Advances in Perceptual Speech Quality Assessment

Abdulhussain E. Mahdi

Department of Electronic and Computer Engineering
University of Limerick, Limerick, Ireland
hussain.mahdi@iul.ie

Abstract

In the context of telecommunications, speech quality is the most visible and important aspects of quality of service (QoS), and the ability to monitor and design for this quality should be a top priority. Speech quality refers to the clearness of a speaker's voice as perceived by a listener. Its measurement offers a means of adding the human end-user's perspective to traditional ways of performing network management evaluation of voice telephony services. Traditionally, measurement of users' perception of speech quality has been performed by expensive and time-consuming subjective listening tests. Over the last three decades, numerous attempts have been made to supplement subjective tests with objective measurements based on algorithms that can be computerised and automated. This paper describes the technicalities associated with speech quality measurement, and presents a review of current subjective and objective speech quality evaluation methods and standards in telecommunications.

1. Introduction

For telecommunication networks, the quality of the communicated speech is one of the most important measuring objects of QoS. Thus, the ability to continuously monitor and design for this quality should always be a top priority to maintain customers' satisfaction of quality. Speech quality, commonly known as voice quality (which is the term used throughout this paper), refers to the clearness of a speaker's voice as perceived by a listener. Voice quality measurement, also known by the acronym VQM, is a relatively new discipline which offers a means of adding the human, end-user's perspective to traditional ways of performing network management evaluation of voice telephony services. The most reliable method for obtaining true measurement of users' perception of speech quality is to perform properly designed subjective listening tests. In a typical listening test, subjects hear speech recordings processed through different network conditions, and rate them using a simple opinion scale such as the ITU-T (International Telecommunication Union-Telecommunication Standardization Sector) 5-point listening quality scale. The average score of all the ratings registered by the subjects for a condition is termed the *Mean Opinion Score* (MOS).

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 of live networks. As an alternative, numerous objective voice quality measures, which provide automatic assessment of voice communication systems without the need for human listeners, have been made available over the last two decades. These objective measures, which are based on mathematical models and can be easily computerised, are becoming widely used particularly to supplement subjective test results. This paper examines some of the technicalities associated with VQM and presents a review of current voice quality measurement methods for telecommunication applications. Following this Introduction, Section 2 discusses what voice quality is and how to measure it. Sections 3 and 4 define the two main categories of metrics used for evaluating voice quality; that is subjective and objective testing, describing and reviewing the various methods and procedures of both, as well as indicating these methods' target applications and their advantages/disadvantages. Section 5 discusses the various approaches employed for non-intrusive measurement of voice quality as required for monitoring live networks, and provides an up-to-date review of developments in the field. Finally, the Conclusions Section gives a summary of the presented review.

2. Voice quality and its measurement in telecommunications

In telecommunications, QoS is thought to be divided into three components [1]. The main component is the speech or voice communication quality, and relates to a bi/multi-directional conversation over the telecommunications network. The second component is the service-related influences, which is commonly referred to as the 'service performance', and includes service support, a part of service operability and service security. The third component of the QoS is the necessary terminal equipment performance. Voice communication quality represents a major component of the overall communication quality perceived by a user and is concerned with the speech transmission from a talker to a listener [2]. Thus, it is user-directed and, therefore, provides close insight in the question of which quality feature results in an acceptability of the service from the user's viewpoint.

Quality can be defined as the result of the judgment of a perceived constitution of an entity with regard to its desired constitution. The perceived constitution contains the totality of the features of an entity. For the perceiving person it is a characteristic of the identity of the entity [1]. Applying this definition to speech, quality can be regarded as the result of a perception and assessment process, during which the assessing subject establishes a relationship between the perceived and the desired or expected speech signal. In other words, voice quality can be defined as the result of the subject's judgment on spoken language, which he/she perceives in a specific situation and judges instantaneously according to his/her experience, motivation and expectation. Regarding voice communication systems, quality is the customer's perception of a service or product, and voice quality measurement (VQM) is a means of measuring customer experience of voice telephony services. The most accurate method of measuring voice quality therefore would

be to actually ask the callers. Ideally, during the course of a call, customers would be interrupted and asked for their opinion on the quality. However, this is obviously not practical. In practice, there are two broad classes of voice quality metrics: *subjective* and *objective*. Subjective measures, known as subjective tests, are conducted by using a panel of people to assess the voice quality of live or recorded speech signals from the voice communication system under test for various adverse distortion conditions. Here, the speech quality is expressed in terms of various forms of a mean opinion score (MOS), which is the average quality perceived by the members of the panel. Objective measures, on the other hand, replace the human panel by an algorithm that compute a MOS value using a sample of the speech in question. Detailed descriptions of both types of methods will be described in the proceeding sections.

