

Agilent Technologies

Methods for Measuring Perceptual Speech Quality

White Paper

By John Anderson
IP Telephony Product Manager

Agilent Technologies
Network Systems Test Division

Table of Contents

Introduction	Page 4
What is Voice-Speech Quality (VSQ)	Page 4
Speech Clarity on Traditional Networks	Page 5
Speech Clarity on Next Generation Networks	Page 6
Traditional Metrics are No Longer Adequate	Page 7
Linearity and Time-Invariance.....	Page 7
Brief Description of Traditional Clarity Metrics	Page 7
Signal-to-Noise Ratio.....	Page 7
Total Harmonic Distortion and Intermodulation Distortion.....	Page 8
Bit Error Rate	Page 8
Modern Networks Require New Metrics	Page 8
Mean Opinion Scores	Page 9
Perceptual Speech Quality Measurement (PSQM)	Page 10
Overview of PSQM Process	Page 11
PSQM Assumptions and Factors	Page 12
Detailed PSQM Process	Page 13
Step 1 - Pre-Processing / Initialization of Signals.....	Page 14
Step 2 - Perceptual Modeling.....	Page 14
Time-Frequency Mapping.....	Page 14
Frequency Warping and Filtering	Page 14
Intensity Warping (Compression)	Page 15
Step 3 - Cognitive Modeling.....	Page 15
Loudness Scaling.....	Page 15
Internal Cognitive Noise.....	Page 16
Asymmetry Processing.....	Page 16
Silent Interval Processing.....	Page 16
Measuring Normalizing Blocks (MNB)	Page 17
Overview of MNB Process	Page 17
Detailed MNB Process	Page 18
Step 1 - Perceptual Transformation.....	Page 18
Step 2 - Compute Frequency Measuring Normalizing Block (FMNB).....	Page 18
Step 3 - Compute Time Measuring Normalizing Block (TMNB).....	Page 18
Step 4 - Generate Auditory Distance Value	Page 19
Step 5 - Mapping AD Values.....	Page 19
Perceptual Speech Quality Measurement Plus (PSQM+)	Page 19
Description of the PSQM+ Improvement	Page 19

Table of Contents (Continued)

Perceptual Analysis Measurement System (PAMS)	Page 20
Overview of PAMS	Page 20
PAMS Assumptions and Factors	Page 22
Detailed PAMS Process	Page 22
Step 1 - Signal Pre-Processing.....	Page 23
Time Alignment	Page 23
Level Alignment.....	Page 23
Equalization.....	Page 23
Step 2 - Auditory Transform.....	Page 23
Step 3 - Error Parameterization.....	Page 24
Step 4 - Regression.....	Page 25
Perceptual Evaluation of Speech Quality (PESQ)	Page 25
Overview of PESQ	Page 25
PESQ Assumptions and Factors	Page 25
Detailed PESQ Process	Page 26
Step 1 - Signal Pre-Processing.....	Page 27
Level Alignment	Page 27
Time Alignment	Page 27
Step 2 - Perceptual Modeling	Page 27
Filtering.....	Page 28
Time-Frequency Mapping.....	Page 28
Frequency Warping	Page 28
Intensity Warping.....	Page 28
Step 3 - Cognitive Modeling.....	Page 28
Time-Frequency Cell Input-Output Difference.....	Page 28
Small Distortion Masking.....	Page 28
Asymmetry Processing.....	Page 28
Frame Disturbances.....	Page 28
Delay Variance Detection.....	Page 29
Time Alignment Re-Assessed	Page 29
Disturbance Values Aggregation and MOS Prediction	Page 29
Comparing Results	Page 29
COM 12-20: PSQM and PSQM+	Page 29
COM 12-58: PSQM, PSQM+, and MNB	Page 30
COM 12-D80: PAMS, PSQM, PSQM+, and MNB	Page 30
NTIA Report 98-347: SNR, CD, BSD, PSQM, and MNB	Page 30
Perceptual Evaluation of Speech Quality	Page 31
What's Next?	Page 31
Agilent Technologies Clarity Measurement Solutions	Page 31
References	Page 33
Agilent Contact Information	Page 36

Introduction

Traditional circuit-switched networks have been designed and optimized over the past century for the time-sensitive delivery of voice traffic. As a result, the Public Switched Telephone Network (PSTN) has provided highly predictable quality of service for voice and, thus, has become the standard-bearer for voice quality. The PSTN delivers an acceptable level of voice quality primarily by allocating dedicated bandwidth and using noncompression analog-to-digital encoding techniques.

Data communication is now seeing a steady trend toward convergent networks, including Voice-over-Internet Protocol (VoIP), voice-over-Asynchronous Transfer Mode (ATM), voice-over-frame relay, wireless, and the PSTN. With the deployment of converged networks, voice quality is neither guaranteed nor predictable. Voice quality has now become a key discriminator for networks, equipment, and services. End users, who ultimately pay the bill for services, have become accustomed to the predictably high quality provided on the PSTN and will not tolerate substandard voice quality. Testing Voice-Speech Quality (VSQ) on new networks is imperative.

Testing can ensure that networks and services are competitive to each other and to the PSTN standard-bearer networks. Measuring overall VSQ, however, is highly subjective, being comprised of several elements. Traditional metrics for VSQ does not provide an accurate measure of subjective quality and no longer apply. Measuring VSQ from the perspective of people is critical to designing and deploying services that will succeed.

One very important aspect of VSQ is the actual *clarity* (clearness and lack of distortion) of the voice signal. During the 1990's, the need for an objective measurement of this very subjective metric quickly arose. Methods for perceptual speech quality measurements, also known as clarity measurements, were developed and evaluated. Many of these measurement methods are now being used in the field. The International Telecommunications Union (ITU) has played an active role in this effort, and provides standards related to the measurement of perceived quality of speech, particularly clarity.

This paper takes a look at the traditional techniques for measuring clarity, their shortcomings, and the new techniques that have been developed in recent years to measure clarity from the perspective of the end users. These methods include Perceptual Speech Quality Measurement (PSQM), Measuring Normalizing Blocks (MNB), Perceptual Speech Quality Measurement Plus (PSQM+), Perceptual Analysis Measurement System (PAMS), and Perceptual Evaluation of Speech Quality (PESQ).

What is Voice-Speech Quality (VSQ)

There are many factors that influence VSQ. Most of these factors can be measured, however, voice quality has the most meaning from the end user's perspective. After all, as mentioned earlier, end users decide what communication services they will pay for and from whom they will buy them. It is from the perspective of the end user that voice quality is best defined.

VSQ is comprised of a number of characteristics including the voice clarity (e.g., sound quality), effects such as echo and time-clipping (dropouts), time delay between spoken phrases, and so on. Of all these parameters, *clarity* and *time delay* are often considered the most important. Other parameters, such as time-clipping and loudness, are actually factors that influence speech clarity, and are often reflected in some kind of clarity measurement or evaluation.

Another key parameter is *echo*, which affects a speaker's perception of quality. Echo is a function of, among other things, network delay. In most cases, however, the listener perceives overall voice quality in terms of speech clarity and time delay.

While time delay is a very objective parameter and can be effectively measured and expressed in milliseconds, speech clarity is a subjective parameter. Clarity can be defined as the clearness, or the fidelity, of voice as it is reproduced by a network and perceived by a listener. In short, clarity is:

- Dependent upon different types of distortion introduced by various network elements.
- Independent of delay; that is, the *clearness* of speech can be maintained and transported with zero, little, or excessive delay across a network. Variance in delay (also known as delay jitter) can, however, influence clarity.
- Independent of echo; since echo is perceived by the speaker and clarity is perceived by the listener.

Several types of distortion in a network can impact clarity. These include:

- Encoding and decoding of voice. This includes waveform encoding with uniform and nonuniform Pulse Code Modulation (PCM), Adaptive Differential Pulse Code Modulation (ADPCM), and low bit-rate voice compression codecs. Linear and non-linear distortions are introduced.
- Time-clipping (also known as front end clipping or FEC) as introduced by Voice Activity Detectors.
- Temporal signal loss and dropouts as introduced by packet or cell loss.
- Delay variance (jitter).
- Environmental noise, including background noise.
- Signal attenuation and gain/attenuation variances.
- Level clipping.
- Transmission channel errors.

While we have defined overall VSQ as a collection of parameters, it is speech *clarity* that is most often used when considering voice quality. In fact, overall voice quality is so dependent on clarity that standards bodies such as the ITU refer to clarity itself as *speech quality*. To avoid ambiguities, however, this paper will adhere to the clear distinction between overall VSQ and speech *clarity*.

Speech Clarity on Traditional Networks

The acceptable standards for speech clarity on a telephone call have been defined for over a century by the PSTN. The original designers of the PSTN went to great lengths to find just the right balance between an acceptable level of quality and cost. Delivering higher speech quality was more expensive (i.e., it required more bandwidth and higher quality phones and encoding processes). Some of the components and characteristics of the PSTN that were intentionally designed to ensure this level of quality include:

- Telephone handset filtering.
- Digital sampling at the oversampled Nyquist rate for voice (8 kHz).
- Mu-law (μ -law) and A-law nonuniform PCM encoding.
- Guaranteed bandwidth at 56/64 kb/s.
- Echo return loss plans (to minimize or eliminate echo).

As a result of these design decisions, PSTN levels of quality (particularly clarity) are quite predictable and reliable. PSTN speech clarity is acceptable, but far from perfect. Several impairments to clarity exist in the PSTN. These include:

- Analog filtering and attenuation in a telephone handset.
- Filtering, attenuation, and possible Electromagnetic Interference (EMI) on analog line transmissions.
- Encoding via nonuniform PCM. Waveform encoders, in particular 64 kb/s PCM (G.711), introduce quantization distortion that has a minimal impact on speech clarity. This impact has been absorbed into the long-accepted standards of clarity expected on the PSTN. In fact, G.711 encoding over a large network actually improves clarity by eliminating impairments introduced by long transmissions over analog lines.
- Bit errors due to channel noise.
- Echo due to many hybrid wire junctions in a call path.

In general, these impairments are easily addressed (e.g., using echo cancellers) or their effects have become accepted as the PSTN clarity standard. Meeting this generally high standard with new packet and wireless networks, however, is a difficult and interesting challenge faced by equipment manufacturers and service providers.

