psytechnics
behind every word

# PESQ: An Introduction

## White Paper

Prepared by:
Psytechnics Limited

23 Museum Street
Ipswich, Suffolk
United Kingdom
IP1 1HN

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

September 2001

## Table of Contents

## Overview

PESQ is the new ITU-T standard for measuring the voice quality of communications networks.

This white paper explains the motivation behind voice quality measurement, describes the development of PESQ, and gives an overview of the components that make up the model. In addition the paper describes some applications of PESQ and presents performance results comparing PESQ with earlier models including PAMS and PSQM. Finally the paper gives some typical PESQ scores for a range of common network conditions.

## Voice quality

### Motivation

End-to-end speech quality is the key measure of voice Quality of Service (QoS). Assessment is essential for equipment selection, monitoring, fault-finding, service level agreements and optimisation of networks. Getting quality right can make a major difference both to customer satisfaction and to the cost of providing a service.

Quality in networks will remain an issue as long as bandwidth and processing power are limited. This applies across networks of all types. In mobile networks, bandwidth to the customer is expensive. Quality measurement means that the network can be engineered to deliver the right quality at the right cost. In Voice over IP (VoIP, Internet telephony), performance is also an issue and operators tend to over-provision. Using the right tools to monitor quality can stop over-provisioning and allow networks to service more customers and therefore make more money.

Factors that affect quality include:

- Low bit-rate coding
- Errors (mobile or packet)
- Background noise
- Silence suppression
- Filtering by handsets or the access network

### Measuring voice quality

The traditional method of determining voice quality is to conduct subjective tests with panels of human listeners. Extensive guidelines are given in ITU-T recommendations P.800/P.830. The results of these tests are averaged to give mean opinion scores (MOS) but such tests are expensive and are impractical for testing in the field.

For this reason the ITU recently standardised a new model, PESQ (ITU-T recommendation P.862), that automatically predicts the quality scores that would be given in a typical subjective test. This is done by making an intrusive test, as shown in Figure 1, and processing the test signals through PESQ.

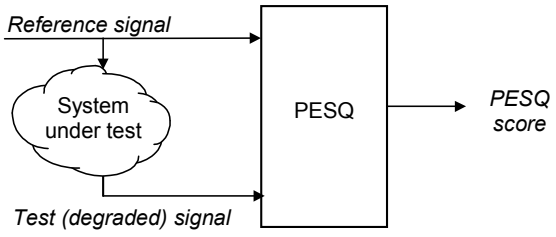


***Figure 1: Use of PESQ***

## Development of PESQ

**Perceptual models for quality assessment**
Modelling perception – specifically human auditory perception – is the core concept behind PESQ and its predecessors. This concept dates back to the late 1970s, when Manfred Schroeder introduced it for speech coding.

Signal compression algorithms, used in modern speech and audio codecs, use perceptual information to decide which parts of a signal to code and which to discard. For example, the MPEG audio codecs use a model of "perceptual masking" to decide how many bits to use for coding each frequency, and which frequencies need not be coded at all.

Simple measures like SNR do not give an accurate measure of the quality of these systems – perceptually masked coding noise, at a typical SNR of 13dB, can be completely inaudible, whereas random noise at the same value of SNR would be extremely disturbing.

Matti Karjalainen first reported the use of a perceptual model for quality assessment in 1985. A perceptual model is used to correctly distinguish between audible and inaudible distortions and this has proven to be the best way of accurately predicting the audibility and annoyance of complex distortions.

Mike Hollier at BT Labs and John Beerends of KPN Research led subsequent innovations in the 1990s on the use of perception for voice quality assessment. Hollier observed that taking account not just of the amount, but also the distribution, of audible distortion could make quality predictions much more accurate. His work was taken up in 1996 by Antony Rix and forms the core of PAMS.

It was not until 1996, following a lengthy international study, that perceptual models for quality assessment were first standardised. The result of this was that Beerends' model, PSQM, became an ITU-T recommendation (P.861) for assessing speech codecs.

**Problems with PSQM**
It soon became clear that PSQM was not suitable for testing networks, where speech codecs are only one part of a complex chain. PSQM was found to correlate very poorly with subjective opinion in some commonly-occurring situations

- speech clipping
- background noise
- packet loss in VoIP networks
- filtering in analogue elements (such as handsets or 2-wire access loops)
- variable delay (common in VoIP).

The extent to which PESQ had problems was illustrated well by one subjective test. The test contained a range of network conditions including filtering and VoIP. The correlation achieved by PSQM against subjective MOS was only 0.26 whereas an ideal model would have a correlation of 1. PESQ, for the same test, has a correlation of 0.93.

**Development of PAMS**
BTs goal was always to produce a model suitable for end-to-end network testing. PAMS was as such designed from the start to include analysis components for level and time alignment. (These are missing from PSQM, had to be provided separately, and could have a significant impact on the model's performance.)

