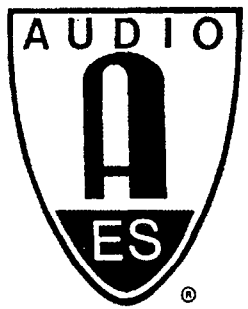
Antony W. Rix[1], John G. Beerends[2], Michael P. Hollier[1], and Andries P. Hekstra[2]

[1]BT Advanced Communications Research, Ipswich IP5 3RE, UK

[2]Royal PTT Nederland NV, NL-2260 Leidschendam, The Netherlands

Presented at
the 109th Convention
2000 September 22-25
Los Angeles, California, USA



AES

AN AUDIO ENGINEERING SOCIETY PREPRINT

PESQ – the new ITU standard for end-to-end speech quality assessment

Antony W. Rix[1], John G. Beerends[2], Michael P. Hollier[1] and Andries P. Hekstra[2]

[1] *BT Advanced Communications Research, B54/86 Adastral Park, Ipswich IP5 3RE, United Kingdom*
[2] *Royal PTT Nederland NV, NL-2260 Leidschendam, The Netherlands*
e-mail: awr@iee.org

This paper describes a new model for perceptual evaluation of speech quality (PESQ). This model is based on an integration of the perceptual speech quality measure (PSQM99) and the perceptual analysis measurement system (PAMS). PESQ is currently a draft ITU-T recommendation P.862, and is expected to replace P.861. PESQ provides a new international standard for objective assessment of speech codecs and end-to-end measurement of telephone networks.

0. Introduction

Speech codecs and other non-linear elements are becoming common in many audio communications systems. Conventional signal processing measures, such as frequency response or signal-to-noise ratio, cannot be reliably used to assess the perceived quality of these complex, non-linear systems. Likewise common synthetic test signals, such as impulses, white noise or sine waves, are inappropriate for testing speech codecs as their behaviour is highly signal-dependent: instead, speech-like signals must be used.

Until recently the only way to measure users' perception of the quality of such systems was to conduct a subjective test [1–3]. However, subjective tests are expensive and slow, and cannot be used in certain applications such as in-service monitoring. Objective models, based on human perception, were therefore developed with the aim of predicting the results of subjective tests.

Perceptual masking in audio coding was first proposed by Schroeder et al. [4], who suggested a model to estimate the audibility of coding noise. Brandenburg extended this to give a measure of noise to masking ratio (NMR) [5]. The concept of comparing internal loudness representations to determine perceived errors was introduced by Karjalainen [6].

Beerends and Stemerdink published several models based on comparison of internal representations to give a single measure of quality for audio or speech codecs [7–11]. Their method for assessing narrowband speech codecs, the perceptual speech quality measure (PSQM) [10], was selected after a competition held by the International Telecommunication Union (ITU-T) and adopted as ITU-T recommendation P.861 [12]. Their method for audio codecs, the perceptual audio quality measure (PAQM) [8, 11] was combined with a number of other audio models to produce a new model known as perceptual evaluation of audio quality (PEAQ), which has recently become ITU-R recommendation BS.1387 [13–15].

Wang et al. also used the comparison of internal representations to give a single measure of speech quality, the Bark spectral distortion (BSD) [16]. It was refined by Hollier, incorporating a bank of linear filters for spectral analysis and by taking account not only of the amount, but also the distribution of audible distortions [17, 18]. This model became the psychoacoustic core of the perceptual analysis measurement system (PAMS) [19, 20], a model designed for assessing telephone networks as well as speech codecs.

Codec assessment models such as PSQM [12] had limitations which made them unreliable when used in certain applications, especially for systems that include linear filtering and/or delay variations. After an ITU-T competition to select a new end-to-end speech quality assessment model, the two algorithms with the highest overall performance in the competition, PAMS and PSQM99 (an updated version of PSQM),

were combined. The new model is known as perceptual evaluation of speech quality (PESQ). Following validation against a large number of subjective tests including real network measurements, PESQ was determined in May 2000 as a new draft ITU-T recommendation P.862 [21]. It is expected that P.862 will replace P.861 early in 2001.

Section 1 of this paper presents the background of subjective and objective quality assessment, introducing PSQM and PAMS and describing the development of models by the ITU-T. In section 2, some conditions in which previous models gave inaccurate results are described, including variable delay and filtering. Section 3 gives an overview of the structure of PESQ. Results comparing PESQ with P.861 are given in section 4. Section 5 describes the range of conditions for which information on the performance of PESQ is currently available and summarises the areas in which further work remains to be done, then conclusions are drawn in section 6.

1. Background and development

1.1 Subjective evaluation of speech quality

The methodology of subjective testing for speech quality is set out in ITU-T recommendations P.800 and P.830 [1, 2]. These describe methods for evaluating one-way listening quality as well as two-way conversational quality. Listening tests are the most common because they are cheaper and easier to conduct than conversational tests; however, listening-only testing cannot take account of factors which only affect conversation, specifically conversational level, sidetone, talker echo and round-trip delay. This paper considers objective models designed to predict the results of listening-only subjective tests.

Several different subjective rating methods are defined in [1]. The simplest is the absolute category rating method (ACR). In this, subjects hear a number of degraded recordings, and are prompted to vote on each one according to an opinion scale such as the 5-point listening quality (LQ) scale shown in Table 1. Degradation and comparison category rating (DCR, CCR) methods are also described in [1]. In the DCR method the subjects hear first an undistorted reference followed by the degraded recording, and then vote on their opinion of the audibility and annoyance of the degradation. In the CCR method the reference may be presented either before or after the degraded recording, and subjects are asked to give their opinion on how much better, or worse, is the second file compared to the first. The ACR method with the LQ opinion scale is the most commonly used method in telecommunications assessment, and was the primary focus of development of PESQ.

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

Table 1: Listening quality opinion scale [1].

There are a number of common factors in subjective listening tests for telecommunications.

- Listening is in a quiet room with a controlled noise level.
- Subjects listen through a telephone handset with a standard response.
- Recordings are typically 8s long and consist of a pair of unrelated sentences.
- Tests are performed with speech from several different talkers (typically two male, two female) for each coding condition.
- Subjects are non-expert.

Once the test is complete the votes are averaged across subjects to give a mean opinion score (MOS). The quality of each condition is given by the condition MOS. In some cases it may also be useful to compute statistics for each file (file MOS) or each talker (talker MOS) within a given condition.

1.2 Perceptual speech quality measure (PSQM)

PSQM was developed by Beerends and Stemerdink from ideas first presented in [7] as a model optimised for the assessment of speech codecs [10, 12]. An overview of the structure of PSQM is shown in Figure 1. The core of the model is an auditory transform that models key psychophysical processes in the human auditory system. This computes a spectrogram-like internal representation of loudness in time and frequency as follows [12]:

- Short-term Fourier transform (STFT) with 50% overlapping Hann windows, 32ms long.
- Frequency warping of short-term power spectrum to 56-band Bark scale.
- Local scaling – partial equalisation of degraded signal to reference based on power of each frame, to account for low-frequency gain modulation.
- Filter with handset receive characteristic.
- Add Hoth noise.
- Loudness warping to a compressed Sone loudness scale.
- Loudness scaling to equalise degraded signal to same total loudness as reference signal in each frame.

