A NEW GENERATION OF PERCEPTUALLY BASED SPEECH QUALITY MEASUREMENTS

M P Hollier Psytechnics Ltd, Ipswich, UK

1 INTRODUCTION

This paper provides an introduction to a new generation of measurements that provide a real-time measurement of customer perception of quality. The area of perceptual engineering is relatively new and the main methods in use today have all been developed in recent years [1-3]. The measurement methods presented here were pioneered over more than a decade by a core team initially within BT and for the last five years within the BT spin-out company Psytechnics Ltd. The work has generated more than 35 patents and 4 ITU-T world standards.

2 FAILURE OF CONVENTIONAL METRICS

2.1 New technologies

With the introduction and standardisation of new technologies for telephony services that introduce new types of distortions, such as:

- Voice over IP (packet loss and variable delay),
- Voice over ATM (cell loss),
- voice over mobile (GSM, frame repeat, front end clipping, comfort noise generation),
- ETSI GSM EFR/AMR coding,
- ITU-T G.728/729/723.1 coding, 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 characterisation, mostly carried out on the basis of a nearly linear, time invariant system, looses meaning with these new technologies.

2.2 Simple network statistics

Engineering metrics such as average packet loss and jitter fail to provide an indication of the IP bearer's ability to deliver acceptable voice quality. For example an average packet loss of 3% could lead "bad" speech quality or "good" speech quality:
scenario 1
Average packet loss is 3%: packet loss is evenly distributed in time: the packet stream is decoded by a soft-phone with a well designed adaptive jitter buffer and packet loss concealment. Speech Quality is "good"; MOS = 4
scenario 2
Average packet loss is 3%: packet loss is bursty: the packet stream is decoded by an IP-phone with a poorly designed jitter buffer and crude error correction. Speech quality is "bad"; MOS = 1

2.3 Planning tools

It is important to note, and is explicitly stated in the standard, that planning tools such as the ITU G.107 E Model are unsuitable for real-time measurement.

Main Menu

Increasingly, and against ITU recommendations, the E-Model is being marketed to the industry as a live speech quality measurement tool for real-time network monitoring.

The ITU-T G.107 Recommendation states at the beginning of the document that,
“ Such estimates are only made for transmission planning purposes and not for actual customer opinion prediction (for which there is no agreed-upon model recommended by the ITU-T). “

When G.107 was written the ITU had yet to approve methods for customer opinion prediction, although it has now done so

- the intrusive method, P.862 PESQ, in February’02
- the non-intrusive method, P.SEAM, approved in March’04 as P.563.
- VoIP method, P.VTQ

3 QUALITY MEASUREMENT PRINCIPLES

3.1 Subjective testing

Subjective tests aim to find the average user’s perception of a system's speech quality by asking a panel of users a directed question and providing a limited choice of responses, e.g. ITU-T P.800 [4]. For example, to determine listening quality users are asked to rate “the quality of the speech” on a five-point scale:

5 Excellent 4 Good 3 Fair 2 Poor 1 Bad	Excellent	5
	Good	4
	Fair	3
	Poor	2
	Bad	1

A mean opinion score, MOS, is calculated for a particular condition by averaging the votes of all subjects. Subjective tests are time consuming to perform, must be very carefully designed and executed and are not a practical solution for most real-world measurement needs.

3.2 Objective measurement

Objective testing techniques measure physical properties of a system in order to predict perceived performance. There are two main classes of measurement: intrusive (active) and non-intrusive (passive). Such measurements are repeatable, efficient and fast.

3.2.1 Intrusive (active) testing

Intrusive techniques inject test signals into a system so they can be captured and assessed at a further point as shown below. To do this the system under test is taken out of service, although in terms of a telephone network this can be limited to a single telephone channel.