To facilitate this development a large database of subjective tests was assembled. The database contains a very wide range of codecs, errors and packet loss, and noise conditions. It is believed to be the largest of its kind in the world and contains over 25,000 distorted speech recordings and over 1/4 million subjective votes.

Version 1 of PAMS was released in August 1998 and already provided greater performance than PSQM in conditions with noise, codecs or packet loss. It was extended to take account of variable delay in Version 2 which was released in December 1998 – the world's first model suitable for assessing VoIP. Finally version 3, released in December 1999, was the first model on the market able to assess the full range of conditions, including VoIP and analogue networks.

**Standardisation of PESQ as P.862**

In parallel with the development of PAMS, BT and a number of other organisations pressed the ITU-T to select a replacement for PSQM that would be more suitable for testing networks. To this end ITU-T study group 12 held a competition from September 1998 to March 2000. The following companies took part: BT (with PAMS), KPN (with PSQM99, an extended and improved version of PSQM), Ascom, Ericsson and Deutsche Telekom.

The outcome of the competition was a clear division of the models into two groups. The winners were PAMS and PSQM99 but unfortunately there was statistically no single winner. PSQM performed better on certain conditions of rapid gain variations and severe temporal clipping whereas PAMS performed better on conditions of VoIP and filtering.

The second group all had significantly lower average correlation and showed shortcomings on many more of the condition types. PSQM, PSQM+ and MNB had poorer performance still.

It was therefore decided to integrate the best two models, PAMS and PSQM99, to produce a single model that would be a best of breed. For this model to be accepted it was decided by the ITU-T that it would need to outperform both PAMS and PESQ by passing even more demanding performance tests. The BT group (who was by then in the process of creating Psytechnics) collaborated with KPN to achieve this. The result was PESQ.

In May 2000 PESQ passed all of the new performance criteria and was submitted for standardisation as P.862. This process completed in February 2001 with the final approval of P.862 and the withdrawal of P.861.

## How PESQ works

### Algorithm overview

PESQ measures one-way voice quality: a signal is injected into the system under test, and the degraded output is compared by PESQ with the input (reference) signal.

The test signals must be speech-like because many systems are optimised for speech and respond in an unrepresentative way to non-speech signals (e.g. tones, noise, ITU-T P.50). The Psytechnics Artificial Speech-like Test Stimulus (ASTS) is specifically designed for this purpose and is provided with PESQ.

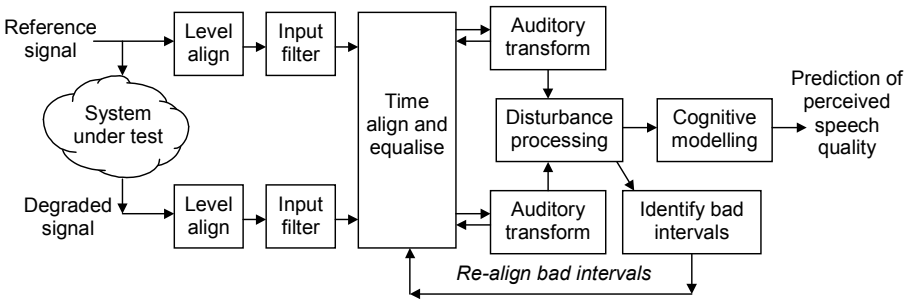


***Figure 2: Structure of PESQ***

The processing carried out by PESQ is illustrated in Figure 2. The model includes the following stages.

Level alignment. In order to compare the signals, the reference speech signal and the degraded signal are aligned to the same constant power level. This corresponds to the normal listening level used in subjective tests.

Input filtering. PESQ models and compensates for filtering that takes place in the telephone handset and in the network.

Time alignment. The system may include a delay, which may change several times during a test - for example Voice over IP often has variable delay. PESQ uses a powerful technique, based on PAMS, to identify and account for delay changes.

Auditory transform. The reference and degraded signals are passed through an auditory transform that mimics key properties of human hearing. This transform removes those parts of the signal that are inaudible to the listener.

Disturbance processing. Disturbance parameters are calculated using non-linear averages over specific areas of the error surface:

- the absolute (symmetric) disturbance: a measure of absolute audible error
- the additive (asymmetric) disturbance: a measure of audible errors that are much louder than the reference

## PESQ outputs

These disturbance parameters are converted to a PESQ score, which ranges from –1 to 4.5. Psytechnics also offer a function to convert this to PESQ-LQ, which gives a P.800 MOS-like listening quality score between 1 and 5 (Table 1).

| Score | Quality of the speech |
|---|---|
| 5 | Excellent |
| 4 | Good |
| 3 | Fair |
| 2 | Poor |
| 1 | Bad |

***Table 1: Listening quality scale***

The Psytechnics release of PESQ also provides extensive diagnostic information computed by the algorithm, such as time-varying delay information, sensation surfaces, and time-varying disturbance values.