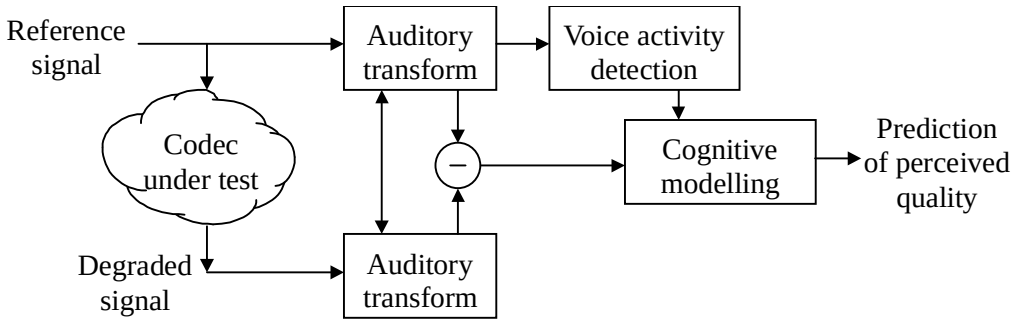


Figure 1: Structure of PSQM

Early approaches to estimating a single quality score were based on the average distance between the transforms of the reference and degraded signal [6, 7, 16]. PSQM introduced a "cognitive model" to interpret the difference between the transforms. This improves correlation between objective score and subjective MOS by modelling two effects: asymmetry and different weighting for speech and silence.

The asymmetry effect is caused by the fact that when a codec distorts the input signal it will in general be very difficult to introduce a new time-frequency component that integrates with the input signal, and the resulting output signal will thus be decomposed into two different percepts, the input signal and the distortion, leading to clearly audible distortion [22]. However, when the codec leaves out a time-frequency component the resulting output signal cannot be decomposed in the same way and the distortion is less objectionable. This effect is modelled in PSQM by multiplying the disturbance by a correction factor using the power ratio between the output signal and the input signal at a certain time-frequency point as a measure of "newness" of this component.

The fact that disturbances that occur during speech active periods are more disturbing than those that occur during silent intervals is modelled by a weighting factor that can be adjusted to the context of the experiment. Details of this procedure can be found in [12].

In the 1993-1996 study period ITU-T study group 12 compared five different objective models, LPC Cepstral Distance, Information Index, Coherence Function and Expert Pattern Recognition and PSQM, on

their ability to predict subjective quality scores. PSQM showed the highest performance on distortions which had not been used during training, and it was accepted in 1996 as ITU-T recommendation P.861 for objective quality measurement of narrowband speech codecs [12]. P.861 includes a detailed definition of its scope, which does not include live network testing or conditions in which variable delay (time warping) can occur. It also describes how the reference and degraded signals must be aligned in time and equalised to a fixed active speech level, corresponding to the standard level used in subjective listening tests.

Further work took place between 1996 and 1999 to improve PSQM and make it suitable for end-to-end testing of real networks, leading to a new version of the model, PSQM99, that incorporates time and level alignment routines.

1.3 Perceptual analysis measurement system (PAMS)

PAMS is an enhanced version of the model proposed by Hollier [17, 18]. This differs from the model of Wang et al. [16] by the use of a bank of linear filters – as opposed to the STFT – to implement time-frequency transformation and by a process termed the perceptual layer to interpret the error surface. Further development took place to provide time alignment, level alignment and equalisation functions essential for use in end-to-end measurement, and the perceptual layer was extended [20].

The model begins with time alignment, using a multi-stage process to align the reference and degraded signals. The signals are divided into sections known as utterances. Delay changes – for example due to packet-based transmission such as IP telephony – are identified. The signals are both equalised to a standard reference listening level corresponding to 79 dB SPL [2]. The auditory transform is then performed as follows.

- Input filter to model the response of the telephone handset, ear coupling and ear canal.
- Bank of linear filters to transform the signal to 19 Bark-spaced perceptual frequency bands.
- Computation of smoothed power envelope for each frequency band in 4ms frames.
- Transfer function estimation and equalisation to account for linear filtering in the system under test; the reference signal is partially equalised to the degraded signal.
- Mapping to phon loudness scale.
- Mapping to a Sone loudness scale (both phon and Sone loudness values are used in the model).

A number of error parameters are computed based on the auditory transforms of the reference and degraded signals, giving a measure of the amount of different classes of distortion. These are averaged in time and are mapped to quality score through a non-linear function, preserving a monotonic relation between each parameter and quality score. This process is outlined in Figure 2. Two quality measures are computed, one on the ACR listening quality opinion scale (Table 1) and the other on the ACR listening effort opinion scale [1].

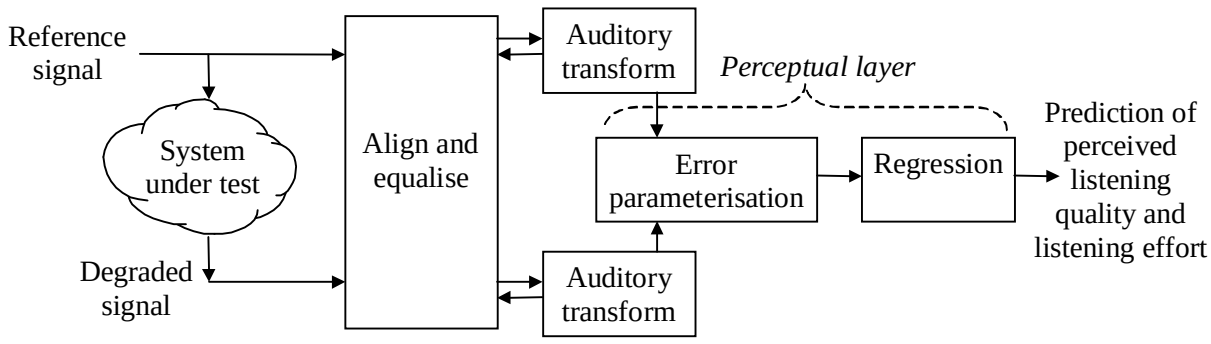


Figure 2: Structure of PAMS

The components that implement variable delay time alignment and transfer function equalisation were designed to enable PAMS to be used in end-to-end measurement applications [19, 20]. With appropriate changes to the input filter to model headphone listening – but with no alterations to the underlying auditory model and perceptual layer – PAMS has also been extended to testing monophonic wideband telephony at 16kHz sample rate [23].

1.4 Further development of perceptual models in the ITU-T

Following the identification of certain conditions in which P.861 PSQM was found to show poor correlation with subjective opinion, another algorithm, Measuring Normalizing Blocks (MNB), was proposed. This was added to P.861 as an informative – but not binding – appendix II in 1998 [12]. However, neither P.861 PSQM nor MNB were found to be suitable for end-to-end measurement of networks. Particular problems were observed with conditions that introduce significant linear filtering, variable delay, or additive background noise. To address these problems, a competition to select a new end-to-end assessment model was conducted by ITU-T study group 12 between September 1998 and March 2000. The competition was unable to identify a single outright winner as it was difficult to distinguish the two models with the highest overall performance. These models, PAMS and PSQM99, were combined and further work was carried out to meet a demanding set of requirements.

