Proceedings of the Institute of Acoustics

INTELLIGIBILITY VS QUALITY IN OBJECTIVE SPEECH QUALITY ASSESSMENT

M P Hollier BT Labs, BT Adastral Park, Martlesham Heath, IPSWICH IP5 3RE

1. INTRODUCTION

This paper describes the range of audio qualities relevant to the communications industry, the different subjective opinion scales used to evaluate these qualities and the development of objective methods to predict speech performance.

The convergence of computing, communications and content is exemplified by the growing capabilities of the Internet and the new services that are possible with the introduction of broadband connections to the home. There is a wide range of audio bandwidths and qualities associated with the radical variety of products and services possible with:

- Broadcast/entertainment
- Global IP networks and Internet 2
- Fixed mobile convergence
- Telephony

Understanding and optimising performance across such a wide range of applications requires a number of different subjective opinion scales. This is now well established in both subjective testing and objective assessment methodologies. Even within the scope of telephony applications the range of performance between toll-quality fixed-networks and data-compressed mobile systems is sufficient to require two opinion scales for characterisation.

There are many situations during design, commissioning and monitoring where subjective testing is impractical or impossible. Even when a subjective test could be used it is expensive and time consuming compared with an objective measurement. Modern communications systems contain complex non-linear processes such as low-bit-rate coding and cannot be adequately characterised with conventional engineering measures. A new generation of measurement methods based on models of human perception, have been developed which can predict subjective performance. A particular feature of BT's PAMS (Perceptual Analysis/Measurement System) is that it is able to predict on two different opinion scales: one predicting perception of quality, and the other predicting listening effort.

The different opinion scales employed to characterise communications systems are described, together with appropriate experimental methodologies, in section 2. Objective performance is introduced in section 3 including an overview of PAMS. Finally in section 4 a number of conclusions are drawn regarding the applicability of different opinion scales.

2. SUBJECTIVE OPINION SCALES

Audio and telephony performance are often evaluated using listening tests – in which subjects are presented with a succession of audio samples and asked to vote. Votes are averaged across listeners and conditions to produce a mean opinion score, or MOS. The methodology for this type of subjective test for telephony is given in ITU-T P.800 [1]. Tests suitable for more

general audio and broadcast tests are described in ITU-R BS.1116 [2]. A more detailed summary of typical experimental methods can be found in [3].

Listening test methodologies include ACR (Absolute Category Rating), where the subject provides their opinion for a given randomly selected speech sample against a fixed scale, and DCR (Degradation Category Rating) where an A-B sample is assessed - where A is always the reference. To illustrate the differences between these scales a common set of conditions were assessed using two ACR and one DCR opinion scales and was reported in [4].
When discussing speech performance across a range of applications it is interesting to concentrate on the two ACR opinion scales: listening quality and listening effort. These scales have been found to be especially useful to assess a range of performance. In low distortion conditions all the samples are intelligible and the quality determines subjective opinion, while with highly degraded conditions quality is uniformly low and intelligibility becomes the determining factor for subjective opinion.

Listening Quality, LQ		**Listening Effort, LE**
Quality of the speech		*Effort required to understand the meanings of sentences*
5	*Excellent*	*Complete relaxation possible; no effort required*
4	*Good*	*Attention necessary; no appreciable effort required*
3	*Fair*	*Moderate effort required*
2	*Poor*	*Considerable effort required*
1	*Bad*	*No meaning understood with any feasible effort*

The two opinion scales cause subjects to behave differently and produce different scores for a given condition. This effect is due to the different psychological targeting of the question and available responses. It can be seen that the task performed by the subject is different in each case: *What is the quality?* vs. *Can you understand what is said?* Hence the effort required to understand the speech depends on the intelligibility of the sample.

Figure 1 shows the subjective opinions of 24 subjects listening to speech samples degraded by MNRU (Modulated Noise Reference Unit) [5] and voting against LQ, DCR and LE scales. MNRU is used as reference conditions in many subjective tests and is intended to represent an increasing amount of quantisation noise as might occur with an increasing number of successive transcodings.

From Figure 1 it is apparent that

- LE and LQ scales produce different scores for a given condition.
- LE saturates first at the high-quality end of the range, since quality can continue to improve after the speech is completely intelligible.
- LQ saturates first at the low-quality end of the range since the intelligibility can continue to degrade after subjects have given the lowest quality score.