Subjective tests can be used to gather firsthand evidence about perceived voice quality, but are often very expensive, time-consuming and labour-intensive. There are many situations, however, where the costs associated with formal subjective tests do not seem to be justified. Examples of these situations are the various design and development stages of algorithms and devices, and the continuous monitoring of telecommunications networks. Hence, an instrumental (non-auditive) method for evaluation of perceived quality of speech is in high demand. Such methods, which have been of great interest to researchers and engineers for a long time, are referred to *as Objective Speech/Voice Quality Measures* [1]. The underlying principle of objective voice quality measurement is to predict voice communication/transmission quality based on objective metrics of physical parameters and properties of the speech signal. Once automated, objective methods enable standards to be efficiently maintained together with effective assessment of systems and networks during design, commissioning, and operation.

A voice communication system can be regarded as a distortion module. The source of the distortion can be background noise, speech codecs, and channel impairments such as bit errors and frame loss. In this context, most current objective voice quality evaluation methods are based on comparative measurement of the distortion between the original and distorted speech. Several objective voice quality measures have been proposed and used for the assessment of speech coding devices as well as voice communication systems. Over the last three decades, numerous different measures based on various perceptual speech analysis models have developed. Most of these measures are based on an input-to-output or intrusive approach, whereby the voice quality is estimated by measuring the distortion between an "input" or a reference speech signal and an "output" or distorted speech signal. Current examples of intrusive voice quality measures include the Bark Spectral Distortion (BSD), Perceptual Speech Quality (PSQM), Modified BSD, Measuring Normalizing Blocks (MNB), PSQM+, Perceptual Analysis Measurement Systems (PAMS) and most recently the Perceptual Evaluation of Speech Quality (PESQ) [3, 4]. In 2004, the ITU-T approved a new non-intrusive voice quality assessment algorithm under its Rec. P.563: "Single ended method for objective speech quality assessment in narrow-band telephony applications" [5].

3. Subjective assessment of voice quality

Voice quality measures that are based on ratings by human listeners are called subjective tests. These tests seek to quantify the range of opinions that listeners express when they hear speech transmission of systems that are under test. There are several methods to assess the subjective quality of speech signals. In general, they are divided in two main classes: a) conversational tests and b) listening-only tests. Conversational tests, whereby two subjects have to listen and talk interactively via the transmission system under test, provide a more realistic test environment. However, they are rather involved, much more time consuming, and often suffer from low reproducibility [1]. Thus listening-only tests are often recommended. Although listening-only tests are not expected to reach the same standard of realism as conversational tests and their restrictions are less severe in some respect, the artificiality associated with them brings with it a strict control of many factors, which in conversational tests are allowed to their own equilibrium.

In subjective testing, speech materials are played to a panel of listeners, who are asked to rate the passage just heard, normally using a 5-point quality scale. All subjective methods involve the use of large numbers of human listeners to produce statistically valid subjective quality indicator. The indicator is usually expressed as a *Mean Opinion Score* (MOS), which is the average value of all the rating scores registered by the subjects. For telecommunications purposes, the most commonly used assessment methods are those standardised and recommended by the ITU-T, and include the *Absolute Category Rating*, *Quantal-Response Detectability*, *Degradation Category Rating* and the *Comparison Category Rating [*6]. Among these methods, the most popular ones are the *Absolute Category Rating (*ACR) and the *Degradation Category Rating* (DCR). In the ACR, listeners are required to make a single rating for each speech passage using a listening quality scale using the 5-point category-judgment scale shown in Fig.1. The rating are then gathered and averaged to yield a final score known as the mean opinion score, or MOS. The test introduced by this method is well established and has been applied to analogue and digital telephone connections and telecommunications devices, such as digital codecs. If the voice quality were to drop during a telephone call by one MOS, an average user would clearly hear the difference. A drop of half a MOS is audible, whereas a drop of a quarter of a point is just noticeable [7]. A typical public switched telephony network (PSTN) would have a MOS of 4.3. DCR involves listeners presented with the original speech signal as a reference, before they listen to the processed (degraded/distorted) signal, and are asked to compare the two and give a rating according to the amount of degradation perceived.

