Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II – Psychoacoustic model

J. G. Beerends (1), A. P. Hekstra (1), A. W. Rix (2), and M. P. Hollier (2)

(1) Royal PTT Nederland NV, P.O. Box 421, NL - 2260 AK Leidschendam, The Netherlands. A. P. Hekstra is now with Philips Research (WY-61), Prof.Holstlaan 4, NL - 5656 AA Eindhoven

(2) Psytechnics Limited, 23 Museum Street, Ipswich IP1 1HN, United Kingdom. Psytechnics was formerly part of BT Laboratories.

Abstract

A new model for perceptual evaluation of speech quality (PESQ) was recently standardised by the ITU-T as recommendation P.862. Unlike previous codec assessment models, such as PSQM and MNB (ITU-T P.861), PESQ is able to predict subjective quality with good correlation in a very wide range of conditions, that may include coding distortions, errors, noise, filtering, delay and variable delay. This paper introduces the psycho-acoustic model that is used in PESQ. An accompanying paper describes the time delay identification technique that is used in combination with the PESQ psychoacoustic model to predict the end-to-end perceived speech quality.

List of Abbreviations

ACELP	Adaptive CELP
ACR	Absolute Category Rating
AMR	Adaptive Multi Rate (GSM codec)
ATM	Asynchronous Transfer Mode
CELP	Code Excited Linear Prediction
CDMA	Code Division Multiple Access
dB	deciBell
EFR	Enhanced Full Rate (GSM codec)
ETSI	European Telecommunications Standards Institute
EVRC	Enhanced Variable Rate Codec
FFT	Fast Fourier Transform
FR	Full Rate (GSM codec)
GSM	Global System for Mobile Communications
HATS	Head And Torso Simulator
HR	Half Rate (GSM codec)
IP	Internet Protocol
IRS	Intermediate Reference System
ITU-R	International Telecommunication Union-Radio sector
ITU-T	International Telecommunication Union-Telecom sector
MNB	Measuring Normalized Blocks [3, appendix II]
MOS	Mean Opinion Score
PAMS	Perceptual Analysis Measurement System
PEAQ	Perceptual Evaluation of Audio Quality
PESQ	Perceptual Evaluation of Speech Quality
PSQM	Perceptual Speech Quality Measure [3]
PSQM99	Perceptual Speech Quality Measure 1999 version
SPL	Sound Pressure Level
TDMA	Time Division Multiple Access
TETRA	Trans European Trunked Radio
VSELP	Vector Sum Excited Linear Predictive Coding

1 INTRODUCTION

With the introduction and standardization of new technologies for telephony services that introduce new types of distortions, like Voice over IP (packet loss and variable delay), Voice over ATM (cell loss), voice over mobile (GSM, UMTS, frame repeat, front end clipping, comfort noise generation) and speech coding (ETSI GSM EFR/AMR, ITU-T G.728/729/723.1 etc) classical quality measurement techniques, using concepts like signal to noise ratio, frequency response functions etc, have become grossly inaccurate.

In fact the whole idea of system characterization, mostly carried out on the basis of a nearly linear, time invariant system, loses meaning with these new technologies. An alternative, perception based, approach has been developed in the last decade. The basic idea of this approach is to take the signal adaptive properties of the system under test into account by feeding it with real world signals and measure the perceptual quality of the output signals. In the case of telephony the signals are usually speech signals, with or without background noise.

If the subjective quality of the output of a non-linear, signal adaptive, time variant system is assessed using the perception based approach one has to be aware that no single number can be attached to the quality of the system under test. Although this is sometimes viewed as a disadvantage one can state that having access to an objective method that can assess the quality under different signal inputs is an advantage over classical approaches because one can exploit the range of signals for which the system under test behaves correctly from a perception point of view.

The first international standard for the perceptual quality measurement of telephone-band (300-3400 Hz) speech signals was PSQM (Perceptual Speech Quality Measure [1–3]) which was benchmarked by the ITU-T. In this benchmark the PSQM method showed the highest correlations between objective and subjective measurements in comparison to four other proposals [2]. The method was standardized as ITU-T recommendation P.861 in 1996 [3]. However the scope of recommendation P.861 was limited to the assessment of telephone-band speech codecs only.

A corresponding international standard for the perceptual quality measurement of wide-band (20-20000 Hz) audio signals is PEAQ (Perceptual Evaluation of Audio Quality) [4]. This method, standardized as ITU-R recommendation BS.1387 [5], resulted from the integration of six different wide-band audio quality measurement systems [6],..[11]. Although from a perceptual point of view a single quality measurement approach should be possible towards both telephone-band speech and wide-band audio (music) signals, no unified method has been presented yet. A first attempt towards such an integrated method is given in [12].

One weakness of the current PSQM standard, from a theoretical point of view, is that the masking model is far too simple. In fact the only masking that is modelled in the PSQM standard is the one in which loud time-frequency localized components mask time-frequency components in the same time-frequency cells. It was expected that the successor of PSQM would include such an extended model of masking but the final model still has this simple approach. During the standardization process of the sucessor of PSQM several extended models of masking proved to be inadequate.

Another limitation of the current PSQM standard, from a practical point of view, is that for some distortions, for which the method was not designed, the correlation between objective and subjective quality scores is very low. The most obvious example for this is

misalignment between original and degraded speech file. Even when original and degraded are time aligned on a global level, modern voice transport techniques like VoIP (Voice over Internet Protocol) can introduce time warping (varying delay) that makes the PSQM algorithm fail completely. Other types of distortion where the PSQM algorithm fails are loud short localised distortions, which are underestimated in their disturbance, and linear filtering distortions, which are overestimated in their disturbance.

During the ITU-T study period 1997-2000 several companies worked on objective speech quality measurements. At KPN, John Beerends and Andries Hekstra made further significant improvements to cope with the weak points of PSQM [13], leading to a new version known as PSQM99. Stephen Voran from NTIA proposed an alternative method that was accepted as an appendix to recommendation P.861, the MNB (Measuring Normalizing Blocks [14], [15]). At BT, Antony Rix and Mike Hollier developed a new method, called PAMS (Perceptual Analysis Measurement System), that could deal with a wide variety of distortions [16]. Several other alternative systems were developed [17], [18], [19] and in 1999 the ITU-T benchmarked five different proposals that claimed to be able to cope with a wide variety of distortions. In this benchmark the best overall results were obtained by PSQM99 and PAMS with an average correlation over 22 speech quality evaluation experiments of 0.93 and 0.92 respectively [19]. None of the proposals however met all of the ITU-T requirements. An integrated method, taking the perceptual model of PSQM99 and the variable delay estimation of PAMS, was able to meet all the requirements. This method, called PESQ (Perceptual Evaluation of Speech Quality), was accepted in February 2001 as the new ITU-T objective speech quality measurement standard P.862 [20], [21].

2 THE BASICS

The basic idea behind the PESQ algorithm is the same as the one used in the development of the PSQM algorithm. Fig. 1 gives an overview of this approach. In PESQ the original and degraded signals are mapped onto an internal representation using a perceptual model. The difference in this representation is used by a cognitive model to predict the perceived speech quality of the degraded signal. This perceived listening quality is expressed in terms of Mean Opinion Score, an average quality score over a large set of subjects. Most of the subjective experiments used in the development of PESQ used the ACR (Absolute Category Rating) opinion scale [22], [23] of table 1. In these types of experiments subjects do not get a reference speech signal to judge the quality and some types of distortion, like missing words, sometimes go unnoticed in such experiments. Experiments in which this missing word phenomenon was clear were used only to a small extent in the optimization of PESQ. In these cases a lower correlation between subjective and objective results is likely.

