We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

1

VoIP Quality Assessment Technologies

Mousa AL-Akhras and Iman AL Momani
The University of Jordan
Jordan

1. Introduction

Circuit Switching technology has been in use for long time by traditional Public Switched Telephone Network (PSTN) carriers for carrying voice traffic. Before users may communicate in circuit switching network, a dedicated channel or circuit is established from the sender to the receiver and that path is selected over the most efficient route using intelligent switches. Accordingly, it is not necessary for a phone call from the same sender to the same receiver to take the same route every time a phone call is made.

During call setup once the route is determined, that path or circuit stays fixed throughout the call and the necessary resources across the path are allocated to the phone call from the beginning to the end of the call. The established circuit cannot be used by other callers until the circuit is released, it remains unavailable to other users even when no actual communication is taking place, therefore, circuit switching is carrying voice with high fidelity from source to destination (Collins, 2003). Circuit switching is like having a dedicated railroad track with only one train, the call, is permitted on the track at one time.

Today's commercial telephone networks that based on circuit switching technology have a number of attractive features, including: Availability, Capacity, Fast Response and High Quality (Collins, 2003). The quality is the main focus of this chapter.

One alternative technology to circuit switching telephone networks for carrying voice traffic is to use data-centric packet switching networks such as Internet Protocol (IP) networks. In packet switching technology, no circuit is built from the sender to the receiver and packets are sent over the most effective route at time of sending that packet, consequently different packets may take different routes from the same sender to the same receiver within the same session.

Transmitting Voice over IP (VoIP) networks is an important application in the world of telecommunication and is an active area of research. Networks of the future will use IP as the core transport network as IP is seen as the long-term carrier for all types of traffic including voice and video. VoIP will become the main standard for third generation wireless networks (Bos & Leroy, 2001; Heiman, 1998).

Transmission of voice as well as data over IP networks seems an attractive solution as voice and data services can be integrated which makes creation of new and innovative services possible. This provides promises of greater flexibility and advanced services than the traditional telephony with greater possibility for cost reduction in phone calls. VoIP also has other advantages, including: number portability, lower equipment cost, lower bandwidth requirements, lower operating and management expenses, widespread availability of IP, and other advantages (Collins, 2003; Heiman, 1998; Low, 1996; Moon et al., 2000; Rosenberg et al.,

1999). VoIP can be used in many applications, including: call centre integration, directory services over telephones, IP video conferencing, fax over IP, and Radio/ TV Broadcasting (Collins, 2003; Miloslavski et al., 2001; Ortiz, 2004; Schulzrinne & Rosenberg, 1999).

VoIP technology was adopted by many operators as an alternative to circuit switching technology. This adoption was motivated by the above advantages and to share some of the high revenue achieved by telecommunication companies. However, to be able to compete with the highly reputable PSTN networks, VoIP networks should be able to achieve comparable quality to that achieved by PSTN networks. Although VoIP services often offer much cheaper solutions than what PSTN does, but regardless of how low the cost of the service is, it is the user perception of the quality what matters. If the quality of the voice is poor, the user of the traditional telephony will not be attracted to the VoIP service regardless of how cheap the service is. This comes from the fact that customers who are used to the high-quality telephony networks, expect to receive a comparable quality from any potential competitor.

IP networks were originally designed to carry non real-time traffic such as email or file transfer and they are doing this task very well, however, as IP networks are characterised by being best-effort networks with no guarantee of delivery as no circuit is established between the sender and the receiver, therefore they are not particularly appropriate to support real-time applications such as voice traffic in addition to data traffic. The best-effort nature of IP networks causes several degradations to the speech signal before it reaches its destination. These degradations arise because of the time-varying characteristics (e.g. packet loss, delay, delay variation (jitter), sharing of resources) of IP networks.

These characteristics which are normal to data traffic, cause serious deterioration to the real-time traffic and prevent IP networks from providing the high quality speech often provided by traditional PSTN networks for voice services. Sharing of resources in IP networks causes no resources to be dedicated to the voice call in contrast to what is happening in traditional circuit switching telephony such as PSTN where the required resources are allocated to the phone call from the start to the end. With the absence of resource dedication, many problems are inevitable in IP networks.

Among the problems is packet loss which occurs due to the overflow in intermediate routers or due to the long time taken by packets to reach their destinations (Collins, 2003). Real-time applications are also sensitive to delay since they require voice packets to arrive at the receiving end within a certain upper bound to allow interactivity of the voice call (ITU-T, 2003a;b). Also, due to their best-effort nature, packets could take different routes from the same source to the same destination within the same session which causes packets interarrival time to vary, a phenomenon known as jitter. Due to the problem of jitter, it is not easy to play packets in a steady fashion to the listener (Narbutt & Murphy, 2004; Tseng & Lin, 2003; Tseng et al., 2004). The above challenges cause degradation to the quality of the received speech signal before it reaches its destination. Many solutions have been proposed to alleviate these problems and the quality of the received speech signal as perceived by the end user is greatly affected by the effectiveness of these solutions.

Another approach is to reserve resources across the path from the sender to the receiver. A mechanism called Call Admission Control (CAC) is needed to determine whether to accept a call request if it is possible to allocate the required bandwidth and maintain the given QoS target for all existing calls, or otherwise to reject the call (Mase, 2004). Among the solutions that have been proposed to implement CAC and to manage the available bandwidth efficiently are: Resource Reservation Protocol (RSVP), Differentiated Service (DiffServ),

MultiProtocol Label Switching (MPLS), and End-to-end Measurement Based Admission Control (EMBAC). Reserving resources is difficult and very expensive proposal as it requires changes to all routers across the network which is inapplicable in non-managed networks such as the Internet.

Therefore, it is important to measure the quality of VoIP applications in live networks and take appropriate actions when necessary. This importance comes from legal, commercial and technical reasons. Measurement of the quality would be a necessity as customers and companies are bound by a service level agreement usually requiring the company to provide a certain level of quality, otherwise, customers may sue the companies for poor quality. Also, measuring the quality gives the chance to network administrators to overcome temporal problems that could affect the quality of ongoing voice calls. Measurement of the quality also allows service providers to evaluate their own and their competitors' service using a standard scale. It is also a strong indicator of users' satisfaction of the service provided (Takahashi et al., 2004; Zurek et al., 2002).

To this end, a specialised mechanism is required for measuring the speech quality accurately. One of driving forces in the world of telecommunication is the International Telecommunication Union (ITU). ITU is the leading United Nations (UN) agency for information and communication technology. As the global focal point for governments and the private sector in developing telecommunication networks and services, ITU's role is to help the world communicate. ITU - Telecommunication Standardisation Sector (ITU-T, http://www.itu.int/ITU-T/) is a permanent organ of the ITU that plays a driving force role toward standardising and regulating international telecommunications worldwide. Toward this goal, ITU-T study technical, operating and tariff questions and produce standards under the name of Recommendations for the purpose of standardising telecommunications worldwide. ITU-T's Recommendations are divided into categories that are identified by a single letter, referred to as the series, and Recommendations are numbered within each series, for example P.800 (ITU-T, 1996b). ITU-T has a formal recognition as it is part of ITU which is a UN Organisation (UNO).

Many ITU-T Recommendations are concerned with standardising the measurement of speech quality for voice services, many of these standards are considered in this chapter. Speech quality in ITU-T standards is expressed as Mean Opinion Score (MOS) which ranges between 1 and 5, with 1 corresponds to poor quality and 5 to excellent quality.

Some standards measure the speech quality or the MOS **subjectively** by setting lab conditions and asking subjects to listen to the speech signal and give their estimation of the quality in terms of MOS. This method is standardised in ITU-T Recommendation P.800 (ITU-T, 1996b). Other methods are **objective** that depend on comparison of the received signal with the original signal to measure the perceived quality in terms of MOS, these methods are known as **intrusive** methods as they require the injection of the original signal to analyse the distortion of the received signal. The most recent method for measuring the speech quality intrusively is known as Perceptual Evaluation of Speech Quality (PESQ). PESQ is standardised as ITU-T Recommendation P.862 (ITU-T, 2001). Yet another **objective** category depends on either the received signal or the networking parameters to estimate the quality **non-intrusively** without the need for the original signal. The two main methods in this category are Recommendation P.563 (ITU-T, 2004) and the E-model as defined in ITU-T Recommendation G.107 (ITU-T, 2009). Many other standards and methods have been proposed by other organisations, other researchers, and the authors of this chapter independent of the ITU-T, these attempts will be discussed in detail later in the chapter.

The selection of a method for VoIP quality assessment should take the characteristics of IP networks and voice calls into consideration. Such characteristics that affect the selection include the requirement to measure the quality of live-traffic while the network is running in a real environment during a voice call. To able to do this, an objective solution that measures the quality without human interference and depending on the received signal at the receiver side without the need for the original speech signal at the sender side; i.e. a non-intrusive measurement is needed.

This chapter aims to serve as a reference and survey for readers interested in the area of speech quality assessment in VoIP networks. The rest of this chapter is organised as follows: Section 2 categorises speech quality assessment techniques and discusses the main requirements of an applicable technique in VoIP environment. Sections 3 and 4 discusses subjective and objective quality assessment technologies, respectively. To avoid ambiguity, different qualifiers are used to distinguish between different quality measurement methods and presented in section 5. Conclusions and possibilites for future work are given in section 6.

2. Categories of VoIP quality assessment technologies

VoIP quality assessment methods can be categorised into either subjective methods or objective methods. Objective methods can be either intrusive or non-intrusive. Non-intrusive methods can be either signal-based or parametric-based. Figure 1 depicts different classifications.

