# Overview on VoIP: Subjective and Objective Measurement Methods

Floriano De Rango, Mauro Tropea, Peppino Fazio, Salvatore Marano
D.E.I.S. Department, University of Calabria, Italy, 87036
e-mail: {derango, mtropea, pfazio, marano}@deis.unical.it

**Summary**
In this paper we have carried out an accurate overview on Voice over IP QoS evaluation techniques. We have described in a detailed way the most important measurement methods that are subdivided in subjective and objective methods. Subjective measurements (e.g. Mean Opinion Score - MOS) are the benchmark for objective methods, but they are slow, time consuming and expensive. Objective measurements can be intrusive or non-intrusive. Intrusive methods (e.g. Perceptual Evaluation of Speech Quality - PESQ) are more accurate, but normally are unsuitable for monitoring live traffic because of the need for a reference data and to utilise the network. Non-intrusive methods, instead, are a measurement based on observing parameters that permit the individuation of voice signal quality. The possibility of having the telephony on Internet presents a lot of advantages, first of all it gives the capability of constructing a single network for transporting voice and data signals. Moreover, it provides the growth of new type of applications that can take advantage of internet network. At last, IP traffic charge is much less than traditional one because it is independent of the path length.

***Key words:***
VoIP, Subjective measurement method, MOS, Objective measurement method, PESQ, E-Model.

## 1. Introduction

Recently, there has been a considerable growth in telephone system: the analog telephone system has been first replaced by the digital network, based on the circuit-switching; then, in the last few years, telephone communication system is living another important revolution, that led to the packet-switching transfer mode: this allows the packets not only to bring data traffic, but also voice traffic over IP networks; this new technology permits the integration of telephone system and IP network and it is called Voice over IP (VoIP).
The VoIP technology allows real time transmission of voice signals, through packets over IP: first, voice information are digitized and, then, transmitted as a packets stream; in this way, information can reach the destination, following the best path on the net and leading to a better use of the resources.
As known, in a packet-switching net, the information transmitted by a single sender can follow different paths to reach the destination, so they can arrive in different times with a different order from the one in the transmission phase; in the worst case, some information can also be lost. At the destination node, information are reassembled in the correct order.
In this work, an overview on VoIP is given, with particular interest to the parameters that characterize the quality of the offered service (Quality of Service, QoS); in section 2 the VoIP related protocols are shown; A/D voice conversion is treated in section 3, while section 4 gives a detailed overview on the measurement techniques used in VoIP network and section 5 concludes the paper.

## 2. VoIP Related Protocols

The VoIP standardization is governed by two main organizations: ITU-T and IETF; they have produced many types of protocol that can be used for utilising VoIP technique. Fig. 1 shows a typical network architecture which need a connection with the traditional PSTN network [1,2].

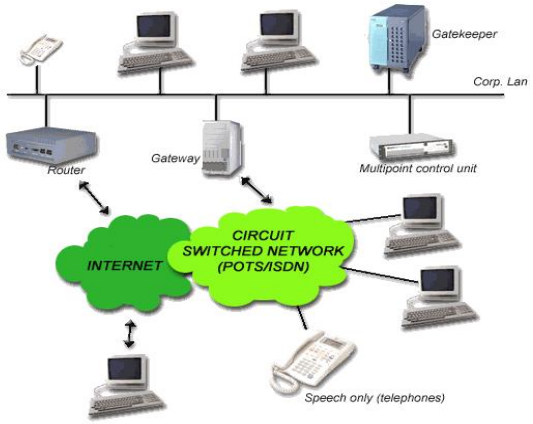


Fig. 1 VoIP network architecture.

These protocols are much important because they are used for initializing the IP voice call. The main protocols are fundamentally of two types, one called peer-to-peer like and second master/slave like. The H323 and SIP protocol belong to first group [3,4], instead MEGACO and MGCP are two important protocol that belong to second group [5-8]. In this work the signalling protocols are not shown and therefore for details to [3-14].

## 3. A/D Voice Conversion

In the VoIP scenario, human voice must be coded in a bit stream, before it can be introduced in the network; voice streams must have homogeneous and compatible structure, as in [15]. The human voice can be considered as a signal with the following characteristics:

a) analog structure: voice signals are functions of time and they assume continuous values, in opposition to digital waveforms, that assume discrete values;
b) limited bandwidth: voice signals occupy a bandwidth that is approximately 4kHz;
c) slowly time-variant: voice signals can be considered stationary in relatively short time intervals (from 5ms up to 100ms).

Human voice must be sampled, quantized and coded before it can be transmitted.

### 3.1 Sampling Phase

This phase permits to transform a continue-time signal to a discrete-time one, with the help of the sampling theorem [15], that can be applied because of the limited bandwidth of voice signals. A waveform *w(t)* is band-limited at B-Hertz, if:

$$W(f)=F[w(t)]=0 \qquad if\ |f|>B \qquad (1)$$

Where, in Eq. 1, *W(f)* is the Fourier transform of *w(t)*.
Since a band-limited function does not have spectral components for frequencies higher than *B*, it cannot vary too rapidly in the time domain. From the sampling theorem, it is clear that a real band-limited waveform *w(t)* can be completely specified by $N=2BT_0$ independent samples, where *B* is the absolute bandwidth and $T_0$ is the considered time interval. So, the lower bound of the required bandwidth to transmit a source signal over an analog communication system is $B=N/2T_0$; the required number of samples to rebuild a signal on a $T_0$ period of time is *N*.

### 3.2 Quantization Phase

This is the second phase of the A/D conversion; the signal is now converted in a discrete-time signal with discrete-values: each sample of the previous phase is transformed, by approximation, into a newer one, chosen from a finite set, with a power of 2cardinality. So, this phase is necessary, because of the limitation of digits that can represent a sample value (usually the used digits are binary). The difference between the original sample value and the quantized one is defined as quantization error (or quantization noise) and, for a better quality of the digital conversion, it must be reduced.

### 3.3 Coding Phase

This is the last phase of the A/D conversion: once the analog signal has been converted in the discrete-time one, with discrete-finite values, it must be coded in order to transmit it on a transmission medium. The choice of the coding scheme is quite important in order to determine the transmission bandwidth. In the following, different audio codecs are treated.

#### 3.3.1 Pulse Code Modulation (PCM)

The Pulse Code Modulation (PCM) [16] is the most used coding scheme for telephony applications, standardized by the ITU-T institute with Rec. G.711. The adopted sampling frequency is *8kHz* and samples are quantized with 8, 12 or 16 bits. In the Nord-America, for the 8-bits quantization the "$\mu$-*law*" compression law with 15 segments and $\mu$=225 is used, while in Europe the "*A-law*" compression law is used, with 13 segments and *A*=87.6. It is an A/D conversion with a logarithmic law with a different number of bits: the most used employs 8 bits (with a rate of 64kbps), but 7 bits can also be used (with a rate of 56kbps) and so on.