Table 1: ACR listening quality opinion scale [22], [23] used in the development of PESQ.

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

An essential difference with the PSQM method [1], [3] is that the time alignment, necessary for the correct comparison of the matching parts of original and degraded, is an integrated part of the new standard. This perception based time alignment algortihm is described in a separate paper [24].

The internal representations, that are used by the PESQ cognitive model to predict the perceived speech quality, are calculated on the basis of signal representations that use the psychophysical equivalents of frequency (pitch measured in Barks) and intensity (loudness measured in Sones). This idea was also used in the PSQM method, however the psycho-acoustic parameters used in the mapping are now more in line with literature [25]. A minor disappointment is that the psychoacoustic model that is used in PESQ, and that will be presented in this paper, still has no correct modelling of masking caused by smearing in the time-frequency plane. Although masking models were implemented and tested in several stages of the development it never improved correlations between subjective and objective scores. This counterintuitive result was already presented in [26] and the first ideas towards incorporating masking into a speech quality model are given in [12]. A final solution to this problem is still under study.

The most important difference, besides the inclusion of a perceptual time alignment, between PSQM and PESQ is found in the cognitive part of the model. In PSQM two major cognitive effects are modelled in order to get high correlations between objective and subjective scores: asymmetry and different weighting of distortions during 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 [27]. 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.

In PESQ the effect is modelled by separately calculating a disturbance caused by introduced components. The introduced components are weighted with an asymmetry similar to the one used in PSQM. Unlike PSQM, which uses a single disturbance, PESQ uses a total and an added disturbance per speech file, which are only combined after they have been aggregated over time.

The second cognitive effect first described in [1] deals with the fact that disturbances that occur during speech active periods are more disturbing than those that occur during silent intervals. In PSQM it is modelled by a weighting factor that can be adjusted to the context of the experiment. However for the ITU-T benchmark no adjustments were allowed for the context and a different time weighting procedure, with optimal performance over a wide range of experimental contexts, was found in using an L_p weighting over time:

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

with N = total number of frames and p>1.0.

Such an L_p weighting emphasizes loud disturbances when compared to a normal, L_1 time averaging, leading to a better correlation between objective and subjective scores [28], [29], [30]. The aggregation of frame disturbances over time is carried in a hierarchy of two layers.

A further difference between PSQM and PESQ is the partial compensation for linear distortions (filtering) as found in the system under test. It is well known that linear distortions are less objectionable than non-linear distortions. Therefore in PESQ minor steady-state differences between original and degraded are compensated. More severe effects, or rapid variations, are only partially compensated so that a residual effect remains and contributes to the overall perceptual disturbance.

The partial frequency response compensation also has an impact on the partial compensation of gain differences in successive frames. This gain compensation is an essential part of any objective speech quality measurement system because slow and/or small gain variations only have a minor impact on the perceived speech quality. Fast and or large gain variations can have a major impact on the perceived speech quality. One of the main problems in designing an objective speech quality measurement system is the way these gain variations are treated and the way they are coupled to the asymmetry effect [31].

The final PESQ algorithm that resulted from the integration of the PSQM99 and PAMS algorithms is given in the next section.

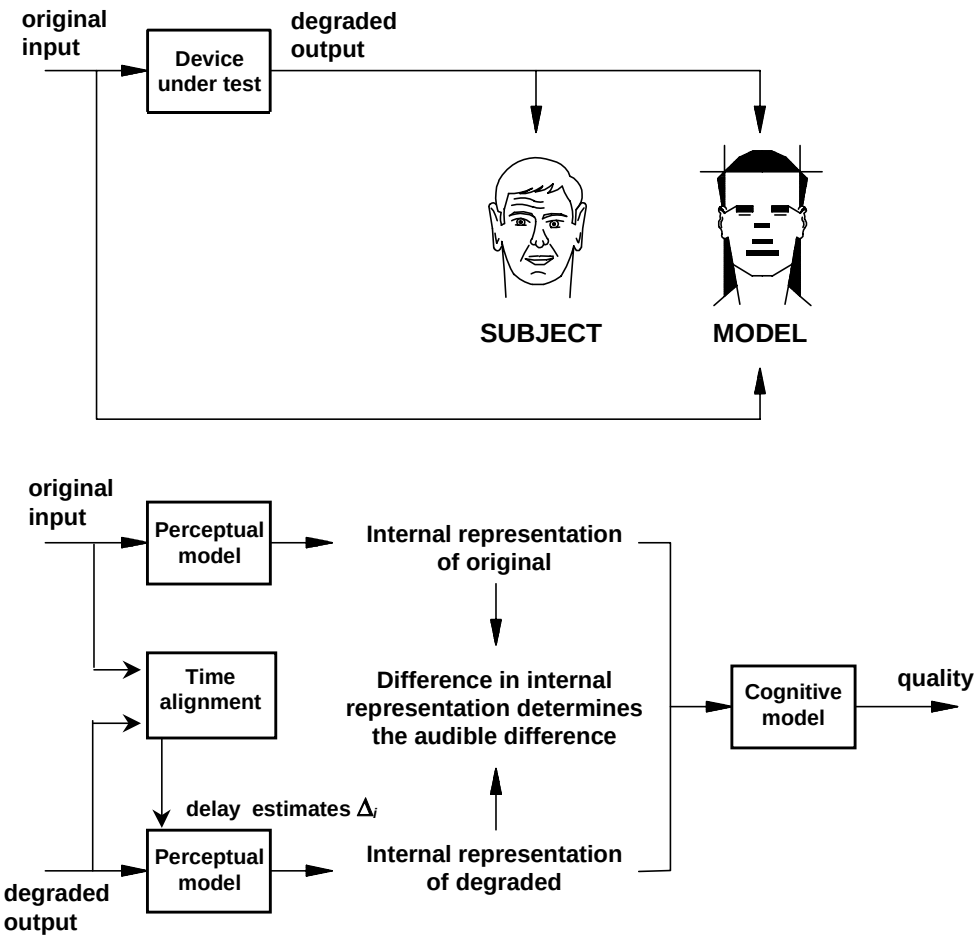

Fig. 1. Overview of the basic philosophy used in PESQ. A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the output of the device under test with the input, using alignment information as derived from the time signals in the time alignment module.

3 DESCRIPTION OF PESQ ALGORITHM

The PESQ algorithm follows the same steps as used in PSQM [1], [3] but with the modifications introduced in the previous section. Each of the consecutive steps is described in the following sections.

3.1 Calibration

The first step in the PESQ algorithm is to compensate for the overall gain of the system under test. This step in combined with a global scaling of the signals to a correct overall level. Both the original *X(t)* and degraded signal *Y(t)* are scaled to the same, constant power level. PESQ thus assumes that the subjective listening level is a constant, about 79dB SPL at the ear reference point (P.830, [23] section 8.1.2), that variations between the levels of the recorded signals within a single subjective experiment are small, and that average level differences between experiments are compensated by the overall level setting in the subjective experiment. The PESQ level alignment is carried out based on the power of bandpass filtered versions (300 - 3000 Hz) of the original and degraded signals.

Besides a level alignment in the time domain it is also necessary to align the level in the frequency domain, after the time-frequency analysis. This is carried out by generating a sine wave with a frequency of 1000 Hz and an amplitude of 40 dB SPL. This sine wave is transformed to the frequency domain using a windowed FFT with 32 ms frame length. After converting the frequency axis to a modified Bark scale the peak amplitude of the resulting pitch power density is then normalized to a power value of 10^4 by multiplication with a power scaling factor S_p.