The new model, perceptual evaluation of speech quality (PESQ), was found to meet the requirements and to have significantly better performance than P.861 PSQM and MNB, even for tests including only speech codecs. PESQ was therefore determined in May 2000 by ITU-T study group 12 as a new draft ITU-T recommendation P.862 [21]. It is anticipated that P.862 will be approved early in 2001, at which point P.861 is expected to be withdrawn.

2. Weaknesses of previous models

2.1 Variable delay (time warping)

2.1.1 Processes causing delay variation

Large delays impair two-way conversation, so it is desirable to minimise end-to-end delay. However packet-based transmission often leads to each packet being delayed by a different amount. A buffer is required to iron out these delay variations and produce a continuous audio stream. It is necessary to balance the length of the buffer – a major addition to end-to-end delay – with the possibility of packet loss.

Two different processes appear to lead to delay variations in voice over IP (VoIP). Dynamic buffer resizing during silence is a common method for dealing with time-varying packet delay by changing the buffer length – and hence delay – during silent intervals. In a less frequent effect, large changes in packet delay can cause buffers to overrun or become empty, leading to delay changes during speech.

The magnitude of delay variations encountered in measurements of delay changes on two packet-based networks are presented in Figure 3. This shows the distribution of changes in end-to-end (audio) delay during a series of 16-second measurements on each type of network. Figure 3(a) gives results from a field trial of a PC-based VoIP system. Figure 3(b) shows the corresponding distribution measured across a PSTN to Internet Gateway system. Delay changes of up to 100ms in magnitude were encountered [19].

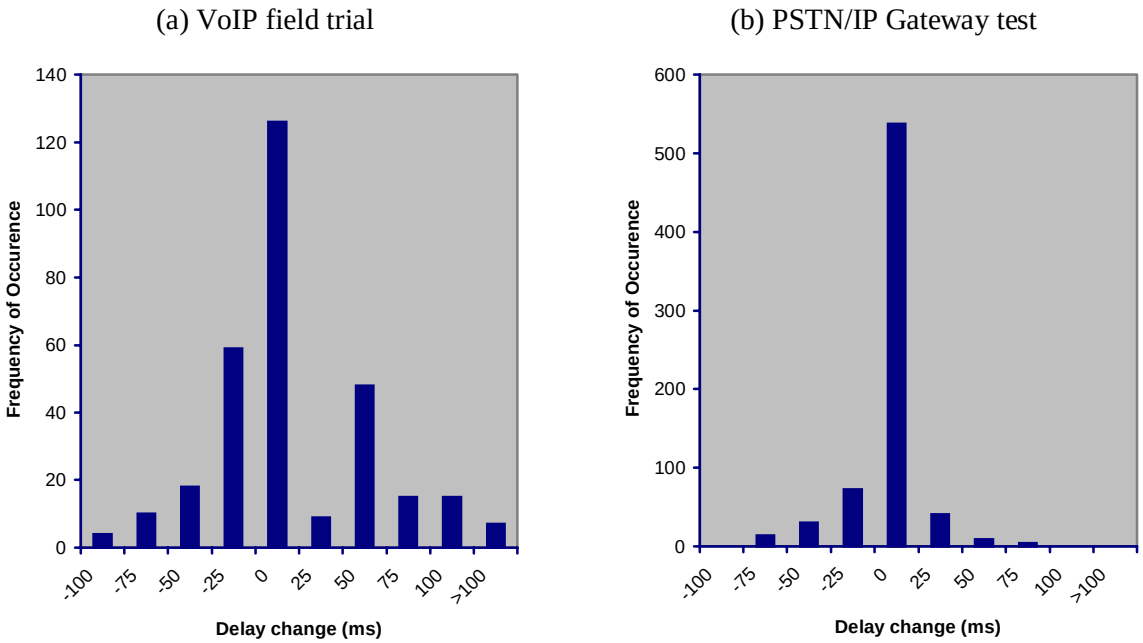


Figure 3: Delay variations measured in voice over IP network connections

2.1.2 Effect of variable delay on perceptual models

The sensitivity to variable delay of codec assessment models such as PSQM and MNB was evaluated in two ways. In the first investigation delay changes in silent periods were introduced. This models the effect of dynamic buffer resizing during silence. Delay changes such as these are not noticed by subjects unless they are very large (e.g. 0.5s) or cause insertion or deletion during speech events. However, ignoring delay variations causes an objective model to compare different sections of the reference and degraded signals. As speech signals are highly time-varying, this causes large false errors to be measured. It was found that a delay change of 20ms is sufficient to cause PSQM [12] to measure a drop in quality equivalent to approximately 1 MOS. MNB [12 appendix II] is even more sensitive, requiring a change of only 5ms to cause an equivalent drop in quality [19].

The second investigation was based on a subjective test in which the distortions included delay changes. The correlation between subjective and objective score was measured using the procedure described in section 4.1 and is summarised in Table 2. The models of P.861, PSQM and MNB, show very low correlation with subjective opinion and clearly give meaningless scores in this test. However, PESQ, which is designed to take variable delay into account, shows a much higher correlation.

Model	*Correlation*
PSQM [12]	0.260
MNB [12]	0.363
PESQ	0.932

Table 2: Correlation between subjective and objective score for VoIP variable delay test. Per condition, after monotonic 3rd-order polynomial mapping.

2.2 Linear filtering

2.2.1 Frequency response of typical network components

Many components used in telephony introduce significant amounts of linear filtering. This can occur in the acoustic path, at 2-wire line interfaces, or even in speech codecs. To give an idea of the magnitude of filtering that occurs in end-to-end measurement, Figure 4 shows the frequency response of two typical network components. Figure 4(a) shows the modified IRS send characteristic [2], which represents the frequency response from mouth to junction through a typical telephone handset. Figure 4(b) shows the response measured by a typical test device, with 2-wire interfaces, on a telephone connection in the UK [19]. These responses are quite typical and show that gain within the 300–3,400Hz passband can vary by ±10dB.