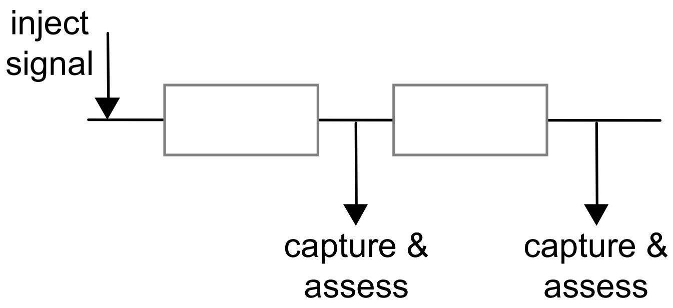

Figure 1, Illustration of an intrusive test

Intrusive assessment involves a comparison between the injected and captured signals. This method:

- enables isolation of the system under test,
- is capable of high accuracy
- can be used during development, commissioning and routine monitoring
- may incur a real cost, e.g. call charges when testing third-party networks
- is possible when no customers are on a system (during development before live traffic is present)
- allows control over external factors

An example intrusive measurement technique is ITU-T P.862 PESQ [5].

3.2.2 Non-intrusive (passive) testing

Non-intrusive techniques monitor live network traffic to determine the perceived quality, as shown below. In consequence network capacity is not lost to testing and the service provider is assessing the performance experienced by customers.

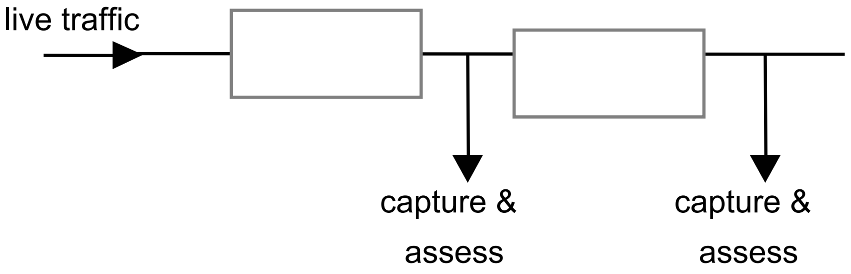

Figure 2, Illustration of a non-intrusive test

Non-intrusive techniques

- allow wider scale and denser testing
- do not use valuable network capacity for test calls
- do not require access to end points for test signal injection
- provide measurement data directly reflecting actual customer use and experience

However, these measurement techniques are complex to develop and are typically less accurate than intrusive techniques. Substantial advances have been made in the development of non-intrusive techniques for general network applications and for packet-networks.

There are two classes of non-intrusive measurement:

- Waveform – where the analysis is based on the transmitted waveform in order to predict the a persons perception of the speech quality, and
- Bearer – where an analysis of the packet bearer (transport) is made in order to predict the impact of the packet network on speech quality.

Two non-intrusive voice quality measurement algorithms, PSM (Psytechnics Speech Monitor – ITU-T P.561. P.562, P.563) and PSI (Psytechnics Speech for IP – ITU-T P.VTQ), represent the state-of-the-art for non-intrusive assessment and are presented below. These methods are currently ITU standards or the subject of work within the ITU for standardisation.

4 OBJECTIVE QUALITY MEASUREMENT METHODS

4.1 Intrusive method: PESQ ITU-T P.862

The world standard for intrusive assessment of end-to-end speech quality is the perceptual evaluation of speech quality (PESQ) model, ITU-T Recommendation P.862 [5].

