XPI Solutions for Multi-Service/Access

Communications Test Equipment

White Paper

Addressing VoIP Speech Quality with Non-intrusive Measurements

by John Anderson
Product Marketing Manager
Agilent Technologies
Network Systems Test Division

Agilent Technologies

Contents

Introduction 3

Traditional Techniques for Non-intrusive Measurement of Speech Quality 6

New Techniques for Non-intrusive Measurement of Speech Quality 8

Summary 10

About the Author 11

Introduction

Integrating commercial or enterprise voice services on a data network can offer significant capital and operational cost reductions, and can enable new competitive services. However, if these business values are to be realized, customers must be won and retained by meeting their expectations for service quality.

To win and retain customers, voice service quality must meet customer expectations set by traditional circuit-switched networks. This can prove very difficult as time-sensitive VoIP packets require a minimum quality of service (QoS) on the data network to meet service quality objectives. VoIP packet loss, jitter, and delay must be tightly controlled by complex network mechanisms. In addition, other network resources such as vocoders, voice activity detectors, jitter buffers, echo cancellers, and various signal processors must be carefully selected and configured to minimize voice quality impairments.

In order to address these challenges, it is essential for a network operator to have visibility into the voice service quality being delivered to customers.

What is Speech Quality?

There are many aspects to voice service quality, but by far the most visible and important is speech quality. Speech quality, also known as clarity, refers to the clearness of a speaker's voice as perceived by a listener. It is the result of the fidelity of speech signal reproduction in a VoIP network. Speech quality is a one-way phenomenon; that is, it is experienced in one direction: from speaker to listener; and it affects listening quality.

Other aspects of voice service quality include delay and speaker echo, which are round-trip phenomenon. Together with speech quality, all of these aspects affect conversational quality.

The International Telecommunications Union (ITU) Recommendation P.800 provides methods for conducting listening and conversational tests, and for producing listening-opinion and conversational-opinion scores. These are more commonly known as Mean Opinion Scores (MOS). MOS scores are often used to quantify the listening and conversational quality of a telephone call.

On a VoIP network, speech quality is impaired by many different systems and network conditions. These include:

- Encoding and decoding of voice. This includes waveform encoding, such as G.711 Pulse Code Modulation (PCM) and G.726 Adaptive Differential Pulse Code Modulation (ADPCM), and low bit-rate voice compression using vocoders.
- Front-end clipping, as introduced by Voice Activity Detectors (VAD) that are used for silence suppression.
- Temporal signal loss and dropouts, as introduced by packet or frame loss.
- Delay variance, also known as jitter.
- Signal attenuation and gain/attenuation variances.
- Signal level clipping.
- Echo cancellation under doubletalk conditions.
- Signal noise introduced by equipment or interference on an analog access network.

Testing Beyond IP Network Performance

By the time a network operator finally determines that service quality has deteriorated, frustrated customers may have already "hung up" on both the call and the service. Network Operators must have an effective means for testing voice service quality if they are to attract and retain customers. This means testing beyond IP network performance measurements. End user speech quality measurements are needed to provide network operators with the level of visibility needed to effectively manage service quality,

It is well known that the IP network performance parameters that impact speech quality are packet loss, delay, and jitter. But the type and degree of impact that these parameters have on speech quality are lesser known. This is because there are many other VoIP processes that impact speech, and these various processes, together with IP network performance conditions, influence each other in complex ways to affect speech quality. It is very easy for the complex interactions of these processes and conditions to get out of control and not only degrade speech quality, but do so in a way that is difficult to diagnose and analyze.

Active Measurement of Speech Quality

Speech quality on traditional circuit-switched networks was never a major issue, so measurement techniques were limited. Electrical signal measurements, such as signal-to-noise ratio or total harmonic distortion, were generically applied to measuring speech quality as a strictly electrical signal.

Today, with the deployment of next generation technologies like VoIP, there are many new techniques for measuring speech quality more accurately, and from the perception of human end-users. Most of these techniques are active measurements. Active measurement techniques comprise injecting a signal as input into a system, and analyzing the output to determine the performance or behavior of the system.

MOS

The most obvious way to measure speech quality is to go right to the source and use human subjects to rate the quality of telephone calls. This "subjective" method is described in two ITU recommendations: P.800 and P.830. The primary results of subjective testing are Mean Opinion Scores (MOS) which rate the quality of a telephone call from an end user perspective. Subjective testing includes both listening-opinion and conversational-opinion tests, and offers different rating methods, including Absolute Category Rating (ACR), Degradation Category Rating (DCR), and Comparison Category Rating (CCR).

