psytechnics™
voice & video quality assessment

Sample PESQ User Guide

This page has been left intentionally blank.

This page has been left intentionally blank.

Important Information

Document issue

This is Issue 2.1 of the PESQ and PESQ Tools sample user guide for Psytechnics Release 2.1 of PESQ.

Intellectual property rights

Software included in this product is protected by copyright and by European, US, and other patents and is provided under licence from Psytechnics Limited.

Warranty

Psytechnics Limited warrants that it has used reasonable commercial efforts prior to packaging and dispatch to make certain that the media on which the software is delivered is error free. In the event that the Licensee discovers any material errors and notifies Psytechnics Limited of the same within 90 days (warranty period) of receiving the software. Psytechnics Limited will at its option either replace the software or fix any material errors, provided any non-compliance has not been caused by any modification, variation or addition by the Licensee. In no circumstance will the existence of any errors constitute a breach of the Licence Agreement.

In addition, Psytechnics Limited warrants that it has used reasonable commercial efforts in the production and dispatch of Documentation and/or Manuals relating to the software. In the event the Licensee discovers a material error and notifies Psytechnics Limited of the same within 90 days (warranty period) of receiving the Documentation and/or Manual. Psytechnics Limited will at its option either replace the Documentation and/or Manual or correct the material error.

The Licensee acknowledges that any and all copyright, trademark and other intellectual property rights subsisting in or used in connection with the software including any Documentation and/or Manual relating thereto are and shall remain the property of Psytechnics Limited and the Licensee shall not during or after expiry or termination of this Agreement in anyway question or dispute the ownership of the Documentation and/or Manuals relating to the software.

Copyright

Trademarks

PESQ™, PESQ Tools™, Psytechnics™ are trademarks of Psytechnics Limited.

Product and company names mentioned herein are trademarks or trade names of their respective companies.

Contact

Psytechnics Limited, Fraser House, 23 Museum Street, Ipswich IP1 1HN, United Kingdom
Tel. +44 (0) 1473 261 800 Fax. +44 (0) 1473 261 880
E-mail: info@psytechnics.com Web: http://www.psytechnics.com

This page has been left intentionally blank.

This page has been left intentionally blank.

Contents

Figures

Tables

This page has been left intentionally blank.

This page has been left intentionally blank.

1. Introduction

1.1 About this document

This document is an overview and user manual for the Psytechnics distribution of PESQ (Perceptual Evaluation of Speech Quality). Licensees of Psytechnics Ltd may customise it for their own end-users, in accordance with their licences and the guidelines provided at the end of this document.

The guidelines should be read by all Licensees of the Psytechnics distribution of PESQ (Perceptual Evaluation of Speech Quality). They contain notes on creating end-user documentation for products that include PESQ or PESQ Tools.

The following documentation is also available on PESQ and PESQ Tools:

- **PESQ and PESQ Tools Code documentation:** Contains detailed documentation of the PESQ and PESQ Tools code and API. It is intended for use by engineers integrating PESQ and PESQ Tools into an end-user product.

1.2 A guide to this document

This sample User Guide is divided into four main divisions:

- The main User Guide (page 15)
- A section of background and advanced material (page 47)
- Supplementary sections, including references and a glossary (page 61)
- Guidelines on how to use the sample information in this document to create end-user documentation for different types of products that include PESQ or PESQ Tools (page 65)

The main User Guide contains three sections:

- Section 2 covers the use of PESQ as a simple measurement device, which returns only a quality score.
- Section 3 covers use as an advanced speech quality analyser, with a full set of features and outputs for use by trained individuals. This section includes descriptions of the diagnostic outputs provided by the PESQ Tools option.
- Section 4 has material specific to the use of Psytechnics PESQ for evaluating Head and Torso (HATS) measurements or wideband telephony. This is an extension to the P.862 standard.

The following sections cover background and advanced material that should be read for specific purposes.

- Section 5 contains instructions for creating speech signals for testing.
- Section 6 summarises techniques used for designing and conducting subjective listening tests, the quality benchmark that PESQ is designed to model.
- Section 7 provides guidance on testing the performance of systems in the presence of background noise.
- Section 8 outlines the methods use to compare objective and subjective scores.
- Some results on the performance of PESQ calculated according to these methods are presented in section 9.

The supplementary material includes:

- References for further reading in section 10
- A glossary of technical terms in section 11

The guidelines includes:

- Classification of different PESQ usage profiles
- Notes on how to use the sample documentation for different application profiles

1.3 Legal notice

Performance results reported in this documentation represent actual results obtained by Psytechnics. Psytechnics does not warrant that the indicated results will be obtained in every test scenario. All warranties with respect to PESQ remain as stated in the applicable licence agreement between Psytechnics and the Licensee. Nothing in this document is to be interpreted as varying the terms of the licence agreement, either expressly or by implication.

This page has been left intentionally blank.

This page has been left intentionally blank.

User Guide

2. PESQ as simple measurement device

2.1 Overview of PESQ

Modern communications networks include elements that cannot reliably be assessed by such conventional engineering metrics as signal-to-noise ratio. Examples of such elements include lossy coding, error-prone channels and voice activity detection. One way to measure customers' perception of the quality of these systems is to conduct a subjective test involving panels of human subjects. However, these tests are expensive and unsuitable for such applications as real-time monitoring.

PESQ provides an objective measure that predicts the results of subjective listening tests on telephony systems. To measure speech quality, PESQ uses a sensory model to compare the original, unprocessed signal with the degraded version at the output of the communications system. This process is shown in Figure 1 and is explained in more detail in the next section.

The result of comparing the reference and degraded signals is a quality score. This score is analogous to the subjective "Mean Opinion Score" (MOS) measured using panel tests according to ITU-T P.800. The PESQ scores are calibrated using a large database of subjective tests.

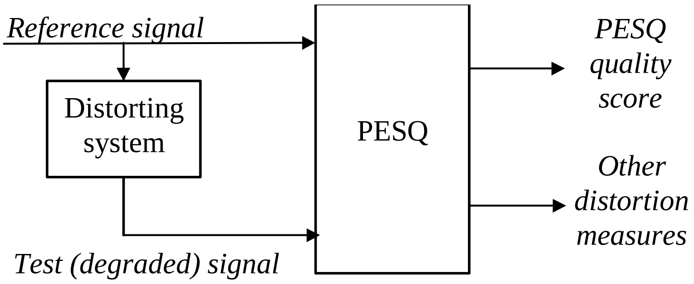
Optionally, PESQ can be used to provide other diagnostic information if required.

PESQ incorporates many new developments that distinguish it from earlier models for assessing codecs, for example, PSQM and MNB (ITU-T P.861). These innovations allow PESQ to be used with confidence to assess end-to-end speech quality as well as the effect of individual elements such as codecs.

This release of Psytechnics PESQ provides a fully conformant implementation of PESQ as defined in ITU-T P.862. Additionally, it provides extensions to allow PESQ to be used with wideband telephony or head and torso simulator (HATS) measurements.

The ITU-T selection process that resulted in the standardisation of PESQ involved a wide range of conditions, with demanding correlation requirements set to ensure that it has good performance in assessing conventional fixed and mobile networks and packet-based transmission systems.

Figure 1: Using PESQ



PESQ takes into account the following sources of signal degradation: coding distortions, errors, packet loss, delay and variable delay, and filtering in analog network components.

PESQ does not take into account the subjective effect of level changes in the network, echo, and the effect of round-trip delay on conversation.

2.2 Inputs

2.2.1 Speech signals

PESQ requires two inputs: the original, the unprocessed test signal and the degraded version that has been passed through the distorting system. In addition, the model needs to know the sampling rate of these files, which may be either 8kHz or 16kHz.

The test signal should be speech-like. This is important, because such technologies as codecs are designed to transmit speech. Simple synthetic signals such as sine waves or white noise may not give results that relate to customers' perception of the system's speech quality.

The reference signal should be filtered with an appropriate send filter before injection in a network under test. This will usually be a modified IRS Send filter [ITU-T P.48]. This filtered reference signal should then be used as an input to the PESQ algorithm.

We recommend that you use the Psytechnics artificial speech-like test signal (ASTS), which is available as an optional addition to PESQ. This reproduces the key temporal, spectral and sequence properties of speech with less redundancy than natural speech, allowing greater confidence with shorter measurements. If you intend to use natural recorded speech, you should first read Section 5. Care should be taken to ensure that the signal has been filtered and is at the correct level before entering the network, so that it is representative of signals transmitted from a telephone handset.

2.2.2 Sampling rate

PESQ is able to process input material at 8kHz or 16kHz sample rates. The 8kHz version of PESQ is faster and requires less memory than the 16kHz version. Both input files must be at the same sample rate. In certain applications, the sampling rate for the PESQ application may be fixed. If the sampling rate can be changed, it is essential that the correct value is specified.

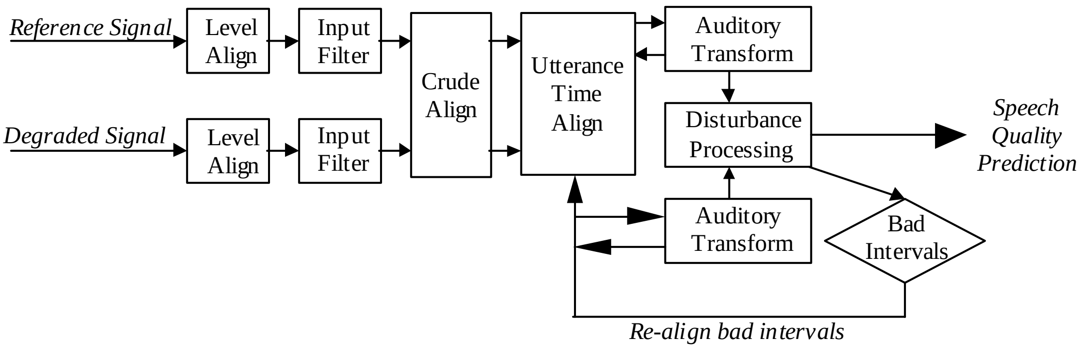
2.2.3 Model specification

Release 1.4 introduced a small modification to the perceptual model that may lead to small changes in PESQ score. This improves the performance of PESQ for cases where the reference signal is very quiet during silent periods. If it is essential to obtain scores that exactly match those obtained by previous releases of PESQ, the version 1.0 model can be selected using the appropriate switch. We recommend that the release 1.4 method is used by default. See section 3.2.1 for more details.

2.3 Operations performed by PESQ

The processing carried out by PESQ is illustrated in Figure 2.

Figure 2: Processing performed in PESQ



The model includes the following stages.

Level alignment. In order to compare the signals, the reference speech signal and the degraded signal should be at the same, constant power level. This is necessary because the reference signal does not have to be to be at a defined level and because the gain of the system under test is unknown before testing.

PESQ assumes that the subjective listening level is a constant 79dB SPL at the ear reference point [ITU-T P.830, section 8.1.2]. A gain is applied to both the reference and degraded signals to bring them to this level.

Input filtering. PESQ models the receive path of the telephone handset using an input filter. This takes account of the effect of the electrical and acoustic components of the handset. The filter used is similar to the IRS receive characteristic [ITU-T P.48].

Time alignment. The system under test may include a delay, which may be variable. In order to compare the reference and degraded signals, they need to be lined up with each other. PESQ applies voice activity detection to the signals to identify those parts of the signal that are speech, ignoring noise.

Time alignment is then done in three stages:

- First, PESQ aligns the overall speech signals (utterances). An utterance is a continuous speech burst identified by the voice activity detector, that does not contain pauses longer than a pre-determined threshold (200ms). This process detects delay over major sections of the degraded signal compared to the reference signal.
- Second, PESQ aligns overlapping sections of the speech (frames). This process detects delay that is variable over the length of an utterance, as this can be significant in packet-based networks.
- The third stage does not occur immediately after the second stage, but is performed after the auditory transform has been calculated. The third stage realigns “bad intervals” (sections of the speech with very large disturbance), and improves the model’s accuracy with a small number of files where delay changes are not correctly identified by the initial time alignment process.

Auditory transform. In order to compare the reference and degraded signals, taking account of how a listener would have heard them, each is passed through an auditory transform that mimics certain key properties of human hearing. This gives a representation in time and frequency of the perceived loudness of the signal, known as the sensation surface.

Equalisation. Part of the auditory transformation equalises certain processes that have little subjective effect. First, the transfer function of the system is estimated, and is used to equalise the reference to the degraded in the auditory transform domain. This takes account of filtering in analogue components of the network such as telephone handsets. Second, the frame-by-frame amplitude gain of the system is estimated and used to equalise the auditory transform of degraded file to the reference. In both cases the equalisation is partial – large amounts of filtering or gain variation are not cancelled, and therefore result in errors being measured.

Disturbance processing. The difference between the sensation surfaces for the reference and degraded files is known as the error surface; this shows any audible differences introduced by the system under test. The error surface is analysed by a process that takes account of the effect that small distortions in a signal are inaudible in the presence of loud signals (masking).

From the positive and negative errors, two disturbance parameters are calculated. They are calculated as non-linear averages over specific areas of the error surface. These disturbance parameters are:

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

This analysis gives two error parameters that summarise the amount of each type of audible error. Finally, the error parameters are converted to a quality score, which is a linear combination of the average symmetric disturbance value and the average asymmetric disturbance value.

2.4 Quality scores

This release of PESQ returns three quality scores:

- PESQ score is calculated according to P.862
- PESQ-LQ gives a quality score on a MOS-like scale
- P.862.1 is the ITU-T standard mapping for PESQ to MOS-like scale
- PESQ-Ie is the impairment factor, *Ie*, which is an input to the E-model

The PESQ-LQ, P.862.1 and PESQ-Ie scores are derived from the PESQ score using simple formulae. The PESQ-LQ mapping was developed by Psytechnics; the P.862.1 mapping is defined as an ITU-T recommendation relating directly to PESQ; the PESQ-Ie mapping is defined in ITU-T Recommendation P.834.