Speech Clarity on Next Generation Networks

The deployment of convergent networks today and in the next few years introduces new challenges to maintaining acceptable standards of speech clarity. Because new technologies are being used to deliver voice services, new impairments are now present in voice networks. Many of these impairments are common across VoIP, ATM, frame relay, and wireless networks.

Most calls traverse hybrid networks that include PSTN-VoIP-PSTN or wireless-PSTN. Hence, the still present traditional impairments introduced by the PSTN are augmented and sometimes exacerbated by impairments introduced by next generation networks. Thus, speech clarity has even more chances to be degraded. Next generation networks in their various forms introduce many impairments to speech clarity. Some key examples are:

- Low-bit rate encoding/decoding of voice. Even high quality codecs like G.726 ADPCM and G.729a do not match G.711 for clarity. Codecs introduce nonlinear distortion and filtering, delay, and under some conditions can introduce severe loss and audio breakup.
- Silence Suppression. The use of Voice Activity Detectors introduce front-end clipping.
- Packet loss which introduces dropouts and time-clipping.
- Packet jitter that can result in 'audio warping' such that speech is not delivered at a constant flow, sometimes causing speech utterances to sound 'warped'. Though VoIP devices have jitter buffers to cancel jitter, perfect reproduction of the audio rate does not always occur.
- Packet delay. While packet delay does not directly impact clarity, increased packet delay can increase loss and jitter.

Although echo does not affect speaker clarity as perceived by a listener, it does impact conversation quality, a key parameter of subjective testing. Perceived echo is made worse by the excessive delay introduced by VoIP processing and transport. Policies for deploying echo cancellers on the PSTN do not always take into account the added processing delay of next generation networks. As a result, echo cancellers may not be deployed where they need to be on a convergent network.

Traditional Metrics Are No Longer Adequate

Traditional metrics used to measure signal quality on traditional circuit-switched networks are most valuable when applied to Linear Time-Invariant (LTI) systems. Some metrics could be adapted to measure non-LTI systems by estimating those systems as LTI in short time segments; however, these metrics are still limited to measuring impairments due to analog transmission and true waveform encoding. Before describing why traditional metrics are less useful in emerging non-LTI convergent networks, an explanation of linearity and time-invariance is useful.

Linearity and Time-Invariance

LTI is a characteristic often ascribed to audio circuits and channels that describes, in a general way, how the audio circuit or channel is likely to behave when it processes an input signal. A system or network is linear if it is both *additive* and *homogenous* [1].

- Additive

The output response resulting from the input $x(t) + y(t)$ is equal to the output response resulting from $x(t)$ plus the output response resulting from $y(t)$. That is, the output function $[F(x+y)](t) = Fx(t) + Fy(t)$. In other words, the output signal from a combined input of signal x and signal y is the same as adding the output signals from individual uncombined inputs of signal x and signal y .

- Homogenous

The output response resulting from the input $a[x(t)]$ is equal to a times the output response resulting from the input $x(t)$, where a is a scalar value. That is, the output function $[F(ax)](t) = a(Fx)(t)$. In other words, a signal multiplied by a and then input into a system would produce the same output as if the signal was input (without multiplying a) and *then* multiplying the output by a .

Linearity can thus be summarized by: $[F(ax+by)](t) = a(Fx)(t) + b(Fy)(t)$

A system or network is time invariant if for any delayed input $x(t - t_0)$, the resulting output response is $y(t - t_0)$ [1]. That is, the shape of the output response waveform is independent of delay. A variation in the delay of an audio signal can cause a time invariant signal to become time variant. Thus, even when using linear encoding techniques, delay jitter introduced by a nondeterministic network (such as a packet network) can produce a time-variant system.

Brief Description of Traditional Clarity Metrics

The measurements traditionally used to evaluate LTI systems are described next.

Signal-to-Noise Ratio

Signal-to-Noise Ratio (SNR) is used to measure relative noise levels on analog signals, and quantization distortion introduced by PCM encoders. SNR can also provide an accurate measurement of the effects bit errors have on a reproduced signal.

For systems with additive noise channels, the received signal is the sum of the transmitted signal plus noise, or $r(t) = s(t) + n(t)$. For systems with bandpass filtering, these are complex functions representing the complex envelopes of the respective signals.

For linear digital encoding systems (e.g., uniform PCM) with possible bit errors, the average-signal to average-noise power output SNR is expressed as [1]:

$$(S/N)_{\text{out}} = \frac{M^2}{1 + 4(M^2 - 1)P_e}$$

Where: *there are M quantization steps in the uniform quantizer,*

P_e is the probability of bit errors,

and the average signal power ratio due to linear quantization distortion is [1]: (S/N)_{out} = M²

For nonuniform quantizers (μ -law and A-law companding), a more complicated and logarithmic expression is used. Output SNR for companding is relatively constant, where it deteriorates rapidly for decreasing input signal levels in uniform PCM systems. Speech pauses can result in falsely poor SNR; therefore, a segmental SNR is used in which SNR is obtained only for intervals of speech and not for the silent periods between utterances.

SNR is useful only when the coding process generally maintains input waveforms at the output. In cases where low bit rate codecs and compression are used, however, it has been found [7] that SNR and segmental SNR measurement results show little correlation to perceived speech clarity. This is one of the reasons why new perceptual clarity algorithms are needed.

Total Harmonic Distortion and Intermodulation Distortion

Total Harmonic Distortion (THD) and Intermodulation Distortion (IMD) measurements are techniques used to evaluate nonlinear distortion introduced by signal processors such as amplifiers. THD is determined from a single-tone input by summing power ratios for outputs of higher orders (>1). IMD is determined from a dual-tone input.

While these metrics are useful for measuring nonlinear distortion for tone inputs, they do not adequately reflect the quality of a signal that has been processed by nonwaveform codecs and packet networks. Again, another reason why new clarity metrics should be employed in emerging convergent networks.

Bit Error Rate

Bit Error Rate (BER) is a very effective measure of quality on physical transmission networks; however, they are obviously insufficient for non-LTI systems. For example, bit pattern may be greatly *errored* by a low bit-rate codec without significantly impacting the subjective speech quality. This is because low bit rate codecs are designed to encode and decode audio characteristics that are important to human perception and not necessarily the exact bit patterns. Thus, clarity metrics that do their work based on human perception should be used when non-LTI systems are part of the telephony path.

Modern Networks Require New Metrics

Next generation networks utilize many technologies such as low bit-rate codecs and packet transport that are non-LTI. As has been described previously, these technologies render many traditional metrics for signal quality (i.e., speech clarity) insufficient. When the transmission path is non-LTI, simple objective measurements, such as those specified in Recommendation G.712 for performance characteristics of PCM systems, are not adequate. In addition, even those traditional metrics that can be applied to non-LTI systems via LTI estimations do not adequately predict a person's perception of speech clarity. For example, segmental SNR and THD will not account for a person's ability to adapt to missing time-frequency energy components as a result of voice encoding.

Traditional measurement results that might otherwise indicate poor speech clarity sometimes expose signal output characteristics that are actually not perceivable to the human listener. For example, slow or local variations in gain or attenuation, slow variations in delay, short time clipping, some filtering, and especially nonwaveform encoding that may distort the waveform (resulting in very poor SNR or BER for example), are all output characteristics that probably would not cause a listener any discomfort.

These output characteristics would, in many cases, be detected by traditional measurements. On the other hand, some perceivable clarity problems may not actually show up in traditional measurements. Remember that nonwaveform codecs attempt to recreate the perceptual audio characteristics and not necessarily the waveform, and so clarity metrics must evaluate signal quality in the same context.

Beginning in the late 1980's, and throughout the 1990's as new technologies were deployed, the industry recognized the need for new measurement techniques that accurately represent clarity the same way humans perceive it. The most important of these methods are described in the sections that follow. These methods are presented in the chronological order in which they were developed and introduced into the industry.

Mean Opinion Scores

The first significant technique used to measure speech clarity was to actually use large numbers of human listeners to produce statistically valid subjective clarity scores. This technique is known as Mean Opinion Scoring (MOS) where the mean value of large volumes of human *opinion scores* are calculated. The techniques for performing MOS testing on networks is generally described in ITU recommendation P.800 [2]. Recommendation P.830 [3] provides more specific methods for subjective testing on speech codecs. Both of these ITU recommendations describe methods for testing, methods for subjective scoring, values of scores, characteristics of speech samples to be used, and other conditions under which testing is to be performed.

MOS testing can be based on two-way conversational tests or on one-way listening tests. Listening tests use standardized speech samples. Listeners hear the samples transmitted over a system or network, and rate the overall quality of the sample, based on *opinion scales*. P.800 specifies several types of subjective testing:

- Conversation Opinion Test
- Absolute Category Rating (ACR) Test
- Quantal-Response Detectability Test
- Degradation Category Rating (DCR) method
- Comparison Category Rating (CCR) method

Each of these methods define corresponding opinion scales. For example, Conversation Opinion and ACR Tests both have a similar scale, called the Conversation Opinion Score and the Listening Quality Score respectively. The Conversation Opinion Test asks subjects their *opinion of the connection you have just been using*. ACR tests ask subjects to rate the *quality of speech*. Scores for both of these scales are as follows:

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

It is this 1-5 scale that is most often cited with reference to MOS testing. Another common example is the ACRs Listening Effort Score. *These ACR tests ask subjects to rate the effort required to understand the meaning of sentences.* Scores are as follows:

Score	Effort required to understand the meaning of sentences
5	Complete relaxation possible; no effort required.
4	Attention necessary; no appreciable effort required.
3	Moderate effort required.
2	Considerable effort required.
1	No meaning understood with any feasible effort.

Obviously, MOS testing has several drawbacks:

- It is subjective because results depend on many uncontrollable attributes of the test subjects including mood, attitude, and culture. Practically speaking, MOS testing is not a repeatable or consistent method for testing.
- It is expensive because it requires a large number of people and elaborate testing setups.
- It is inefficient and impractical to use to perform frequent testing such as that needed for network design and configuration changes and for routine network monitoring.

MOS testing drawbacks suggest that objective, automated, and repeatable testing methods are needed for measuring subjective speech clarity.