The same 40 dB SPL reference tone is used to calibrate the psychoacoustic (Sone) loudness scale. After warping the intensity axis to a loudness scale using Zwicker's law [25] the integral of the loudness density over the Bark frequency scale normalized to 1 Sone using the loudness scaling factor S_l.

3.2 IRS-Receive Filtering

It is assumed that listening is carried out using a handset with a frequency response that follows an IRS receive [32] or a modified IRS [23] receive characteristic. A perceptual model of the human evaluation of speech quality must take account of this, to model the signals that the subjects actually heard. Therefore IRS-like receive filtered versions of the original speech signal and degraded speech signal are computed. In PESQ this is implemented by an FFT over the length of the file, filtering in the frequency domain with a piecewise linear response similar to the (unmodified) IRS receive characteristic (P.48, [32]), followed by an inverse FFT over the length of the speech file. This results in the filtered versions $X_{IRSS}(t)$ and $Y_{IRSS}(t)$ of the scaled input and output signals $X_S(t)$ and $Y_S(t)$. A single IRS-like receive filter is used within PESQ irrespective of whether the real subjective experiment used IRS or modified IRS filtering. The reason for this approach was that in most cases the exact filtering is unknown, and that even when it is known the coupling of the handset to the ear is not known. It was therefore an ITU-T requirement that the objective method should be relatively insensitive to the filtering of the handset. Furthermore no adjustments for filtering were allowed within the ITU-T benchmark and thus the best overall filtering compromise had to be implemented.

3.3 Calculation of the Active Speech Time Interval

If the original and degraded speech file start or end with large silent intervals, this could influence the computation of certain average distortion values over the files. Therefore, an estimate is made of the silent parts at the beginning and end of these files. The sum of five successive absolute sample values must exceed 500 from the beginning and end of the original speech file in order for that position to be considered as the start or end of the active interval. The interval between this start and end is defined as the active speech time interval. In order to save computation cycles and/or storage size, some computations can be restricted to the active interval.

3.4 Time-Frequency Decomposition, Time Axis Modification

The human ear performs a time-frequency transformation. In PESQ this is modelled by a short term FFT with a Hann window over 32 ms frames. The overlap between successive frames is 50%. The power spectra – the sum of the squared real and squared imaginary parts of the complex FFT components – are stored in separate real valued arrays for the original and degraded signals. Phase information within a single frame is discarded in PESQ and all calculations are based on only the power representations $PX_{WIRSS}(f)_n$ and $PY_{WIRSS}(f)_n$.

The startpoints of the frames in the degraded signal are shifted over the delay observed by the variable delay estimator [24]. The time axis of the original speech signal is left as is. If the delay increases, parts of the degraded signal are omitted from the processing, while for decreases in the delay parts of the degraded signal are repeated. This time axis modification gave best results in terms of correlation with the subjectively perceived overall speech quality. A minor extension to this strategy is given in section 3.12.

3.5 Calculation of the Pitch Power Densities

The Bark scale reflects that at low frequencies, the human hearing system has a finer frequency resolution than at high frequencies. This is implemented by binning FFT bands and summing the corresponding powers of the FFT bands with a normalization of the summed parts. The warping function that maps the frequency scale in Hertz to the pitch scale in Bark approximates the values given in the literature. The resulting signals are known as the pitch power densities $PPX_{WIRSS}(f)_n$ and $PPY_{WIRSS}(f)_n$.

3.6 Compensation of the Linear Frequency Response

To deal with filtering in the system under test, the power spectrum of the original and degraded pitch power densities are averaged over time. This average is calculated over speech active frames only using time-frequency cells whose power is more than 30 dB above the absolute hearing threshold. Per modified Bark bin, a partial compensation factor is calculated from the ratio of the degraded spectrum to the original spectrum. The maximum compensation is never more than 20dB. The original pitch power density $PPX_{WIRSS}(f)_n$ of each frame n is then multiplied with this partial compensation factor to equalise the original to the degraded signal. This results in a filtered version of the original pitch power density $PPX'_{WIRSS}(f)_n$.

This partial compensation is used because severe filtering is disturbing to the listener while mild filtering effects hardly influence the perceived overall quality, especially if no reference is available to the subject. The compensation is carried out on the original signal because the degraded signal is the one that is judged by the subjects in an ACR experiment.

3.7 Compensation of the Time Varying Gain

Short-term gain variations are partially compensated by processing the pitch power densities frame by frame. For the original and the degraded pitch power densities, the sum in each frame *n* of all values that exceed the absolute hearing threshold is computed. The ratio of the power in the original and the degraded files is calculated and bounded to the range $\{3 \cdot 10^{-4}, 5\}$. A first order low pass filter (along the time axis) is applied to this ratio. The time constant of this filter is approximately 16ms. The distorted pitch power density in each frame, *n*, is then multiplied by this ratio, resulting in the partially gain compensated distorted pitch power density $PPY'_{WIRSS}(f)_n$.

3.8 Calculation of the Loudness Densities

After partial compensation for filtering and short-term gain variations, the original and degraded pitch power densities are transformed to a Sone loudness scale using Zwicker's law [25].

$$LX(f)_n = S_l \cdot \left(\frac{P_0(f)}{0.5} \right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{PPX'_{WIRSS}(f)_n}{P_0(f)} \right)^{\gamma} - 1 \right]$$

with $P_0(f)$ the absolute hearing threshold and S_l the loudness scaling factor.

Above 4 Bark, the Zwicker power, γ, is 0.23, the value given in the literature. Below 4 Bark, the Zwicker power is increased slightly to account for the so called recruitment effect. The resulting two dimensional arrays $LX(f)_n$ and $LY(f)_n$ are called loudness densities.

3.9 Calculation of the Disturbance Density

The signed difference between the distorted and original loudness density is computed. When this difference is positive, components such as noise have been added. When this difference is negative, components have been omitted from the original signal. This difference array is called the raw disturbance density.

Masking is modelled by applying a deadzone in each time-frequency cell, as follows. The per cell minimum of the original and degraded loudness density is computed for each time-frequency cell. These minima are multiplied by 0.25. The corresponding two dimensional array is called the mask array. Next the following rules are applied in each time-frequency cell:

- If the raw disturbance density is positive and larger than the mask value, the mask value is subtracted from the raw disturbance.
- If the raw disturbance density lies in between plus and minus the magnitude of the mask value the disturbance density is set to zero.
- If the raw disturbance density is more negative than minus the mask value, the mask value is added to the raw disturbance density.

The net effect is that the raw disturbance densities are pulled towards zero. This represents a deadzone before an actual time-frequency cell is perceived as distorted. This models the process of small differences being inaudible in the presence of loud signals (masking) in each time-frequency cell. The result is a disturbance density as a function of time (frame number *n*) and frequency, $D(f)_n$.

3.10 Modelling of the Asymmetry Effect

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 [2]. 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 by calculating an asymmetrical disturbance density $DA(f)_n$ per frame by multiplication of the disturbance density $D(f)_n$ with an asymmetry factor. This asymmetry factor equals the ratio of the distorted and original pitch power densities raised to the power of 1.2. If the asymmetry factor is less than 3 it is set to zero. If it exceeds 12 it is clipped at that value. Thus only those time-frequency cells remain, as nonzero values, for which the degraded pitch power density exceeded the original pitch power density.