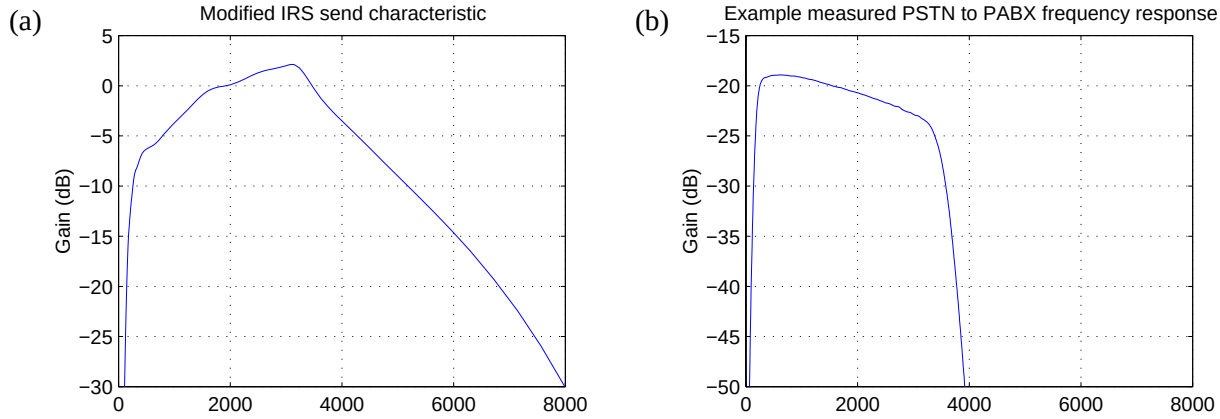


Figure 4: Frequency responses of telephone network components

2.2.2 Effect of filtering on perceptual models

Although linear filtering does have some subjective effect, it is generally much less significant than non-linear coding distortion. The early perceptual models such as BSD [16], PSQM and MNB [12] made no distinction and therefore measure large errors due to filtering alone. It is recognised that perceptual models for use with end-to-end audio systems must give less weight to linear distortions. This is usually achieved by equalising the reference signal to the degraded signal. Note that standard linear transfer function estimation/equalisation techniques cannot normally be applied because they are unstable with low bit-rate speech codecs [19]. Several approaches can ensure that audible filtering is not completely ignored. Partial compensation eliminates most of the effect but leaves some residual distortion to be measured by the perceptual model; this is the method used in PAMS, PSQM99 and PESQ. If the filtering is fully compensated, separate linear distortion measures can be used as part of the final regression to subjective MOS, the method used in PEAQ [13, 14].

The effect of filtering is illustrated by the correlation, given in Table 3, between subjective and objective MOS for a subjective test on low bit-rate mobile codecs. In half of the conditions in this test an IRS send filter similar to that of Figure 4(a) was applied; in the other conditions there was no filtering. PSQM and MNB [12] perform quite badly because the objective scores given to the filtered conditions are very different from the unfiltered conditions. PESQ take good account of the filtering, resulting in a much higher correlation than the models of P.861 [12].

Model	*Correlation*
PSQM [12]	0.616
MNB [12]	0.626
PESQ	0.914

Table 3: Correlation between subjective and objective score for mobile codec test with filtering. Per condition, after monotonic 3rd-order polynomial mapping.

2.3 Gain variation

Although uncommon in today's telephone networks, it is possible for speech to be subject to forms of low-frequency amplitude modulation. This can occur with automatic gain control (AGC), in which speech levels are dynamically adjusted towards a standard level. The intention is to cancel the effect of variable losses in subscriber equipment or level variations between different countries' networks. Sometimes, however, undesirable gain changes occur as a consequence of background noise or normal variations in vocal level. Most current AGC systems change level slowly, and only in silent periods, but some systems have been found to operate during speech.

The subjective effect of AGC is usually quite limited. Because speech is naturally time-varying, it is difficult to detect gain changes of less than 3dB, and gain changes of up to 10dB are not generally very disturbing as long as audible discontinuities are avoided. However, as perceptual models are based on a comparison of internal loudness representations, it is necessary to track and equalise gain variations otherwise large errors would be measured for a relatively inaudible effect.

Models differ widely in their approach to modulation. It is essentially ignored by MNB. PAMS identifies and cancels out gain changes only if they occur during silence, but gain changes during speech cause large errors to be measured. PSQM, PSQM99 and PESQ adaptively track envelope changes frame-by-frame, cancelling the effect with a lag to ensure that some errors are measured due to gain changes.

A number of tests used in the validation of PESQ contained gain variation; some of these results are reproduced in section 4.2. It appears that the adaptive method employed in PESQ takes good account of the subjective impact of gain variation.

2.4 Temporal clipping

A general term for the replacement of speech by silence, temporal clipping can occur with many different networks, including international, mobile and VoIP. The most common type is front-end clipping, where the first few milliseconds of speech are not transmitted. Back-end clipping is the corresponding effect where the end of a speech utterance is truncated. This is usually the result of voice activity detection errors with discontinuous transmission in silence (DTX), where transmission ceases when speech becomes inactive to free up capacity. DTX is usually accompanied by comfort noise insertion, which aims to re-create background noise of similar spectrum so that the listener does not notice the effect. Clipping may also occur during speech, for example if a packet is lost and is replaced by silence.

Different forms of temporal clipping appear to have very different subjective effect. Front-end clipping often has little perceived effect; in some tests, even 50ms of front-end clipping was not noticed by the subjects. However, in other tests where semantically important parts of speech were deleted by front-end clipping or by packet loss, as little as 10ms of clipping is found to be perceptually significant. It is clearly impossible for an objective model to predict these conflicting effects. Most current models predict that temporal clipping is significant, giving a drop of quality in the region of 0.5 MOS for around 50ms of front-end clipping, and this is probably the best that can be done with this type of distortion.

3. Structure of PESQ

An overview of PESQ is shown in Figure 5. The model begins by level aligning both signals to a standard listening level. They are filtered (using an FFT) with an input filter to model the telephone handset. The signals are aligned in time and are then processed through an auditory transform similar to that of PSQM. Part of the transformation involves equalising the signals for the frequency response of the system and for gain variation. The difference between the transforms of the reference and degraded signals is known as the disturbance. This is processed to extract two distortion parameters, which are aggregated in frequency and time and mapped to a prediction of subjective MOS. The details of the time alignment, transformation and disturbance processes are discussed in the following sections.

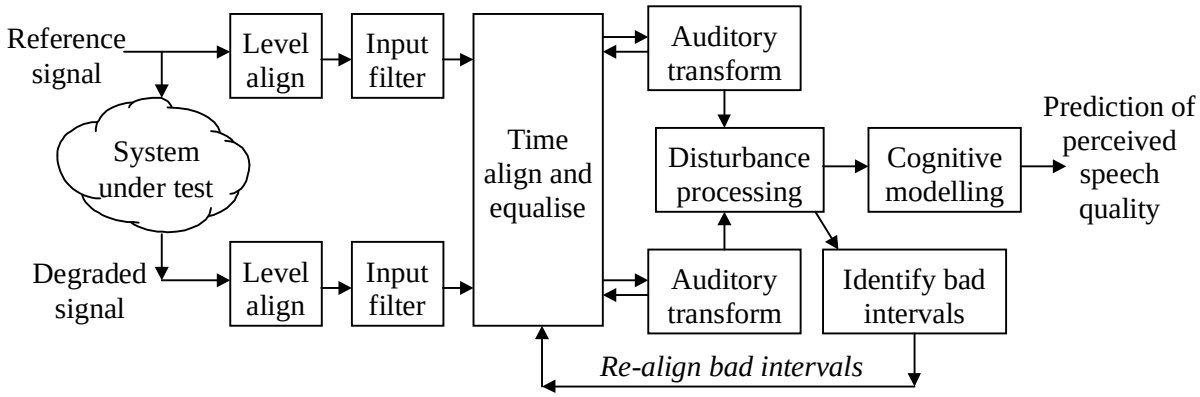


Figure 5: Structure of PESQ

3.1 Time alignment

The time alignment of PESQ assumes that the delay of the system is piecewise constant. Delay changes are allowed in silent periods (where they will normally be inaudible) and in speech (where they are usually audible). The signals are aligned using the following steps [20].

