

Chapter 2

Speech Quality Measurement Methods

This chapter deals with measurement methods particularly relevant to any assessment of the perceived quality of voice and speech. Such voice and speech quality measurement methods are employed in several scientific fields, such as medicine (e.g. the evaluation of voice-related problems), linguistics or speech technology (e.g. the evaluation of speech transmission systems or their components). Each field has its own assessment paradigm. This chapter makes use of two statistical parameters: (i) the Pearson correlation coefficient, ρ , and (ii) the standard deviation of the prediction error, σ . Both are defined in Sec. 5.1.3.2.

2.1 Definitions

In *metrology*, which is the science of measurement, **measurement** is generally defined as (BIPM Guides in metrology, 2008):

a process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity.

and a **measurement result** is (BIPM Guides in metrology, 2008):

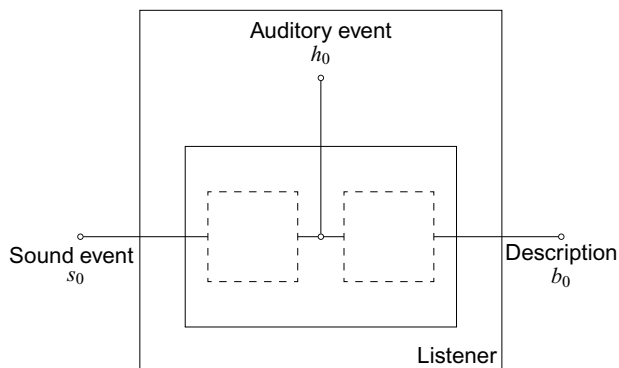
a set of quantity values being attributed to a measurand together with any other available relevant information.

In a voice and speech quality measurement, a **measurand** is (Jekosch, 2005):

a feature of the perceived speech event which can numerically be described on a measurement scale.

A generic description of an auditory experiment was proposed by Blauert (1997). A corresponding schematic representation of a listener involved in such an auditory

Fig. 2.1 Auditory test: schematic representation of a listener (Blauert, 1997). s_0 : sound event; h_0 : auditory event; b_0 : description of the sensation by the listener



experiment is shown in Fig. 2.1. In a first step, a sound event (i.e. an acoustic signal), s_0 , reaches the listener's ear. After the perception process, this acoustic signal results in an auditory event, h_0 (i.e. a *sensation*), in the listener's sensory system (see Sect. 1.1.3). Except for the subject himself (by *introspection*), the auditory event is hardly accessible to the experimenter. Only a description of the sensation by the listener, b_0 , is accessible to other persons. In psychophysics, the subject is asked to define b_0 , which relates at the best the auditory impression, h_0 . The output of the assessment process, b_0 , corresponds to either a linguistic description of h_0 or a quantification on a psychophysical measurement scale (i.e. the amount of the sensation), see Stevens (1957). Such measurement scales used in auditory experiments are consequently a major part of the speech perception and assessment process. Given the relationship: $b_0 = f\{s_0\}$, the loss of information related to the quality judgment should be small in order to get a benefit from the auditory experiment.

Consequently, according to Jekosch (2005), the goal of a **measurement scale** is such that:

[...] that the numerical relational system forms an accurate copy of the structures and features of the speech that has been perceived and judged.

Recently, Durin and Gros (2008) investigated the impact of speech quality on human behavior in communication tasks. The dual task method they employed avoids the use of measurement scales in auditory test methods. However, such method is taxing for the test subjects and is still under developments.

Quality measurement methods must be designed so that the parameter really quantified is the user's perception. Following Osgood (1952), a satisfactory measurement method meets the six following characteristics:

Objectivity b_0 is reproducible (verifiable) over different listeners (i.e. *inter-subjectivity*).

Reliability	The results provided by the method show no large scattering when s_0 is repeated to the same listener (i.e. <i>intra</i> -subjectivity).
Validity	The parameter measured by the method is the one intended to be measured (i.e. one element of the semantic triangle, see Sec. 1.1.3).
Sensitivity	The distinctions enabled by the method are as fine as those made by the listener.
Comparability	The method is applicable to a wide range of perceived qualities and makes possible comparisons between groups of conditions.
Utility	The pieces of information provided by the method are useful.

These six characteristics are congruent with the definition of measurement in use in metrology.

In the literature, the terms, *assessment* and *evaluation*, both refer to measurement methods. The methodologies under focus here deal with assessment. Indeed, an overall evaluation of a system would include the measurement of many characteristics that are not into the scope of this book (e.g. cost). On the other hand, the term, assessment, is related to the performances of a system for comparison purposes (Möller, 2005).

In addition, the “tool” or measurement apparatus used to measure the speech signal can be either a test subject (auditory method) or a physical equipment (instrumental method). In auditory methods, a test subject is asked to judge the quality of the speech signal. Since the perception process (described in Sec. 1.1.3) and the resulting perceived quality are internal to the user and not accessible from the outside, auditory experiments are the only method that meets the six characteristics introduced by Osgood (1952). But as they are costly and time-consuming, instrumental measurement methods have been developed. They provide one with a quality estimation from physically measured values. In this sense, the term *model* is used in computer sciences instead of the instrumental measurement method. In the literature, these two approaches are improperly referred to as *subjective* and *objective* methods (Blauert and Guski, 2009), respectively. This terminology is even used by the ITU-T organization (ITU–T Rec. P.800.1, 2006). However, the objectivity or degree of objectivity of the auditory results refers to the amount of consistency between listeners in perception.

For Jekosch (2005) the **objectivity** is:

the invariance with respect to the totality of what is individually perceived. [...]
Objectivity is the extent of inter-individual agreement.

In this book, several statistical metrics, which determine the degree of objectivity, will be introduced later (Sec. 5.1.3).

The present chapter gives an overview of different auditory methods (see Sec. 2.2) and instrumental methods (see Sec. 2.3). In speech quality, a specific unit, called Mean Opinion Score (MOS), is employed to define the resulting quality scores. It corresponds to an average of the individual scores for each processing condition. Then suffix notations are added to the term MOS (ITU–T Rec. P.800.1, 2006) in order to indicate:

- The *test modality*: listening (MOS_{LQ}), talking (MOS_{TQ}) or conversation (MOS_{CQ}) quality.
- The *measurement method*: instrumental (referred to as objective MOS_{LQO} for signal-based models or estimated MOS_{CQE} for parameter-based models) or auditory (referred to as subjective MOS_{LQS}).
- The *context of measurement*: a Narrow-Band context (MOS_{LQON}), WideBand (MOS_{LQOW}) or a mix of both bandwidths (MOS_{LQOM}). No specific notation has been defined for a Super-WideBand context so-far.

2.2 Auditory methods

The most accurate auditory measurement method would be an assessment by customers in natural environments. Theoretically, the customer should be able to assess the quality of an ongoing call through use of his phone keypad. In practice, such “in field” tests are hardly implemented, and speech quality is assessed thanks to artificial auditory quality tests carried out in laboratories (i.e. under designed and controlled conditions).

According to Jekosch (2005), a **speech quality test** is:

a routine procedure for examining one or more empirically restrictive quality features of perceived speech with the aim of making a quantitative statement on these features.

The ears of any individual are permanently submitted to a flow of acoustic signals. However, only the characteristics that are source of information for the listener are analyzed. In “undirected” speech perception processes (e.g. an everyday conversation) the interlocutors exchange pieces of information (i.e. meaning of the spoken sentences). However, during an auditory experiment the speech perception process is “directed” by the experimenter, i.e. the test subject is oriented throughout the experiment by means of directives. The directives are part of the modifying factors introduced in Sec. 1.2¹. In directed communications, the test subjects do not expect the same type of information as in undirected communications. In this case, subjects may focus on the sign carrier (i.e. the form of the speech signal, see Fig. 1.2)

¹ For an example of such directives see ITU–T Rec. P.800 (1996).

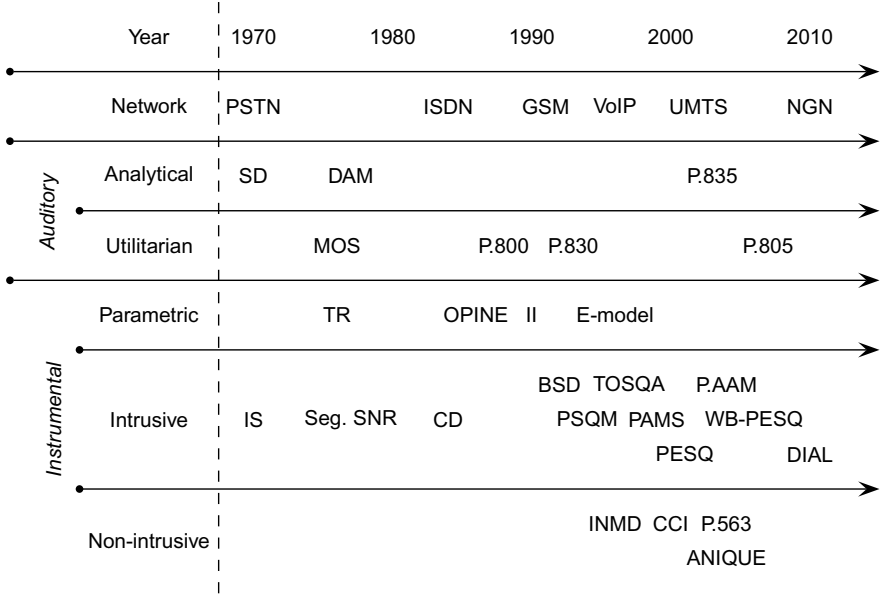


Fig. 2.2 List of auditory and instrumental speech quality measurement methods

which may bias the quality perception. However, speech quality tests must reflect the quality perception of users during undirected communications. Consequently, the directives and the measurement scale have to be carefully design by the experimenter.

Following the tests classification introduced by Letowski (1989), speech quality tests can be classified into four categories according to two dichotomies: (i) *analytic / utilitarian*, and (ii) *subject-oriented / object-oriented* (see Table 2.1). By using specific directives, the experimenter can, in directed communications, adjust the influence of each quality feature. Listener’s attention can, thus, be focused on a group of speech quality features, or on a single one. The selected quality features and the corresponding perceived quality, are all stored as auditory memory traces in the Short-Term Memory (see Sec. 1.1.3.1). In *utilitarian* test methods, subjects assess the integral quality of speech transmission systems thanks to a single quality score on a one-dimensional rating scale. It consequently permits a comparison between different processing conditions. On the other hand, in *analytic* test methods, the perceptual features of the integral speech quality are identified and then quantified. Two different approaches are available: either a single one-dimensional scale is used and the listeners are asked to focus on a given feature, or several scales, one per quality feature, are employed. In the latter, the subjects’ judgments may be decomposed into orthogonal quality features on the basis of a multidimensional analysis. In addition to this first dichotomy, the speech quality test may lead to two different analyses: (i) an *object-oriented* analysis about the perceived quality of processing

conditions, or (ii) a *subject-oriented* analysis based on the role of the test subjects in the perception process.

Table 2.1 Quality test classification following Letowski (1989)

	Subject-oriented tests	Object-oriented tests
Utilitarian judgments	Psychoacoustic research	Speech quality assessment
Analytical judgments	Audiological evaluation	Diagnostic quality assessment

Implementing such speech quality tests is a complex task. According to Stevens (1957), Möller (2000) and Raake (2006b), five main characteristics defines the exact test results. The experimenter selects the appropriate characteristics based on the number and the type of assessed processing conditions.

The presentation method	Paired comparison or absolute assessment of stimuli
The scale level	A ratio-, interval-, ordinal- or nominal-scale ²
The scaling method	A single- or multi-scale rating process
The test modality	Listening-only, talking-only or conversation test
The analysis method	Simple average or multidimensional analysis

Several examples of utilitarian and analytical methods focusing on standard measurement methods which are widely used by telecommunication providers will be briefly presented hereafter. Moreover, Fig 2.2 (p. 41) proposes an exhaustive list of auditory test methodologies.

2.2.1 Test subjects

The selection of the test subjects should be consistent with the test purpose. Indeed, they may be classified according to their knowledge about the selection of the processing conditions under test. The corresponding two groups are, namely *expert* subjects (or trained) and *naïve* subjects (or untrained). The learning and adaptation effects of trained subjects on the quality judgments were observed in IEEE Standards Publication 297 (1969). Utilitarian test methods are usually aimed at getting the speech quality as perceived by the “average” user population. Consequently such tests are commonly carried out with naïve subjects as recommended by the ITU-T organization.

² The heard speech samples are ranked by the test subjects on an ordinal scale rank. From an interval (resp. ratio) scale, the difference (resp. ratio) between two categories is quantified by a numerical value.

According to ITU-T Rec. P.800 (1996), the definition of a **naïve** test subject is as follows:

Subjects taking part in listening tests are chosen at random from the normal telephone using population, with the provisos that:

1. they have not been directly involved in work connected with assessment of the performance of telephone circuits, or related work such as speech coding;
2. they have not participated in any subjective test whatever for at least the previous six months, and not in any listening-opinion test for at least one year; and
3. they have never heard the same sentence lists before.

The work, presented in the Chap.s 3, 4 and 5, is based on auditory tests carried out with naïve subjects.

In speech quality tests, both expert and naïve test subjects must be free of any hearing impairment. A subject's hearing ability is commonly evaluated from his hearing threshold determined by an audiometric test. In addition, the mother tongue of the subjects must correspond to the language in use in the experiments. Both characteristics (lack of hearing impairment and native speaker) can be seen as inconsistent with the definition of an average user. Therefore, auditory tests used for market-planning purposes have different requirements. For instance, if a telecommunication service is designed for a specific segment of the population (e.g. age range, disability ...), the selected test subjects have to be representative of this specific category of users. Raake et al. (2008) carried out an exhaustive quality test with different groups of test subjects: a subject-oriented analysis revealed several group-dependencies in the quality judgments by the subjects. For instance, "IP-expert" users gave a significantly lower quality rating than the other users under conditions with a high rate of packet-loss, probably because of their past experience with VoIP systems.