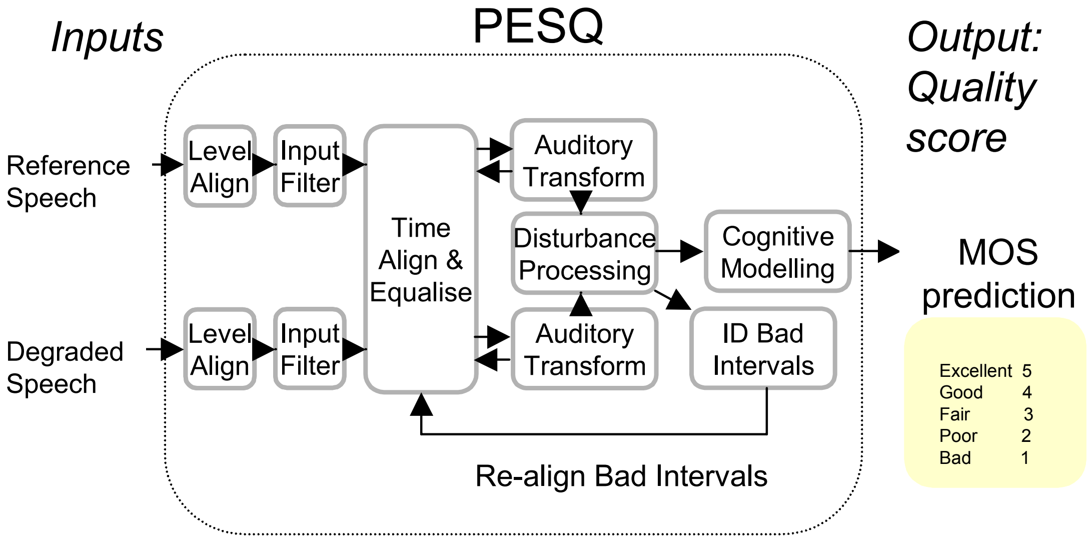

Figure 3, Functional block diagram of ITU-T P.862 PESQ

An overview of PESQ is shown above. 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 the telephone handset. The signals are aligned in time and are then processed through an auditory transform similar to that of PSQM [1]. Part of the transformation involves equalising the signals for the frequency response of the system and for gain variation. The difference between the transforms of the reference and degraded signals is known as the disturbance. This is processed to extract two distortion parameters, which are aggregated in frequency and time and mapped to a prediction of subjective MOS.

PESQ was extensively tested with “unknown” subjective test data during the competitive ITU selection process. It achieved remarkable performance with typical correlation between objective and subjective measurements of greater than 0.9 for a wide range of network types, codecs, background noise and real-world filtering characteristics [6,7].

4.2 Non-intrusive method: Waveform

PSM is a non-intrusive measurement algorithm that predicts the listening quality of voice channels by directly analysing the live voice signals being transported.

PSM seeks to identify whether a human vocal tract could produce particular speech sounds and speech sound sequences. The speech stream under assessment is reduced by an acoustic tube vocal tract model into a set of parameters that are sensitive to the type of distortion to be assessed. The parameters are chosen to ensure operation across a wide range of talker characteristics, including gender. Once parameterised, the data is used to generate a set of physiologically based rules for error identification. The resulting error identification is perceptually weighted and yields results that can be mapped to a listener's perception of speech quality [8].

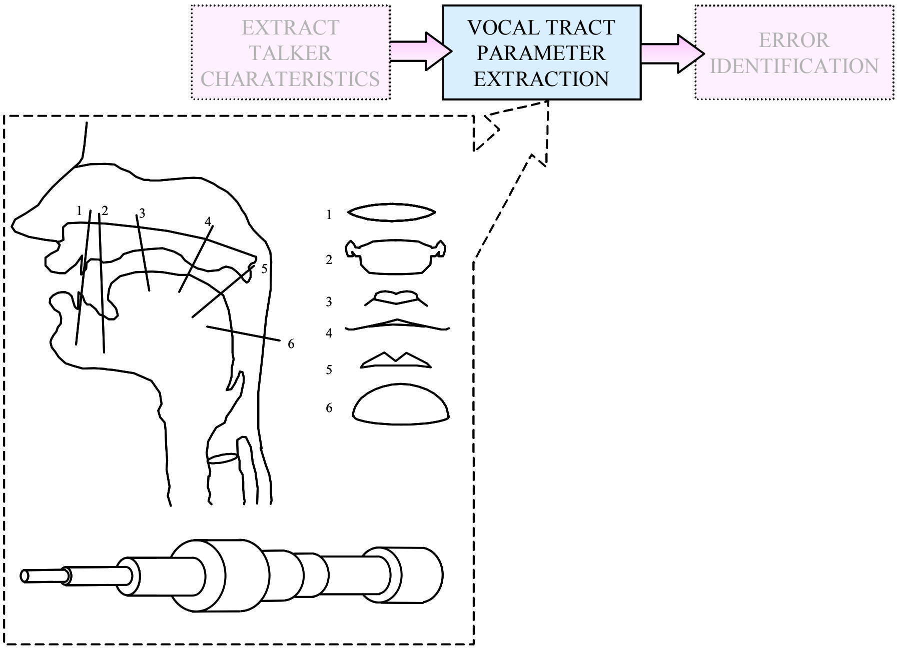
EXTRACT TALKER CHARATERISTICS
VOCAL TRACT PARAMETER EXTRACTION
ERROR IDENTIFICATION
1
2
3
4
5
6
1
2
3
4
5
6