Speech quality is most effectively tested with Listening-opinion tests using the ACR method. The most frequently used MOS scale is the Listening Quality scale defined as follows:

Quality of speech	*Score*
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

It is MOS scores on this scale that objective speech quality measurement techniques attempt to reproduce.

PSQM, PAMS, PESQ

Subjective testing is obviously limited in its application. It is very expensive and impractical to use in most network testing environments. So a variety of objective measurements have been developed based on mathematics that can be computerized and automated.

The first of these to be widely accepted was the Perceptual Speech Quality Measurement (PSQM), which was adopted by the ITU in 1996 as recommendation P.861. Two other widely used techniques soon followed; the Perceptual Analysis Measurement System (PAMS) and the Perceptual Evaluation of Speech Quality (PESQ™), the latter of which was adopted by the ITU in 2001 as recommendation P.862, replacing P.861.

These techniques are based on psycho-acoustic science, and use a common approach in which a sample of voice is input into a network, and the subsequent output is recorded. The output sample is then compared to the input sample to produce a score that represents how well (or poorly) the network reproduced at the output the original speech. The two key features of these techniques are that the input and output signal are both modeled in a "perceptual" domain first, and then the comparison determines audio-perceptual distances or disturbances as a human would perceive them. The objective of each technique is to produce scores, like MOS, that reliably predict the results of subjective tests.

VoIP Packet Emulation

Techniques like PSQM, PAMS, and PESQ are very accurate, but tend to be costly in terms of implementation because they require the establishment of a telephony channel across the network-under-test. This often means specialized equipment and software must be used to support telephony interfaces and signaling.

Another class of active measurement techniques has been developed to reduce this cost but retain the objective of predicting the results of subjective tests. These techniques, known as VoIP packet emulation, use software agents to transmit VoIP packets from an IP interface, receive those packets on a distant end, and measure the fundamental IP network performance parameters such as packet loss and jitter. From these network performance measurements, a speech quality score is predicted. The concept is that a VoIP network will deliver a certain level of speech quality for a limited set of network performance conditions (i.e., packet loss, jitter, delay). In some instances, a certain codec is taken into account, and the previously known impact that the measured network performance conditions have on that codec is considered when predicting speech quality.

While this technique can be implemented with relatively low cost, it is limited in its accuracy and therefore its application. Unlike PSQM, PAMS, and PESQ, VoIP packet emulation cannot measure from an end-user perspective. It omits the impacts that all other processes and conditions in a network have on speech quality. These include jitter buffers, VADs, packet loss concealment, echo cancellers, transcoding, and any proprietary signal processing that occurs in a network.

Non-intrusive Measurement of Speech Quality

Active measurements are valuable in that they actually measure the quality of speech from an end-user perspective, and therefore are highly accurate. However, the equipment needed to perform such measurements tends to be expensive. It can also be costly to use because it requires careful control to place telephone calls and generate test signals, and it uses network resources by generating traffic.

Non-intrusive measurements comprise passive techniques for performing measurements on traffic generated by actual users. These measurements can be applied in the packet domain (to measure packet loss and jitter, for example) as well as the circuit domain (to measure signal and noise levels, for example).

Non-intrusive testing is simple and relatively inexpensive because it can be software-based and does not utilize network bandwidth or traffic resources. It can offer visibility into network performance characteristics (e.g., packet loss and jitter), and some end user characteristics, such as loudness, noise, and echo. Furthermore, new developments now enable speech quality to be accurately predicted based on non-intrusive testing. Finally, unlike active testing, it can measure actual customer traffic. Therefore, it can be used for many Operations Support Systems (OSS) and Business Support Systems (BSS) applications.

Traditional Techniques for Non-intrusive Measurement of Speech Quality

Non-intrusive techniques for measuring speech quality can vary from simply reporting measurements made by endpoints (from RTCP reports), to performing packet loss and jitter measurements, to actually predicting end-user speech quality. This last technique is currently receiving much focus, and significant developments are being made to predict speech quality scores, such as Mean Opinion Scores (MOS), from non-intrusive measurements.