#### 3.3.2 Differential PCM (DPCM)

In the audio and video signals sampling, it is often observed that adjacent samples assume similar values; it means that there can be a lot of redundancy in the signal samples and, so, a wastage of bandwidth when the same values are transmitted. A possible solution to minimize the transmission redundancy is to transmit a PCM signal not directly related to the samples values, indeed to the difference between adjacent samples values; this coding method is called Differential Pulse Code Modulation [17]. On the receiver side, the signal is rebuild by adding the differential value to the previous one; DPCM offers, for voice signals, the same performances in terms of signal-to-noise ratio of a PCM system, with a 3 or 4 bits of higher resolution. For this reason, DPCM systems use a bit-rate from 24kbps up to 32kbps, instead of the 64kbps of the standard PCM. For example, if 8 bits are necessary to

transmit a sample, only 5 or 6 bits are necessary to transmit the difference from the previous one.

### 3.3.3 Adaptive DPCM (ADPCM)

In this coding scheme [17] a different approach is adopted in order to reduce the number of needed bits, adapting the quantization level to the instantaneous power of the audio signal; in other words, the amplitude of the quantization intervals is dynamically changed, but the number of quantization levels is fixed. The ITU-T Rec. G.722 is an improvement of the previous G.711. The specifications consider audio signals with 7kHz of bandwidth and a rate of 64kbps. The coding scheme SB-ADPCM (sub-band ADPCM) works with a sampling frequency of 16kHz and each sample is represented with 14 or 16 bits, with rates of 64, 56 and 48kbps.
Another coding scheme derives from the combination of the previous ones and it is standardized by ITU-T with Rec. G.726; the performance can be compared with those of the PCM scheme, with a variable bit-rate of 40, 32, 24 and 16kbps. Other codecs are based on this scheme, with different sampling frequencies, as the ITU-T Rec. G.722, with ADPCM coding for 7kHz of bandwidth signals and 63kbps.

### 3.3.4 Linear Predicting Coding (LPC)

LPC coding drastically reduces the transmission rate, but the quality of the signal is heavy affected: the most used LPC scheme has a bit-rate of 4.8kbps with a higher grade of compression than other standard codecs, allowing to contain only a percentage of 60% of the original signal [18]. This codec processes the input signal, obtaining a set of parameters that describe the signal; these parameters are then send to the decoder, that uses them to generate a synthesized voice signal, similar to the original one (it can be heard as an artificial voice generated by a machine).

### 3.3.5 Codebook Excited Linear Prediction (CELP)

In these compression schemes, the input signal is subdivided in blocks of samples (vectors), that are treated as an unique unit [19]. The coding is made by using an analysis and synthesizing algorithm, with a weighted quantization of vectors, with linear prediction. The transmitter analyzes the voice signal, comparing it with different models, so for each voice component, it sends a certain code that corresponds to the appropriate model, in addition to some information that can describe the variations of the real voice in comparison to the chosen model. The receiver combines the model with these additional information, then synthesizes the obtained voice signal. A good CELP codec can reproduce a voice signal with the same quality of a PCM signal with 64kbps, using only 16kbps. In addition, it can use a lower bandwidth, generating an artificial sound; the most used CELP codec in VoIP applications are named Low Delay CELP or LD-CELP and they are described in the ITU-T Rec. G.728 [20].

### 3.3.6 Adaptive Multi-Rate (AMR)

AMR speech codec was developed by ETSI and it has been standardized for GSM. It has been chosen by 3GPP as the mandatory codec. The AMR is a multi-mode codec with 8 narrow band modes with bit rates of 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2 and 12.2 Kb/s. Mode switching can occur at any time (frame-based) [21]. AMR speech codec represents a new generation of coding algorithms which are developed to work with inaccurate transport channels. The flexibility on bandwidth requirements and the tolerance in bit errors of AMR codecs are not only beneficial for wireless links, but are also desirable for VoIP applications.

## 4. Qos Measurement Methods

The Quality of Service (QoS) term indicates the capacity of a network to offer some services, guaranteeing the characteristics needed to obtain the optimal service. The most important problem in VoIP applications regards the possibility to offer to customers a service that can be compared with the one of the traditional telephony system [2,22,23]. Differently from the PSTN, where an end-to-end connection is established when originating a call, packets networks use the “statistical multiplexing” of net resources. Although the sharing of net resources between a multitude of users offers restricted and contained costs (that is the first prerogative of VoIP traffic), the use of shared resources on the networks heavily affects the QoS received by users.

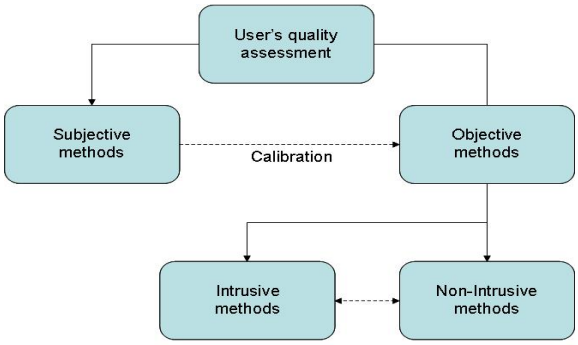


Fig. 2 Overview of VoIP measurement methods.

## 4.1 QoS Parameters

Different parameters are used to describe the service offered by a considered system, as delay, jitter, packet loss, echo and call setup delay. Here they are described in order to have an idea on how they must be considered to give an evaluation of the quality of service:

- **Delay.** Delay effects represent a primary parameter in the project of a VoIP net; they are treated in the ITU-T Rec. G.114 [24]. Before describing the delay effects, it is useful to know what cause a delay:
  a) *Algorithmic delay*: it is introduced by the audio codec of the used coding algorithm; Table 1 resumes the delay introduced by every codec;

Table 1: Codecs delay resume

| ***Coding Standards*** | ***Algorithmic Delay (ms)*** |
|---|---|
| G.711 | 0.125 |
| G.726 | 1 |
| G.728 | 3-5 |
| G.729 | 15 |
| G.723.1 | 37,5 |

  b) *Packetization Delay*; in the RTP [25], voice samples are often crowded in order to reduce the overheads; the RTP standard previews that the packetization time must be 20ms, so the G.711 codec can crowded 160 voice samples in every frame, while with the G.723.1 [26] codec a single frame is generated every 30ms and only one voice frame is transmitted in a single RTP packet;
  c) *Serialization Delay*: it is the necessary time to transmit IP packets; for example, by using the G.711 codec with a packetization time of 20ms (that correspond to a RTP payload of 160bytes), the entire frame can be of 206bytes. In order to transmit the frame 1.1ms are needed on a T1 line, 3.2ms in a 512kbps line and 25.8ms in a 64kbps line. This type of delay can be affected by the number of routers or switches that a packet must traverse on the path (for example, crossing of 10 routers causes a delay of 10ms);
  d) *Propagation Delay:* it represents the time necessary to cross networks that are very distant between them. An example can be the voice traffic in satellites networks, where the delay is 100ms for a satellite placed in orbit at an altitude of 14000km and it is 260ms for an altitude of 36000km;
  e) *Component Delay:* it is caused by the different components of the network: for example, the crossing of the packet through a router from the input interface to the output interface; it is negligible if compared with the previous ones.
- **Jitter.** When a frame is transmitted over an IP network, the delay may change, because it depends on the times of arrival of single packets that are related to the load of the system; an overload situation does not permit the right receiving of the packets in the bounded time intervals; there are many methods to decrease such type of delay, like the use of long buffers, that permit to slower frames to arrive in the correct order.
- **Packet Loss.** Voice packets that cross the IP network, may be lost [23], because of a possible high level of congestion; in a real time voice transmission, the re-transmission is not allowed, because the delay would be higher; the effects of the lost packets on the quality of service depend on their terminal management: a terminal can leave an empty gap in order to substitute lost frames. In order to better face this situation, many methods can be employed, like the retransmission of the previous correctly received samples or the prediction of the lost sample (this method is known as Packet Loss Concealment, PLC).
- **Echo.** This effect, that consists of hearing a delayed repetition of the voice signal, can be encountered if there are more than 50ms of round-trip delay; in order to avoid this undesired effect, codecs must implement some echo-erasers, that permit to preserve for a while the voice signal that will be subtracted from the echoed signal.