The primary criterion for voice and video communication is subjective quality, the user's perceptions of service quality. A subjective quality assessment method is used to measure the quality. Subjective quality factors affect the quality of service of VoIP, among those factors are: packet loss, delay, jitter, loudness, echo, and codec distortion. To measure the subjective quality, a subjective quality assessment method is used, the most widely accepted metric is the Mean Opinion Score (MOS) as defined by ITU-T Recommendation P.800 (ITU-T, 1996b).

However, although subjective quality assessment is the most reliable method, it is also time-consuming and expensive as any other subjective test. Thus other methods to automatically estimate quality objectively should be considered. This can be done intrusively by comparing the reference signal with the degraded signal or non-intrusively utilising physical quality parameters or the received signal without using the reference signal.

The applicability of any solution for measuring the speech quality in VoIP networks should take into consideration the nature of IP networks and the characteristics of voice traffic. Among the desired features for a VoIP speech quality assessment solution are:

1. Automatic: It should provide measurement of speech quality online while the network is running.
2. Non-intrusive: It should be able to provide measurement of the speech quality depending on the received speech signal or network parameters without the need for the original signal.
3. Accurate: It should provide accurate measurement of speech quality to reflect how the quality is perceived by the end-user.
4. With the changing world, it should be applicable to new and emerging applications and networking conditions. As such it should avoid the subjectivity in estimating parameters. The E-model (section 4.2) for example depends on subjective tests to estimate packet loss parameters which hinders its applicability for new networking conditions.

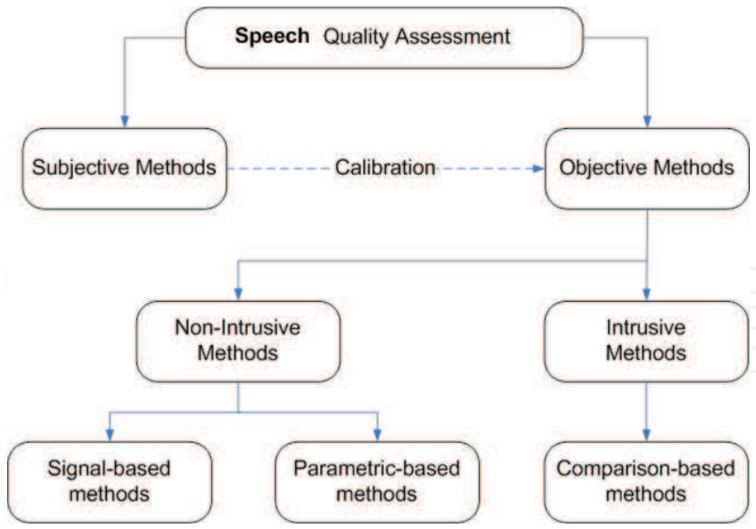


Fig. 1. Overview of VoIP measurement methods (Sun, 2004)

Based on the above requirements and from the previous discussion, the subjective and intrusive solutions that will be discussed in sections 3 and 4.1 respectively cannot be used for such task as they are manual and intrusive, respectively. The non-intrusive objective solutions that will be discussed in section 4.2 are candidates for such task. The most famous and widely used non-intrusive subjective solutions for measuring the speech quality are P.563 (ITU-T, 2004) and the E-model (ITU-T, 2009).

3. Subjective assessment of quality

The most widely used subjective quality assessment methodology is opinion rating defined in ITU-T Recommendation P.800 in which a panel of users (test subjects) perform the subjective tests of voice quality and give their opinions on the quality (ITU-T, 1996b). Subjective tests could be conversational or listening-only tests. In conversational test, two subject share a conversation via the transmission system under test, where they are placed in separated and isolated rooms to report their opinion on the opinion scale recommended by ITU-T and the arithmetic mean of these opinions is calculated. In listening tests, one subject is listening to pre-recorded sentences (ITU-T, 1996b).

To conduct a subjective experiment according to the ITU-T Recommendation P.800, strict lab conditions should be in place. Such conditions concerns the room size, noise level, and the use of sound-proof cabinet in a room with a volume not less than 20 m^3. In case of recording the test room must be a volume of 30 m^3 up to 120 m^3, with an echo duration lower than 500 ms (200-300 ms is preferred) and a background noise lower than 30 decibel (dB). The recording system must be of high quality and recorded voice signals must consist of simple, meaningful, short phrases, taken from newspapers or non-technical lectures, randomly ordered (phrases of 3-6 seconds of length or conversations of 2-5 minutes of length), all the used material must be recorded with a microphone at a distance of 140-200 mm from the speakers mouth. Also, the sound pressure level should be measured from a vertical position above the subjects seat while the furniture in place (ITU-T, 1996b).

Recommendation P.800 also specifies other conditions regarding the subjects who participate in the test such as they have not been directly involved in work connected with assessment of

the performance of telephone circuits, or related work such as speech coding, also they have not participated in any subjective test whatever for at least the previous six months, and not in a conversational/listening test for at least one year. In case of listening-test they have never heard the same sentence lists before (ITU-T, 1996b).
In opinion rating methodology the performance of the system is rated either directly (Absolute Category Rating, ACR) or relative to the subjective quality of a reference system as in (Degradation Category Rating, DCR), or Comparison Category Rating (CCR) (ITU-T, 1996b; Takahashi, 2004; Takahashi et al., 2004).
The most common metric in opinion rating is Mean Opinion Score (MOS) which is an ACR metric with five-point scale: (5) Excellent, (4) Good, (3) Fair, (2) Poor, (1) Bad (ITU-T, 1996b). MOS is internationally accepted metric as it provides direct link to the quality as perceived by the user. A MOS value is obtained as an arithmetic mean for a collection of MOS scores (opinions) for a set of subjects. When the subjective test is listening-only, the results are in terms of listening subjective quality; i.e. MOS - Listening Quality Subjective or MOS_{LQS}. When the subjective test is conversational, the results are in terms of conversational subjective quality; i.e. MOS - Conversational Quality Subjective or MOS_{CQS} (ITU-T, 1996b; 2006). Although the overall quality of VoIP must be discussed in term of conversational quality, listening quality assessment is also quite helpful in analysing the effect of individual quality factors such as distortion due to speech coding and packet loss.
In DCR test two samples (A and B) are present: A represents the reference sample with the reference quality, while B represents the degraded sample. The subjects are instructed to acoustically compare the two samples and rate the degradation of the B sample in relation to the A sample according to the following five-point degradation category scale: degradation is (5) inaudible, (4) audible but not annoying, (3) slightly annoying, (2) annoying, and (1) very annoying. The samples must be composed of two periods, separated by silence (for example 0.5 seconds), firstly sample A then sample B.
The results (opinions) are averaged as Degraded MOS (DMOS). Each configuration is evaluated by means of judgements on speech samples from at least four talkers. DCR test affords higher sensitivity and used with high-quality voice samples, this is especially useful when the impairment is small and a sensitive measure of the impairment is required as ACR is inappropriate to discover quality variations as it tends to lead to low sensitivity in distinguishing among good quality circuits (ITU-T, 1996b; Takahashi et al., 2004).
The CCR method is similar to the DCR method as subjects are presented with a pair of speech samples (A and B) on each trial. In the DCR procedure, a reference sample is presented first sample (A) followed by the degraded sample (B). In the DCR method, listeners always rate the amount by which sample B is degraded relative to sample A. In the CCR procedure, the order of the processed and unprocessed samples is chosen at random for each trial. On half of the trials, the unprocessed sample is followed by the processed sample. On the remaining trials, the order is reversed. Listeners use the following scale: (3) Much Better, (2) Better, (1) Slightly Better, (0) About the Same, (-1) Slightly Worse, (-2) Worse, and (-3) Much Worse (ITU-T, 1996b). In this technique listeners provide two judgements with one response where the advantage of the CCR method over the DCR procedure is the possibility to assess speech processing that either degrades or improves the quality of the speech. The quantity evaluated from the scores is represented as Comparison MOS (CMOS).
Results of MOS scores should be dealt with care as results may vary depending on the speaker, hardware platform, listening groups and test data and slight variation between different subjective tests should be expected although the above rigid conditions should guarantee

minimisation of such cases.
Although opinion rating methods are the most famous subjective quality assessment methodology, but other methods have also been proposed. Diagnostic Rhyme Test (DRT) is an intelligibility measure where the subject task is to recognise one of two possible words in a set of rhyming pairs (e.g. meat-beat). Diagnostic Acceptability Measure (DAM) scores are based on results of test methods evaluating the quality of a communication system based on the acceptability of speech as perceived by a trained normative listener (Spanias, 1994). Li (2004) proposed the use of intelligibility index as an additional parameter that can be used along with the commonly used MOS score. Opinion rating methods are still the most famous and widely used method.
Although subjective quality measurement is the most accurate and reliable assessment method to measure the quality as it reflects the user's perceptions of service quality, but there are few problems associated with subjective tests. It is apparent from the strict conditions associated with opinion rating methods as mentioned above that the inherent problems in subjective MOS measurement are that it is: time-consuming, expensive, lacks repeatability , and inapplicable for monitoring live voice traffic as commonly needed for VoIP applications. This has made objective methods very attractive to estimate the subjective quality for meeting the demand for voice quality measurement in communication networks to avoid the limitations of the subjective tests.

4. Objective assessment of quality

Objective speech quality assessment simulates the opinions of human testers algorithmically or using computational models to automatically evaluate the transmitted speech quality over IP networks to replace the human subjects, where the aim is to predict MOS values that are as close as possible to the rating obtained from subjective test and to avoid the limitations of subjective assessment methods. However, as subjective methods are the most accurate and reliable methods for measuring speech quality, they are used to calibrate objective methods. Therefore the accuracy, effectiveness and performance evaluation of objective methods are determined by their correlation with the subjective MOS scores.
Objective assessment of speech quality is based on objective metrics of speech signal or properties of the carrier network. Objective quality assessment methodologies can be categorised into two groups: Intrusive speech-layer models and Non-Intrusive models (Signal-based and parametric-based). Figure 2 shows the three main types of objective measurement.