Contrary to utilitarian methods, analytical methods may use a complex measurement process which implies a specific training of the test subjects. The expert subjects may then have a common understanding of the auditory quality features involved in the experiment. This training process can significantly improve the objectivity and reliability of the test method. Example of a subject-selection process is available in Isherwood et al. (2003). In addition, the greater production of diagnostic information (e.g. through a linguistic description of the impairments) by expert subjects compared to naïve ones can lead to a reduction of the experiment cost. However, the outputs of an experiment conducting with few trained subjects should be considered as an informal test since it will not give a representative account of the quality perceived by the final users.

2.2.2 Speech material

According to the perception process introduced in Sec. 1.1.3, the linguistic information is extracted and stored in the Short-Term Memory for a few seconds. But, as the auditory memory traces used in the quality judgment are fading very rapidly, the length of the speech samples used in quality measurements is limited to 8 seconds. The speech material used in speech quality tests must consist of simple, meaningful and phonetically balanced sentences so as to reflect the phonemic frequencies in the subjects' language. For an example of phonetically balanced sentences in French language see Combescuré (1981). Since the speech material has a strong influence on the perceived quality, the sentences must create an equivalent meaning in the mind of all of the test subjects. In listening quality tests, the speech samples should be ideally made of two sentences separated by a silent gap. Thus, the subjects can evaluate the background noise introduced by the transmission system without any masking effect from the speech signal. There should be no obvious connection of meaning between the two sentences.

As described in Sec. 1.1.3.3, the perceived quality may be affected by the talker characteristics (e.g. gender, accent). Consequently, the ITU-T recommends that a test be conducted with, at least, 2 male and 2 female voices per processing condition.

2.2.3 Utilitarian methods

Utilitarian test methods are employed to assess the integral quality of speech stimuli as perceived by an end-user. In these tests, a speech sample is played to a group of subjects, who are asked to rate the quality of the sample on a one-dimensional rating scale. Then, a unique quality value, comprising the effect of all quality features, is calculated for each processing condition. Such test methods are widely used for the assessment of new speech processing applications or the comparison of different versions of a single application (e.g. a speech coding algorithm).

2.2.3.1 Intelligibility tests

Intelligibility is only one specific attribute of the perceived integral speech quality (Voiers, 1977; Volberg et al., 2006). Even if modern transmission systems lead to an almost perfect intelligibility of the far-end speaker, specific speech processing systems such as hearing aids are still evaluated through intelligibility experiments. In such tests, the subject is asked to rewrite the understood parts of the listened speech sample. The output score then corresponds to a percentage of correct recognition.

In the phonetic process introduced in Sec. 1.1.3.2, the first stage, comprehensibility, is specifically assessed by Vowel–Consonant–Vowel (VCV) tests or by modified

versions of this test (CV, VC, CCVC, ...): the subject has to recognize a particular phoneme. Examples of such tests are the Standard Segmental Test (S_{AM}) or CLuster Identification test (CLID) (Jekosch, 1993). Then, intelligibility is assessed in tests such as the Diagnostic Rhyme Test (DRT) or the Modified Rhyme Test (MRT). Here, the subject has to recognize whole words. Other tests are targeted to the assessment the final stage of the phonetic process, the comprehension. Here, the whole (speech) message has to be recognized by the subject. For an example of such comprehension test, see Raake and Katz (2006). In addition, the ITU-T standardized a test method dedicated to the assessment of the effort made by the subject to understand the meaning of the sentence. The rating scale used is called *listening-effort scale* (see Table 2.2). An average over all listeners results in a mean listening-effort opinion score MOS_{LE} .

Table 2.2 Listening-effort scale

Effort required to understand the meaning of sentence	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

2.2.3.2 Conversation tests

Conversation test methods are described in ITU–T Rec. P.800 (1996) and ITU–T Rec. P.805 (2007). Such experiments try to simulate a natural use of telephone services and are consequently the most relevant (Dimolitsas, 1993). The conversational quality assesses the interlocutors’ ability to communicate throughout a call. This ability is dependent upon the transmission quality and by conversation effectiveness factors such as echoes at the talking side, transmission delays and sidetone distortion (see Sec.1.2.3). Two or more test subjects are asked to achieve a task specified by an interactive communication scenario. After a short conversation of about 5 minutes, the subjects assess different aspects of the connection thanks to e.g. a listening quality rating scale and a talking quality rating scale. Seven scales are presented in ITU–T Rec. P.805 (2007). In addition, it is usual for the experimenter to ask the subject to describe in his/her own words the nature of the degradation (e.g. echo, noise). In conversation tests, the arithmetic mean over all test subjects of the quality judgments is called the MOS–Conversational Quality Score, and is denoted by MOS_{CQS} (ITU–T Rec. P.800.1, 2006). An overview of conversational quality tests is available in Möller (2000). However, the design and conduct of conversational tests are more complex than those of listening tests. In practice, only few conditions are assessed in conversation tests. For an overview of the relationships between listening and conversational quality see Guéguin et al. (2008).

2.2.3.3 Listening-Only tests

Listening-only experiments are carried out to gather the most important quality features. Their realism is lower than that of conversational tests since only the speech transmission quality can be assessed. The P-Series of Recommendations published by the ITU-T such as ITU-T Rec. P.800 (1996) and ITU-T Rec. P.830 (1996) describe a general framework of measurement methods used in assessments of speech quality. In a listening quality test (referred to as Listening-Only Test (LOT) by the ITU-T), the listeners rate on a measurement scale a set of short speech samples, called *stimuli*, transmitted by different speech transmission systems. In such listening tests, the listening level is alike whatever the stimulus and set to 79 dB_{SPL} (dB rel. 20 μ Pa), which corresponds to the preferred listening level in a NB context (ITU-T Handbook on Telephonometry, 1992).

Absolute Category Rating (ACR)

In telecommunications, the most widely used speech quality test is an Absolute Category Rating (ACR) test that uses the 5-point integral quality scale presented in Table 2.3. An arithmetic mean over all listeners of the quality judgments is called a MOS_{LQS} value.

Table 2.3 Absolute Category Rating (ACR) Listening-quality scale

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

Degradation Category Rating (DCR)

However, the sensitivity of such methods is insufficient for the comparison of speech processing systems of similar integral quality. In such cases, a Degradation Category Rating (DCR) method is more appropriate. For small impairments, a paired comparison (A–B) method is more sensitive than an ACR method (Combescur et al., 1982). In DCR tests, for each trial, the subjects listen to both a reference (i.e. non degraded) and a degraded speech signal. The listener is asked to rate on the 5-point rating scale presented in Table 2.4 the perceived degradation in quality of the processed (i.e. second) signal from comparison to the reference (i.e. first) signal. The resulting quality value is referred to as the *Degradation Mean Opinion Score*

(DMOS). The DCR method is part of the quality test framework defined by ITU–T Rec. P.800 (1996) and ITU–T Rec. P.830 (1996).

Table 2.4 Degradation Category Rating (DCR) scale

Score	The degradation is . . .
5	inaudible
4	audible but not annoying
3	slightly annoying
2	annoying
1	very annoying

Comparison Category Rating (CCR)

Another type of standard quality test uses a reference speech sample, which may be of lower quality than the rated sample. The scale in use is, thus, the 2-sided rating scale presented in Table 2.5. This method, called Comparison Category Rating (CCR) can be seen as a refinement of DCR tests where the reference can be presented in the first or the second position (A–B and B–A). The resulting quality value of CCR test is a *Comparison Mean Opinion Score* (CMOS). The CCR method is also part of the quality test framework defined by ITU–T Rec. P.800 (1996) and ITU–T Rec. P.830 (1996).

Table 2.5 Comparison Category Rating (CCR) scale

Score	Quality of the second stimulus compared to the first one
3	Much better
2	Better
1	Slightly better
0	About the same
–1	Slightly worse
–2	Worse
–3	Much worse

2.2.3.4 High-quality listening tests

ITU–T Rec. P.800 (1996) has been defined for the assessment of Narrow-Band telephony. With the introduction of WideBand transmissions, the ITU-T published a specific recommendation for the evaluation of WB speech codecs ITU–T Rec. P.830

(1996). It slightly differs from the ITU–T Rec. P.800 (1996). For instance, in WB tests, the listening terminal should reproduce at least the WB bandwidth: the typical IRS-type user terminal is replaced by a high-quality headphone. In addition, the listening mode, which is usually “monotic” in NB tests, is replaced by a “diotic” presentation of the stimuli³. Nowadays, signals can be transmitted with a wider bandwidth than in the past (e.g. S-WB telephony). Unfortunately such methodologies are not suited to this range of quality. Consequently, high-quality speech processing systems are assessed by methodologies used in the audio world and published by the Radiocommunication sector of the ITU (ITU-R) organization. Two of these methodologies are presented below.

Assessment of small impairments

In WB and S-WB telephony, an auditory method for the assessment of small impairments in audio systems is usually employed (ITU–R Rec. BS.1116–1, 1997; ITU–R Rec. BS.1284–1, 2003). In this method, three audio samples are presented, *A*, *B* and *X*, are presented to the listener, who is asked to select which of the samples, *A* and *B*, is identical to the reference, *X*. Then, the listener rates also the degradation of the other sample through comparison to *X*. This method is appropriate to the assessment of small impairments such as those introduced by high quality audio coding algorithms.

MULTI Stimulus test with Hidden Reference and Anchor (MUSHRA)

Another ITU-R standardized method used in audio quality test is the MUSHRA (ITU–R Rec. BS.1534–1, 2003). In this test, several speech sample, including a known reference and hidden anchors, are presented together through a multi-scale interface. The subject is asked to rate on a continuous scale defined from 0 (i.e. lowest quality) to 100 (i.e. best quality) the whole set of stimuli, except the known reference. This scale is divided into five equal intervals and labeled with the same adjectives as the listening-quality scale shown in Table 2.3: bad, poor, fair, good and excellent. The resulting scores quantify the degradation of the conditions under test from comparison to known reference.

³ A *monotic*, or *monaural*, mode corresponds to a presentation of the speech stimuli to only one ear (left or right, depending on the subject). A *diotic* mode corresponds to a presentation of the same signal to both ears. A diotic mode differs from a *dichotic* presentation, where signals sent to the right and left ears are different. Such a dichotic mode can be either stereo or binaural when recorded with an artificial head (ITU–T Rec. P.58, 1996).

2.2.4 Analytical methods

A utilitarian test method quantifies the integral speech quality as it is perceived by an end-user. However, the information provided by a single quality value is insufficient to allow comparisons between very different processing conditions. Indeed, two communication systems may have the same integral quality but a totally different behavior. Analytical test methods give diagnostic information about the assessed processing conditions. Such quality tests rely on either a multi-scale rating process (e.g. SD) or a multidimensional analysis of the auditory results (e.g. MDS).

2.2.4.1 Diagnostic Acceptability Measure (DAM)

Voiers (1977) developed a specific multidimensional scaling method called Diagnostic Acceptability Measure (DAM) which assesses several quality features of speech samples. The subjects evaluate the speech samples on 20 continuous rating scales. Each scale is dedicated to the assessment of a given quality feature from negligible (0) to extreme (100). This auditory method has the advantage of providing the individual differences in taste and preference. The scales are divided into three categories: (i) features related to the speech signal (e.g. interrupted, rasping), (ii) features related to the background noise (e.g. hissing, babbling), and (iii) features covering both speech and background noise (e.g. intelligibility, acceptability). However, such a test is expensive and time-consuming since the listeners are trained beforehand (experienced). Finally, on the basis of a linear relationship between the quality features (related to the speech signal and the background noise) and the acceptability, the auditory results can also be used for diagnostic purpose.

2.2.4.2 Semantic Differential (SD)

The Semantic Differential (SD) method developed by Osgood (1952) was first applied to the definition of a semantic space related to “words”. It uses a set of opposite attributes, i.e. pairs of antonym terms (e.g. small/large and wet/dry). Each pair of antonyms defines the poles of a continuous rating scale. This method relies upon the following hypotheses (Osgood, 1952):

The process of description or judgment can be conceived as the allocation of a concept to an experiential continuum, definable by a pair of polar terms.

A limited number of such continua can be used to define a semantic space within which the meaning of any concept can be specified.

The subject is asked to judge the intensity and the polarity of the feature underlying the pair of antonyms (e.g. volume and wetness respectively). Using such pairs for characteristics related to voice- and speech-quality features (e.g. low/high) makes possible the application of the SD method for diagnostic purposes, e.g. see McGee (1964). This measurement method is sometimes referred to as *attribute scaling*.

2.2.4.3 Evaluation of Noise Reduction (NR) algorithm

An important perceptual dimension of speech communication quality is the amount of noise in the transmitted signal. Communication systems where background noise can be present, e.g. mobile phones or hands-free terminals, are more and more frequent. A real background noise brings information about the environment of the far-end talker, especially in speech-free periods. As the integral quality is highly degraded in a speech signal polluted by noise, Noise Reduction (NR) systems have been integrated to user terminals. Such NR systems are designed to increase the SNR, but they may degrade the speech signal. Recently, the ITU-T published an analytical measurement method for the evaluation of noise reduction algorithms (ITU-T Rec. P.835, 2003). This methodology uses 3 5-point rating scales to assess the quality of the speech signal alone (*speech signal distortion*), the background noise alone (*background noise intrusiveness*) and the integral quality. The subject is asked to listen to the same speech sample three times; a silent pause consecutive to each listening allows him to score the sample on one of the three rating scales.