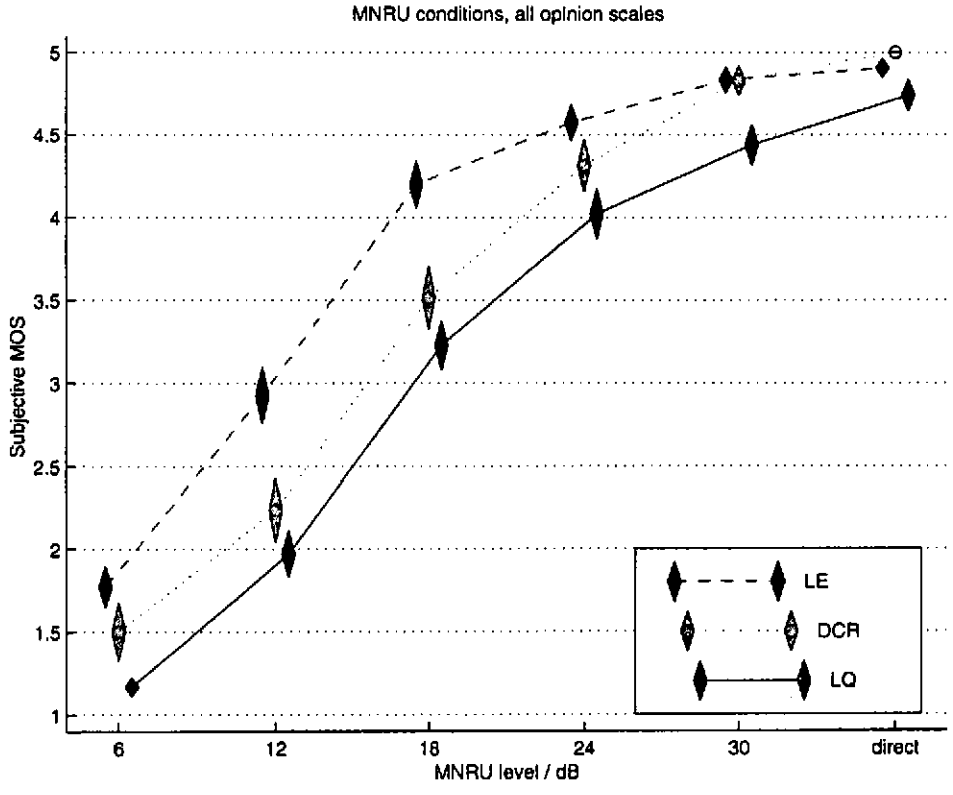

Figure 1: MNRU conditions, from [4].

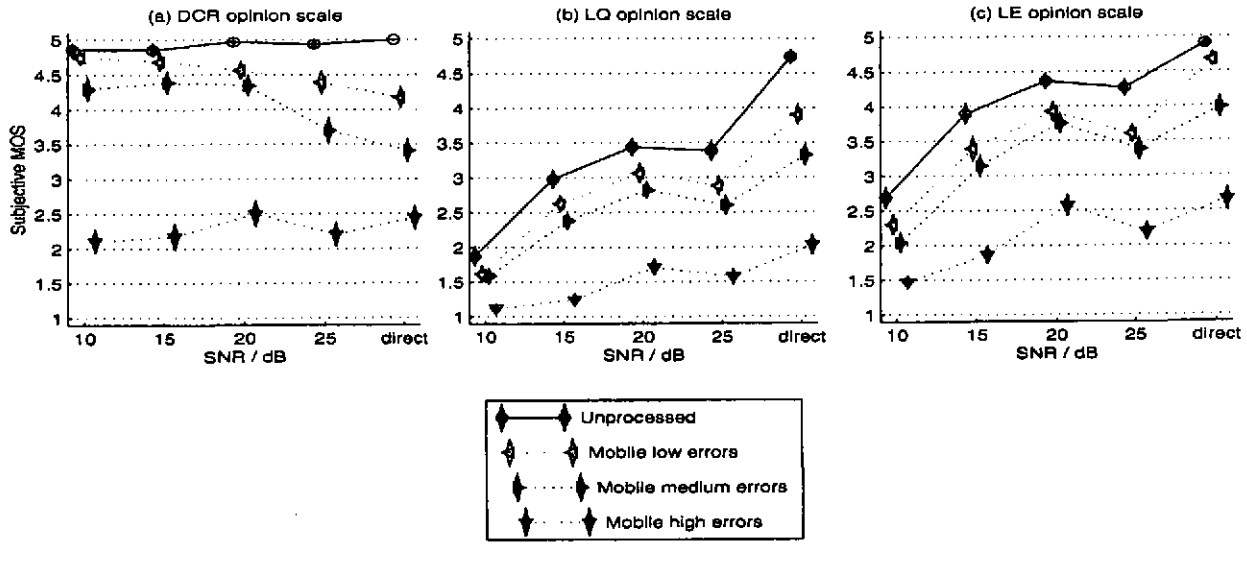

Figure 2: Mobile conditions, from [4]

Figure 2 shows a series of representative real-world conditions assessed against different opinion scales. The conditions cover a wide range of performance in different background noise conditions. It can be seen that as the SNR (due to background noise) improves LE

increases more quickly and is better able to differentiate between conditions at low SNR. Similarly, LQ is relatively saturated for the high error conditions but is better able to differentiate between conditions at high SNR.

