PERCEPTUAL SPEECH QUALITY ASSESSMENT IN ACOUSTIC AND BINAURAL APPLICATIONS

Tom Goldstein and Antony W. Rix

Psytechnics Limited, Fraser House, 23 Museum Street, Ipswich IP1 1HN, United Kingdom

E-mail: tom.goldstein@psytechnics.com, antony.rix@psytechnics.com

ABSTRACT

Perceptual models such as perceptual evaluation of speech quality (PESQ, ITU-T P.862) are now in common use for estimation of listening quality mean opinion score (MOS) of telephone networks and equipment. PESQ was originally designed for evaluation of narrowband telephony, with electrical and/or digital connections to the systems under test. This paper discusses extending PESQ to measurements of terminals, such as mobile or hands-free telephones, using acoustic interfaces, under the working title of ITU-T P.AAM (acoustic assessment model). The changes to the model, and results comparing the extended model with subjective test data, are presented.

1. INTRODUCTION

Speech quality assessment algorithms have evolved considerably in recent years. Early methods, such as PSQM (ITU-T P.861), were limited to the assessment of narrow-band speech codecs. A major work programme in ITU-T Study Group 12 recently standardised the perceptual evaluation of speech quality (PESQ) model as ITU-T Recommendation P.862 [1][2]. P.862 has a much wider scope than P.861, encompassing speech codecs and end-to-end narrowband telephone networks. The structure and current scope of P.862 are outlined in section 2.

A major category of measurement is, however, currently outside the scope of P.862: testing of handsets and other terminals using acoustic interfaces. A project is under way in ITU-T SG12 to create both a new subjective testing method, and an objective acoustic assessment model (AAM), which are suitable for this purpose. The authors have collaborated in this work with Beerends and Berger [3]–[6].

Section 3 of this paper gives an overview of the new categories of acoustic measurement and introduces the subjective test methods that are used. In section 4, the standardisation process and the modifications that have been made to PESQ to create AAM, are summarised. Finally, section 5 gives results to illustrate the performance of the extended model that was submitted to the ITU in June 2003.

2. OVERVIEW AND SCOPE OF PESQ

2.1 Overview

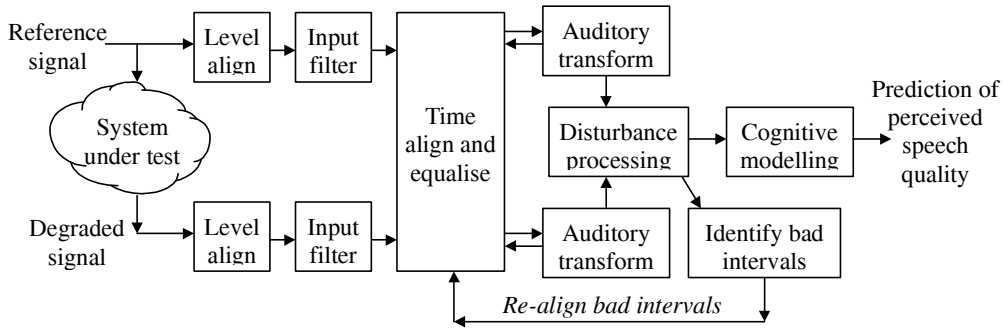
The structure of PESQ is shown in Figure 1 [2]. The model begins by level aligning both the reference and degraded signals to a standard listening level. The signals are then passed through an input filter which models a standard telephone handset. After time alignment, the signals are processed through an auditory transform, including partial equalisation of linear filtering and gain variations in the system under test. Two distortion parameters are computed and averaged in frequency and time. These are then used to compute PESQ score, a prediction of subjective mean opinion score (MOS).

2.2 Scope

PESQ was developed for P.862 with the following assumptions.

- Measurements are made using electrical (2-wire or 4-wire) or digital interfaces to the system under test, at 8kHz or 16kHz sampling rate.
- The user listens through narrowband telephone handset with a 300–3400Hz receive response.
- The equivalent subjective quality scale is P.800 listening quality (excellent, good, fair, poor, bad).

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



PESQ met the ITU's accuracy requirements, based on correlation coefficient, for a wide range of types of network, fixed, mobile, or voice over Internet Protocol (VoIP) – with distortions due to codecs, transcoding, channel errors, delay variations, and linear components such as analog connections. The ITU's test data also included other key factors such as background noise, noise suppression, voice activity detection and error concealment [1].

2.3 Extending PESQ for wider applications

PESQ has good correlation with subjective listening quality across a very large corpus of tests covering a wide range of narrowband telephony applications. However, subjective test data for acoustic applications is relatively scarce. It is therefore highly desirable to alter the model as little as possible when extending it to these wider applications, and to test it across the existing narrowband data as well as new acoustic tests. In P.AAM the only changes between the narrowband and acoustic modes of the model are to take account of the listening equipment (wideband headphones or narrowband handset) and whether listening is monaural or binaural. Using a single core model for both narrowband and acoustic applications reduces the risk of over-training to small data sets and makes maximum use of the breadth of the data for narrowband telephony.