Perceptual Speech Quality Measurement (PSQM)

To answer the need for an objective, automated, and repeatable speech clarity testing method that takes into account clarity's subjective nature and human perception, a technique was developed by John G. Beerends and J. A. Stemerdink of KPN Research in The Netherlands. It was called the Perceptual Speech Quality Measurement or PSQM [4].

From 1993-1996, several methods for objective speech clarity measurement, including PSQM, were investigated by NTT and the ITU. These organizations compared the results of these methods to determine which gave the most accurate estimates of subjective clarity. The ITU concluded that PSQM best correlated with the subjective test results (such as those produced by MOS testing).

PSQM was subsequently approved by ITU-T Study Group 12 and published by ITU as Recommendation P.861 in 1996 [5]. It has since gained wide acceptance as a consistent and accurate measurement of speech clarity based on human perception factors.

Overview of PSQM Process

PSQM is a mathematical process that provides a measurement of the subjective quality of speech. The objective of PSQM is to produce scores that reliably predict the results of subjective tests, particularly those methods spelled out in P.830 (MOS). PSQM scores, however, are on a different scale, and reflect a *perceptual distance measure*. That is, PSQM scores reflect the amount of divergence from a *clean* signal that a distorted signal exhibits once it has presumably been processed by some telephony system. This will be discussed in more detail later in this paper.

PSQM is designed to be applied to telephone band signals (300-3400 Hz) processed by speech codecs and measures the distortion introduced by speech codecs according to human perception factors. It is particularly useful for measuring speech clarity when low bit-rate voice compression codecs or vocoders are used. The PSQM testing process is shown in Figure 1.

To perform a PSQM measurement, a sample of recorded human speech is input into a system and is processed by whatever codec is used. The characteristics of the input signal follow those used for MOS testing and are specified in ITU P.830. Input signals may be real human speech or artificial speech per ITU recommendation P.50. The ITU-T recommends that input signals are filtered according to the modified IRS (Intermediate Reference System specified in ITU P.48) receiving characteristics [5] defined in Annex D/P.830. This emulates the receiving frequency characteristics of a telephone handset.

The output signal is recorded as it is received. It is then time-synchronized with the input signal, and the output and input signals are compared by the PSQM algorithm. This comparison is performed on individual time segments (or frames) in the frequency domain (known as time-frequency components), acting on parameters derived from the spectral power densities of the input and output time-frequency components. The comparison is based on factors of human perception, such as frequency and loudness sensitivities, rather than on simple Spectral Power Densities (SPD). PSQM analysis is described in much more detailed later.

The resulting PSQM scores range from 0 to infinity, representing the *perceptual distance* between the input and output signals. For example, a 0 score indicates a perfect match between the input and output signals, often interpreted as perfect clarity. Higher PSQM scores indicate increasing levels of distortion, often interpreted as lower clarity. In practice, upper limits of PSQM scores range from 15-20.

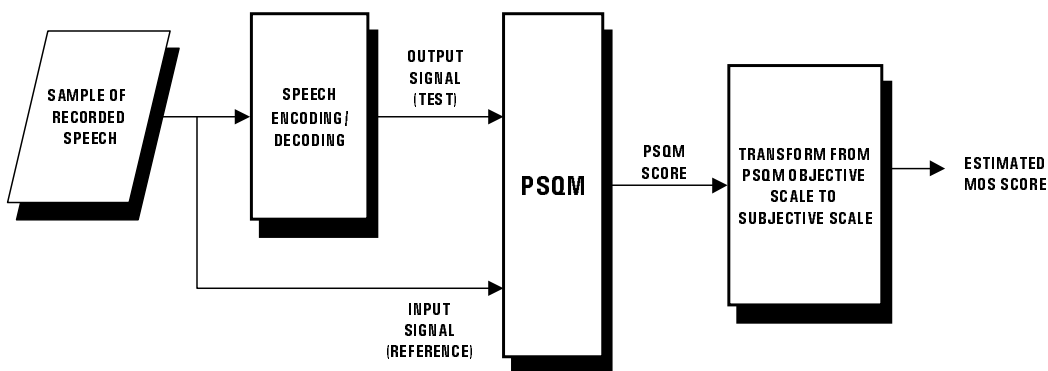


Figure 1. PSQM Testing Process

PSQM Assumptions and Factors

The PSQM algorithm assumes the input signal, transmission systems, and processing systems exhibit certain characteristics. These are:

- Input and output signals must be time-synchronized prior to PSQM analysis. Accurate time-synchronization is critical because PSQM performs frequency-domain comparisons of the input and output signals in time segments (256 samples for 8kHz sampled signals). If the two signals are not accurately synchronized, the comparison will be performed as if the output signal was synchronized with the input signal and simply exhibited poor quality. False poor quality scores (high PSQM scores) will result.

NOTE

There are several methods among PSQM test tools for time-synchronization, and only the most robust methods will produce accurate scores. Time-synchronization is beyond the scope of this paper.

- If accurate time-synchronization is performed, any existing network delay will have no impact on PSQM analysis. If network delay, however, changes greatly while the output signal is being recorded, higher PSQM scores will result for at least part of the speech sample even if time-synchronization was performed.
- The input signal is free of background noise, however, because PSQM compares what is present in the output signal with what is present in the input signal, existing background noise in the input signal will have no impact on PSQM results if it is present in the output signal.
- There are no channel degradations such as bit errors, packet loss, and time clipping. PSQM does not attempt to measure the impacts of these types of degradations. Again, PSQM simply compares what is present in the output with what is present in the input. PSQM will produce lower scores (higher quality) than subjective tests when clipping or loss is present due to level scaling that is performed prior to analysis. The scaling process is described later.

PSQM accurately predicts subjective speech clarity test results when those results are impacted by the following processes or parameters [5]:

- Waveform codecs (e.g., G.711, G.726).
- CELP-based codecs ≥ 4 kpbs (e.g., G.729a, G.723.1 5.3 and 6.3 rates, G.728).
- Multiple bit rates of a codec.
- Transcodings (the conversion from one digital format to another).
- Speech input levels to codecs.
- Talker dependencies (e.g., languages, phrases).

The accuracy of PSQM is currently unknown for, or PSQM is not intended to measure the impacts of, the following parameters [5]:

- Delay
- Delay variance
- Time-clipping

- Level clipping
- Overall system gain/attenuation
- Multiple simultaneous talkers
- Transmission channel errors (BER, packet loss)
- Bit-rate mismatching between encoder and decoder for multiple bit rates
- Background noise
- Music as input signal
- CELP codecs < 4kbs

PSQM compensates for overall system gain/attenuation by global scaling prior to signal comparison. Thus, gain/attenuation should not be reflected in PSQM scores. However, due to the level scaling performed by PSQM, noise may be scaled upward along with the voice component in an output signal when the output is attenuated, resulting in higher PSQM scores due to the amplified noise. This effect is typically present when output signal attenuation is greater than 10dB. Also, slow variations in gain/attenuation that are normally not perceived by human listeners will not be reflected in PSQM scores. Fast variations that *are* often perceived, however, will be reflected.

With regard to speech input levels and talker dependencies, note the effects of large signals through μ -law and A-law PCM: greater quantization errors at higher levels. Under severe conditions, nonlinear vocoders can produce level clipping. These impairments will result in additive or subtractive distortion during coding that will be measured by PSQM. Thus, a single voice sample played across a codec may produce PSQM scores that vary in time and that are dependent on utterances within the sample (e.g., an accented emphasis on a word may produce a higher PSQM score than nonemphasized words).

Detailed PSQM Process

This section provides a detailed description of PSQMs pre-processing and signal comparison algorithm [5]. Many of the concepts described here will be referred to later in the paper as other perceptual algorithms are presented. The PSQM model is shown in Figure 2.

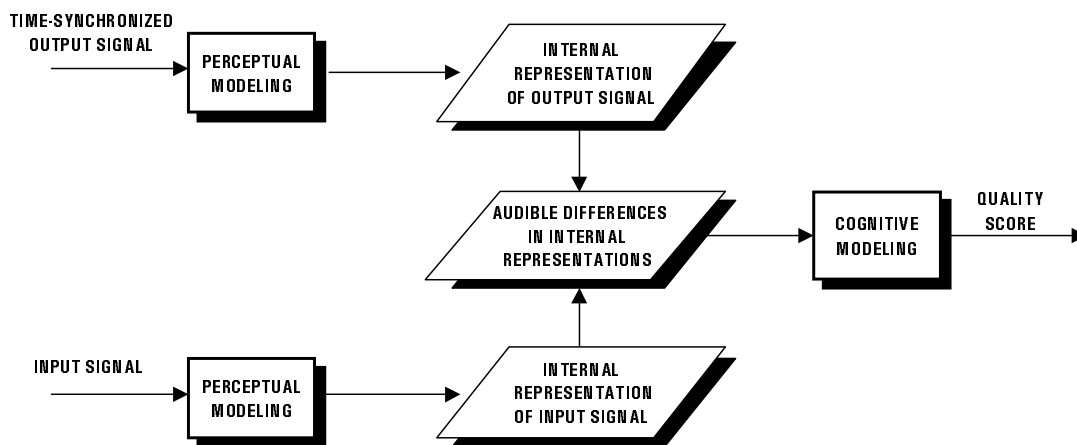


Figure 2. The PSQM Model

Referring to Figure 2, the PSQM process steps are:

Step 1 - Pre-processing / Initialization of Signals

Prior to PSQM's perceptual modeling process, and prior to the cognitive modeling process that produces the actual PSQM score, global initialization is performed. PSQM assumes that input signals are 8-kHz or 16-kHz sampled, 16-bit linear PCM encoded signals. P.861 specifies three initialization processes:

- Time alignment in which input and output signals are aligned in time (described earlier). The PSQM algorithm itself does not perform time alignment, but rather assumes the input and output signals are already aligned.
- Global scaling in which the output signal is scaled to compensate for overall gain of the system under test.
- Global speech loudness calibration in which a calibration factor between an assumed listening level and loudness is calculated. This factor helps determine the loudness of the output signal based on the assumed listening level (for example, the audio level an ear would receive the speech signal over a telephone) and human hearing thresholds at different frequencies. This factor is used later to calculate loudness densities (i.e., in the Intensity Warping process described later).

