P.563 a Breakthrough in Single Ended Voice Quality Testing

Dipl. Ing. Christian Schmidmer
OPTICOM, D-91058, Germany, Email: info@opticom.de

Introduction

Until recently voice quality was measured intrusive, which means that a known reference signal was injected into the system under test and the degraded output signal of the system under test was then compared to the injected reference signal, taking human perception into account. Such measurement methods are widely used today and standardized e.g. in ITU-T P.862 [1]. While these methods are very accurate and reliable, they always suffer from the disadvantage that a known reference signal is required. In some applications, like e.g. network monitoring systems, such a reference signal is simply not available. Other methods are therefore required which can assess the voice quality using the degraded signal only. The ITU recognized this need and selected a method for non-intrusive voice quality assessment after a competitive phase in 2003 under the working title P.SEAM (**S**ingle **E**nded **A**ssessment **M**ethod). The new method is expected to become draft recommendation P.563 [2] in March 2004. P.563 was jointly developed by OPTICOM, Psytechnics and Swissqual. This paper will outline the basic structure of the algorithm as well as its performance compared to subjective test results.

Overview of P.563

P.563 Starts by pre-processing the input signals. This pre-processing begins with a model of the receiving handset. Following it, a voice activity detector (VAD) is used to identify portions of the signal that contain speech and the active speech level is calculated. Finally, a speech level adjustment to -26 dBov is applied.

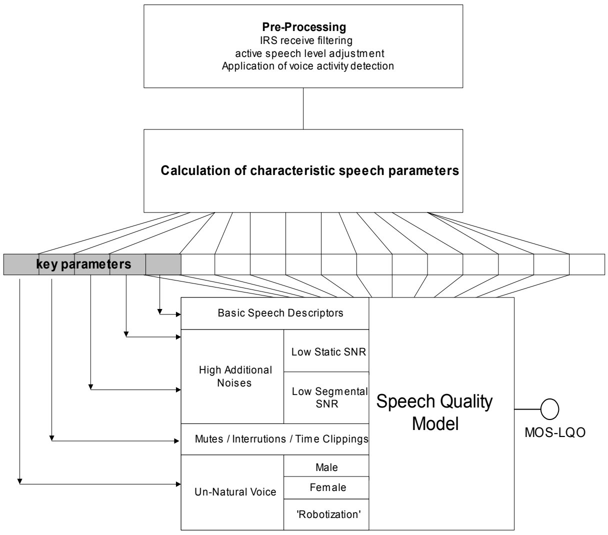

Figure 1: Block scheme of P.SEAM

The pre-processed speech signal will be assessed by several separate analysis blocks, which detect – like a sensor layer – a set of characterizing **signal parameters**. Based on a restricted set of **key parameters** an assignment to a **dominant distortion** class will be made.

The key parameters and the assigned distortion class are used for the adjustment of the speech quality model. This provides a perceptual based weighting where several distortions are occurring in one signal but one distortion class is more prominent than the others. The basic block-scheme of P.563 is shown in Figure 1.

Calculation of Characteristic Speech Parameters

The main block, *Calculation of Characteristic speech parameters* can be further subdivided into three main distortion type detectors, which are explained in the following. All three blocks produce a number of parameters that identify specific signal characteristics.

Vocal Tract Analysis and Unnaturalness of Speech

This functional block contains a speech production model for extracting signal parts that could be interpreted as voice and separates them from the non-speech parts. Furthermore, high order statistical analysis gives additional information about how human-like the speech is.

The unnaturalness of speech will be rated separately for male and female voices. Furthermore, in the case of strong robotization[1], another separate rating is made, which is gender-independent.

In this section the signal is investigated for the occurrence of tones like DTMF-tones or similar highly periodic signals that are not speech.

Other very annoying disturbances are repeated speech frames. In packet based transmission systems a typical error that can occur is the loss of packets. Some speech codecs employ error concealment methods in order to increase the received speech quality. In fact some error concealment methods use packet (frame) repetitions that simply replace a lost packet by, for example, a previously successfully transmitted packet, tend to decrease the quality of the signal rather than to increase it.

A more general description of the received speech quality is given by comparing the input signal with a pseudo reference signal generated by a speech enhancer. The speech enhancer is based on LPC analysis of the degraded signal, shaping the LPC coefficients and a subsequent inverse LPC. This

[1] Robotization is descibed by a voice signal that contains too much periodicity.

synthesized reference signal is then compared to the degraded signal using an algorithm known as PSQM99, which is also part of PESQ (ITU-T P.862) [1].