4.1 Intrusive objective assessment of quality

Intrusive measures, often referred to as input-to-output measures or comparison-based methods, base their quality measurement on comparing the original (clean or input) speech signal with the degraded (distorted or output) speech signal as reconstructed by the decoder at the receiver side, this is shown in Figure 2 (a). Intrusive objective assessment of speech quality or speech-layer objective models are full-reference methods for measuring the quality. They provide an accurate method for measuring speech quality as they require the original or reference speech signal as input and produce measurement of listening MOS by comparing the post-transmitted signal with the original one (double-ended) using a distance measure, based on this comparison the quality of the degraded signal is measured in comparison with the quality of the original signal. However, such methods are inapplicable in monitoring live traffic because it is difficult or impossible to obtain actual speech samples as the

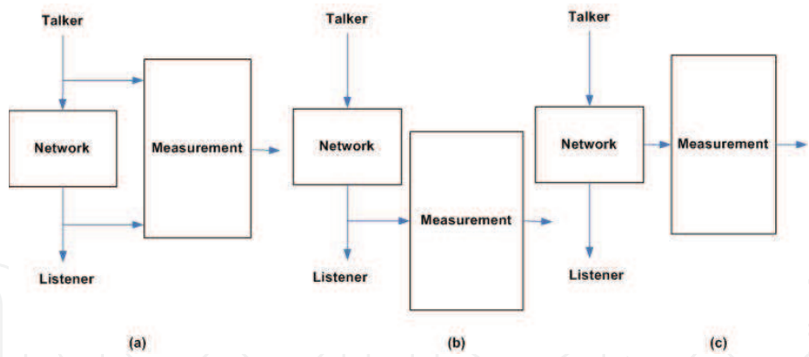


Fig. 2. Three main categories of objective quality measurement: (a) Comparison-based intrusive method, (b) Signal-based non-intrusive method, (c) Parametric-based non-intrusive method (Sun, 2004)

reference signal is not available at the receiver side. Some intrusive algorithms are used in time-domain, other objective quality assessment methods make use of spectral distortion (frequency-domain) to evaluate the performance of LBR codecs. None of these standards were accurate enough to be adopted by ITU-T. Later, perceptual domain measures were introduced and standardised.

Perceptual domain measures are based on models of human auditory perception. These measures transform the speech signal into a perceptually relevant domain such as bark spectrum or loudness domain, and incorporate human auditory models (Sun, 2004). In perceptual measure the original and the degraded signal are both transformed into a psychophysical representation that approximates human perception or simulate 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. Then the perceptual difference between the original and the degraded signal is mapped into estimation of perceptual quality difference as perceived by the listener. Perceptual domain measures have shown to be highly accurate objective performance measures because many modern codecs are nonlinear and non-stationary making the shortcomings of the previous objective measures even more evident. Perceptual domain measures include: Measuring Normalising Block (MNB), Perceptual Analysis Measurement System (PAMS), Perceptual Speech Quality Measure (PSQM), and Perceptual Evaluation of Speech Quality (PESQ) which is the latest ITU-T intrusive standard for assessing speech quality for communication systems and networks (ITU-T, 1998; 2001; Sun, 2004; Voran, 1999a;b).

4.1.1 Signal to noise ratio (SNR) and segmental signal-to-noise-ratio (SegSNR)

Time domain measures are the simplest intrusive measures that consist of an analogue or waveform-comparison algorithms in which the target is to reproduce a copy of input waveform such that the original and distorted signals can be time-aligned and noise can be accurately calculated, SNR and SegSNR are the most important method of this category. Signal refers to useful information conveyed by some communications medium, and noise refers to anything else on that medium. SNR gives a measure of the signal power improvement related to the noise power calculated for the original signal and the degraded signal. In SNR sample-by-sample comparison is performed. SEGmental SNR (SegSNR) can also be utilised where SNR is computed for each N-point segment of speech to detect

temporal variations. As time-domain measures, SNR and SegSNR can be used for evaluation of non-speech signals (Quackenbush et al., 1988; Mahdi and Picoviciv, 2009). SNR is defined as the ratio of a signal power to the noise power corrupting the signal:

$$SNR = 10\log_{10}\frac{\sum_n x^2(n)}{\sum_n (x(n)-d(n))^2}, \quad (1)$$

where x(n) represents the original (undistorted) speech signal, d(n) represents the distorted speech reproduced by a speech processing system and n is the sample index (determined points on time domains). SegSNR calculates the SNR for each N-points segment of speech. The result is an average of SNR values of segments, and can be computed as follows:

$$SSNR = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\left(\frac{\sum_{n=Nm}^{Nm+N-1} x^2(n)}{\sum_{n=Nm}^{Nm+N-1}[d(n)-x(n)]^2}\right), \quad (2)$$

Where x(n) represents the original speech signal, d(n) represents the distorted speech signal, n is the sample index, N is the segment length, and M is the number of segments in the speech signal. Classical windowing techniques are used to segment the speech signal into appropriate speech segments.
SNR and SegSNR algorithms are easy to implement, have low computational complexity, can provide good performance measure of voice quality of waveform codecs.
Although SNR is the most common method, its main problem is that it cannot be used with Low-Bit-Rate (LBR) codecs as these codecs do not preserve the shape of the signal. As such SNR cannot compare the pre-encoded signal with the post-decoded signal as they show little correlation to perceived speech quality when applied to LBR codecs such as vocoders as in these codecs the shape of the signal is not preserved and they become meaningless (Kondoz, 2004). These measures are also sensitive to a time shift, and therefore require precise signal alignment such that to achieve the correct time alignment it may be necessary to correct phase errors in the distorted signal or to interpolate between samples in a sampled data system.

4.1.2 Spectral domain measures

Spectral domain measures or frequency-domain measures are known to be significantly better correlated with human perception, but still relatively simple to implement. One of their critical advantages is that they are less sensitive to signal misalignment and phase shift between the original and the distorted signals than time domain measures. Most spectral domain measures are related to speech codecs design and use the parameters of speech production models. Their capability to effectively describe the listeners auditory response is limited by the constraints of the speech production models. Some of the most popular frequency domain techniques are the Log-Likelihood (LL) (Itakura, 1975), the ItakuraSaito (IS) (Itakura & Saito, 1978), and the Cepstral Distance (CD) (Kitawaki et al., 1988).

4.1.3 Measuring normalising blocks (MNB)

In this algorithm, both the input and the output speech signals are perceptually transformed and a distance measure that consists of a hierarchy of Measuring Normalising Blocks (MNB) is then calculated. Each MNB integrates two perceptually transformed signals over some time or frequency interval to determine the average difference across the interval. This difference is then normalised out of one signal to provide one or more measurements.
MNB algorithm starts by estimating the delay between the input speech signal and the output

speech signal due to the device or system (possibly IP network) under test. This is done using cross-correlation of speech envelops because LBR codecs do not preserve the speech waveform, therefore waveform cross correlation gives misleading estimation for the delay. Once the delay is estimated and compensated for, MNB proceed to the next step which is perceptual transformation.
In perceptual transformation, the representation of the audio signal is modified in such a way it is approximately equivalent to the human hearing process and only perceptual information is retained. The following steps are performed by Voran in (Voran, 1999a;b) on the speech signal sampled at a rate of 8000 samples/s before perceptual transformation. The speech signal is divided into frames of size 128 samples with 50% frame overlap. Each frame is multiplied by a Hamming window and transformed using FFT transform and only the squared magnitudes of the FFT coefficients are preserved. As Voran pointed out, the nonuniform ear's frequency resolution on the Hertz scale and nonlinear relation between loudness perception and signal intensity are the most important perceptual properties to model (Voran, 1999a).
For modelling the nonuniform frequency resolution, the Hertz frequency scale is replaced by a psychoacoustic frequency scale such as the bark frequency scale using the relation:

$$b = 6.\sinh^{-1}\left(\frac{f}{600}\right) \tag{3}$$

where b is the Bark frequency scale variable. f is Hertz frequency scale variable.

Figure 3 shows the transformation from Hertz to Bark scale. In bark scale, roughly equal frequency intervals are of equal importance. From the figure it can be seen that on the band 0-1 kHz in Hertz scale (corresponding to 0-7.703 Bark) is given equal importance by Bark scale as the band 1-4 kHz. It is worth noting that bark scale is used recently for measuring speech quality for wideband speech coding (Haojun et al., 2004). To model the nonlinear relation between loudness perception and signal intensity, the logarithmic function is used to convert signal intensity to perceived loudness.
The distance between the two signals is calculated using a hierarchy of Time MNB (TMNB) and Frequency MNB (FMNB). The hierarchy structure works from larger time and frequency scales down to smaller time and frequency scales. Each block integrates the perceptually transformed signals over time or frequency to determine the average difference between the two signals. Once all the measurements of the hierarchy are calculated on different levels, these measurement are linearly combined to calculate the Auditory Distance (AD) between the two signals. Finally the AD can be mapped using a logistic function into a finite set of values from 0 to 1 to increase correlation with subjective tests (Voran, 1999a;b).