- Narrowband filter applied to both signals to emphasise perceptually important parts. These filtered signals are only used for time alignment.
- Envelope-based delay estimation.
- Division of reference signal into utterances.
- Envelope-based delay estimation for each utterance.
- Fine correlation histogram-based delay identification for each utterance.
- Utterance splitting and re-alignment to test for delay changes during speech.

The result is a number of sections with a given start, end, and delay. This is translated to a frame-by-frame delay for use in the auditory transform.

3.2 Auditory transform

The auditory transform of PESQ is a psychoacoustic model which maps the signals into a representation of perceived loudness in time and frequency.

Bark spectrum. A STFT with a Hamming window is used to calculate the instantaneous power spectrum in each frame, for 50% overlapping frames of 32ms duration. This is grouped without smearing into 42 bins, equally spaced in perceptual frequency on a modified Bark scale similar to that of PSQM [12].

Frequency equalisation. The mean Bark spectrum for active speech frames is calculated. The spectral difference gives an estimate of the transfer function, assuming that the system under test has a constant

frequency response. The reference is equalised to the degraded signal using this response, with bounds to limit the equalisation to ±20dB.

Equalisation of gain variation. The ratio between the audible power of the reference and the degraded in each frame is used to identify gain variations. This is filtered with a first-order low-pass filter, and bounded, then the degraded signal is equalised to the reference.

Loudness mapping. The Bark spectrum is mapped to (Sone) loudness, including a frequency-dependent threshold and exponent. This gives the perceived loudness in each time-frequency cell.

3.3 Disturbance processing and cognitive modelling

The absolute difference between the degraded and the reference signals gives a measure of audible error. In PESQ, this is processed through several steps before a non-linear average over time and frequency is calculated.

Deletion. If deletion occurs (a negative delay change) there will be a section which overlaps in the degraded signal. If the deletion is longer than half a frame, the overlapping sections are discarded.

Masking. Masking in each time-frequency cell is modelled using a simple threshold below which disturbances are inaudible; this is set to the lesser of the loudness of the reference and degraded signals, divided by four. The threshold is subtracted from the absolute loudness difference, and values less than zero are set to zero. Methods for applying masking over distances larger than one time-frequency cell were examined with earlier versions of PSQM and PSQM99, but did not improve overall performance [9], and were not used in PESQ.

Asymmetry. Unlike P.861, PESQ computes two different error averages, one without and one with an asymmetry factor. The PESQ asymmetry factor is calculated from a stabilised ratio of the Bark spectral density of the degraded to the reference signals in each time-frequency cell. This is raised to the power 1.2 and is bounded with an upper limit of 12.0. Values smaller than 3.0 are set to zero. The asymmetric weighted disturbance, obtained by multiplying by this factor, thus measures only additive distortions.

3.4 Aggregation of disturbance in frequency and time

Following the concept that localised errors dominate perception [18, 24], PESQ integrates disturbance over several scales using a method designed to take optimal account of the distribution of error in time and amplitude. The disturbance values are aggregated using an *Lp* norm, which calculates a non-linear average using the following formula:

$$L_p = \left(\frac{1}{N} \sum_{m=1}^{N} disturbance[m]^p \right)^{1/p}$$

The disturbance is first summed across frequency using an *Lp* norm, giving a frame-by-frame measure of the perceived distortion. This frame disturbance is multiplied by two weightings. The first weight is inversely proportional to the instantaneous energy of the reference, raised to the power 0.04, giving slightly greater emphasis on sections for which the reference is quieter. This process replaces the silent interval weighting used in P.861. After this, the frame disturbance is bounded with an upper limit of 45. The second weight gives reduced emphasis on the start of the signal if the total length is over 16s, modelling the effect of short-term memory in subjective listening. This multiplies the frame disturbance at the start of the signal by a factor decreasing linearly from 1.0 (for files shorter than 16 seconds) to 0.5 (for files longer than 60 seconds).

After weighting, the frame disturbance is averaged in time over split second intervals of 20 frames (approx 320ms, accounting for the overlap of frames) using *Lp* norms. These intervals overlap 50%, and no window function is used. The split second disturbance values are finally averaged over the length of the

speech files, again using *Lp* norms. Thus the aggregation process uses three *Lp* norms – in general with different values of *p* – to map the disturbance to a single figure. The value of *p* is higher for averaging over the split second intervals to give greatest weight to localised distortions. The symmetric and asymmetric disturbance are averaged separately.

3.5 Realignment of bad intervals

In certain cases the first time alignment may fail to correctly identify a delay change, resulting in large errors for each section with incorrect delay. These are identified by labelling bad frames (which have a symmetric disturbance of more than 45) and joining together bad sections in which bad frames are separated by less than 5 good frames.

Each bad section is then realigned and the disturbance recalculated. Cross-correlation is used to find a new delay estimate. The auditory transform of the degraded signal is recalculated and the disturbance found. For each frame, if the realignment results in a lower disturbance value, the new value is used. Aggregation over split second intervals and the whole signal is performed after realignment.

3.6 MOS prediction and model calibration

To train PESQ a large number of different symmetric and asymmetric disturbance parameters were calculated by using different values of *p* for each of the three averaging stages. A linear combination of disturbance parameters was used as a predictor of subjective MOS. A further regression is required for each subjective test to account for context and voting preferences of different subjects, as discussed in section 4.1; for calibration a linear mapping was also used at this stage. Parameter selection was performed for all candidate sets of up to four disturbance parameters. The optimal combination – giving the highest average correlation coefficient – was found. This enabled the best parameters to be chosen from the full set of several hundred candidate disturbance parameters.

The partial compensation method used in PESQ means that it is not necessary to use many separate parameters to predict quality, for example to take account of filtering, modulation, or distribution of errors. Thus it was found that a combination of only two parameters – one symmetric disturbance and one asymmetric disturbance – gave a good balance between accuracy of prediction and ability to generalise. However, as this low-dimension model depends on earlier stages to incorporate complex perceptual effects, it was necessary to perform several design iterations. Coefficients in the auditory transform and disturbance processing were optimised then the optimal parameter combination was found, and the process repeated several times. The final training was performed on a database of 30 subjective tests.

The output mapping used in PESQ is given by

$$PESQMOS = 4.5 - 0.1\, disturbance_{SYMMETRIC} - 0.0309\, disturbance_{ASYMMETRIC}$$

For normal subjective test material the values lie between 1.0 (bad) and 4.5 (no distortion). In cases of extremely high distortion the *PESQMOS* may fall below 1.0, but this is very uncommon.

4. Performance results

4.1 Evaluation of performance

Condition MOS is the most common measure of subjective quality: this is the average MOS for four or more recordings of at least 8s in duration. These recordings are usually different sentence pairs spoken by two male and two female talkers; the condition MOS is therefore a material-independent measure of the quality of the connection. For comparison between objective and subjective score it is usual to compare the condition MOS with the condition average objective score.