2.2.4.4 Assessment of speech quality dimensions

In the SD method, the poles of the scales are labeled with a pair of antonym quality features (e.g. Continuous—Discontinuous). However, in SD and DAM tests, all scales are presented simultaneously. In a recent analytical method dedicated to diagnostic purpose and developed by Wältermann et al. (2010b), speech samples are assessed on three continuous rating scales dedicated, respectively, to the following perceptual speech quality dimension, i.e. *Discontinuity*, *Noisiness* and *Coloration*, and previously derived by Wältermann et al. (2008). The measurement takes place in two steps: (i) an LOT/ACR “overall quality” test according to ITU-T Rec. P.800 (1996) and (ii) a “dimension assessment”. In the interval between them, the meaning and use of the three scales are described in detail to the subjects by means of directives. Moreover, for each scale, examples are proposed for training.

2.2.4.5 Single quality feature

In an auditory test, focus may be on a single specific quality feature. For instance, the listening level has a strong influence on the integral speech quality. Consequently, the ITU-T recommends an ACR scale for the specific assessment of the *preferred* listening level, see Table 2.6 and ITU-T Rec. P.800 (1996). The output quality score (mean loudness-preference opinion score) is denoted by MOS_{LP} .

Further to an exhaustive campaign of auditory experiments conducted in the 1980s, the preferred listening level for a monaural listening situation was found to be 79dB_{SPL}. This finding led the ITU-T Handbook on Telephonometry (1992) to recommend the use of this specific level in every monaural speech quality experiments. However, according to ITU-T Contrib. COM 12–11 (1993), a level higher

Table 2.6 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

than the preferred listening level leads to a higher MOS_{LQS} value. The maximum speech quality is obtained at the *optimum* listening level. But, at levels higher than the optimum listening level, the integral speech quality is decreased. In addition, ITU-T Contrib. COM 12–11 (1993) showed that the difference between the preferred and the optimum listening levels is dependent upon other quality features such as the bandwidth. For instance, the difference is more marked under WB conditions than under NB ones.

2.2.4.6 Multi-Dimensional Scaling (MDS)

A general description of this statistical analysis method is available in Kruskal (1964). Contrary to the other analytical methods, the Multi-Dimensional Scaling (MDS) focuses only on the perceptual “differences” between stimuli. This method requires dissimilarity data between several speech stimuli. They are acquired from a similarity test performed on a continuous scale labeled with the two attributes “very similar” and “not similar at all”. In the ideal case, the similarity of all $\frac{N \times (N-1)}{2}$ possible pairs of stimuli (given N speech samples) is judged by the subjects. Then, a multidimensional similarity space can be derived from the auditory results. The number of dimensions that defines the derived space is a compromise between the covered variability of the original similarity results and the possibility to interpret each dimension.

The different MDS techniques are distinguished through use of the following criteria:

Classical / Nonclassical

In classical MDS, a single dissimilarity space is derived for all subjects whereas in nonclassical MDS several spaces are derived. For instance, in *weighted* MDS (also known as *individual difference scaling* INDSCAL Carroll and Chang, 1970) a subject space is derived in addition to the similarity space. The subject space shows the weight given to the dimensions by each subject.

Metric / Nonmetric

In metric MDS, test subjects are required to quantify the dissimilarity. On the other hand, in nonmetric MDS, they judge about the rank order dissimilarity. For

instance, the results issued from a triadic comparison test can be analyzed by a nonmetric MDS.

The interpretation of the derived dimensions is relatively complex. A first possibility is the “arbitrary” selection of an attribute through an experts’ exhaustive evaluation of the degradation differences along one specific dimension. A second possibility is the comparison of the derived space with other auditory test results. For instance, the dimensions can be described by the degree of correlation with the antonym pairs of attributes used in the SD test.

2.2.4.7 Preference mapping

The preference mapping is a multidimensional statistical analysis of preference judgments. In this case, a preference test such as an ACR listening quality test or a paired-comparison preference test is conducted at first. It is followed with a factor analysis made through application of a Principal Component Analysis (PCA) algorithm, for example, to the test results. This permits one to distinguish two types of preference mapping: the *internal* and the *external* preference mapping methods.

Internal preference mapping provides a multidimensional representation of the speech stimuli where test subjects or group of test subjects are represented as vectors. Carroll (1972) developed an internal preference mapping algorithm called multidimensional preference scaling (MDPREF). An external preference mapping uses a pre-existing multidimensional representation of the speech stimuli. Then, the relationship between the preference of the speech stimuli and each dimension (e.g. a quality feature) is derived. The external preference mapping is widely used in the food industry in order to adapt or create new products for each segment of the population.

2.2.5 Relativity of subjects’ judgments

Perception in real world is about perception in context.

(Lotto and Sullivan, 2008)

Many factors can influence the way the user is perceiving the speech sample under test. Quality scores are “relative” to the test characteristics and consequently not “absolute”. The following section will briefly review several aspects of an auditory test liable to affect the subject’s judgment. For an exhaustive review of all influencing factors see Möller (2000); Poulton (1979); Zieliński et al. (2008). According to

the description made by Jekosch (2005), these aspects may be classified into three groups:

- The scaling-effect** induced by the use of the measurement scale as an interface with the subject.
- The subject-effect** generated by the use of human listeners as an instrument of measurement.
- The context-effect** due to the relationship between the context and the use of speech as an object of measurement.

This means that getting an absolute quality value based on the subject's judgment is not possible. Following some simple guidance rules, the experimenters try to get a quality score as absolute as possible in order to differentiate and compare the processing conditions under study in the test. These rules are aimed at reducing biases in subjects' judgments.

2.2.5.1 Scaling-effect

Jekosch (2005) assumed that the scale has a strong influence on the results. It should enable each subject to encode the different features he has perceived. On the other hand, rating by a subject should not represent his own interpretation of the speech message (i.e. *meaning*) but rather a common perception of the acoustic quality features (i.e. *form*, see Sec. 1.1.3). A review of all of the measurement scale effects is available in Poulton (1979).

- **Intervals between categories**

In category scales (e.g. ACR method), the intervals between two categories (i.e. the quality scale labels) may be unequal, and this inequality leads in non-linear measurement scales. Such scales are referred to ordinal scales. However, simple statistical parameters such as arithmetic mean have been developed for interval and ratio scales (Möller, 2000). These nonlinearities are attenuated by introducing numerical value in front of each category, see Table 2.3.

- **Language**

The translation into another language of the category names influences the MOS values. For instance, Zieliński et al. (2008) showed that the semantic difference between the English words, "Fair" and "Poor", and the one between their French equivalents, "Assez bon" and "Médiocre", are not alike. Quality assessments are, thus, language-dependent.

- **Sensitivity of category scales**

Even though a discrete 5-point scale seems to be the preferred scale in terms of "ease of use", a 5-point MOS scale has a relatively low sensitivity (Preston and Colman, 2000). On the other hand, sensitivity is increased when a continuous scale is employed for rating speech quality (ITU-T Contrib. COM 12-39, 2009) since the standard deviation of the processing conditions is reduced. However,

ITU-T Contrib. COM 12-120 (2007) showed that subjects mostly use notches on continuous scales (e.g. numbers or category names) to judge the stimuli.

- **Saturation-effect**

Among the other scale-effects let us cite the *saturation-effect*. The extreme categories of the scale are neglected by naïve subjects, which introduces nonlinearities. In ITU-T Contrib. COM 12-39 (2009), the authors showed the saturation-effect of the discrete ACR scale, see Table 2.3.

2.2.5.2 Subject-effect

In the specific case of a listening-only test, the subject-effect and the context-effect can be described by Fig. 2.3. This diagram shows the temporal relationship between both effects and their impact on the subject's judgment about the current speech stimulus (i.e. t_0). Three specific parts on the time scale are defined. The context-effect corresponds to the last two parts but it is also dependent on the test situation.

The left part corresponds to the subject-effect; it is caused by the differences in the internal reference of each subject. The internal reference corresponds to his overall experience in telecommunications. He has his own opinion about the importance of each quality feature involved in the integral speech quality. However, the test reliability is decreased by the variations in judgments with the personal internal reference: indeed, the subjects expects the auditory event to have a perceptual quality similar to his own internal reference. According to Takahashi et al. (2005b) subjects show a preference for the one or the other of the specific NB or WB bandwidths. To reduce this effect, the ITU-T proposed the two solutions described hereafter:

- The introduction text, read to the subjects at the beginning of the auditory test (i.e. “directives”, see introduction of Sec. 2.2), should include a “question” in relation with the quality scale in use. This question defines how the subjects have to judge the speech samples. It has a strong influence on the features involved in the quality judgment and finally on the utility of the test results. The ITU-T Handbook on Telephonometry (1992) recommended some specific questions/scales such as; “Please give your opinion on whether you or your partner had any difficulty in talking or hearing over the connection according to the following rating

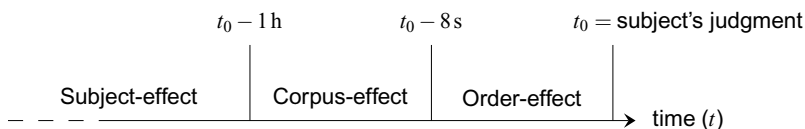


Fig. 2.3 Influences of the subject-effect and the context-effect (i.e. corpus-effect and order-effect) on the subject's judgment versus time (Côté et al., 2009)

scale”: *Yes* (Difficulty), *No* (No difficulty).

- The number of subjects should be large enough to get a certain degree of objectivity in the quality judgments since the resulting average over all subjects corresponds to the inter-individual agreement. This led the ITU–T Handbook on Telephonometry (1992) to recommend sets of 30 subjects. In addition, to reduce this subject-effect, some stimuli are usually presented over a training period prior to the conduct of the experiment. These stimuli include the highest and lowest qualities of the test corpus, which are then seen like an anchor by the subjects.

At last, to avoid any fatigue effect, and a consequent decrease in the accuracy of subjects’ judgments, it is recommended to interrupt the test procedure by short breaks at regular intervals (e.g. 15–20 minutes).

2.2.5.3 Context-effect

The context-effects represent the influence of assessment situation. One of the most influencing bias in auditory results corresponds to the listening environment. Assessments by subjects are made through in-laboratory tests, which are quite different from a real-life situation. For Guski and Blauert (2009), the highest bias of auditory judgments obtained in laboratory tests compared to real-life situations is their restriction to one signal modality. Perception in the real world is multimodal and in the case of speech signs, two modalities are perceived: *vision* and *sound*. However, in the specific case of telephony studies, the lack of visual cues reduces the gap between real-life and laboratory environments.

Corpus-effect

The central part of Fig. 2.3 corresponds to the *corpus-effect*, and $t_0 - 1$ hour refers to the test start. As subjects’ judgments are affected by both the range of degradations within the test corpus and their distribution over the quality range, the interpretation of MOS values is dependent on the corpus-effect. Hardy (2003) stated that conditions included in the test corpus should be realistic and, thus, account for the quality range met in an “ecologically valid” situation, i.e. under real telephony situation.

Several studies have dealt with the influence of the context-effect on MOS values, and more specially with the influence of the corpus-effect. For instance, further to their investigations about the impact by bandwidth restriction, Möller et al. (2006) found that an uncoded NB (i.e. 300–3 400 Hz) condition obtains a higher MOS value in a purely NB corpus than in a mixed-band one where high quality WB (i.e. 50–7 000 Hz) conditions are introduced. Côté and Durin (2008) demonstrated that all perceptual dimensions are dependent on this corpus-effect, which can, however, be reduced by introducing several reference conditions equally distributed over the judgment scale (Barriac et al., 2004). For instance, the systematic introduction

of a WB condition in a purely NB test corpus may reduce the corpus-effect for the perceptual dimension *coloration*. Ideally, all corpora should include NB, WB and S-WB conditions, but such a requirement is not possible in practice. In addition, studies by Takahashi et al. (2005b) about the subject's sensitivity in a purely NB context and in a mixed-band one suggested that the introduction of WB conditions caused no decrease of his sensitivity.

In order to compare new auditory tests to those previously carried out in a NB context, Barriac et al. (2004) proposed to define a mapping function from the NB MOS scale to the WB MOS one. Therefore, the ITU-T introduced in ITU-T Rec. P.800.1 (2006) a specific label for the context-related MOS values: this means that the NB context gives MOS_{LQSN} values, whereas the WB one provides MOS_{LQSW} values and the mixed-band one, NB/WB, leads to MOS_{LQSM} values (see Sec. 2.1).

Order-effect

The right part of Fig. 2.3 corresponds to the influence of the preceding stimulus on how the current stimulus is judged, given that the most recent stimulus has the strongest influence. This is called the *order-effect*. This effect is attenuated using different listening orders for each subject (or group of subjects).

2.2.6 Reference conditions and normalization procedures

Some of the biases introduced in the previous section are reduced by setting reference conditions with their corresponding normalization procedures for application to the MOS values. Such reference conditions correspond to processing schemes defined by a known-in-advance parameter, which quantifies the introduced degradation. These reference units are added to the auditory test corpora and cover a defined quality range so that the quality of the conditions under study falls within this range (i.e. one reference unit has the lowest quality and another reference unit has the best quality of the test corpus). In addition, for a given auditory test, the MOS values of the conditions under study can be transformed to an “absolute” scale through a normalization procedure aimed at ruling out some of the test-specific effects. Consequently, the introduction of reference conditions enables one to compare the results across tests (and even across laboratories) despite the differences in languages, methodologies, etc.

The MOS values obtained for all processing conditions can be normalized to a specific range (e.g. 1–4.5) by using a simple linear mapping function computed by:

$$MOS_{\text{norm},i} = \frac{MOS_i - MOS_{\min}}{MOS_{\max} - MOS_{\min}}(MOS_{\text{lim}} - 1) + 1, \quad (2.1)$$

where i corresponds to the current condition, $MOS_{\min}$ and $MOS_{\max}$ are, respectively, the lowest and highest MOS values of the corpus, $MOS_{\lim}$ is the highest MOS value wanted (e.g. 4.5) and $MOS_{\text{norm},i}$ is the resulting normalized MOS value.