## 4.2 Subjective Measurement Methods

Subjective measurements of the QoS are carried out by a group of people (test subjects) [27]; a test phrase is recorded and, then, test subjects listen to it in different conditions. These tests are performed in special rooms, with background noises and other environment factors, that are kept under control for test executions. Some examples are: conversational opinion test, listening opinion test, interview and survey test; but there are also some disadvantages for this kind of tests, like the experience, the humour and the culture of test subjects; these tests are very expensive and they are not used in practice, because of the high number of persons and runs that must be employed to obtain truthful results.

### 4.2.1 Listening Tests

This kind of test is used for unidirectional transmissions and it is based on a transmission test carried out with conversations or pre-recorded phrases; the goal of this test is the evaluation of single performances of terminals and

algorithms under different conditions; some of the well known listening tests are: ACR, DCR and CCR.

**Absolute Category Rating (ACR)**

This type of test is used to obtain the absolute quality of the voice sample, through the direct hearing of the sample, without a reference sample; the results are represented by numerical values, under the subjective rating system Mean Opinion Score (MOS) [28]. The MOS represents the average value of the numerical values associated to the evaluation (numerical values go from 1 to 5). The MOS gives an indication of the mean impression of the present test persons; in order to obtain a truthful result, many persons must be present, because the single opinions may be widely vary. The ITU-T Rec. P.800 (1996) [28] shows how a MOS test must be carried out:

- source recordings:
  - the test room must be a volume of $30m^3$ up to $120m^3$, with an echo duration lower than 500ms (200÷300ms is preferred) and a background noise lower than 30dB;
  - all receiving systems, local phones or an Intermediate Reference Systems (IRS, ITU-T Rec. P.48), must be calibrated following the specifications of ITU-T Rec. P.64 [29] and the sensibility test of the system must be performed at the beginning and at the end of test;
  - the recording system must be of high quality (2 ways tape, 2 digital channels audio microprocessor or computer-driven recording system);
  - recorder voice signals must consist of few short phrases, taken from papers or non-technical lectures, casually 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÷200mm from speaker's mouth.
- listening test procedure:
  - the test room must satisfy the same conditions of recording rooms, except from the background noise, that must be lower than 50dB; also in this case, noise spectrums must be evaluated two times;
  - all reproduction systems, local phone, IRS or loudspeakers, must be calibrated following the specifications of ITU-T Rec. P.64 [29];
  - Hearing persons must not be interested in telephone works and they must have never heard the phrases used in the tests;
  - the quality must be evaluated through different opinion scales (like illustrated in Table. 2-3-4):
    - Listening quality scale (MOS):

Table 2: Listening quality scale

| ***Quality of the speech*** | Score |
|---|---|
| Excellent | 5 |
| Good | 4 |
| Fair | 3 |
| Poor | 2 |
| Bad | 1 |

- Listening-effort scale ($MOS_{LE}$):

Table 3: Listening effort scale

| ***Effort required to understand the meanings of sentences*** | Score |
|---|---|
| Complete relaxation possible: no effort required | 5 |
| Attention necessary: no appreciable effort required | 4 |
| Moderate effort required | 3 |
| Considerable effort required | 2 |
| No meaning understood with any feasible effort | 1 |

- Loudness-preference scale ($MOS_{LP}$):

Table 4 Loudness-preference scale

| ***Loudness preference*** | Score |
|---|---|
| Much louder than preferred | 5 |
| Louder than preferred | 4 |
| Preferred | 3 |
| Quieter than preferred | 2 |
| Much quieter than preferred | 1 |

**Degradation Category Rating (DCR)**

The Degradation Category Rating is used when there high quality are voice samples and the ACR results inappropriate to discover quality variations. In this listening test two samples (A and B) are present: A represents the reference sample with the reference quality, while B represents the degraded sample; listeners must compare the B sample with a degradation scale and the results are summarized as Degraded MOS (Degraded Mean Opinion Score). The ITU-T Rec. P.800 (1996) Annex D [28], shows how the DCR test must be carried out. Each configuration is evaluated by almost 4 talkers; the samples must be composed of two periods, separated by silence (for example 0.5 seconds). DCR is different from ACR for the types of used samples; at this point, subjects use a degradation scale, composed of five points to indicate the degradation of sample B, referring to sample A (as shown in Table 5).

Table 5 Degradation category scale

| | |
|---|---|
| 5 | Degradation is inaudible |
| 4 | Degradation is audible but not annoying |
| 3 | Degradation is slightly annoying |
| 2 | Degradation is annoying |
| 1 | Degradation is very annoying |

**Comparison Category Rating (CCR)**

The Comparison Category Rating (described in the ITU-T Rec. P.800 Annex D) is similar to DCR, except for the type of used samples. In the DCR procedure, the reference sample is first presented, followed by the degradated sample, while in CCR the order of samples presentation is casually chosen for every iteration; listeners do not know when the reference or degradated sample is transmitted, so they must give an opinion on the quality of the second heard sample in respect of the first one. Table 6 shows the scale used in CCR.

Table 6 Opinion Scale for Comparison Category Rating

| | |
|---|---|
| 3 | Much better |
| 2 | Better |
| 1 | Slightly better |
| 0 | About the same |
| -1 | Slightly worse |
| -2 | Worse |
| -3 | Much worse |

### 4.2.2 Conversational opinion tests

Conversational opinion tests are laboratory tests that aim to reproduce the real conditions experimented by customers. It is important that the simulating conditions in the test are correctly specified, reproduced and accurately measured before and after experiments. Conversational opinion test s description is defined by the ITU-T Rec. P800 Annex A [28]:

- Two subjects, that perform the test, are placed in separated and isolated rooms with a not inferior volume of 20 $m^3$ and a echo lower than 500 ms (comprised between 200 and 300 ms). The room construction would have to be such to allow that performance of sufficient sound in order to simulate the external atmosphere, with a level of noise (when there are not noises intentionally introduced) such from being able to be held more low possible (standard ISO 9996, i.e. hospitals and libraries);
- The subjects choice is operated in a random manner, on condition that:
    - they are not involved directly with works inherent to telephony or voice coding.
    - they are never participated to other subjective experiments in the precedent six months and conversational experiments in the last year.
- The subjects have to report their opinion on a opinion scale chosen between those recommended by ITU-T and the arithmetic mean of this results is called Mean Conversation-opinion Score (MOS).
- Each conversation generates two opinions and once verified anomalies absence it is possible go on with next experiments.