PESQ score, PESQ-LQ and P.862.1 are output in the file `pesqlog.txt`, and are quoted to two decimal places.

2.4.1 PESQ score

PESQ returns a quality score, known as PESQ score, which conforms to ITU-T P.862. PESQ score lies on a scale from –0.5 to 4.5, though in most cases it is between 1 and 4.5. PESQ score correlates with subjective quality.

2.4.2 PESQ to MOS mappings

It has been found that PESQ score is consistently higher than subjective MOS for poor quality conditions. In order to deliver an objective MOS score which is more closely aligned with subjective MOS, a simple mapping can be applied. This mapping aligns the PESQ output scale to the subjective test scale obtained from ITU-T P.800 listening quality tests.

This is reproduced in Table 1 along with the prompt that is given to subjects. Listening quality scores lie between 1 and 5. PESQ-LQ score lies between 1.0 and 4.5. This is because 4.5 is usually the maximum obtained in a subjective test.

Table 1: Listening quality scale

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

The score gives a measure of customers' perception of quality. The highest score, 4.5, means that no distortion is measured. As the amount of distortion increases the quality falls. For more information on how to compare PESQ scores to subjective test data, see section 8.

2.4.3 PESQ-LQ

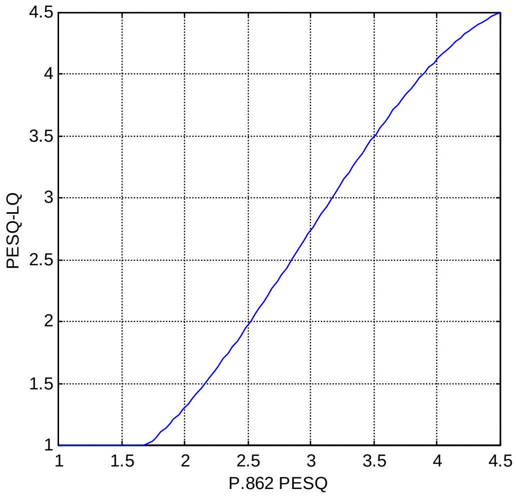
Psytechnics have analysed this using a very large number of subjective tests. To make it easier to compare PESQ score with MOS, a second quality value, PESQ-LQ, has been introduced.

PESQ-LQ scores are closer to the listening quality subjective opinion scale, which is standard in the industry and is defined in [ITU-T P.800].

2.4.4 Relationship between PESQ score and PESQ-LQ

The function which is used to calculate PESQ-LQ is shown in Figure 3.

Figure 3: Mapping from raw PESQ score to PESQ-LQ



The mapping from PESQ score to PESQ-LQ can be computed as follows:

if $pesq\_score < 1.7$ then $pesq\_lq = 1.0$

else $pesq\_lq =$

$$-0.157268\, pesq\_score^3 + 1.386609\, pesq\_score^2 - 2.504699\, pesq\_score + 2.023345$$

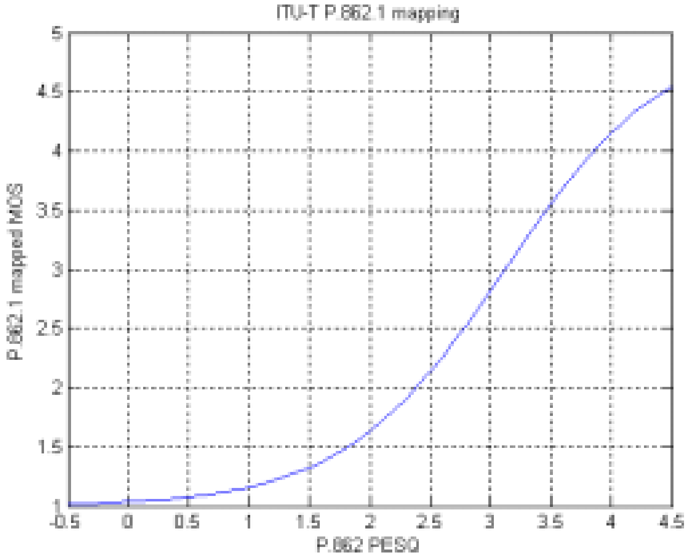
2.4.5 P.862.1

The ITU has standardised a universal PESQ to MOS mapping. This was created from a shared pool of subjective test results covering wireless, VoIP, fixed and codec-only conditions, including Japanese, British English, American English, French, German, Italian, Swedish, Dutch and Finnish.

2.4.6 Relationship between raw PESQ score and P.862.1

This mapping is continuous from PESQ –0.5 to 4.5 and MOS 1 to 4.55. It takes the form of a logistic with 4 parameters, and is shown below:

Figure 4: Mapping from raw PESQ score to P.862.1 MOS



The mapping from PESQ score to P.862.1-MOS can be computed as follows:

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945 * x + 4.6607}}$$

The graph of the P.862.1 function is presented in Figure 4.

For more information on this mapping, please see ITU-T recommendation P.862.1.

2.4.7 Typical quality scores

Based on simulations and real measurements, Table 2 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.

Table 2: Typical PESQ scores for a range of conditions

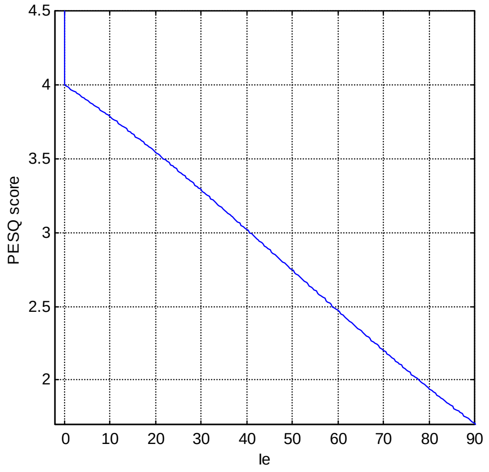
Network condition	Typical PESQ score	Typical PESQ-LQ score
Clean ISDN network	4.3	4.4
Analogue network (G.711)	4.1	4.2
G.728 codec (16kbit/s)	3.8	3.9
G.729 codec (8kbit/s)	3.6	3.7
G.723.1 codec (6.3kbit/s)	3.5	3.4
GSM EFR codec (12.2kbit/s)	3.9	4.0
GSM FR codec (13kbit/s)	3.5	3.5
GSM-EFR mobile network 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

Note: 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.

2.4.8 PESQ-Ie mapping

The PESQ-Ie score is the impairment factor, *Ie*, which is an input to ITU-T G.107 E-model. The PESQ-Ie score uses a scale from 0 to 140, and is calculated from the PESQ score using the relationship shown by the following graph.

Figure 5: Mapping between PESQ score and PESQ-Ie.



The mapping from PESQ score to PESQ-Ie is defined in ITU-T Recommendation P.834.

2.4.9 PESQ Usage warnings

These warnings are designed to indicate when the scores returned by P.862 maybe unreliable. Psytechnics has implemented these warnings in the pesqmain.c, main module. This section of the code can be used in its original form for standalone executables, or as an example of how to generate the warning for a PESQ library build.

Possible time alignment failure warning

There are certain situations – for example when the degraded file does not originate from the reference and therefore contains different speech or just circuit noise – where the time alignment in PESQ will incorrectly estimate the delay between the reference and degraded signals. When this happens it is possible that PESQ will returns scores that are inappropriately high. The following test is used to assess whether the reference and degraded files may not be related:

Delay confidence < 0.3 and standard deviation of delay > 0.05 and raw PESQ score > 1.5.

Standard deviation of delay is computed in units of seconds from the frame-by-frame delay used in PESQ.

File duration problem

PESQ has been validated in the ITU-T for use with signals up to 30 seconds. Due to the precision available to the floating point arithmetic in PESQ, once the signals being processed reach a certain length errors will start to be introduced in the signal energy calculation. From our analysis, it was found that signals with more than about 1 million samples will start to cause problems. 60 seconds of 16kHz mono signal contains 960,000 samples and this would be a sensible threshold at which to apply a warning. If the signal is at 8kHz then potentially twice the length could be used. However since P.862 has only been validated up to 30 seconds, two separate warnings should be displayed, one if the reference signal length exceeds 35 seconds and a second if the number of samples in the reference or degraded exceeds 960,000.

Potential level alignment problems

This issue has two effects depending on whether an utterance has been deleted or added to the degraded, and whether a large amount of silence padding has been added to the degraded.

When an utterance has been deleted from, or a large amount of silence padding has been added to the degraded signal, the signal will be level shifted to a value above the optimum.

When an utterance has been added to the degraded signal, the signal will be level shifted to a value below the optimum.

These will have an effect on the amount of disturbance measured in the degraded signal and will therefore effect the PESQ score. This issue be addressed by displaying a warning if the degraded and reference signals vary in length or active speech level by more than 20%.

Speech activity warning

P.862 states that the speech activity in a test signal to be used with PESQ should be between 40% and 80%. A low speech activity could cause the PESQ score to be inaccurate. Although the typical speech activity for a test signal can vary depending on the language used in the signal, A warning should be shown if PESQ detects that the speech activity in the reference or degraded is below 35% or above 85%.

3. Advanced use

Before reading this section, all users should read the material of Section 2. Advanced applications may offer a full set of features and outputs for use by trained individuals. This section includes descriptions of the diagnostic features available in the PESQ Tools option.

3.1 Input options

The input options are:

- Sampling rate (section 2.2.2)
- Choice of version 1.4 or 1.0 models (sections 2.2.3 and 3.2.1).
  Psytechnics recommends the version 1.4 model.

3.2 Outputs

The following section describes the outputs that may be obtained from PESQ or PESQ Tools. Plots of some outputs are provided giving example results from interesting network conditions.

Notation

A number of outputs are returned using the dBov scale. This is defined such that a square wave, of amplitude equal to the maximum possible value of a 16-bit PCM signal, has a level of 0dBov. A difference between two dBov quantities has the units dB.

In some cases, a value in dBov or dB cannot be computed, for example if the degraded file contains digital silence. In these cases a value of –999.0 is returned.

The following outputs are available in PESQ and PESQ Tools:

- Frame-by-frame delay (section 3.2.2)
- Transfer function and signal spectra (section 3.2.3)
- Perceptual parameters (section 3.2.4)
- Frame-by-frame score (section 3.2.5)
- Signal waveforms (section 3.2.6)
- Sensation surfaces (section 3.2.7)
- Error surface (section 3.2.8)

The following outputs are only available in PESQ Tools:

- Frame-by-frame delay statistics (section 3.2.9)
- Utterance-by-utterance delay measures (section 3.2.9)
- Signal level and gain (section 3.2.9)
- Utterance-by-utterance level (section 3.2.11)
- Bark signal spectra (section 3.2.13)
- Linear spectra (section 3.2.14)
- Transfer function estimation (section 3.2.15)
- Signal spectrograms (section 3.2.16)
- LPC excitation (section 3.2.17)
- Speech activity related outputs (section 0)
- Speech outputs (section 3.2.19)

3.2.1 PESQ Score

In release 1.4, a small modification was introduced to the PESQ perceptual model, which affects PESQ score in some cases. The new model gives identical scores to the old model in most circumstances where natural speech recordings are used. However, the new model gives higher, and more accurate, scores in cases where the reference signal is very quiet during silent periods, for example if it includes digital silence. In these cases the difference in PESQ score between the two models has been found to be as large as 0.25.

Psytechnics recommend that for normal network measurement purposes, the new model introduced in PESQ release 1.4 should be used. However there may be circumstances for which the results obtained with previous versions of PESQ must be reproduced exactly. For these cases the old (backwards-compatible) model may be used by making the appropriate switch. The default option is the PESQ release 1.4 model.

3.2.2 Frame-by-frame delay

An overview of the PESQ time alignment operation is given in section 2.3. It generates two sets of results, the utterance-by-utterance and frame-by-frame delay values.

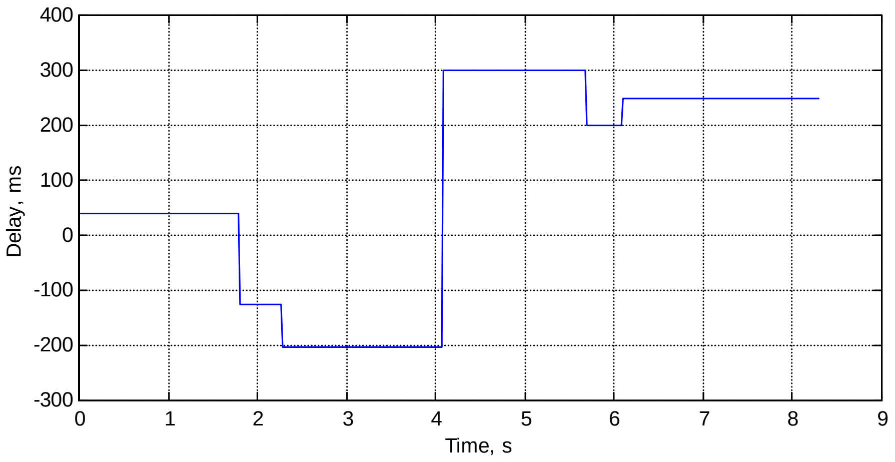
Frame-by-frame delay is the delay measure used in calculating the PESQ quality score. Utterances are broken up into frames of 32 ms duration. Frames use a window function that gives greater weight to the central 16ms of each frame, and there is an overlap between successive frames of 50%. Effectively, therefore, each frame is 16ms long; this can be thought of as “sampling” the values every 16ms. PESQ calculates the delay in each frame, based on the nearest utterance.

Because it models the processing used in PESQ, the frame-by-frame delay is the best way of tracking how delay varies during the signal.