In May 2003, ITU-T approved Rec. P800.1 [8] that provides a terminology to be used in conjunction with voice quality expressions in terms of MOS. This new terminology is motivated by the intention to avoid misinterpretation as to whether specific values of MOS are related to listening quality or conversational quality, and whether they originate from subjective tests, from objective models or from network planning models. Accordingly, the following identifiers are recommended to be used together with the abbreviation MOS in order to distinguish the area of application: LQ to refer to *Listening Quality*, CQ to refer to *Conversational Quality*, S to

refer to *Subjective* testing, O to refer to *Objective* testing using an objective model, and E to refer to *Estimated* using a network planning model.

Figure 1 *The ITU-T listening quality scale.*

4. Objective voice quality measures

Objective voice quality metrics replace the human panel by a computational model or an algorithm that compute a MOS value by observing a sample of the speech in question [2]. The aim of objective measures is to predict MOS values that are as close as possible to the ratings obtained from subjective tests for various adverse speech distortion conditions. The accuracy and effectiveness of an objective metric is, therefore, determined by its correlation, usually the Pearson correlation, with the subjective MOS scores. If an objective measure has a high correlation, typically >0.8, it is deemed to be effective measure of perceived voice quality, at least for the speech data and transmission systems with the same characteristics as those in the test experiment [9].

Starting from late 1970, researchers and engineers in the field of objective measures of speech/voice quality have developed different objective measures based on various speech analysis models. Based on the measurement approach, objective measures are classified into two classes: intrusive and non-intrusive, as illustrated in Fig. 2. Intrusive measures, often referred to as input-to-output measures, base their measurement on computation of the distortion between the original (clean or input) speech signal and the degraded (distorted or output) speech signal. Non-intrusive measures (also known as output-based or single-ended measures), on the other hand, use only the degraded signal and have no access to the original speech signal.

4.1. Intrusive objective voice quality measures

Although there are different types of intrusive (or input-to output) objective voice quality measures, they all share a similar measurement structure that involves two main processes, as shown in Fig. 3. As shown, the first process involves pre-processing of the speech signal and extraction of relevant speech parameters. Here, the original (input) speech signal and the signal degraded by the system under test, i.e. the output signal, are transformed into a relevant domain such as temporal, spectral or perceptual domain. The second process involves a distance measure, whereby the distortion between the input and output speech signals is computed using an appropriate quantitative measure.

Depending on the domain transformation used, objective measures are often classified into: *Time Domain Measures*, *Spectral Domain Measures* and *Perceptual Domain Measures*. Time domain measures are generally applicable to analogue or waveform coding systems in which the target is to reproduce the waveform. *Signal-to-Noise Ratio* (SNR) and *Segmental SNR* (SNRseg) are typical time domain measures [2]. Spectral domain measures are more credible than time-domain measures as they are less susceptible to the occurrence of time misalignments and phase shift between the original and the distorted signals. Several spectral domain measures have been proposed in the literature, including the *log likelihood ratio*, *Itakura-Saito distortion measure* and the *cepstral distance measure* [10, 11]. However, as most of the spectral domain measures use the parameters of speech production models used in codecs, their performance is usually limited by the constraints of those models. In contrast to the spectral domain measures, perceptual domain measures are based on models of human auditory perception and, hence, have the best potential of predicting subjective quality of speech. In these measures, speech signals are transformed into a perception-based domain using concepts of the psychophysics of hearing, such as the critical-band spectral resolution, frequency selectivity, the equal-loudness curve and the intensity-loudness power law to derive an estimate of the auditory spectrum [12]. In principle, perceptually relevant information is both sufficient and necessary for a precise assessment of perceived speech quality. The perceived quality of the coded speech will, therefore, be independent of the type of coding and transmission, when estimated by a distance measure between perceptually transformed speech signals. The following sections give descriptions of some of the recent and commonly used perceptual voice quality measures.

4.1.1. Bark Spectral Distortion measure (BSD)

The Bark spectral distortion (BSD) measure was developed by Wang and co-workers [9] as a method for calculating an objective measure for signal distortion based on the quantifiable properties of auditory perception. The overall BSD measurement represents the average squared Euclidean distance between spectral vectors of the original and coded utterances. The main aim of the measure is to emulate several known features of perceptual processing of speech sounds by the human ear, especially frequency scale warping, as modelled by the Bark transformation, and critical band integration in the cochlea; changing sensitivity of the ear as the frequency varies; and difference between the loudness level and the subjective loudness scale.

