PERCEPTUAL SPEECH QUALITY ASSESSMENT – A REVIEW

Antony W. Rix

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

E-mail: antony.rix@psytechnics.com

ABSTRACT

This paper reviews the development of perceptually-motivated models for quality assessment of speech transmission/storage systems. The aim is to predict subjective mean opinion score (MOS) for non-linear, time-variant distortions such as lossy coders, channel errors or noise reduction, particularly for telecommunications applications.

Because linear methods have proven unsuitable for this purpose, many researchers studied perceptual quality assessment, using a comparison of auditory transforms to estimate quality. This work has led to several ITU standards. Non-intrusive models, arguably more suited to network monitoring, are the focus of much current interest. Intrusive, signal-based non-intrusive, and parametric non-intrusive models are discussed.

1. INTRODUCTION

Telecommunications networks have become increasingly complex and non-linear since the early 1990s, as low-bit rate speech coders and links that may be subject to transmission errors have become widespread, in particular for long-distance, mobile and voice over IP (VoIP) connections. This paper describes the development of methods that can be used to measure the quality of these networks, as perceived by the end-user. Such methods can be used in turn to optimise the networks for quality, capacity or cost, and in network management to monitor the quality in terms of customer experience.

As well as distorting the signal by low bit-rate perceptual speech coding, telephone networks may include error processes such as bit errors or packet loss, error concealment, discontinuous transmission with comfort noise insertion during silent periods, and noise reduction. These can seriously affect the quality – in a way that cannot be accurately modelled by simple objective measures such as root mean squared error (RMSE), noise level, or frequency response.

For example, if white noise is added to speech or music at a signal-to-noise ratio (SNR) of 13dB, it is clearly audible and disturbing. However, if the noise is shaped in time and frequency to be below the masked threshold [1], it may be almost impossible for a human to detect any disturbance, even in stringent listening conditions. Yet in these two cases, the noise level, SNR, RMSE and linear frequency response of the system under test are identical.

Because objective metrics such as these correlate badly with users' perceptions of quality, subjective testing became the key measure of quality of these systems. In a typical listening test, subjects hear recordings processed through about 50 network conditions, and vote on a simple opinion scale such as the common 5-point listening quality (LQ) scale [2]. The average score for a condition, across all subjects, is termed mean opinion score (MOS). These methods are described further in section 2.

Subjective tests are, however, slow and expensive to conduct, making them accessible only to a small number of laboratories and unsuitable for real-time monitoring. Computational models that accurately predict MOS would be more suitable for field applications, motivating the research described below. Section 2 also describes methods used to evaluate model performance.

Section 3 of this paper provides an introduction to intrusive models, which use both the input and output of the system for quality assessment and generally require an active measurement. Section 4 describes parametric non-intrusive models that use measurements or estimates of network properties to predict the quality. Section 5 describes signal-based non-intrusive models, which require only the processed signal and can be applied to monitor general, live traffic.

2. SUBJECTIVE TESTS

A very wide range of audible distortions can be caused by the systems and processes introduced above. Subjective testing methods have been developed to provide an overall score of the quality of a system or service from the customer's viewpoint, independent of the underlying technology used in the network.

2.1 Listening quality

In listening tests, subjects hear a number of distorted recordings, and vote on their opinion of the quality. The absolute category rating (ACR) LQ method has been the most commonly used subjective test procedure in telecommunications [2]. This is a 5-point discrete scale, using the labels excellent, good, fair, poor, and bad, assigned integer values from 5 to 1. Early perceptual model research considered other opinion scales, in particular the diagnostic acceptability measure (DAM) [3]. Other methods used in audio quality tests are described in ITU-R BS.1116 [4].

Careful test design can control some undesirable factors that influence the voting process, such as dependence on presentation order. One important variable that is only partially controlled in the ACR LQ method is the subjective scale itself: depending on the range of conditions, and the subjects' cultural interpretation of terms such as excellent, there can be systematic offsets as large as 1.0 MOS between tests or from different countries.

Methods that use anchors, such as comparison category rating (CCR) where pairs of signals are presented, one of which is usually a clean reference, can be used to fix the point of zero audible distortion [2]. These methods are less commonly used than ACR LQ, partly because they take longer to conduct, and perhaps also because only one point is anchored – the way that scores are allocated across the scale is still influenced by subjective and cultural interpretations and by the distribution of quality of the conditions in the test.

2.2 Conversation quality

Because they do not include any interaction between users, listening tests cannot model some important effects that emerge only in two-way communication. In conversational tests, pairs of subjects talk over a test connection before voting on its quality, often using the standard 5-point quality scale (excellent…bad). This can take account of the whole link, including network, handsets and sidetone, echo, level and delay impairment [2][5]. As they are generally more expensive, and can investigate fewer conditions than listening tests, conversational tests are relatively rare. Parametric models such as the E-model, which is described in section 4, are sometimes used in their place.

2.3 Performance assessment of models

If perceptual models are to be used in place of subjective tests, their accuracy must be evaluated by comparison to subjective test data. However, this is potentially difficult because subjective tests themselves use small sample sizes (typically 24 subjects), sometimes only partially control variables such as material dependence, and exhibit variations in the voting scale between tests as described above. Because of this, relatively simple correlation-based methods have become the main figure of merit in recent work in the ITU [6][7].

The preferred method evaluates model performance separately for each subjective test. Offset and non-linearity in the relationship is eliminated by applying a monotonic function (typically fitted for minimum RMSE) to map the objective scores onto the subjective scale. Performance is measured using the Pearson correlation coefficient. The residual error distribution, computed after the mapping function, may also be evaluated. Normally these measures are calculated per condition, as this reduces the influence of talker and material dependence [6].

Much initial work used the logistic function for this monotonic mapping [8][9]. The logistic can become very flat – artificially improving the performance of models with poor prediction power at the extremes – but can only take on a small range of curvature modes which may not match those in subjective tests. Because of these limitations, the monotonic function used in most recent standards work is the monotonic cubic polynomial, which has the same number of degrees of freedom (4) as the logistic. A monotonic function is necessary because order must be preserved; this is usually achieved by fitting the polynomial using a gradient descent method with a cost constraint.

The range of conditions that may be encountered in telecommunications networks is enormous (see for example [6]). Perceptual models often include tens or hundreds of internal coefficients, making over-training a strong possibility. It is therefore of vital importance to use a large number of subjective tests to evaluate model performance. The selection of ITU-T P.862 [6] used 22 subjective tests known to the authors, and 8 unknown tests run by independent laboratories, containing about 1300 conditions in total.

It may seem like an obvious point, but researchers and end-users should be highly sceptical of the accuracy of models where few subjective test results are reported and especially where there is no independent validation.

3. INTRUSIVE MODELS

Intrusive test methods pass a known (reference) signal through the system under test, capture the processed (degraded) signal, and compare the two to derive a quality score that should correlate well with MOS.

3.1 Masked-error models

One of the first applications of perceptual models for quality assessment was proposed by Schroeder et al, who used a simple masking method to estimate the audibility of coding noise in a speech coder [10]. This was extended by Brandenburg to give a measure of the mean noise to masking ratio (NMR) [11]. These methods basically assume that any (time-domain) difference between the original and processed signals is noise, leading to poor performance when this does not hold, for example in filtering, phase jitter, or re-synthesis.

3.2 Models based on comparison of auditory transforms

Karjalainen introduced a more general technique for estimating error audibility based on auditory spectrum distance (ASD), a comparison of audible time-frequency-loudness representations [12]. This approach can be adapted to simulate a much wider range of perceptual effects, and has been much more successful. Although a successful implementation was demonstrated by Karjalainen, many later authors did not cite this work.

A wide range of models for extracting distortion parameters were described by Quackenbush, who developed models for predicting DAM-derived quality scores [3] using measures such as cepstral distance. Although he did consider using a Bark spectrum, and mentioned the problem of spectral tilt that was later re-interpreted as transfer function equalisation, Quackenbush did not strongly pursue the perceptual analogy. Similar objective metrics were used in many models (see [8] for references), and as recently as 1998 in the measuring normalizing blocks model (MNB), which used a multi-scale method to compute a quality score from the difference between logarithmic spectrograms of the signals [8][9]. For intrusive applications, however, the perceptual approach has become dominant.

Several new perceptual models for measuring the quality of speech and audio coders emerged in the early 1990s. Wang et al. took an approach similar to that of Karjalainen, although without temporal masking, to compute loudness on a Sone scale in Bark bands, and evaluate the mean squared Bark spectral distance (BSD) [13]. This approach was generalised by Hollier, who noted that multiple distortion parameters must be computed for a more general prediction, to model not only the amount but also the distribution of errors [14].

The perceptual method was also explored for quality assessment of audio coders and systems [15]–[19], leading in 1999 to an ITU-R standard model, perceptual evaluation of audio quality (PEAQ) [20]. Although audio quality is not the focus of this paper, some of these authors introduced concepts that were later used in the speech quality models described below.

Beerends and Stemerdink's perceptual audio quality measure (PAQM) introduced the asymmetry factor, weighting the difference in each time-frequency cell by the power ratio of the reference and degraded signals. This amplifies loud additive distortions, emulating perceptual streaming [15]. This was adapted, including the removal of masking, into a method for speech coder evaluation known as the perceptual speech quality measure (PSQM) [21]. After a competition, PSQM was adopted as ITU-T P.861 in 1996 [8].

3.3 Models for network testing

Most of the models described above were developed for testing speech or audio coders. Real networks introduce level changes, unknown delay and/or linear filtering – all of which may vary dynamically. If these are not taken into account, they may lead to large false errors being observed in intrusive models that use the method of comparison of auditory transforms, causing highly inaccurate quality scores. From the mid-1990s the focus of intrusive model research shifted to solving these problems and using large databases of training data, to make models that remained accurate when used in the field.

Linear filtering may occur in many places in audio or speech transmission systems, and is generally less disturbing than non-linear coding distortions. This was modelled by Thiede for audio quality assessment, by estimating the frequency response smoothed in time and frequency, and equalising the signals to eliminate the error due to steady-state differences [18][20]. A similar approach was mentioned for speech quality assessment by Berger [22], although few details were given. Rix used a combination of phaseless cross-spectrum-based transfer function equalisation and spectral difference, for partial equalisation in a model based on that of Hollier, known as the perceptual analysis measurement system (PAMS) [23]. A similar, unpublished method was developed by Beerends and Hekstra in an improvement of PSQM known as PSQM99.

Time-delay proved to be a significant challenge as VoIP became widespread in the late 1990s. Variations in packet delay or dynamic jitter buffer re-sizing may lead to a change in the end-to-end audio delay [23]. The comparison method requires the reference and degraded signals to be aligned, but few early models provided an algorithm for delay assessment, and none could deal with time-varying delay. Rix and Reynolds introduced a set of methods in PAMS to identify delay changes between and during speech utterances, and used this to improve model accuracy for conditions including delay variation [23].

Because it lacked these processes, PSQM [8] was not suitable for network testing, and a competition was held to replace it. This was jointly won by PSQM99 and PAMS, which were then integrated – using the time alignment of PAMS and the auditory transform of PSQM99 – to produce a new model known as perceptual evaluation of speech quality (PESQ) that was standardised as ITU-T P.862 [6] in 2001, and P.861 was withdrawn. The average correlation of PESQ with MOS on both known and unknown subjective test data was found to be 0.935 in the ITU-T evaluation [6]. A separate paper at this conference describes recent work by Goldstein, Rix, Beerends and Berger to extend PESQ for acoustic and binaural applications for testing both telephone networks and terminals such as handsets.

4. PARAMETRIC MODELS

Simple computational models have been proposed to estimate the quality of a network without the need to run subjective or intrusive tests, for network planning [5] or non-intrusive measurement [24]. Methods based on this parametric approach have been developed by two companies for non-intrusive, real-time quality assessment of VoIP systems.

4.1 The E-model

Developed as a tool for network designers, the E-model has become a popular framework for estimating the quality of networks [5]. It produces a transmission rating *R*, which can be used to estimate conversation quality. An additive relationship between component factors is assumed. It must be stressed that while this makes the model simple and easy to understand, this assumption is known to be wrong in some cases. Because of this and the limited validation of the E-model for modern network configurations, it is recommended for network planning only.

4.2 Parametric conversation quality measures

For trunk network links that are subject to minimal coding distortions or channel errors, the main factors that affect conversation quality are speech level, noise level, talker echo and delay. These parameters may be measured using proprietary in-service non-intrusive measurement devices (INMDs), typically at international switches, and can then be used in parametric models to estimate conversation quality. The use of the E-model for this purpose, and an alternative model known as the call clarity index (CCI), are described in ITU-T P.562 [24].

4.3 Parametric models of VoIP quality

Two competing parametric approaches to real-time assessment of VoIP quality have emerged in the last three years. Both use parameters from the RTP voice packet stream to compute speech quality impairments, can estimate delay from the RTCP stream, if available, and may be integrated into the E-model framework. Clark proposed a method, based on the Gilbert-Eliot hidden Markov model for bursty packet loss, to calculate the coding/ error impairment factor *Ie* that is an input to the E-model [25].

Broom, Reynolds, Hollier and Barrett have argued that this does not take into account large differences between VoIP devices such as gateways and IP phones, for example in the implementation of the jitter buffer and error concealment. They have proposed a method to calibrate, using PESQ, a proprietary multi-parameter model for a specific edge device to allow a more accurate quality estimate [26]. An ITU-T competition to select one of these methods for non-intrusive parametric evaluation of VoIP is currently in progress [27].

5. SIGNAL-BASED NON-INTRUSIVE MODELS

The parametric models described above can only be used with certain types of network, such as VoIP. General non-intrusive (also known as no-reference or single-ended) methods require only the processed signal and can be used with live traffic. Unlike the parametric models, these process the audio stream to extract distortion indicators, which are used to estimate MOS.

Work in this area was pioneered by Gray, who used a vocal tract model to identify distortions, for example through physiologically implausible shapes or transitions. Gray also described a calibration procedure using PAMS [28]. Beerends and Hekstra considered a method based on distortion detection by integrating the perceptual model of PESQ into the non-intrusive model [29]. An alternative approach using a single-ended auditory model for distortion identification using analysis of frequency modulation and articulation was proposed by Kim [30].

ITU-T SG12 has recently selected a non-intrusive model based on the work of Gray, Beerends and others, after a competition under the working title P.SEAM [7]. The winning candidate showed average correlation of 0.884 with 22 known subjective tests, and 0.814 with 6 unknown tests. This exceeds the performance of PSQM, which has an average correlation of 0.81 over a similar dataset of 22 tests, indicating that this non-intrusive model compares favourably with the first generation of intrusive perceptual models.

6. REFERENCES

[1] Moore, B. C. J. *An introduction to the psychology of hearing*. 4th edition, Academic Press, 1997.

[2] *Methods for subjective determination of transmission quality.* ITU-T Rec. P.800, Aug. 1996.

[3] Quackenbush, S. R., Barnwell III, T. P., Clements, M. A. *Objective measures of speech quality.* Prentice Hall, 1988.

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

[5] *The E-model, a computational model for use in transmission planning.* ITU-T Rec. G.107, July 2002.

[6] *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.

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

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

[9] 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.

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

[11] Brandenburg, K. *Evaluation of quality for audio encoding at low bit rates*. 82nd AES Convention, London, preprint 2433, March 1987.

[12] Karjalainen M. *A new auditory model for the evaluation of sound quality of audio system.* IEEE ICASSP, Tampa, Florida, 608–611, March 1985.

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

[14] 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.

[15] 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, Dec. 1992.

[16] Paillard, B., Mabilleau, P., Morisette, S. and Soumagne, J. PERCEVAL: perceptual evaluation of the quality of audio signals. *Journal of the Audio Engineering Society*, 40 (1/2), 21–31, Jan. 1992.

[17] Colomes C., Lever M., Rault J.B. and Dehery Y.F. A perceptual model applied to audio bit-rate reduction. *Journal of the Audio Engineering Society*, 43 (4), 233-240, April 1995.

[18] Thiede, T. and Kabot, E. *A new perceptual quality measure for bit rate reduced audio.* 100th AES Convention, Copenhagen, preprint 4280, May 1996.

[19] Sporer T. *Objective audio signal evaluation – applied psychoacoustics for modeling the perceived quality of digital audio.* 103rd AES Convention, New York, preprint 4512, Oct. 1997.

[20] *Method for objective measurements of perceived audio quality.* ITU-R Rec. BS.1387, Jan. 1999.

[21] 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, March 1994.

[22] Berger, J. *TOSQA – Telecommunication objective speech quality assessment,* ITU-T COM12-34, Dec. 1997.

[23] 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.

[24] *Analysis and interpretation of INMD voice-service measurements.* ITU-T Rec. P.562, May 2000.

[25] Clark, A. "Description of VQmon algorithm." ITU-T COM12-D105, Jan. 2003.

[26] Broom, S. "High Level Description of Psytechnics ITU-T P.VTQ candidate." ITU-T COM12-D175, Sept. 2003.

[27] Barriac, V. "Status report of P.VTQ model selection." ITU-T COM12-D166, Sept. 2003.

[28] Gray, P., Hollier, M. P. and Massara, R. Non-intrusive speech quality assessment using vocal-tract models. *IEE Proc. Vision, Image and Signal Processing,* 147 (6), 493–501, Dec. 2000.

[29] Beerends, J. G., Gray, P., Hekstra, A. P. and Hollier, M. P. "Call for proposals for a single-ended speech quality measurement method for non-intrusive measurements on live voice traffic." ITU-T COM12-C11, Nov. 2000.

[30] Kim, D.-S. "ANIQUE: Lucent Technologies' candidate algorithm for ITU-T single-ended speech quality assessment model." ITU-T COM12-D181, Sept. 2003.