Step 2 - Perceptual Modeling

The next pre-processing step, perceptual modeling, is the transformation from the physical (external) domain to the psychophysical (internal) domain. In other words, the PSQM algorithm takes a mathematical representation of the actual physical signal and converts it to a mathematical representation that accounts for the physiological realities of human perception. This is performed by three operations:

- Time-Frequency Mapping

A Fast Fourier Transform (FFT) is performed on the input and output time domain signals (power vs. time) to transform them into the frequency domain. This is done in 32-millisecond time frames resulting in time-frequency components called *cells* (see Figure 3). A cell, therefore, represents a specific frequency band in a specific time frame (256 digital samples per time frame for signals with 8-kHz sampling frequency).

NOTE

The 32-millisecond time frame is chosen because it corresponds to the *window length* or time threshold for human hearing.

- Frequency Warping and Filtering

The traditional Hertz frequency scale is warped to take into account human frequency sensitivities. The scale is warped to specific critical bands such that the scale is no longer strictly linear. PSQM defines 56 critical bands. Input and output signals are also filtered based on telephone handset characteristics.

- Intensity Warping (Compression)

The intensity scale, which is based on power densities, is warped to a loudness scale to represent human loudness sensitivities. This is necessary because humans perceive distortion differently depending on the loudness of the audio signal in which noise is present (noise in loud signals is less noticeable than noise in soft signals). PSQM calculates some parameters, such as loudness, for an entire frame and compares this value with that for the same parameter for individual time-frequency cells, applying local scaling where necessary. This enables PSQM to distinguish codec distortion (which acts on individual time-frequency cells) from overall signal attenuation or gain.

The output of this Perceptual Modeling process is often referred to as an *internal representation of the input and output signals*. Restated another way, this process produces a mathematical representation of acoustic signals that takes into account human physiology and auditory sensitivities.

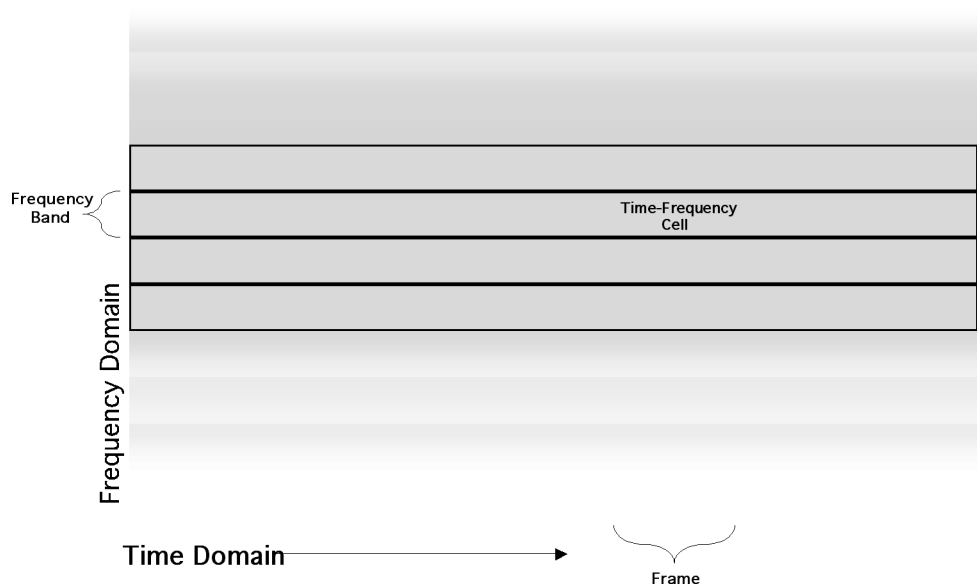


Figure 3. Time Frequency Domain of Signals

Step 3 - Cognitive Modeling

With the completion of two pre-processing steps, the final step, called Cognitive Modeling, is where the input and output signals are directly compared and the actual PSQM score is produced. Cognitive Modeling evaluates the audible errors in the output signal by essentially computing the noise disturbance for individual time-frequency cells. Remember, scaling of the entire frame is first performed to partition disturbances within individual cells from that of the overall signal due to system attenuation or gain. The average noise disturbance is directly related to the quality of coded speech. Cognitive Modeling is performed by four operations:

- Loudness Scaling

Within each frame, the loudness density of the output signal is scaled relative to the input signal.

- Internal Cognitive Noise

Noise disturbance is calculated as the difference in loudness density between the input and output signals as would be perceived by a human listener.

- Asymmetry Processing

Human perception of subjective speech clarity is asymmetrical. When a time-frequency component is not coded (signal loss in the output signal), it affects subjective quality less than those cases when an unrelated time-frequency component is introduced (added distortion). That is, small additive codec distortion is more noticeable to a human listener than small signal loss due to codec distortion. PSQM addresses this asymmetry effect by scaling the noise disturbance for time-frequency components (cells) differently. If a disturbance is caused by additional energy in a cell, PSQM scales the noise disturbance with a factor greater than 1, resulting in a higher PSQM score.

If a disturbance is caused by missing energy in a cell, PSQM scales the noise disturbance with a factor less than 1, resulting in a lower PSQM score. As a result of this asymmetry processing, PSQM scores correlate better with subjective results with regard to added energy versus missing energy in cells of coded speech.

NOTE

The above process is addressed differently by PSQM+ as an improvement to PSQM. PSQM+ is discussed later in the paper.

- Silent Interval Processing

To better match human perception, differences between the input and output signals during intervals of silence should have less impact on PSQM scores than differences during speech. PSQM computes average noise loudness for silence frames and for speech frames separately, applying different weighting factors for each.

The output of the Cognitive Modeling process is an objective value called the PSQM score, which ranges from 0 (perfect) to 15 (extremely poor) or higher. Some important final points to note about PSQM:

- Differences between the input and output signals, if inaudible, will not result in higher PSQM scores (lower clarity/quality).
- If the input and output signals are identical, PSQM will predict perfect quality regardless of the quality of the input signal. In other words, if a noisy signal is faithfully reproduced by the system under test, a comparison between the noisy input and equally noisy output signals will produce very good or perfect PSQM scores.
- Different models of PSQM test tools may give different PSQM scores for the same call. This may be due to two reasons:
 - Different methods of time-synchronization may have different levels of accuracy, resulting in one method having more of a mismatch than the other. Remember that timing mismatches will result in higher PSQM scores.
 - Some tools perform PSQM while others perform PSQM+. When testing a network with transmission impairments or silence suppression, PSQM and PSQM+ can produce different scores.

Measuring Normalizing Blocks (MNB)

In 1997, and based on a report [7] by Stephen D. Voran of the Institute for Telecommunications Sciences, ITU Study Group 12 Contribution 24 (COM 12-24-E) [6] was published as a proposed annex to P.861 (PSQM). This annex was accepted in 1998 as appendix II to P.861. It describes an alternative technique to PSQM for measuring the perceptual distance between the perceptually transformed (internal representations) of the input and output signals described in the last section. This technique is known as Measuring Normalizing Blocks (MNB).

Overview of MNB Process

It was observed [6] that "listeners adapt and react differently to spectral deviations that span different time and frequency scales." MNBs address this issue by analyzing perceptual distance across multiple time and frequency scales, working from larger scales to smaller ones. The MNB technique is recommended for use for measuring the impact of the following on speech clarity [7]:

- Transmission channel errors
- CELP and hybrid codecs with bitrates less than < 4 kb/s
- Vocoders

There are two types of MNBs: Time Measuring Normalizing Blocks and Frequency Measuring Normalizing Blocks. The algorithm creates a single, nonnegative value called Auditory Distance (AD), which is a measure of perceptual distance between the input (reference) and output (test) signals in order to predict subjective quality. The MNB model is shown in Figure 4.

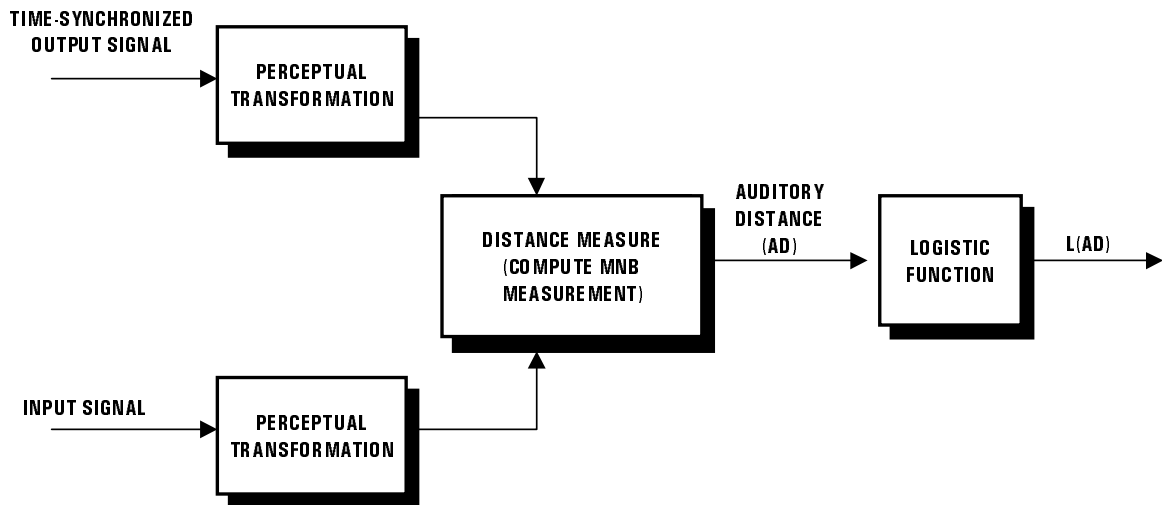


Figure 4. The MNB Model

Detailed MNB Process

The MNB process [6] is applied to the perceptually modified input and output signals as described in the last section; that is, the *internal* representations of the input and output signals. Referring to Figure 4, the MNB process steps are:

Step 1 - Perceptual Transformation

The time-synchronized input (reference) and output (test) signals are input into the model. The input and output signals are level-aligned by removing the DC component of each signal. Both signals are mapped to the frequency domain. Silence frames are detected and removed. The power scale is transformed to a perceived loudness scale.