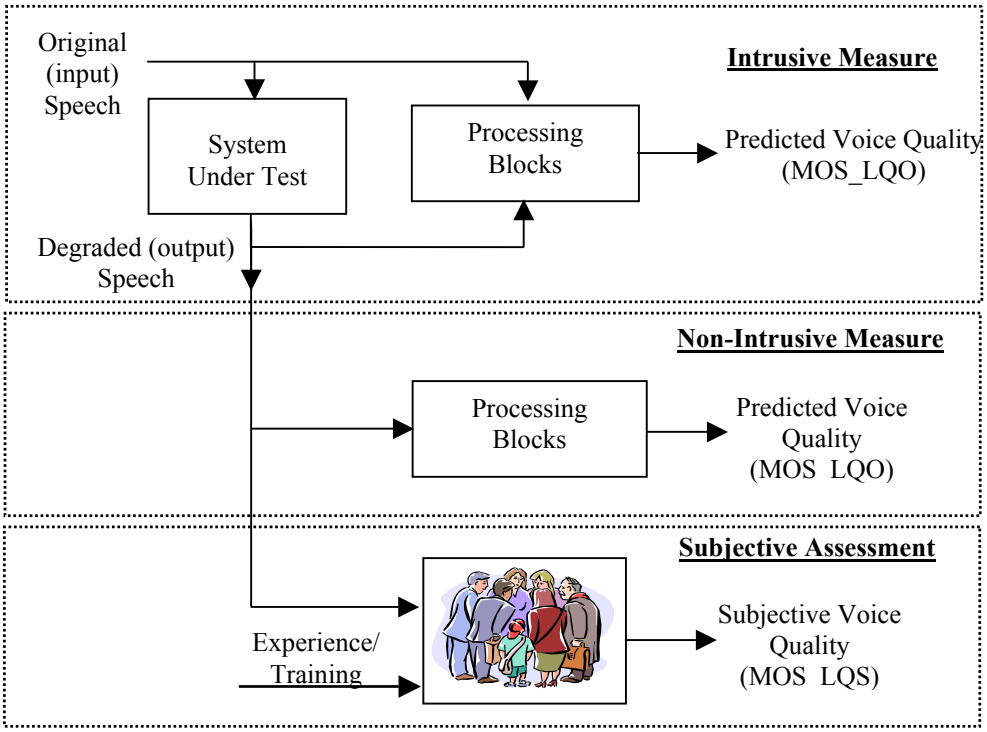

Figure 2 *Intrusive and non-intrusive voice quality measures.*

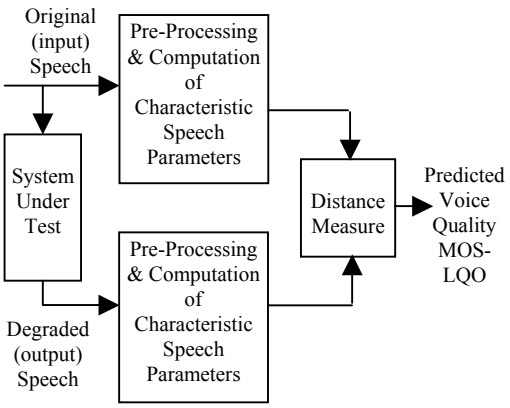

Figure 3 *Basic structure of an intrusive (input-to output) objective voice quality measure.*

4.1.2. Perceptual Speech Quality Measurement (PSQM)

To address the continuous need for an accurate objective measure, Beerends and Stemerdink from KPN Research - Netherlands, developed a voice quality measure which takes into account the clarity's subjective nature and human perception. The measure is called the Perceptual Speech Quality Measurement or PSQM [13]. In 1996, the PSQM was approved by ITU-T and published Rec. P.861 [22]. The PSQM, as shown in Fig. 4 is as a mathematical process that provides an accurate objective measurement of the subjective voice quality. The main objective of PSQM is to produce scores that reliably predict the results of the recommended ITU-T subjective tests. PSQM is designed to be applied to telephone band signals (300-3400 Hz) processed by low bit-rate voice compression codecs and vocoders.