## Applications of PESQ

PESQ can be used in a wide range of measurement applications. Being fast and repeatable, PESQ makes it possible to perform extensive testing over a short period and also enables the quality of time-varying conditions to be monitored.

**Codec development.** The impact of changes to a coding algorithm can be quickly investigated using the objective model, even if their effect is small. The model can also be used to explore how quality varies with bit rate, input level or channel errors.

**Equipment selection.** Codecs or other communications systems can be compared using PESQ. For example, PESQ has been successfully used to compare technologies and distortion scenarios for mobile networks, VoIP, and speech codecs.

**Equipment optimisation.** It can be very difficult for a user to find the "correct" values given a choice of coder, input level, bit rate or buffer length. Using an objective model allows the optimum to be found quickly, and is able to work on much smaller differences than could be measured in a conventional subjective test.

**Monitoring.** With a network of test devices to make regular measurement calls, PESQ can be used to benchmark the call quality of communications networks. As well as tracking quality over time or in varying conditions, the model can even help to identify problems before the customers notice.

## PESQ compared with PAMS, PSQM, PSQM+ and MNB

### Correlation scores and error distribution

PESQ was compared with PSQM and MNB models using methodology similar to that used by the ITU-T in the selection of recommendation P.862. See the AES 109th convention paper for more details (reference given below).

The evaluation used correlation coefficient and residual error distribution to quantify the performance of different models at predicting subjective MOS. These metrics are calculated for each subjective test separately. Objective scores are mapped to subjective scores using a monotonic third-order polynomial, aiming to minimise the squared error for that test. This mapping ensures that the comparison is made in the MOS domain whilst allowing for normal variations in subjective voting between tests. Table 2and Table 3 show correlation and residual error distributions for PESQ and other quality assessment models (PAMS, PSQM, PSQM+ and MNB) for the 38 subjective tests that were available to the developers of PESQ. Tests are grouped according to whether conditions were predominantly from mobile, fixed, VoIP and multiple type networks. These included a wide range of simulated and real network measurements.

Table 2 and Table 5 present the results, for PESQ only, of an independent evaluation that was conducted after development was complete.

All of this data relates to subjective listening tests carried out on the Absolute Category Rating (ACR) listening quality opinion scale. Test material consists of natural speech recordings of 8–12s in duration, with four talkers (two male, two female) for each condition. The results are calculated per condition unless otherwise stated.

| No. tests | Type | Corr. coeff. | PESQ | PAMS | PSQM | PSQM+ | MNB |
|---|---|---|---|---|---|---|---|
| 19 | Mobile | average | 0.962 | 0.954 | 0.924 | 0.935 | 0.884 |
| | network | worst-case | 0.905 | 0.895 | 0.843 | 0.859 | 0.731 |
| 9 | Fixed | average | 0.942 | 0.936 | 0.881 | 0.897 | 0.801 |
| | network | worst-case | 0.902 | 0.805 | 0.657 | 0.652 | 0.596 |
| 10 | VoIP/ | average | 0.918 | 0.916 | 0.674 | 0.726 | 0.690 |
| | multi-type | worst-case | 0.810 | 0.758 | 0.260 | 0.469 | 0.363 |

***Table 2: Average and worst-case correlation coefficient for 38 subjective tests known during PESQ development***

| Absolute error range | <0.25 | <0.5 | <0.75 | <1.0 | <1.25 |
|---|---|---|---|---|---|
| % errors in range, PESQ | 74.7 | 93.9 | 99.3 | 99.9 | 100.0 |
| % errors in range, PAMS | 74.4 | 93.3 | 98.3 | 99.7 | 100.0 |
| % errors in range, PSQM | 54.6 | 82.3 | 92.1 | 96.7 | 98.7 |
| % errors in range, PSQM+ | 59.6 | 84.5 | 93.7 | 97.2 | 98.9 |
| % errors in range, MNB | 46.1 | 74.5 | 89.4 | 96.1 | 98.9 |
| Absolute error range | <0.25 | <0.5 | <0.75 | <1.0 | <1.25 |

***Table 1: Error distribution across all 38 known subjective tests.***

| Test | Type | | Corr. coeff. |
|---|---|---|---|
| 1 | Mobile; real network measurements | | 0.979 |
| 2 | Mobile; simulations | | 0.943 |
| 3 | Mobile; real networks, per file only | | 0.927 |
| 4 | Fixed; simulations, 4–32 kbit/s codecs | | 0.992 |
| 5 | Fixed; simulations, 4–32 kbit/s codecs | | 0.974 |
| 6 | VoIP; simulations | | 0.971 |
| 7 | Multiple network types; simulations | | 0.881 |
| 8 | VoIP frame erasure concealment; simulations | | 0.785 |
| | | average | 0.932 |
| | | worst-case | 0.785 |