Delay changes are most likely to be caused by jitter buffer adaptation in VoIP telephony edge devices. This adaptations occur when there is a large change in the jitter on an IP network. As jitter on the VoIP network increases, the delay measured by PESQ Tools will typically increase as the jitter buffer grows in size. As the jitter decreases, the delay measured will typically decrease as the jitter buffer decreases in size.

Figure 6 plots the frame-by-frame delay for the same condition as shown in Figure 13.

Figure 6: Frame-by-frame delay

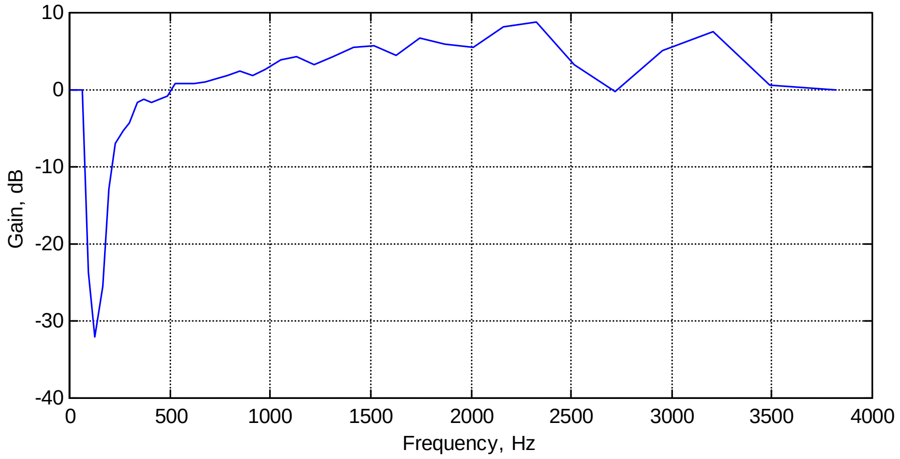


3.2.3 Bark scale transfer function

The system's transfer function, in dB, is estimated for each of the 42 perceptual frequency bands at 8kHz sample rate (49 bands at 16kHz sample rate). A typical transfer function is shown in Figure 7.

Note that the transfer function is calculated after level alignment has been performed. Constant gain in the system under test will therefore not be shown in the transfer function estimate. The overall dB gain of the system can be found using the signal level measures (section 3.2.9).

Figure 7: Transfer function



3.2.4 Perceptual parameters

PESQ computes two parameters that describe the amount and distribution of audible errors:

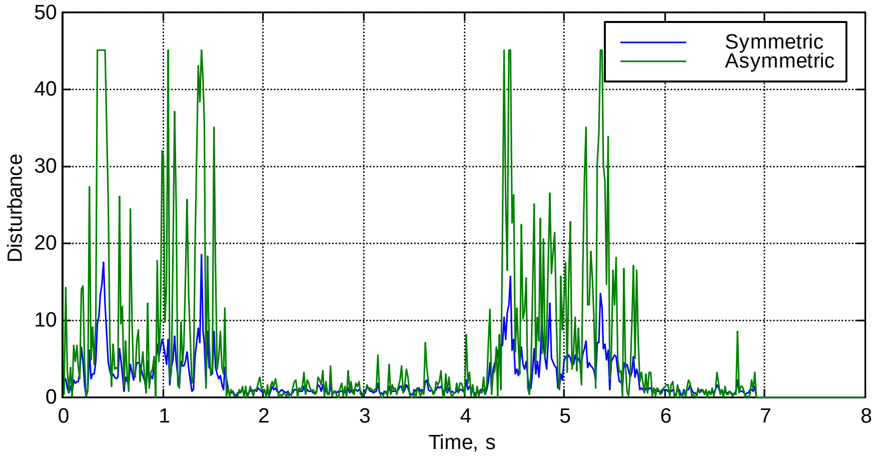
- Symmetric disturbance
- Asymmetric disturbance

These values are returned both frame-by-frame and as averages.

Both types of disturbance range between 0 (no distortion) and 45 (maximum).

An example plot of the frame-by-frame disturbance parameters is shown in Figure 8. Note that PESQ usually ignores the silent periods at the start and end of any signal, which is why both disturbance values go to zero at the end of this example.

Figure 8: Frame-by-frame disturbance



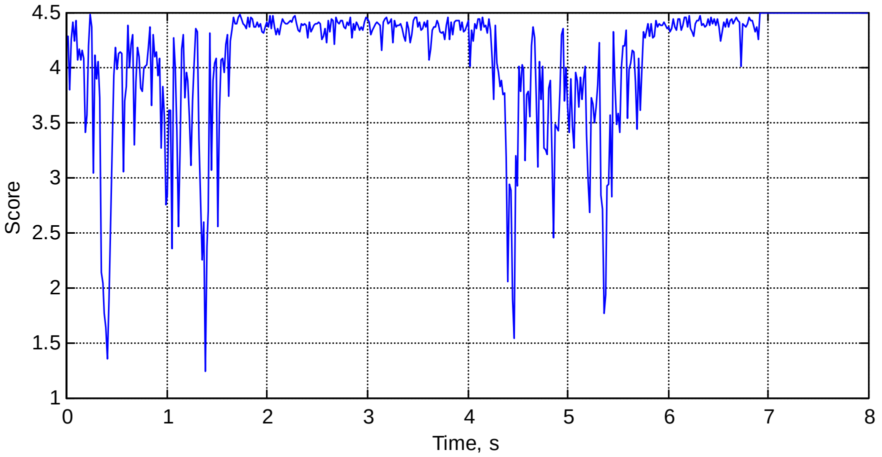
3.2.5 Frame-by-frame score

Frame-by-frame quality score is calculated from the frame-by-frame symmetric and asymmetric disturbance values, to provide a simpler way to interpret distortions.

An example of frame-by-frame score is shown in Figure 9, corresponding to the same condition as Figure 8. Note that the PESQ score is not a simple average of the frame-by-frame score. A complex non-linear averaging process is applied separately to obtain the average symmetric and asymmetric errors, and the PESQ score is derived from these.

Frame-by-frame score should only be taken as a rough guide to the location and relative magnitude of distortions – it is not meaningful to talk about subjective quality on such short time intervals.

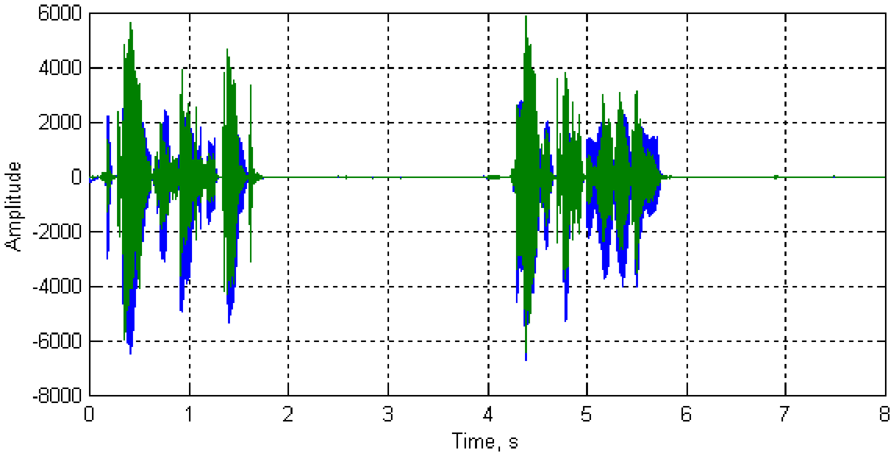
Figure 9: Frame-by-frame score



3.2.6 Signal waveforms

The signal waveforms plot the amplitude of each signal over time, as shown in the example in Figure 10.

Figure 10: Signal waveforms



3.2.7 Sensation surfaces

The sensation surfaces show the perceived loudness, on the Sone scale, of the signals in time and frequency. The frequency scale is a modified Bark scale, and time interval between successive samples is 16ms. The sensation surfaces are very useful, clearly showing the content of the signals.

The sensation surfaces are available both pre-equalisation (before either transfer function equalisation or equalisation for time-varying gain have been applied) and post-equalisation. The error surface and the PESQ disturbance parameters are calculated post-equalisation has been applied.

An example sensation surface is shown in Figure 11.

Figure 11: Degraded sensation surface

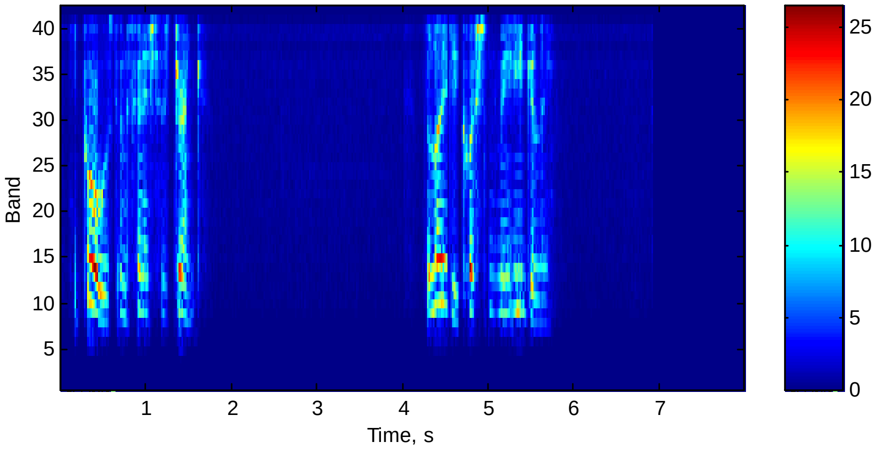
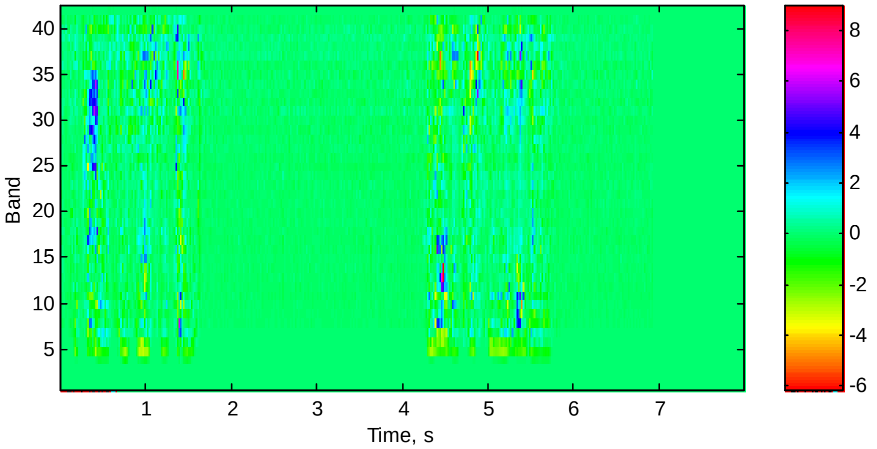


Figure 12: Error surface



3.2.8 Error surface

The error surface is the degraded sensation surface minus the reference sensation surface. This means that errors that have added to the signal (for example, noise) have positive values, while parts of the signal that have been attenuated or muted have negative values. The amplitude of errors is related to how audible and annoying they will be.

Examples of errors that may occur are listed here.

- **Front-end clipping** causes large but short negative errors at the start of speech bursts.
- **Muting** can be seen as prolonged negative errors during speech, where the degraded sensation surface falls close to zero.
- **Addition of background noise** shows up as positive error, and is most obvious in silent periods.
- **Coding distortion** generally causes low-level errors throughout speech bursts, although this is very codec-dependent.
- **Bit or frame errors** tend to cause localised distortion, which may be positive or negative. This effect is dependent on the codec and any error concealment algorithm used.

An example error surface is shown in Figure 12.

3.2.9 Frame-by-frame delay statistics

PESQ Tools Only.

PESQ Tools provides statistics for the frame-by-frame delay values described in section 3.2.2. These statistics are:

- mean delay
- maximum delay
- minimum delay
- standard deviation of delay
- delay histogram

The histogram comprises ten uniformly spaced bins, which are optimised to best reflect the spread of the data.

3.2.10 Utterance-by-utterance delay measures

PESQ Tools Only.

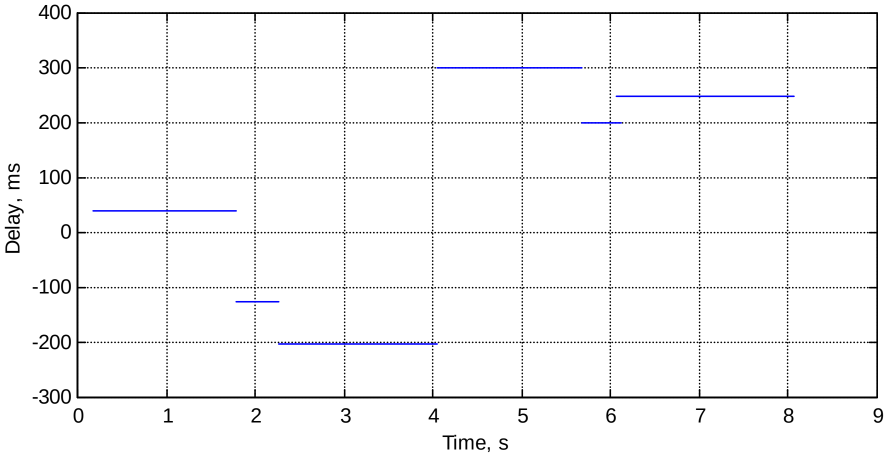
An overview of the PESQ time alignment operation is given in section 2.3. It generates two sets of results, the utterance-by-utterance and frame-by-frame delay values.