To perform a PSQM measurement, a sample of recorded speech is fed into a speech encoding/decoding system and processed by whatever communication system is used. Recorded as it is received, the output signal (test) is then time-synchronised with the input signal (reference). Following the time-synchronisation the PSQM algorithm will compare the test and reference signals. This comparison is performed on individual time segments acting on parameters derived from spectral power densities of the input and output time-frequency components. The comparison is based on factors of human perception, such as frequency and loudness sensitivities, rather than on simple spectral power densities. The resulting PSQM score representing a perceptual distance between the test and reference signals can vary from 0 to infinity. As an example, 0 score suggests a perfect correlation between the input and output signals, which most of the time is classified as perfect clarity. Higher scores indicate increasing levels of distortion, often interpreted as lower clarity. In practice upper limits of PSQM scores range from 15 to 20. At the final stage, the PSQM scale is mapped from its objective scale to the 1-5 subjective MOS scale. One of the main drawbacks of this measure is that it does not accurately report the impact of distortion caused by packet loss or other types of time clipping. In other words, human listeners

reported higher speech quality score than PSQM measurements for such errors.

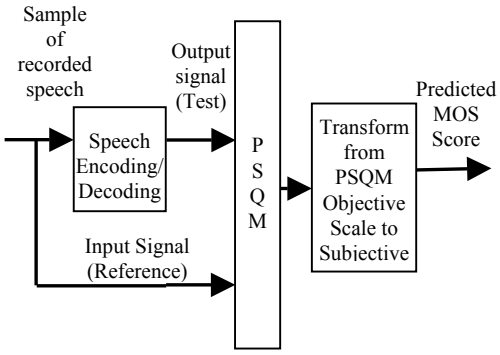

Figure 4 *PSQM testing process.*

4.1.3. Perceptual Speech Quality Measurement Plus (PSQM+)

Taking into account the drawbacks of the PSQM, Beerends, Meijer and Hekstra developed an improved version of the conventional PSQM measure. The new model, which became known as PSQM+, was reviewed by ITU-T Study Group 12 and published in 1997 under COM 12-20-E [14]. PSQM+. It is based directly on the PSQM model, represents an improved method for measuring voice quality in network environments. For systems comprising speech encoding only both methods give identical scores. PSQM+ technique, however, is designed for systems which experience severe distortions due to time clipping and packet loss. When a large distortion, such as time clipping or packet loss is introduced (causing the original PSQM algorithm to scale down its score), the PSQM+ algorithm applies a different scaling factor that has an opposite effect, and hence produces higher scores that correlate better with subjective MOS than the PSQM.

4.1.4. Perceptual Evaluation of Speech Quality (PESQ)

In May 2000, a collaborative draft from KPN Research and British Telecommunications was submitted to ITU-T describing a new measurement technique called Perceptual Evaluation of Speech Quality (PESQ). In February 2001, ITU-T approved the PESQ under Rec. P.862 [4]. PESQ is directed at narrowband telephone signals and is effective for measuring the impact of the following conditions: waveform and non waveform codecs, transcodings, speech input levels to codecs, transmission channel errors, noise added by system (not present in input signal), and short and long term warping.

The PESQ is designed for use with intrusive tests: a signal is injected into the system under test, and the distorted output is compared with the input (reference) signal. The difference is then analysed and converted into a quality score. As a result of this process, the predicted MOS as given by PESQ varies between 0.5, which corresponds to a bad distortion condition, and 4.5 which corresponds to no measurable distortion. The PESQ model, which is shown in Fig.5, was developed to accurately estimate the listening speech quality performed by wireless, VoIP and fixed networks. It can be used in a wide range of measurement applications, such as codecs development, equipment optimisation and regular network monitoring. Being fast and repeatable, PESQ makes it possible to perform extensive testing over a period of only few days, and also enables the quality of time-varying conditions to be monitored.

In order to align with the new MOS terminology, a new ITU-T Recommendation, Rec. P.862.1 [15] was published. This Recommendation defines a mapping function and its performance for a single mapping from raw P.862 scores to the MOS LQO (Rec. P.800.1). In 2005, the ITU-T issued Rec. P.862.2 [16], which describes a simple extension to the PESQ algorithm defined in P.862. The P.862.2 is mainly intended for use with wideband audio systems (50 – 7000 Hz). The wideband extension describes two main additions to the P.862: (a) the replacement of the input filter applied to both the reference and degraded speech signals by an IIR filter with a flatter response above 100 Hz and a gentle roll-off below this point, modelling the attenuation of the headphones and ear at low frequencies, and (b) a new mapping function is defined to be used with wideband applications. In addition to the above extensions, the ITU-T also issued Rec. P.862.3, which is an application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2 [17].

