommended duration for testing is 10 seconds.

If ASTS is to be used the following sequence of utterances provide a test signal of just over 10 seconds.

ASTS Utterance	Accumulated duration (s)
Joed	1.41
This	2.42
Oatlet	4.26
Woaner	5.01
Nullow	5.69
Illindge	6.92
Lowlant	8.70
Else	10.19

Table 1: ASTS sequence definition

If testing for longer durations is required it is recommended that the test signal be split into 10-15 second segments and processed separately through PESQ. These segment results can then be analyzed to find the average, min and max quality for that period. It should be noted that the splits should be made in silence intervals and not in the middle of a speech event.

Determining Levels

An important consideration for voice quality testing is the input level of the test signal. Having the signal too loud or too quiet will lead to MOS results that are unrepresentative of what a user hears. Generally when testing it is best to test at an “average” network level of –20dBm0 (as referenced to the four-wire point of the network). If testing to see how a network performs for different talker levels it is important to use the “average” network level and then adjust from this point.

[1] Assuming 16-bit linear PCM original recordings

When connected digitally into a network the setting of level is simple since –20dBm in the PCM stream is equivalent to –20dBm0 at the 4-wire reference point. For analog testing this is more complex since from a level perspective the 2-wire and handset connection points are not defined reference points. With handset connections there are two methods by which to determine the test level, the first is the most accurate and practical, and the second although often used, is not recommended.

Method 1: If the Send Loudness Rating (SLR) of a phone is known then the network level can be determined. If the level is not known, assume it to be around 8 dB which is the international recommendation for phones. For an "average" speech level of 89dB SPL input into a phone with an SLR of 7dB a network level of -20dBm0 is expected. By monitoring the RTP stream on the IP network and extracting the level out of the RTP voice stream the level of the input test signal can be adjusted until it reaches the expected level.

Method 2: Injecting a test signal which is too loud or too quiet will generally result in a drop in voice quality and hence a drop in PESQ scores. By adjusting the input test signals to achieve maximum voice quality it would suggest that the optimum level is found. Although this may be true, the optimum level may actually relate to someone shouting down a phone if a phone input is very quiet. This effect is inappropriate.

Similar approaches can be taken for the analog 2-wire connections; the most practical approach is equivalent to method 1, where test levels are adjusted until the network-measured level is around –20dBm0. Although Method 2 is not recommended it should be noted that this is a very useful diagnostics technique to understand if a drop in voice quality on a network is due to a level issue.

Base-lining Equipment

Before running a series of tests it is important to validate that a test set-up is working correctly. This is most easily achieved by performing a voice quality test over a known connection and checking that the MOS score returned is as expected.

The simplest example of this would be a PSTN analog or PSTN digital connection back to another PSTN analog or digital connection. The call will be carried by G.711 and so you should expect to see a PESQ LQ score of at least 4.2. If scores are returned below this then check the following:

Check	Solution
The noise floor in the reference is above 75dBov	Seek a new reference signal with a lower noise floor
The noise floor in the degraded is high	If using an analog connection this could be either the noise floor of the tester or the telephone connection itself. Try connecting to another analog line. If using a digital connection this could be a problem with an incorrect codec being selected
The reference has frequency content below 200 Hz	Select a pre-filtered reference signal or if not available apply a narrowband IRS (ITU P.48 Smj) pre-filter to the reference signal
The reference signal has less than 5 seconds of active speech	Use a reference signal that is longer or has a higher speech content
The degraded signal contains less speech than the reference	The test system has a synchronization issue or the network you are testing has a long delay. If you have missed the end of the reference signal maybe try increasing the record time
The active speech level of the degraded file is below 60 dBov	Check level at which the signal is played into the network as well as checking the amplification settings on the testers record capability

Sequencing Tests

Having established a PESQ VoIP test capability it is important to decide what connection scenarios to test and when to perform these tests. Testing should be carried out that reflects usage patterns and should as far as possible include tests during periods of high traffic – such as a busy hour.

Consider scenarios that include:

- IP-phone to IP-phone
- IP-phone to PSTN connection
- IP-phone to legacy PBX
- Legacy PBX to PSTN (benchmark)

A VoIP network is dynamic in nature and so a single test should not be used as a basis to draw conclusions. Instead at least five PESQ tests should be performed and an average and minimum of these measurements presented. In addition multiple calls should be made for each scenario to demonstrate call stability.