2.2.6.1 Modulated Noise Reference Unit (MNRU)

Over the first half of the twentieth century, transmitted speech was mainly degraded on the specific perceptual dimension *Noisiness*. This led to the definition by Rothauser et al. (1968) of a “reference unit” as a signal of the same nature as the samples included in the test corpus. Nowadays, a consensus seems to have reached about the conduct of comparison of the references and the conditions under tests along the same perceptual dimensions. For Rothauser et al. (1968), the easiest way to produce reference conditions is, therefore, the introduction of noise in speech samples. This additive noise can be white or shaped (e.g. pink noise), stationary or modulated with the speech signal amplitude (Law and Seymour, 1962). An example of the latter type was standardized as the Modulated Noise Reference Unit (MNRU) in ITU–T Rec. P.810 (1996) and further used quite extensively in the assessment of speech codecs. In this specific case the noise is correlated to the speech signal. The degradation introduced is similar to the quantizing noise produced by the logarithmic PCM technique used by waveform speech codecs. A detailed description of the MNRU normalization approach is available in App. A.

However, nowadays, MNRUs neither give account for the current diversity of degradations, nor reduce the corpus-effect. For instance, a MDS analysis enabled Hall (2001) to demonstrate that signal-correlated noises are perceptually different from the non-linear degradations introduced by low bit-rate speech codecs. The perceptual dimension, *Noisiness*, is far more affected by MNRU conditions than by low bit-rate speech codecs.

2.2.6.2 Standard speech codecs

A single auditory test rarely includes the whole perceptual space defined by modern speech transmission systems. In the extreme case where the subject of focus is only the quality degradation introduced by a specific speech codec, a test corpus including strong degradations like MNRU conditions will likely introduce a corpus-effect. It will prevent both an exact quantification of the speech codec quality and the inter-comparisons of speech codecs. In the last decade, these considerations let the ITU-T to adopt a different type of reference units. The quality of several common speech codecs was quantified by a parameter called “equipment impairment factor” I_e from previously carried out auditory tests. ITU–T Rec. G.113 (2007) defines I_e values for several speech codecs introduced in Sec. 1.3.4.1. Then, some of these common speech codecs were proposed as reference conditions in auditory tests (ITU–T Rec. P.833, 2001) or in a pool of stimuli whose the quality is estimated by an instrumen-

tal model (ITU–T Rec. P.834, 2002).

Wideband versions of the normalization procedures were recently published as the ITU–T Rec. P.833.1 (2008) and ITU–T Rec. P.834.1 (2009). The latter procedure is described in detail in Sec. 3.2.3.

2.3 Instrumental methods

As described in the introduction of this chapter, auditory methodologies rely on judgments by subjects, who are asked to give their opinion about the quality of a speech signal. Since auditory tests are costly and time-consuming, instrumental methods, referred to as *quality models*, have been developed. They consist in a computer program designed to automatically estimate the perceived quality of speech signals. Such a method is based on a mathematical model, which establishes a relationship between a sensation and a physical magnitude. A first “psychophysics” model has been developed by Fechner (1860) and referred in the literature as the “Weber-Fechner law” states that:

$$S = \alpha \cdot \ln \left(\frac{\phi}{\phi_0} \right), \quad (2.2)$$

where S is the sensation (i.e. perceived intensity), ϕ is a physical parameter and ϕ_0 is the perception threshold. However, such mathematical models provide an *estimation* of the sensation perceived by human subjects.

Since auditory methods are the most reliable way to assess the perceived quality of a system under study, the development of an instrumental model must be based

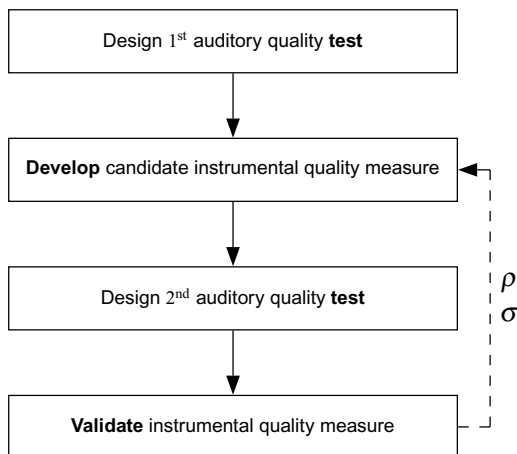


Fig. 2.4 Development procedure of an instrumental model from Wolf et al. (1991)

on auditory results and, according to Wolf et al. (1991), should consist of four main steps (Wolf et al., 1991): (i) the design of a first auditory test, (ii) the development of a candidate instrumental measure based on the auditory test results, (iii) the design of a second auditory test, and (iv) the validation of the instrumental method on the auditory results issued from the second auditory test through use of statistical parameters, e.g. the Pearson correlation coefficient, ρ , and the prediction error, σ , (see Sec. 5.1.3.2). In the case where the candidate instrumental measure fails to the validation phase, additional developments, including the design of a new auditory test, may be necessary. Consequently, enhancements of instrumental models have always been related to the development of new speech processing systems. The accuracy of a candidate model is quantified by comparison of the quality estimations with the auditory speech quality ratings. This accuracy is used as the main criterion for the validation of candidate models. A detailed validation procedure is described in Sec. 5, and the four steps are depicted in Fig. 2.4. The resulting instrumental quality measurement is highly dependent upon the design of the first auditory test (e.g. the scale level, the presentation method and the test conditions). Instrumental measures are restricted to specific applications contrary to auditory methods. In this sense, an over-generalization of instrumental measurement and calculation methods are heavily criticized by scaling experts (Jekosch, 2005).

The instrumental models can be characterized by six criteria adapted from Rabiner (1995):

Completeness	All of the speech processing systems already in-use throughout the world fall within the scope of the model. This criterion shows that the development of speech quality models has been intimately related to the historical evolution of the speech processing systems.
Accuracy	The most widely used criterion. The estimated scores are correlated with human perception.
Credibility	The estimation is easily interpretable.
Extensibility	The scope of the model can increase.
Manipulability	The model is easily employed. The model must be totally self-sufficient: there is no need for fine tuning by the users.
Consistency	The relationship between the estimations and the auditory results is monotonic (<i>internal consistency</i>). The absolute estimated values have approximately the same magnitude as the auditory results (<i>external consistency</i>).

Instrumental methods have different applications such as the daily monitoring of transmission networks or the optimization of speech processing systems. Instrumental quality models are classified in three different groups from their assessment paradigm (Takahashi et al., 2004):

- **Parameter-based models**

The quality elements of the transmission path are characterized by parameters,

- to plan future transmission networks.
- **Signal-based models**
They use signals either transmitted through a telephony network or degraded by a speech processing system to evaluate under-development and in-use transmission networks and speech processing systems.
 - **Packet-layer models**
They analyze the parameters provided by transmission networks such as the pattern of transmitted packets by VoIP networks to further monitor in-use packet-switched networks.

The choice of an instrumental model depends on the current state of the speech processing system under study. The development phase of such systems is described by a *quality loop* (Jekosch, 2005) such as the one presented in Table 2.7. The quality elements are selected and evaluated over a network-planning phase (i.e. the network is not yet set up). During this *planning* phase, parameter-based models inform the telecommunication companies about the quality of the future transmission system. The parameter-based models can, thus, only *predict* the perceived quality of the future system. During the *execution* phase, signal-based models compare different versions of a speech processing system or different network configurations. Then, in the *usage* phase, telecommunication companies monitor, and maintain when needed, in-service networks or optimize some algorithm. Such estimations are provided by signal-based and packet-layer models. After a short recall of the historical evolution, the next section will give typical examples of applications of the three types of instrumental models and highlight their limitations (see Fig. 2.2 (p. 41 for an exhaustive list of instrumental quality models).

Table 2.7 Quality loop (Jekosch, 2005)

Phase	Name	Description
1 st	Planning	Market research, design, testing and production planning
2 nd	Execution	Production, final testing and distribution
3 rd	Usage	Monitoring and maintenance

2.3.1 Parameter-based models

The integral quality of an entire transmission path can be assessed from the characteristics of each element of the network. Then a relationship between the physical characteristics of the elements and the corresponding perceived quality is established through use of a set of parameters defining each element of the transmission

system from the talker's mouth to the listener's ear. From this set of parameters, parameter-based models are able to predict the speech communication quality of future networks, before the implementation of the system under study. In the next paragraphs, the historical evolution of network-planning models will be briefly recalled so as to further describe the corresponding relationships between the physical parameters of either the transmission network or user terminal and the expected speech communication quality.

2.3.1.1 Loudness Rating

Fletcher and Galt (1950) developed an assessment procedure of the loudness loss mainly induced by electro-acoustic devices in telephone networks. The resulting measurement, expressed in dB, is referred to as the Loudness Rating (LR). This parameter is used in telephonometry to express the sum of attenuations by the transmission path in each frequency band. The transmission path is compared to a reference system. Different reference systems have been developed such as the *orthotelephonic reference position* (ITU-T Handbook on Telephonometry, 1992) (i.e. 1 m air path) and the IRS (ITU-T Rec. P.48, 1988). The LR model was published as ITU-T Rec. P.76 (1988). The specific acoustic procedure in use in LR measurements is available in ITU-T Rec. P.79 (2007). The overall transmission loss generated by the entire transmission path is the *Overall Loudness Rating* (OLR). It corresponds to the sum of three parameters:

- SLR** Send Loudness Rating: from the talker's mouth to the handset microphone output.
- JLR** Junction Loudness Rating: linear and non-linear distortion in the transmission network.
- RLR** Receive Loudness Rating: from the handset loudspeaker input to the listener's ear.

2.3.1.2 Opinion models

In the 1960s and 1970s, several national telecommunication companies carried out auditory tests so as to evaluate and facilitate the extension and maintenance of their telephony networks. From auditory tests results, they derived algorithms predicting the opinion expressed by users about the phone connection, i.e. their own auditory "reaction" to the telephone network. Such parameter-based models of first generation are often termed opinion models. They cover almost all of the degradations observed in analog telephony networks: attenuation of the transmission path, circuit noise, environmental noise, quantizing noise, talker echo and sidetone. However, the principles in use in each of these models are different. In 1993, four different parameter-based models that all designed from the Loudness Rating (LR) model were proposed in the ITU-T Suppl. 3 to P-Series Rec. (1993). Among them, the Bellcore Transmission Rating (TR) model developed by Cavanaugh et al. (1976) re-

lies on the use of the OLR value of the transmission system whereas the three other ones employ an auditory perception model mainly based on the LR model.

- TR** From mainly mapping functions between the opinions expressed by subjects and empirical data, the Bellcore TR model (Cavanaugh et al., 1976) predicts the quality of telephone network on a “transmission rating scale” (R -scale). As this scale is anchored at two points, the produced scores are much less context-effect dependent. The combination of input scalar parameters leads to a transmission rating factor, R , which is increasing monotonically with the transmission quality.
- CATNAP** The British Telecom Computer-Aided Telephone Network Assessment Program (CATNAP) was proposed in ITU-T Suppl. 3 to P-Series Rec. (1993) and based on a previous model called SUBjective MODEL (SUBMOD). This model was described, at first, by Richards (1974). From a theoretical model of human auditory perception, CATNAP simulates this perception process through use cause-and-effect relationships between the input parameters and output values. These input parameters are frequency-dependent quantities, e.g. transmission path sensitivity, listener’s hearing and speaker’s talking features, room noise spectra and sidetone characteristics. CATNAP provides one with two opinion scores: (i) the conversation opinion score (Y_C) and (ii) the listening-effort score (Y_{LE}).
- II** The Information Index (II) method developed by Lalou (1990) predicts the transmission quality using both scalar parameters and frequency-dependent ones. Two scores are provided, (i) the listening information index (I_l), and (ii) the information index in a conversation context (I_c).
- OPINE** The Overall Performance Index model for Network Evaluation (OPINE) developed by Osaka and Kakehi (1986) predicts the quality of a speech communication. It relies on the additivity theorem established by Allnatt (1975), which defines that all psychological factors are additive on a psychological scale. The OPINE model uses scalar parameters and frequency-dependent parameters.

All of these four models were developed to predict the quality of fixed-line networks such as PSTN. One should note that none of them considers the distortions introduced by digital systems such as low bit-rate speech codecs.

2.3.1.3 E-model

In 1997, the different opinion models were integrated in a new parameter-based model, the E-model, by Johannesson (1997): this means that for instance, the transmission rating scale (R -scale) and the “additivity property” of impairment factors

used in the OPINE model are taken into account by the new algorithm. The author also included the effects related to modern digital networks. The E-model is thus adapted to both traditional impairments such as echo and transmission delay, and degradations introduced by modern transmission scenarios (e.g. non-linear distortions introduced by low bit-rate codecs). This parametric model was published at first as the ETSI ETR 250 (1996) by the European Telecommunications Standards Institute (ETSI). Then, it was published as the ITU-T Rec. G.107 (1998). The listening and talking terminals (e.g. LR), the transmission network (e.g. delay, circuit noise) and several environmental factors (e.g. environmental noise) are characterized by 21 input parameters. The transmission quality rating, R -value (i.e. rating), is computed as:

$$R = R_0 - I_S - I_d - I_e + A, \quad (2.3)$$

where R_0 is the “highest” Signal-to-Noise Ratio (SNR) in absence of other impairments. This SNR is based on the basic noise parameters (real background noises and circuit noises). Then, each impairment factor quantifies a specific degradation. For example, I_S represents the impairments that occur simultaneously with the speech signal, I_d encompasses the impairments related to the conversational effectiveness (impact of delay and echo) and I_e corresponds to the equipment impairment factor introduced by a low bit-rate codec. The advantage factor A allows a compensation of the impairment factors in terms of “advantage of access” (e.g. cordless handset).