### 4.2.3 Quantal-Response Detectability Tests

Quantal-response detectability tests, defined by ITU-T Rec. P800 Annex C [28], represent the best method for obtaining information on the analog sound properties (that are parameters that influence QoS) and they offer quality through the vote on a scale called *Detectability opinion scale*, so to provide an opinion on parameters that are tested.

The ITU-T recommendation introduces an evaluation scale so that:

- Objectionable;
- Detectable;
- Not detectable.

This type of scale can be used for different types of quantal-response tests; for example, it can be used for echo evaluating, for tones interferences, voce-switching mutilation and sidetone.

In some situations it can be considered this vote as an opinion score, respectively 2,1 and 0 for being associate to opinion score of listening and conversation tests.

This association type is not always possible because quantal-response uses a reduced scale in respect of the classic opinion score that is formed by 5 points. For this motive, it is possible to use other scales that allow to utilize more opinion score:

- Inaudible;
- Just audible;
- Slight;
- Moderate;
- Rather loud;
- Loud;
- Intolerable.

## 4.3 Objective Measurement Methods

The subjective methods are not practicable during the network planning phase. These methods are limited, impracticable and too expensive. In order to avoid these problems, new methods that permit the calculation of values representing the different damaging factors combinations of the network have been developed. The

quality estimation, providing results as near as possible to MOS values, is the primary task of these methods.
ITU proposes an objective, automatic and replicable testing method that accounts the perceptual QoS [30,31]. The objective measurement techniques development uses an approach where a voice sample represents the input signal to produce a score, representing the original signal produced by the network. Three objective evaluation methods can be distinguished: *comparison methods,* that compare a signal with a reference, *absolute methods,* that are based on the absolute quality estimation and the *transmission methods,* that obtain a value through the network study and analysis in order to in advance know the audio quality [30,31]. Moreover, another classification that it is possible done on this measurement method is in intrusive measurement and non-intrusive measurement as it is possible to see in the follow.

### 4.3.1 Intrusive Speech Quality Measurement

This audio quality measurement system typically uses two input signals: a reference signal (original signal) and a degraded (or distorted) signal taken from the network or system to test output [30]. These tests are much accurate for the perceptual end-to-end quality of service, while they result to be inadequate for other tasks, such as the traffic monitoring. There are many objective quality measurement methods that are classified in three wide groups:

- *Time domain groups*: Signal to Noise Ratio (SNR) and Segmental to Noise Ratio (SNRseg) can represent some examples of these groups, that are easy to implement, but they are not suitable for modern networks and low bit rate codec. Novel codecs on the market are developed to reduce the original signal using audio production model, rather than reproducing the original signal. This approach does not permit the time dependent measure.
- *Spectral domain measures:* Linear Predictive Coding (LPC) and Cepstral distance Measure (CD) [32]. These ones, calculated on variable length segments, are more reliable than time domain measures and certainly less sensible to the time misalignment measures between the original signal and distorted one. The spectral measure is based on the same working principles of the modern codec. However, the production parameters audio usage limits the prediction performance in function of the applied speech production model (or codec).
- *Perceptual domain measures*: they are based on the human audio perceptual model showing to be the best objective voice quality evaluation method [33]. Perceptual Speech Quality Measure (PSQM), Perceptual Analysis Measurement System (PAMS), Measuring Normalizing Block (MNB), Enhanced Modified Bark Spectral Distortion (EMBSD) and Perceptual Evaluation of Speech Quality (PESQ), standardized by ITU-T for the voice quality evaluation on communication systems and networks represent some typical perceptual domain measures [30,31].
  An example of a perceptual method structure, that includes perceptual transform and cognition/judge modules, is depicted in Fig. 3. The first module transforms the signal in an approximated human perceptual representation, while the second one maps the difference between the original signal (the reference signal) and distorted signal (or degraded signal) in order to estimate the perceived distortion.

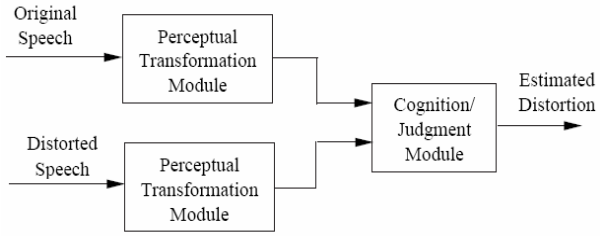

Fig. 3 Perceptual domain measures basic structure

**Perceptual Speech quality measurement method (PSQM)**
The PSQM method is developed by John G.Beerends and J.A. Stemerdink at the KPN Research and it has been later standardized by the ITU-T Rec. P.861 in 1998 [34]. The method consists of a mathematical algorithm that evaluates the difference between a telephone system distorted signal and a reference signal (a clean signal). The difference is used to calculate the presumed noise on the network in order to use this value in the measuring of the voice quality.

The mathematical algorithm can be separated in three blocks depicted in Fig.4:

1. Pre-processing: this phase is executed before the measurement. It verifies whether input and output signals are temporarily aligned and resizes the signals in order to compensate the network gain.
2. Perceptual modelling: it transforms the values in a perceptual domain; in other terms the math model obtains a physical signal representation that is converted in order to realize a real human perception signal.
3. Cognitive modelling: it is used to compare the input signal with the output signal evaluating the perceived error and calculating the noise disturbance.

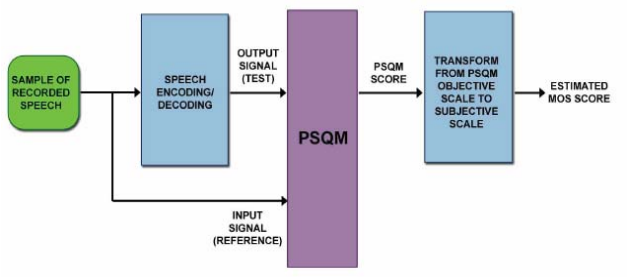


Fig. 4 PSQM model

In order to execute a PSQM measurement, the real and artificial audio samples are processed through the codec applied by the under testing system. A score of 0 represents a perfect alignment between the reference and distorted signal interpreted as a perfectly clear signal; an high score value shows the distortion amount in a connection and, typically, a score in the range [28] indicates that the connection is really degraded. The PSQM method is not suitable for the parameters evaluation such as delay and packet loss because they can damage the measurement results. Other more advanced version of the PSQM such as PSQM+ have been developed later becoming very popular in the IP telephony systems [34].

**Measuring normalizing blocks (MNB)**

The Measuring Normalizing Blocks method (MNB) has been developed by the US department of Commende in 1997 and it has been proposed as an alternative technique to the PSQM. The MNB technique [35] is recommended for the impact evaluation of some parameters that affect the signal such as error due to the communication channel or to the lower than 4kbps bit rate codes. Two MNB techniques can be applied: *Time Measuring Normalizing Blocks* and *Frequency Measuring Normalizing Blocks* In Fig. 5 a MNB architectural model is shown.