Measuring IP Network Performance Parameters

Traditional non-intrusive testing techniques for VoIP focus on measuring RTP packet loss, jitter, and delay. These measurements are important because they indicate the performance of the IP network that is relevant for delivering voice service. However, these measurements alone do not indicate the end user experience of speech quality. This is because there are many other VoIP processes that impact speech quality, and these various processes, together with IP network performance conditions, influence each other in complex ways to affect overall voice service quality. For example, a random packet loss of 1% could have anywhere from a negligible to a significant impact on speech quality, depending on the codec and packet loss concealment technique used, and where in a voice stream packets are dropped.

As another example, a VoIP network may be experiencing zero packet loss, but excessive jitter could cause the destination jitter buffer to drop packets. The interaction of the IP network's jitter, the gateway's jitter buffer, and the codec's compression scheme and packet loss concealment will affect speech quality in ways that are not directly apparent by packet loss and jitter measurements alone.

Add to these complexities the various other gateway processes such as echo cancellation and gain control, and RTP packet loss and jitter measurements become only one piece in a complicated puzzle.

Recent techniques have been developed to fill-in one of the missing pieces of the non-intrusive speech quality measurement puzzle: the voice codec. By relating RTP packet measurements to known impacts that packet loss has on various codecs, a slightly more accurate picture of speech quality score can be presented.

The ITU E-model

Another technique that is used is known as the "E-model", and is described in ITU Recommendations G.107 and G.108. The E-model was developed as a transmission planning tool, and was not designed as a measurement tool. It is based on assumed "impairment factors" which are assigned to generic network systems (e.g., multiplexers, switches, codecs).

A primary output of the E-model is the Rating Factor, R, also known as the "R Factor". The R Factor is calculated based on adding assumed impairment factors in a call path. It is not based on any measured parameters of an actual network. In fact, G.107 explicitly states:

"It must be emphasized that the primary output from the model is the "Rating Factor" R but this can be transformed to give estimates of customer opinion. Such estimates are only made for transmission planning purposes and not for actual customer opinion prediction (for which there is no agreed-upon model recommended by the ITU-T)."

It should be noted this statement from G.107 is dated 1998, and PESQ was approved after that for the purpose of customer opinion prediction via active measurement.

Another limitation of the E-model for measurement purposes is that it is additive; it simply adds the impairment factors for multiple systems, to arrive at the R Factor. This makes the E-model erroneous when applied to VoIP networks, which are known to be non-additive and non-linear. In this regard, G.107 Annex A states:

"The E-model supposes that different kinds of impairments are additive on the scale of the transmission rating factor R. This feature has not been checked to a satisfying extent. Especially, very few investigations are available regarding the interaction of low bit rate codecs with other kinds of impairments, e.g. with room noise. Additionally, the order effects when tandeming several low bit rate codecs remain uncertain."

Despite these known limitations, the E-model is often used as a basis for network measurement and quality rating. E-model measurement methods vary greatly, but typically comprise performing a small set of actual measurements to determine some impairment factors, using assumptions of the E-model for other non-measured impairment factors, and then adding these to calculate the R Factor.

Due to the variance in methods performed, accuracy will also vary. In general, accuracy is limited to the assumptions of the E-model, which was intended only for planning purposes. Those methods that perform the most measurement (and therefore use the least assumptions) will provide the greater accuracy.

Perhaps the most important factor in the E-model is the Equipment Impairment Factor, Ie, which represents impairments due to network equipment such as VoIP gateways. If an E-model measurement method for VoIP networks is to provide an adequate level of accuracy, it must be able to account for all of the VoIP gateway processes that will contribute to Ie.

Limitations of Traditional Techniques

As previously described, there are many VoIP processes that affect speech quality. These include dynamic jitter buffers, packet loss concealment, echo cancellers, comfort noise generation, automatic gain control, noise reduction, and other proprietary signal processors. Even implementation of standard codecs like G.729 and G.723.1 can vary with regard to load handling capabilities. Some gateways do not perform as well as others under traffic loads when compressing voice.

Techniques for calculating speech quality scores (e.g., MOS) based on RTP packet measurements alone will be limited in their accuracy, because they omit the effects of these other processes. Even techniques that augment the calculation with assumptions about the impact of a specific codec still omit many important factors.

New Techniques for Non-intrusive Measurement of Speech Quality

If a VoIP network is to be designed, deployed, and managed to meet performance objectives, how the actual performance compares with the performance objectives must be known. This requires accurate performance measurements. The higher the level of accuracy, the closer the objectives can be met.