4.1.4 Perceptual analysis measurement system (PAMS)

Developed by Psytechnics, a UK-based company associated with British telecommunications (BT). The PAMS process uses an auditory model that combines a mathematical description of the psychophysical properties of human hearing with a technique that performs a perceptually relevant analysis taking into account the subjectivity of the errors in the received signal. PAMS extracts and selects parameters describing speech degradation addressed by damaging factors such as time clipping, packets loss, delay and distortion due to the codec usage and constrained mapping to subjective quality. PAMS compares the original and the received signals and produces two scores, listening quality score (Ylq) and listening effort

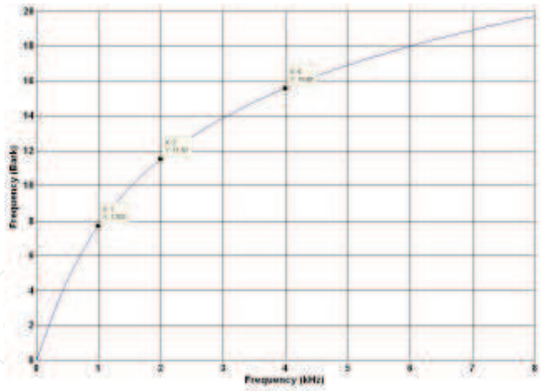


Fig. 3. Transformation from Hertz scale to Bark scale

score (Yle). Both scores are in the range 1 to 5 and MOS score can be estimated using a linear combination of both scores (Duysburgh et al., 2001; Zurek et al., 2002).

4.1.5 Perceptual speech quality measure (PSQM)

PSQM originally developed by KPN, Netherlands and then standardised by ITU-T as Recommendation P.861. PSQM transforms the speech signal into the loudness domain, applies a nonlinear scaling factor to the loudness vector of distorted speech. The scaling factor is obtained by calculating the loudness ratio of the reference and the distorted speech. The difference between the scaled loudness of the distorted speech and loudness of the reference speech is called Noise Disturbance (ND). The final estimated distortion is an average ND over all the frames processed where a small weight is given to silence portions during calculations. PSQM computes the distortion frame by frame, with the frame length of 256 samples with 50% overlap. The result is shown in ND as a function of time and frequency. The average ND is directly related to the quality of coded speech. There are two meaningful scores in the PSQM measure: one is a distortion measure and the other is a mapped number such as MOS. PSQM measurements are in the range 0 to 6.5 with lower values means lower distortion which indicates better quality. PSQM scores can be mapped into MOS scores using a nonlinear mapping (ITU-T, 1998; Sun, 2004; Zurek et al., 2002).

PSQM was designed to work under error-free coding conditions, therefore it is inapplicable for VoIP environment which suffers from packet loss especially in mobile communications that suffer from bit errors. PSQM+ was proposed by KPN to improve the performance of PSQM for loud distortions and temporal clipping. PSQM+ uses the same perceptual transformation module as PSQM. Comparing to PSQM, an additional scaling factor is introduced when the overall distortion is calculated. This scaling factor makes the overall distortion proportional to the amount of temporal clipping distortion. Otherwise, the cognition module is the same as PSQM.

4.1.6 Perceptual evaluation of speech quality (PESQ)

PESQ is the latest ITU-T standard for objective evaluation of speech quality in narrowband telephony network and codecs. It was a result of a collaboration project between KPN and BT by combining the two speech quality measures PSQM+ and PAMS. Later it was standardised by ITU-T as Recommendation P.862 (ITU-T, 2001; Rix et al., 2001). Upon its standardisation, PSQM in Recommendation P.861 was withdrawn by ITU-T (ITU-T, 2001; 2005b; Rohani & Zepernick, 2005; Zurek et al., 2002).

Real systems may include filtering and variable delay, as well as distortions due to channel errors and LBR codes. PSQM was designed to assess speech codec and is not able to take proper account of filtering, variable delay, and short localised distortions. PESQ was specifically developed to be applicable to end-to-end voice quality testing under real network conditions, such as VoIP, ISDN etc. The results obtained by PESQ was found to be highly correlated with subjective tests with correlation factor of 0.935 on 22 ITU benchmark experiments, which cover 9 languages (American English, British English, Dutch, Finnish, French, German, Italian, Swedish and Japanese).

In PESQ the original and the degraded signals are time-aligned, then both signals are transformed to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system, taking account of perceptual frequency (Bark) and loudness (Sone). After this transformation to the internal representation, the original signal is compared with the degraded signal using a perceptual model. This is achieved in several stages: level alignment to a calibrated listening level, compressive loudness scaling, and averaging distortions over time as illustrated in Figure 4 (ITU-T, 2001; Rix et al., 2001).

PESQ score lies in the range -0.5 to 4.5, to make such score comparable with ACR MOS score, a function is provided in Recommendation P.862.1 to map these values to the range 1 to 5. The function in equation (4) do the conversion from a PESQ score to a MOS - Listening Quality Objective or MOS_{LQO} which makes the comparison with other MOS results very convenient independent of the implementation of ITU-T Recommendation P.862 (ITU-T, 2005b).

$$MOS_{LQO} = 0.999 + \frac{4.999 - 0.999}{1 - e^{(-1.4945 * PESQ + 4.6607)}} \tag{4}$$

ITU-T Recommendation P.862.1 (ITU-T, 2005b) also provides a formula to move back to PESQ score from an available MOS_{LQO} score. The equation is:

$$PESQ = \frac{4.6607 - \ln\left(\frac{4.999 - MOS_{LQO}}{MOS_{LQO} - 0.999}\right)}{1.4945} \tag{5}$$

In 2005, the ITU-T issued Recommendation P.862.2 ITU-T (2005c). P.862.2 extends the application of P.862 PESQ to wideband audio systems (50-7000 Hz). The definition of a new output mapping function, which is a modification to that recommended in P.862.1, to be used with wideband applications is as follows:

$$MOS_{LQO} = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.3669 PESQ + 3.8224}} \tag{6}$$

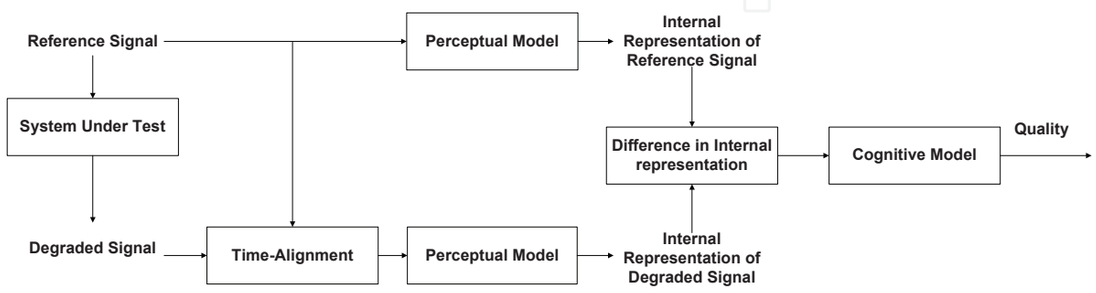


Fig. 4. Conceptual diagram of PESQ philosophy (ITU-T, 2001)

4.1.7 Other methods

Many other intrusive objective methods have been proposed by other researchers. Fu et al. proposed using an ANN model where the feature vector of the input vector and output vector corresponding to the original and degraded signal respectively are fed to the ANN model in addition with MOS subjective score as calculated using set of subjects as a target for the ANN. Utilising the proposed method and using the error between the original input signal and the output signal as input to the ANN model, MOS score can be estimated directly using one-step rather than the usual approach of estimating the distortion and then mapping the quality (Fu et al., 2000).

4.2 Non-intrusive objective assessment of quality

The intrusive methods described in section 4.1 need the input signal as a reference for comparison with the output signal which is a problem in live network as it is difficult to obtain the original speech signal at the receiver side. Additionally, in some situations the input speech may be distorted by background noise, and hence, measuring the distortion between the input and the output speech does not provide an accurate indication of the speech quality of the communication system.

On the other hand, non-intrusive measures (also known as output-based or passive measures) use only the degraded signal without access to the original signal. Non-intrusive methods provide a convenient measure for monitoring of live networks. Real-time quality assessment is important for instance if the application is to perform some form of dynamic quality control, e.g. by changing encoding or redundancy parameters to optimise quality when network conditions worsen.

Different methods have been proposed for objectively estimating the speech quality non-intrusively. These methods vary in complexity and accuracy from very simple techniques to very complex ones. It is identified that the non-intrusive objective assessment of quality is the most appropriate method for monitoring the speech quality in VoIP networks. This section discusses different non-intrusive methods for measuring the speech quality objectively. Non-intrusive methods are also divides into to subcategories, signal-based and parametric-based methods.

Signal-based methods are based on digital signal processing techniques that estimate the speech quality when the envelope of the speech signal may have suffered from degradation overtime due to LBR coding or transmission over noisy wireless links, in other words signal-based methods process the audio stream that is decoded after buffer playout to extract relevant information for estimating the voice quality. ITU-T Recommendation P.563 or 3SQM (Single-Sided Speech Quality Measurement) that achieves a correlation coefficient with subjective tests of around 0.8 defined to be a standard for this type of measure (ITU-T, 2004).

On the other hand, parametric-based methods based their results on various properties relevant to telecommunication network parameters for example packet loss, delay and jitter. This makes parametric model to be more specific for a particular type of communications network by depending their prediction on the parameters of that network, which makes parametric-based methods to be more accurate than signal-based methods for that network which are more suitable for general prediction for a wider variety of networks and conditions. The E-model which is one of the most widely used parametric-based methods defined according to ITU-T Recommendation G.107 (ITU-T, 2009). The details of the ITU-T Recommendation P.563 or 3SQM and the E-Model are discussed in the next sections.

4.2.1 ITU-T recommendation P.563 or 3SQM

In 2004 ITU-T standardised its P.563 Recommendation (ITU-T, 2004) for single-ended objective speech quality assessment in narrow-band telephony applications. Recommendation P.563 approach is the first ITU-T Recommendation for single-ended signal-based non-intrusive measurement application that takes into account perceptual distortions to predict the speech quality on a perception-based scale to produce MOS - Conversational Quality Objective or MOS_{CQO}. This Recommendation is not restricted to end-to-end measurements; it can be used at any arbitrary location in the transmission Path. The basic block diagram of P.563 is shown in Figure 5.