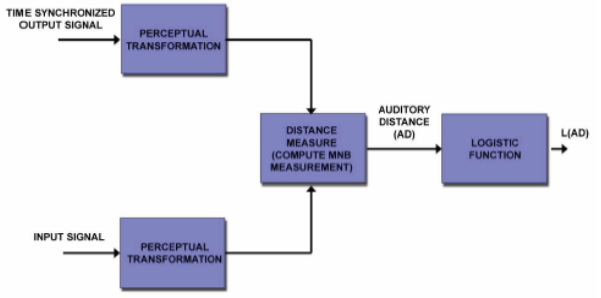


Fig. 5 MNB model

The MNB can be structured as follows:

- The time-synchronized reference and test signals are inserted in the model. They are aligned on the frequency domain inside the model and the silent frames are eliminated;
- Signals are transmitted as input to the *Compute Frequency Measurung Normalizing Block* (FMNB) and they are transmitted to the next *Time Measuring Normalizing Block*, (TMNB) with the task to provide a parameters set to be used by the signal in order to obtain different measures;
- The algorithm, at this point, merges the two MNB signals of output in a single value called *Audity Distance* (AD) that represents the quality measurement on a two signals comparison basis. This value is then mapped on the quality scale in order to obtain a subjective quality prediction [27].

**Perceptual Analysis Measurement system (PAMS)**

The *Perceptual Analysis Measurement System* task is the providing of a voice quality objective measurement in a system, addressed by damaging factors such as time clipping, packets loss, delay and distortion due to the codec usage. PAMS uses a model (Fig. 6) based on human perceptual factors, in order to measure the signal clearness for comparison with a reference signal [36,37]. Two signals are inserted in a model that aligns themselves and remove the delay effects.

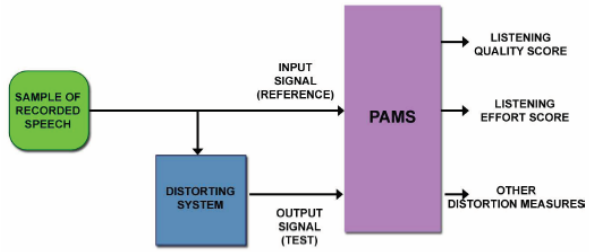


Fig. 6 PAMS model

The model can be divided as follows:

- Pre-processing: in this block, two signals are time aligned in a unique time segment to compensate the delay;
- Auditory transform: at this point the signals are modelled to include the human expressions;
- Error parameterisation: the difference between the two signals is evaluated in order to detect the errors presence. The perceptual errors are mapped in a subjective quality scale. In particular, the PAMS produces a Listening Quality Score and a Listening Effort Score that correspond to the ACR opinion scale presented in the previous paragraph.

The accurate subjective perceptual QoS is based on some tests where the voice can be affected by some parameters such as noise, excessive delay, delay due to the different codec conversion and real-time loss.

**Perceptual Evaluation of Speech Quality (PESQ)**

The PESQ is the last standard proposed by ITU (Rec.862, Feb.2001) [38] for the objective evaluation of the coded voice signal that goes through the telephone network. The model, shown in Fig. 7, includes errors such as filtering delay jitter, distortion and low bit rate.

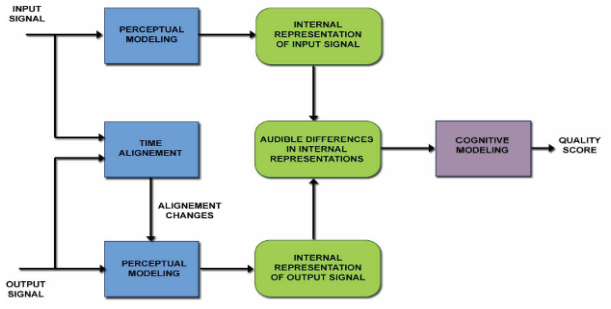


Fig. 7 PESQ model

PESQ adds novel factors and methods to calculate signal distortion offering the possibility to use natural and artificial audio samples [38-41]. The model is recommended for the impact evaluation of a codec on quality and for the prototype networks testing.

It is applied for the evaluation of the following factors:

- Transcoding
- Transmission Errors on the transmission channel;
- Codec Errors;
- Noise introduced by the system;
- Packet loss;
- Time clipping;

The PESQ method can be structured as follows [36,38-41]:

- Step 1 – *Signalling Pre-processing*: this step includes the input signal frequency and time alignment.
- Step 2 – *Perceptual Modelling*: this step concerns the input and output transformation in understandable by men representations. This transformation includes the mapping in the time and frequency domain (32ms or 256 samples for 8kHz in a way similar to the PSQM) and a signal filtering for the bandwidth typical of the telephone network (in order to not affect the PESQ measurement).
- Step 3 – *Cognitive Modelling*: in this phase the values that represent noise computation are evaluated. These values are then combined to provide a MOS score prediction. A difference between reference signal and distorted signal for each time-frequency cell is calculated. A positive difference indicates the presence of noise, while a negative difference indicates a minimum noise presence such as codec distortion. The model permits to discover the time jitter and to identify which frames are involved and which frames affected by the delay are erased in order to prevent a bad score.

### 4.3.2 Non Intrusive Speech Quality Measurement

Non intrusive or passive methods are developed to execute real-time traffic measurements. Differently from the intrusive methods; this measurement type is executed without knowing the reference signal and the traffic effect [32,42,43]. The non-intrusive traditional methods perform their prediction directly from varying IP network impairment parameters, for example jitter, delay, packet loss. Fig. 8 shows a block diagram of this technique.

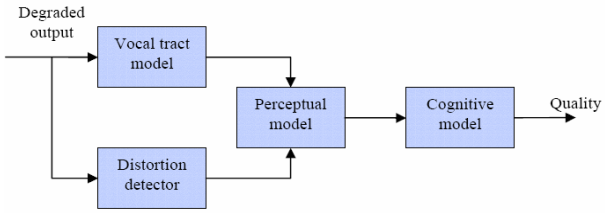


Fig. 8 Non-Intrusive technique

**In Service Non-intrusive Measurement Device, INMD**

Initially, this approach has been developed to measure parameters such as voice, noise, eco and loss levels in a circuit switched network. Now it has been improved to support also packet switched networks. INMD has been standardized by ITU in 2000 with the P.562 recommendation [44] and with the purpose of localizing and analyzing the voice performance damaging. The measured parameters in the INMD model are often the same of the parameters applied in the E-Model.

The INMD model is used for network with single distortions. The lack of a listed parameters effects combination led to the development of a new method called Call Clarity Index (CCI).

**Call Clarity Index (CCI)**

The CCI model, shown in Fig. 9, has been developed by the British Telecom to integrate the INMD model [45]. The method combines the measurement of the effects on a single call using a perceptual human model that is calibrated to obtain as result a quality subjective score called *clarity index*.