The E-model is mainly used in network-planning where it helps planners to make sure that users will be satisfied with the overall transmission system. It determines a conversational quality on the R -scale ranging from $R = 0$ (the poorest possible quality) to $R = 100^4$ (the best quality). Then, the following mapping function (2.4) enables one to transfer the predicted R -value to an opinion scale for transformation into a MOS_{CQEN} value (i.e. in a NB context):

$$\begin{aligned} \text{For } R < 0 : \quad & MOS = 1 \\ \text{For } 0 \leq R \leq 100 : \quad & MOS = 1 + 0.035R + R(R - 60)(100 - R)7 \cdot 10^{-6} \\ \text{For } R > 100 : \quad & MOS = 4.5 \end{aligned} \quad (2.4)$$

The current algorithm has been developed for NB connections and is currently extended to WB scenario (Raake et al., 2010). An exhaustive description of the achieved steps will be given in Sec. 3.2.1. A set of default values for the 21 input parameters has been published. These default values correspond to a standard ISDN connection defined by: (i) user terminals corresponding to the IRS (ITU-T Rec. P.48, 1988), (ii) a certain amount of environmental noises, and (iii) an ITU-T Rec. G.711 (1988) speech codec. This channel obtains a relatively high R -value ($R = 93.2$).

⁴ Several parameter values, e.g. $N_{for} > -64$ dBmp, lead to R values greater than 100, which are outside the permitted range defined in ITU-T Rec. G.107 (1998) and are thus normalized to 100.

A recent update of the E-model (ITU–T Del. Contrib. COM 12–44, 2001) was aimed at predicting the communication quality when random transmission errors occur in packet-switched networks. The equipment impairment factor, I_e , was adjusted towards an $I_{e,\text{eff}}$ value, which quantifies the impact of packet-loss on the speech quality as follows:

$$I_{e,\text{eff}} = I_e + (95 - I_e) \cdot \frac{P_{pl}}{P_{pl} + B_{pl}}, \quad (2.5)$$

where $I_{e,\text{eff}}$ is the “effective” equipment impairment factor impaired by packet losses, I_e is the equipment impairment factor in the error-free case, P_{pl} is the percentage of lost packets, and B_{pl} is a factor describing the robustness of the codec against packet loss. The higher B_{pl} is, the lower the artifacts associated to packet-loss are. The constant of 95 in (2.5) represents approximately the R -value of an “optimal” NB condition.

According to ETSI ETR 250 (1996), the standard deviation of E-model prediction errors on the MOS scale is about $\sigma = 0.7$. An exhaustive evaluation of the E-model by Möller (2000) showed that, in the case of individual degradations for both traditional network and low bit-rate speech codecs, the quality predictions are reliable. On the other hand, the assessment of combined degradations may lack of reliability and, thus, in this specific case, signal-based models are more accurate (Möller and Raake, 2002).

2.3.2 Signal-based models

Whereas the parameter-based models introduced in the previous section are mainly used over the first stage of the quality loop, they do not provide useful information during the last stage, i.e. the usage phase of the transmission system (see Sec. 1.2.2), where telecommunication providers are interested in the quality assessment of in-use networks to detect problems liable to occur. In addition, parameter-based models are less reliable in that case of combined degradations, and thus a different type of instrumental models, the “signal-based” models, is employed. These models are split into two groups (see Fig. 2.5):

Intrusive models : also known as double-ended measurement methods since they use a *reference* (clean or system input) speech signal, $x(t)$, and a corresponding *degraded* (distorted or system output) speech signal, $y(t)$ ⁵.

Non-intrusive models: also known as single-ended or output-based measurement methods since they use only the *degraded* speech signal, $y(t)$.

⁵ In practice, the model input signals correspond either to digital signals, $x(k)$ and $y(k)$, or electrical ones, $x(t)$ and $y(t)$.

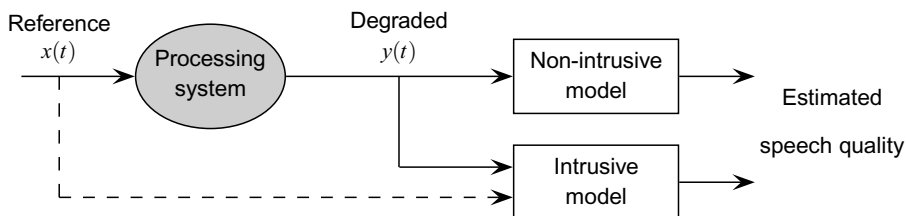


Fig. 2.5 Intrusive and non-intrusive speech quality models

Early studies on speech quality estimation were mainly based on intrusive models. It is only since the 1990s that the estimations by non-intrusive models are reliable. Intrusive models consist of three main components:

1. a pre-processing step,
2. a component that transforms the speech signal(s),
3. an assessment unit.

For instance, in frame-by-frame comparisons, intrusive methods need a re-alignment of the two speech signals. In this case, the pre-processing step includes a specific algorithm proceeds to a precise time-alignment, which may be a difficult task especially with packet-switched networks. In addition, for improvement of intrusive models, a Voice Activity Detector (VAD) algorithm may be used in the pre-processing step. The degradation is, thus, estimated from only the active frames. Then, different types of signal transformation exist. The first measurement methods are defined in the time domain and represent a simple way to characterize the performance of a system under test using a single number. The other measurement methods are defined in either the frequency domain or in the perceptual one and enable a more sophisticated approach to estimate the perceived speech quality. Comparison studies of several signal-based quality models are reported in Quackenbush and Barnwell (1985), Lam et al. (1996), Au and Lam (1998) and Côté et al. (2008).

2.3.2.1 Input signals

As set above, instrumental methods have been introduced for reliable prediction of quality scores assessed through auditory tests. As seen in Sec. 2.2.2, the source material in use in auditory tests corresponds to speech records (i.e. *natural speech*). However, instrumental measurement methods may use different input signals. For instance, an intrusive model uses a reference signal and its corresponding degraded version. The former, $x(t)$, is a “clean” signal with a linear PCM coding scheme (16bit quantization) and frequency components within the bandwidth context: NB, WB or S-WB. The latter, $y(t)$, is any signal processed by a speech processing system or transmitted through a network.

In some specific cases, the variability of natural speech leads to undesirable effects on the quality estimations. As, by definition, the perceived quality is dependent upon both the speech message (i.e. meaning) and talker's characteristics (see Sec. 1.1.3), this leads to confusion when comparing instrumentally measured quality values. To avoid such a bias, the set of natural speech input signals used by the experimenter has to be large. Another solution is the use of a simpler signal such as *pink noise* or *tones*. For instance, in the audio domain, the usual analysis corresponds to the Total Harmonic Distortion (THD), and the stimulus is a white noise. However, the temporal and spectral properties displayed by such signals and human speech are not alike.

Therefore, *artificial voice* signals have been developed (Billi and Scagliola, 1982; Brehm and Stammer, 1987; Hollier et al., 1993). Artificial voices recommended for instrumental evaluations of speech processing systems such as transmission networks or telecommunication devices are defined in ITU-T Rec. P.50 (1999) and ITU-T Rec. P.59 (1993). They are of two kinds so as to reproduce the temporal and spectral characteristics of human female and male voices. Kitawaki et al. (2004) used the ITU-T Rec. P.50 (1999) artificial voices with the intrusive quality model PESQ (ITU-T Rec. P.862.2, 2005). The authors employed two artificial voices and four real voices. Since the quality estimations made from the real voices proved to be highly correlated with the artificial voices ($\rho > 0.95$), they concluded that artificial voices may facilitate instrumental quality assessments of a whole transmission network.

2.3.2.2 Time analysis

These methods use time-domain (i.e. waveform) differences between the reference and the degraded signals denoted by $x(k)$ and $y(k)$, respectively. A widely used instrumental parameter of speech quality, easy to compute and well understood, is the Signal-to-Noise Ratio (SNR). Through use of discrete signals of length N , it calculates the ratio between the energy of the input signal, $x(k)$, and the transmission system-introduced noise, $n(k) = y(k) - x(k)$ as follows:

$$SNR = 10 \log_{10} \frac{\sum_{k=1}^N x(k)^2}{\sum_{k=1}^N [y(k) - x(k)]^2}, \quad (2.6)$$

where k is the sample index. Since this parameter is computed over the entire speech signal, the result of Eq. (2.6) is called either *long-term SNR* or *global SNR*. The estimated value (in dB) is decreasing with reduction in the perceived quality.

The *long-term SNR* is a poor estimator of the perceived speech quality. Since speech sounds are considered as stationary within a time interval of approximately 20ms, speech processing algorithms work on short segments, l , composed of M samples and called a *frame*. A typical frame length lies in the range 10–40ms, see Table 1.2. Following this principle, Zelinski and Noll (1977) defined a *segmental SNR* (SNR_{seg}) as an arithmetic mean of the SNR values (in dB) calculated for individual speech segments.

$$SNR_{\text{seg}} = \frac{1}{L} \sum_{l=0}^{L-1} 10 \log_{10} \left(\frac{\sum_{k=1}^M x(lM+k)^2}{\sum_{k=1}^M [y(lM+k) - x(lM+k)]^2} \right), \quad (2.7)$$

where L is the number of frames. To reach a higher accuracy with the auditory results, SNR_{seg} is usually restricted to the range 0–40dB within each frame. Even if this parameter is more reliable than the *long-term SNR*, it fails to predict a correct ranking between different speech processing systems (Mermelstein, 1979).

These first two SNR-based parameters employ the speech waveform, they are, thus, useful to estimate distortions introduced by additive noise (due to analog transmission or real background noise) or signal-correlated noise (generated by the waveform-coder). However, they are not reliable for other degradations such as filtering or phase distortions (Tribolet et al., 1978).

2.3.2.3 Frequency analysis

As seen in Section 1.1.2.2, the human auditory system performs a spectral analysis. The perceived quality of a speech processing system depends, among other aspects, on the frequency distribution of the degradation over the short-term speech spectrum. SNR-based measurements needs further enhancements in order to estimate such a degradation distribution. A Power Spectral Density (PSD) $\hat{\Phi}_{xx}(l, e^{j\Omega m})$ with $\Omega m = m2\pi/M$ is estimated from the l^{th} segment waveform through application of an M -point Discrete Fourier Transform (DFT) analysis. A time-window analysis such as a Hann-window is usually employed. The resulting spectrum is defined on $m = 1 \dots M/2$ points, i.e. up to half the Nyquist frequency. The third parameter from the SNR family is a frequency-weighted SNR (SNR_{FW}) that uses a frequency-band filtering as follows:

$$SNR_{\text{FW}} = \frac{2}{L \times M} \sum_{l=0}^{L-1} \left(\frac{\sum_{m=1}^{M/2} W(e^{j\Omega m}) 10 \log_{10} \frac{\hat{\Phi}_{xx}(l, e^{j\Omega m})}{\hat{\Phi}_{nn}(l, e^{j\Omega m})}}{\sum_{m=1}^{M/2} W(e^{j\Omega m})} \right), \quad (2.8)$$

where $W(e^{j\Omega m})$ is a long-term frequency-weighting function. Several examples of global and segmental SNR_{freq} are given in Tribolet et al. (1978). One specific global SNR_{FW} parameter is the so-called Articulation Index (AI), which is still used as an intelligibility parameter (French and Steinberg, 1947).

According to Hansen and Pellom (1998), the SNR-based measurements are poor predictors of speech quality; moreover, they are highly dependent on the time-alignment and phase shift between the reference and degraded speech signals. A second type of instrumental measurement methods uses the vocoder technique employed by low bit-rate speech codecs, see Sec. 1.3.4.1. The Linear Predictive Coding (LPC) coefficients are estimated from the reference and the degraded signals. Then, the estimated LPC coefficients are compared with a specific assessment unit. The speech signal spectrum can be transformed to a slightly different frequency scale to better correlate with the perceived distortion. For instance, the Cepstral Distance (CD), developed by Kitawaki et al. (1984), compares the logarithmic spectra of $x(k)$ and $y(k)$. It is calculated as follows:

$$CD \cong 10 \log_{10} \sqrt{2 \sum_{i=1}^p [c_x(i) - c_y(i)]^2}, \quad (2.9)$$

where $c_x(i)$ and $c_y(i)$ are the “cepstrum” coefficients of the reference and the degraded signals, respectively, and p is the prediction order usually in the range 10–15. Other typical LPC-based parameters are:

- Itakura–Saito (IS) distance developed by Itakura and Saito (1968)
- Log Likelihood Ratio (LLR) developed by Itakura (1975)
- Line Spectrum Pair (LSP) distance developed by Coetzee and Barnwell (1989)

In a study by Kitawaki et al. (1982) about the performance of seven time-domain and frequency-domain instrumental measurement methods in quality measurements of NB speech codecs, CD proved to be the best ($\sigma_{CD} = 0.458$). Even though frequency-domain measurement methods are more reliable than time-domain ones, they are not good enough for predicting the auditory quality of a wide range of distortions, in particular they are unsuited to speech quality prediction in case of simultaneous distortions (Tribolet et al., 1978). Some new LPC-based parameters have been significantly enhanced by using masking properties of human auditory system (Chen et al., 2003).

2.3.2.4 Perceptual analysis

With the increasing introduction of non-linear speech processing systems in telephony network, studies are mainly focused on accurate representations of speech signals in the time-frequency domain. The input speech signal is transformed into an auditory nerve excitation through several psychoacoustic processes (see Sec. 1.1.2.2) to further simulate the peripheral hearing system. Usually, this trans-

formation follows the psychoacoustic model for loudness calculation developed by Zwicker and Fastl (1990) and published as the ISO Standard 532-B (1975). Therefore, the term *model* is used for perception-domain measurement methods whereas the term *parameter* is used for time- and frequency-domain ones.

Beerends and Stemerdink (1994) assumed that, in speech quality estimation, the simulation by the perceptual model has not to be alike the activity by the human hearing system. However, they considered that a simulation of simple cognitive processes is necessary. In that case, the assessment unit mimics the “reflexion” phase described in Sec. 1.2.1. The integral quality is calculated from a combination of several quality features. However, the cognitive processes are less developed in current quality models than the perceptual signal transformation, mainly because of the little number of available studies devoted to the cognitive processes performed in the human auditory cortex.