***Table 4: Correlation coefficient, 8 unknown subjective tests (PESQ only)***

| **Absolute error range** | <0.25 | <0.5 | <0.75 | <1.0 | <1.25 |
|---|---|---|---|---|---|
| % errors in range, PESQ | 72.3 | 91.1 | 97.8 | 100.0 | 100.0 |

***Table 5: Error distribution, 7 unknown subjective tests (PESQ only). Test 3 was excluded from this comparison as data for this test was per-file only.***

### Summary of results

The conclusions of this evaluation can be summarised as follows.

- PESQ has higher accuracy than any other model both on average and in the worst case.
- PESQ is highly robust and gives accurate predictions of quality for a very wide range of conditions.
- PSQM, PSQM+ and MNB all have areas of poor correlation with subjective MOS, in particular with conditions that include VoIP, packet loss, background noise and/or filtering.
- The accuracy of PAMS is close to that of PESQ but there are some situations in which PAMS is not quite as accurate and this is reflected in the worst-case performance.

## PESQ scores for typical network conditions

Based on simulations and real measurements, Table 6 presents the results of a number of typical networks and codecs with no errors or packet loss. In addition, it gives the scores that can be expected in some mobile network conditions where errors are significant.

Please note that results can be affected by a number of factors; for example the test signal used. We averaged the scores from measurements with different speech material in four languages. Each measurement was 8s long and used clean speech. The speech signals at the input to the network were MIRS send filtered and were at an active speech level of –26 dBov.

| Network condition | Typical PESQ score | Typical PESQ-LQ score |
|---|---|---|
| Clean ISDN network | 4.3 | 4.4 |
| Analogue network (G.711) | 4.1 | 4.2 |
| G.728 codec (16kbit/s) | 3.8 | 3.9 |
| G.729 codec (8kbit/s) | 3.6 | 3.7 |
| G.723.1 codec (6.3kbit/s) | 3.5 | 3.4 |
| GSM EFR codec (12.2kbit/s) | 3.9 | 4.0 |
| GSM FR codec (13kbit/s) | 3.5 | 3.5 |
| GSM-EFR mobile network in typical operating range | 3.6 to 3.1 | 3.6 to 2.9 |
| GSM-EFR mobile network in very poor conditions | 2.2 | 1.6 |

***Table 6: Typical PESQ scores for a range of network conditions***

## Glossary

| | |
|---|---|
| EFR | Enhanced full-rate GSM codec |
| FR | Full-rate GSM codec |
| GSM | Global system for mobile |
| MIRS | Modified intermediate reference system, a model of a representative telephone handset (ITU-T recommendation P.830) |
| ITU-T | International Telecommunication Union, Telecommunication standardisation sector |
| ISDN | Integrated services digital network |
| MNB | Measuring normalizing blocks (Appendix II to ITU-T recommendation P.861, January 1998, withdrawn February 2001) (developed by Voran) |
| MOS | Mean opinion score |
| PAMS | Perceptual analysis measurement system (developed by BT) |
| PESQ | Perceptual evaluation of speech quality (ITU-T recommendation P.862, February 2001) |
| PSQM | Perceptual speech quality measure (ITU-T recommendation P.861, January 1997, withdrawn February 2001) (developed by KPN) |
| PSQM+ | An extended version of PSQM developed by Beerends but never standardised |
| SNR | Signal-to-noise ratio |
| VoIP | Voice over Internet Protocol |

## Further reading

Recommendation P.862 is published by the ITU (http://www.itu.int):

Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P.862, February 2001.

The following papers can be obtained from Psytechnics on request:

Rix, A. W. PESQ white paper, May 2001.

Rix, A. W., Beerends J.G., Hollier, M. P. and Hekstra A.P., "Perceptual evaluation of Speech Quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs." IEEE Signal Processing Society International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2001.

Rix, A. W., Beerends, J.G, Hollier, M. P. and Hekstra, A. P. "PESQ – the new ITU standard for end-to-end speech quality assessment". 109th Audio Engineering Society Convention, pre-print no. 5260, September 2000.

Rix, A. W., Bourret, A. and Hollier, M. P. "Modelling human perception", BT Technology Journal, 17 (1), 24–34, January 1999.

Rix, A. W., Hollier, M. P. and Gray, P. "Predicting speech quality of telecommunications systems in a quality differentiated market", 6th IEE Conference in Telecommunications (ICT'98). IEE conference publication 451, 156–160, 1998.

Psytechnics website: http://www.psytechnics.com

psytechnics
behind every word



PRD_PESQ00_XXX_ES0089_02