3. CLASSES OF ELECTRICAL OR ACOUSTIC MEASUREMENTS

Acoustic measurements of terminals are made using head-and-torso simulators (HATS), which provide a representative and repeatable physical model of the human user [7]. With either acoustic or electrical interfaces at the send and receive ends of the connection, there are four possible test scenarios. In practice we have found that these can be reduced to two, since the receive interface (electrical or acoustic), is the dominant factor.

3.1 Electrical measurements

This category of measurement includes **electrical-electrical** and **acoustic-electrical**. This is a simple extension of the existing scope of P.862. Material can be sent over either an electrical or acoustic interface, or a filter can be used to model the handset send response. PESQ is already able, without changes, to take account of a range of send filters through its equalisation process. However, its accuracy in acoustic-electrical cases has not previously been studied, and further developments mean that model accuracy can be enhanced in both scenarios.

3.2 Acoustic measurements

This category includes **electrical-acoustic** and **acoustic-acoustic** measurements. Here the signals are recorded through the relevant HATS ear (or potentially, both ears), and are presented to subjects over headphones. Because the signals have already been recorded in the acoustic domain, it is necessary to present them using wideband headphones with flat equalisation. This aims to reproduce, for each subject, the sound as if he/she had been using that terminal in the given position and environment. Because the subjects are unable to move the handsets, effects such as noise shielding and coupling are controlled.

The listening quality (LQ) opinion scale is normally used for these acoustic tests. Other than the presentation equipment, the main modifications to the subjective testing method are to explain to subjects that the signal is presented as if they are using the telephone, and if necessary to introduce the noise environment. In other respects these subjective tests are the same as those used for electrical measurements or for PESQ [1].

Environmental noise at the listener can be included in the test by making recordings "live" or by using techniques to approximately recreate the desired sound field. This makes it possible to assess interactions between the noise and the signal processing in the terminal, for example voice activity detection, echo control or sidetone.

Some aspects of this type of subjective test are described in ITU-T P.832 [8], but the ITU-T is now in the process of updating its recommendations on subjective testing and it is expected that these procedures will be included in a future revision. For this work, the authors and others have published example test plans covering the main classes of acoustic subjective test [5][9].

4. MODEL EXTENSIONS IN P.AAM

4.1 Standardisation process

Work on the development of AAM began in March 2002 with the initiation by ITU-T SG12 of a new competition to produce an objective model for assessment of the quality of terminals as well as networks. Three proponents entered this competition: KPN Research, T-systems and Psytechnics (KPN and Psytechnics co-authored PESQ). After a period of separate study, the three proponents joined forces to share subjective test data and produce a single model, which was submitted in June 2003 [4][6]. This paper describes the results of a candidate model (2a) produced in this collaborative development [6].

At the time of writing, the plans for approval or further development of AAM are subject to ongoing discussions in ITU-T SG12. Points of debate include the subjective testing process for hands-free terminals and their comparison with handsets/headsets, the relative status of AAM and PESQ, and the mapping between AAM score and subjective listening quality.

4.2 Extensions to scope

The main extensions to the scope of P.AAM compared to PESQ can be summarised as follows.

- At the transmit side acoustic, electrical or digital interfaces may be used.
- At the receive side, either (a) an acoustic interface with free field equalisation [7], or (b) an electrical/digital interface may be used. This option is input to the model.
- For case (a), the recorded signal may be monaural or binaural. In the binaural case, both ear signals are used in the model.
- Also for case (a), subjective listening levels may differ between handsets, binaural headsets, and hands-free devices, and the appropriate calibrated level is input to the model.

The recommended sampling rate for acoustic recordings is 16kHz; for narrowband electrical/digital recordings, 8kHz or

16kHz may be used in the same way as PESQ.

4.3 Overview of changes in the model

The following are the main changes between PESQ and AAM.

4.3.1 Level alignment

Level alignment now takes place after time alignment but before the auditory transform. This means that, unlike in PESQ, the signals have been input filtered before level alignment. In the acoustic case, level alignment takes account of the calibrated listening level and binaural audition. Subjective testing of hands-free telephones uses calibrated listening levels of 63–69dB SPL at ERP, and binaural headsets are typically calibrated to 73dB SPL; these compare with 79dB SPL for monaural handset listening [7][8]. Due to binaural effects, stereo listening also produces the perception that the signal is typically 3–6dB louder. These factors are combined to give a new calibrated level that is used for alignment of both reference and degraded signals.