5. Non-intrusive objective voice quality measures

All objective measures presented in Section 4.1 are based on an input-to-output approach and, hence, classified as intrusive. Intrusive voice quality measures have few drawbacks. Firstly, in all these measures the time-alignment between the input and output speech vectors, which is achieved by automatic synchronisation, is a crucial factor in deciding the accuracy of the measure. In practice, perfect synchronisation is difficult to achieve due to fading or error burst that are common in wireless systems, and hence degradation in the performance of the measure is inevitable. Secondly, there are many applications where the original speech is not available, as in cases of wireless and satellite communications. In most situations it is not always possible to have access to both ends of a network connection to perform speech quality measurement using an input-to-output method.

Intrusive voice quality measures are more accurate, but normally are unsuitable for monitoring real-time traffic in live networks. An objective measure which can predict the quality of the transmitted speech using only the output (or degraded) speech signal, i.e. one end of the network, would therefore cure all the above problems and provide a convenient non-intrusive measure for monitoring of live networks. Ideally what is required for a non-intrusive objective voice quality measure is to be able to assess the quality of the distorted speech by simply observing a sample of the speech in question with no access to the original speech. However, due to non-availability of the original (or input) speech signal such a measure is very difficult to realise.

Over the last decade, a number of non-intrusive voice quality measures based on either vocal tract modelling [18], statistical models [19, 20], or comparison between signal under test and an artificial matching reference extracted from appropriately formulated speech codebook [21], have been

reported. Currently, however, the only standard non-intrusive voice quality algorithm is the ITU-T P.563 [5].

5.1. ITU -T P.563

In 2004, the ITU-T approved a new non-intrusive voice quality assessment algorithm under its Rec. P.563: "Single ended method for objective speech quality assessment in narrow-band telephony applications" [5]. The algorithm represents the first internationally recognised method for single-ended non-intrusive voice quality measurement applications that takes into account the full range of distortions occurring in public switched telephony networks (PSTN), and able to predict the voice quality on a perception-based scale MOS–LQO according to ITU-T Rec. P.800.1.

The basic block diagram of P.563 is shown in Fig. 6 (adopted from [5]).The P.563 approach could be visualized as an expert who is listening to a real call with a test device like a conventional handset into the line in parallel. The quality score predicted by P.563 is related to the perceived quality by linking a conventional handset at the measuring point. Hence, the listening device has to be part of the P.563 approach. To achieve this, each speech signal to be assessed is first pre-processed, beginning with the model of the receiving handset. This is followed by a voice activity detector (VAD) to identify portions of the signal that contain speech and the speech level is calculated and adjusted. The pre-processed speech signal is then investigated by several separate analyses to detect a set of characterising signal parameters. The signal parameterisation is divided into three independent functional blocks corresponding to three main classes of distortion, namely: vocal tract analysis and unnaturalness of speech; analysis of strong additional noise; and speech interruptions, mutes and time clipping. Accordingly, a total of 51 characteristic signal parameters are computed, and a dominant distortion class based on restricted set of 8 key parameters is selected. The key parameters and the selected distortion class are then used to adjust the speech quality model. In addition, a linear combination of parameters for each distortion class is used to produce an intermediate quality rating, which together with the other signal features is combined to compute the raw objective quality score.

ITU-T recommended that P.563 be used for voice quality measurement in narrow-band telephony applications only, under the following scenarios [5]: Live network monitoring using digital or analogue connection to the network; Live network end-to-end testing using digital or analogue connection to the network; and Live network end-to-end testing with unknown speech sources at the far end side. Target coding technologies of P.563 are: waveform codecs, such as G.711,G.726 and G.727; CELP and hybrid codecs at bit rates $\geq$ 4 kbit/s, such as G.728, G.729 and G.723.1; as well as other codecs, such as GSM-FR, GSM-HR, GSM-EFR, GSM-AMR, CDMA-EVRC, TDMA ACELP, TDMA-VSELP, TETRA [5]. P. 563 is, however, known to provide inaccurate predictions when used in conjunction with the following variables/technologies: listening levels, loudness loss, sidetone, effect of delay in conversational tests, talker echo and music or network tones as input signal, and LPC vocoder technologies at bit rates < 4.0 kbit/s, such as IMBE, AMBE, LPC10e [5].