New non-intrusive measurement techniques have now been developed that take into account all VoIP gateway processes that affect speech quality. These new techniques provide a more accurate prediction of end user speech quality because they can measure IP network performance and "know" how a VoIP gateway or phone will react to measured performance in delivering speech quality.

Methods for Accurate Non-intrusive Measurement of Speech Quality

In an attempt to predict speech quality by accounting for all VoIP gateway processes, there are two approaches: analytical and empirical.

The analytical approach attempts to predict speech quality by analyzing the effect each VoIP process has on speech quality, under different IP network conditions, in a qualitative manner. Then, when a certain IP network condition is measured, calculations try to relate the impact of each VoIP process on each network condition, and predict speech quality. As previously described, VoIP processes and their interactions with network conditions are very complicated, and therefore it is extremely difficult to analyze and accurately predict their impact on speech quality. This fact is evident in some of the new analytical methods for non-intrusive speech quality measurement. These methods limit their scope to just one or two VoIP processes under a small set of network conditions. This is because any greater scope becomes exponentially more difficult to analyze and predict.

The empirical approach relies on using actual measurement data of speech quality for different VoIP processes under different IP network conditions. With this approach, speech quality is determined for a certain set of VoIP processes under a specific set of network conditions. Then, when a specific set of network conditions is measured for a call, and a certain set of VoIP processes is known to be used for that call, the speech quality rating can be reproduced. In effect, such a method is calibrated to produce a speech quality rating for specific combinations of VoIP processes and network conditions.

This new empirical approach is offered by the PsyVoIP speech quality measurement from Psytechnics. Psytechnics were the original inventors of PAMS and co-inventors of PESQ, and have calibrated PsyVoIP to PESQ. The output of the PsyVoIP measurement are predictive MOS speech quality scores. This output is truly predictive MOS because the measurement attempts to include all VoIP processes that will affect the experience of the end user.. Through extensive calibration to specific sets of VoIP processes and network conditions, the complex interactions of these processes and conditions are accurately accounted for in the predictive MOS speech quality rating. The result is a speech quality rating for a VoIP call that more accurately correlates with how a human would rate the call.

How New Techniques Address Limitations of Traditional Techniques

These new measurements remove much of the guesswork from speech quality measurement by building upon established measurement techniques like PESQ. PESQ attempts to produce a MOS-like score by modeling speech detection in the human perception domain, so as to account for all impairments that a human would perceive. In turn, the PsyVoIP non-intrusive Predictive MOS measurements attempt to produce PESQ-like scores, and therefore MOS-like scores, by being calibrated to PESQ measurements.

This offers a significant improvement in accuracy over previous methods of simply measuring network conditions, and accounting for only the impact of a codec based on the impacts of a few network conditions. These new predictive MOS methods can account for all VoIP gateway processes, including specific implementations of codecs, dynamic jitter buffers, packet loss concealment, echo cancellers, comfort noise generation, automatic gain control, noise reduction, other proprietary signal processors. Furthermore, they can account for the complex interactions these processes have with a large set of IP network performance conditions.

Use of New Techniques in the E-model

These new techniques also have a use in the E-model method for network measurement. A shortfall of the E-model when used as a measurement technique is the extent of assumptions upon which a result is based. In previous applications of the E-model, the Equipment Impairment Factor Ie, a primary component in the R Factor, was simply assigned from a list of values for generic network equipment. Now, an accurate predictive MOS score can be calculated based on measurements, and an Ie can be derived from that score. This will result in an R Factor that is based more on measurement than assumption.

Applications for New Techniques

New techniques for non-intrusive speech quality measurements have many applications for VoIP networks.
These include:

- ***Certify and qualify new VoIP deployments***
  When a new VoIP service, site, or customer is deployed, non-intrusive predictive MOS scores can be obtained on the new traffic to determine if service performance meets stated objectives in terms of end user quality.

- ***Troubleshoot VoIP service quality impairments***
  Non-intrusive measurements can be made on live VoIP traffic streams. Predictive MOS scores bundled with diagnostics of contributing factors, such as packet loss and jitter analysis, can indicate which traffic is experiencing degraded speech quality, where in the network impairments are occurring, and what is the cause of those impairments.

- ***Baseline network performance***
  Non-intrusive predictive MOS scores on live VoIP traffic can be used to benchmark the nominal performance of a network in terms of speech quality. These benchmark values can then be used as alarm thresholds for network monitoring and as terms in SLAs.