4.3.2 Time alignment and transfer function equalisation

These processes have been optimised for the wider range of filtering that may be present in acoustic conditions. The time alignment algorithm used in PAMS and PESQ [2] may be influenced by the signal spectra, particularly in reverberant cases. Internal parameters such as the pre-filter were adjusted to improve robustness in these cases.

As in PESQ, transfer function equalisation takes place after the first stages of the auditory transform have been performed, as it works in the pitch power domain. However, the method used in AAM is more similar to PAMS [10] than to PESQ.

- The reference VAD is used so that only active speech periods are processed for transfer function estimation, reducing the bias due to noise.
- The bark spectra are smoothed in each frame using a first-order filter at about –30dB/bark, both up and down in frequency, controlling the effect of notches in the response.
- A local coherence factor is also used to reduce the weight during periods of high distortion.
- The weighted phase-less cross-spectrum [10] and reference spectrum are evaluated in bark bands, and are used to compute a stabilised transfer function estimate, further limited to ±15dB in the main speech band.

The smoothed transfer function estimate is then used to equalise the reference signal to the degraded signal.

4.3.3 Binaural processing

The degraded signal may be stereo, in which case both channels are processed through the auditory transform. A binaural masking decision distinguishes whether the signals present in each ear will mask each other, or reinforce each other; this determines whether the ear signals will be treated essentially separately, or be used to model binaural masking by the noise.

The decision uses the output of the time alignment. Essentially, if one ear contains mainly noise, it will not align well with the reference signal, giving highly variable delay estimates and low delay confidence [4]. In this case the decision is to perform noise masking; otherwise the two ear signals are processed independently and the distortion parameters are combined to produce the quality estimate. Noise masking is performed on the outputs of the auditory transform. If the level of the noise exceeds that of the distortion in a given time-frequency cell, the error level is reduced by 50% in distortion parameters D1 and D2, and by 75% in distortion parameters D3 and D4.

4.3.4 Auditory transform

The auditory transform models the time-frequency resolution, masking and loudness perception of the human hearing system. In AAM the auditory transform has been modified to include forward masking – which is not modelled in PESQ – and to improve the modelling of short-term gain variations.

Forward temporal masking is simulated in AAM by first-order smoothing in each bark band, in the power domain. A decay rate of –15dB per 16ms frame period was found to be optimal. Transfer function equalisation is performed after this process.

Short-term gain equalisation is then performed using a soft scaling approach. A scale factor is derived from a regularised ratio of the power in given frame of the reference and degraded signals. This is bounded and processed through a non-linearity, so that small gain differences are fully compensated, but larger differences are only partially compensated. The scale factor is smoothed using a third-order filter and then used to equalise the degraded signal to the reference signal.

4.3.5 Distortion parameters

Because of the number of changes to the model and the increased range of subjective listening conditions, it has been necessary to extend the disturbance processing and to extract two additional parameters. A muting boost process is used to increase the distortion measured during periods of low-level, continuous muting. This was found to improve performance for certain packet loss conditions and is included in all parameters.

In AAM, four distortion parameters are used. Two of these, D1 and D2, are evaluated over speech periods only – ignoring silent periods. The other parameters, D3 and D4, are calculated over the whole signals. D1, D2 and D3 are all symmetric error measures, taking equal account of both positive and negative errors. D4 is an asymmetric error measure.

4.3.6 Cognitive model

The process used to train AAM was broadly the same as for PESQ [2]: an iterative approach was used to select and optimise coefficients and L_p powers for the three stages of parameter averaging, with linear regression at the output stage. However, the final model used four parameters, rather than two in PESQ, and was calibrated onto an arbitrary 0–100 scale rather than a MOS-like scale. The error parameters and output function are the same whether the model is in electrical or acoustic modes.

A further mapping could be applied to estimate MOS on the conventional 1–5 scale [10]. Variation of this mapping to model dependence on subjective context (e.g. hands-free, PSTN, mobile) is under consideration, as a way to deal with large and systematic offsets between subjective tests of different designs.

5. RESULTS

The performance of AAM was assessed using a large database of electrical and acoustic subjective tests. 19 of the electrical tests from the standardisation of PESQ were made available for testing AAM. In addition, a further 9 acoustic tests (4 acoustic-electric, 4 electric-acoustic, and one acoustic-acoustic) were assembled. About half of these electrical and acoustic tests were used in model training, with the remainder held back for validation. A number of other proprietary subjective tests were used for model training, but these did not conform to ITU-T recommendations and are not analysed below. All acoustic tests included both network factors (such as codecs or channel errors) as well as acoustic factors such as handset type/position or noise.