3.11 Aggregation of the Disturbance Densities over Frequency and Silent Interval Processing

The disturbance density $D(f)_n$ and asymmetrical disturbance density $DA(f)_n$ are integrated (summed) along the frequency axis using two different *Lp* norms and a weighting on soft frames (having low loudness):

$$D_n = M_n \sqrt[3]{\sum_{f=1,..Number of\ Barkbands} (\ |D(f)_n|\ W_f\ )^3}$$

$$DA_n = M_n \sum_{f=1,..Number of\ Barkbands} (\ |DA(f)_n|\ W_f\ )$$

with M_n a multiplication factor equal to *((power of original frame + 10^5)/10^7)$^{-0.04}$*, resulting in an emphasis of the disturbances that occur during silences in the original speech fragment, and W_f a series of constants proportional to the width of the modified Bark bins. After this multiplication the frame disturbance values are limited to a maximum of 45. These aggregated values, D_n and DA_n, are called frame disturbances.

If the distorted signal contains a decrease in the delay larger than 16 ms (half an FFT frame) the repeat strategy as mentioned in 3.4 is applied. It was found to be better to ignore the frame disturbances during a decrease in delay in the computation of the objective speech quality. As a consequence frame disturbances are zeroed when this occurs. The resulting frame disturbances are called D'_n and DA'_n.

3.12 Realignment of Bad Intervals

Consecutive frames with a frame disturbance above a threshold are called bad intervals. In a minority of cases the objective measure predicts large distortions over a minimum number of bad frames due to incorrect time delays observed by the preprocessing. For those so called bad intervals a new delay value is estimated by locating the maximum of the cross correlation between the absolute original signal and absolute degraded signal precompensated with the delays observed by the preprocessing. When the maximimal cross correlation is below a threshold, it is concluded that the interval is matching noise against noise and the interval is no longer called bad, and the processing for that interval is halted. Otherwise, the frame disturbance for the frames during the bad intervals is recomputed and, if it is smaller, replaces the original frame disturbance. The result is the final frame disturbances D''_n and DA''_n that are used to calculate the perceived overall speech quality.

3.13 Aggregation of the Disturbances over Time

First the frame disturbances are aggregated over split second intervals. Next the split second disturbances are aggregated over the complete active time interval. For the split second time aggregation the frame disturbance values and the asymmetrical frame disturbance values are L_6 aggregated over 20 frames (accounting for the overlap of frames: approx. 320 ms). These split second intervals also overlap 50% and no window function is used. Over the speech file length an L_2 norm is used.

The split second disturbance values and the asymmetrical split second disturbance values are aggregated over the active interval of the speech files (the corresponding frames) now using L_2 norms. The higher value of p for the aggregation within split second intervals as compared to the lower p value of the aggregation over the speech file is due to the fact that when parts of the split seconds are distorted, that split second loses meaning, whereas if a first sentence in a speech file is distorted the quality of other sentences remains intact.

3.14 Computation of the PESQ Score

The final PESQ score is a linear combination of the average disturbance value and the average asymmetrical disturbance value. This linear combination was optimized on a large set of subjective experiments and after the mapping the range of the PESQ score is –0.5 to 4.5, although for most cases the output range will be a MOS-like score between 1.0 and 4.5, the normal range of MOS values found in an ACR subjective experiment.

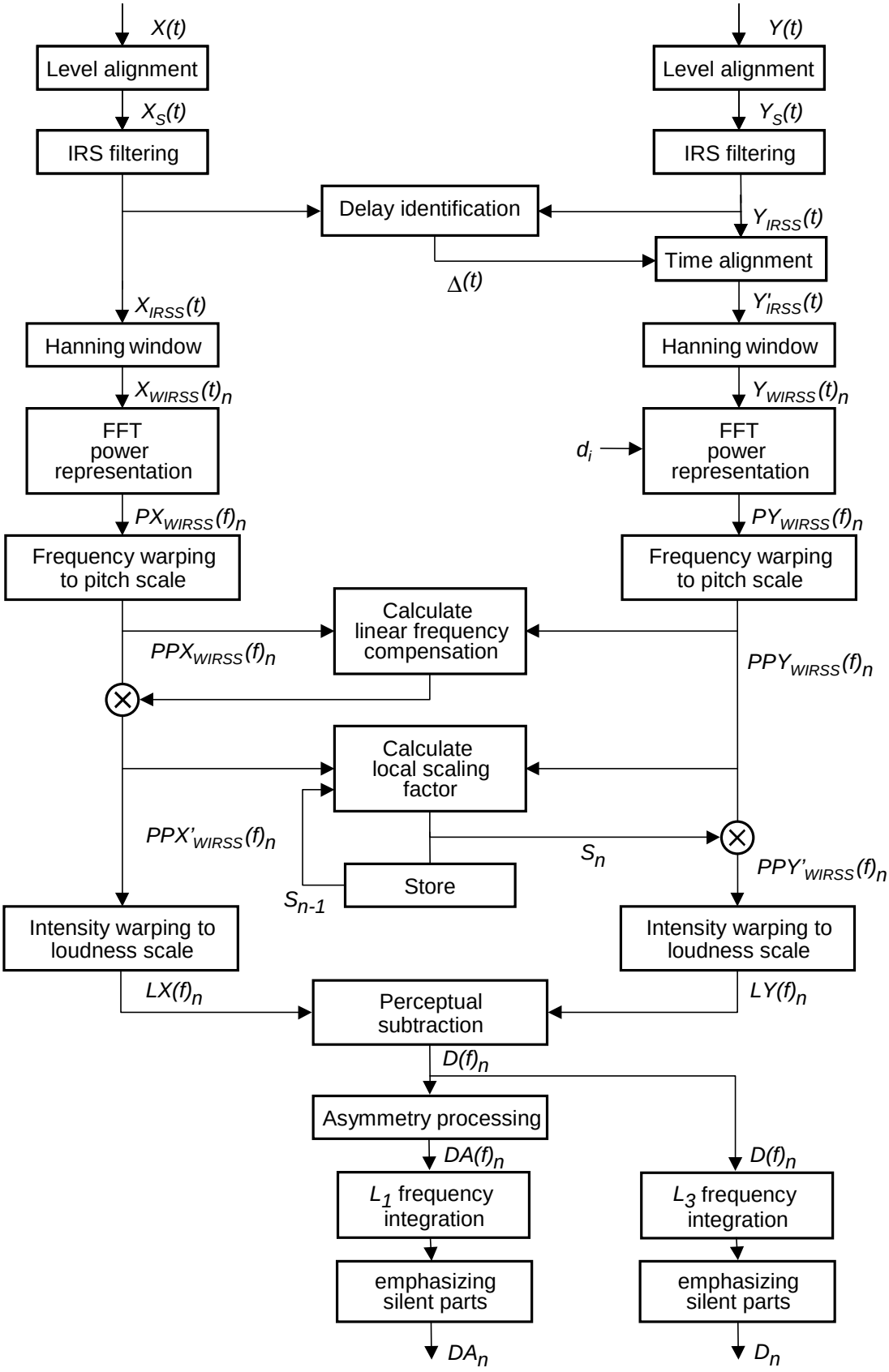

Fig. 2. Overview of the perceptual model. The distortions per frame D_n and DA_n have to be aggregated over time (index n) to obtain the final disturbances (see Fig. 3).

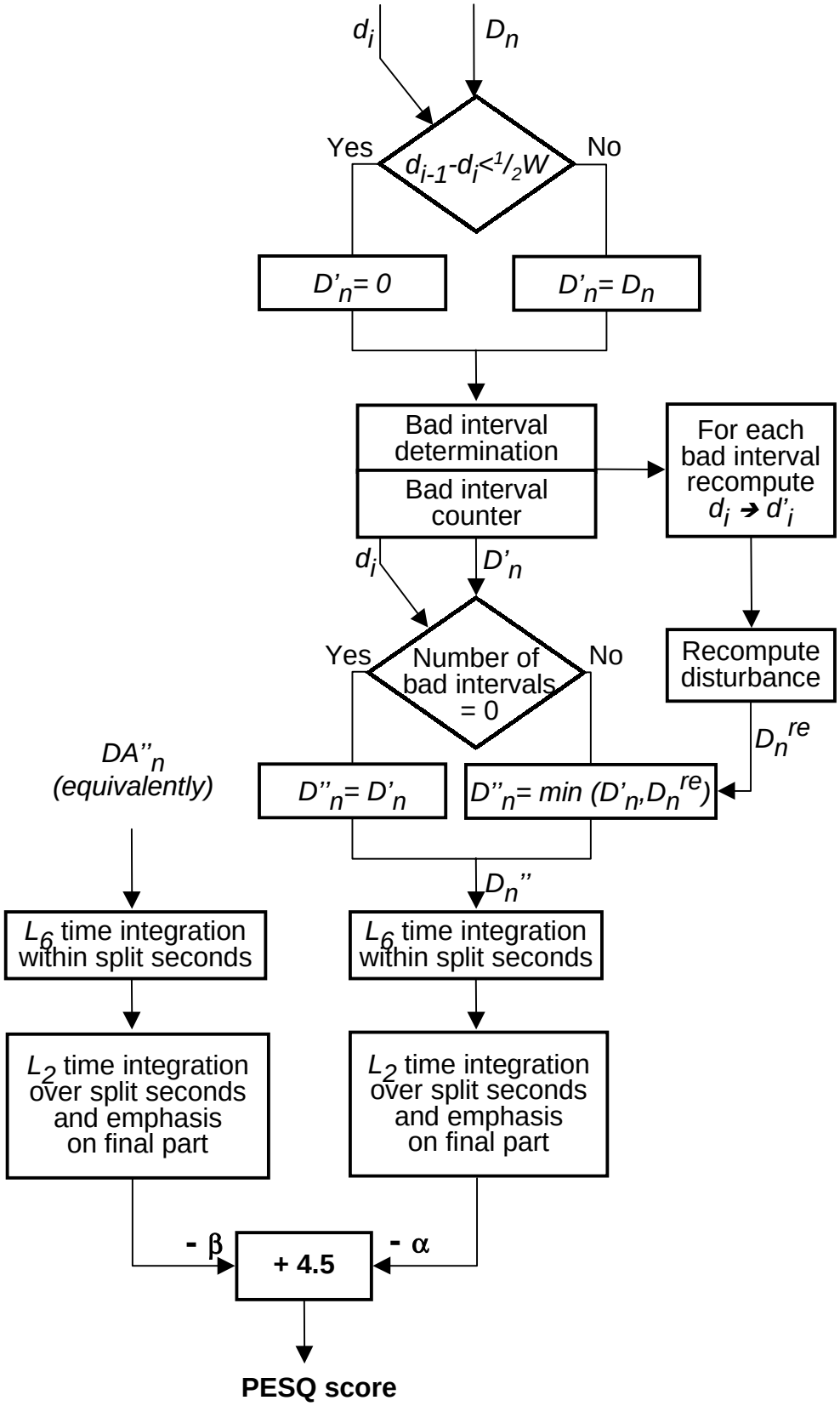

Fig. 3. Overview of the perceptual model. After re-alignment of the bad intervals the distortions per frame D''_n and DA''_n are integrated over time and mapped to the PESQ score. W is the FFT frame length in samples.

4 TRAINING AND PERFORMANCE RESULTS OF PESQ

It is important that test signals for use with PESQ are representative of the real speech signals carried by communications networks. Networks may treat speech and silence differently and coding algorithms are often highly optimised for speech – and so may give meaningless results if they are tested with signals that do not contain the key temporal and spectral properties of speech. Further pre-processing is often necessary to take account of filtering in the send path of a handset, and to ensure that power levels are set to an appropriate range.

4.1 Source Speech Material

At present all official performance results for PESQ relate to experiments conducted using the same natural speech recordings in both the subjective and objective test. The use of artificial speech signals and concatenated real speech test signals is recommended only if they represent the temporal structure (including silent intervals) and phonetic structure of real speech signals. Artificial speech test signals can be prepared in several ways. A concatenated real speech test signal may be constructed by concatenating short fragments of real speech while retaining a representative structure of speech and silence [34]. Alternatively, a phonetic approach may be used to produce a minimally redundant artificial speech signal which is representative of both the temporal and phonetic structure of a large corpus of natural speech [33]. Test signals should be representative of both male and female talkers. In preliminary tests, high quality artificial speech and concatenated real speech both showed good results with PESQ. In these tests the objective scores for the test signals in each condition served as a prediction for the subjective condition MOS values. This approach makes it possible to determine the quality of the system under test with the least possible effort [33], [34].

Most of the experiments used in calibrating and validating PESQ contained pairs of sentences separated by silence, totalling 8s in duration; in some cases three or four sentences were used, with slightly longer recordings (up to 12s). Recordings made for use with PESQ should be of similar length and structure. Thus if a condition is to be tested over a long period it is most appropriate to make a number of separate recordings of around 8-20 seconds of speech and process each file separately with PESQ. This has additional benefits: if the same original recording is used in every case, time variations in the quality of the condition will be very apparent; alternatively, several different talkers and/or source recordings can be used, allowing more accurate measurement of talker or material dependence in the condition. Note that the non-linear averaging process in PESQ means that the average score over a set of files will not usually equal the score of a single concatenated version of the entire set of files.

Signals should be passed through a filter with appropriate frequency characteristics to simulate sending frequency characteristics of a telephone handset, and level-equalized in the same manner as real voices. ITU-T recommends the use of the Modified Intermediate Reference System (IRS) sending frequency characteristic as defined in Annex D of Recommendation P.830 [23]. Level alignment to an amplitude that is representative of real traffic should be performed in accordance with section 7.2.2 of Recommendation P.830.

In some cases the measurement system used (for example, a 2-wire analogue interface) may introduce significant level changes. These should be taken into account to ensure that the signal passed into the network is at a representative level.

The prepared source material after handset (send) filtering and level alignment is normally used as the original signal for PESQ.

4.2 Addition of background noise

It is possible to use PESQ to assess the quality of systems carrying speech in the presence of background or environmental noise (e.g. car, street, etc). Some of the PESQ training and validation material contained background noise of different types and PESQ performed well on these databases.

Noise recordings should be passed through an appropriate filter similar to the modified IRS sending characteristic – this is especially important for low-frequency signals such as car noise which are heavily attenuated by the handset filter – and then level aligned to the desired level for the test. For PESQ to take account of the subjective disturbance in an ACR context, due to the noise as well as any coding distortions, the original signal used with PESQ should be clean, but the noise should be added before the signals are passed to the system under test. This process is shown in Figure 4.

(a) testing with clean speech (b) testing with noisy speech

Fig. 4. Methods for testing quality with and without environmental noise using the PESQ algorithm.

4.3 Training of PESQ

A large database of subjective tests was assembled to enable PESQ to be trained over as wide a range of conditions as possible, and to minimise the risk of over-training. 30 subjective tests were used in the final training of the model.

The training process was iterative. A large number of different symmetric and asymmetric disturbance parameters were calculated for each condition by using different values of p for each of the three averaging stages. Subsets of these disturbance parameters were combined using linear regression to give a predictor of subjective MOS. A further regression is needed for each subjective test to account of context and voting preferences of different subjects. During the training process a linear mapping was also used at this stage. The regression was performed for all candidate subsets of up to four disturbance parameters, and the optimal combination – giving the highest average correlation coefficient – was found. This enabled the best disturbance parameters to be chosen from several hundred candidates. Further checks were carried out by training on a subset and prediction on the remaining set of approximately 30 additional subjective tests. Finally, manual adjustments were made to components of the model and the process repeated a number of times.