This visualisation explains also the main application and allows the user to rate the scores gained by P.563. 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, the algorithm combines 4 processing stages as illustrated in Figure 6: preprocessing; basic distortion classes and speech parameters extraction; detection of dominant distortion; and mapping to final quality estimate. Brief overview of the main steps is given here:

– **Preprocessing**: The first preprocessing step in the Intermediate Reference System (IRS) filtering, where the speech signal to be assessed is filtered to simulate a standard receiving telephone handset. This is followed by a Voice Activity Detector (VAD) to separate speech from silence. The speech level is then calculated and adjusted to -26 dBov.

– **Extraction of basic distortion classes and speech parameters**: The preprocessed speech signal is analysed to detect a set of characterising signal parameters. In total there are 51 distortion parameters that are divided up into 3 independent functional blocks, namely: vocal tract analysis and unnaturalness of speech; analysis of strong additional noise; and speech interruptions, mutes and time clipping. All of these distortion classes are based on very general principles that make no assumptions about the underlying network or distortion types occurring under certain conditions. Additionally, a set of basic speech descriptors like active speech level, speech activity and level variations are used, mainly for adjusting the pre-processing and the VAD. Some of the signal parameters calculated within the pre-processing stage are used in these 3 functional blocks.

– **Detection of dominant distortion**: This analysis is applied at first to the signal. Based on a restricted set of key parameters, an assignment to a main distortion class will be made. The key parameters and the assigned distortion class are used for the adjustment of the speech

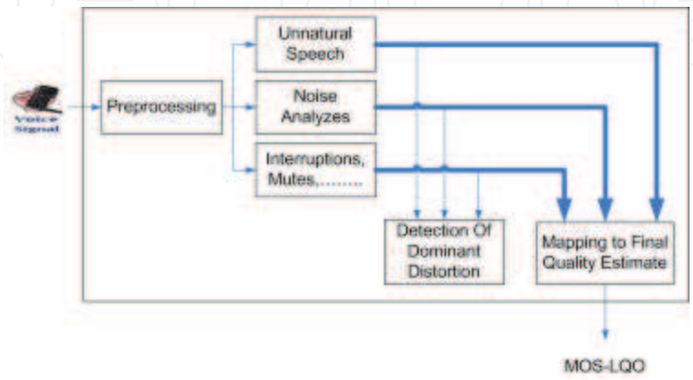


Fig. 5. Basic block diagram of P.563 overall structure (ITU-T, 2004)

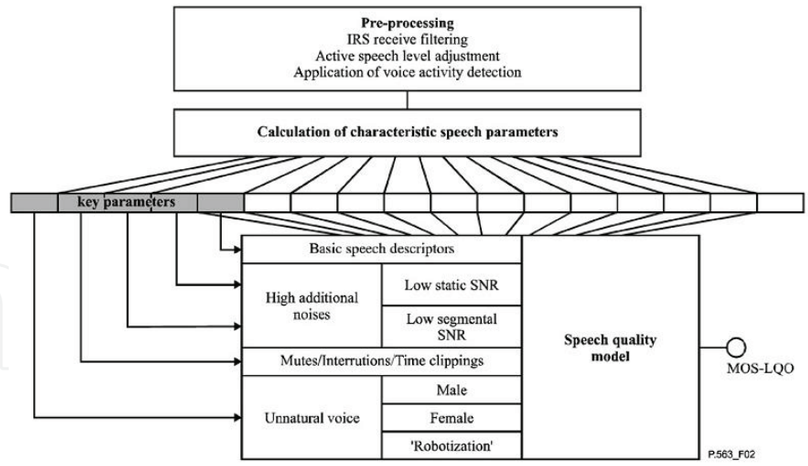


Fig. 6. Block diagram of P.563 algorithm detailing the various distortion classes used (ITU-T, 2004)

quality model. This provides a perceptual based weighting where several distortions occur in the signal but one distortion class is more prominent than the others. The process models the phenomenon that any human listener focuses on the foreground of the signal stream; i.e. the listener would not judge the quality of the transmitted voice by a simple sum of all occurred distortions but because of a single dominant noise artifact in the signal.
In the case of several distortions occurring to the signal, a prioritisation is applied on the distortion classes according to the distortions relevance with respect to the average listeners opinions. This is followed by estimation of an intermediate speech quality score for each class distortion. Each class distortion uses a linear combination of parameters to generate the intermediate speech quality. The final speech quality estimate is calculated by combining the intermediate quality results with some additional signal features.

– **Final quality estimate**: In this stage, a speech quality model is used to map the estimated distortion values into a final quality estimate equivalent to MOS_{CQO}. The speech quality model is composed of 3 main blocks:
  – decision on a distortion class.
  – speech quality evaluation for the corresponding distortion class.
  – overall calculation of speech quality.

4.2.2 The E-model

One of the most widely used methods for objectively evaluating the speech quality non-intrusively is opinion modelling. In opinion models subjective quality factors are mapped into manageable network and terminal quality parameters to automatically produce an estimate of subjective quality. The most famous standard for opinion modelling is the E-model which is defined according to ITU-T Recommendation G.107 (ITU-T, 2009; Takahashi et al., 2004).
The E-model, abbreviated from the European Telecommunications Standards Institute (ETSI), was developed by a working group within ETSI during the work on ETSI Technical Report ETR 250 (ETSI, 1996). It is a computational tool originally developed as a network planning tool, but it is now being used for objectively estimating voice quality for VoIP applications

using network and terminal quality parameters. In the E-model, the original or reference signal is not used to estimate the quality as the estimation is based purely on the terminal and network parameters. Network parameters such as packet loss rate can be estimated from information contained in the headers of Real-time Transport Protocol (RTP) and Real-time Transport Control Protocol (RTCP). The E-model is a non-intrusive method of measuring the quality as it does not require the injection of the reference signal (ITU-T, 2009; Sun, 2004; Takahashi et al., 2004).

In the E-model, the subjective quality factors are mapped into manageable network and terminal quality parameters. Among the network quality parameters are: network delay and packet loss. Among the terminal quality parameters are: jitter buffer overflow, coding distortion, jitter buffer delay, and echo cancellation. Example of mapping is the mapping of delay subjective quality parameter into network delay and jitter buffer delay.

The fundamental principle of the E-model is based on a concept established by J. Allnatt around 35 years ago (Allnatt, 1975):

Psychological factors on the psychological scale are additive

It is used for describing the perceptual effects of diverse impairments occurring simultaneously on a telephone connection. Because the perceived integral quality is a multidimensional attribute, the dimensionality is reduced into one-dimension so-called transmission rating factor, *R*-Rating Factor. Based on Allnatt's psychological scale all the impairments are - by definition - additive and thus independent of one another.

In the E-model all factors responsible for quality degradation are summed on the psychological scale. Due to its additive principle, the E-model is able to describe the effect of several impairments occurring simultaneously.

The E-model is a function of 20 input parameters that represent the terminal, network, and environmental quality factors (quality degradation introduced by speech coding, bit error, and packet loss is treated collectively as an equipment impairment factor).

The E-model starts by calculating the degree of quality degradation due to individual quality factors on the same psychological scale. Then the sum of these values is subtracted from a reference value to produce the output of the E-model which is the *R*-Rating Factor. The *R*-Rating Factor lies in the range of 0 and 100 to indicate the level of estimated quality where R=0 represents an extremely bad quality and R=100 represents a very high quality. The *R*-Rating Factor can be mapped into a MOS score based on the G.107 ITU-T's Recommendation (ITU-T, 2009) as explained later in this section. The reference model that represents the E-model is depicted in Figure 7 (ITU-T, 2009). The input parameters to the E-model, beside their default values and permitted range are listed in Table 1.

By following the additive principle, the E-model is able to describe the effect of several impairments occurring simultaneously, the *R*-Rating Factor combines the effects of various transmission parameters such as (packet loss, jitter, delay, echo, noise). The *R*-Rating Factor is calculated according to the following formula which follows the previous summation principle:

$$R = R_0 - Is - Id - Ie\text{-}eff + A \tag{7}$$

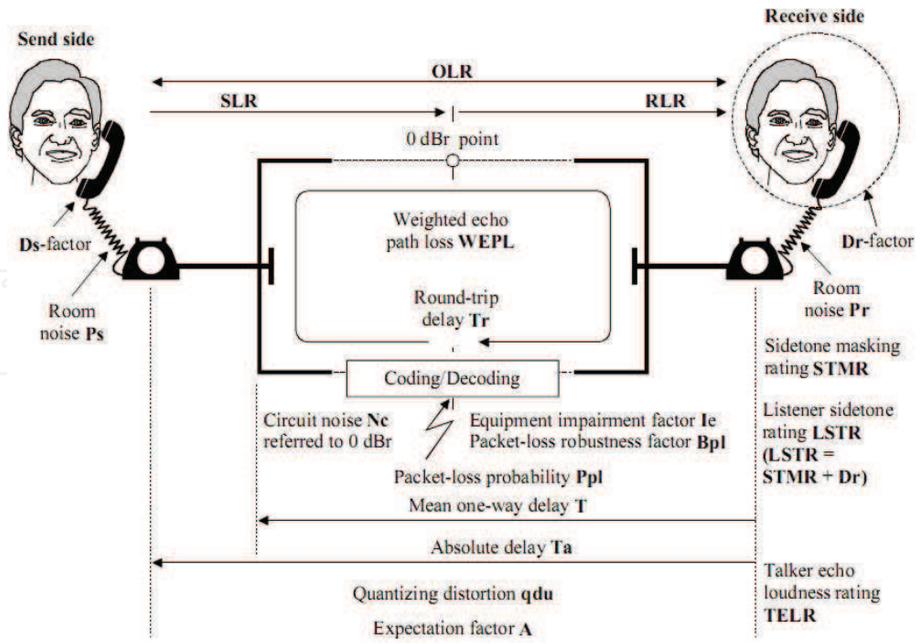


Fig. 7. Reference connection of the E-model (ITU-T, 2009)