- ***Monitor network performance and SLAs***
  Network performance can be monitored in terms of end user quality by performing non-intrusive predictive MOS measurements on live VoIP traffic.

- ***Optimize VoIP networks and services***
  Speech quality on a VoIP network can be optimized with different network routing and QoS prioritization models. The results can be measured using non-intrusive predictive MOS to determine the optimal network design.

Summary

Due to the efficiency and relatively low cost of non-intrusive testing, and the value of obtaining metrics on live customer traffic, non-intrusive speech quality measurements offer an attractive value proposition. They are simple and relatively inexpensive because they can be software-based and do not utilize network bandwidth or traffic resources. They offer visibility into both network performance characteristics and end-user speech quality. Finally, they can measure actual customer traffic and hence can be applied to many OSS and BSS applications.

However, accuracy is a key issue. If a VoIP network is to be designed, deployed, and managed to meet performance objectives, how the actual performance compares with performance objectives must be known. This requires accuracy in performance measurements.

The evolution of speech quality measurement has progressed from the costly MOS methods, to less expensive computerized methods like PSQM and PESQ, to non-intrusive methods. All along, the focus has been to maximize measurement accuracy. The latest techniques for non-intrusive speech quality measurements are realizing significant improvements in predicting MOS results by using empirical data to accurately account for all VoIP gateway processes and their complex interactions with various IP network performance conditions.

Non-intrusive speech quality measurement is currently a study topic for standards activities. The International Telecommunications Union (ITU) is considering new standards for such measurements, along with improvements to the E-model using these measurement techniques.

Agilent Technologies offers test solutions for IP telephony networks, including both active and non-intrusive speech quality testing. The Agilent Telephony Network Analyzer delivers eXtreme Productivity Improvement, through the provision of simple and precise diagnostics of VoIP Quality of Service (QoS) via non-intrusive measurements, including PsyVoIP speech quality measurement technology from Psytechnics. PsyVoIP is the only non-intrusive speech quality measurement that accounts for VoIP gateway processes and network conditions through extensive calibration for many VoIP gateway and phone models. Using PsyVoIP predictive MOS measurements, along with detailed RTP and RTCP performance analysis, the Telephony Network Analyzer provides visibility into both service quality and network performance.

About the Author

John Anderson is the IP Telephony Product Manager for Agilent Technologies Network Systems Test Division. John is responsible for developing strategies for testing IP Telephony systems and networks. John has over thirteen years experience in the telecommunications industry, including network and systems engineering assignments at MCI Telecommunications and Level 3 Communications. John holds a Bachelor of Science Degree in Electronic Engineering from Iowa State University.

PESQ™ is a registered trademark of Psytechnics, Inc

www.agilent.com

Online assistance: www.agilent.com/find/assist
By internet, phone or fax, get assistance with all your test and measurement needs.

Argentina	+54 11 5811 7115
Australia	1 800 629 485
Austria	+43 (01) 25 125 7006
Belgium	+32 (0) 2 404 9340
Brazil	+55 11 4197 3600
Canada-English	877 894 4414
Canada-French	877 894 4414
China	800 810 0189
Denmark	+45 70 13 15 15
Finland	+358 (0) 10 855 2100
France	+33 (0) 825 010 700
Germany	+49 (0) 18 05 24 63 33
Hong Kong	800 930 871
India	1600 112 929
Ireland	+353 1890 924 204
Israel	+972 3 6892 500
Italy	+39 02 92 60 8484
Japan	0120 421 345
Luxemborg	+32 (0) 2 404 9340
Malaysia	1800 888 848
Mexico	+52 01800 506 4800
Netherlands	+31 (0) 20 5472111
Norway	+47 23 25 3720
Philippines	1800 1651 0170
Poland	+48 22 723 0066
Russia	+7 095 797 3963
Singapore	1800 375 8100
South Korea	080 769 0800
Spain	+34 91 631 3300
Sweden	0200 88 22 55
Switzerland-German	+41 (0) 1 735 9300
Switzerland-Italy	+39 (0) 2 92 60 8484
Switzerland-French	+33 (0) 825 010 700
Taiwan	0800 047 866
Thailand	1800 226 008
United Kingdom	+44 (0) 7004 666666
USA	800 452 4844

Product specifications and descriptions in this document are subject to change without notice.

Printed in U.S.A. September 12, 2002

5988-7878EN

Together with Agilent, gain the Extreme Productivity Improvements that your business demands!

www.agilent.com/comms/XPI

Agilent Technologies