In order to make PESQ as robust as possible, it was desired to keep the number of disturbance parameters used to two, symmetric disturbance and asymmetric disturbance. This avoids a risk of over-training if a large number of separate parameters are used – for example, to take account of modulation, clipping, filtering, etc. – but it relies on earlier components of the model to include the perceptual effect of these phenomena. This made it necessary to use the iterative design process to jointly optimise the components of the model and the final mapping to subjective quality.

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

4.4 Performance results

Condition MOS is one of the most common measures of subjective quality used in speech quality evaluation. It represents the average MOS for four or more recordings for a single network condition. 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 device under test.

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 [22], [23], 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 function is used, for each subjective test, to map the objective PESQ MOS scores onto the subjective scores. It is then possible to calculate correlation coefficients and residual errors, between objective and subjective scores.

4.4.1 Correlation results

The performance of PESQ is compared to PSQM [1], [3] and MNB [3 appendix II], [14] in Figures 5–8 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. Fig. 5 presents 19 tests containing mainly mobile codecs and/or networks. Fig. 6 gives results from 9 tests on predominantly fixed networks or codecs. Fig. 7 shows 10 tests containing VoIP conditions on a wide range of codec/error types. Finally, Fig. 10 gives the results for 8 tests conducted on PESQ by independent laboratories using data unknown in the development of the model.

The different tests were conducted in a number of different languages, and eight of the tests included conditions with background noise. For the 22 known ITU benchmark experiments the average correlation was 0.935. For the set of 8 independent experiments used in the final validation (plotted in Fig. 10) – experiments that were unknown during the development of PESQ – the average correlation was also 0.935. The fact that the average correlation on both the trained and unknown set is the same shows the stability of the model.

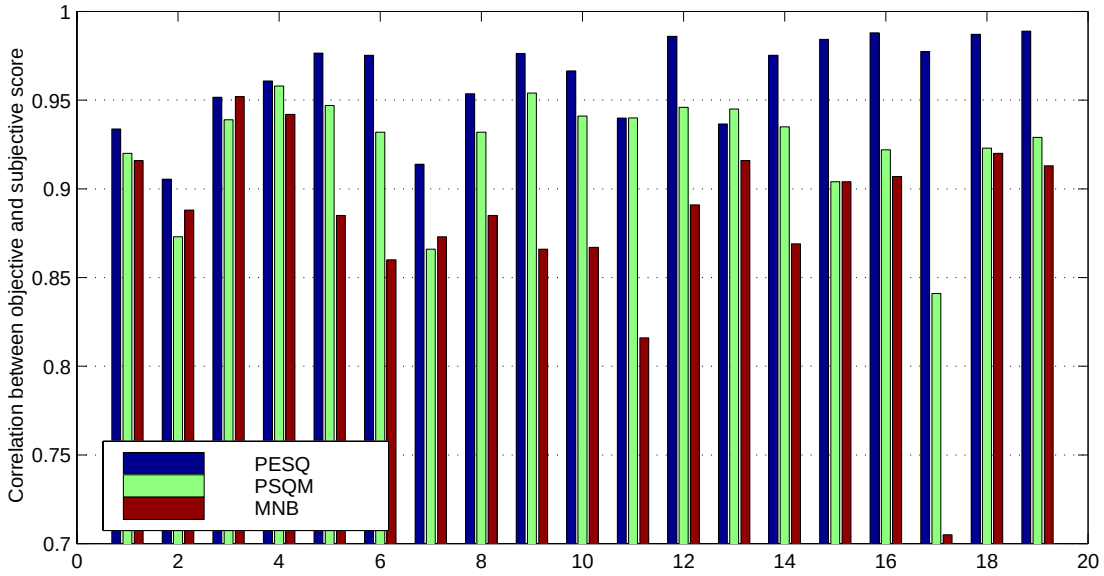

Fig. 5. Mobile network performance results for PESQ, PSQM [1], [3] and MNB [3], [14]. Condition correlation coefficient, per experiment, after monotonic 3rd-order polynomial mapping.

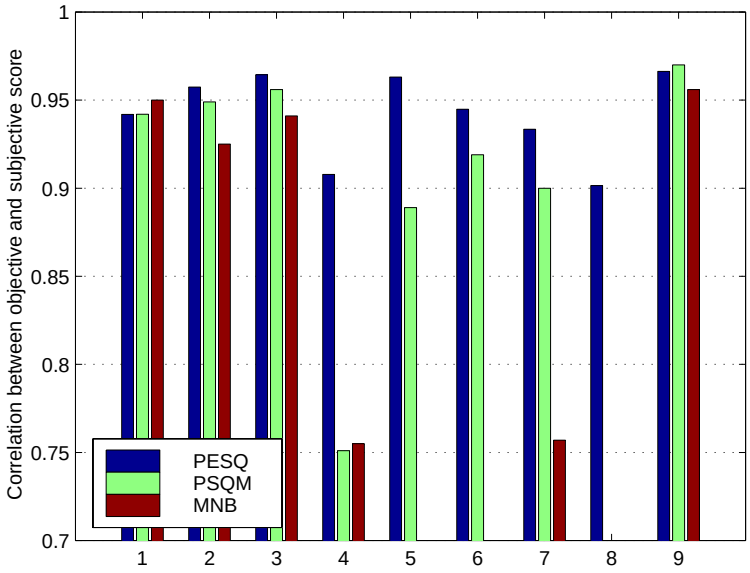

Fig. 6. Fixed network performance results for PESQ, PSQM [1], [3] and MNB [3], [14]. Condition correlation coefficient, per experiment, after monotonic 3rd-order polynomial mapping. In tests 5, 6 and 8 the scores for MNB (and PSQM in test 8) are off the bottom of the scale.

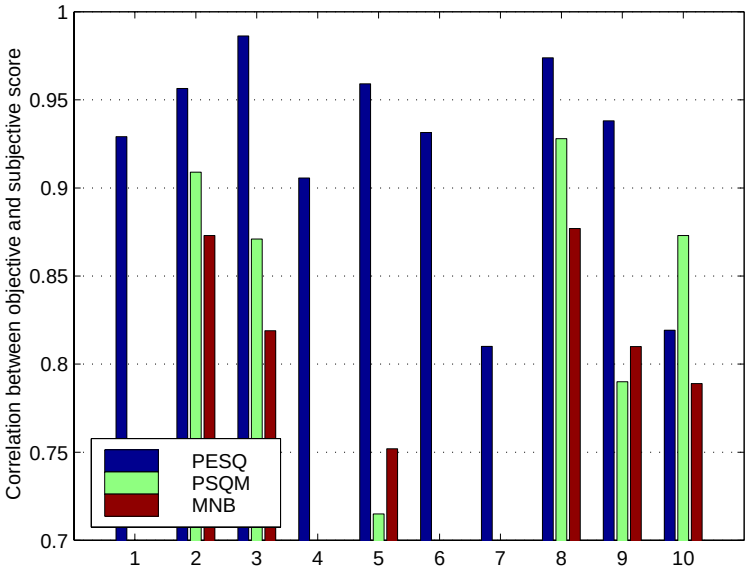

Fig. 7. VoIP and multi-type test results for PESQ, PSQM [1], [3] and MNB [3], [14]. Condition correlation coefficient, per experiment, after monotonic 3rd-order polynomial mapping. In tests 1, 4, 6 and 7 the scores for MNB and PSQM are off the bottom of the scale.

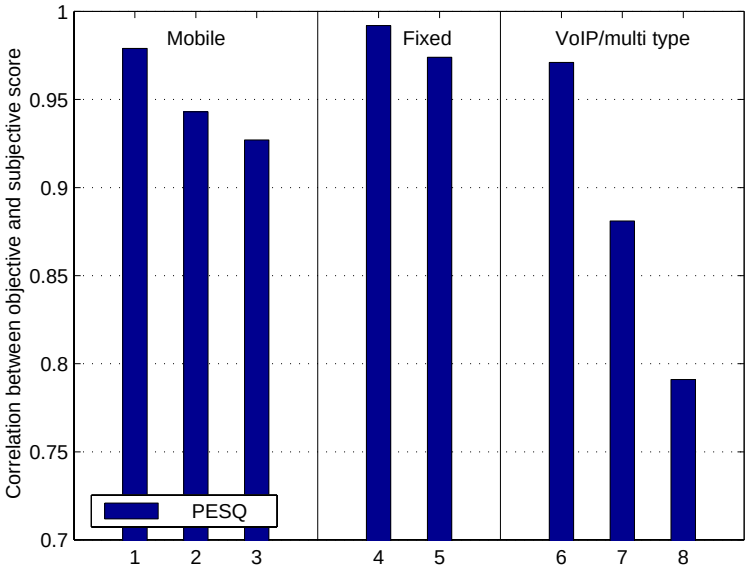

Fig. 8. Independent results for unknown subjective tests (PESQ only). Condition correlation coefficient, per experiment, after monotonic 3rd-order polynomial mapping.

4.4.2 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. Figures 9 plots the cumulative distribution of errors for PESQ, PSQM [1], [3] and MNB [3 appendix II], [14], 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.

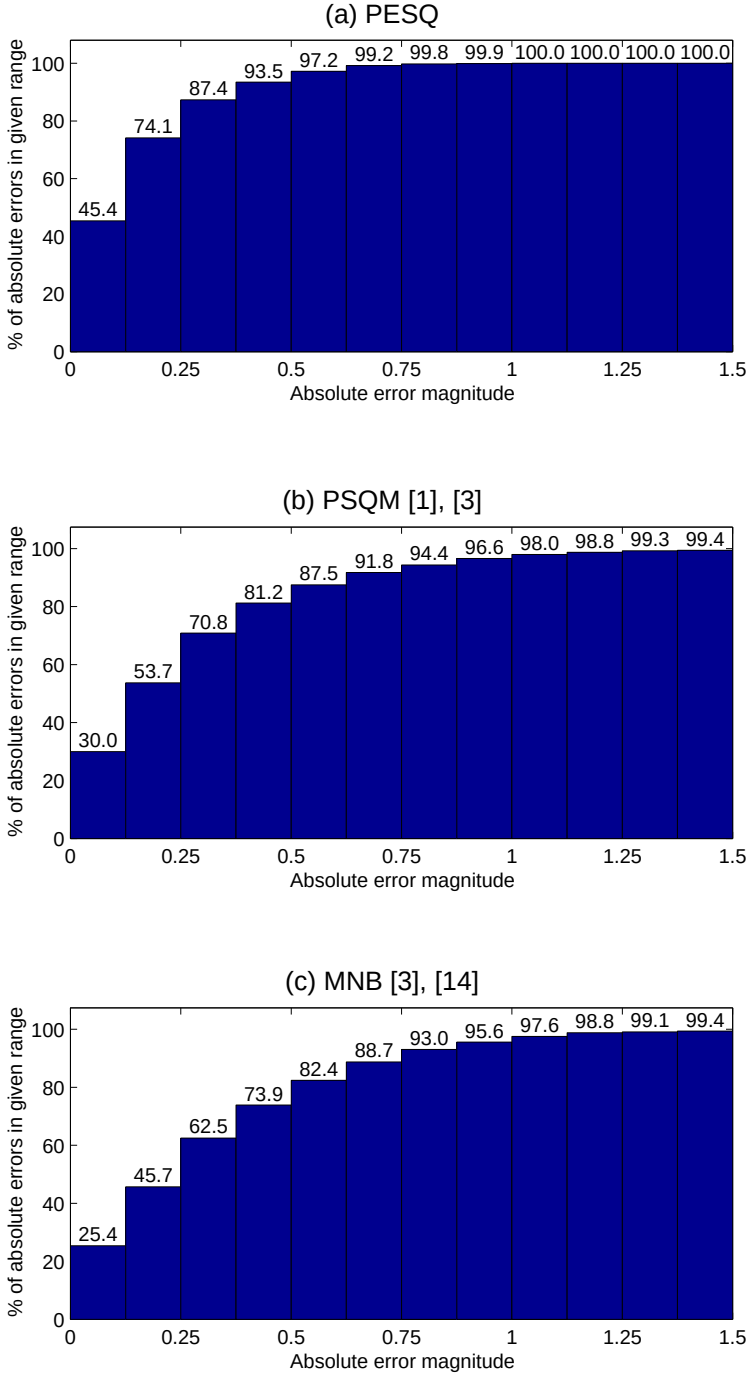

Fig. 9. Residual error distribution for PESQ, PSQM [3], and MNB [3, 14]. Per condition, after monotonic 3rd-order polynomial mapping.

5 Using PESQ now and in the future

Although PESQ was developed for a wide range of distortions it still not the ultimate perceptual measurement technique. As stated in section 2 the psychoacoustic model that is used in PESQ does not model masking caused by smearing in the time-frequency plane. PESQ may therefore give inaccurate scores with music signals. In this section an overview is given of where PESQ can be applied and where it fails.

Table 2 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].

Table 2. Factors for which PESQ can be used for objective speech quality measurement.

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

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

PESQ is not intended to be used to assess:

- effect of listening level
- conversational delay
- talker echo, where a subjects hears his own voice delayed
- talker sidetone, where a subjects may hear its own voice distorted
- non-intrusive measurements, where only output signals are available from the system
- music

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. The most extreme examples have been found in cases where complete words or even sentences are omitted from the speech signal. In this case the subjective test methodology in the form of ACR testing is questionable, because sometimes subjects are unable to notice missing words. However, systems which leave out words and sentences should be avoided in telecommunications.

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

- listener echo
- very low bit-rate speech vocoders (below 4kbit/s)
- systems where the assessments have to be made in the acoustic domain, like head and torso simulator (HATS) measurements on handsets and/or hands-free telephones

- wideband speech, with bandwidth significantly about 4kHz and listening with wideband headphones, although this may be made possible by an appropriate change of filter [36].

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 using subjective tests conducted according to ITU-T P.800 or P.830 [22], [23] – i.e. "telephone quality" speech signals with a frequency response that rapidly falls off below 300 and above 3400 Hz. PEAQ [4], [5] is able to measure the quality of audio codecs – "audio quality" – for applications such as broadcast, with headphone or loudspeaker listening [35]. In between these two ranges is the so-called "intermediate quality" [36] where no standardized perceptual quality measurement system can be used. It is hoped that PESQ can be extended to provide assessment of systems at this intermediate quality level. A first attempt to integrate the ideas from speech quality measurement and music quality measurement into a single quality measurement system that can deal with the complete range of qualities is given in [12].

6 Conclusions

For quality assessment of telephone band speech signals (300-3400 Hz) PESQ performs much better than earlier speech codec assessment models such as P.861 PSQM and MNB. In February 2001, PESQ replaced these models and became new ITU-T recommendation P.862. The major advantages of PESQ over PSQM and MNB are:

- inclusion of a dynamic, perceptual, time alignment that allows for assessments under a wide variety of time axis distortions (see accompanying paper [24])
- inclusion of an L_p weighting over time that correctly models the higher weight that subjects give on short loud disturbances
- a better modeling of the asymmetry effect, the difference in disturbance between time-frequency components that are introduced versus time-frequency components that are omitted
- the ability to correctly deal with linear frequency response distortions
- an improved local power scaling that deals with the perceptual influence of gain variations

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. On both a training set of 22 benchmark experiments and on a set of 8 validation experiements the average correlation was 0.935, showing the stability of the model.

PESQ represents a significant step forward in the accuracy and range of applicability of objective speech quality assessment methods.

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. The authors would like to acknowledge the assistance of many of their colleagues at BT and KPN, and thank the companies who acted as independent validation laboratories: AT&T, Lucent Technologies, Nortel Networks, and especially France Telecom R&D.

8 References

[1] J. G. Beerends and J. A. Stemerdink, “A Perceptual Speech Quality Measure Based on a Psychoacoustic Sound Representation,” *J. Audio Eng. Soc.,* vol. 42, pp. 115-123, (1994 March).

[2] ITU-T Study Group 12, “Review of Validation Tests for Objective Speech Quality Measures,” Document COM 12-74 (1996 March).

[3] ITU-T Rec. P.861, “Objective Quality Measurement of Telephoneband (300-3400 Hz) Speech Codecs,” International Telecommunication Union, Geneva, Switzerland (1996 Aug.).

[4] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B. Feiten, “PEAQ - The ITU-Standard for Objective Measurement of Perceived Audio Quality,” *J. Audio Eng. Soc.*, vol 48, pp. 3-29, (2000 Jan./Feb.).

[5] ITU-R Rec. BS.1387, “Method for Objective Measurements of Perceived Audio Quality,” International Telecommunication Union, Geneva, Switzerland (1998 Dec.).

[6] T. Thiede and E. Kabot, “A New Perceptual Quality Measure for Bit Rate Reduced Audio,” presented at the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 44, p. 653 (1996 July/Aug.), preprint 4280.

[7] J. G. Beerends and J. A. Stemerdink, “A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation,” *J. Audio Eng. Soc.,* vol. 40, pp. 963-978 (1992 Dec.).

[8] B. Paillard, P. Mabilleau, S. Morisette, and J. Soumagne, "PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals," *J.Audio Eng. Soc.*, vol. 40, pp. 21-31, (1992 Jan./Feb.).

[9] J. Herre, E. Eberlein, H. Schott, and Ch. Schmidmer, "Analysis Tool for Real Time Measurements using Perceptual Criteria," In *Proc.* AES 11th Int. Conf. (Portland, Or, USA, 1992), pp. 180-190.

[10] T. Sporer, “Objective Audio Signal Evaluation -- Applied Psychoacoustics for Modeling the Perceived Quality of Digital Audio,” presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 45, p. 1002 (1997 Nov.), preprint 4512.

[11] C. Colomes, M. Lever, J.B. Rault, and Y.F. Dehery, "A perceptual model applied to audio bit-rate reduction," *J.Audio Eng. Soc.*, vol. 43, pp. 233-240, (1995 Apr.).

[12] J. G. Beerends, “Measuring the Quality of Speech and Music Codecs, an Integrated Psychoacoustic Approach,” presented at the 98th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 43, p. 389 (1995 May), preprint 3945.

[13] ITU-T Study Group 12, “Improvement of the P.861 Perceptual Speech Quality Measure,” Document COM 12-20 (1997 Dec.).

[14] S. Voran, “Objective Estimation of Perceived Speech Quality - Part I: Development of the Measuring Normalizing Block Technique”, *IEEE Trans. on Speech and Audio Processing.,* vol. 7, pp. 371-382 (1999 July).

[15] S. Voran, “Objective Estimation of Perceived Speech Quality - Part II: Evaluation of the Measuring Normalizing Block Technique”, *IEEE Trans. on Speech and Audio Processing.,* vol. 7, pp. 383-390 (1999 July).

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

[17] M. Hansen and B. Kollmeier, “Objective Modeling of Speech Qualtiy with a Psychoacoustically Validated Auditory Model,” *J. Audio Eng. Soc.*, vol. 48, pp. 395-408 (2000 May.).

[18] ITU-T Study Group 12, “TOSQA – Telecommunication Objective Speech Quality Assessment,” Document COM 12-34 (1997 Dec.).

[19] ITU-T Study group 12, “Report of the question 13/12 rapporteur’s meeting, Solothurn, Switzerland,” Document COM 12-117, (2000 March).

[20] ITU-T Study Group 12, “Performance of the Integrated KPN/BT Objective Speech Quality Assessment Model,” Delayed Contribution D.136 (2000 May) (equivalent to KPN Research publication 00-32201a).

[21] ITU-T Rec. P.862, “Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs”, International Telecommunication Union, Geneva, Switzerland (2001 Feb.)

[22] ITU-T Rec. P.800, “Methods for Subjective Determination of Transmission Quality,” International Telecommunication Union, Geneva, Switzerland (1996 Aug.).

[23] ITU-T Rec. P.830, “Subjective performance assessment of telephone-band and wideband digital codecs,” International Telecommunication Union, Geneva, Switzerland (1996 Feb.).

[24] A. W. Rix, M. P. Hollier, A. P. Hekstra and J. G. Beerends, “Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I – Time Alignment”, *J. Audio Eng. Soc.*

[25] E. Zwicker and R. Feldtkeller, “Das Ohr als Nachrichtenempfänger,” S. Hirzel Verlag, Stuttgart (1967).

[26] J. G. Beerends and J. A. Stemerdink, “The optimal time-frequency smearing and amplitude compression in measuring the quality of audio devices,” presented at the 94th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 41, p. 409 (1993 May) preprint 3604.

[27] J. G. Beerends, “Modelling Cognitive Effects that Play a Role in the Perception of Speech Quality,” in DEGA, ITG and EURASIP, editors, *Speech Quality Assessment,* pp. 1-9, Bochum, Germany (1994 Nov.).

[28] S. R. Quackenbush, T. P. Barnwell III, M. A. Clements, “Objective measures of speech quality,” Prentice Hall Advanced Reference Series, New Yersey USA (1988).

[29] ETSI/TM/TM5/TCH-HS, “Correlation of a Perceptual Speech Quality Measure with the Subjective Quality of the GSM Candidate Half Rate Speech Codecs,” Technical Document 92/44, (1992 Dec.).

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

[31] ITU-T Study Group 12, "Improvement of the P.861 Perceptual Speech Quality Measure," Document COM 12-20 (1997 Dec.).

[32] ITU-T Rec. P.48, "Specification for an Intermediate Reference System," International Telecommunication Union, Geneva, Switzerland (1989).

[33] M. P. Hollier, M. O. Hawksford and D. R. Guard, "Characterisation of communications systems using a speech-like test stimulus," *J.Audio Eng. Soc.*, vol. 41, pp. 1008-1021, (1993 Dec.).

[34] ITU-T Study Group 12, "Results of the PESQ (Perceptual Evaluation of Speech Quality) algorithm using speech like test signals," Delayed Contribution D.141 (2000 May).

[35] ITU-R Rec. BS.1116, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunication Union, Geneva, Switzerland (1994 March).

[36] A.W. Rix and M. P. Hollier, "Perceptual speech quality assessment from narrowband telephony to wideband audio", presented at the 107th Convention of the Audio Engineering Society, (2000 Sep.), preprint 5018.