Following the description made by Grancharov and Kleijn (2007), in quality models, mimics of the auditory peripheral system can be achieved through two approaches:

- | | |
|----------------------------------|--|
| Masking threshold concept | The reference is used to compute the degradation-masking threshold, which is in turn used by the assessment unit to calculate the perceptual difference between the reference and degraded speech signals. |
| Perceptual representation | Here, both signals are transformed into signals that contain only the pieces of information essential for the auditory cortex. Then, the assessment unit compares the transformed signals. |

From the late 1970s, perception-domain models have been developed to optimize the quality of waveform coders. Such a model was introduced, at first, by Schroeder et al. (1979), and then extended by Brandenburg (1987), which developed the Noise-to-Masking Ratio (NMR) model. This model assessed audible noise in the degraded speech signal through use of a frequency-masking model. The new requirements by networks have led to the extensive development of perceptual models and to their application to a wider range of distortions. Psychoacoustic features are employed to transform reference and degraded speech signals according to the peripheral auditory system. The first model based on this concept was introduced by Klatt (1982) in the Weighted Spectral Slope (WSS) algorithm to measure a phonetic distance and by Karjalainen (1985) in the Auditory Spectrum Distance (ASD) to compare the audible time (ms) pitch (Bark) amplitude (dB) representations of reference and degraded signals. Intrusive models similar to the ASD method have become predominant in speech quality assessment. Figure 2.2 (p. 41) gives an overview of all intrusive models.

Bark Spectral Distortion (BSD)

The BSD was developed by Wang et al. (1992). The perceptual transformation emulates several auditory phenomena such as (i) the critical band integration in the cochlea (restricted to the first 15 Bark units in the range 50–3400 Hz) and (ii) the loudness compression. The algorithm begins with a Power Spectral Density (PSD) estimation of each 10 ms frame. The frames are overlapped by 50%. The disturbance is computed as the square of the Euclidean distance between both transformed speech signals:

$$BSD = \frac{1}{L} \sum_{l=1}^L \sum_{z=1}^Z [N_x(l, z) - N_y(l, z)]^2, \quad (2.10)$$

where l and z are the frame and Bark index, respectively, L is the number of frames, Z is the number of critical bands, and N_x and N_y are the loudness densities of the reference and the degraded speech signals, respectively. In the study by Wang et al. (1992), a relatively high Pearson correlation coefficient between the predicted quality scores, MOS_{LQON} , and the auditory quality scores, MOS_{LQSN} , was achieved ($\rho = 0.85$) for the BSD algorithm. Then, the Modified Bark Spectral Distortion (MBSD) model was developed by Yang et al. (1998) to incorporate a noise-masking threshold in order to differentiate the audible distortions from the inaudible ones. When the absolute loudness difference between the reference and the degraded speech signals is less than the loudness of the noise-masking threshold, this difference is considered as imperceptible and, thus, set to 0. A comparison of the correlation coefficients respectively produced by the MBSD and the conventional BSD models ($\rho_{MBSD} = 0.956$ against $\rho_{BSD} = 0.898$) showed that the former worked better (Yang et al., 1998). Then, Yang (1999) developed an Enhanced Modified Bark Spectral Distortion (EMBSD) from the MBSD model by including a new cognitive model to simulate a time-masking effect by the perceptual distortion. The resulting improvement was confirmed by the finding of a higher correlation coefficient: $\rho_{EMBSD} = 0.98$ and $\rho_{MBSD} = 0.95$.

Perceptual Speech Quality Measure (PSQM)

The Perceptual Audio Quality Measure (PAQM) and the Perceptual Speech Quality Measure (PSQM) were both developed by Beerends and Stemerdink in 1992 and 1994, respectively. The former is devoted to audio assessments whereas the latter is dedicated to speech assessments. Both of them employ a high level psychoacoustic model. Moreover, in the PSQM model, the perceptual transformation is optimized for speech signals. In addition, Beerends (1994) assumed that a perceptual model was unable to model speech quality and, thus, that a cognitive model was needed. For instance, signal components added by the system under study are much more annoying than components which are attenuated. This effect is quantified by an “asymmetry” factor which has been included in the PSQM model (Beerends and

Stemerding, 1994).

The following four intrusive quality models were studied by the Comité Consultatif International Téléphonique et Télégraphique (CCITT), the predecessor of the International Telecommunication Union (ITU), for possible recommendations:

- Information Index (II) (Lalou, 1990),
- Cepstral Distance (CD) (Kitawaki et al., 1984),
- Coherence Function (CHF) (Benignus, 1969) and
- Expert Pattern Recognition (EPR) (Kubichek et al., 1989).

But, as none of them achieved a minimum amount of accuracy on auditory results, no recommendation was published by this organization. However, after evaluation, the PSQM model was approved as an ITU-T standard and further published as the ITU-T Rec. P.861 (1996)⁶. It was developed and tested on numerous auditory tests carried out during the development of the ITU-T Rec. G.729 (2007) speech coding algorithm (published as ITU-T Suppl. 23 to P-Series Rec. (1998)). The PSQM model gives reliable estimations for low bit-rate speech codecs. An improved version of the PSQM model, called PSQM+ (ITU-T Contrib. COM 12-20, 1997), was further developed. Its scope is wider than that of the PSQM. This updated version predicts distortions due to transmission channel errors, such as time clipping and packet loss.

Measuring Normalizing Blocks (MNB)

Following the idea introduced by Beerends and Stemerding (1994), Voran (1999a) developed a model called Measuring Normalizing Blocks (MNB) with a rather simple perceptual transformation and a more sophisticated assessment unit. From the assumption of differences in the opinions expressed by listeners about short- and long-term spectral deviations, the author developed a family of perceptual analyses that covers multiple frequency- and time-scales. The resulting parameters were linearly combined to form the perceptual distance between the original and the degraded signal. A comparison against ITU-T Rec. P.861 (1996) showed that the correlation coefficients yielded by the MNB model were higher for transmission errors and low bit-rate speech codecs than those issued by the PSQM (Voran, 1999b). Consequently, in 1998, the ITU-T published an alternative intrusive quality measure to the PSQM quality model as an annex to the ITU-T Rec. P.861 (1996).

Telecommunication Objective Speech-Quality Assessment (TOSQA)

To assess speech coding algorithms, Berger (1996), developed a quality model called Deutsche Telekom—Speech Quality Estimation (DT-SQE). But, contrary to

⁶ In 2001, the ITU-T Rec. P.861 (1996) was replaced by the ITU-T standard ITU-T Rec. P.862 (2001).

the other models, DT-SQE does not compute a distance, but a similarity between the reference speech signal, $x(k)$, and the degraded speech signal, $y(k)$, according to:

$$BSA = \sqrt{\frac{1}{L} \sum_{l=1}^L r_l \{N_x(l, z), N_y(l, z)\}^2}, \quad (2.11)$$

where r_l is the correlation parameter between the loudness densities, and the *BSA* stands for “Barkspektrale Ähnlichkeit” (Bark-Spectral Similarity). Its basic structure matches the common structure and components of perceptual-based models. The pre-processing stages and those of the PSQM are alike. However, the DT-SQE model considers additional characteristics not taken into account in previous intrusive speech quality models. Firstly, $x(k)$ and $y(k)$ are both filtered by a standard 300–3400 Hz bandpass filter to simulate the listener’s terminal. Because the model may also be applied to the acoustic signals available at the talker’s and listener’s terminals, the input signal, $x(k)$, can be additionally filtered with a modified IRS sending characteristic. Then, the delay κ between the reference and the degraded signals is compensated. However, the variable delay is estimated from the highest possible similarity between reference and degraded frames. Then, the reference and the degraded signals are split and Hann-windowed into 16 ms-long segments. The spectrum of each segment is calculated by a DFT. The spectra are transformed to the perceptual domain by using the loudness calculation model developed by Zwicker and Fastl (1990). The resulting loudness pattern of the reference speech file is modified to reduce inaudible effects, known to only very slightly affect the perceived integral quality in a Listening-Only Test (LOT). For instance, the linear spectral distortions due to the frequency response of the system under study have a small effect on the perceived quality⁷ and are consequently eliminated. The computed similarity is the main speech quality result of the DT-SQE model. The simple arithmetic mean value of all short-time similarities yields the final quality score.

An updated version of the DT-SQE model, called Telecommunication Objective Speech-Quality Assessment (TOSQA) is available in the ITU-T Contrib. COM 12–34 (1997). However, the standard deviations of the prediction errors by TOSQA are slightly higher than those by PSQM: $\sigma_{TOSQA} = 0.31$ and $\sigma_{PSQM} = 0.28$. In ITU-T Contrib. COM 12–19 (2000), the TOSQA model was extended in the so-called “2001 version” of TOSQA. Furthermore, a variable gain compensation, an adaptive threshold for the internal VAD as well as a modified background noise calculation to take into account CNG algorithms were included in this improved version. A comparison of the two models on the database “Sup23 XP1” described in ITU-T Suppl. 23 to P-Series Rec. (1998) and App. B showed that the TOSQA-2001 model performed slightly better than TOSQA with, for example, $\rho_{TOSQA-2001} = 0.961$ against $\rho_{TOSQA} = 0.953$ for the auditory test “Sup23 XP1-D”. In addition, the replacement of the IRS Receive filter by a 200–7000 Hz passband filter and that of the modified IRS Send filter applied to $x(k)$ by a flat filter both permitted its adaptation to

⁷ Only when it falls within a small dynamic range (e.g. ± 10 dB).

WB transmissions. Since this improved version can work with either electrically- or acoustically-recorded input signals, it can assess the impact of terminals (i.e. electro-acoustic interfaces), such as degradation due to non-ideal frequency responses or non-linear transducer distortions. In ITU-T Contrib. COM 12–20 (2000), an evaluation of TOSQA-2001 on an auditory database including WB conditions and acoustic recordings led to a correlation coefficient of $\rho = 0.91$ between the TOSQA-2001 MOS_{LQOM} estimations and the auditory MOS_{LQSM} values.

Asymmetric Specific Loudness Difference (ASLD)

Hauenstein (1997) developed a new quality model called ASLD model, which includes a psychoacoustic model developed especially for speech signals. The structure of the ASLD model includes several pre-processing stages: (i) both the overall delay and overall gain of the system are compensated, (ii) the pauses are eliminated by a VAD, and (iii) two filters simulate the NB telephone bandwidth and the frequency response of the average telephone handset, respectively. Then, the perceptual transformation includes a fine tuned temporal and frequency masking effect. The computing effort required by this model is, therefore, high. The quality score is computed as a weighted sum of positive and negative disturbances. This algorithm follows the asymmetric factor principle developed by Beerends (1994). A comparison of ASLD to several intrusive models in ITU-T Contrib. COM 12–34 (1997) showed that it yielded a higher correlation coefficient than PSQM and TOSQA.

PERception Model—Quality assessment (PEMO-Q)

Hansen and Kollmeier (1997) applied an advanced model of auditory perception developed by Dau et al. (1996) to speech quality estimations. This “effective” auditory signal model simulates the transformation of acoustic signals into neural activity patterns by the human ear. This model contains a Gammatone filter bank, a model of the human hair cells and an adaptation loop to model critical band integration (on 19 Bark units) and temporal masking effects. The resulting speech quality score, called q_C , corresponds to a similarity measure between the weighted perceptual representations of the reference and the degraded speech signals. However, this model is not fully adapted to any type of degradations because of a lack of pre-processing stages (e.g. speech activity detection function is missing). Therefore, the evaluation of q_C is worse than those by PSQM, TOSQA and ASLD in ITU-T Contrib. COM 12–34 (1997): $\rho_{q_C} = 0.90$ against $\rho_{PSQM} = 0.93$, $\rho_{TOSQA} = 0.92$ and $\rho_{ASL} = 0.96$. Huber and Kollmeier (2006) expanded Hansen’s model to Full-Band (FB) audio signal assessments. New developments on the model of auditory perception by Dau et al. (1997) were included in this expanded model called PEMO-Q⁸. Conversely to

⁸ The acronym comes from the perception model developed by Dau et al. (1996).

the q_C model, this extended version estimates the perceived quality of any kind of distortion and any kind of audio signal (including speech).

Perceptual Ascom Class Enhanced (PACE)

The perceptual quality model developed by Juric (1998) is called PACE and was designed, at first, for quality estimation of overall transmissions, especially in the field of mobile communications. In addition to the components included in the other intrusive models (e.g. time-alignment, critical band filtering, ...), the PACE model contains an algorithm dedicated to “importance-weighted” comparisons of the perceptually transformed original and degraded signals: this algorithm assumes that signal parts with a high energy are more important for the perceived speech quality. This model is integrated to the *Qvoice* equipment evaluation framework developed by the telecommunication company “Ascom”. The evaluation of PACE, in ITU-T Contrib. COM 12–62 (1998), on the “Sup23 XP1” and “Sup23 XP3” databases described in ITU-T Suppl. 23 to P-Series Rec. (1998) and App. B highlighted its capability by resulting in high correlation coefficient values on both databases: $\rho_{XP1} = 0.96$ and $\rho_{XP3} = 0.94$.

Perceptual Analysis Measurement System (PAMS)

Hollier et al. (1994) developed a specific description of audible errors introduced by the system under study. The differences between the original and the degraded speech signals are represented by an *error surface*. Briefly, the error entropy (i.e. distribution of errors) enables one to extract several error descriptors and to quantify the total amount of errors. This description led to a first model used for quality estimations of speech coding algorithms (Hollier et al., 1995).

For extension of this version to quality estimations of overall transmissions (referred to as “end-to-end”) including electro-acoustic interfaces, further developments were made by Rix et al. (1999). This quality model, called PAMS was developed for the assessment of recent voice technologies, including packet-switched networks where the time-delay between the reference and the degraded speech signals is liable to vary. Its estimation was a significant challenge in the late 1990s. This is why the PAMS model contains a robust time-alignment algorithm to precisely estimate the time-delay introduced by the transmission system. It is worth underlining that, in case of variable delay over the whole speech file, especially variation during a pause, the other quality models do not re-align the original and the degraded speech signals. In addition, following Berger’s concept, the PAMS model compensates the linear degradations of low impact on the perceived quality.