The core presented through a quality scale represents a connection test estimation. The score calculation algorithm can be loaded in a test or positioning machine in a network device such as a router or a switching platform.

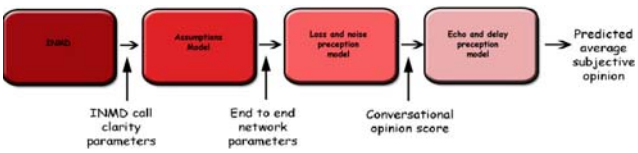


Fig. 9 CCI model

The first step executed by the model consists of the estimation of the network parameters through the consideration of the parameters that cannot be measured by the INMD. Then, perceptual factors are considered such as sidetone, environmental noise in order to produce a first score that presents a conversional speech quality. Disturb factors such as eco and delay are added to produce a final score trough the math calculus.

**Non-intrusive Quality Assessment (NIQA)**

NIQA method has been developed by the Psytechincs [46] as an extension to the CCI model to manage all distortion types (delays, silent frames, low bit-rate).

NIQA can be implemented inside a gateway, a switch, a test architecture or an architecture for the quality improvement. The possibility to include a huge distortion range permits to produce a score for each transport and codec type used by the modern digital networks. The key elements of NIQA architecture are presented in Fig. 10.

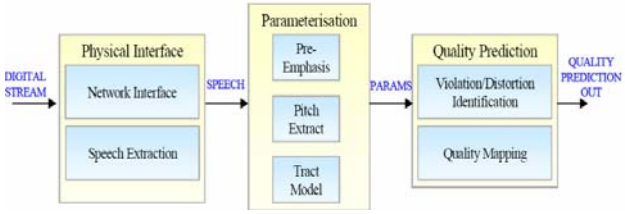


Fig. 10 NIQA architecture

NIQA algorithm can schematized as follows:

- The signal, that are not produced by the human voice, are identified;
- The distortion impact (similar to that seen for the PESQ) on the quality is predicted through a cognitive model;
- The voice signal distortions are collected and a quality score is calculated in Violation/Distortion Identification block in order to be able to correlate it through the *quality mapping* to a subjective MOS score.

**PsyVoIP**

Another developed method is PsyVoIP, similar to INMD model but dedicated to VoIP systems [47]. It is a non-intrusive method that can evaluate the voice quality on a call-by-call system, monitor the calls and measure the quality for providing a prediction of the users evaluation (Fig.11).

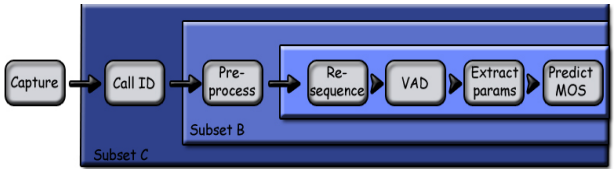


Fig. 11 PsyVoIP Model

This model can be implemented in a whichever point of VoIP network and can be so articulated as in the following:

- *Capture*: Captures all packets or captures only relevant packets, such as those belonging to a particular call;
- *Call_id*: Identifies if the packet belongs to a call and if so, which call;
- *Pre-process*: It extracts information required by the rest of model so packets can be discarded;
- *Re-sequence*: It accounts for out-of-sequence packets;
- *VAD*: It enables PsyVoIP to distinguish between speech and silence intervals during a conversation. Packet loss and jitter have less effect on quality during silence;
- *Extract Params*: It extracts the statistical descriptors required to predict MOS;
- *Predict MOS*: When enough packets have arrived to make a quality prediction valid, this performs its MOS calculation.

A very similar model to PsyVoIP is Vqmon [48] that, before of being standardized, represents an extension of E-Model standard. It permits the monitoring of the RTP traffic and the incorporating of the loss effects and packet dropping allowing so the task of obtaining a R factor (function that permits to calculate the loss effects and disturbances like jitter) that can be used for prediction a MOS score. The PsyVoIP represents the ideal candidate for the future for evaluating the voice quality on IP systems.

**Perceptual Single ended Objective Measure (PSOM)**

Perceptual Single ended Objective Measure method (PSOM) has been developed by France Telecom R&D and aims to offer a traffic quality prediction evaluation [42].

In the PSOM method speech distortions are separately measured and analyzed through comparison with a statistical model for clean speech. The output is a likelihood of each parameter that also can be characterized as a numerical distance between a reference and a degraded output signal. The model, shown in Fig. 12, does not provide any further degradations analysis than the output quality score.

The PSOM algorithm works as follows:

- An auditory transform is first applied to the input audio signal. The auditory transform output is a time-frequency representation that approximates fundamental psychoacoustics properties.
- A speech segregation stage separates the speech stream from the additive noise and detects the occurrence of non-speech components like music or signalling tones.
- Speech and noise streams are separately processed by a characterization stage. In the case of speech signal,

the likelihood measures of various perceptual parameters values are estimated from the statistical model of the speech. Furthermore, the characterization stage computes a set of disturbance measures from the noise signal, like the energy level and the spectral characteristics of background noise.

- The likelihood measures of speech and the disturbance measures of noise are merged into a quality prediction stage, which maps them to an objective speech quality grade.

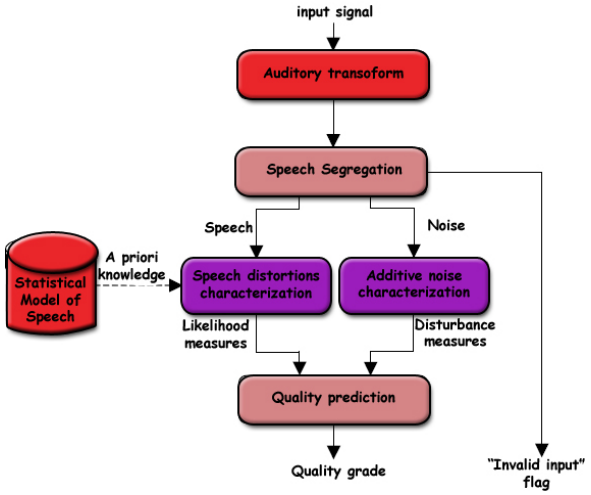


Fig. 12 PSOM model

## E-Model

The most popular objective measurement method is E-Model that belongs to non intrusive methods, that are measurement techniques that do not require the help or original speech signal.

E-Model, that is the abbreviation of *"European Telecommunications Standards Institute* (ETSI) *Computation Model* [49,50], was developed by a ETSI work group chosen by ITU [49]. It is different from other methods because it represents also a network simulation tool. The key elements of E-Model architecture are presented in Fig. 13.

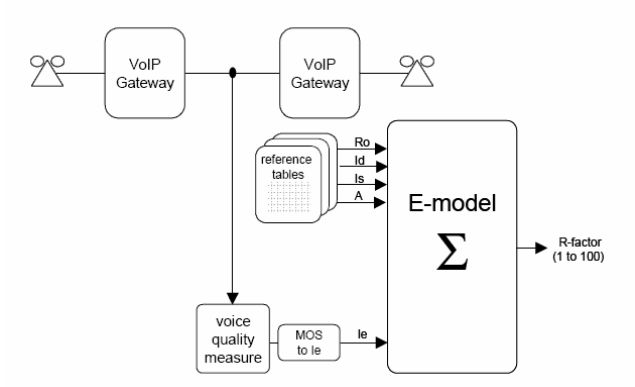


Fig. 13 E-Model architecture

E-Model is a computational method based on the assumption that each quality degradation type is associated to a certain type of damaging factor.