Parameter	Default value	Permitted range
Send Loudness Rating	8	0...+18
Receive Loudness Rating	2	-5...+14
Sidetone Masking Rating	15	10...20
Listener Sidetone Rating	18	13...23
D-Value of Telephone, Send Side	3	3...+3
D-Value of Telephone, Receive Side	3	-3...+3
Talker Echo Loudness Rating	65	5...65
Weighted Echo Path Loss	110	5...110
Mean one-way Delay of the Echo Path	0	0...500
Round-Trip Delay in a 4-wire Loop	0	0...1000
Absolute Delay in echo-free Connections	0	0...500
Number of Quantisation Distortion Units	1	1...14
Equipment Impairment Factor	0	0...40
Packet-loss Robustness Factor	1	1...40
Random Packet-loss Probability	0	0...20
Burst Ratio	1	1 2
Circuit Noise referred to 0 dBr-point	-70	-80...-40
Noise Floor at the Receive Side	-64	
Room Noise at the Send Side	35	35...85
Room Noise at the Receive Side	35	35...85
Advantage Factor	0	0...20

Table 1. Default values and permitted ranges for the E-model's parameters (ITU-T, 2009)

where

R_0	Basic signal-to-noise ratio (groups the effects of noise)
Is	Impairments which occur more or less simultaneously with the voice signal e.g. (quantisation noise, sidetone level)
Id	Impairments due to delay, echo
Ie-eff	Impairments due to codec distortion, packet loss and jitter
A	Advantage factor or expectation factor (e.g. 10 for GSM)

The advantage factor captures the fact that users might be willing to accept some degradation in quality in return for the ease of access, e.g. users may find the speech quality is acceptable in cellular networks because of its access advantages. The same quality would be considered poor in the public circuit-switched telephone network. In the former case A could be assigned the value 10, while in the later case A would take the value 0 (Estepa et al., 2002; Markopoulou et al., 2003).

Each of the parameters in equation (7) except the Advantage factor (A) is further decomposed into a series of equations as defined in ITU-T Recommendation G.107 (ITU-T, 2009). When all parameters set to their default values (Table 1), R-Rating Factor as defined in equation (7) has the value of 93.2 which is mapped to an MOS value of 4.41.

When the effect of delay is considered, the estimated quality according to the E-model is conversational; i.e. MOS - Conversational Quality Estimated MOS_{CQE}. When the effect of delay is ignored and Id is set to its default value the estimation is listening only; i.e. MOS - Listening Quality Estimated MOS_{LQE}.

Packet loss as defined in equation (7) is characterised by packet loss dependent Effective Equipment Impairment Factor (*Ie-eff*), *Ie-eff* is calculated according to the following formula (ITU-T, 2009):

$$Ie\text{-}eff = Ie + (95 - Ie).\frac{Ppl}{\frac{Ppl}{BurstR} + Bpl} \tag{8}$$

where

Ie	Codec-specific Equipment Impairment Factor
Bpl	Codec-specific Packet-loss Robustness Factor
Ppl	Packet loss Probability
$BurstR$	Burst Ratio (BurstR-to count for burstiness in packet loss)

Ie-eff -as defined in equation (8) - is derived using codec-specific values for Ie and Bpl at zero packet-loss. The values for Ie and Bpl for several codecs are listed in ITU-T Recommendation G.113 Appendix I (ITU-T, 2002) and they are derived using subjective MOS test results. For example for the speech coder defined according to the ITU-T Recommendation G.729 (ITU-T, 1996a), the corresponding Ie and Bpl values are 11 and 19 respectively. On the other hand Ppl and $BurstR$ depend on the packet loss presented in the system. BurstR is defined by the latest version of the E-model as (ITU-T, 2009):

$$BurstR = \frac{\text{Average length of observed bursts in an arrival sequence}}{\text{Average length of bursts expected for the network under random loss}} \tag{9}$$

When packet loss is random; i.e., independent, $BurstR$ = 1 and when packet loss is bursty; i.e., dependent, $BurstR > 1$.

The impact of packet loss in older versions of the E-model (prior to the 2005 version) was characterised by Equipment Impairment (Ie) factor. Specific impairment factor values for

codec operating under random packet loss have been previously tabulated to be packet-loss dependent. In the new versions of the E-model (after 2005), Bpl is defined as codec-specific value and Ie is replaced by the $Ie\text{-}eff$.
R-Rating Factor from equation (7) can be mapped into an MOS value. Equation (10) (ITU-T, 2009) gives the mapping function between the computed R-Rating Factor and the MOS value.

$$MOS = \left\{ \begin{array}{cc} 1 & R < 0 \\ 1 + 0.035R + R(R-60)(100-R).7.10^{-6} & 0 < R < 100 \\ 4.5 & R > 100 \end{array} \right\} \quad (10)$$

ITU-T Recommendation G.107 (ITU-T, 2009) also provides a formula to move back to R-Rating Factor from an available MOS score. The equation is:

$$R = \frac{20}{3}\left(8 - \sqrt{226}\left(h + \frac{\pi}{3}\right)\right) \quad (11)$$

with

$$h = \frac{1}{3} atan2\left(18566 - 6750MOS, 15\sqrt{-903522 + 1113960MOS - 202500MOS^2}\right) \quad (12)$$

where

$$atan2(x,y) = \left\{ \begin{array}{cc} atan\left(\frac{x}{y}\right) & for\ x \geq 0 \\ \pi - atan\left(\frac{y}{-x}\right) & for\ x < 0 \end{array} \right\} \quad (13)$$

The calculated R-Rating Factor and the mapped MOS value can be translated into a user satisfaction as defined by ITU-T Recommendation G.109 (ITU-T, 1999) and listed in Table 2. Connections with R values below 50 are not recommended. Understanding the degree of user's needs and expectations and having a direct measurement of user's satisfaction is important for commercial reasons as a network that does not satisfy user's expectations is not expected to be a commercial success. If the quality of the network is continuously low, more percentage of users are expected to look for a an alternative network with a consistent quality.

The E-model is a good choice for non-intrusive estimation of voice quality non-intrusively, but it has some drawbacks. It depends on the time-consuming, expensive and hard to conduct subjective tests to calibrate its parameters (Ie and Bpl), consequently, it is applicable to a limited number of codecs and network conditions (because subjective tests are required to derive model parameters) and this hinders its use in new and emerging applications. Also, it is less accurate than the intrusive methods such as PESQ because it does not consider the contents of the received signal in its calculations which rises questions about

R-Rating factor	**MOS**	**Quality**	**User Satisfaction**
$90 \leq R < 100$	$4.34 \leq MOS < 4.50$	Best	Very Satisfied
$80 \leq R < 90$	$4.02 \leq MOS < 4.34$	High	Satisfied
$70 \leq R < 80$	$3.60 \leq MOS < 4.02$	Medium	Some users dissatisfied
$60 \leq R < 70$	$3.10 \leq MOS < 3.60$	Low	Many users dissatisfied
$50 \leq R < 60$	$2.58 \leq MOS < 3.10$	Poor	Nearly all users dissatisfied

Table 2. User satisfaction as defined by ITU-T Recommendation G.109

its accuracy. Consequently, the E-model as standardised by the ITU-T satisfies only the first two requirements but does not satisfy the other two requirements from the list of desired requirements of speech quality assessment solutions.

Several efforts have been going on to extend the E-model based on the intrusive-based *PESQ* speech quality prediction methodology (Ding & Goubran, 2003a;b; Sun, 2004; Sun & Ifeachor, 2003; 2004; 2006). These studies, despite their importance, but they focused on a previous version of the E-Model (ITU-T, 2000) where burstiness in packet loss was not considered although Internet statistics according to several studies have shown that there is a dependency in packet loss; i.e. when packet loss occurs, it occurs in bursts (Borella et al., 1998; Liang et al., 2001). These and similar studies illustrate the importance of taking burstiness into account. In the current version of the E-model (ITU-T, 2009) burstiness is taken into account.

The authors of this book chapter has avoided these limitations by taking burstiness into consideration in their previous publications as newer versions of the E-model (ITU-T, 2005a; 2009) are used in the extension. Utilising the intrusive-based PESQ solution as a base criterion to avoid the subjectivity in estimating the E-model's parameters, the E-model was extended to new network conditions and applied to new speech codecs without the need for the subjective tests. The extension is realised using several methods, including: linear and nonlinear regression (AL-Akhras, 2007; ALMomani & AL-Akhras, 2008), Genetic Algorithms (AL-Akhras, 2008), Artificial Neural Network (ANN) (AL-Akhras, 2007; AL-Akhras et al., 2009), and Regression and Model Trees (AL-Akhras & el Hindi, 2009). In these implementations the modified E-model calibrated using PESQ is compared with the E-model calibrated using subjective tests to prove their effectiveness.

Another extension implemented by the authors to improve the accuracy of the E-model in comparison with the PESQ, analyses the content of the received degraded signal and classifies packet loss into either Voiced or Unvoiced based on the received surrounding packets. An emphasis on perceptual effect of different types of loss on the perceived speech quality is drawn. The accuracy of the proposed method is evaluated by comparing the estimation of the new method that takes packet class into consideration with the measurement provided by PESQ as a more accurate, intrusive method for measuring the speech quality (AL-Akhras, 2007).

The above two extensions for quality estimation of the E-model were combined to offer a complete solution for estimating the quality of VoIP applications objectively, non-intrusively, and accurately without the need for the time-consuming, expensive, and hard to conduct subjective tests (AL-Akhras, 2007). In other words a solution that satisfies all the requirements for a good VoIP speech quality assessment solution. Complete details about these extensions can be found and downloaded (AL-Akhras, 2007).