The number of tests per scenario very much depends on what the information will be used for. If providing installation validation then 20 bi-directional measurements over 5 calls will be ample. However, if performing a Service Level Agreement (SLA) test then Psytechnics recommended that at least 100 tests per scenario be performed. To provide a balance between having a representative number of calls and keeping overall testing of a scenario within a manageable time (a call set-up reduces the time available to perform voice quality measures) the testing should be split into 20 bi-directional PESQ tests over 5 phone calls.

To understand how each scenario performs during busy hours it would be nice to test all scenarios throughout the day, however this is time consuming. A set of 20 bi-directional PESQ measurements is likely to take around 20 minutes. To perform 100 tests is therefore going to take around 1 hour 40 minutes. In a working day it is therefore possible to test between 3 and 4 scenarios.

As each scenario consists of 5 calls with the recommended 4 measurements in each call, it is possible to schedule calls to alternate between scenarios so that each scenario has one call during a busy hour.

Note: The 20 minutes for 20 bi-directional PESQ measurements may sound a little generous but scheduling this time per call allows for equipment set-up and teardown as well as coping with unexpected situations.

Interpreting Results

Test results may be given as PESQ scores or PESQ LQ values; PESQ LQ is an adjustment to the PESQ ITU-T P.862 scale to translate PESQ results to standard Listening Quality MOS scores. The PESQ LQ scale ranges from 1 to 5, with 1 representing bad quality and 5 excellent quality. It should be noted however that PESQ will not produce values greater than 4.5 as this is the maximum score expected from a well-designed subjective test. This application note recommends and assumes that PESQ LQ scores are being used.

As a guideline on a newly installed G.729 based VoIP system you should expect to see PESQ LQ scores around 3.8 and for a G.711 system PESQ LQ scores around 4.2. If your scores are more than 0.1 below these numbers you may have problems on your network. In addition, a G.729 system would be unlikely to achieve a score of above 4.0 although a G.711 system may return scores at the maximum of 4.5.

In terms of what is acceptable for communications purposes a number of Enterprises have specified an acceptable PESQ LQ performance threshold of 3.7. If the quality drops slightly below this figure the connection is still usable but it is likely that the user will start to become aware of network degradations.

When analyzing results it is useful to decide on a pass threshold and then calculate the following simple statistics for each scenario tested:

- Mean PESQ LQ
- Min PESQ LQ
- Percentage above threshold
- Number of measurements

From these a table of test scenarios can be drawn up and problems quickly identified. It is also worth keeping at this stage the two directions as separate entries in order to spot network asymmetries.

The table below contains a results summary for three connection scenarios taken from different voice quality audits performed by Psytechnics.

Label	Mean PESQ LQ	Min PESQ LQ	% above 3.7	# measurements
1 – PSTN to IP	3.20	2.71	15	100
1 – IP to PSTN	3.93	3.78	100	100
2 – IP to IP	3.84	3.75	100	100
2 – IP to IP	3.84	3.78	100	100
3 – PSTN to IP (backup)	3.75	3.46	92	100
3 – IP to PSTN (backup)	3.76	3.42	92	100

All tests are from G.729 systems. The first two lines represent a test performed on a connection from a VoIP network out to a PSTN connection where the precedence settings where incorrectly set on the media gateway. The next two lines were taken from a test of two IP phones in different buildings while the final set of two lines show a test of a branch connected to an IP-Centrex system via an ISDN back-up connection.

The table clearly shows that there is no problem with the second scenario as all scores are above the threshold. The first test shows consistent asymmetry that suggested a configuration issue was more likely than a congestion problem. The third showed generally a lower MOS score with the average around 3.75 and occasional drops below the 3.7 thresholds down to around 3.5, this occurred to both directions and is more consistent with congestion.

In addition to this type of analysis it is also useful to perform a time of day review. In order to do this place all test data on a graph in chronological order and look for general trends, especially around the expected busy hour times of 10am and 1pm. If an area of time shows a consistent drop in quality, which is not attributable to a fault, it could indicate a capacity issue.

Glossary

ASTS	Artificial Speech-like Test Stimulus
IP	Internet Protocol
ITU	International Telecommunications Union
MOS	Mean Opinion Score
PESQ	Perceptual Evaluation Speech Quality
RTP	Real Time Protocol
SLR	Send Loudness Rating
SNR	Signal to Noise Ratio
SPL	Sound Pressure Level
THD	Total Harmonic Distortion
VoIP	Voice over Internet Protocol

psytechnics
behind every word

PRD_PESQ00_XXX_ES0111_01