Step 2 - Compute Frequency Measuring Normalizing Block (FMNB)

The perceptually transformed input and output signals are process inputs to an FMNB. The process outputs of the FMNB are a set of integrated difference measurements, and a normalized output signal. This is described as follows:

- The perceptually transformed input and output signals are in the time-frequency domain; that is, they are functions of both time and frequency. These input and output signal functions are each mathematically integrated over the time scale of the input signal.
- The integrated input signal is subtracted from the integrated output signal. The result is a function of frequency (for a specific value of time) that represents the difference in the signals.
- The measured difference (above) is subtracted from the output signal (at different frequencies), producing a normalized output signal.
- The positive and negative portions of the difference measurement are mathematically integrated over four frequency bands that span the telephony band. The results are four FMNB measurements.

Step 3 - Compute Time Measuring Normalizing Blocks (TMNB)

The perceptually transformed input and normalized output signals are process inputs to a TMNB. The process outputs of the TMNB are a set of integrated difference measurements, and the normalized output signal. MNBs are idempotent, which means that a second pass of the output signal through an MNB will not further alter it by normalizing it; if the TMNB follows the FMNB, the output signal going into the TMNB will be the same as the output signal coming out of the TMNB. TMNBs are computed across different frequency bands, in progressive iterations. This is described as follows:

- The perceptually transformed input and output signals are in the time-frequency domain; that is, they are functions of both time and frequency. These input and output signal functions are each mathematically integrated over the frequency scale of the input signal.
- The integrated input signal is subtracted from the integrated output signal. The result is a function of time (for a specific value of frequency) that represents the difference of the signals.
- The measured difference is subtracted from the output signal (at different times), producing a normalized output signal.
- The positive and negative portions of the difference measurement are mathematically integrated over time.

Nine TMNBs are computed, each with possibly two measurement results (positive and negative). Since MNBs are computed at decreasing time and frequency scales, when modeling how human listening adapts the spectral deviations at each scale are removed before the next scale is measured.

Step 4 - Generate Auditory Distance Value

Linearly-independent FMNB and TMNB measurements are combined, with weighting factors, to generate an Auditory Distance (AD) value.

Step 5 - Mapping AD Values

A logistic function maps AD values into a finite range to provide correlation with MOS scores.

Perceptual Speech Quality Measurement Plus (PSQM+)

As the rapid deployment of next generation networks drove a strong need for speech quality testing (particularly speech *clarity*), PSQM gained widespread acceptance. PSQM became popular as a way to measure voice clarity not only across vocoders but also across entire networks, including VoIP. One PSQM drawback, however, is that it does not accurately report the impact of distortion when that distortion is caused by packet loss or other types of time clipping. In other words, PSQM would report better clarity under these conditions than a human listener would.

Recognizing the need to account for these network transmission impairments, J. G. Beerends, E. J. Meijer, and A. P. Hekstra, developed an improvement to the PSQM model and submitted it as a contribution to the ITU-T. The model was reviewed by ITU Study Group 12 and published in 1997 under COM 12-20-E [8]. This model became known as PSQM+, and quickly became a preferred method for measuring speech quality in network environments. PSQM+ (which is based directly on the ITU P.861 PSQM model and, in reality, represents only a slight deviation) improves the way the PSQM technique is applied to a system which experiences severe distortions due to time clipping and packet loss. For systems comprising speech encoding only, PSQM and PSQM+ give identical scores.

Description of the PSQM+ Improvement

Remember in the description of the asymmetry effect in the cognitive modeling process of the PSQM section earlier, PSQM treats a distortion represented by signal gain within a time-frequency cell differently than distortion represented by signal loss within a cell. Because additive signal distortion (added energy) has a larger impact on perceived clarity than subtractive distortion (missing energy), PSQM *scales up* the additive disturbance to result in higher PSQM scores (i.e., poorer quality scores) and *scales down* the subtractive disturbance to result in lower PSQM scores (i.e., higher quality scores) [8]. For small distortions that are likely due to a codec, PSQM provides excellent correlation with subjective test results. For large distortions due to time-clipping and packet loss in which all cells within a frame experience severe loss of signal energy, PSQM produces far too low of scores in comparison with subjective test results.

To account for this problem in PSQM's asymmetric processing, and to make sure PSQM+ correlates more accurately with subjective test results than PSQM, PSQM+ adds a second scaling factor that counters the PSQM scaling factor under conditions of severe distortions represented by large loss of energy in a time-frequency cell [8]. This new factor is applied to each frame. When input and output signal power are about the same in a frame, the new factor is about 1.0, and thus PSQM+ produces about the same score as PSQM.

When a large distortion such as time clipping or packet loss is introduced (causing the original PSQM algorithm to *scale down* the noise disturbance), the PSQM+ algorithm applies another scaling factor that has an opposite effect, and scales up the noise disturbance. This results in higher PSQM+ scores, which correlate with subjective test results more accurately.

Within PSQM+, the second scaling factor is always applied. But when distortions are small, it equals 1 and has little or no impact. As distortions increase, it counters the first scaling factor. For loud distortions (added energy), the second scaling factor leads to lower scores; for severe loss (missing energy), it leads to higher scores. So for small distortions due to codecs, both PSQM and PSQM+ produce about the same scores, which correlate well with subjective testing. For severe loss and time-clipping distortions, PSQM+ will produce higher scores that correlate better than PSQM. For loud distortions, PSQM+ will produce lower scores that correlate better than PSQM.

Note that PSQM and PSQM+ distinguish distortions within time-frequency cells from those for an entire time frame. If distortions are time-frequency cell dependent, they are likely due to codec distortion. If a distortion affects all cells within a frame similarly, it is likely due to time-clipping or packet loss.

Perceptual Analysis Measurement System (PAMS)

An effort independent of those at KPN Research (where PSQM was developed) was undertaken at British Telecommunications to address the problem of objectively measuring subjective speech clarity. The result was a technique called the Perceptual Analysis Measurement System (PAMS). PAMS was developed by the PsyTechnics group within British Telecommunications and was first issued in August 1998, with version 3.0 being released in February 2000 [9].

PAMS offers a different model than PSQM+ but with the same goal: to objectively predict results of subjective speech quality tests for systems on which coding distortions as well as time-clipping and packet loss are potentially problems. PAMS first gained wide acceptance in Europe, and is experiencing an increase in use throughout North America.

Overview of PAMS

PAMS uses a model based on factors of human perception to measure the perceived speech clarity of an output signal as compared with the input signal. Although similar to PSQM in many aspects, PAMS uses different signal processing techniques, and a different perceptual model. The PAMS testing process is shown in Figure 5.

To perform a PAMS measurement, a sample of recorded human speech is input into a system or network. The characteristics of the input signal follow those that are used for MOS testing and are specified in P.830. Though natural speech samples may be used, PAMS is optimized for proprietary artificial-like speech samples.

The output signal is recorded as it is received. The input and output signals are then input into the PAMS model. PAMS performs time alignment, level alignment, and equalization to remove the effects of delay, overall systems gain/attenuation, and analog phone filtering. Time alignment is performed in time segments, so that the negative effects of large delay variations (that cause problems for PSQM) are removed. The perceivable effects of delay variation, however, are preserved and reflected in PAMS scores.

PAMS then compares the input and output signal in the time-frequency domain, comparing time-frequency cells within time frames. This comparison is based on human perception factors. Note that the definition of *time-frequency cell* and *time frame* are similar to those used in the PSQM case.

The results of the PAMS comparison are scores that range from 0-5, and that correlate on the same scale as MOS testing. In particular, PAMS produces a Listening Quality Score and a Listening Effort Score that correspond with both the ACR opinion scale in P.800 and the opinion scale in P.830 [2,3,9].

Listening Quality Scale

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

Listening Effort Scale

Score	Effort required to understand the meaning of sentences
5	Complete relaxation possible; no effort required.
4	Attention necessary; no appreciable effort required.
3	Moderate effort required.
2	Considerable effort required.
1	No meaning understood with any feasible effort.

Several other measured distortion parameters are produced by PAMS, including the calculation of an Error Surface.

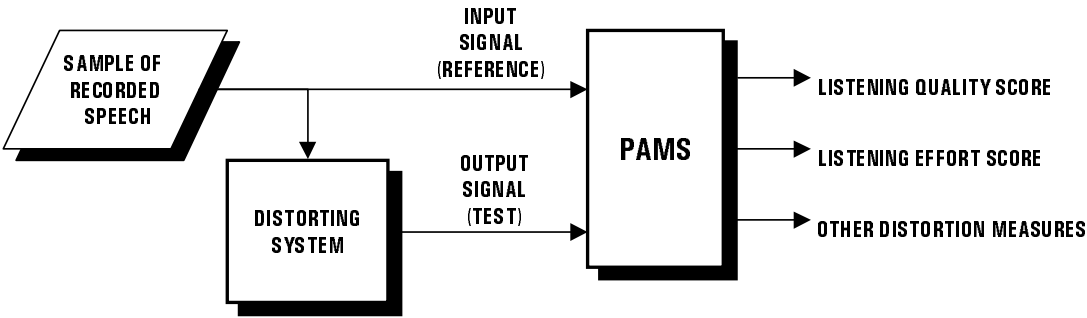


Figure 5. PAMS Testing Process

PAMS Assumptions and Factors

The perceptual modeling within PAMS assumes that there is no delay, no large delay variations, no overall system gain/attenuation, or no analog telephone characteristic filtering. The PAMS algorithm performs unique signal pre-processing to remove the effects of these conditions. As a result, PAMS focuses on measuring speech clarity based on the effects of coding distortion, time-clipping, packet loss, and jitter.

PAMS also assumes no background noise that will impact speech clarity. PAMS compares what is in the output signal with what is in the input signal. Noise that is not present in the input signal can be added by a system under test, and its impact measured by PAMS.

PAMS assumes a constant listening level of about 79dB SPL (per P.830). Again, signal pre-processing performed by PAMS accounts for this, so that input and output signals of different levels (as a result of overall system gain or attenuation) may be input into PAMS.

PAMS accurately predicts subjective speech clarity test results when speech clarity is affected by the following processes or parameters [9]:

- Waveform codecs (e.g., G.711, G.726)
- Nonwaveform Vocoders, including those with multiple bit rates
- Transcodings (the conversion from one digital format to another)
- Speech input levels to codecs
- Talker dependencies (e.g., languages, phrases)
- Fast delay variances
- Time-clipping
- Level clipping
- Added noise

PAMS is not intended to measure the impact of the following processes or parameters on speech clarity [9]:

- Delay
- Slow delay variances
- Overall system gain/attenuation
- Analog phone filtering
- Background noise present in input signal
- Music as input signal

Detailed PAMS Process

The following section describes in more detail the pre-processing and analysis performed by the PAMS algorithm [9]. The PAMS model is shown in Figure 6.

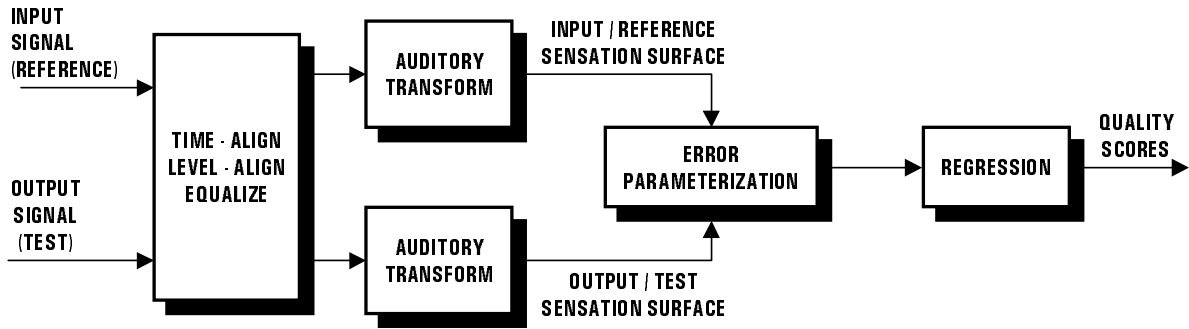


Figure 6. The PAMS Model

Referring to Figure 6, the PAMS process steps are:

Step 1 - Signal Pre-Processing

PAMS pre-processes the input and output signals by performing the following operations:

- Time Alignment

Input and output signals are time-synchronized in individual time segments, or utterances, that vary in length. This is to remove the effects of delay and slow delay variation. Fast delay variation that is perceptible is preserved and is measured by PAMS.

- Level Alignment

Input and output signals are level-aligned in time segments to remove the effects of overall system gain/attenuation and slow variances.

- Equalization

Input and output signal spectrums are equalized to remove the effects of telephone band (300-3400 Hz) filtering.

Unlike PSQM, the PAMS algorithm performs its own time alignment and equalization.

Step 2 - Auditory Transform

PAMS performs a perceptual modeling process to transform the input and output signals into the time-frequency domain. PAMS uses Sekey filter banks to filter audible signals down to the relevant perceptual domain and places the signals in 19 frequency bands. It performs frequency shaping (using a Bark scale) to reflect human frequency sensitivities. The Bark Scale is a nonuniform frequency scale that represents, better than a uniform Hertz scale would, how humans perceive audio energies (e.g., speech) differently at different frequency ranges. The Bark scale reflects that at low frequencies, the human hearing system has a higher frequency resolution than at high frequencies. In perceptual modeling in PSQM, PSQM+, PAMS, and PESQ, the input and output test signals are transformed from a uniform frequency scale to a Bark scale to better represent how humans perceive audio signals in the frequency domain. Time intervals are in 4-millisecond samples.

The result is a representation in time and frequency of perceived loudness, known as the Sensation Surface, that is analogous to the spectral power density, but is based on how a human would perceive the signal in each time-frequency cell. PAMS calculates Sensation Surfaces for both the input and the output signals.

Step 3 - Error Parameterization

PAMS determines the audible differences in the Sensation Surfaces of the input and output signals by subtracting the input signal Sensation Surface from the output signal Sensation Surface. The result is another time-frequency representation known as the Error Surface. The Error Surface represents the audible errors, in time-frequency cells, found in the output signal when compared to the input signal.

Errors that represent added signal energy (e.g., noise or added codec distortion) have positive values in a cell in the Error Surface. Errors that represent lost signal energy (muting, packet loss, time-clipping) have negative values. The amplitude of each cell in the Error Surface is related to the level of human perception or annoyance.

PAMS analyzes the Error Surface in several ways. It calculates the average positive distortion (added signal energy in a cell) and the average negative distortion (lost signal energy in a cell). Several error parameters are calculated that indicate the amount of audible errors.

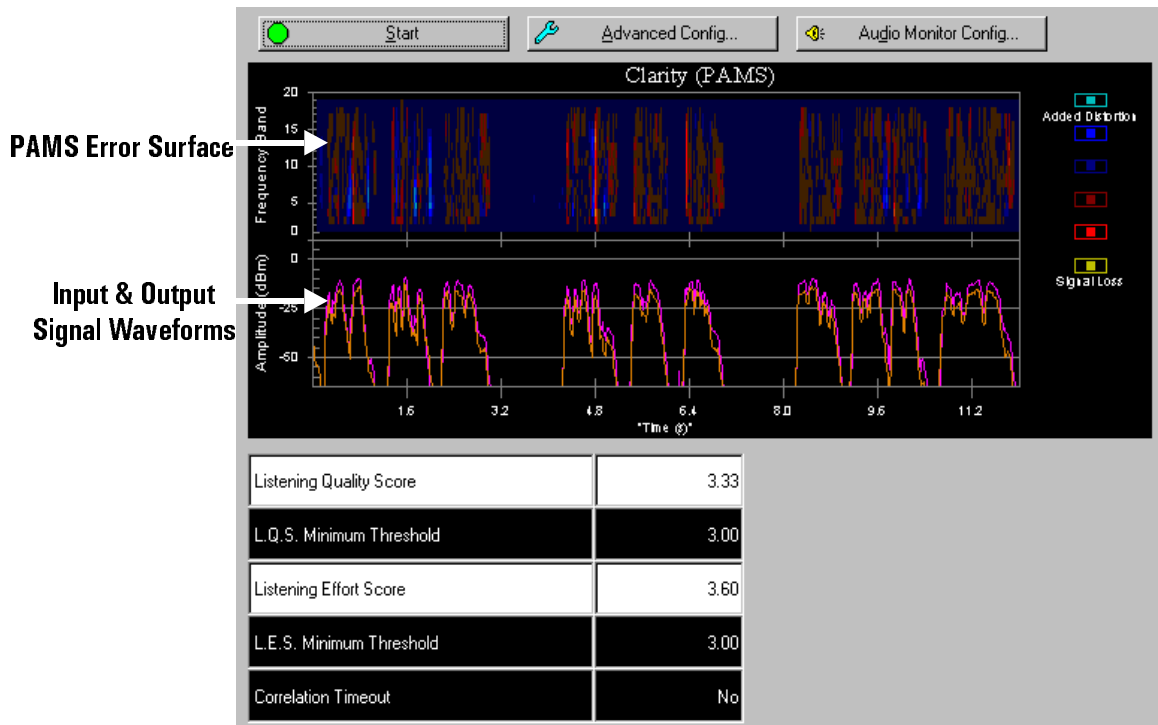


Figure 7. PAMS Error Surface

Step 4 - Regression

Finally, audible errors are evaluated and mapped to predictive scores based on correlation with a large database of subjective testing results. This database of subjective test results indicates how a person would score specific audible errors. The error parameters calculated by the PAMS Error Parameterization process can thus be mapped to predictive scores to reflect the results produced had human test subjects performed this mapping.

Perceptual Evaluation of Speech Quality (PESQ)

PSQM was again improved upon by KPN Research in 1999 to so that it would correlate better with subjective tests under network conditions. This improvement is known as PSQM99.

In a study period from 1998-2000, the ITU reviewed five submissions of drafts for new perceptual speech quality measurements. Included in this review were PSQM99 and PAMS, both of which proved to match subjective testing the best. It was determined that each had significant merits and that it would be beneficial to the industry to combine the merits of each into a new measurement technique [10,11].

A collaborative draft was submitted to the ITU in May 2000 by John G. Beerends and Andries P. Hekstra of KPN Research, and by Anthony W. Rix and Mike Hollier of British Telecommunications [10]. The new perceptual speech clarity algorithm is called the Perceptual Evaluation of Speech Quality (PESQ), and is currently under review by the ITU, with expected approval as Recommendation P.862 in 2001. It is widely anticipated that PESQ will eventually replace previous methods for objective speech quality testing.

Overview of PESQ

Like PSQM and PAMS before it, PESQ is still directed at narrowband telephone signals. It is applicable to systems with speech coding (including low bit-rate vocoders), variable delay, filtering, packet or cell loss, time-clipping, and channel errors. PESQ scores predict listening quality scores for ACR listening tests.

PESQ leverages the best features of PAMS and PSQM99. It combines the robust time-alignment techniques of PAMS with the accurate perceptual modeling of PSQM99, and it adds new methods including transfer function equalization and a new method for averaging distortion over time.

The time-alignment methodology of PAMS cancels effects of delay and slow delay variances while preserving the effects of perceivable impairments due to fast delay variances. The perceptual modeling of PSQM99 differs from PSQM and PSQM+ in the asymmetry processing and scaling. PSQM99 provides more accurate correlation with subjective test results than PSQM or PSQM+ [10,11].

PESQ Assumptions and Factors

PESQ is effective for measuring the impact of at least the following conditions or processes on subjective speech clarity [10,11]:

- Waveform codecs (e.g., G.711, G.726)
- Nonwaveform codecs ($\geq 4\text{kb/s}$), including those with multiple bit rates
- Transcodings (the conversion from one digital format to another)
- Speech input levels to a codec
- Transmission channel errors
- Noise added by system (not present in input signal)
- Short- and long-term time warping
- Packet loss and time-clipping (it has been noted [10] that PESQ is more sensitive to front-end clipping and less sensitive to time clipping than subjective testing; correlation may vary)

The accuracy of PESQ is currently unknown for, or PESQ is not intended to measure the impacts of, the following parameters [10,11]:

- Delay (is canceled by time alignment)
- Listening levels and overall system gain/attenuation (canceled by level alignment)
- Talker dependencies (vocal characteristics such as tonal range, language, gender, age, etc.)
- Multiple simultaneous talkers
- Bit-rate mismatching between encoder and decoder
- Background noise present in input signal
- Artificial speech as input signal
- Music as input signal
- Codecs $< 4\text{kb/s}$
- Level clipping
- Listener echo and sidetone

Detailed PESQ Process

The following section describes the pre-processing and analysis performed by the PESQ algorithm [10,11]. The PESQ model is shown in Figure 8.

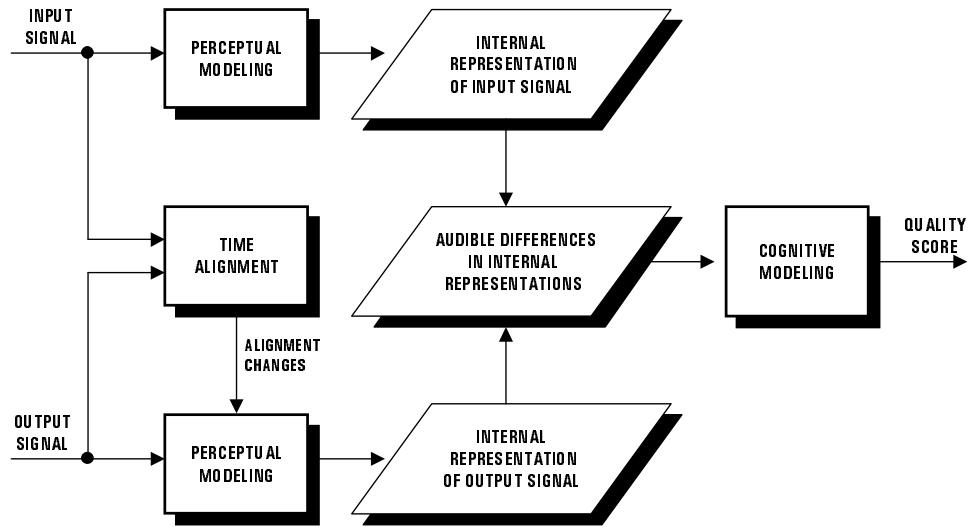


Figure 8. The PESQ Model

The input signal characteristics for PESQ are consistent with those for PSQM and PAMS. Natural speech samples can be used, and should follow P.830 guidelines, including specifications for listening level and filtering. Artificial speech may be used but should represent temporal and phonetic structures of natural speech. Signals should be stored as 8-kHz or 16-kHz sampled, 16-bit linear PCM encoded signals. Referring to Figure 8, the PESQ process steps are:

Step 1 - Signal Pre-Processing

Signal processing prior to the signals being analyzed by the Perceptual Model includes these operations:

- Level Alignment

The input and output signals are level-aligned to the same constant power level to account for overall system gain/attenuation and slow gain/attenuation changes. First the signals are filtered, then their average power values computed, and finally gains are applied to align both signals.

- Time Alignment

The output signal is time shifted to align with the input signal. This is performed in individual time segments, or speech utterances, that vary in length. In fact, many delay calculations within an utterance are performed, and an utterance may be divided if delay calculations vary greatly. Delay during both speech and silence is accounted for by performing time alignment.

Step 2 - Perceptual Modeling

Perceptual Modeling transforms the input and output signals into internal (human perception) representations. This modeling includes time-frequency mapping (similar to PSQM), frequency warping (using a modified Bark scale), and compressive loudness scaling in which a straight dBm scale and warping in a nonlinear way to reflect human sensitivities.

- Filtering

Filtering is applied to conform the signals to telephone band characteristics. Thus, filtering characteristics of a telephone handset do not impact the PESQ measurement.

- Time-Frequency Mapping

Time-frequency mapping. Like PSQM, an FFT with a window size of 32-millisecond, or 256 samples for 8-kHz sampling rate, is used to segment the input and output signals into individual time-frequency cells.

- Frequency Warping

Frequency warping to a modified Bark scale is performed to reflect human frequency sensitivities.

- Intensity Warping

Intensity warping of spectral power to a Sone loudness scale is performed.

The Perceptual Modeling process produces internal representations of the input and output signals. That is, the output is a time-frequency representation of input and output signals that take into account human perception characteristics.

Step 3 - Cognitive Modeling

A Cognitive Modeling process is performed to calculate two types of average noise disturbance values. These two values are combined at the end to produce a predicted MOS score.

- Time-Frequency Cell Input-Output Difference

For each time-frequency cell, a difference between output and input signal is calculated. A positive difference indicates components such as noise have been added. A negative difference indicates components have been omitted, for example, due to coding distortion or signal loss.

- Small Distortion Masking

A threshold of sorts is applied to distortion levels within each cell, and a corresponding scaling towards zero disturbance is performed to mask the impact of small distortions that are not perceivable in the presence of loud signals.

- Asymmetry Processing

Asymmetry processing is performed similar to P.861 PSQM, calculating asymmetrical disturbance using a scaling factor to apply different weightings to positive and negative disturbances. Asymmetrical disturbance is such that only cells with positive disturbances remain. The counter-scaling that is part of PSQM+ is not performed in PESQ. The results of PESQ asymmetry processing are values for asymmetrical disturbances, which are for positive disturbances only, and normal disturbances, which include positive and negative disturbances.

- Frame Disturbances

Normal (nonasymmetrical) disturbances and asymmetrical disturbances are calculated and aggregated across frequency bands, resulting in frame disturbances.

- Delay Variance Detection

From the initial time alignment process, PESQ can detect when delay varies and identify those frames involved. Small frame disturbances during delay variances are canceled to prevent false poor scores.

- Time Alignment Re-Assessed

Another assessment of time alignment is performed for consecutive frame disturbances above a threshold. If time alignment is determined to be inaccurate resulting in large frame disturbances, time alignment is repeated and frame disturbances are recomputed. Perceivable effects of delay variance are preserved in frame disturbances.

- Disturbance Values Aggregation and MOS Prediction

Frame disturbance values and asymmetrical frame disturbance values are aggregated over time in progressive levels. A predicted MOS score is calculated as the linear combination of the average disturbance value and the average asymmetrical disturbance value.

As a result of this computation, the range for predicted MOS scores is actually 0.5 to 4.5. It is noted [10] that in most cases the MOS score predicted by PESQ will be between 1.0 and 4.5.

Comparing Results

Various validation tests have been performed to determine how accurately different objective speech clarity measurements (i.e. PAMS, PSQM+, etc.) correlate with subjective test results (e.g. MOS). These tests produce correlation coefficients that provide a convenient tool for assessing the accuracy of a measurement. For example, for each objective measurement that was evaluated, measurement results were compared to subjective test results to obtain a correlation coefficient. It is important to note that the results of each objective measurement technique vary with different types of networks, impairments, and languages. It can be said that for each type of network or impairment, a certain measurement technique may be more appropriate, but it cannot be said that a single technique is always better. Some of these validation tests are identified here.

COM 12-20: PSQM and PSQM+

COM 12-20-E [8] publishes results of validation tests performed using different databases of speech samples. These validation tests compare the correlation results of PSQM vs. subjective testing and PSQM+ vs. subjective testing. The speech databases contained samples obtained from different types of networks and exhibiting various impairments. It is noted the correlation will vary depending on the type of network under test and the impact of certain impairments, such as time-clipping. It was also noted that caution must be used when comparing results obtained via mathematical algorithms versus subjective human testing ³/₄ the algorithms often weight silence intervals differently than human listeners.

In general, PSQM+ offered an improvement in its correlation to subjective testing for loud distortions and time-clipping from 0.58 to 0.87. PSQM+ offered an improvement in correlation to subjective testing for time-clipping only from 0.85 to 0.97.

COM 12-58: PSQM, PSQM+, and MNB

COM 12-58-E [12] publishes results of validation tests in which ten different databases of speech samples were used, and correlation results of PSQM, PSQM+, and MNB are compared. Databases differed with languages, codecs, and impairments. Different silence interval weighting factors were used, with the best results obtained when a silence interval weighting of 0.0 was used. Correlation results vary with each database, and no method was near-perfect for all databases. PSQM+ offered the best average correlation over all ten databases of 0.85.

PSQM+ provided an overall average correlation of 0.85, with a worst-case correlation of 0.66.

PSQM provided an overall average correlation of 0.80, with a worst-case correlation of 0.60

MNB provided an overall average correlation of 0.75, with a worst-case correlation of 0.17.

It is important to note that these results were over a wide variety of databases. For more specific network conditions and languages, a particular method may provide better correlation than other methods. Refer to COM 12-58-E for the detailed results.

COM 12-D80: PAMS, PSQM, PSQM+, MNB

COM 12-D80-E [13] publishes results of validation tests for PAMS, PSQM, PSQM+, and MNB. Correlation tests used over 16,000 speech samples from 30 subjective listening tests. The document above contains an excellent description of the validation test conditions, how the tests were performed, and the results.

Many different implementations and conditions were tested. Each test was performed for three different silence interval weighting factors for PSQM and PSQM+. Different tests were performed for PAMS Listening Quality Scores and Listening Effort Scores. Each test was performed for two versions of MNB. It is noted that results for MNB vary depending on its implementation, and that the document above should not be used as a guide for MNB performance.

For clean speech samples, PAMS is shown to provide better correlation results than the other algorithms tested. PAMS also provides better results for those cases when listening effort was specifically asked for. When background noise is present in the input signal, PSQM+ provides the best results. PSQM+ provides better results than PSQM in most tests, and a silence interval weighting factor of 0.2 produced the most consistent results. Refer to COM 12-D80-E for the detailed information.

NTIA Report 98-347: SNR, CD, BSD, PSQM, MNB

NTIA Report 98-347 [7] publishes results of tests comparing many traditional metrics with PSQM and MNB. Metrics include:

- Signal-to-Noise Ratio (SNR)
- Segmental Signal-to-Noise Ratio (SNRseg)
- Perceptually Weighted SNRseg (PWSNRseg)
- Cepstral Distance (CD)
- Bark Spectral Distortion (BSD)
- PSQM Noise Disturbance (ND)
- MNB Auditory Distance (AD)

Different tests, including those outside the scope of PSQM, were conducted. In general, SNR and SNRseg performed poorly, with correlations of less than 0.5 on most of the tests. PWSNRseg performed better, but still with correlations of less than 0.65 on all of the tests. CD and BSD performed better than any SNR metric. PSQM ND performed better than BSD on all tests, and better than CD on all but one test, including tests outside the defined scope of PSQM.

When comparing PSQM ND with two variations of MNB, results were very similar. MNB AD performed better than PSQM ND on two tests, and test results nearly the same on the other tests.

Perceptual Evaluation of Speech Quality (PESQ)

The developers of PESQ at British Telecom provide a comparison of PESQ with PSQM and MNB. In these tests, PESQ offers significant improvements over PSQM and MNB, especially for VoIP networks. While further validation tests with PSQM+ and PAMS would be interesting, it is expected that PESQ offers improvements over both of these techniques, since it combines the merits of each.

What's Next?

In addition to the standard and de facto standard methods described in this paper, other proprietary methods for objectively measuring subjective speech clarity are being developed. To enable meaningful comparisons between networks and systems of differing technologies and locales, a single standard method will need to emerge.

PSQM, PSQM+, and PAMS are all relatively new, and further validation in the proving grounds of network testing is needed. PESQ is especially new, but is already anticipated to become a preferred method and standard. PESQ is expected to be approved by the ITU as Recommendation P.862 early in 2001.

Going forth, acceptance of a standard and an establishment of the grounds for its global acceptance will continue to be the domain of the ITU.

It is unlikely that the Internet Engineering Task Force (IETF) will enter the voice clarity/quality picture. The IETF is focused on interoperability and IP protocol systems. Speech quality, particularly *clarity*, is a much broader issue within telecommunications, and thus the domain of the ITU.

Independent of particular standards and methods, the need for an objective measurement of speech clarity continues to grow as next generation networks built on converging technologies and grow exponentially in their deployment and use. Speech clarity, along with other measures of overall voice quality, has become the key differentiating factor in competing networks, systems, and technologies. Objectively assessing this parameter will indeed enable the success of voice services on next generation networks.

Agilent Technologies Clarity Measurement Solutions

Agilent Technologies VQT (Voice Quality Tester) system provides clarity measurement solutions that use both the PSQM+ and PAMS algorithms. This powerful tool allows you to directly and objectively quantify voice clarity on telephony devices and systems either end-to-end using a single test device or end-to-end in a distributed manner using two test devices at separate locations.

The VQT eliminates the need for the large numbers of subjective human listeners that have traditionally been used for this purpose (MOS testing), and provides measurement capabilities appropriate for the conditions found in emerging voice-over-packet environments. VQT clarity measurements can be performed using a single voice/audio sample, performed multiple times over longer time periods for trending, or performed post-process using voice/audio files captured from the network under test.

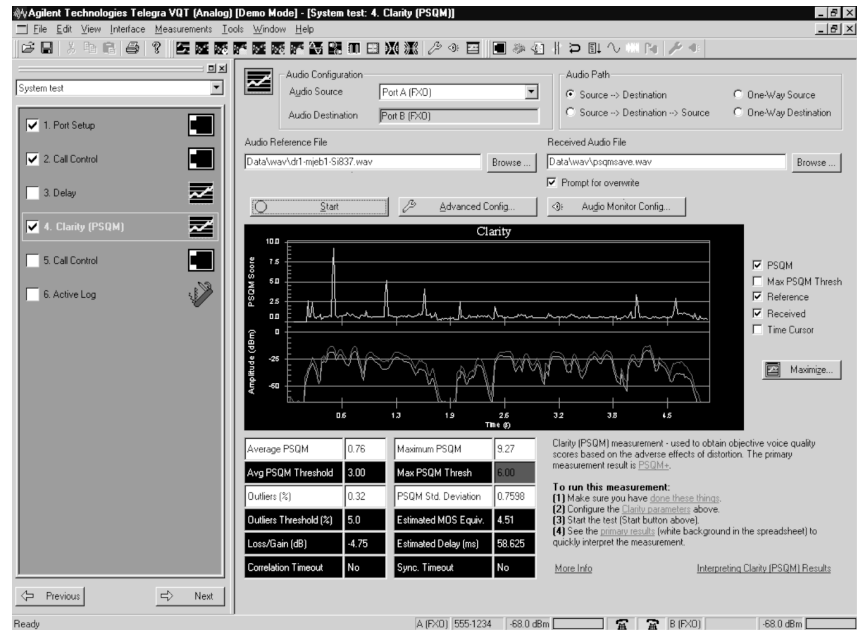


Figure 9. VQT Clarity Measurement

In addition to clarity measurements, the VQT also provides a direct measure of end-to-end delay, echo and echo canceller performance, dual-tone modulation frequency (DTMF) tones, voice activity detector performance, signal loss, and impulse response. It also provides numerous other transmission and simulation tools including file play and record and tone/noise generation to evaluate conditions that affect voice quality. The VQT supports and operates on multiple telephony interfaces including FXO/E&M and T1/E1.

References

- [1] Couch, Leon W., "Digital and Analog Communications Systems", Macmillan Publishing, 1987
- [2] ITU-T Recommendation P.800, "Methods for Subjective Determination of Transmission Quality", International Telecommunication Union, Geneva, Switzerland, August 1996
- [3] ITU-T Recommendation P.830, "Subjective Performance Assessment Of Telephone-Band And Wideband Digital Codecs", International Telecommunication Union, Geneva, Switzerland, February 1996
- [4] J. G. Beerends and J. A. Stemerdink, "A Perceptual Speech Quality Measure Based on a Psychoacoustic Sound Representation", J. Audio Eng. Soc. 42:115-123, March 1994
- [5] ITU-T Recommendation P.861, "Objective Quality Measurement of Telephone Band (300-3400 Hz) Speech Codecs," International Telecommunication Union, Geneva, Switzerland, August 1996
- [6] Atkinson, D.J., "Proposed Annex A to Recommendation P.861", ITU-T Study Group 12 Contribution 24 (COM 12-24-E), International Telecommunication Union, Geneva, Switzerland, December 1997
- [7] Voran, Stephen D., "Objective Estimation of Perceived Speech Quality Using Measuring Normalizing Blocks", NTIA Report 98-347, National Telecommunications and Information Administration, U.S Dept of Commerce, April 1998
- [8] J. G. Beerends, E. J. Meijer and A. P. Hekstra, "Improvement of the P.861 Perceptual Speech Quality Measure", Contribution COM 12-20 to ITU-T Study Group 12, December 1997
- [9] PsyTechnics Group, British Telecommunications, "PAMS Usage Guidelines", February 2000
- [10] J. G. Beerends, A. W. Rix, A. P. Hekstra, M. P. Hollier, "Proposed Draft Recommendation P.862: Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs", Delayed Contribution D.140 to ITU-T Study Group 12, May 2000
- [11] Antony W. Rix, John G. Beerends, Michael P. Hollier, Andries P. Hekstra, "PESQ - the new ITU Standard for End-to-end Speech Quality Assessment", AES 109th Convention, September 2000
- [12] J. G. Beerends and A. P. Hekstra, "Comparison of the ITU-T P.861 PSQM, PSQM+, and MNB Objective Speech Quality Measurement Methods", ITU-T Study Group 12 Contribution 58 (COM 12-58-E), International Telecommunication Union, Geneva, Switzerland, September 1998
- [13] A. W. Rix and M. P. Hollier, "Comparison of Speech Quality Assessment Algorithms: BT PAMS, PSQM, PSQM+ and MNB", ITU-T Study Group 12 Delayed Contribution 80 (COM 12-D80), International Telecommunication Union, Geneva, Switzerland, November 1998

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, including the award winning and successful Telegra VQT (Voice Quality Tester).

John has over ten 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.

Notes _____

**Agilent Technologies’
Test and Measurement Support,
Services, and Assistance**

Agilent Technologies aims to maximize the value you receive, while minimizing your risk and problems. We strive to ensure that you get the test and measurement capabilities you paid for and obtain the support you need. Our extensive support resources and services can help you choose the right Agilent products for your applications and apply them successfully. Every instrument and system we sell has a global warranty. Support is available for at least five years beyond the production life of the product. Two concepts underlie Agilent’s overall support policy: “Our Promise” and “Your Advantage.”

Our Promise

Our Promise means your Agilent test and measurement equipment will meet its advertised performance and functionality. When you are choosing new equipment, we will help you with product information, including realistic performance specifications and practical recommendations from experienced test engineers. When you use Agilent equipment, we can verify that it works properly, help with product operation, and provide basic measurement assistance for the use of specified capabilities, at no extra cost upon request. Many self-help tools are available.

Your Advantage

Your Advantage means that Agilent offers a wide range of additional expert test and measurement services, which you can purchase according to your unique technical and business needs. Solve problems efficiently and gain a competitive edge by contracting with us for calibration, extra-cost upgrades, out-of-warranty repairs, and on-site education and training, as well as design, system integration, project management, and other professional engineering services. Experienced Agilent engineers and technicians worldwide can help you maximize your productivity, optimize the return on investment of your Agilent instruments and systems, and obtain dependable measurement accuracy for the life of those products.

By internet, phone or fax, get assistance with all your Test and Measurement needs.

Online assistance:

<http://www.agilent.com/find/assist>

United States:

(Tel) 1 800 452 4844

Canada:

(Tel) 1 877 894 4414

(Fax) (905) 282 6495

China:

(Tel) 800-810-0189

(Fax) 1-0800-650-0121

Europe:

(Tel) (31 20) 547 2323

(Fax) (31 20) 547 2390

Japan:

(Tel) (81) 426 56 7832

(Fax) (81) 426 56 7840

Korea:

(Tel) (82-2) 2004-5004

(Fax) (82-2) 2004-5115

Latin America:

(Tel) (305) 269 7500

(Fax) (305) 269 7599

Taiwan:

(Tel) 080-004-7866

(Fax) (886-2) 2545-6723

Other Asia Pacific Countries:

(Tel) (65) 375-8100

(Fax) (65) 836-0252

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

©Agilent Technologies, Inc. 2000-2001

Printed in U.S.A. October 22, 2001



5988-2352EN

Use this link to go directly to our network troubleshooting solutions:

<http://www.agilent.com/comms/onenetworks>