4.2.3 Other methods

Wide range of non-intrusive methods for non-intrusive VoIP quality assessment have been proposed, next reference to some attempts are mentioned, including: (Kim & Tarraf, 2006; Raja et al., 2006; Raja & Flanagan, 2008; Sun, 2004; Sun & Ifeachor, 2002; AL-Khawaldeh, 2010; Picovici & Mahdi, 2004; Mohamed et al., 2004; Da Silva et al., 2008). Many other attempts can be found in (AL-Akhras, 2007; AL-Khawaldeh, 2010).

5. Relationship among different subjective and objective assessment techniques

To avoid ambiguity, different qualifiers used to distinguish among different quality measurement methods are presented. Careful selection of terminology is used and

differentiation among different terms used to describe the quality is clearly stated. A qualifier is added to the terms used to make sure of no vagueness in the meaning of the term. ITU-T Recommendation P.800.1 (ITU-T, 2006) gives a clear terminology distinction among different MOS terms whether the test is listening or conversational and whether it a result of subjective or objective test by adding an appropriate qualifier. This section shows how different quantifiers are obtained and how they are related to each other. In the recommendation it is stated that the identifiers in the following Table are to be used:

LQ	Listening Quality
CQ	Conversational Quality
S	Subjective
O	Objective
E	Estimated

Table 3. MOS Qualifiers

It is recommended to use these identifiers together with the MOS to avoid confusion and distinguish the area of application. The result of such qualification is (ITU-T, 1996b; 2001; 2004; 2006; 2009):

- **Subjective Tests**
  - **Listening Quality:** For the score collected by calculating the arithmetic mean of listening subjective tests conducted according to Recommendation P.800, the results are qualified as MOS - Listening Quality Subjective or MOS_{LQS}.
  - **Conversational Quality:** For the score collected by calculating the arithmetic mean of conversational subjective tests conducted according to Recommendation P.800, the results are qualified as MOS - Conversational Quality Subjective or MOS_{CQS}.
- **Network Planning Estimation Tests**
  - **Listening Quality:** For the score calculated by a network planning tool to estimate the listening quality according to Recommendation G.107 and then transformed into MOS, the results are qualified as MOS - Listening Quality Estimated or MOS_{LQE}.
  - **Conversational Quality:** For the score calculated by a network planning tool to estimate the conversational quality according to Recommendation G.107 and then transformed into MOS, the results are qualified as MOS - Conversational Quality Estimated or MOS_{CQE}.
- **Objective Tests**
  - **Listening Quality:** For the score calculated by an objective model to predict the listening quality according to Recommendation P.862 and then transformed into MOS, the results are qualified as MOS - Listening Quality Objective or MOS_{LQO}.
  - **Conversational Quality:** For the score calculated by an objective model to predict the conversational quality according to Recommendation P.563 and then transformed into MOS, the results are qualified as MOS - Conversational Quality Objective or MOS_{CQO}.

The relation between different listening MOS qualifiers is depicted in Figure 8 where the related speech signal and the MOS from the subjective tests, PESQ and the E-model are related together.

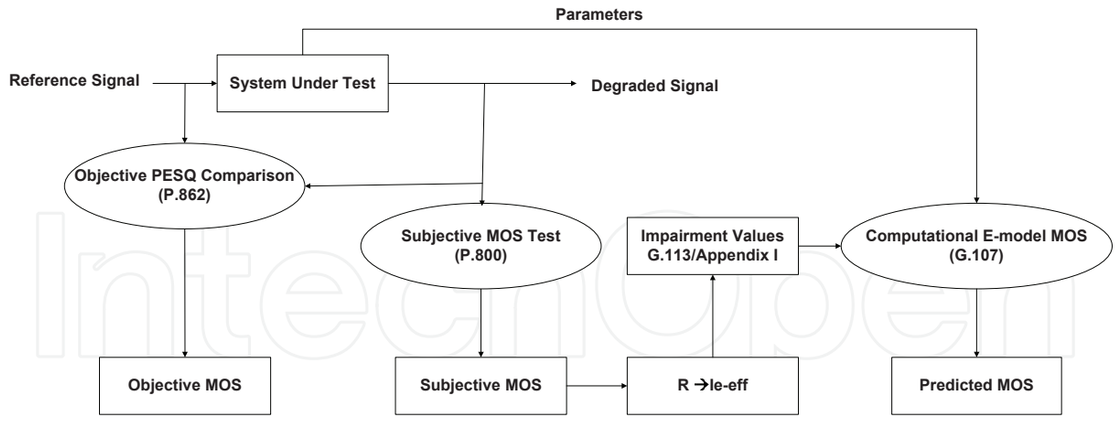


Fig. 8. Relationship between MOS qualifiers (ITU-T, 2006)

6. Conclusions and future work

Measuring the quality of VoIP is important for legal, commercial and technical reasons. This chapter presented the requirements for a successful VoIP quality assessment technology. The chapter also critically reviewed different VoIP quality assessment technologies. Sections 3 and 4 discussed subjective and objective speech quality measurement methods, respectively. In objective measurement methods both intrusive (section 4.1) and non-intrusive (section 4.2) methods were discussed.

Based on the requirements of measuring the speech quality non-intrusively and objectively, it can be concluded that objective and non-intrusive methods such as P.563 and the E-Model are the best methods for VoIP quality assessment. Still the accuracy of these methods can be improved to make their estimation of the quality as accurate as possible.

7. References

AL-Akhras, M. (2007). *Quality of Media Traffic over Lossy Internet Protocol Networks: Measurement and Improvement*, PhD thesis, Software Technology Research Laboratory (STRL), School of Computing, Faculty of Computing Sciences and Engineering, De Montfort University, U.K.
URL: *http://www.tech.dmu.ac.uk/STRL/research/theses/thesis/40-thesis-mousa-secure.pdf*

AL-Akhras, M. (2008). A genetic algorithm approach for voice quality prediction, *The 5th IEEE International Multi-Conference on Systems, Signals & Devices, 2008. IEEE SSD' 08, Amman, Jordan* pp. 1–6.

AL-Akhras, M. & el Hindi, K. (2009). Function approximation models for non-intrusive prediction of voip quality, *IADIS International Conference Informatics 2009, Algarve, Portugal* .

AL-Akhras, M., Zedan, H., John, R. & ALMomani, I. (2009). Non-intrusive speech quality prediction in voip networks using a neural network approach, *Neurocomputing* 72(10-12): 2595 – 2608. Lattice Computing and Natural Computing (JCIS 2007) / Neural Networks in Intelligent Systems Designn (ISDA 2007).

AL-Khawaldeh, R. (2010). *Ant colony optimization for voip quality optimization*, Master's thesis, Computer Information Systems Department, King Abdullah II School for Information Technology (KASIT), The University of Jordan, Jordan.

Allnatt, J. (1975). Subjective Rating and Apparent Magnitude, *International Journal Man -Machine Studies* 7: 801–816.

ALMomani, I. & AL-Akhras, M. (2008). Statistical speech quality prediction in voip networks, *The 2008 International Conference on Communications in Computing (CIC'8), Las Vigas* .

Borella, M., Swider, D., Uludag, S. & Brewster, G. (1998). Internet Packet Loss: Measurement and Implications for End-to-End QoS, *Architectural and OS Support for Multimedia Applications/Flexible Communication Systems/Wireless Networks and Mobile Computing: Proceedings of the 1998 ICPP Workshops on*, pp. 3–12.

Bos, L. & Leroy, S. (2001). Toward an All-IP-Based UMTS System Architecture, *IEEE Network* 15(1): 36–45.

Collins, D. (2003). *Carrier Grade Voice over IP*, 2nd edn, McGraw-Hill Companies.

Da Silva, A., Varela, M., de Souza e Silva, E., Rosa, L. & G.Rubino, G. (2008). Quality assessment of interaction voice applications, ***Computer Networks*** 52(6): 1179–1192.

Ding, L. & Goubran, R. (2003a). Assessment of Effects of Packet Loss on Speech Quality in VoIP, *Proceedings. of the 2nd IEEE Internatioal Workshop on Haptic, Audio and Visual Environments and their Applications, 2003. HAVE 2003*, pp. 49–54.

Ding, L. & Goubran, R. (2003b). Speech Quality Prediction in VoIP Using the Extended E-Model, *IEEE Global Telecommunications Conference, 2003. GLOBECOM '03.*, Vol. 7, pp. 3974–3978.

Duysburgh, B., Vanhastel, S., De Vreese, B., Petrisor, C. & Demeester, P. (2001). On the Influence of Best-Effort Network Conditions on the Perceived Speech Quality of VoIP Connections, *Proceedings. Tenth International Conference on Computer Communications and Networks, 2001.*, pp. 334–339.

Estepa, A., Estepa, R. & Vozmediano, J. (2002). On the Suitability of the E-Model to VoIP Networks, *Proceedings of Seventh International Symposium on Computers and Communications,2002. ISCC 2002.*, pp. 511–516.

ETSI (1996). ETSI Tech. Report (ETR) 250 - Speech Communication Quality from Mouth to Ear of 3.1 kHz Handset Telephony Across Networks, *Technical report*, European Telecommunications Standards Institute.

Fu, Q., Yi, K. & Sun, M. (2000). Speech Quality Objective Assessment Using Neural Network, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP '00.*, Vol. 3, pp. 1511–1514.

Haojun, A., Xinchen, Z., Ruimin, H. & Weiping, T. (2004). A Wideband Speech Codecs Quality Measure Based on Bark Spectrum Distance, *Proceedings of 2004 International Symposium on Intelligent Signal Processing and Communication Systems, 2004. ISPACS 2004.*, pp. 155–158.

Heiman, F. (1998). A Wireless LAN Voice over IP Telephone System, *Northcon/98 Conference Proceedings*, pp. 52–54.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing* 23(1): 67 – 72.