In order to deal with variable delay, PESQ sub-divides the signal into a number of utterances. Each utterance is time-aligned separately. The calculation returns, for each utterance:

- the estimated delay in samples
- a delay confidence between 0 (no confidence) and 1 (full confidence)
- the utterance start sample index
- the utterance end sample index

These quantities enable the variation of delay throughout the recording to be plotted. An example is shown in Figure 13. The utterance-by-utterance results are a preliminary set of values and the frame-by-frame delay values (section 3.2.2) are the values actually used in calculating the quality score.

Figure 13: Utterance by utterance delay



3.2.11 Utterance-by-utterance level

PESQ Tools Only.

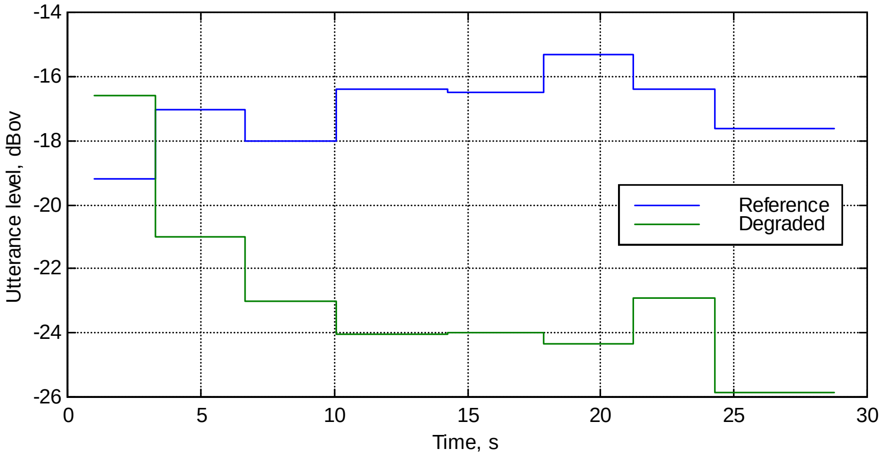
To analyse the effect of time-varying processes, such as automatic level control, PESQ Tools includes measurements of the active speech level of each speech utterance in the reference and degraded signals. This is given in dBov.

By comparing the level of each utterance in the reference and degraded signals, it is clear if the gain is changing during the measurement. Gain variation can appear as a consequence of any of the following:

- automatic level control (ALC)
- dynamic noise reduction
- strong filtering (e.g. in an analog connection).

An example plot showing the effect of ALC is shown in Figure 14.

Figure 14: Utterance-by-utterance level



3.2.12 Signal level and gain

PESQ Tools Only.

PESQ Tools provides various measures of level (amplitude), which are calculated separately for the reference and degraded signals. The measures, and a description of each, are shown in Table 3. From these values, some additional quantities are derived, which are shown in Table 4.

Note 1. For computing most PESQ Tools parameters such as level, spectrum, transfer function and speech parameters, the degraded signal is aligned in time with the reference signal. The computation is based on the parts of the two signals that overlap. This means that some measures of the degraded signal will give slightly different results from measures computed without this time alignment. However, this process makes it much more convenient to compare the reference and degraded signals.

Note 2. Level measures are computed using a voice activity decision based on the reference signal VAD. This can produce different results from a VAD applied only to the degraded signal, for example if the addition of noise alters the classification of speech and noise. Voice activity decision is sometimes ambiguous, so you may encounter unexpected results with the MNL of the reference and degraded signals if the reference signal is hard to classify.

Table 3: Signal level measures calculated separately for reference and degraded signals

Measure	**Units**	**Meaning**	**Typical value**	**Typical range**
Active speech level (ASL)	dBov	Power (RMS) level during speech active periods	–26	(–35, –15)
Mean noise level (MNL)	dBov	Power (RMS) level during silent periods only	–70 (clean speech)	(–80, –15)
RMS mean level	dBov	Power (RMS) level of the entire signal	–30	(–40, –15)
Estimated signal-to-noise ratio (SNR)	dB	The relative loudness of speech to noise, i.e. ASL–MNL.	45 (clean speech)	(10, 60)
DC offset	PCM units	The DC offset of the input signal	0	(–32, 32)

Table 4: Level measures of the system under test

Measure	**Units**	**Meaning**	**Typical value**	**Typical range**
Insertion gain	dB	Power gain of the system under test. Calculated as (ASL of degraded signal) minus (ASL of reference signal).	0 (digital) –12 (analog)	(–20, 6)
Noise gain	dB	Gain calculated for noise in silent periods. Calculated as (MNL of degraded signal) minus (MNL of reference signal). May differ from the system gain if noise is added or suppressed.	0	(–20, 20)

3.2.13 Bark signal spectra

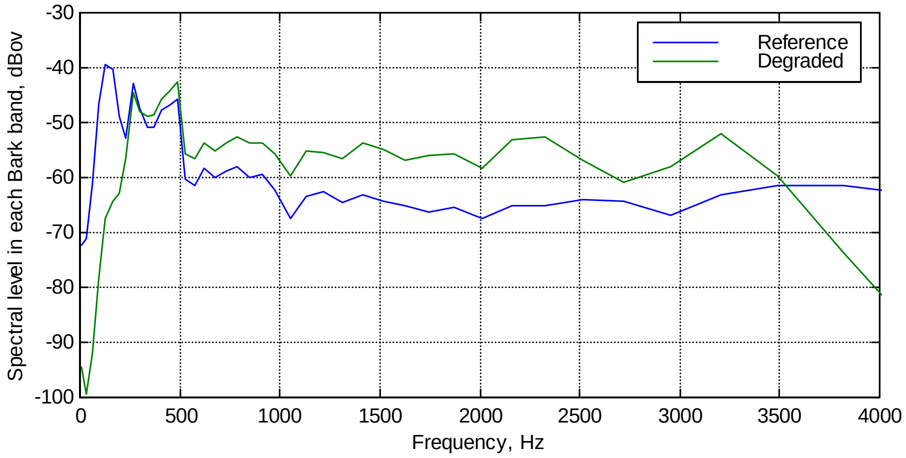
PESQ Tools Only.

The Bark signal spectra are calculated using the Bark frequency scale. These measures can be used to compare the spectrum of different signals and compare speech and noise. The spectra returned, for the reference and degraded signals, are as follows:

- speech spectrum (speech active periods only)
- noise spectrum (silent periods only)
- average spectrum of the whole signal.

The level of each frequency band is in dBov. The centre frequency in Hz of each Bark band is also returned and this can be used to plot the data on a linear frequency scale, as shown in Figure 15.

Figure 15: Speech spectrum of reference and degraded signals



3.2.14 Linear spectra

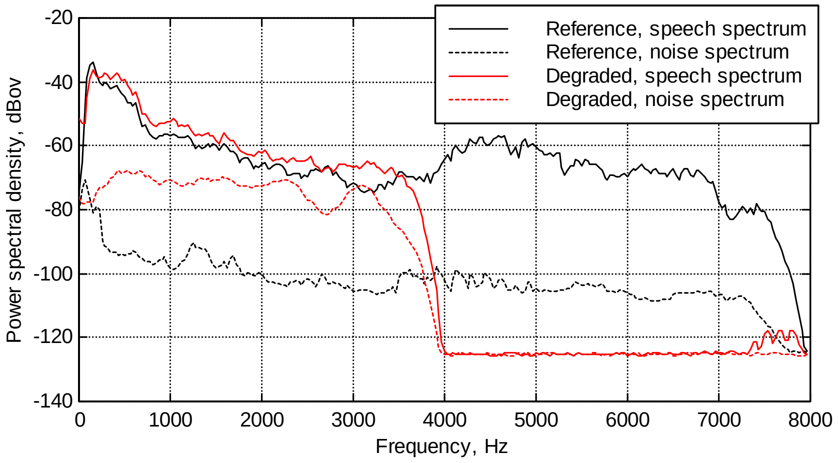
PESQ Tools Only.

The linear signal spectra are calculated using a linear frequency scale. These measures can be used to compare the spectrum of different signals and compare speech and noise. The spectra returned, for the reference and degraded signals, are as follows:

- speech spectrum (speech active periods only)
- noise spectrum (silent periods only)
- average spectrum of the whole signal.

The level of each frequency band is in dBov. Examples of the speech and noise spectrum for reference and degraded signals are shown in Figure 16.

Figure 16: Linear spectrum of reference and degraded signals



3.2.15 Transfer function estimation

PESQ Tools Only.

Five different transfer function estimates are provided: four long-term spectra, and one time-domain signal. The frequency scale of the spectra is linear; the values provided for each frequency band in the spectra represent gain, and are given in dB.

- The linear transfer function is an estimate of the transfer function between the input and output of the system under test. The value provided for each frequency band is the modulus of the mean complex gain for that band. The complex gain values used in the averaging process are calculated every 16ms using a Fourier transform.
- The phaseless transfer function is similar to the linear transfer function, but uses the modulus of the gain in the averaging process, rather than the complex value.
- The spectral difference transfer function estimate is derived from the ratio of the power of the output signal to the power of the input signal in each frequency band. An example of linear, phaseless and spectral difference transfer function estimates is shown in Figure 17.
- The coherence spectrum provides an indication of the linearity of the system under test in each frequency band, as shown in Figure 18.
- The time-domain transfer function is an estimate of the impulse response of the system under test. It is derived by taking the inverse Fourier transform of the linear transfer function described above, but using the complex value rather than the modulus of the mean gain for each band. An example of a time-domain transfer function estimate is shown in Figure 19. Figure 16–Figure 19 all relate to the same test condition.

Figure 17: Transfer function estimates

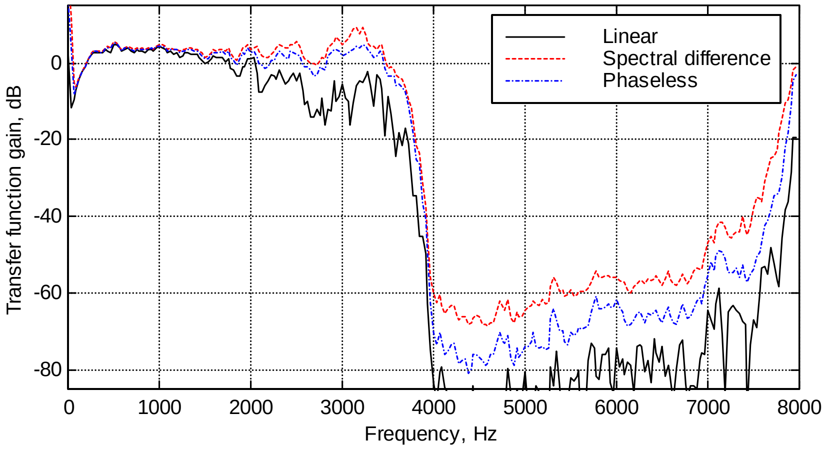


Figure 18: Coherence function

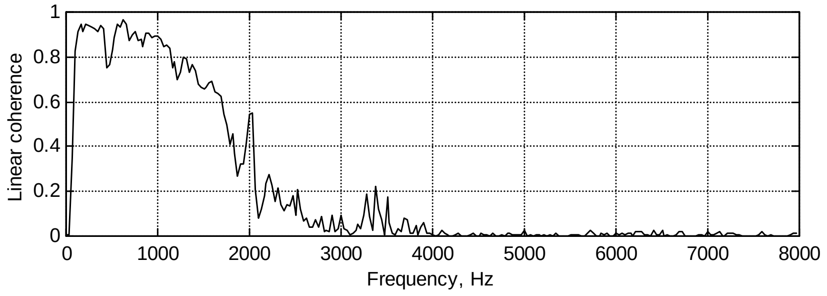
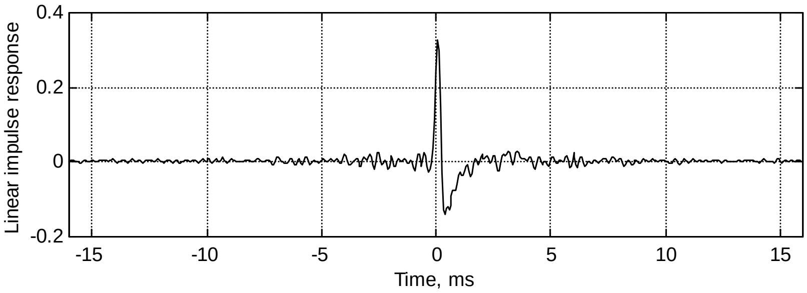


Figure 19: Impulse response estimate



3.2.16 Signal spectrograms

PESQ Tools Only.

A spectrogram is a two-dimensional output that comprises a time sequence of frequency spectra. PESQ Tools provides two spectrograms for the reference and the degraded signals.

The linear spectrogram is a sequence of Fourier transform spectra, which are calculated every 16ms using overlapping 32ms Hann windows. An example linear spectrogram is shown in Figure 20.

The linear predictive coding (LPC) spectrogram is a sequence of spectra derived by calculating the Fourier transform of 16th order LPC coefficients. The LPC coefficients are generated from the input signals every 16ms using a Hamming Window. An example LPC spectrogram is shown in Figure 21.

Figure 20: Linear spectrogram of degraded signal

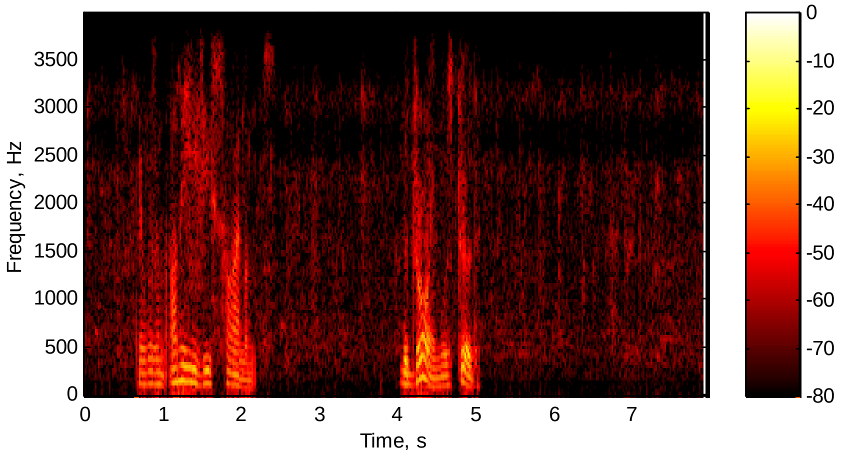
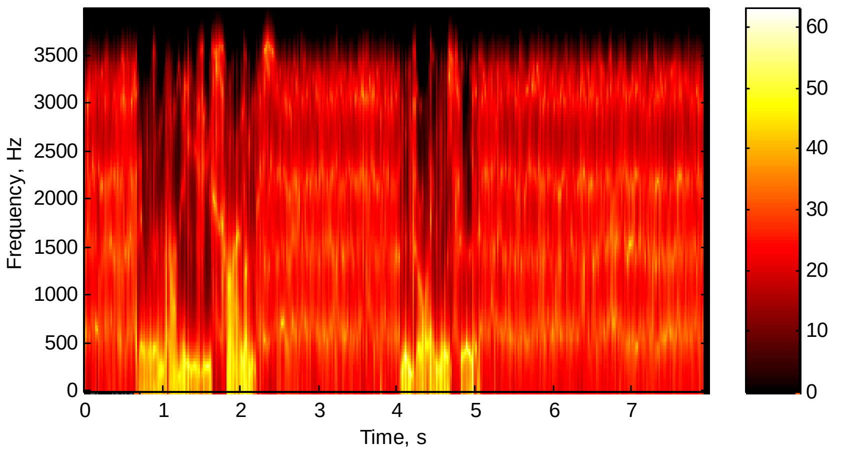


Figure 21: LPC spectrogram of degraded signal



3.2.17 LPC excitation

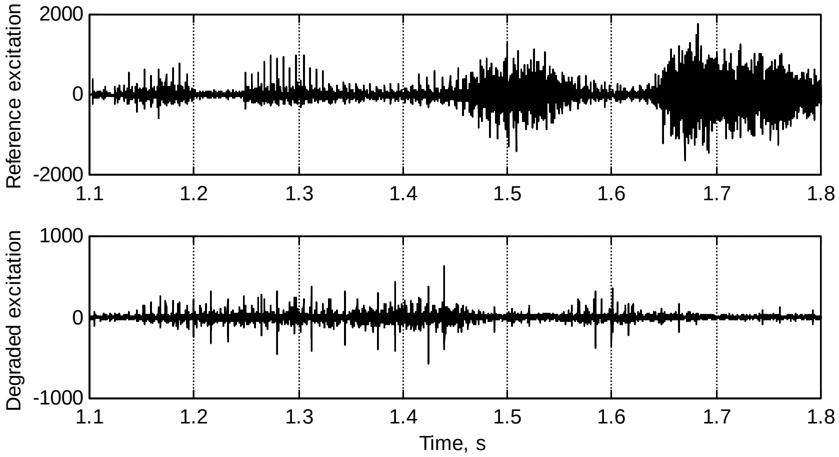
PESQ Tools Only.

PESQ Tools generates a time-domain excitation signal for both the reference and degraded input. The excitation of a speech signal is the residual signal generated by filtering it with a time varying linear predictive coding (LPC) filter. In PESQ Tools, the excitation of an input signal is produced by dividing the input signal into segments, calculating a set of 16th order LPC coefficients for each segment, and then filtering each signal segment with the corresponding coefficients.

The excitation signal is a valuable tool in speech analysis because it approximates the speech at the point of excitation, i.e. before the signal spectrum is modified by the effects of the vocal tract and lip radiation. Voiced sounds are generated from pulses produced by the periodic opening and closing of the vocal cords. The time between two pulses is the pitch period for that section of speech. Unvoiced sounds are generated by forcing air through a constriction in the vocal tract, for example that created by placing the upper teeth in the lower lip, and is typically noise-like in nature.

An example plot of a sequence of voiced sounds followed by an unvoiced sound is shown in Figure 22. In this example, the voiced part runs from about 1.1–1.65s; the unvoiced sound from 1.65s onwards.

Figure 22: Excitation of reference and degraded signals



3.2.18 Speech activity related outputs

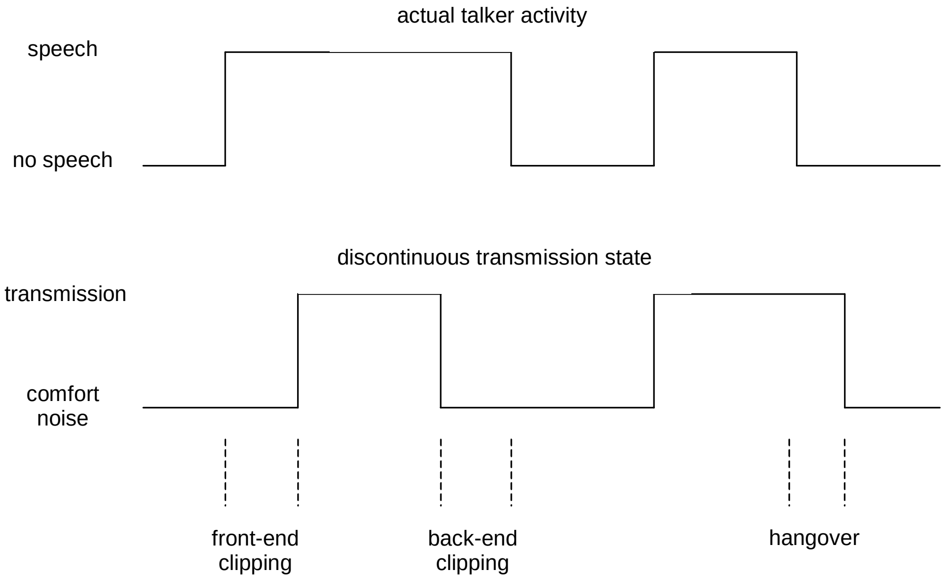
PESQ Tools Only.

PESQ Tools provides a number of diagnostic outputs that relate to the use of muting algorithms and discontinuous transmission. These outputs are generated by comparing the degraded signal to the reference signal.

Muting of a signal typically occurs when an error concealment algorithm at a receiver has insufficient information to replace missing or corrupted data. The muting estimate is provided in terms of the proportion of signal frames that have been muted by the system under test.

Discontinuous transmission (DTX) schemes aim to increase transmission efficiency by ceasing transmission during periods of talker inactivity. Applications of DTX include increasing battery life, reducing interference, or freeing transmission capacity. Temporal clipping occurs when the voice activity detection (VAD) algorithm in a DTX system misclassifies part of a speech utterance as noise, and replaces it with comfort noise at the receiver. Front-end clipping refers to the case where the start of an utterance has been clipped; back-end clipping refers the case where the end of an utterance has been clipped. Hangover is a term applied to the period after the end of an utterance when a discontinuous transmission scheme continues to transmit as normal, rather than generating comfort noise. These different events are shown diagrammatically in Figure 23.

Figure 23: Discontinuous transmission events



Statistics are provided for the following clipping events:

- All types of clipping
- All types of clipping, excluding front-end
- Front-end clipping only
- Back-end clipping only

The statistics are:

- The proportion of speech subject to clipping as a value between 0 and 1
- The number of clipping events
- The total duration of clipping events in seconds
- The mean duration of clipping events in seconds

In addition, the total duration of speech, the total duration of hangover and the number of instances of hangover are also returned.

PESQ Tools also provides an output that divides the input signal into 1ms frames, and sets various classification flags for each 1ms frame according to any speech activity events. The following flags may be set:

- Reference signal is active.
- Reference signal is active at the P.56 criterion.
- Degraded signal is active at the P.56 criterion.
- Clipping has been detected (reference is active, but degraded is not).
- Clipping classified as front-end.
- Clipping classified as back-end.
- Hang-over period (degraded is active, but reference is not).
- Comfort noise period (neither reference nor degraded is active).

Note 1. A signal is defined to be active according to the P.56 criterion if its level in the frame is greater than (ASL–15.9dB), where ASL is the active speech level of that utterance.

Note 2. The labelling of speech, clipping and noise is dependent on the voice activity decision and other classifiers – different classifiers may give different results.

3.2.19 Speech outputs

PESQ Tools Only.

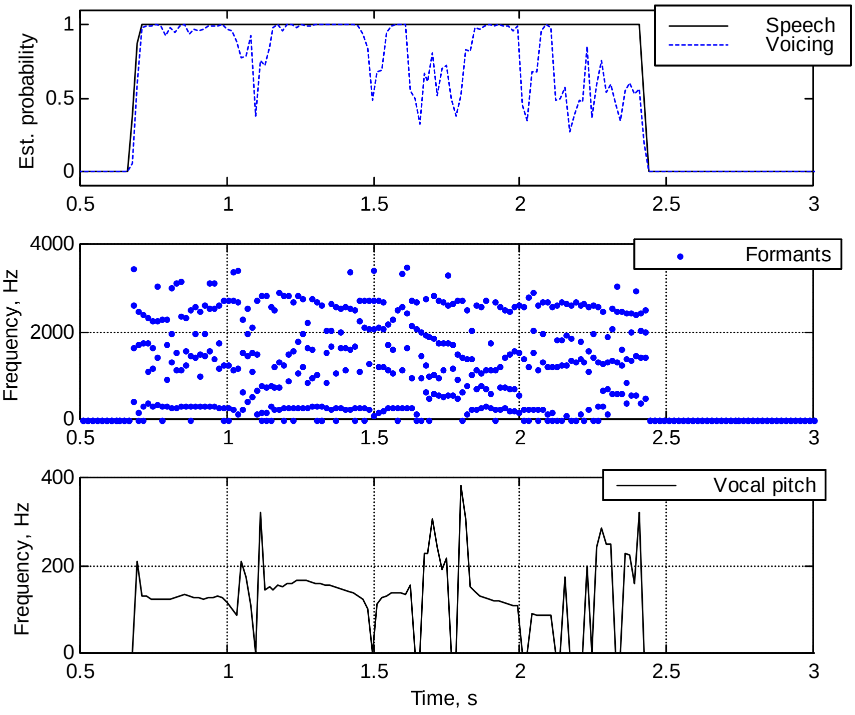
PESQ Tools provides a number of speech diagnostics for both the reference and degraded signals. These outputs relate to the production of the speech signal, and are complementary to the excitation signal discussed in section 3.2.17. The outputs are calculated using overlapping 32ms Hann windows and are updated every 16ms.

The following outputs are provided:

- vocal pitch in Hertz
- frequency of first four formants, $f_1 - f_4$, in Hertz
- power of 32ms window (absolute value – not dBov)
- probability of voicing
- probability of speech (calculated from the reference signal)

The formants are only calculated during periods of speech activity, while the pitch is only calculated during periods of voiced speech.

Figure 24: Example speech and voicing probability, pitch and formant estimates



4. Extensions to P.862

4.1 Choice of model

The models that PESQ can implement are:

- PESQ release 1.4 model (narrowband handset on reference and degraded signals)
- Backwards-compatible PESQ version 1.0 model (narrowband handset on reference and degraded signals)
- HATS ear recording on degraded signal, unprocessed (wideband) reference signal
- Wideband model (headphone listening)

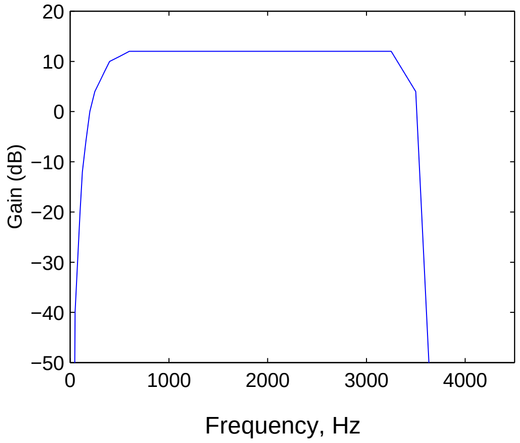
The default is the PESQ release 1.4 model.

4.2 PESQ input filters

Depending on the choice of model (section 4.1), PESQ determines internally which input filter to apply.

In the standard narrowband PESQ measurements, an input filter is applied to both the reference and degraded files before time alignment and psychoacoustic processing. The filter used, which is similar to the modified IRS receive filter specified in P.830, is shown in Figure 25. This is an approximation to the filter characteristic of a telephone handset.

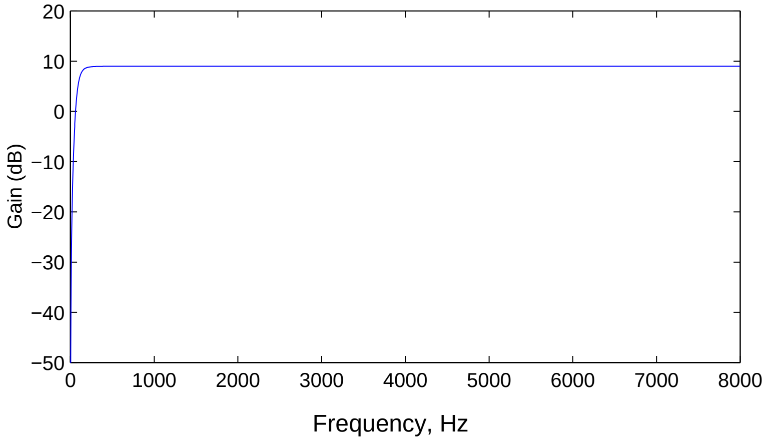
Figure 25: PESQ narrowband input filter characteristic



For wideband measurements, a filter with a flat response above 100Hz and a gentle roll-off below this point is used. This models the attenuation of the headphones and ear at low frequencies. The response of the 16kHz implementation is shown in Figure 26. The 8kHz implementation has the same gain (within 0.1 dB) in the 1Hz–4kHz range.

For HATS measurements using a telephone handset, the standard narrowband input filter is applied to the reference (to model a telephone handset), and a wideband filter is applied to the degraded file as the HATS recording will automatically include the handset path. The wideband filter used for HATS measurements has a lower gain than the filter used in the wideband model, but its frequency response otherwise has the same shape.

Figure 26: PESQ wideband input filter characteristic



This page has been left intentionally blank.

This page has been left intentionally blank.

Background and Advanced Information

5. Notes on speech signals

This section provides background material on speech signals and essential information on creating and using your own test files.

5.1.1 Properties of test signals

Networks may treat speech and silence differently, and often behave in a way that is dependent on the signals passing through them. In designing a test signal it is essential to consider the following factors:

- Temporal structure—speech and silent periods
- Level and frequency content
- Source material—natural or artificial speech
- Duration of an individual recording
- Requirement for multiple measurements of the same condition

5.1.2 Temporal structure

Test signals should include speech bursts separated by silent periods, to be representative of natural pauses in speech. Speech bursts should normally be 1–3 seconds in duration. To test certain types of voice activity detector, silent periods should be at least 300ms in duration. As a guide, speech should be active for between 40% and 80% of the time.

5.1.3 Level and frequency content

A key factor in speech quality is the level (the signal power), usually quoted in dB. In digital speech files, a typical level is –26dBov. Signals injected into the network should normally be at the appropriate calibrated level, which may vary depending on the national standards and the impedance of the circuit.

As telephone handsets and analog networks both introduce filtering, it is important that the test signals have a representative frequency content. In other words, they must be pre-filtered in an appropriate way. For fixed network measurements, the modified IRS send filter is normally applied to the speech before injection into the network [ITU-T P.830]. This attenuates strongly below 300Hz and also provides a small boost of about +10dB per decade within the passband. Level is measured *after* the filtering has been applied.

5.1.4 Source material

Natural recorded speech or the artificial speech supplied with PESQ may be used as test signals. Natural speech recordings should contain a representative and balanced range of speech sounds. If different recordings are to be concatenated, the joins must be made in silent periods to avoid discontinuities.

Signals that are not speech-like should not be used with PESQ for several reasons. They may cause the network to behave in an unrepresentative way, they cannot fully test the quality of speech codecs, and they do not reproduce the temporal structure of speech that may be exploited by elements such as voice activity detectors.

5.1.5 Duration of an individual recording

PESQ is optimised for recordings of 8s in duration containing at least 4s of active speech. As a guide, the minimum length for a measurement to give a representative PESQ score is about 6s, containing at least 3s of active speech. Recordings of 16s or longer in duration should be split into shorter sections and each processed separately through PESQ.

The reference and degraded signals do not have to be of exactly the same length. PESQ aligns and processes only the sections for which data is available from both reference and degraded signals. If a measurement introduces significant unknown delay, it is a good idea to extend the recording at both the start and end to ensure that the entire test signal is captured.

5.1.6 Multiple measurements

Whenever possible, more than one measurement should be made of a given condition to allow time-varying quality or material dependence to be assessed. The PESQ scores for all recordings on a given condition can be averaged to give a view of the overall quality, and the individual scores show quality variation over the condition due to material or time dependence.

If artificial speech is being used, the measurement should be at least 28s long. It is recommended that this be split into three or four files. If natural recorded speech is used as a test signal, 32s should be regarded as a minimum (8s for each of four talkers) and, if possible, up to two minutes (16 recordings of 8s duration) should be used.

5.1.7 Reference signal

The reference provides PESQ with information on how the original, unprocessed signal should sound. The file must contain samples at 8kHz or 16kHz sample rate. This data should normally be stored as 16-bit integers.

The reference should be distortion-free so that PESQ can assess the quality of the system under test. The reference can often be exactly the same file that is passed through the distorting system. Certain types of pre-processing make little difference in practice to PESQ scores, especially filtering with the modified IRS send characteristic, or level adjustment (as long as quantisation errors remain small).

Various types of noise may be added to evaluate the system's performance at transmitting noisy speech. In these situations, the reference that is used with PESQ should be the original file before any processing was applied and with no background noise added. See section 7 for more information on using PESQ to test quality in the presence of noise.

It's important to note that real speech signals that are passed through networks are not usually completely (digitally) silent during pauses between speech utterances. PESQ is able to detect the effect of small amounts of added noise if the reference signal is very quiet in silent periods. This means that a measurement of a system that adds noise (such as an analogue connection), using a reference signal that includes digital silence, may give a slightly lower quality score than a measurement of the same system using a noisy reference. In effect the noisy reference "masks" the noise added by the system.

5.1.8 Degraded signal

The degraded signal is the distorted version of the reference signal, measured at the output of the system under test. As little further degradation as possible should be introduced before this signal is input to PESQ, as the model would not be able to separate this from the distortion introduced by the system. Ideally the degraded signal should be recorded at 16kHz sample rate, though for certain applications the use of 8kHz sample rate might be unavoidable. The signal should be stored with at least 16 bits of precision, at a level that avoids amplitude clipping and unnecessary quantisation.

6. Overview of subjective testing

This section describes the subjective testing methods used to obtain the opinion scores that PESQ is calibrated to predict. It is beyond the scope of this document to provide a full guide on designing and conducting subjective tests.

For more information, you should consult the references listed in section 10. This gives the ITU-T recommendations concerning subjective testing. However, it should be noted that there are certain differences between these recommendations and the methods in current use in the standards bodies such as ETSI. What we describe here is focused towards the subjective methods used to gather data for calibrating PESQ, based on best practice in standards-related work.

6.1 Listening and conversational testing

Subjective testing aims to obtain a key benchmark of network performance based on the customers' perception of speech quality. Examples of the behaviours considered include low bit-rate coding, transcoding (multiple coding stages), and channel errors due to mobile or packet-based transmission.

There are two distinct classes of telephony subjective test: listening and conversational. In listening tests, subjects hear various distorted recordings, and vote on their opinion of the quality after hearing each one. Because there is no two-way element of communication, listening tests cannot fully model the effect of listening level, talker echo, delay or handset sidetone.

In conversational tests, pairs of subjects hold a conversation over a test network connection before voting on its quality. These measurements take account of the whole link, including handsets and sidetone, echo, level and delay impairment. Conversational tests are generally more expensive than listening tests, and a single conversational test is only able to investigate a small number of conditions.

PESQ on its own is a listening model, so PESQ quality scores do not normally take account of the conversational factors: level, talker echo, delay and sidetone. However, information on level and delay may be gained from the PESQ level and delay values if the measurement setup is appropriately calibrated. Other techniques can be used to estimate level, echo and delay. Sidetone can often be assumed constant based on typical equipment used in a given country.

Conversational factors may be important in some circumstances. In particular, if a network introduces significant level changes (attenuation or gain), or if it has audible talker echo or large delays, it may be appropriate to consider measurements of these factors as well as PESQ scores. For example, voice over IP transmission equipment may often improve listening quality by increasing buffer length, introducing greater delay. This causes greater conversational impairment and, since the network is most likely to be used for two-way communication, this change in delay should also be considered before conclusions on overall quality are made.

In the remainder of this document we consider only subjective listening tests.

6.2 Design of a subjective test

Subjective perception of quality depends on a large number of factors. In designing a subjective test it is essential to control many extraneous variables by choosing appropriate values or averaging over a typical population distribution. These variables are examined in this section.

6.2.1 Opinion scales

The most common technique in listening testing for telephony is known as the Absolute Category Rating (ACR) method. In this type of test, subjects hear only the processed conditions. After hearing each recording the subjects are prompted to vote. PESQ produces listening quality scores that are analogous to the ACR listening quality opinion scales.

The votes given by subjects for each file are then averaged to give a *file* mean opinion score (MOS). The average of all votes given to all files for a given network condition is known as the *condition* MOS.

There are some alternative test structures in use for specific applications. These include Degradation Category Rating (DCR) and Comparison Category Rating (CCR) methods. Because these methods use a different quality question, they will not normally give the same results as an ACR test. Indeed, there is evidence to suggest that asking a different quality question may result in different conclusions being reached when comparing one type of communications technology with another.

Where subjective test results are to be compared with PESQ scores, we strongly recommend that the ACR listening quality method is used.

6.2.2 Conditions

A typical listening test allows up to about 50 network conditions to be evaluated, assuming that an Absolute Category Rating (ACR) method is used with speech material from four talkers (see below for more details). At least six of these conditions should normally be given over to MNRU references [P.810] that cover the full range of quality. It is also a good idea to include standard network conditions such as G.711 so that quality scores can be compared against them.

At the start of each test all subjects hear the same set of 6–8 preliminary conditions, covering a range of distortion types, and vote on their quality using the same procedure for voting as the main set of conditions. The votes for the preliminaries are discarded; they serve as an anchor to ensure that all subjects start the test with the same idea of what the range of quality and the types of distortions will be.

6.2.3 Other factors

A test aims to obtain a measure of the subjective quality of a number of network conditions. However there are usually many other variables. The design of a subjective test should attempt to control these to prevent them from influencing the condition opinion score. The following are the most important of these variables.

Talker dependence. Because different people’s speech may be distorted in different ways, it is usual to pass speech from four different talkers – two adult male, two adult female – through each condition. Subjects hear each condition four times at different stages in the test, with speech from each of the four talkers.

Material dependence. Different sections of speech may be distorted in different ways. For example, a frame erasure event may be less audible if it coincides with an unvoiced part of speech, as opposed to a voiced part of speech. Recently, practice in subjective testing has moved to control this effect by using partially or fully factorial designs, which evaluate three or more different recordings from each talker for a given condition. Different groups of subjects hear a different combination of source speech material and condition. This appears to give more consistent results than using only one recording from each talker for a given condition.

Order dependence. A subject’s vote for a given condition will depend to some extent on the last few conditions heard. This effect may be partially controlled by scrambling the order. Ideally, a different order should be used for every subject, otherwise there is danger that the subjective results could show a bias that is due not to the quality of the conditions but to their presentation order.

Language dependence. Subjects are normally native speakers of the language used in a test. If language dependence is to be evaluated subjectively, it is necessary to use a pool of subjects of different nationalities or to conduct tests in several countries. On the available evidence it appears that PESQ performs well for subjective tests conducted in several languages, or language groups.

Number and population of subjects. A telephony listening test normally uses at least 16 subjects; 24 is a common number of subjects in standards work. They should be untrained, and should not have participated in another test within the last year. Typically, subjects are selected at random from an adult population, and should ideally cover a representative range of ages and be approximately equally split between the sexes. Averaging across the votes of this population aims to control possible preference effects, for example due to gender or age.

Balance. The conditions in the test should cover a broad range of quality. Although MNRU references help to ensure this, the other conditions should be chosen to include several different levels of audible distortion.

6.3 Processing of speech material

Although the methods used to process material for a subjective test are beyond the scope of this document, examples are given here of the processing stages for two types of condition.

Simulated condition

1. Record original speech material using high-quality microphone in quiet conditions.
2. Send filtering (e.g. modified IRS) and level alignment (e.g. to –26dBov).
3. Add environmental noise at appropriate level if required.
4. Downsample to 8kHz, the sample rate at which the codec simulations operate.
5. Apply coder.
6. Channel error insertion.
7. Apply decoder.
8. If multiple transcodings are simulated, a filter and an arbitrary delay may be inserted to make the transcodings asynchronous, then the coder/error/decoder stages are repeated.
9. Upsample to 16kHz for presentation in subjective test, checking for clipping.
10. Verify that active speech level lies within bounds.

Measured condition

1. Record original speech material using high-quality microphone in quiet conditions.
2. Send filter and level align to calibrated level for measurement system.
3. Set up connection.
4. Play out original signal at 16kHz sample rate.
5. Record degraded output of system at 16kHz sample rate.
6. Adjust level to calibrated active speech level (e.g. –26dBov).

6.4 Analysis of results

6.4.1 Condition mean opinion score

The key measure of quality is the average of votes, across all subjects and all files, given to each condition. This is known as the condition mean opinion score, often abbreviated to MOS, and is the figure most commonly used to describe a condition.

6.4.2 Other MOS measures

It is also possible to average votes to obtain an MOS for each file (‘file MOS’) and/or each talker (‘talker MOS’) in a given condition. Though less commonly used than condition MOS, these scores given an indication of quality dependence on material or talker.

6.4.3 Further statistical analysis

Many statistical techniques may be applied to analyse the distribution of votes and investigate the influence of factors such as talker or subject. For example, the following methods may be useful.

- **Confidence interval** provides an estimate of the range in which the ‘true’ mean may lie given the distribution of observations (votes).
- **T-tests** allow the votes from two different conditions to be compared to assess whether there is evidence that any differences between them are significant or merely stem from randomness in the voting process.
- **ANOVA** (analysis of variance) is a technique for testing, and ideally eliminating, the influence of many factors that cannot be fully controlled, for example, talker dependence, listening order, and individual subjects.

6.4.4 Further reading

More information on these and other statistical methods may be obtained by following the references listed in section 10.

7. Noise testing

For certain types of network – especially mobile – it may be important to evaluate the quality of transmission when the original signal is corrupted by background noise. For example, a low bit-rate coder optimised to transmit speech may produce strange-sounding distortions when noise is present.

PESQ may be used for testing transmission quality in the presence of noise as described in this section. We would like to emphasise that conducting subjective tests with background noise conditions is more difficult than conventional subjective testing, and should be approached with caution.

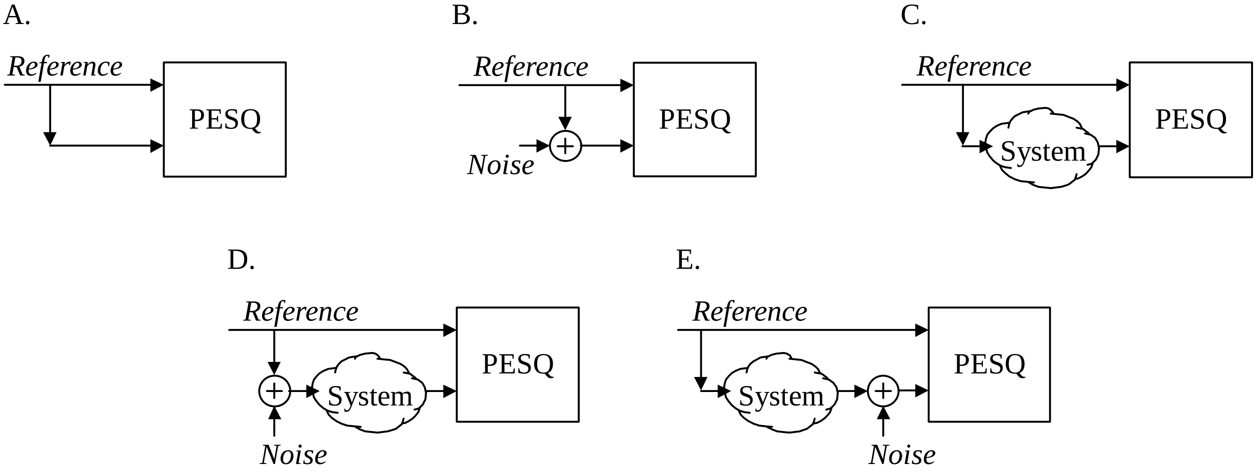
7.1 Background noise testing with PESQ

Five different tests can be made with PESQ to evaluate the effect of noise on the quality of a given codec or system:

1. **No noise or coding.** This gives the baseline quality with no distortion. PESQ scores are normally 4.5 in this case.
2. **Noise only, no coding.** This gives the effect on quality of the noise alone, and is important because the presence of the noise itself may be the largest factor.
3. **Coding only, no noise.** This gives the quality of the system with clean speech.
4. **Noise added before coding** (at input to system). This gives the quality of the system when transmitting noisy speech.
5. **Noise added after coding** (at output of system). This separates the effect of the noise from the effect of noise on the system.

In all cases the reference signal supplied to PESQ should be the clean speech. These permutations are shown in Figure 27.

Figure 27: Evaluation of quality with background noise



This makes it possible to investigate: the effects of: the noise alone (B), the performance of the transmission system alone (C), or the performance of the system while transmitting noisy speech (D). (A) provides a simple check of the baseline quality and may be omitted with PESQ. Comparison of (E) with (B) and (D) provides another way for establishing the effect of the system on the noise.

7.2 Subjective testing with background noise

It is not possible for this document to fully describe the methods used in subjective testing with noisy speech. We can only summarise some of the available techniques and outline a typical test design.

The choice of subjective opinion scale and voting method is critical. This usually affects the results of a test because, as noted above, the quality prompt can influence the votes given to different conditions, even changing their ordering.

The ACR listening quality method may be used for background noise testing. In this case the noise is one type of degradation that the subjects vote on. The listening quality method appears to be more sensitive to noise (compared to coding distortions) than listening effort. In an ACR test, the effect of noise on a transmission system may be tested by including the following conditions in the test (analogous to those described in the previous section):

1. clean speech, unprocessed
2. noisy speech, unprocessed
3. clean speech, coded
4. noisy speech, coded

These conditions allow tests of the subjectivity of several factors: the noise alone (B compared to A), the system with clean speech (C compared to A), and the system with noisy speech (D compared to B). Several different types or levels of noise can be assessed in a test, although of course there is only need for a single set of clean speech conditions. This type of test normally uses the MNRU as a reference with clean speech only.

DCR and CCR methods may also be used. In this case the reference signal that subjects hear may be the noise-free, unprocessed speech (standard methods) or the noisy, unprocessed speech (the so-called 'modified' methods). These methods allow a comparison similar to that possible with the ACR methods, although ACR requires a shorter listening time for each condition.

The background noise tests used in PESQ calibration were all conducted with the ACR methods. In this case the reference signal presented to PESQ is the clean, unprocessed speech. If the results of a subjective test including environmental noise are to be compared with PESQ scores, it is strongly recommended that either the ACR listening quality or the ACR listening effort method is used.

8. Comparison between objective and subjective results

8.1 Mapping PESQ scores to subjective MOS

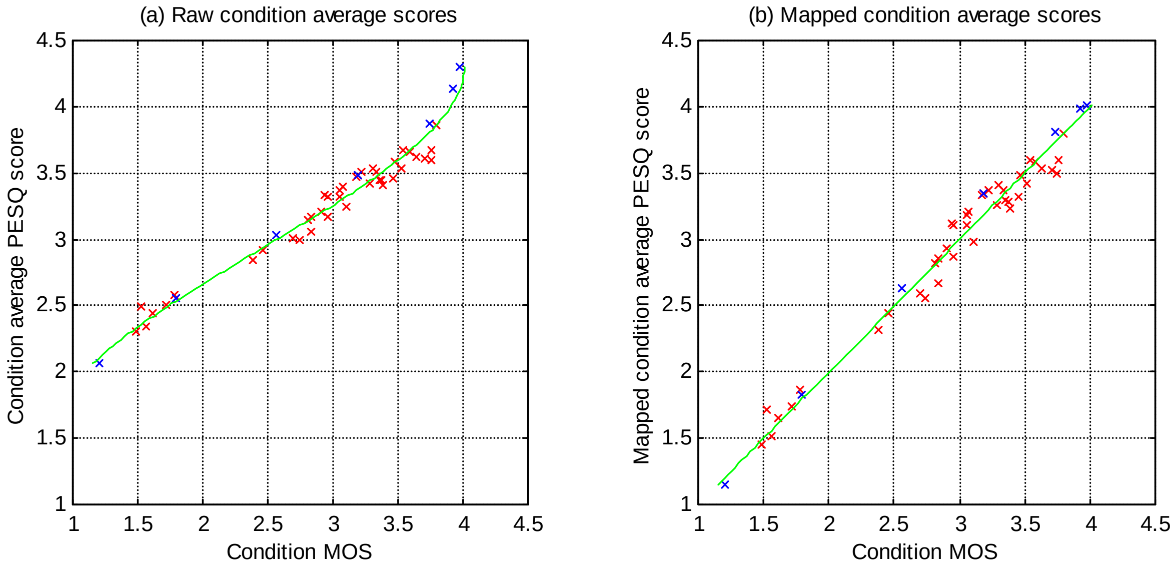
Scores given to identical conditions in two subjective tests will not generally be equal. It is necessary to take account of this fact in comparing subjective and objective scores. Subjective votes are affected by such factors as the balance of the other conditions in a test or the individual preferences of each subject. Since one subjective test cannot be directly compared with another, it is impossible for an objective model such as PESQ to give exactly the same scores as every subjective test.

However, the difference between two sets of scores for the same conditions is usually no more than a smooth curve, plus small (ideally random) errors. This curve can be thought of as a function that can approximately map one set of scores on to the other. To preserve order, this mapping should be monotonic (one-to-one). This section illustrates how PESQ scores may be mapped to subjective MOS using this method.

The techniques outlined in this section apply equally to PESQ score and PESQ-LQ. PESQ-LQ is generally closer to listening quality MOS than PESQ score, but the comparison between either value and MOS is affected by the same variability in subjective votes, and hence MOS, that is outlined above.

Figure 28(a) plots the subjective condition MOS against the condition-averaged PESQ quality score for each condition for a subjective test on mobile codecs. This clearly shows that there is a simple relationship between PESQ score and MOS.

Figure 28: Mapping between PESQ score and subjective condition MOS



This relationship between PESQ score and MOS is modelled using a monotonic cubic polynomial. The solid line in Figure 28(a) shows this polynomial function. The polynomial can then be applied to map the PESQ scores for each condition onto the same scale as MOS in this test. Figure 28(b) shows the same subjective condition MOS plotted against the mapped PESQ scores, illustrating how the mapping works.

All of this analysis is normally performed with condition averages of both objective and subjective scores. The mapping should be constrained to be monotonic across the range of the data, otherwise it will not preserve the ordering of the objective scores. A different mapping is required for each subjective test to take account of the differences outlined above.

Psytechnics recommend this method of using a monotonic cubic polynomial, optimised for minimum mean squared error, to map between subjective and objective scores. This method has been accepted in the ITU-T as giving the most relevant comparison between objective models and subjective scores, and it was used in the calibration of PESQ.

8.2 Correlation coefficient

The closeness of the fit between PESQ and the subjective scores may be measured by calculating the correlation coefficient. Normally this is performed on condition averaged scores, after mapping the objective to the subjective scores; in other words, with data of the form plotted in Figure 28(b). The correlation coefficient is calculated with Pearson's formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

In this formula, x_i is the condition MOS for condition *i*, and $\bar{x}$ is the average of $x_1 \Lambda\ x_N$. y_i is the mapped condition-averaged PESQ quality score for condition *i*, and $\bar{y}$ is the average of $y_1 \Lambda\ y_N$. For the data shown in Figure 28(b), the correlation coefficient *r*=0.988. Correlation coefficients for a number of subjective tests are given in the next section.

8.3 Residual errors

The mapping removes any systematic offset between the PESQ scores and the subjective MOS, minimising the mean square of the residual errors

$$e_i = x_i - y_i$$

Various measures may be applied to the residual errors to given an alternative view of the closeness of PESQ scores to subjective MOS.

9. Performance of PESQ

9.1 Narrowband measurements

PESQ was compared to PSQM and MNB, the previous standards, using methodology similar to that of the ITU-T competition that resulted in recommendation P.862. See Rix, Beerends, Hollier, and Hekstra, reference in Section 10.

The test used correlation coefficient and residual error distribution to quantify the performance of models at predicting subjective MOS. These metrics are calculated for each subjective test separately, after mapping the objective scores to the subjective scores for that test in a minimum squared error sense using monotonic third-order polynomial regression. This mapping ensures that the comparison is made in the MOS domain whilst allowing for normal variations in subjective voting between tests.

Tests are grouped according to whether conditions were predominantly from mobile, fixed, voice over IP (VoIP) and multiple type networks. Table 5 and Table 6 show correlation and residual error distribution for PESQ, PSQM and MNB for 38 subjective tests that were available to the developers of PESQ. These included a wide range of simulated and real network measurements. Table 7 and Table 8 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.

Table 5: Average and worst-case correlation coefficient for 38 subjective tests known during PESQ development, sub-divided by test type

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

Table 6: Error distribution across all 38 known subjective tests.

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, 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

Table 7: Correlation coefficient, 8 unknown subjective tests (PESQ only)

Test	**Type**	**Corr.**
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

Table 8: Error distribution, 7 unknown subjective tests (PESQ only).

Note: test 3 was excluded as the data for this test was per-file 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

9.2 Wideband measurements

Wideband PESQ is a Psytechnics extension. Results of tests on wideband PESQ were reported in Rix, “Proposed modification to Draft P.862” — see reference in Section 10. The results below summarise the correlation between measurements using PESQ and subjective tests. In all cases the subjects listened binaurally through wideband headphones.

The performance of wideband PESQ was assessed against four subjective experiments:

1. Narrowband and wideband MNRU conditions and CELP codecs.
2. Narrowband (8kHz sample rate) conditions only: MNRU, CELP codecs and three packet loss conditions for each CELP codec.
3. The same structure as experiment 2, but with all of the conditions wideband.
4. Four different families of codecs at between 8 and 64 kbit/s, along with MNRU references, at three different sample rates (8kHz, 11.025kHz and 16kHz).

The results are summarised in Table 9, which presents the correlation of wideband PESQ with subjective MOS for each of the four wideband speech experiments. For all of these experiments wideband PESQ shows high correlation with subjective quality. It should be noted that wideband PESQ has not been validated with any conditions containing additive background noise.

Table 9: Overall correlation of wideband PESQ with subjective test results

Experiment	**P905/1**	**P905/2a**	**P905/2b**	**AES107**
Per condition correlation coefficient between wideband PESQ and subjective MOS, per condition, after third order mapping	0.952	0.981	0.977	0.949

Supplementary Information

10. References

10.1 Objective speech quality assessment

Wang, S., Sekey, A. and Gersho, A. “An objective measure for predicting subjective quality of speech coders”. *IEEE Journal on Selected Areas in Communications,* 10 (5), 819–829, 1992.

Hollier, M. P., Hawksford, M. O. and Guard, D. R. “Characterisation of communications systems using a speech-like test stimulus”, *Journal of the Audio Engineering Society*, 41 (12), 1008–1021, 1993.

Beerends, J. G. and Stemerdink, J. A. “A perceptual speech-quality measure based on a psychoacoustic sound representation”. *Journal of the Audio Engineering Society,* 42 (3), 115–123, 1994.

Hollier, M. P., Hawksford, M. O. and Guard, D. R. “Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain”. *IEE Proceedings – Vision, Image and Signal Processing,* 141 (3), 203–208, 1994.

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

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.

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

Rix, A. W., Reynolds, R. and Hollier, M. P. “Perceptual measurement of end-to-end speech quality over audio and packet-based networks”. *106th Audio Engineering Society Convention*, pre-print no. 4873, May 1999.

Rix, A. W. and Hollier, M. P. “Perceptual speech quality assessment from narrowband telephony to wideband audio”, *107th Audio Engineering Society Convention*, pre-print no. 5018, September 1999.

Rix, A. W., Reynolds, R. and Hollier, M. P. “Robust perceptual assessment of end-to-end audio quality”, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* 39–42, October 1999.

Rix, A. W., “Proposed modification to Draft P.862 to allow PESQ to be used for quality assessment of wideband speech”, ITU-T Study Group 12 Delayed Contribution COM12-D007 (February 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 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2001.

Rix, A. W., “Proposed Annex B to Recommendation P.862: Application of PESQ to speech quality assessment of wideband telephone networks and speech codecs”, ITU-T Study Group 12 Contribution COM12-36 (August 2001).

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

10.2 Subjective testing

Methods for subjective determination of transmission quality. ITU-T Recommendation P.800, 1996.

Modulated noise reference unit (MNRU). ITU-T Recommendation P.810, 1996.

Subjective performance assessment of telephone-band and wideband digital codecs. ITU-T Recommendation P.830, 1996.

10.3 Statistics

Kreyszig, E. *Advanced engineering mathematics.* McGraw-Hill, 8th edition, 1998.

Peebles, P. *Probability, random variables and random signal principles.* McGraw-Hill, 3rd edition, 1993.

Dunn, O. J. *Applied statistics: analysis of variance and regression.* Wiley, 2nd edition, 1987.

Snedecor, G. W. and Cochran, W. G. *Statistical methods.* Iowa State University Press, 8th edition, 1989.

11. Glossary

ACR	Absolute Category Rating: a method for subjective rating of quality in tests
ASL	Active Speech Level
CCR	Comparison Category Rating: a method for subjective rating of quality in tests
DCR	Degradation Category Rating: a method for subjective rating of quality in tests
DTX	Discontinuous transmission
HATS	Head And Torso Simulator
IRS	Intermediate Reference System
ITU-T	International Telecommunications Union – Telecommunication Standardisation Sector
LPC	Linear predictive coding
LQ	Listening Quality
MIRS	Modified IRS
MNB	Measuring Normalising Blocks: an earlier model for assessing speech quality of codecs.
MNL	Mean Noise Level
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
PESQ	Perceptual Evaluation of Speech Quality. An algorithm described in ITU-T recommendation P.862.
PSQM	Perceptual Speech Quality Measure: an earlier model for assessing speech quality of codecs.
RMS	Root Mean Square
SNR	Signal to Noise Ratio
VAD	Voice activity detector

End of sample userguide

Guidelines

Not to be included in end user documentation – for licensee use only

A. Guidelines for the use of the sample user guide by the licensee

These guidelines should be read by all Licensees of the Psytechnics distribution of PESQ (Perceptual Evaluation of Speech Quality). They contain notes on creating end-user documentation for products that include PESQ or PESQ Tools.

The guidelines begin by defining three PESQ usage profiles. The guidelines then follow the structure of the user guide to describe which inputs and outputs are recommended for use in the different profiles. Cross-references to the appropriate sections of the sample PESQ User Guide are provided.

Provided that a Licensee continues to pay PESQ royalties to Psytechnics, the Licensee may include the text of the User Guide in their own User Guides as appropriate. The text should be modified, as described in these guidelines, to suit the requirements of the Licensee's product.

A.1 Introduction

PESQ can be used in different types of application, and by people with different requirements and levels of knowledge. For simplicity, we have identified three usage profiles. In creating your own user guide, you should base it on the profile that your own application most closely matches.

The usage profiles are:

- Profile 1: PESQ as a simple measurement device, where there may be a choice to process speech sampled at either 8kHz or 16kHz. This returns only quality scores.
- Profile 2: PESQ as an advanced speech quality analyser, with a full set of features and outputs for use by trained individuals.
- Profile 3: Use of Psytechnics PESQ in Head and Torso (HATS) measurements and in wideband telephony. This is an extension to the P.862 standard.

The concept of profiles is provided for guidance and it is left to licensees to choose the appropriate profile.

A.2 PESQ Tools

The PESQ Tools option greatly extends that range of diagnostic outputs provided by PESQ. The use of PESQ Tools is therefore highly recommended for Profile 2.

A.3 Inputs and Outputs for basic use of PESQ

(User Guide, section 2)

You are recommended to include the information referred to in this section as part of your user documentation for all profiles (1–3). For a device according to profile 1, you may choose to provide only this information and omit the background and advanced use information referred to in the following sections.

A.3.1 Input option: Speech signals

(User Guide, section 2.2.1)

Implementations may vary on how test signals are stored and passed to PESQ. Your documentation should explain the format(s) that you choose to offer in your product.

A.3.2 Input option: Sampling Rate

(User Guide, section 2.2.2)

Implementations may vary in how the sample rate is specified and whether there is any default value. For measurement devices that only operate at one sample rate, there is no need to offer this option or to provide information on the choice of sample rate. If the model detects sample rate by other means, for example from a .wav format file header, the documentation should still discuss the issues related to choice of sample rate.

A.3.3 Input option: Model specification

(User Guide, section 2.2.3)

You may offer a switch to select the PESQ version 1 operation mode (model –1). However we recommend that the default processing should be the new process in PESQ release 1.4 (model 0), and we request that you make model 0 the default.

A.3.4 Level alignment

(User Guide, section 2.3)

Level alignment is integral to PESQ and there must be no option to change this.

A.3.5 Time alignment

(User Guide, section 2.3)

You must not offer any option to alter the way in which time alignment is performed.

Although a small improvement in processing speed may be gained by preventing PESQ from testing for delay changes during speech, this could cause the PESQ scores to be significantly in error if delay changes do actually occur. If processing speed is a major problem even after fully optimising your code, you should contact Psytechnics.

A.3.6 Results: quality scores

(User Guide, section 2.4)

You are encouraged to offer PESQ score, PESQ-LQ and PESQ-Ie as the outputs of the model. You may use the descriptions in section 2.4 of the User Guide, including the formula for PESQ-LQ and the reference to ITU-T Recommendation P.834, in your documentation.

A.4 Advanced use (including PESQ Tools)

(User Guide, section 3)

The information referred to in this section will be provided in addition to the basic information described in A.3, and will typically apply to profiles 2 and 3. It should be included in your user documentation when your product offers the corresponding input or output to the user. Unless otherwise stated, you may provide the following information in profiles 2 and 3.

The PESQ Tools option greatly extends that range of diagnostic outputs provided by PESQ. The use of PESQ Tools is therefore highly recommended for Profile 2.

A.4.1 Input option: Model specification

(User Guide, section 3.1)

You may include the further discussion of changes in release 1.4 that are presented in section 3.2.1of the User Guide.

A.4.2 Results option: Frame-by-frame Delay

(User Guide, section 3.2.2)

We recommend that you present a graph showing the frame-by-frame delay, for example as shown in section 3.2.2 of the User Guide. Alternatively you may plot a histogram of the frame-by-frame delay.

The words “time offset” may be used in your documentation instead of, or in addition to, “delay”.

A.4.3 Results option: Bark scale transfer function

(User Guide, section 3.2.3)

You should display the Bark scale transfer function estimate for products in profile 3, and you may wish to offer it for profile 2.

A.4.4 Results option: Perceptual parameters

(User Guide, section 3.2.4)

The symmetric and asymmetric disturbance values may be presented graphically frame-by-frame. Alternatively, you may also present the average symmetric and asymmetric disturbance as single values for each condition.

A.4.5 Results option: Frame-by-frame quality score

(User Guide, section 3.2.5)

As an alternative to the symmetric and asymmetric disturbance values, you may wish to present the simpler frame-by-frame quality score. In this case you must include comments on the limitations of this output, as given in section 3.2.5 of the User Guide.

A.4.6 Results option: Signal waveforms

(User Guide, section 3.2.6)

The signal waveforms show the amplitude and timing of the signals. You may choose whether or not to display them. You are encouraged, where possible, to provide an option to play back the original and degraded files.

A.4.7 Results option: Sensation surfaces

(User Guide, section 3.2.7)

You are encouraged to show the sensation surfaces and error surface (see section A.4.8). It is often useful to present either one or both of the sensation surfaces, or the reference signal waveform, alongside the error surface so that the location of error events may be easily seen. The format that you use (for example, an image, where colour is related to loudness, or as a 3-D surface) is left to you.

You may also wish to include in your documentation some sample sensation and error surfaces so that users can learn how to interpret the images. Examples of different types of distortion are listed in the User Guide, section 3.2.8.

A.4.8 Results option: Error surface

(User Guide, section 3.2.8)

The error surface is not provided explicitly by PESQ, but is simply calculated as the degraded surface minus the reference surface. You are encouraged to show the error surface along side the sensation surfaces. See section A.4.7 for further information.

A.4.9 Results option: Frame-by-frame delay statistics

(User Guide, section 3.2.9; PESQ Tools only)

If PESQ Tools is available, you may wish to provide the frame-by-frame delay statistics for profile 1 in addition to profiles 2 and 3.

A.4.10 Results option: Utterance-by-utterance delay

(User Guide, section 3.2.10; PESQ Tools only)

You may wish to offer a view of the utterance-by-utterance delay using the utterance delay, start, end and confidence information.

A.4.11 Results option: Utterance-by-utterance level

(User Guide, section 3.2.11; PESQ Tools only)

You may wish to offer a view of the utterance-by-utterance level using the utterance level, start, end and confidence information. This can be useful in diagnosing some advanced network processes such as adaptive level control.

A.4.12 Results option: Signal level and gain measures

(User Guide, section 3.2.12; PESQ Tools only)

The levels of speech and noise, and the gain of the system, are interesting for many applications and you are encouraged to offer them to users for profiles 2 and 3. You should edit the description of the outputs to reflect how you calculate them and present them to the user.

(Note for developers: the level outputs are returned from PESQ using the `LV_Params` array; your own implementation should calculate SNR using the simple formula given in section 3.2.12 of the User Guide.)

A.4.13 Results option: Bark signal spectra

(User Guide, section 3.2.13; PESQ Tools only)

For profile 3, and possibly also for profile 2, you may wish to show the signal, speech and noise spectra that are calculated for both reference and degraded signals. PESQ Tools returns these measures on a perceptual (Bark) frequency scale and a linear frequency scale (see A.4.14).

A.4.14 Results option: Linear spectra

(User Guide, section 3.2.14; PESQ Tools only)

For profile 3, and possibly also for profile 2, you may wish to supplement the Bark signal spectra (see A.4.13) with the equivalent linear frequency spectra.

A.4.15 Results option: Transfer function estimation

(User Guide, section 3.2.15; PESQ Tools only)

You may display the four linear frequency transfer function estimates (TFE) and a time-domain TFE for products in profile 3, and you may wish to offer it for profile 2. These complement the Bark scale TFE discussed in section A.4.3.

A.4.16 Results option: Signal spectrograms

(User Guide, section 3.2.16; PESQ Tools only)

You may display the linear signal spectrogram for products in all profiles. The LP spectrogram is more specialised, and therefore more suitable to profiles 2 and 3. The format that you use (for example, an image, where colour is related to loudness, or as a 3-D surface) is left to you.

A.4.17 Results option: LP excitation

(User Guide, section 3.2.17; PESQ Tools only)

The theory behind LP analysis requires an advanced understanding of digital signal processing theory. The LP excitation is therefore recommended for products where the user is likely to be interested in the properties of the speech signal.

A.4.18 Results option: Speech activity related outputs

(User Guide, section 0; PESQ Tools only)

These outputs will be of use to anyone interested in diagnosing the effects of error concealment algorithms or discontinuous transmission systems. The clipping statistics can be shown directly, whereas it is recommended that the clipping flags be plotted alongside the reference and degraded signals.

A.4.19 Results option: Speech diagnostic outputs

(User Guide, section 3.2.19; PESQ Tools only)

These outputs will be of use in products where the user may be interested in the properties of the speech signal, for example a tool to aid the development of speech coding algorithms. It is recommended that the pitch information and formants be plotted on a time-frequency axis pair. The power output and speech and voicing probabilities can be plotted alongside the time-frequency representation, using the same time axis.

A.5 Extensions to P.862

(User Guide, section 4)

You are recommended to include the information in section 4 of the User Guide when your users can use the additional facilities in PESQ for Head and Torso Simulator (HATS) ear measurements and wideband telephony measurements, for example in profile 3.

Please note that relatively little testing has been done on the performance of PESQ with these alternative models; performance results from 4 wideband telephony experiments are given in the User Guide. HATS and wideband applications should therefore be approached with care and after appropriate training.

A.6 Notes on speech signals

(User Guide, section 5)

You are recommended to include the information in section 5 of the User Guide as part of your user documentation for products designed for profile 2 or profile 3, unless the users are only going to be applying a test signal that was prepared by you according to the guidelines given in this section.

A.7 Overview of subjective testing

(User Guide, section 6)

You are recommended to include the information in section 6 of the User Guide as part of your user documentation for products designed for profile 2 or profile 3.

A.8 Noise testing

(User Guide, section 7)

You are recommended to include the information in section 7 of the User Guide as part of your user documentation for products designed for profile 2 or profile 3.

A.9 Comparison between objective and subjective results

(User Guide, section 8)

You are recommended to include the information in section 8 as part of your user documentation for products designed for profile 2 or profile 3.

A.10 Performance of PESQ

(User Guide, section 9)

You are recommended to include the information in section 9.1 as part of your user documentation for all profiles. Section 9.2 (wideband telephony model) is appropriate only to profile 3.

A.11 References

(User Guide, section 10)

You are recommended to include the information in this section as part of your user documentation for products designed for profile 3. You may also wish to include some of the references in documentation for products in profile 1 or profile 2.

A.12 Glossary

(User Guide, section 11)

You are recommended to include some or all of these terms, as appropriate, in a glossary in your own user documentation.

A.13 Document details

Your documentation should show your own company details. As specified in your license agreement, you must personalise your implementation of PESQ and all accompanying documentation with your own identity.

End of guidelines