Figure 4, Acoustic tube vocal tract model for waveform analysis

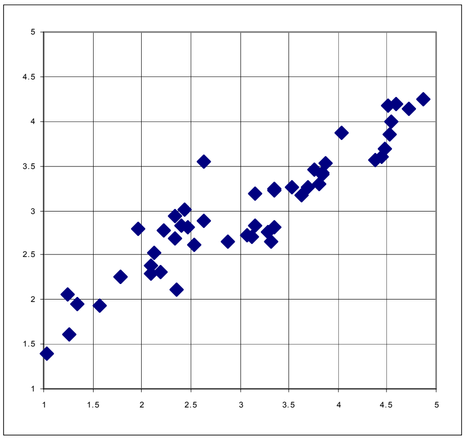

Figure 5, Example performance - subjective MOS vs. PSM MOS prediction for ETSI VoIP measurement test.

PSM has been selected as the core IPR for the new ITU-T P.SEAM standard, P.563

In addition to analyzing the waveform to predict listening quality, PSM can also include ITU P.562 call clarity index (CCI). CCI uses a number of link performance parameters, including level, echo and delay, to predict the conversational quality of a connection.

4.3 Non-intrusive method: IP-Bearer

Core 3G networks and NGNs (Next Generation Networks) are packet based, and hence methods that can accurately predict network performance from packet statistics are highly desirable – such a method has been developed, PSI, and has already been adopted by leading industry players. The PSI algorithm predicts the listening quality of a live VoIP call [9]. Simple quantities such as average packet loss are unable to predict MOS since a given average packet loss may yield good or bad speech quality depending on the distribution of the packet loss and choice of edge device, as illustrated in section 2.2.

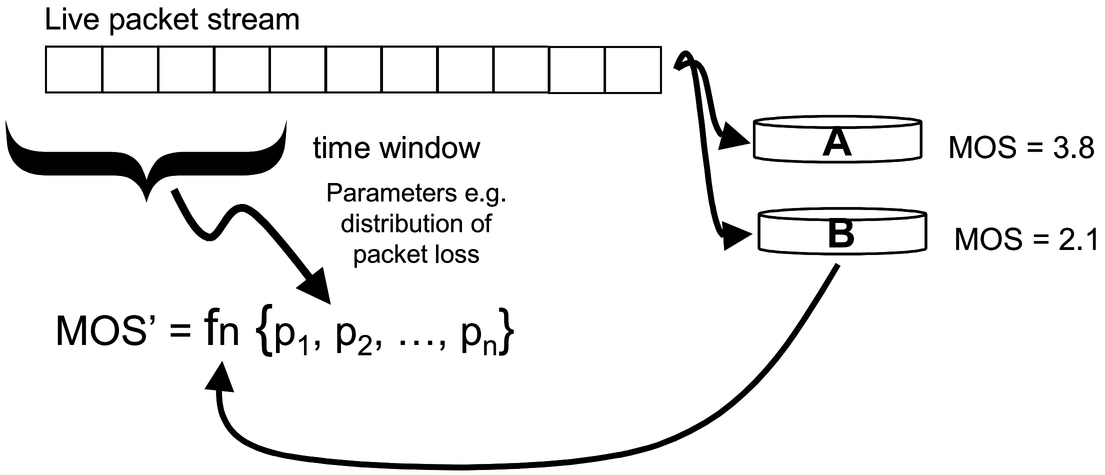
Live packet stream
time window
Parameters e.g. distribution of packet loss
A
MOS = 3.8
B
MOS = 2.1
MOS' = fn {p_1, p_2, …, p_n}