It uses transmission parameters to predict the subjective speech quality of packetized voice. The primary output from the E-Model is the "Rating Factor" R [36,50,51], and R can be further transformed to give estimates of customer opinion by mapping it to the MOS scale. The model calculates a base value for evaluating the quality determined by network factors. Each damage factor is then expressed in a value that is, later, subtracted from the base value.

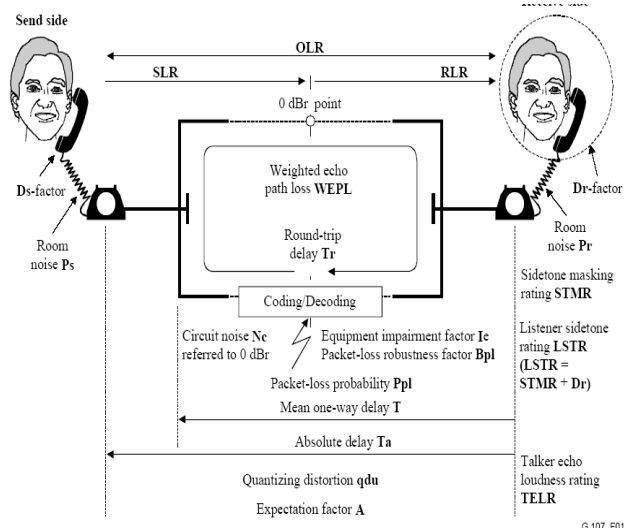


Fig. 14 Damages factor for R factor computation

The damage factors involved in the R factor computation are represented in the Fig. 14. A detailed list follows below:

- Room noise PS and Room noise PR, background noise
- Noises derived from the microphone and loudspeaker use, Ds-Factor and Dr-Factor
- Sending Loudness Rating SLR, Receiving Loudness Rating (RLR), Overall Loudness Rating (OLR)
- Quantizing Distortion (qdu)
- Equipment impairment factor (Ie)
- Packet-loss probability (Ppl)
- Mean one-way delay (T)
- Absolute Delay (T)
- Expectation factor (A)
- Parameters receiver side: Side tone masking rating (STMR), listener masking rating (LSTR), Talker echo loudness rating (TELR)
- Weighted echo path loss (WEPL), Round-trip delay (Tr)

E-Model input is composed of parameters that are valid in network design and installation. Some of these parameters are measured while other parameters are extrapolates by standard reference samples.

The first output of E-Model is *Transmission rating factor R* that is used for evaluating the quality perceived by the network. The formula in order to calculate R factor is shown follow (Eq. 2):

$$R = R_0 - Is - Id - Ie + A \qquad (2)$$

- $R_0$ represents the signal to noise ratio, including the noise generated by the circuit and the background noise;
- *Is* factor is obtained from the combination of all damage factors that almost simultaneously affect the voice signal;
- *Id* coefficient represents the damages caused from delay;
- Ie coefficient represents the damages caused from codecs with low bit rate;
- *A* represents a factor that adapts the quality value.

The *R* factor result is represented from a scale that goes from 10 to 100, but typically the range used goes from 50 to 90 as it is possible to view in the table 7 where S represent Satisfied and DS Dissatisfied.

Table 7 R factor quality classes

| ***R-value range*** | ***100-90*** | ***90-80*** | ***80-70*** | ***70-60*** | ***60-50*** |
|---|---|---|---|---|---|
| Speech transmission quality category | Best | High | Medium | Low | (very) Poor |
| User's satisfaction | Very S | S | Some users DS | Many users DS | Nearly all users DS |

The last step of E-Model is that of mapping the R factor value in an equivalent MOS value shown in the Fig. 15.

The ITU-T Rec. G.107 standard [49] defines the relation between R factor and MOS through the follow equation (Eq. 3):

$$MOS = \begin{cases} 1 & for\ R \le 0 \\ 1 + 0.035R + R(R-60)(100-R)7x10^{-6} & for\ 0 < R < 100 \\ 4.5 & for\ R \ge 100 \end{cases} \qquad (3)$$

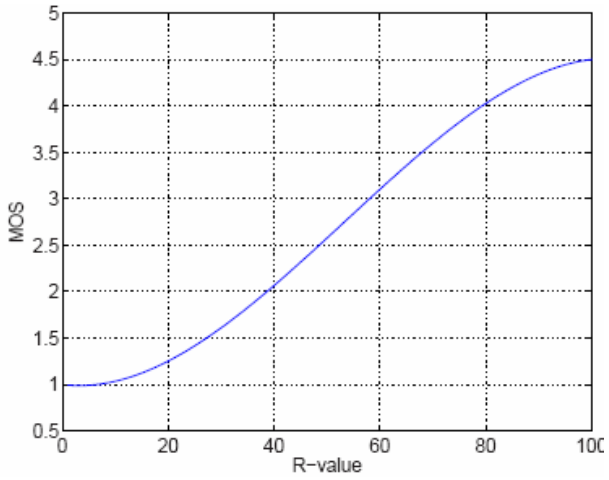


Fig. 15 Relation between R factor and MOS

## 5. Conclusions

In this paper we have conducted a detailed study on VoIP service. We have focused on QoS problem in the VoIP realization. The quality of VoIP services in accordance with changing network conditions and to control the QoS and manage the utilisation of available resources can be optimized through the use of quality metric. The use of this metric is more important in the actual internet network, that has been projected for a typology of service called best-effort and it is not designed to support real-time voice communications. We have described what are the main QoS parameters like delay, jitter, packet loss and so on. Then we have effectuated a classification of the QoS measurement technique that are divided in subjective and objective methods. We have shown the main methods for both classes of measurement.
This document can represent a base document for the research of a better utilization of VoIP services that, still today, are incapable of guarantee a satisfactory quality of service to its users.