Rix and Hollier (1999) extended it to WB transmissions assessment by calibrating the perceptual layer that extracts the error descriptors so as to produce quality scores on the MOS_{LQOM} quality scale, which is called W_{LQ} by the authors.

Perceptual Evaluation of Speech Quality (PESQ)

Since the publication of the PSQM model as the ITU-T Rec. P.861 (1996) for instrumental quality measurements of transmitted speech, telephony networks have broadly changed: indeed, further to the introduction of highly non-linear degradations, quality is no longer kept at a constant level over an entire call. Unfortunately, ITU-T Rec. P.861 (1996) is unsuited to these networks and poorly correlated to the perceived integral quality (Thorpe and Yang, 1999). Consequently, the ITU-T has been working on the development of an overall speech quality model as a successor to the ITU-T Rec. P.861 (1996). Five measurement algorithms were, thus, been proposed, namely PACE (*Ascom*), PAMS (*British Telecom*), TOSQA (*Deutsche Telekom*), VQI⁹ (*Ericsson*) and PSQM99¹⁰ (*KPN*). Across 22 auditory experiments, the PAMS and PSQM99 models gave the highest average correlation coefficient between the model estimations and the auditory quality scores (i.e. $\rho_{PAMS} = 0.92$ and $\rho_{PSQM99} = 0.93$). But, as the five models failed to fulfill all of the requirements for minimum performances, none of them was declared overall winner. The whole statistical evaluation was published in ITU-T Contrib. COM 12-117 (2000). Consequently, the strongest components of both PAMS and PSQM99 model were integrated into a new algorithm denoted by PESQ. They consisted of (i) the perceptual transformation of the PSQM99 model (Beerends et al., 2002) and (ii) the time-alignment algorithm of the PAMS model (Rix et al., 2002). The average correlation coefficient found, over the same 22 databases, between the PESQ estimations and the auditory quality scores was $\rho = 0.935$ (Beerends et al., 2002).

This new model was standardized as the ITU-T Rec. P.862 (2001). Thus, the different parts of this quality model are the result of evolutions over more than ten years. Since 2001, PESQ is the most widely used instrumental model. As the original input file (and its related degraded version) of the PESQ model need to follow some simple rules so as to avoid inconsistently estimated MOS values, the ITU-T published an application guide termed as the ITU-T Rec. P.862.3 (2005). These guidelines provide one with relevant pieces of information required to get, in practice, stable, reliable and meaningful instrumental measurement results. In addition, to check whether the PESQ model has been correctly implemented, a procedure that uses the ITU-T Suppl. 23 to P-Series Rec. (1998) databases is described in detail in the ITU-T Rec. P.862.3 (2005).

⁹ VQI stands for Voice Quality Index (VQI).

¹⁰ During the selection phase the name PSQM99 (PSQM, 1999 version) was used instead of PSQM+. Several improvements were included in the PSQM+ before its submission.

The PESQ model estimates the perceived quality of transmitted speech for the classical NB telephone bandwidth. In 2005, the PESQ model was extended to the evaluation of WB transmissions, and this WB mode of PESQ, called Wideband-Perceptual Evaluation of Speech Quality (WB-PESQ), was standardized as the ITU-T Rec. P.862.2 (2005). The algorithm of WB-PESQ is very similar to the one used by PESQ for NB signals. However, the WB-PESQ uses exclusively speech signals with a sampling frequency of $f_s = 16\,000\text{ Hz}$.

Project—Acoustic Assessment Model (P.AAM)

An acoustic model, like TOSQA, can work on acoustic recordings of speech signals transmitted over an electro-acoustic interface (e.g. handset, headset or loudspeaking telephones). Such a model is able to assess the influence of the transducers (i.e. microphone and loudspeaker) on the speech quality. Acoustic recordings are made by an artificial head (ITU-T Rec. P.58, 1996), which simulates the speech production and perception processes. Impairments introduced by acoustic components are outside the scope of the ITU-T Rec. P.862 (2001). This model uses electrical signals which means that the two input speech signals are electrically recorded in the transmission system. However, the user terminals may include complex signal processing systems (especially in mobile and DECT telephones, see Sec. 1.3.3). Studies have been devoted to updates of the PESQ model in order to further estimate the quality of acoustic path. For instance, in ITU-T Del. Contrib. COM 12–6 (2001) the PESQ model was extended to quality measurements of monaural acoustic interfaces in listening environments with background noise. In addition, in ITU-T Del. Contrib. COM 12–41 (2001) a post-processing tool for intrusive models such as PESQ was proposed for quality measurement of binaural recordings.

The ITU-T has worked on the selection of a new standard dedicated to acoustic quality measurements. The requirements of this ITU-T project, called P.AAM, were described in ITU-T Contrib. COM12–42 (2002). Three measurement algorithms have been proposed by *Psytechnics*, *TNO* and *Deutsche Telekom*. A report on the statistical evaluation of the three models is available in the ITU-T Del. Contrib. COM 12–109 (2003).

Then, the three proponents announced their co-operation to develop one single model integrating the best components of each individual model. The three models are all based on changes in the PESQ model (ITU-T Rec. P.862, 2001). The resulting integrated P.AAM model was described in Rix et al. (2003) and Goldstein and Rix (2004). The P.AAM model differs from the PESQ model by the following improvements:

- Quality estimation in noisy listening environments.
- Extension to acoustic quality measurements.
- Extension from monaural to binaural acoustic interfaces.

Unfortunately, the submitted integrated model failed to meet the minimum requirements (i.e. correlation coefficients) in all cases. Consequently, the development was stopped in September 2003 (ITU-T Temp. Doc. TD.10 Rev.1, 2003).

Perceptual Objective Listening Quality Analysis (POLQA)

In 2007, ITU-T launched a new standardization program, Perceptual Objective Listening Quality Analysis (POLQA) (ITU-T Temp. Doc. TD.52 Rev.1, 2007), aimed at selecting an intrusive speech quality model suitable for NB to S-WB connections and electro-acoustic interfaces and able to compensate for the defects observed in the PESQ model. The future ITU-T standard POLQA is expected to predict reliably the integral “speech transmission quality” for fixed, mobile and IP based networks. Speech quality models have been proposed by six proponents: *Opticom*, *SwissQual*, *TNO*, *Psytechnics*, *Ericsson* and a consortium formed by *France Télécom* and *Deutsche Telekom*. The model developed by *Ericsson* was published by Ekman et al. (2011). The DIAL model developed by the latter proponent is presented in Chap. 4. According to the selection criteria, detailed in Sec. 5.1.3, three proponents *OPTICOM*, *SwissQual* and *TNO* met the requirements. They agreed to combine their proposed algorithms into a joint model called POLQA. This joint model led to an improved version selected as a new ITU-T recommendation. A draft version of this new recommendation is available in ITU-T Contrib. COM 12–141 (2010).

2.3.2.5 Diagnostic measures

The intrusive quality measurement methods introduced in the previous Sects. 2.3.2.2, 2.3.2.3 and 2.3.2.4 give a single estimated quality score MOS_{LQO} which represents the integral perceived quality of the assessed speech signal. The possible occurrence of two degradations at the same time and in such a way that the integral quality is kept unchanged makes necessary the development of *diagnostic* measurements based on a decomposition of the integral quality into several attributes. For instance, the system under study can be characterized by several physical attributes such as its overall gain, its frequency response and its SNR. However, these information data are useless for the end user and give no insight into the influence of system parameters involved in user’s perception.

According to Jekosch (2005), a **diagnosis** is:

the production of a system performance profile with respect to some taxonomization of the space of possible inputs.

Such diagnostic measurement should rely on quality-attributes or -features defined, here, in Sec. 1.2.1. These quality features can be derived from a multidimensional analysis of the auditory results, see Sec. 2.2.4.6. The following paragraphs make a brief recall on several instrumental measurement methods in use.

- Quackenbush et al. (1988) investigated the correlation between auditory results from a DAM experiment and several quality estimators based on time-analysis (e.g. segmental SNR and frequency-dependent SNR) and frequency-analysis (e.g. LPC-based spectral distances). From the ten original quality features assessed in the DAM experiment, four quality features were selected:

SL Speech Lowpass: muffled, smothered.
BN Background Noisy: hissing, rushing.
SI Speech Interrupted: irregular, interrupted.

Here, the acronyms refer to the DAM scales, see Voiers (1977). For each feature a corresponding estimator was developed (O_{SL} , O_{BN} , O_{SI} and O_{SH}). Their linear combination into an estimation of the integral speech quality led to a correlation coefficient of $\rho = 0.75$ with the auditory results. More recently, Sen (2004) derived a set of three orthogonal dimensions from a PCA analysis applied to DAM auditory results. The author developed a new set of three estimators for the corresponding perceptual dimensions: **SH** (Speech Highpass), **SL** and **BNH** (Background Noise Hiss).

- Halka and Heute (1992) decomposed the degradation into two quality features: (i) the linear distortion described by an average spectrum of the degraded signal $\overline{\Phi}_{yy}(e^{j\Omega})$, and (ii) the nonlinear distortion by an average “noise” spectrum $\overline{\Phi}_{nn}(e^{j\Omega})$. Then, an integral quality score d_{nlsd} was calculated from the spectral distance issued from the two signal spectra. One should note that these authors used a specific reference signal. In order to reduce the impact by the talker’s characteristics, they recommend employing an artificial signal instead of natural speech signals. This test signal corresponds to a Spherically Invariant Random Process (SIRP).
- The model of loudness calculation developed by Zwicker and Fastl (1990) is based on the Bark scale concept. It is used by several speech quality models, e.g. MNB and TOSQA. Glasberg and Moore (2002) developed a model based on a similar concept called Equivalent Rectangular Bandwidth scale, ERB_N (N for normal hearing). Both scales reflect the concept of critical band filters introduced in Sec. 1.1.2.2. Contrary to the former used for steady sounds, Glasberg and Moore (2002)’s model can be used for time-varying stimuli, e.g. speech or music.

From this loudness model, Moore et al. (2004) developed a perception-oriented approach to model how the perceived speech and music quality is affected by mixtures of linear and nonlinear distortions. For this purpose, a linear estimator

and a nonlinear one, respectively denoted by S_{lin} and S_{nonlin} , were developed to assess the impact of linear distortions (Moore and Tan, 2004) and that of nonlinear distortions (Tan et al., 2004). Then, the two predicted scores are combined as follows:

$$S_{\text{overall}} = \alpha S_{\text{lin}} + (1 - \alpha) S_{\text{nonlin}} , \quad (2.12)$$

where $\alpha = 0.3$. S_{lin} measures the *coloration* or *naturalness* as a function of the changes induced by the system under study in the “excitation pattern”, i.e. the image of the sound spectrum on the ERB_N scale. S_{nonlin} measures the *harshness*, *roughness*, *noisiness* or *crackling* in the audio signals. These degradations correspond to the introduction of frequency components missing in the original signal. This estimator uses an array of 40 Gammatone filters (1 ERB_N bandwidth) so as to cover the whole audible frequency range (50–19 739 Hz). Then a similarity value is calculated between the perceptually transformed reference and degraded signals. The final model was evaluated on both music and speech stimuli (Moore et al., 2004) through auditory quality tests taking into account artificial conditions (i.e. digital filters) and acoustic recordings of transducers. Correlation coefficients of $\rho = 0.85$ and $\rho = 0.90$ are obtained for speech stimuli and music stimuli, respectively, between the estimations of the overall quality, S_{overall} , and the auditory scores.

- The quality features instrumentally measured by either Quackenbush et al. (1988), or Halka and Heute (1992) or Moore et al. (2004) and the perceptual quality dimensions described in Sec. 1.5, which are by definition orthogonal, are not alike. In addition, Heute et al. (2005) showed that the composite scores issued from a combination of quality estimators are outperformed by almost all of the intrusive quality models that give a single integral quality score. The finding let them to assume that a reliable diagnostic model must rely on quality features that must:
 - be in a small number,
 - correspond to perceptual dimensions,
 - be estimated by a reliable instrumental measure,
 - combined to give an integral quality score,

Then, from the perceptual quality space derived by Wältermann et al. (2006b) and described in Sec. 1.5, a set of three quality estimators has been developed by Scholz and Heute (2008). Briefly, each estimator quantifies the perceived quality on one out of the three next perceptual dimensions:

- *Directness/Frequency Content (DFC)*
- *Discontinuity*
- *Noisiness*

The estimator of the quality dimension, *DFC*, was defined in Scholz et al. (2006) to measure the linear frequency degradation introduced by a transmission system. From a perceptual representation, the bandwidth and the slope, β , of the system frequency response are expressed in terms of ERB and dB per Bark, respectively.

The estimator of the dimension *Noisiness*, defined in Scholz et al. (2008), quantifies two types of noise: (i) the *additive noise*, estimated by its level, $NL^{(add)}$, and its center of gravity, $f_G^{(add)}$, and (ii) the amount of *signal-correlated noise*, estimated and quantified by the parameter $N^{(cor)}$. A third estimator for the perceptual dimension *Discontinuity* was defined in Scholz (2008) and uses three parameters: (i) the Interruption Rate (*IR*), which quantifies the percentage of silence insertion in speech segments, (ii) the Clipping Rate (*CR*), which defines the percentage of silence insertion at the start or at the end of speech segments (also known as Front/End Clipping), and (iii) the Additive Instationary (*AI*) distortions, which quantify the influence of the musical noises contained in the speech signal. One should note that the packet losses concealed by PLC algorithm are not considered by this last estimator. Finally, an integral quality estimator was developed by Scholz and Heute (2008) from these three estimators. It relies on the linear combination of the perceptual dimensions introduced by Wältermann et al. (2006b). Its evaluation led to a correlation coefficient of $\rho = 0.862$ between the estimated MOS_{LQON} values and the auditory MOS_{LQSN} values.

