PERCEPTUAL EVALUATION OF SPEECH QUALITY (PESQ) – A NEW METHOD FOR SPEECH QUALITY ASSESSMENT OF TELEPHONE NETWORKS AND CODECS

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

[1] PsyTechnics, B54/86 Adastral Park, Ipswich IP5 3RE, United Kingdom

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

E-mail: awr@iee.org

ABSTRACT

Previous objective speech quality assessment models, such as bark spectral distortion (BSD), the perceptual speech quality measure (PSQM), and measuring normalizing blocks (MNB), have been found to be suitable for assessing only a limited range of distortions. A new model has therefore been developed for use across a wider range of network conditions, including analogue connections, codecs, packet loss and variable delay. Known as perceptual evaluation of speech quality (PESQ), it is the result of integration of the perceptual analysis measurement system (PAMS) and PSQM99, an enhanced version of PSQM. PESQ is expected to become a new ITU-T recommendation P.862, replacing P.861 which specified PSQM and MNB.

1. INTRODUCTION

The motivation for using perceptual models to assess non-linear and error-prone audio communications systems is well-established and models have been proposed by many authors.

Beerends and Stemerdink's model, the perceptual speech quality measure (PSQM) [1], was adopted in 1996 as International Telecommunication Union (ITU-T) recommendation P.861 [2]. An alternative system based on measuring normalizing blocks (MNB) [3], proposed by Voran, was added in 1998 as an appendix to P.861. Another model by Beerends and Stemerdink, the perceptual audio quality measure (PAQM) [4], was combined with several different audio models to produce a method known as perceptual evaluation of audio quality (PEAQ), which became ITU-R recommendation BS.1387 in 1999 [5, 6].

Hollier's extensions to the bark spectral distortion (BSD) model [7] led to the development of the perceptual analysis measurement system (PAMS) [8–11]. This was the first model in the literature to focus on end-to-end behaviour, including the effects of filtering and variable delay [10, 11].

These effects, along with certain types of coding distortion, packet loss and background noise, were found to cause earlier models – such as BSD, PSQM and MNB – to produce inaccurate scores [10–12]. A competition was therefore held by ITU-T study group 12 to select a new model with good performance across a very wide range of codecs and network conditions. The two algorithms with the highest performance in this competition, PAMS and PSQM99 (an updated and extended version of PSQM), were combined to produce a new model known as perceptual evaluation of speech quality (PESQ). This was selected in May 2000 as draft ITU-T recommendation P.862, and is expected to replace P.861 early in 2001 [12, 13].

The next section of this paper presents a description of the structure of PESQ and the key processes that it includes. This is followed by results from 38 known and 8 unknown subjective tests. The scope and limitations of PESQ are also discussed and conclusions are drawn.

2. DESCRIPTION OF PESQ

2.1 Overview

The structure of PESQ is shown in Figure 1. 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 a standard telephone handset. The signals are aligned in time and then processed through an auditory transform similar to that of PSQM. The transformation also involves equalising for linear filtering in the system and for gain variation. Two distortion parameters are extracted from the disturbance (the difference

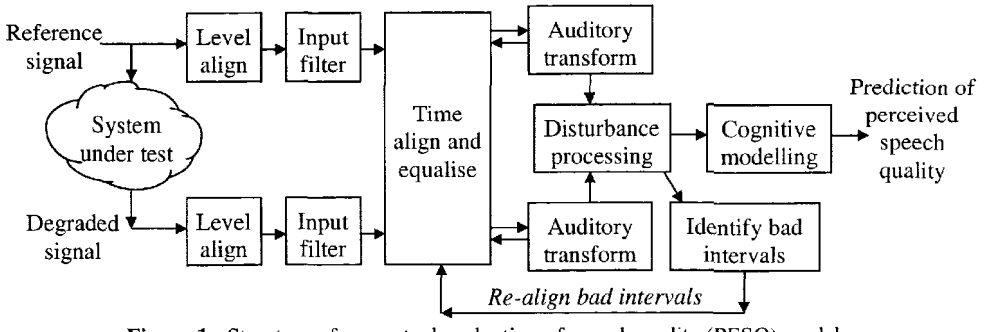

Figure 1: Structure of perceptual evaluation of speech quality (PESQ) model.