As the case with PESQ, the ITU-T emphasises that the P.563 algorithm cannot be used to replace subjective testing but it can be applied for measurements where auditory tests would be too expensive or not applicable at all.

6. Conclusions

In this paper, we have presented a detailed review of currently used metrics and methods for measuring user's perception of the voice quality of telephony systems. Descriptions of various internationally standardised subjective tests that are based on ratings by humans were presented, with particular emphasis on those approved by the ITU-T. Limitations of subjective testing were then discussed, paving the ground for a comprehensive review of various objective voice quality measures highlighting in a comparative manner their historical evolution, target applications and performance limitations. In particular, two main categories of objective voice quality measures were described: intrusive or input-to-output measures and non-intrusive or single-ended measures, providing an insight into advantages/disadvantages of each.

7. References

[1] Moller, S, Assessment and Prediction of Speech Quality in Telecommunications, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[2] Quackenbush, S. R., Barnawell, T. P. and Clements, M. A., Objective Measures of Speech Quality, Prentice Hall, NJ, USA, 1988.

[3] Anderson, J., Methods for Measuring Perceptual Speech Quality - White Paper, Agilent Technologies, USA, 2001 (available at http://www.agilent.com).

[4] ITU-T Rec. P.862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, ITU-T, Geneva, Switzerland, 2001.

[5] ITU-T Rec. P.563, Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications, ITU-T, Geneva, Switzerland, 2004.

[6] ITU-T Rec. P.800, Methods for Subjective Determination of Transmission Quality, ITU-T, Geneva, Switzerland, 1996.

[7] Psytechnics, Mobile Quality Survey, Case Study Report, Psytechnics Ltd., Ipswich, UK, 2003 (available at http://www.psytechnics.com/psy_frm01.html).

[8] ITU-T Rec. P.800.1, Mean Opinion Score (MOS) Terminology, ITU-T, Geneva, Switzerland, 2003.

[9] Wang, S., Sekey, A., Gersho, A., An objective measure for predicting subjective quality of speech coders, IEEE J. Select. Areas Comm., 10 (5): 819-829, 1992.

[10] Itakura, F. and Saito, S., Analysis synthesis telephony based on the maximum likelihood method, in Proc. 6th Int. Congr. Acoust. Tokyo, Japan, 1978, pp. C17 -C-20.

[11] Kitawaki, N., Nagabuchi, H., Itoh, K, Objective quality evaluation for low-bit-rate speech coding systems, IEEE J. Select. Areas Comm., 6 (2): 242-248, 1988.

[12] Quatieri, T. E., Discrete-Time Speech Signal Processing: Principles and Practice, Prentice Hall, NJ, USA, 2002.

[13] Beerends, J. G. and Stemerdink, J. A., A perceptual speech quality measure based on a psychoacoustic sound representation, J. Audio Eng. Soc., 42 (3): 115-123, 1994.

[14] Beerends, J. G., Meijer, E. J. and Hekstra, A. P., Improvement of the P. 861 Perceptual Speech Quality

Measure: Contribution to COM 12-20, ITU-T Study Group 12, Geneva, Switzerland, 1997.
[15] ITU-T. Rec. P.862.1, Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO, ITU-T, Geneva, Switzerland, 2003.
[16] ITU-T Rec. P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs, ITU-T, Geneva, Switzerland, 2005.
[17] ITU-T Rec. P.862.3, Application Guide for Objective Quality Measurement Based on Recommendations P.862, P.862.1 and P. 862.2, ITU-T, Geneva, Switzerland, 2005.
[18] Gray, P., Hollier, M. P. and Massara, R. E., Non-intrusive speech quality assessment using vocal-tract models, IEEE Proc.-Vis. Image Signal Process., 147 (6): 493-501, 2000.
[19] Chen, G. and Parsa, V., Bayesian model based non-intrusive speech quality evaluation, in Proc. of IEEE Intl. Conf. Acoustics, Speech, and Signal Process., ICASSP, PA, USA, March 2005, pp. I-385-388.
[20] Kim, D.-S., ANIQUE: An auditory model for single-ended speech quality estimation, IEEE Trans. Speech and Audio Process., 13(5): 821-831, 2005.
[21] Picovici, D. and Mahdi, A. E., New output-based perceptual measure for predicting subjective quality of speech, in Proc. Intl. Conf. on Acoustics, Speech, and Signal Process., ICASSP, Montreal, Canada, May 2004, pp. 633-636.
[22] ITU-T. Rec. P.861, Objective Quality Measurement of Telephone-band (300-3400 Hz) Speech Codecs, ITU-T, Geneva, Switzerland, 1996.