As was the case with PESQ, the main figure of merit used to evaluate AAM is correlation coefficient, evaluated per condition after 3rd-order monotonic polynomial regression to normalise MOS variations between experiments. ITU-T SG12 set required performance thresholds that had to be achieved for AAM to proceed to standardisation. For the electrical tests, required correlation was set 0.01 lower than for PESQ, to allow for the wider range of conditions that AAM must process. For the acoustic tests, minimum correlation was set at 0.90 for all tests.

Table 1 presents average correlation results for PESQ and AAM across the 28 subjective tests introduced above [6], along with the average required correlation for AAM. This table also shows the results for each category of subjective test.

Table 1: Model performance

Dataset	Req.	PESQ	AAM
A. Electric-electric (19)	0.9021	0.9318	0.9424
B. Acoustic-electric (4)	0.9	0.9231	0.9426
C. Electric-acoustic (4)	0.9	0.8449	0.9188
D. Acoustic-acoustic (1)	0.9	0.9318	0.9144
E. Overall (28)	0.9014	0.9181	0.9381

In all but the acoustic-acoustic test, AAM appears to have higher average correlation with MOS than PESQ. The difference between the models for dataset A (electric-electric), 0.0106, is not significant ($P(T<t)=0.08$ using a paired two-sided T-test). However, the difference on the acoustic tests is much larger, and on average, for the overall dataset E, AAM has correlation 0.02 higher than PESQ and this is significant ($P(T<t)=0.015$). Other, unpublished acoustic test data, particularly for hands-free conditions, also indicates that AAM has higher correlation with MOS than PESQ.

The difference between the models is further shown by comparing the results of each subjective test with the ITU requirement. PESQ fails four acoustic cases by between 0.0415 and 0.1512 – the latter leading to a correlation of only 0.7488. AAM fails one (electrical-electrical) case by 0.0006 – it was agreed that this is not significant – and meets the requirement for all of the acoustic tests [6].

6. CONCLUSION

By making a number of changes to the input filter, equalisation, masking and perceptual model, it has been possible to extend PESQ to create an acoustic model, AAM, with a much wider scope. AAM was found to show good correlation with subjective MOS for both electrical and acoustic subjective tests, and has met the ITU's performance requirements for a model for testing terminals and networks. As the perceptual processing is unchanged whether the model operates in electrical or acoustic modes, it should generalise well to other types of distortion.

7. ACKNOWLEDGEMENTS

We would like to thank Lars Birger Nielsen at Brujel & Kjaer for assistance with HATS equipment and making HATS measurements. We have collaborated in the P.AAM development with John Beerends at TNO Telecom (formerly KPN Research), and Jens Berger at T-systems. In particular, several innovations described in section 4 are due to Beerends. Colleagues at Opticom provided assistance with implementation of the P.AAM software. Further data for the development and evaluation of P.AAM has been provided by a number of participants in ITU-T SG12, including Lucent Technologies, France Telecom R&D, Swissqual, Opticom and BTexact. Finally, we wish to thank our colleagues in Psytechnics for their help in many aspects of this work.

8. REFERENCES

[1] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.* ITU-T Rec. P.862, Feb. 2001.

[2] Rix, A.W., Beerends, J.G., Hollier, M.P. and Hekstra, A.P. "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs." *IEEE ICASSP,* Vol. 2, 749-752, May 2001.

[3] Beerends, J.G., Hekstra, A.P., Rix, A.W. and Hollier, M.P. "Proposal for the use of draft recommendation P.862, the perceptual evaluation of speech quality (PESQ), for measurements in the acoustic domain with background masking noise." ITU-T SG12 COM12-D6, Feb. 2001.

[4] Goldstein, T., Klaus, H., Beerends, J. G. and Schmidmer, C. "Draft Recommendation P.AAM – An objective method for end-to-end speech quality assessment ofnarrow-band telephone networks including acoustic terminal(s)." ITU-T SG12 COM12-C64, July 2003.

[5] Beerends, J. G., Berger, J. and Rix, A. W. "Testplan for the benchmarking of mouth-to-ear speech quality assessment algorithms including terminals." ITU-T SG12 COM12-D127, Jan. 2003.

[6] Berger, J. "Report of the Rapporteur's Meeting in Berlin, 25-28 June 2003." ITU-T SG12 COM12-C58, July 2003.

[7] *Head and torso simulator for telephonometry.* ITU-T Rec. P.58, Aug. 1996.

[8] *Subjective performance evaluation of hands-free terminals.* ITU-T Rec. P.832, May 2000.

[9] Veaux, C., Juric, P., Kim, D.-S., Gray, P., and Schmidmer, C. "Joint Test Plan for Single-Ended Assessment Models." ITU-T SG12 COM12-D121, Jan. 2003.

[10] Rix, A. W. and Hollier, M. P. "The perceptual analysis measurement system for robust end-to-end speech quality assessment." *IEEE ICASSP,* Vol. 3, 1515–1518, June 2000.