Figure 6, Principle of parametric MOS prediction for packet network

PSI is distinct from PSM in that it uses packet arrival times and information from the packet headers, rather than their payloads, to the predict the effect of an IP link on perceived quality. For example parameters are calculated that describe the distribution of packet loss over perceptually relevant intervals of time. While certain assumptions must be made about the voice signal being transported, for example its level, this approach has the advantage that it is very low complexity, and it can work with encrypted data streams, such as secure RTP. Another important feature is that knowledge of the downstream device, i.e. the gateway or IP-phone, can be used to calibrate measurements made at any point in the IP network to accurately predict speech quality from an IP packet stream (or even a post jitter-buffer packet stream).

The following graphs show performance data for two commercially available IP gateway products. Note the different x-axis ranges and that:

- the large dot is the average MOS performance
- the light dots are the range of MOS performance
- the graphs represent a packet loss variation of just 0 to 1%

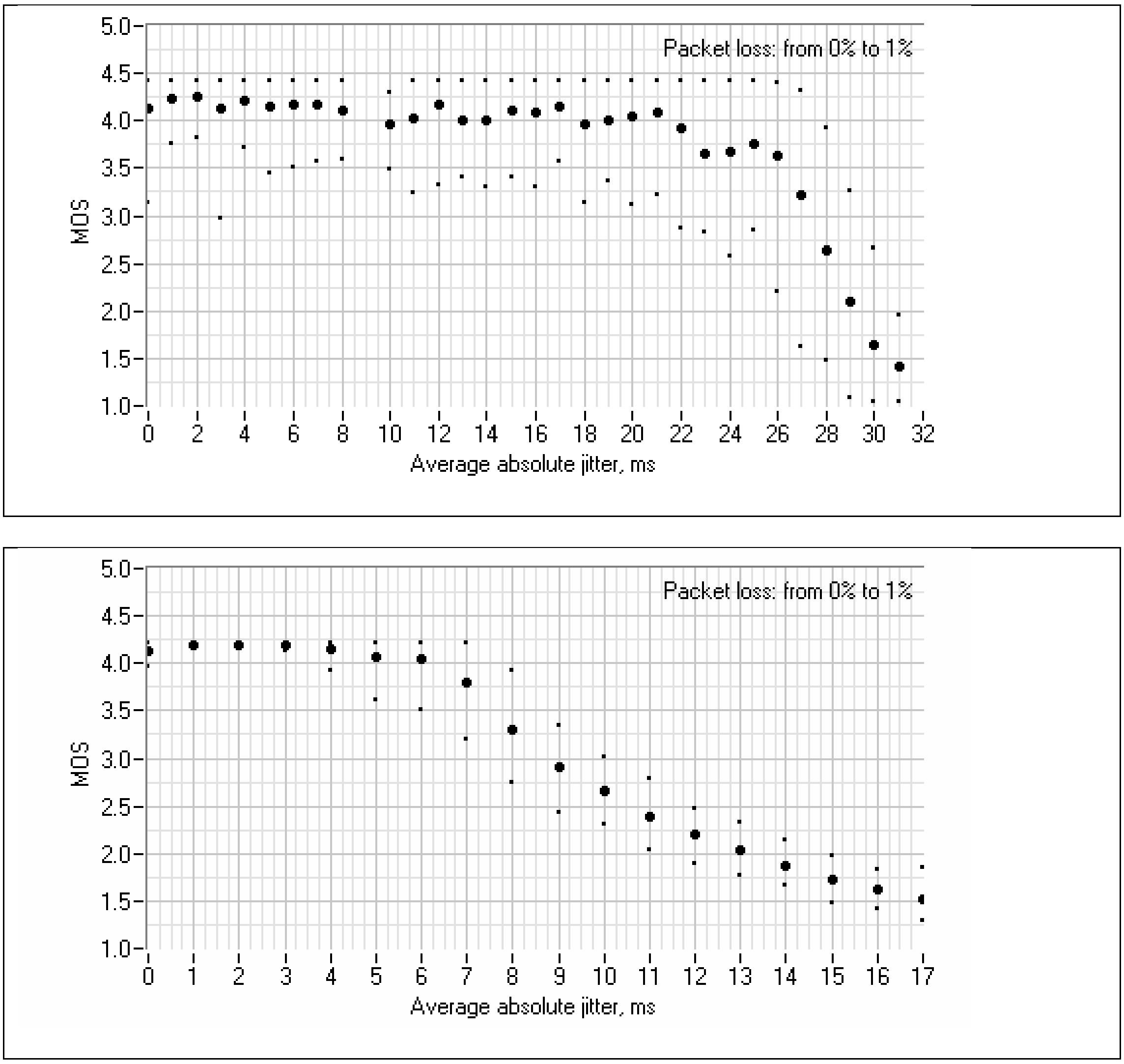

Figure 7, MOS performance of two commercially available VoIP gateways.

It is apparent that any prediction algorithm that is unable to take account pf the specific edge device cannot be as accurate as one that does. For example the MOS performance achieved by the two gateways with identical network behaviour varies by more than 2.0 MOS scores. Hence, consideration of the particular edge device is certainly required to predict the speech quality experienced by a particular end user.

When calibrated to a particular edge device PSI is capable of very high accuracy, as shown by the example results below. Further more this high calibrated accuracy also serves to prove that the underlying prediction parameters are highly sensitive to the MOS performance of the IP-Bearer. Hence the method can be used in confidence with a “reference calibration” to plot performance trends when the specific edge device is not known.