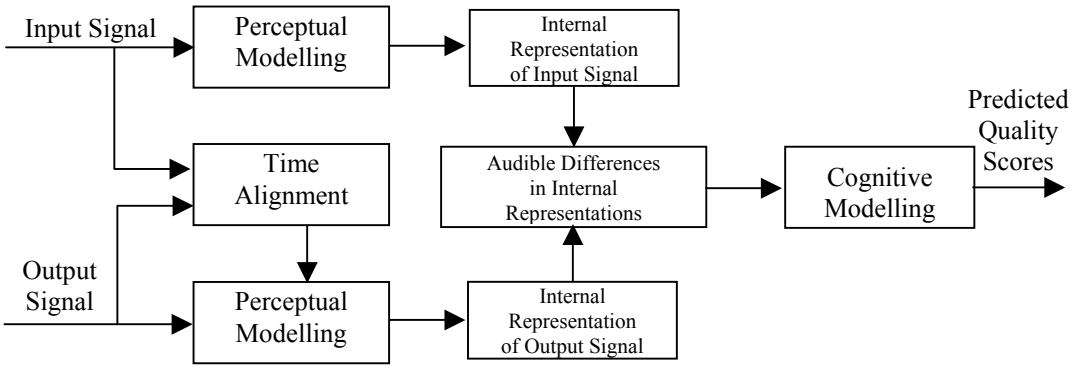

Figure 5 *The PESQ model.*

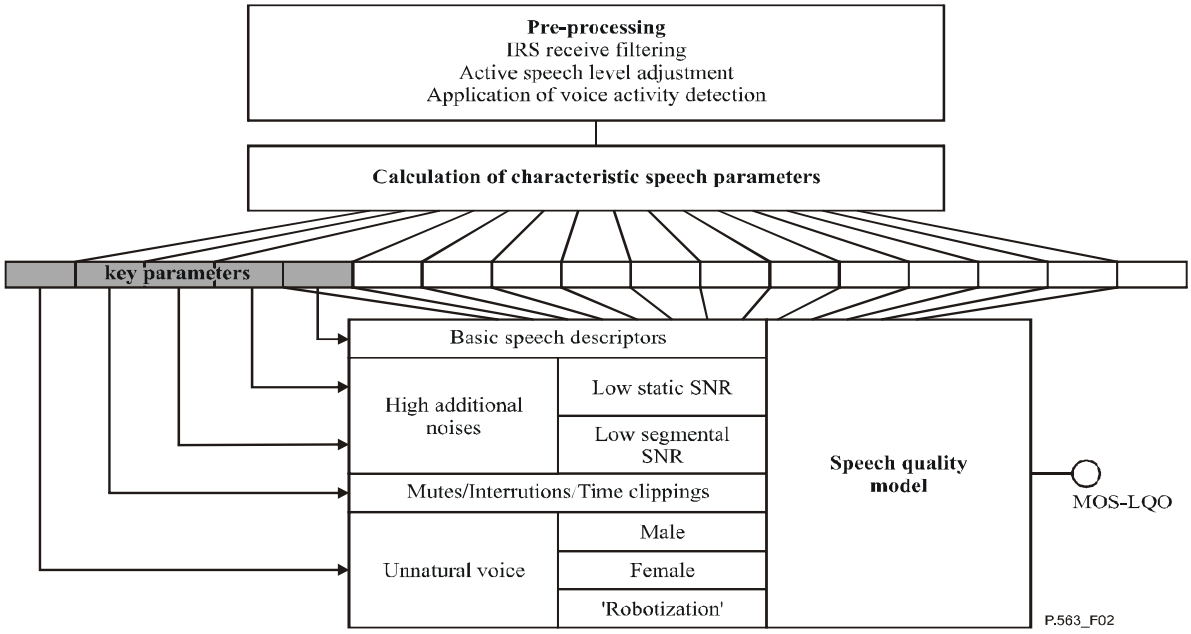

Figure 6 *Block diagram of the P.563 (adopted from Ref. [5]).*