- A second set of three estimators for the same perceptual dimensions was developed by Huo et al. (2008a,b, 2007) to cover a wider scope of estimated transmission systems. Indeed, the former is restricted to only NB transmissions whereas the latter can estimate the quality of NB and WB transmissions. The estimator for the perceptual dimension *DFC*, defined in Huo et al. (2007), includes two new parameters: *sharpness* (S) and *reverberation time* (T_{30}). Sharpness is used in place of the center of gravity, z_G , and reverberation time, T_{30} , estimates the impact by the room in the case where an HFT is used. The estimator for *Noisiness*, defined in Huo et al. (2008a), uses Cepstral Distances (d_{cep}) and differentiates the contributions by high-frequency, n_{hf} , and low-frequency, n_{lf} , additive noises. From the Weighted Spectral Slope (WSS) distances and signal temporal loss, the estimator for *Discontinuity*, defined in Huo et al. (2008b), derives three parameters: (i) the interruption rate, r_I , (ii) the artifact rate, r_A , and (iii) the clipping rate, r_C . A comparison against the estimator developed by Scholz (2008) showed its greater efficiency for packet-loss impairment.
- From a model of human perception developed by Sottek (1993), Genuit (1996) developed an instrumental measurement method, termed Relative Approach (RA), for specific assessment of acoustic quality. This method is applicable to acoustic recordings of environmental noise (e.g. within an office or a car). According to Genuit (1996), a characteristic of human hearing is that humans are more affected by quite fast level variations than by slow changes. The RA compares the instantaneous signal with a “smooth” estimated version of the signal. Since temporal and spectral structures have both an influence on the perceived noise-induced annoyance, the comparison is made within each frequency band over time and within each time window over the whole frequency range. The model gives a three dimension representation of the spectrum displaying the amount of annoyance for each time-frequency cell. In addition, contrary to other

diagnostic models, the RA does not require a reference signal.

Gierlich et al. (2008a) updated the RA to the diagnostic assessment of NB and WB communications in the presence of background noise. This model is based on the ITU-T Rec. P.835 (2003) auditory method which uses three rating scales. The resulting diagnostic model estimates three quality values, the *Speech MOS* (*S-MOS*), the *Noise MOS* (*N-MOS*) and the *Global MOS* (*G-MOS*, i.e. the integral quality including speech and background noise). In addition to the reference and the degraded signals used by intrusive quality models, this diagnostic model needs a third signal referred to as “unprocessed” and corresponding to the talker’s terminal input before any processing or transmission. This third signal is a mixture of both the speech and the background noise. This method is applicable to assessment of user terminals (at the talker’s side only), environmental noises, Noise Reduction (NR) algorithms, WB speech codecs and VoIP networks. In both WB and NB contexts and for all the three estimated scores, it proved to be accurate ($\rho_{\text{GMOS, WB}} = 0.935$ and $\rho_{\text{GMOS, NB}} = 0.932$). This diagnostic model was published by the ETSI as the ETSI EG 202 396-3 (2007).

- Beerends et al. (2007) proposed a diagnostic model mainly based on the intrusive quality model, PESQ. Three quality features close to the perceptual dimensions derived by Wältermann et al. (2006b): the first degradation indicator quantifies the impact by additive noise in silent frames. Mapping of the estimated parameter, *Noise Distortion* (*ND*), to the MOS scale gives an objective MOS noise value, $OMOS_{\text{NOI}}$. The second indicator quantifies the effect generated by deviation in the linear frequency response from computation of the frequency response of the system under study for only speech segments with a high loudness. After a noise and an overall gain compensation, an $OMOS_{\text{FRQ}}$ value is obtained by mapping the global *Frequency Distortion* (*FD*) on a MOS scale. The last indicator quantifies the time-varying impairments such as packet losses and pulses by using two mapped MOS values: an $OMOS_{\text{TIM-CLIP}}$ value (i.e. time clipping) and an $OMOS_{\text{TIM-PULSE}}$ value, which respectively give account for the local decrease and the local increase. Both values are computed after noise and frequency response compensation. The overall $OMOS_{\text{TIM}}$ value is defined as the minimum over the two $OMOS_{\text{TIM-CLIP}}$ and $OMOS_{\text{TIM-PULSE}}$ values.

Beerends et al. (2007) evaluated the three degradation indicators using a specific auditory test procedure close to the one described in ITU-T Contrib. COM 12–82 (2009). In this auditory test, the speech stimuli have been degraded by a mixture of noise, frequency response and time-varying distortions. Moreover, they assumed that the integral quality was dominated by the worst degradation. Therefore, they derived the integral quality from the three degradation indicators as the minimum over the 3 MOS values as follows:

$$OMOS_1 = \min \{ OMOS_{\text{NOI}}, OMOS_{\text{FRQ}}, OMOS_{\text{TIM}} \} , \quad (2.13)$$

The correlation coefficient between the resulting $OMOS_1$ estimations and the integral quality scores was $\rho = 0.82$. Use of the aggregation expressed in Eq. (2.14)

$$OMOS_2 = \left(\frac{\sqrt[10]{OMOS_{NOI}} + \sqrt[10]{OMOS_{FRQ}} + \sqrt[10]{OMOS_{TIM}}}{3} \right)^{10}, \quad (2.14)$$

led to a slight improvement with a correlation coefficient of $\rho = 0.85$ obtained between the $OMOS_2$ values and the integral quality scores.

- Current work in ITU-T is focusing on the development of a new diagnostic model, called Perceptual Approaches for Multi-Dimensional Analysis (P.AMD), which is able to estimate the four quality features introduced in Sec. 1.5, namely, *discontinuity*, *coloration*, *noisiness*, and non-optimum *loudness* (ITU-T Contrib. COM 12–143, 2010). Additional dimensions might be included, such as low-frequency coloration, high-frequency coloration, fast-varying time-localized distortions and slowly-varying time-localized distortions.

2.3.2.6 Non-intrusive models

A crucial step in intrusive models is the alignment of the reference and the degraded speech signals. Indeed, a perfect alignment is difficult to achieve with signals transmitted by packet-switched networks introducing a variable delay. A wrong synchronization results in a dramatic decrease of the model accuracy (Rix et al., 2002). In addition, the signal under study and the reference signal are both required by intrusive models. But, in some important applications (e.g. network monitoring), the reference signal is unavailable. Therefore, it has made necessary the development of “non-intrusive” models.

Non-intrusive measurement methods rely on two different approaches: (i) a priori-based approach, and (ii) source-based approach. In both cases, several parameters are derived from the degraded speech signal and may describe either perceptual features (e.g. LPC coefficients) or physical characteristics (e.g. speech level in dB).

A priori-based approach

In the *a priori*-based approach, at first, a set of known distortions is characterized by several parameters. Then, a relationship between this finite set of distortions and the perceived speech quality is derived. This approach is usually based on machine learning techniques such as Gaussian mixture models or artificial neural networks. In this case, the parameters characterizing the set of known distortions are stored in

an optimally clustered codebook.

For instance, Au and Lam (1998) inspected visual characteristics of the speech spectrogram to detect noise or frequency distortions. Another example corresponds to the ITU-T Rec. P.561 (2002) and ITU-T Rec. P.562 (2004). The former recommendation defines an In-service Non-intrusive Measurement Device (INMD) that quantifies physical characteristics in live call traffic such as speech level, noise level, echo loss and echo path delay. ITU-T Rec. P.562 (2004) shows how one can use INMDs to predict perceived speech quality through use of two parametric models: the Call Clarity Index (CCI) and the E-model previously introduced in Sec. 2.3.1.3. It is worth noting that none of them is based on machine learning techniques. Briefly, CCI provides an estimating MOS_{CQE} value on the conversational quality scale defined in ITU-T Rec. P.800 (1996), whereas the parametric E-model gives an estimated R -value on the Transmission Rating scale defined in ITU-T Rec. G.107 (1998).

Another type of *a priori*-based non-intrusive models uses the likelihood that the degraded speech signal has been produced by the human vocal system. The speech signal is reduced to few speech features related to physiological rules of voice production. The derived parameters are then combined and mapped to a quality scale. This approach was followed by Gray et al. (2000) who ran a vocal tract model to detect distortions in the transmission system. A non-intrusive model, including the vocal tract model developed by Gray et al. (2000), was published by the ITU-T as the ITU-T Rec. P.563 (2004). It relies on the association of three principles: (i) a derivation, from the degraded signal, of several parameters related to the voice production mechanism, (ii) after reconstruction of a reference signal from a degraded signal, both signals are assessed by an intrusive model, and (iii) detection of specific distortions in the degraded signal. Then, the derived parameters are linearly combined to predict a speech transmission quality. Over this aggregation step, the perceptual impact of each parameter is quantified through a distortion-dependent weighting operation. In an evaluation of ITU-T Rec. P.563 (2004) on the “Sup23 XP1” and “Sup23 XP3” databases (Malfait et al., 2006) described in ITU-T Suppl. 23 to P-Series Rec. (1998) and App. B of this book, the correlation coefficients obtained by the PESQ model for all the seven auditory tests were higher, for example, $\rho_{PESQ} = 0.957$ against $\rho_{P.563} = 0.842$, for the auditory test “Sup23 XP1-D”. Kim et al. (2004) developed a non-intrusive model called Auditory Non-Intrusive Quality Estimation (ANIQUE), where the naturalness in degraded speech signals is detected by a machine learning technique. The performances of this model proved to be less than those of ITU-T Rec. P.563 (2004).

Source-based approach

In the *source*-based approach, an artificial reference signal is selected from parameters characterizing the degraded speech signal. Then, the selected artificial reference

is compared to the degraded signal. Like the *a priori*-based approach, this kind of non-intrusive quality models usually relies on machine learning techniques. Moreover, here, the codebook saves parameters derived from a large set of reference speech materials, and the range of estimated distortion types is wider than the one by *a priori*-based models.

An example of a *source*-based approach was applied by Liang and Kubichek (1994). After derivation of Perceptual Linear Prediction (PLP) coefficients from the degraded speech signal, the authors selected, from these PLP coefficients, an artificial reference signal to further calculate the Euclidean distance between the artificial reference and the degraded speech signals. Falk and Chan (2006) developed a *source*-based non-intrusive model that uses a neural network and showed improvements with respect to ITU-T Rec. P.563 (2004).

2.3.3 *Packet-layer models*

A quality model can be designed to assess certain processing conditions or networks, e.g. for network monitoring purposes. However, the high number of assessment requests needed to monitor the whole transmission system implies simplifications in the algorithm complexity. Packet-layer models enable one to consider both these constraints and the need of a reliable instrumental model. These methods measure, in gateways, or at the listener's side, several network-related and IP packet pattern-based parameters such as transmission delay, packet-loss percentage and burst ratio. Compared to intrusive methods, packet-layer models are less complex and require less memory. In addition, packet-layer quality measurement methods associate the advantages of parametric models to those of signal-based models. They use several parameters provided by the packet-switched network and estimations from simple non-intrusive models. The current ITU-T standard is the ITU-T Rec. P.564 (2007).

2.4 Summary and Conclusion

This chapter reviewed in detail the auditory and instrumental measurement methods dedicated to assessments of perceived speech quality. In particular, it highlighted the standards published by organizations such as ITU-T, ETSI or ISO. One should be aware that, as the described auditory methods are usually set in practice within laboratories, the opinions expressed by subjects are affected by this specific listening environment. This means that getting an absolute quality value is by nature impossible, but thanks to the procedures made available by these organizations, biases on the quality judgments can be reduced. In other aspects, this chapter presented the historical evolution of instrumental quality models from the first one defined by

Fletcher and Arnold (1929) to the future POLQA standard (ITU–T Contrib. COM 12–141, 2010). This description also included the target applications with the corresponding performances of the instrumental quality models.

Amongst the three instrumental quality measures recommended by the ITU-T, which are: (i) the parameter-based E-model (ITU–T Rec. G.107, 1998), (ii) the non-intrusive ITU–T Rec. P.563 (2004), and (iii) the intrusive PESQ (ITU–T Rec. P.862, 2001), the third one, PESQ, proved to be the most accurate on auditory quality scores (Falk and Chan, 2009). Many intrusive quality models, described in Sec. 2.3.2.4, simulate the human peripheral auditory system (i.e. represent the signal at the output of the inner ear). But, this paradigm shows limitations. A model of cognitive processes is, thus, a must to increase the accuracy of instrumental measurement methods. Ideally, as done by human subject, the instrumental measurement methods should interpret the perceptual dimensions involved in the assessment process. However, some cognitive effects are already simulated by instrumental models:

- Linear distortion is generally less objectionable than nonlinear distortion (Thorpe and Rabipour, 2000).
- Speech correlated distortion has a greater impact on the perceived quality than uncorrelated noise (Leman et al., 2008).
- Distortions on time-spectrum components that carry information (e.g. formants) have a high impact on the perceived quality (Beerends and Stemerdink, 1994).

Cognitive processes are generally modeled by machine learning techniques in both intrusive and non-intrusive models. For instance, Pourmand et al. (2009) used a Bayesian modeling to estimate the quality of Noise Reduction algorithms. An intrusive speech quality model developed by Chen and Parsa (2007) includes a high-level psychoacoustic and a cognitive model. This non-intrusive method combines the model of loudness calculation developed by Moore et al. (1997) with a Bayesian modeling and a Markov Chain Monte Carlo (MCMC).

Since defects have been observed in the PESQ model recommended by the ITU-T organization (ITU–T Rec. P.862, 2001), in the next chapter, after a detailed description of these shortcomings in the PESQ estimations, ways to enhance PESQ reliability will be proposed.

Integral and Diagnostic Intrusive Prediction of Speech
Quality

Côté, N.

2011, XVIII, 250 p., Hardcover

ISBN: 978-3-642-18462-8