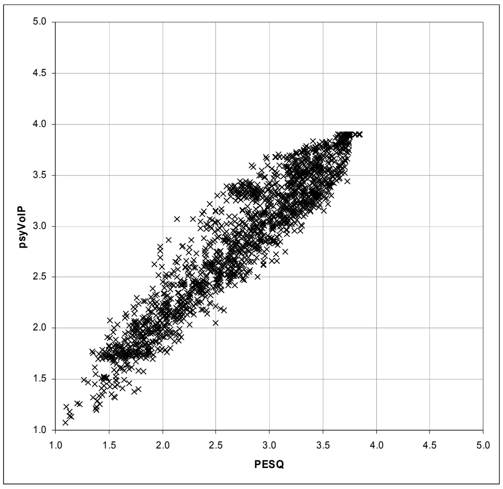

Figure 8, PESQ score vs. PSI prediction Per-condition (5 files per condition) scatter plot for Nortel ITG gateway calibration file (G.729 codec 20ms packet size) (1838 points)

PSI is currently under consideration in the ITU-T P.VTQ competition.

5 DISCUSSION

This paper explains why conventional engineering metrics have become inadequate for predicting the speech quality performance of digital communications systems. New methods based on models of human hearing and taking account of the subjectivity of errors have been developed in recent years. These methods are highly innovative and are now gaining wide acceptance in the industry, for example as new ITU-T standards. The two main classes of measurement were introduced and example methodologies presented with indicative results.

6 REFERENCES

1. J G Beerends, J A Stemerdink, “A perceptual audio quality measure based on a psychoacoustic sound representation”, Journal of the Audio Engineering Society, 10 (5), 963-978, June 1992.
2. 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.
3. 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.
4. *Methods for subjective determination of transmission quality,* ITU-T recommendation P.800, August 1996.
5. *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 Recommendation P.862, February 2001.
6. 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-delay compensation”, Journal of the Audio Engineering Society, 50 (10), 755-764, October 2002.
7. J G Beerends, A P Hekstra, A W Rix and M P Hollier. “Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II – Psychoacoustic model”, Journal of the Audio Engineering Society, 50 (10), 765-778, October 2002.
8. P Gray, M P Hollier, M E Massara, “Non-Intrusive Speech-Quality Assessment using Vocal-Tract Models”, IEE Proc.-Vis. Image Signal Process., Vol.147, No.6, Dec 2000..
9. A W Rix, S R Broom and R J B Reynolds. “Non-intrusive monitoring of speech quality in voice over IP networks”, ITU-T study group XII delayed contribution COM12-D049, October 2001.