3. OBJECTIVE PERFORMANCE ASSESSMENT

Subjective testing is expensive and time consuming compared with objective measurement. Further it is not always possible to perform a subjective test during practical commissioning and monitoring operations. The need for new objective measures to allow reliable assessment of non-linear systems has been extensively discussed elsewhere [6]. Perceptually motivated audio assessment is relatively advanced and numerous models have been proposed for objectively assessing both high quality audio [7,8] and telephone speech quality [9,10].

PAMS is an auditory-model based speech-quality assessment system designed for robust end-to-end measurements. This distinguishes PAMS from other models and from the current ITU-T standard P.861 which is intended for codec assessment and is not therefore recommended for network assessment where unknown non-linear and linear distortions may occur. Figure 3 shows the basic structure of PAMS: the reference and degraded signals are passed through an auditory transform to model the acoustic transfer functions of the outer and middle ear and reproduce the main functions of the inner ear using a filter bank model.

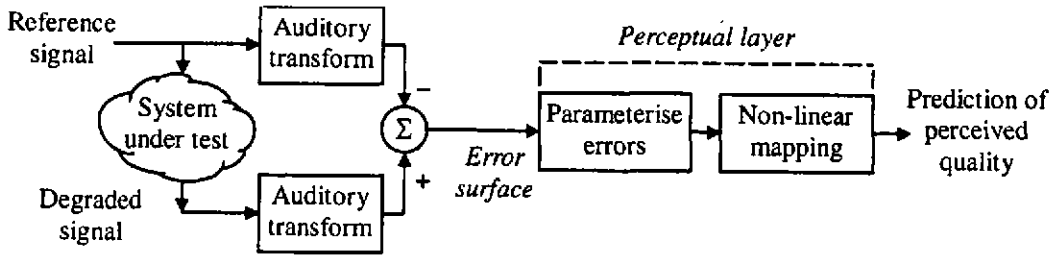

Figure 3: Structure of BT's PAMS perceptual model

The prediction of audible differences between a degraded and reference signal can be thought of as the *sensory layer* of a perceptual analysis, while the subsequent categorisation of audible errors can be thought of as the *perceptual layer*. Models for assessing high quality audio have tended only to predict the probability of detection of audible errors since any audible error is deemed to be unacceptable, while early speech models have tended to predict the presence of audible errors and then employ simple distance measures to categorise their subjective importance [9,10,11].

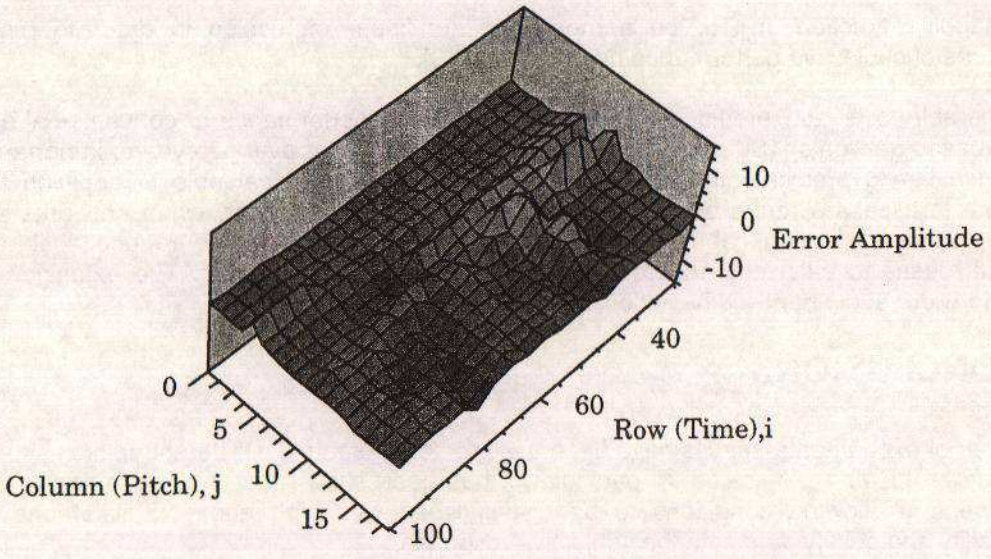

Figure 4, Fragment of audible error surface.