0-7803-7041-4/01/$10.00 ©2001 IEEE

between the transforms of the signals), and are aggregated in frequency and time and mapped to a prediction of subjective mean opinion score (MOS). Some details are discussed below.

2.2 Time alignment

The time alignment of PESQ assumes that the delay of the system is piecewise constant. This assumption appears to be valid for a wide range of systems, including packet-based transmission such as voice over IP (VoIP) [10, 11]. 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 [11].

- 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.

These give a delay estimate for each utterance, which is used to find the frame-by-frame delay for use in the auditory transform.

2.3 Auditory transform

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

Bark spectrum. An FFT 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 [2].

Frequency equalisation. The mean Bark spectrum for active speech frames is calculated. The ratio between the spectra of reference and degraded gives a transfer function estimate, assuming that the system under test has a constant frequency response. The reference is equalised to the degraded signal using this estimate, 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.

2.4 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. A deletion (a negative delay change) leaves 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 [14], and were not used in PESQ.

Asymmetry. Unlike P.861 PSQM [2], 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.

2.5 Aggregation of disturbance in frequency and time

Following the understanding that localised errors dominate perception [9], PESQ integrates disturbance over several time-frequency 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 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.

2.6 Realignment of bad intervals

In certain cases the time alignment described in section 2.2 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.

2.7 MOS prediction and model calibration

To train PESQ a large number of different symmetric and asymmetric disturbance parameters were calculated by using multiple 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 3; 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 use of partial compensation in PESQ, for example in equalising for gain modulation, avoids the need for using a large number of parameters to predict quality. A combination of only two parameters – one symmetric disturbance (d_{SYM}) and one asymmetric disturbance (d_{ASYM}) – 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, several design iterations were required. Coefficients in the auditory transform and disturbance processing were optimised then the optimal parameter combination was found, and the process repeated several times. Final training was performed on a database of 30 subjective tests, giving the following output mapping used in PESQ:

$$PESQMOS = 4.5 - 0.1\, d_{SYM} - 0.0309\, d_{ASYM}$$

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

3. PERFORMANCE RESULTS

Following the methodology of the ITU-T competition, we used correlation coefficient and residual error distribution to quantify the performance of models at predicting subjective MOS. These metrics are calculated for each subjective test separately, after mapping the objective scores to the subjective scores for that test in a minimum squared error sense using monotonic 3rd-order polynomial regression. This mapping ensures that the comparison is made in the MOS domain whilst allowing for normal variations in subjective voting between tests.

Tests are grouped according to whether conditions were predominantly from mobile, fixed, voice over IP (VoIP) and multiple type networks. Tables 1 and 2 show correlation and residual error distribution for PESQ, PSQM and MNB [2] for 38 subjective tests that were available to the developers of PESQ. These included a wide range of simulated and real network measurements. Tables 3 and 4 present the results, for PESQ only, of an independent evaluation that was conducted after development was complete. All of this data relates to subjective listening tests carried out on the absolute category rating (ACR) listening quality (LQ) opinion scale. Test material consists of natural speech recordings of 8–12s in duration, with four talkers (two male, two female) for each condition. The results are calculated per condition unless otherwise stated.

No. tests	*Type*	*Corr. coeff.*	*PESQ*	*PSQM*	*MNB*
19	Mobile network	average	0.962	0.924	0.883
		worst-case	0.906	0.841	0.705
9	Fixed network	average	0.942	0.881	0.802
		worst-case	0.902	0.657	0.596
10	VoIP/ multi-type	average	0.921	0.679	0.694
		worst-case	0.810	0.260	0.363

Table 1: Average and worst-case correlation coefficient for 38 subjective tests known during PESQ development, sub-divided by test type.

Absolute error range	<0.25	<0.5	<0.75	<1.0	<1.25
% errors in range, PESQ	74.7	93.9	99.2	99.9	100.0
% errors in range, PSQM	54.6	82.3	92.1	96.7	98.7
% errors in range, MNB	46.1	74.5	89.4	96.1	98.9

Table 2: Error distribution across all 38 known subjective tests.

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

Table 3: Correlation coefficient, 8 unknown subjective tests (PESQ only).

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

Table 4: Error distribution, 7 unknown subjective tests (PESQ only). Test 3 excluded as data was per-file only.

4. SCOPE AND APPLICATIONS

Using results such as those above, a range of applications and test conditions have been identified for which PESQ is believed to give accurate predictions of quality [13]. These include the following.

Codec and error distortions: waveform codecs (e.g. G.711, G.726), CELP/hybrid codecs at or above 4kbit/s (e.g. G.728),

mobile codecs/systems including GSM FR, EFR, HR, AMR, CDMA EVRC, TDMA ACELP, VSELP, and TETRA; transcodings of various codecs; random, burst, and packet loss errors. PESQ can be used for applications such as codec and/or system evaluation, selection and optimisation.

Network behaviours: filtering e.g. due to analogue interfaces; time warping (variable delay) such as packet-based transmission in VoIP. This enables PESQ to be used in a wide range of end-to-end measurement applications with live and simulated networks. Background (environmental) noise, and noise processing, can be assessed by presenting PESQ with the clean, unprocessed original and the coded, noisy degraded signal.

One distortion type, replacement of speech by silence, causes all perceptual models difficulty in predicting MOS. Up to about 50ms of front- and back-end clipping (due to voice activity detection) can have little to no subjective impact. However clipping during speech, e.g. packet loss concealment by silence, is often rated harshly by subjects – with a drop of over 1 MOS for 50ms of clipping. PESQ scores in between these extremes: 50ms clipping typically causes *PESQMOS* to fall by around 0.5 regardless of location. PESQ may thus correlate poorly with subjective MOS if this is a factor – such as test 8 in Table 3.

As a listening-only model with a fixed assumed listening level, PESQ should not be used to assess the effect of listening level, sidetone/talker echo, or conversational delay, and it is not intended for non-intrusive measurements. Certain other applications have not yet been fully characterised or may need parts of the model to be changed. These include: music quality; wideband telephony; the so-called "intermediate audio quality"; listener echo; very low bit-rate vocoders below 4kbit/s; acoustic and head-and-torso simulator measurements.

In contrast, PSQM and MNB were only recommended for use in narrowband codec assessment [2], and were known to produce inaccurate predictions with certain types of codec, background noise, and end-to-end effects such as filtering and variable delay. The scope of PESQ is therefore very much wider than P.861.

5. CONCLUSION

The result of a major international collaboration, PESQ provides significantly higher correlation with subjective opinion than the models of P.861, PSQM and MNB. Results indicate that it gives accurate predictions of subjective quality in a very wide range of conditions, including those with background noise, analogue filtering, and/or variable delay. We believe that PESQ is suitable for many applications in assessing the speech quality of telephone networks and speech codecs.

6. ACKNOWLEDGEMENTS

Thanks are due to ITU-T study group 12 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.

7. REFERENCES

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

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

[3] Voran, S. "Objective estimation of perceived speech quality — part I: development of the measuring normalizing block technique". *IEEE Trans. Speech and Audio Processing*, 7 (4), 371–382, July 1999.

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

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

[6] 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 Audio Engineering Society*, 48 (1/2), 3–29, January/February 2000.

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

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

[9] 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 Proc. Vision, Image and Signal Processing*, 141 (3), 203–208, 1994.

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

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

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

[13] *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.

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