## References


[1] "Overview of the PSTN and Comparisons to Voice over IP", CH01, White paper, October 2001;
[2] H. M. Chong, H. S. Matthews, "Comparative Analysis of Traditional Telephone and Voice-over-Internet Protocol (VoIP) Systems", IEEE ISEE 2004;
[3] "H.323 Technology", Ixia, White Paper, 2004;
[4] IETF RFC 3261, "SIP: Session Initiation Protocol", June 2002;
[5] IETF RFC 3435, "Media Gateway Control Protocol (MGCP) Version 1.0", January 2003;
[6] "Media Gateway Control Protocol (MGCP) Technology", Ixia, White Paper, 2004;

[7] "The role of Megaco/H.248 in media gateway control: A protocol standards overview", Nortel network, White paper;
[8] "Megaco Tecnology", Ixia, White Paper, 2004;
[9] T. Taylor, "Megaco/H248: A New Standard for Media Gateway Control", IEEE Communication Magazine Oct. 2000;
[10] B. Goode, "Voice Over Internet Protocol (VoIP)", IEEE Communication Magazine Sept. 2002;
[11] H. Liu, P. Mouchtaris, "Voice over IP Signaling: H.323 and Beyond", IEEE Communication Magazine Oct. 2000;
[12] "Voice over IP (VoIP)", Spirent Communications, P/N 340-1158-001 REV A, 8/01, 2001;
[13] VoIP Problem Detection and Isolation", Qovic, Inc, Nov. 2003;
[14] "Voice Over Internet Protocol", White Paper by DSQ Software LTd;
[15] Leon W., Couch II, "Digital and Analog Communication Systems", Macmillan;
[16] ITU-T Recommendation G.711, "Pulse codec Modulation (PCM) of voice frequencies", Nov. 1988;
[17] ITU-T Recommendation G.726, "Adaptive Differential Pulse Codec Modulation", Dic. 1990;
[18] J. Bradbury, "Linear Predictive Coding", Dec 2000.
[19] ITU-T Recommendation G.729, "Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)", March 1996;
[20] ITU-T Recommendation G.728, "Coding of Speech at 16Kbit/s Using Low-Delay Excited Linear Prediction", Sept. 1992;
[21] ETSI EN 301 704 V7.2.1 (2000-04), Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi-Rate (AMR) Speech Transcoding (GSM 06.90 version 7.2.1 Release 1998);
[22] E. Bernex, A. Gatineau, "Quality of service in VoIP environments", White Paper, www..neotip.com;
[23] Lijing Ding, Rafik A. Goubran, "Assessment of effects of packet loss of speech quality in voip", IEEE HAVE 2003;
[24] ITU-T Recommendation G.114, "One-way transmission time", May 2003;
[25] IETF RFC 3550, "RTP: Transport Protocol for Real-Time Application", July 2003;
[26] ITU-T Recommendation G.723.1, Dual Rate Speech Coder for Multimedia Communication Transmitting at 5.3 and 6.3 kbit/s, March 1996;
[27] A. Lakaniemi, J. Rosti, V. I. Raisanen, "Subjective VoIP speech quality evaluation based on network measurements", IEEE ICC 2001;
[28] ITU-T Recommendation P.800, "Methods for Subjective Determination of Transmission Quality", 1996;
[29] ITU-T Recommendation P.64, "Determination of sensitivity/frequency characteristics of local telephone systems";
[30] J. Anderson, "Methods for Measuring Perceptual Speech Quality", White Paper, Agilent Technologies, Oct. 2001;
[31] F. Hammer, P: Reichi, T. Ziegler, "Where Packet Traces Meet Speech Samples: An Instrumental Approach to Perceptual QoS Evaluation of VoIP", IEEE IWQOS 2004;
[32] ETSI Final Draft EG 201 377-3 v1.1.1, "Non-intrusive objective measurement methods applicable to networks and links with classes of services", STQ Specification and Measurement of Speech Transmission Quality, April 2003;
[33] E.E.Zurek, J.Leffew, W.A Moreno, "Objective Evaluation of Voice Clarity Measurement For VoIP Compression Algorithms", Fourth IEEE International Caracas Conference on Devices. Circuits and Systems. Aruba. April 17-19.2002;
[34] ITU-T Recommendation P.861, Objective quality measurement of telephone-band (300-3400 Hz) speech codecs, February 1998
[35] ITU-T Recommendation P.861, App. II, "Objective quality measurement of telephone-band (300–3400 Hz) speech codecs using measuring normalizing blocks (MNB's)", Genevra, Switzerland, 1998;
[36] www.psytechnics.com;
[37] "PAMS: Measuring speech quality over networks...as the customers hear it", White Paper, Psytechnics May 2001;
[38] ITU-T Recommendation 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", Feb. 2001;
[39] "PESQ: An Introduction", White Paper, Psytechnics Sep. 2001;
[40] A.E. Conway, "Output-Based Method of Applying PESQ to Measure the Perceptual Quality of Framed Speech Signals", WCNC 2004;
[41] "PESQ: Measuring speech quality over networks, as the customers hear it", ", White Paper, Psytechnics 2001;
[42] J. Anderson, "Addressing VoIP Speech Quality with Non-Intrusive Measurement", White Paper, Agilent Technologies;
[43] L. Sun and E. Ifeachor, "Learning Models for Non-intrusive Prediction of Voice Quality for IP Networks," IEEE Transactions on Neural Networks, 2004.
[44] ITU-T Recommendation P.562, "Analysis and Interpretation of INMD Voiceservice Measurements", May 2000.
[45] "CCI: Getting the message loud and clear - measuring the clarity of speech over networks", White Paper, Psytechnics 2001.
[46] "Non Intrusive Quality Assessment", Psytechnics, Jan. 2003.
[47] "PsyVoIP in IP Phones", Product Description Psytechnics Jun. 2003
[48] http://www.telchemy.com/vqmonep.html;
[49] ITU-T Recommendation G.107, "The E-model, a computational model for use in transmission planning", Mar. 2005;
[50] "The E-Model, R Factor and MOS, Overview", Psytechnics Dec. 2003;
[51] www.itu.int/ITU-/studygroups/com12/emodelv1/index.htm;

**Floriano De Rango** was born in Cosenza, CS, Italy, in 1976. He received the degree in computer science engineering in October 2000, and a Ph.D. in electronics and communications engineering in January 2005, both at University of Calabria, Italy. From January 2000 to October 2000 he worked in the Telecom Research LAB C.S.E.L.T. in Turin as visiting scholar student. From March 2004 to November 2004 he was visiting researcher at the University of California at Los Angeles (UCLA). Since November 2004 he joined the D.E.I.S. Department, University of Calabria as Research Fellow. He served as reviewer of VTC'03, ICC'04, WCNC'05, Globecom'05, WTS'05, WirelessCOM'05, IEEE Communication Letters, JSAC. His interests include Satellite networks, IP QoS architectures, Adaptive Wireless Networks and Ad Hoc Networks.

**Mauro Tropea** graduated in computer engineering at the University of Calabria, Italy, in 2003. Since 2003 he has been with the telecommunications research group of D.E.I.S. in the University of Calabria. In 2004 he won a regional scholarship on Satellite and Terrestrial broadband digital telecommunication systems. Since November 2005 he has a Ph.D student in Electronics and Communications Engineering at University of Calabria. His research interests include satellite communication networks, QoS architectures and interworking wireless and wired networks, mobility model.

**Peppino Fazio** received the degree in computer science engineering in May 2004. Since November 2004 he has a Ph.D student in Electronics and Communications Engineering at University of Calabria. His research interests include mobile communication networks, QoS architectures and interworking wireless and wired networks, mobility model.

**Salvatore Marano**, graduated in electronics engineering at University of Rome in 1973. In 1974 he joined the Fondazione Ugo Bordoni. Between 1976 and 1977 he worked at ITT Laboratoriy in Leeds, United Kngdom. Since 1979 he has been an associate professor at the University of Calabria, Italy. His research interests include performance evaluation in mobile communication systems