However, a one-to-one comparison between objective and subjective MOS is not normally possible with tests conducted according to the ITU-T testing method [1, 2], because subjective votes are affected by factors such as the voting preferences of each subject or the balance of conditions in a test. This makes it impossible to directly compare results from one subjective test with another; some form of mapping between the two is required. The same is true for comparing objective scores with subjective MOS.

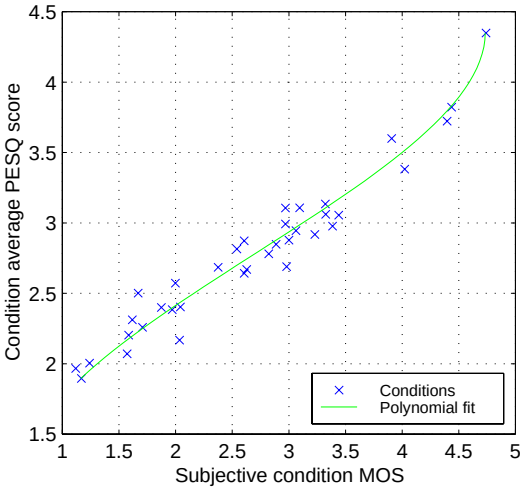
However, it is reasonable to expect that order should be preserved, so the difference between two sets of scores should be a smooth, monotonically increasing (one-to-one) mapping. The function used in ITU-T evaluation of objective models is a monotonic 3rd-order polynomial. This is applied, for each subjective test, to map the objective score onto the subjective score. It is then possible to calculate correlation coefficient and residual errors.

This process is illustrated by the following example, a subjective test on the performance of fixed and mobile networks with errors, noise and noise suppression. Figure 6(a) shows a scatter plot between subjective MOS and PESQ score, along with the monotonic 3rd-order polynomial fit with minimum mean squared error. The PESQ score is mapped by this polynomial to give a prediction of subjective quality, shown in Figure 6(b). The correlation coefficient for this test is 0.974, given by the following equation

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

where x_i is the condition MOS for condition i, $\bar{x}$ is the average of x_i across all conditions, y_i is the mapped condition-averaged PESQ score for condition i, and $\bar{y}$ is the average of y_i across all conditions.

(a) Raw scores and monotonic polynomial fit



(b) Scores after application of mapping

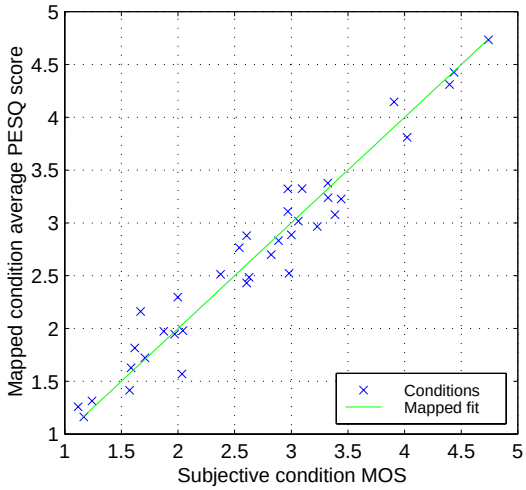


Figure 6: Mapping between objective and subjective MOS.

4.2 Correlation results

The performance of PESQ is compared to PSQM [12] and MNB [12 appendix II] in Figures 7–10 using correlations calculated according to the process described in the previous section. The figures plot the correlation coefficient between each model and subjective MOS for a number of ACR listening quality tests. Figure 7 presents 19 tests containing mainly mobile codecs and/or networks. Figure 8 gives results from 9 tests on predominantly fixed networks or codecs. Figure 9 shows 10 tests containing VoIP conditions on a wide range of codec/error types. Finally, Figure 10 gives the results for 8 tests conducted on PESQ by independent laboratories using data unknown in the development of the model. The tests

were conducted in a number of different languages, and 8 of the tests included conditions with background noise.

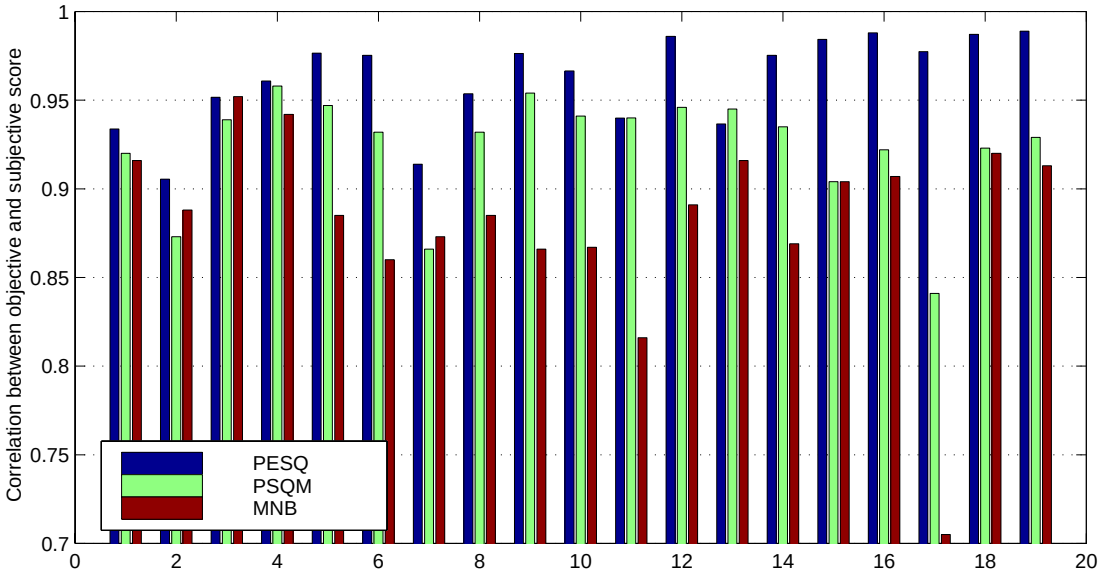


Figure 7: Mobile network performance results for PESQ, PSQM [12], and MNB [12]. Correlation coefficient, per condition, after monotonic 3rd-order polynomial mapping.

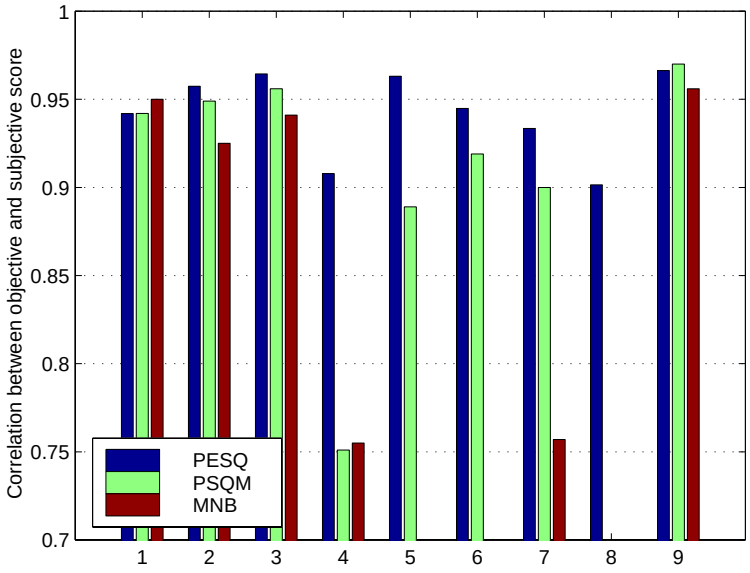


Figure 8: Fixed network performance results for PESQ, PSQM [12], and MNB [12]. Correlation coefficient, per condition, after monotonic 3rd-order polynomial mapping. In tests 5, 6 and 8 the scores for MNB (and PSQM in test 8) are off the bottom of the scale.

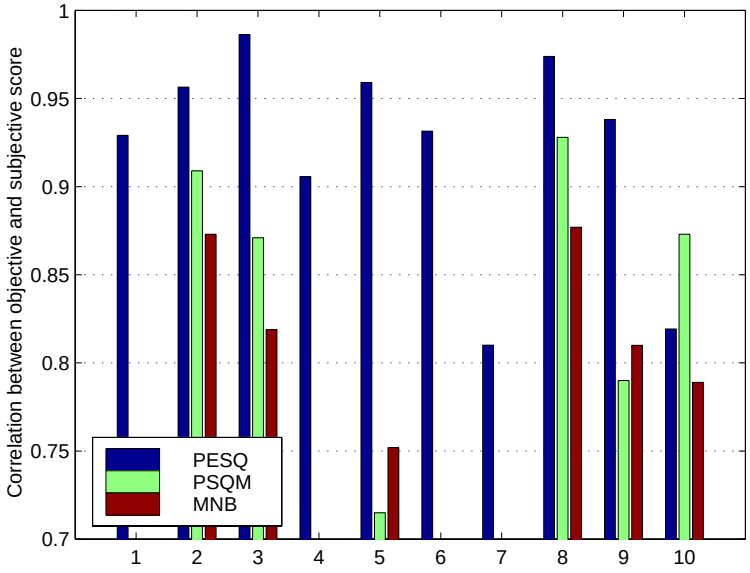


Figure 9: VoIP and multi-type test results for PESQ, PSQM [12], and MNB [12]. Correlation coefficient, per condition, after monotonic 3rd-order polynomial mapping. In tests 1, 4, 6 and 7 the scores for MNB and PSQM are off the bottom of the scale.

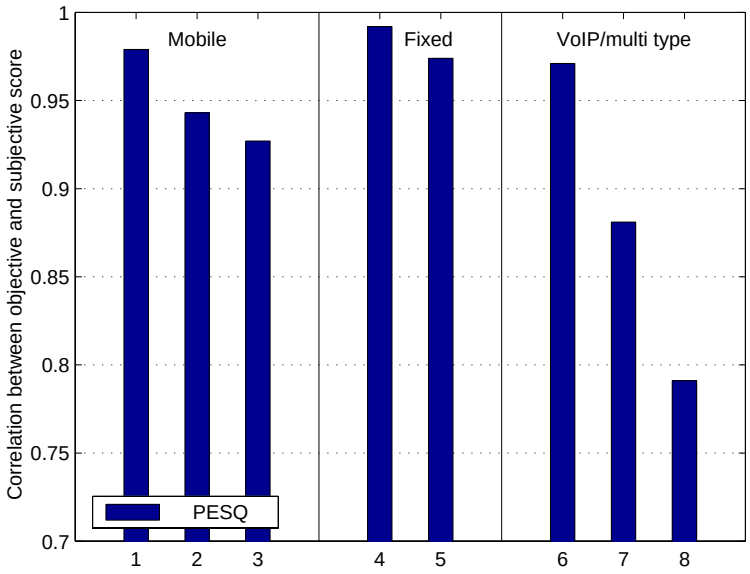
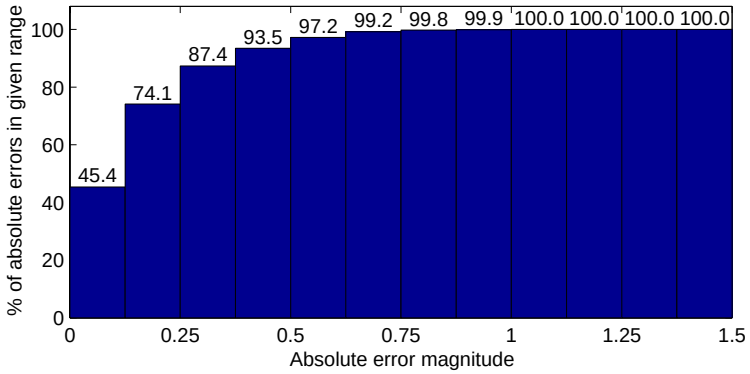


Figure 10: Independent results for unknown subjective tests (PESQ only). Correlation coefficient, per condition, after monotonic 3rd-order polynomial mapping.

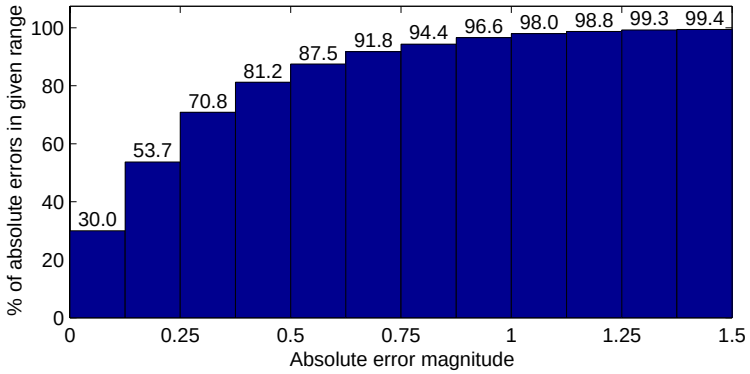
4.3 Residual error distribution

A further method for measuring model performance is to plot the distribution of the absolute residual errors $|x_i - y_i|$ after the mapping described in section 4.1. Figures 11 plots the cumulative distribution of errors for PESQ, PSQM [12] and MNB [12 appendix II], calculated across 40 ACR listening quality tests containing a total of 1921 conditions. This shows, for example, that 93.5% of PESQ scores were within 0.5 MOS of the subjective score, and 100% of PESQ scores were within 1.125 MOS of the subjective score for these 40 tests.

(a) PESQ



(b) PSQM [12]



(c) MNB [12]

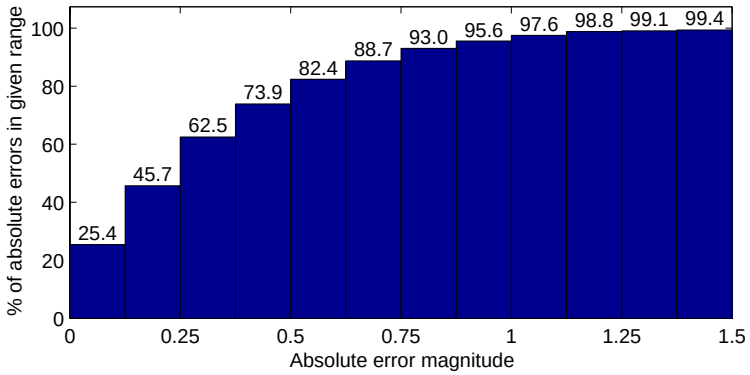


Figure 11: Residual error distribution for PESQ, PSQM [12], and MNB [12]. Per condition, after monotonic 3rd-order polynomial mapping.

5. Scope of PESQ

5.1 Conditions successfully tested

Table 4 presents a summary of the range of conditions for which PESQ has been tested and found to give acceptable performance. Full details of the scope of the model may be found in P.862 [21].

Test factors	Coding/network technologies	Measurement applications
Coding distortions Transmission/packet loss errors Multiple transcodings Environmental noise * Time warping (variable delay)	Waveform codecs (e.g. G.711, G.726, G.727) CELP/hybrid codecs at 4kbit/s and above (e.g. G.728, G.729, G.723.1) Mobile codecs and systems (e.g. GSM FR, EFR, HR, AMR; CDMA EVRC, TDMA ACELP, VSELP; TETRA)	Live network testing Network planning Codec evaluation/selection Equipment selection Codec/equipment optimisation

Table 4: Factors for which PESQ can be used for objective speech quality measurement.

* Note: for testing the effect of environmental noise, PESQ should be presented with the clean, unprocessed original and the noisy, coded, degraded signal.

5.2 Problems and areas for which PESQ is not applicable

PESQ is not intended to be used to assess:

- effect of listening level
- conversational delay
- talker echo/sidetone
- non-intrusive measurements.

Additionally, problems have been found with measurements on systems that replace speech with silence, for example front-end clipping or packet loss concealment with silence. See section 2.4 for more discussion of this.

5.3 Areas for further work

Certain applications of PESQ are currently under study or may require changes to the model, for example:

- wideband telephony/conferencing (16kHz sample rate)
- listener echo
- very low bit-rate speech vocoders (below 4kbit/s)
- head and torso simulator (HATS) measurements of handsets and/or hands-free telephones
- assessment of music.

One goal of further development is to extend the range of signal types and quality levels that a model can be used to assess. At present PESQ is calibrated to predict subjective tests conducted according to ITU-T P.800 or P.830 [1, 2] – i.e. “telephone quality”, where subjects listen through a standard narrowband telephone handset. PEAQ [13, 14] is able to measure the quality of audio codecs – “audio quality” – for applications such as broadcast, with headphone or loudspeaker listening [3]. In between these two ranges is the so-called “intermediate quality” [23]. It is hoped that PESQ can be extended to provide assessment at both telephone and intermediate quality.

6. Summary and conclusion

PESQ performs much better than earlier codec assessment models such as P.861 PSQM and MNB, and is expected to replace them in early 2001 as a new ITU-T recommendation P.862. PESQ has been evaluated on a very wide range of speech codecs and telephone network tests. It has been found to produce accurate predictions of quality in the presence of diverse end-to-end network behaviours such as filtering and variable delay. PESQ represents a significant step forward in the accuracy and range of applicability of speech quality assessment models. It can be used for development, selection and optimisation of telephone network equipment and codecs, as well as for measurement applications such as network monitoring.

7. Acknowledgements

Thanks are due to ITU-T study group 12 question 13 for organising and driving the recent competition, and in particular the other proponents (Ascom, Deutsche Telekom and Ericsson) who contributed valuable test data and provided stiff competition. We would also like to thank the companies who acted as independent validation laboratories: AT&T, Lucent Technologies, Nortel Networks, and especially France Telecom R&D. We acknowledge the assistance of many of our colleagues at BT and KPN. Antony Rix is also supported by the Royal Commission for the Exhibition of 1851.

8. References

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

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

[3] Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. ITU-R Recommendation BS.1116, July 1998.

[4] Schroeder, M. R., Atal, B. S. and Hall, J. L. "Optimizing digital speech coders by exploiting masking properties of the human ear". Journal of the Acoustical Society of America, 66 (6), 1647–1652, December 1979.

[5] Brandenburg, K. "Evaluation of quality for audio encoding at low bit rates". 82nd AES Convention, pre-print no. 2433, 1987.

[6] Karjalainen, J. "A new auditory model for the evaluation of sound quality of audio systems", IEEE ICASSP, 608–611, 1985.

[7] Beerends, J. G. and Stemerdink, J. A. "Measuring the quality of audio devices". 90th AES Convention, pre-print no. 3070, 1991.

[8] Beerends, J. G. and Stemerdink, J. A. "A perceptual audio quality measure based on a psychoacoustic sound representation". Journal of the AES, 40 (12), 963–974, December 1992.

[9] Beerends, J. G. and Stemerdink, J. A. "The optimal time-frequency smearing and amplitude compression in measuring the quality of audio devices". 94th AES Convention, pre-print no. 3604, 1993.

[10] Beerends, J. G. and Stemerdink, J. A. "A perceptual speech-quality measure based on a psychoacoustic sound representation". Journal of the AES, 42 (3), 115–123, March 1994.

[11] Beerends, J.G. "Measuring the quality of speech and music codecs, an integrated psychoacoustic approach". 98th AES Convention, pre-print no. 3945, 1995.

[12] Objective quality measurement of telephone-band (300–3400 Hz) speech codecs. ITU-T Recommendation P.861, February 1998.

[13] Method for objective measurements of perceived audio quality. ITU-R Recommendation BS.1387, January 1999.

[14] Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K. and Feiten, B. "PEAQ–The ITU standard for objective

measurement of perceived audio quality". Journal of the AES, 48 (1/2), 3–29, January/February 2000.

[15] Treurniet, W. C. and Soulodre, G. A. "Evaluation of the ITU-R objective audio quality measurement method". Journal of the AES, 48 (3), 164–173, March 2000.

[16] 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, June 1992.

[17] Hollier, M. P., Hawksford, M. O. and Guard, D. R. "Characterisation of communications systems using a speech-like test stimulus", Journal of the AES, 41 (12), 1008–1021, December 1993.

[18] 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, June 1994.

[19] Rix, A. W., Reynolds, R. and Hollier, M. P. "Perceptual measurement of end-to-end speech quality over audio and packet-based networks". 106th AES Convention, pre-print no. 4873, May 1999.

[20] Rix, A. W. and Hollier, M. P. "The perceptual analysis measurement system for robust end-to-end speech quality assessment", IEEE ICASSP, June 2000.

[21] Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Draft Recommendation P.862, May 2000.

[22] Beerends, J.G. "Modelling cognitive effects that play a role in the perception of speech quality", in Proc. Int. Workshop on Speech Quality Assessment, Bochum, pages 1-9, November 1994.

[23] Rix, A. W. and Hollier, M. P. "Perceptual speech quality assessment from narrowband telephony to wideband audio", 107th AES Convention, pre-print no. 5018, September 1999.

[24] Quackenbush, S.R., Barnwell III, T.P., Clements, M.A. Objective measures of speech quality. Prentice Hall Advanced Reference Series, New Jersey USA, 1988.