Analysis of strong additional noise

The noise analysis calculates different characteristics of noise. Based on two key parameters the decision will be made if additional noise is the main degradation. If additional noise is detected as the main degradation class a decision is made for the type of noise. Either it is static and present over all the signal (at least during speech activity) such that the noise power is not correlated with the speech signal, or the noise power shows dependencies on the signal power envelope.

If there was noise found that is likely to be static, several detectors try to quantify the amount of noise 'locally' and 'globally'. The expression 'local' noise as it is used here, describes the signal parts found especially between phonemes, whereas 'global' noise was defined as the signal between utterances such like sentences. Distinguishing between those noise types is important as, for example, in mobile communications often different settings for speech active parts and non-active parts are applied, e.g. comfort noise insertion.

Interruptions, mutes and time clipping

Mutes and Interruptions also form a separate distortion class. Such distortions can only partly be described by outcomes of the vocal tract investigation. Hence, a separate analysis is made to detect and to rate time clippings and unnatural mutes in the signal. Signal interruption can occur in two variants i.e. as temporal speech clipping or as speech interruption. Both lead to a loss of signal information. Temporal clipping may occur e.g., when voice activity detection is used or the signal becomes interrupted. This clipping is an annoying phenomenon that cuts off a bit of speech in the instant it takes for the transmitter to detect presence of speech. It is possible to detect the interruptions of the speech signal, which occur during the active speech intervals. The algorithms used in P.SEAM are able to distinguish between normal word endings and abnormal signal interruptions as well as unnatural silent intervals in a speech utterance.

Distortion classification

Some of the distortion detectors produce key parameters which proofed to classify certain types of distortions. These key parameters are prioritized and define the dominant distortion type. Depending on the dominant distortion, a subset of the characteristic speech parameters is selected and the linear combination of these is used to form the final result.

Result

The result of P.563 is a number indicating the listening quality on a MOS scale between 1 and 5 (MOS-LQO) and it is directly comparable the results of subjective ACR experiments.

Performance of P.563

Table 1 gives an overview of the accuracy that can be expected by P.563. The correlation coefficient for a number of subjective speech databases is given after calculating the condition MOS and a 3'rd order polynomial mapping per database. The column *Required* indicates the minimum requirement as it was defined by ITU-T SG12 for the standardisation of the model. A part of these databases was already used for the standardisation of P.862. Interested parties can register for a field trial with live measurements at www.3sqm.com/onlinetest (popup windows must be enabled).

Test	Required	3rd order corr. coeff per cond.	diff to Req.
ITU Supp23, 1A	0.80	**0.885**	+ 0.085
ITU Supp23, 1D	0.80	**0.842**	+ 0.042
ITU Supp23, 1O	0.80	**0.902**	+ 0.102
ITU Supp23, 3A	0.80	**0.867**	+ 0.067
ITU Supp23, 3C	0.80	**0.854**	+ 0.054
ITU Supp23, 3D	0.80	**0.929**	+ 0.129
ITU Supp23, 3O	0.80	**0.917**	+ 0.117
P.862_BT_1st	0.80	**0.851**	+ 0.051
P.862_dtag_1st	0.80	**0.846**	+ 0.046
P.862_ascom_1	0.80	**0.903**	+ 0.103
P.862_ascom_2	0.80	**0.876**	+ 0.076
P.862_dtag_share	0.80	**0.872**	+ 0.072
P.862_KPN_1st	0.80	**0.813**	+ 0.013
P.862_bgn_e	0.80	**0.889**	+ 0.089
P.862_bgn_g	0.80	**0.888**	+ 0.088
P.862_g_etsi_voip	0.80	**0.961**	+ 0.161
P.862_net_emul_d	0.75	**0.821**	+ 0.071
P.862_net_emul_e	0.75	**0.806**	+ 0.056
P.862_net_meas_d	0.80	**0.812**	+ 0.012
P.SEAM_Swissqual	0.80	**0.893**	+ 0.093
P.SEAM_Opticom	0.80	**0.921**	+ 0.121
P.SEAM_Psytechnics	0.80	**0.932**	+ 0.132
P.SEAM_Lucent	0.80	**0.934**	+ 0.134
P.SEAM_FT	0.80	**0.847**	+ 0.047
Average		**0.8776**	**+ 0.0817**

Table 1: Correlation coefficients between MOS and P.SEAM, per condition, after monotonic 3rd order polynomial mapping

[1] ITU-T Recommendation P.862, PESQ an objective method for end-to-end voice quality assessment of narrowband telephone networks and speech codecs, February 2001

[2] ITU-T Delayed Contribution COM12-D183-E, Draft new Recommendation P.SEAM (for consent). OPTICOM, Psytechnics, Swissqual, March 2004