It has been previously shown [12] that a more sophisticated description of the audible error provides an improved correlation with subjective performance. In particular, the amount of error, distribution of error, and correlation of error with original signal have been shown to provide an improved prediction of error subjectivity. Figure 4 shows a hypothetical fragment of an error surface. The error descriptors used to predict the subjectivity of this error are necessarily multi-dimensional, i.e. no single dimensional metric can be contrived to map between the error surface and the corresponding subjective opinion. The error descriptors, E_d, are typically in the form:

$$E_{d1} = fn_1\{e(i,j)\}$$

where fn_1 is a function of the error surface element values for descriptor 1. For example the error descriptor for the distribution of the error, Error-entropy (Ee), proposed in [12] was given by:

$$Ee = \sum_{i=1}^{n} \sum_{j=1}^{m} a(i,j) \ln a(i,j) \qquad \text{where } a(i,j) = |e(i,j)| / Ea$$

and Ea is the sum of |e(i,j)| with respect to time and pitch. More sophisticated error descriptors may also operate on subsets of the filter bank output and distinguish between additive errors and signal loss.

$$\text{Opinion prediction} = fn_2\{E_{d1}, E_{d2}, ..., E_{dn}\}$$

where fn_2 is the mapping function between the n error descriptors and opinion scale of interest.

It has been shown that a judicious choice of error descriptors can be mapped to a number of different subjective opinion scales [6]. This is an important result since the error descriptors can be mapped to different opinion scales that are dominated by different aspects of error subjectivity. It has also been shown that it is highly advantageous to constrain the

final mapping function, fn_2, to be a monotonic non-linear regression in order to make predictions of subjective performance highly robust [13].

PAMS is able to reliably predict the subjective end-to-end performance of conventional and packet-based networks [13], and has been in commercial use for over two years. Importantly PAMS is able to predict both LE and LQ and this has proved valuable in applying the algorithm to assess a range of communications systems. In particular, when networks are constructed using leased capacity data compression must be used to reduce cost and practical means to efficiently ensure adequate quality is highly valuable. The extension of PAMS to wider audio bandwidths is described in [3].

4. CONCLUSIONS

A number of experimental results have been used to demonstrate the different opinion scores produced by LE and LQ scales. In particular, it has been shown that these two scales are complimentary allowing conditions to be distinguished with high and low distortions as intelligibility and quality dominate the subjective response.

The value of objective speech performance measurement, to reduce time and cost, has also been introduced. An overview of BT's PAMS speech quality assessment system has been given including the capability of the system to predict LE and LQ scales.

REFERENCES

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

[2] ITU-R BS.1116 "Methods for the subjective assessment of small impairments in audio systems including multi-channel sound systems", July 1998

[3] Rix A W, Hollier M P, "Perceptual speech quality assessment from narrowband telephony to wideband audio", Presented to the 107th AES Convention in New York, Preprint No.5018, September 1999.

[4] Rix A W, "Comparison Of Opinion Scales For Subjective Listening Tests", Delayed Contribution to ITU-T 15,22/12, D.102, Geneva, September 1999.

[5] ITU-T Recommendation P.810 "Modulated noise reference unit (MNRU)", Feb. 1996.

[6] Hollier M P, Sheppard P J, "Objective speech quality assessment: towards an engineering metric ", Presented at the 100th AES Convention in Copenhagen, Preprint No.4242, May 1996.

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

[8] Paillard B, Mabilleau P, Morissette S, Soumagne J, "PERCEVAL: Perceptual Evaluation of the Quality of Audio Systems.", J. Audio Eng. Soc., Vol.40, No.1/2, Jan/Feb 1992.

[9] Hollier M P, Hawksford M O, Guard D R, "Characterisation of Communications Systems Using a Speech-Like Test Stimulus", J. Audio Eng. Soc., Vol.41, No.12, December 1993.

[10] Beerends J, Stemerdink J, "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation", J. Audio Eng. Soc., Vol.40, No.12, December 1992.

[11] Wang S, Sekey A, Gersho A, "An Objective Measure for Predicting Subjective Quality of Speech Coders", IEEE J. on Selected areas in Communications, Vol.10, No.5, June 1992.

[12] Hollier M P, Hawksford M O, Guard D R, "Error-activity and error entropy as a measure of psychoacoustic significance in the perceptual domain", IEE Proc.-Vis. Image Signal Process., Vol.141, No.3, June 1994.

[13] Rix A W, Reynolds R, Hollier M P, "Perceptual Measurement of end-to-end speech quality over audio and packet-based networks", Presented to the 106th AES Convention, Munich, Preprint 4873, May 1999.