Itakura, F. & Saito, S. (1978). Analysis synthesis telephony based on the maximum likelihood method, *Acoustics, Speech and Signal Processing* pp. C17–C20.

ITU-T (1996a). *Recommendation G.729 - Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (1996b). *Recommendation P.800 - Methods for Subjective Determination of*

Transmission Quality, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (1998). *Recommendation P.861 - Objective Quality Measurement of Telephoneband (300-3400 Hz) Speech Codecs*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (1999). *Recommendation G.109 - Definition of Categories of Speech Transmission Quality*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2000). *Recommendation G.107 - The E-model, a Computational Model for use in Transmission Planning*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2001). *Recommendation P.862 - Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2002). *Recommendation G.113 Appendix I - Provisional Planning Values for the Equipment Impairment Factor Ie and Packet-Loss Robustness Factor Bpl*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2003a). *Recommendation G.114 - One-Way Transmission Time*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2003b). *Recommendation G.114 Appendix II - Guidance on One-Way Delay for Voice over IP*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2004). *Recommendation P.563 - Single-ended method for objective speech quality assessment in narrow-band telephony applications*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2005a). *Recommendation G.107 - The E-model, a Computational Model for use in Transmission Planning*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2005b). *Recommendation P.862.1-Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2005c). *Recommendation P.862.2-Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2006). *Recommendation P.800.1 - Mean Opinion Score (MOS) Terminology*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

ITU-T (2009). *Recommendation G.107 - The E-model, a Computational Model for use in Transmission Planning*, International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

Kim, D.-S. & Tarraf, A. (2006). Enhanced Perceptual Model for Non-Intrusive Speech Quality Assessment, *IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006*, Vol. 1, pp. I–I.

Kitawaki, N., Nagabuchi, H. & Itoh, K. (1988). Objective quality evaluation for low-bit-rate speech coding systems, *IEEE Journal on Selected Areas in Communications* 6(2): 242–248.

Kondoz, A. M. (2004). *Digital Speech Coding for Low Bit Rate Communication Systems*, 2nd edn, John Wiley and Sons Ltd, New York, NY, USA.

Li, F. (2004). Speech Intelligibility of VoIP to PSTN Interworking - A Key Index for the QoS, *IEE Telecommunications Quality of Services: The Business of Success, 2004. QoS 2004.*, pp. 104–108.

Liang, Y., Steinbach, E. & Girod, B. (2001). Multi-stream Voice over IP Using Packet Path Diversity, *IEEE Fourth Workshop on Multimedia Signal Processing, 2001*, pp. 555–560.

Low, C. (1996). The Internet Telephony Red Herring, *IEEE Global Telecommunications Conference, 1996. GLOBECOM '96.*, pp. 72–80.

Mahdi and Picoviciv (2009). Advances in voice quality measurement in modern telecommunications, *Digital Signal Processing* 19: 79–103.

Markopoulou, A., Tobagi, F. & Karam, M. (2003). Assessing the Quality of Voice Communications over Internet Backbones, *IEEE/ACM Transactions on Networking* 11(5): 747–760.

Mase, K. (2004). Toward Scalable Admission Control for VoIP Networks, *IEEE Communications Magazine* 42(7): 42–47.

Miloslavski, A., Antonov, V., Yegoshin, L., Shkrabov, S., Boyle, J., Pogosyants, G. & Anisimov, N. (2001). Third-party Call Control in VoIP Networks for Call Center Applications, *2001 IEEE Intelligent Network Workshop*, pp. 161–167.

Mohamed, S., Rubino, G. & Varela, M. (2004). Performance Evaluation of Real-Time Speech Through a Packet Network: A Random Neural Networks-Based Approach, *Performance Evaluation* 57(2): 141–161.

Moon, Y., Leung, C., Yuen, K., Ho, H. & Yu, X. (2000). A CRM Model Based on Voice over IP, *2000 Canadian Conference on Electrical and Computer Engineering*, Vol. 1, pp. 464–468.

Narbutt, M. & Murphy, L. (2004). Improving Voice over IP Subjective Call Quality, *IEEE Communications Letters* 8(5): 308–310.

Ortiz, S., J. (2004). Internet Telephony Jumps off the Wires, *Computer* 37(12): 16–19.

Picovici, D. & Mahdi, A. (2004). New Output-based Perceptual Measure for Predicting Subjective Quality of Speech, *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (ICASSP '04)*, Vol. 5, pp. V–633–6.

Quackenbush, S., Barnawell, T. & Clements, M. (1988). *Objective Measures of Speech Quality*, Prentice Hall, Englewood Cliffs, NJ.

Raja, A., Azad, R. M. A., Flanagan, C., Picovici, D. & Ryan, C. (2006). Non-Intrusive Quality Evaluation of VoIP Using Genetic Programming, *1st Bio-Inspired Models of Network, Information and Computing Systems, 2006.*, pp. 1–8.

Raja, A. & Flanagan, C. (2008). *Genetic Programming*, chapter Real-Time, Non-intrusive Speech Quality Estimation: A Signal-Based Model, pp. 37–48.

Rix, A., Beerends, J., Hollier, M. & Hekstra, A. (2001). Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs, *Proceedings. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. (ICASSP '01)*, Vol. 2, pp. 749–752.

Rohani, B. & Zepernick, H.-J. (2005). An Efficient Method for Perceptual Evaluation of Speech Quality in UMTS, *Proceedings Systems Communications, 2005.*, pp. 185–190.

Rosenberg, J., Lennox, J. & Schulzrinne, H. (1999). Programming Internet Telephony Services, *IEEE Network* 13(3): 42–49.

Schulzrinne, H. & Rosenberg, J. (1999). The IETF Internet Telephony Architecture and Protocols, *IEEE Network* 13(3): 18–23.

Spanias, A. (1994). Speech Coding: A Tutorial Review, *Proceedings of the IEEE* 82(10): 1541–1582.

Sun, L. (2004). *Speech Quality Prediction for Voice over Internet Protocol Networks*, PhD thesis, School of Computing, Communications and Electronics, University of Plymouth, U.K.

Sun, L. & Ifeachor, E. (2002). Perceived Speech Quality Prediction for Voice over IP-Based Networks, *IEEE International Conference on Communications, 2002. ICC 2002.*, Vol. 4, pp. 2573–2577.

Sun, L. & Ifeachor, E. (2003). Prediction of Perceived Conversational Speech Quality and Effects of Playout Buffer Algorithms, *IEEE International Conference on Communications, 2003. ICC '03.*, Vol. 1, pp. 1–6.

Sun, L. & Ifeachor, E. (2004). New Models for Perceived Voice Quality Prediction and their Applications in Playout Buffer Optimization for VoIP Networks, *IEEE International Conference on Communications, 2004*, Vol. 3, pp. 1478–1483.

Sun, L. & Ifeachor, E. (2006). Voice Quality Prediction Models and their Application in VoIP Networks, *IEEE Transactions on Multimedia* 8(4): 809–820.

Takahashi, A. (2004). Opinion Model for Estimating Conversational Quality of VoIP, *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04).*, Vol. 3, pp. iii–1072–5.

Takahashi, A., Yoshino, H. & Kitawaki, N. (2004). Perceptual QoS Assessment Technologies for VoIP, *IEEE Communications Magazine* 42(7): 28–34.

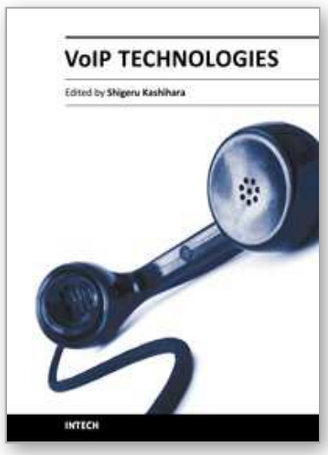
Tseng, K.-K., Lai, Y.-C. & Lin, Y.-D. (2004). Perceptual Codec and Interaction Aware Playout Algorithms and Quality Measurement for VoIP Systems, *IEEE Transactions on Consumer Electronics* 50(1): 297–305.

Tseng, K.-K. & Lin, Y.-D. (2003). User Perceived Codec and Duplex Aware Playout Algorithms and LMOS-DMOS Measurement for Real Time Streams, *International Conference on Communication Technology Proceedings, 2003. ICCT 2003.*, Vol. 2, pp. 1666–1669.

Voran, S. (1999a). Objective Estimation of Perceived Speech Quality-Part I: Development of the Measuring Normalizing Block Technique, *IEEE Transactions on Speech and Audio Processing* 7(4): 371–382.

Voran, S. (1999b). Objective Estimation of Perceived Speech Quality-Part II: Evaluation of the Measuring Normalizing Block Technique, *IEEE Transactions on Speech and Audio Processing* 7(4): 383–390.

Zurek, E., Leffew, J. & Moreno, W. (2002). Objective Evaluation of Voice Clarity Measurements for VoIP Compression Algorithms, *Proceedings of the Fourth IEEE International Caracas Conference on Devices, Circuits and Systems, 2002.*, pp. T033–1–T033–6.



VoIP Technologies

Edited by Dr Shigeru Kashihara

ISBN 978-953-307-549-5
Hard cover, 336 pages
Publisher InTech
Published online 14, February, 2011
Published in print edition February, 2011

This book provides a collection of 15 excellent studies of Voice over IP (VoIP) technologies. While VoIP is undoubtedly a powerful and innovative communication tool for everyone, voice communication over the Internet is inherently less reliable than the public switched telephone network, because the Internet functions as a best-effort network without Quality of Service guarantee and voice data cannot be retransmitted. This book introduces research strategies that address various issues with the aim of enhancing VoIP quality. We hope that you will enjoy reading these diverse studies, and that the book will provide you with a lot of useful information about current VoIP technology research.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen