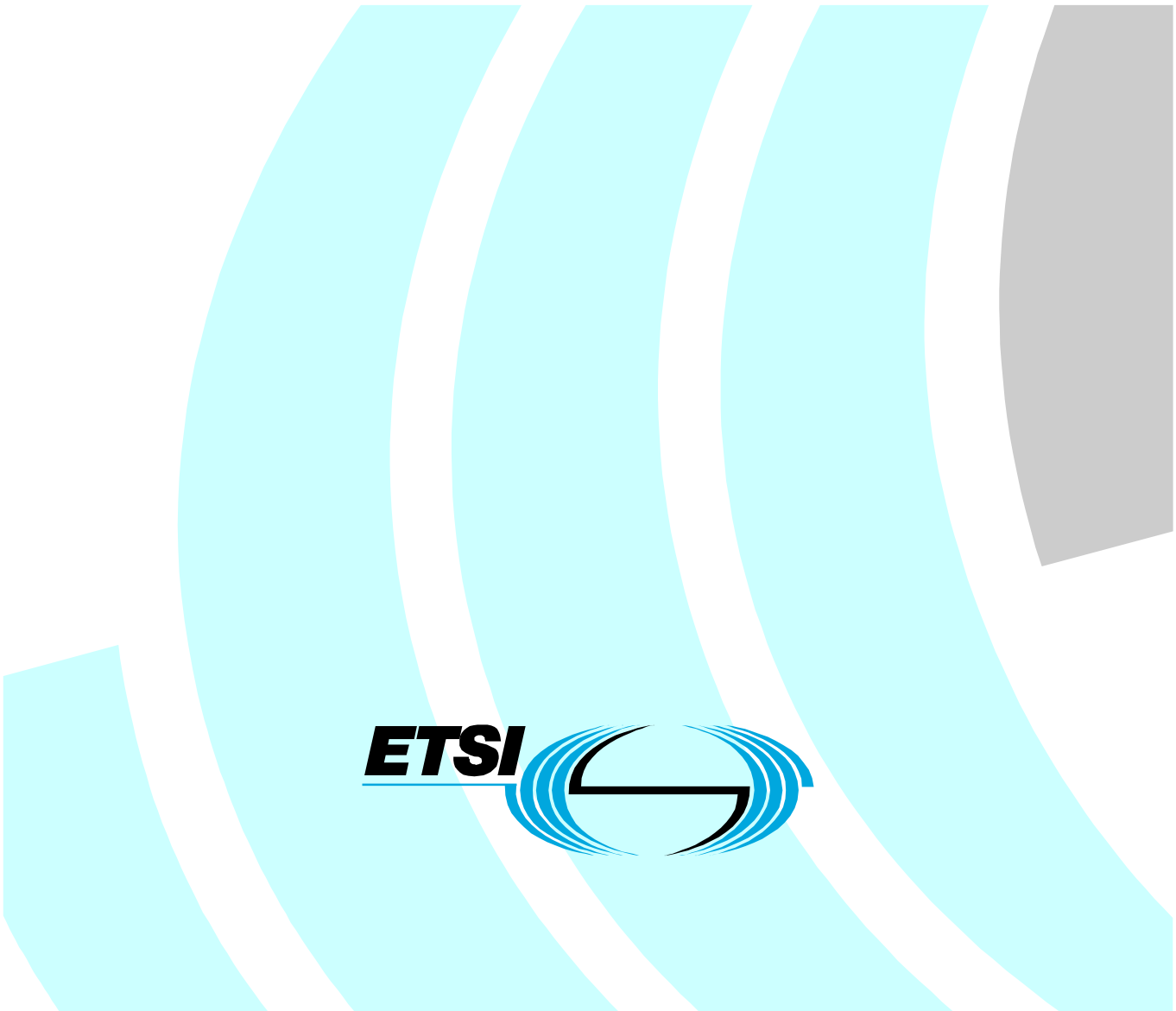
ETSI EG 201 377-1 V1.2.1 (2002-12)

ETSI Guide

**Speech processing, Transmission and Quality aspects (STQ);
Specification and measurement of
speech transmission quality;
Part 1: Introduction to objective comparison measurement
methods for one-way speech quality across networks**



Reference
REG/STQ-00033

Keywords
interworking, quality, speech, testing,
transmission, voice

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Important notice

Individual copies of the present document can be downloaded from:
http://www.etsi.org

The present document may be made available in more than one electronic version or in print. In any case of existing or perceived difference in contents between such versions, the reference version is the Portable Document Format (PDF). In case of dispute, the reference shall be the printing on ETSI printers of the PDF version kept on a specific network drive within ETSI Secretariat.

Users of the present document should be aware that the document may be subject to revision or change of status.
Information on the current status of this and other ETSI documents is available at
http://portal.etsi.org/tb/status/status.asp

If you find errors in the present document, send your comment to:
editor@etsi.org

Copyright Notification

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://webapp.etsi.org/IPR/home.asp).

All published ETSI deliverables shall include information which directs the reader to the above source of information.

Foreword

This ETSI Guide (EG) has been produced by ETSI Technical Committee Speech processing, Transmission and Quality aspects (STQ).

The present document is part 1 of a multi-part deliverable covering the specification and measurement of speech transmission quality, as identified below:

Part 1: "Introduction to objective comparison measurement methods for one-way speech quality across networks";

Part 2: "Mouth-to-ear speech transmission quality including terminals";

Part 3: "Objective measurement methods applicable to networks and links with classes of services".

1 Scope

The present document is part 1 of a series of documents on the specification and measurement of mouth-to-ear (also end-to-end) speech transmission quality. Its main objective is to describe objective comparison-based methods and systems for measuring mouth-to-ear speech quality in networks. Apart from this, it gives an overview on other important aspects of mouth-to-ear speech quality. As the need arises, these other aspects will be covered in more detail in subsequent parts of the present document.

The present document gives an overview of the methods available for measuring one-way speech transmission quality. Its purpose is to give information and guidance primarily for operators, users, consumer organizations and regulators who wish to measure or compare the speech transmission quality provided by different networks. The need for the present document has been increased by:

- the liberalization of voice services, which has introduced alternative competing providers of voice services;
- the introduction of new mobile and IP based technologies,

which has increased the range of services and cost/quality options for users.

The present document applies to both fixed and mobile networks with or without terminal equipment connected to the network. It applies only for narrowband (i.e. between 300 and 3 400 Hz) communications. In principle, comparison methods can be used for IP-based (internet protocol-based) networks, but further work is needed on the calibration of the methods for such networks. The present document describes:

- methods for measurements of individual impairments or combinations of impairments to be made at acoustic or electrical interfaces;
- methods for combining measures of different impairments into a single objective measure;
- methods for predicting the subjective effect of impairments that would be perceived by users.

The methods in the present document assume that subjects with normal hearing have been involved in the test. Therefore, the instrumental methods estimate the perceived speech quality of persons with normal hearing. For each method, the guide contains a general description to highlight the main points, and provides references for more detailed information. The present document does not contain detailed specifications of the individual methods.

The present document concentrates on *one-way* speech quality in networks. It gives no guidance on how to evaluate systems that include equipment such as echo cancellers or in which interactive impairments such as talker echo are significant. The perceived quality in such cases depends not only on the one-way performance, but very much on the behaviour of the equipment under duplex conditions; specifically, the influence of double-talk and delay shall be considered.

Although all assessments of overall speech quality are ultimately subjective because they depend on the user's opinion, a distinction is made between:

- subjective methods, which involve real time user assessment; and
- objective methods, which use stored information on the user's assessment and therefore involve some degree of calibration.

Objective methods for the evaluation of speech quality fall into three categories:

a) *Comparison Methods:* Methods based on the comparison of transmitted speech signal and a known reference.

b) *Absolute Estimation Methods:* Methods based on the absolute estimation of the speech quality (i.e. there is no known reference signal); e.g. INMD (ITU-T Recommendation P.561 [9]).

c) *Transmission Rating Models:* Methods that derive a value for the expected speech quality from knowledge about the network; e.g. ETSI Model (ETR 250 [1], ITU-T Recommendation G.107 [7]).

The classification of assessment methods is depicted in figure 1.

Practical implementations of test equipment may include combinations of these methods. The focus of the present document is on comparison methods (intrusive methods), which currently yield the most accurate results. The other categories are only covered in short overviews, although they may be preferable for certain applications.

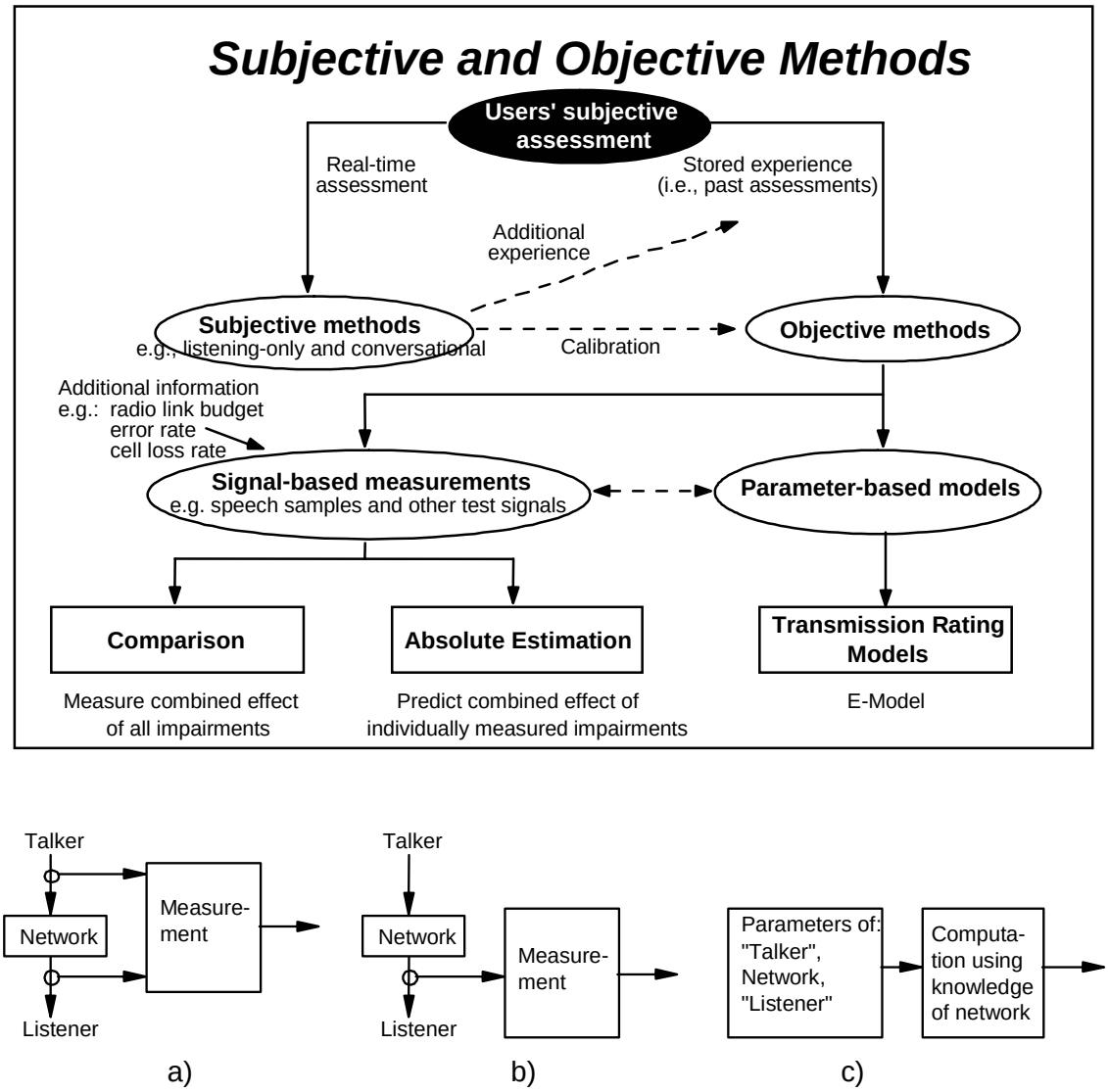


Figure 1: Classification of assessment methods showing: a) Comparison methods, b) Absolute estimation methods, c) Transmission rating models

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication and/or edition number or version number) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies.

Referenced documents which are not found to be publicly available in the expected location might be found at http://docbox.etsi.org/Reference/.

[1] ETSI ETR 250: "Transmission and Multiplexing (TM); Speech communication quality from mouth to ear for 3,1 kHz handset telephony across networks".

[2] ETSI EG 201 050: "Speech Processing, Transmission and Quality Aspects (STQ); Overall Transmission Plan Aspects for Telephony in a Private Network".

[3] ETSI TR 102 082: "Speech Processing, Transmission and Quality Aspects (STQ); Guidance on writing specifications and tests for non-linear and time variant telephony terminals ".

[4] EURESCOM Project P603 vol.1: "Quality of Service: Measurement Method Selection; Deliverable 2: Measurement Method; Volume 1 of 2: Main Report".

[5] EURESCOM Project P603 vol.2: "Quality of Service: Measurement Method Selection; Deliverable 2: Measurement Method; Volume 2 of 2: Annexes".

[6] ISO 532 (1975): "Acoustics - Method for calculating loudness level".

[7] ITU-T Recommendation G.107 (2002): "The E-model, a computational model for use in transmission planning".

[8] ITU-T Recommendation P.501: "Test signals for use in telephonometry".

[9] ITU-T Recommendation P.561 (2002): "In-service, non-intrusive measurement device - voice service measurements".

[10] ITU-T Recommendation P.800 (1996): "Methods for subjective determination of transmission quality".

[11] ITU-T Recommendation P.830 (1996): "Subjective performance assessment of telephone-band and wideband digital codecs".

[12] ITU-T COM12-20: "Improvement of the P.861 perceptual speech quality measure".

[13] ITU-T COM12-24: "Proposed Annex A to Recommendation P.861".

[14] ITU-T COM12-34: "TOSQA - Telecommunication objective speech quality assessment".

[15] ITU-T COM12-62 : "Results of Processing ITU speech database supplement 23 with the end-to-end quality assessment algorithm "PACE"".

[16] Beerends J.G., Stemerdink J.A. (1992): "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation", Journal of the Audio Engineering Society, vol. 40, no. 12, pp. 963-978.

[17] Void.

[18] Broom, S.; Coackley, P.; Sheppard, P. (1998): "Getting the message loud and clear: quantifying call clarity", BT Engineering Journal, Vol. 17, p. 66-72.

[19] De A., Kabal P. (1994): "Auditory Distortion Measure for Speech Coder Evaluation - Discrimination Information Approach", Speech Communication, 14(3):205-229.

[20] Deller J.R., Proakis J.G., Hansen J.H.L. (1993): "Discrete Time Processing of Speech Signals", McMillan Publishing Company, Eaglewood Cliffs NJ.

[21] Gabrielsson A. (1979): "Statistical treatment of data from listening tests on sound-reproducing systems", Report TA No. 92, KTH Karolinska Institutet, Department of Technical Audiology, S-10044 Stockholm, Sweden.

[22] Hogg R.V., Craig A.T. (1995): "Introduction to Mathematical Statistics", Prentice Hall Press, Eaglewood Cliffs.

[23] Hollier, M.P.; Hawksford, M.O.; Guard, D.R. (1994): "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain", IEEE Proceedings-Vision, Image and Signal Processing 141 (3), pp. 203-208.

[24] Irii H.: "Comparison of four objective speech quality assessment methods based on international subjective evaluations of universal codecs ", Proceedings of IEEE ICC '91, pp. 1726-1730.

[25] Juriæ P.: "An Objective Speech Quality Measurement in the QVoice", Proceedings of IEEE 5th International Workshop on Systems, Signals and Image Processing IWSSIP'98, pp. 156-163".

[26] Void.

[27] Zwicker E., Fastl H. (1990): "Psychoacoustics, facts and models", Springer-Verlag, Berlin, Heidelberg.

[28] ITU-T Recommendation P.862 (2001): "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs".

[29] ITU-T Recommendation G.168: "Digital network echo cancellers".

[30] ITU-T Recommendation P.831: "Subjective performance evaluation of network echo cancellers".

[31] ITU-T COM12-6: "Subjective evaluation of hands-free telephones using conversational test, specific double talk test and listening only test".

[32] Gierlich, H.W. (1996): "The Auditory Perceived Quality of Hands-Free Telephones: Auditory Judgements, Instrumental Measurements and Their Relationship", Speech Communication 20 (1996) 241-254.

[33] ITU-T Recommendation P.58 (1996): "Head and torso simulator for telephonometry".

[34] ITU-T Recommendation P.64 (1999): "Determination of sensitivity/frequency characteristics of local telephone systems".

[35] ITU-T Recommendation P.57 (2002): "Artificial ears".

[36] ITU-T Recommendation P.340 (2000): "Transmission characteristics of hands-free telephones".

[37] Gierlich, H.W.; Kettler, F., Diedrich, E.: "Speech Quality Evaluation of Hands-Free Telephones During Double talk: New Evaluation Methodologies"; EUSIPCO '98, Rhodos, Greece, Conference Proceedings, vol. 2, pp. 953 - 956, 1998.

[38] CCITT Supplement No. 5 to Recommendation P.74: "The SIBYL Method of Subjective Testing", Red Book, Volume V.

[39] ITU-T Recommendation P.82 (1988): "Method for evaluation of service from the standpoint of speech transmission quality".

[40] Zoubir A.M., Boashash B.: "The bootstrap and its application in signal processing", IEEE Signal Processing Magazine, pp. 56-76, Jan. 1998.

[41] ETSI TBR 008: "Integrated Services Digital Network (ISDN); Telephony 3,1 kHz teleservice; Attachment requirements for handset terminals".

[42] ETSI TBR 009: "European digital cellular telecommunications system; Attachment requirements for Global System for Mobile communications (GSM) mobile stations; Telephony".

[43] ETSI TBR 010: "Digital Enhanced Cordless Telecommunications (DECT); General Terminal Attachment Requirements; Telephony Applications".

[44] ETSI TBR 038: "Public Switched Telephone Network (PSTN); Attachment requirements for a terminal equipment incorporating an analogue handset function capable of supporting the justified case service when connected to the analogue interface of the PSTN in Europe".

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the following terms and definitions apply:

bark: frequency unit in the perceptual domain; e.g. frequencies at 3, 4, and 5 Bark are perceived as equally-spaced

cepstrum: cepstrum of a signal is defined as the inverse Fourier transform of the logarithm of the power spectrum of that signal

NOTE 1: See figure 5.

NOTE 2: Linear distortions of a signal (e.g. delay, echo) are additive in the cepstral domain.

cognitive: pertaining to higher layers of human reception; e.g. interpretation of speech

perceptual: pertaining to lower layers of human reception; e.g. processing of sound signals

psycho-acoustic: pertaining to acoustic processing particular to the human sound perception system; e.g. masking of adjacent frequency components

3.2 Abbreviations

For the purposes of the present document, the following abbreviations apply:

ANOVA	ANalysis Of VAriances
ACR	Absolute Category Rating
ATM	Asynchronous Transfer Mode
CCR	Comparison Category Rating
CD	Cepstral Distance
CDI	Cochlear Discrimination Information
CMOS	Comparison Mean Opinion Scores
DC	Direct Current
DCME	Digital Circuit Multiplication Equipment
DCR	Degradation Category Rating
DFT	Discrete Fourier Transform
DMOS	Degradation Mean Opinion Scores
FFT	Fast Fourier Transform
FMNB	Frequency Measuring Normalizing Block
GSM	Global System for Mobile communication
INMD	In-service, Non-intrusive Measurement Device
IP	Internet Protocol
ISDN	Integrated Services Digital Network

LAR	Log-Area Ratios
LPC	Linear Prediction Coefficient
MNB	Measuring Normalizing Blocks
MOS	Mean Opinion Score
PAMS	Perceptual Analysis/Measurement System
PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
POTS	Plain Old Telephony Service
PSQM	Perceptual Speech Quality Measure
PSTN	Public Switched Telephone Network
QoS	Quality of Service
QSDG	Quality of Service Development Group
SNR	Signal-to-Noise Ratio
TMNB	Time Measuring Normalizing Block
TOSQA	Telecommunication Objective Speech Quality Assessment

4 Overview

Today, telecommunication is strongly influenced by three major facts:

- the liberalization of telecommunication, i.e. the separation between regulatory bodies and operators;
- the splitting of operations into network providers and service providers; and
- the increase of international traffic due to the internationalization of trade and business.

In addition to these facts, there is also a strong influence due to technical evolution. The most important trends are the move from fixed networks to mobile networks, but also from conventional switched PSTN and ISDN networks to packet-based networks such as the Internet. These technical trends will make it necessary to extend the applicability of the methods described below in order to cover speech quality impairments from "new" types of degradations, such as packet losses and variable delay.

The liberalization as well as the splitting of operations lead to new legal/commercial/technical interfaces, which need a definition both in the contractual and technical sense:

- regulators need a measurement basis in order to specify the requirements which "their" network operators have to fulfil;
- operators of private networks (e.g. corporate networks, closed user groups) need a measurement basis as well for double-checking transmission planning issues for the interconnection of private networks with the public ISDN/PSTN; and
- service providers want to compare different network providers concerning their price/performance ratio.

In all cases the traditional methods for speech quality assessment based on subjective rating of speech samples are far too expensive, too slow and lack the precise repeatability.

The internationalization of traffic as well as the multitude of network providers lead to the fact that in many cases a phone call is routed through several networks, where these networks are based on different technologies (fixed analogue or digital, ATM, Internet, mobile networks, satellite links, etc.). The concatenation of multiple different networks is no longer restricted, and the resulting effects on speech quality are not well covered up to now.

4.1 Objective

The aim of the present document is to give:

- general information on mouth-to-ear speech quality, and the factors to be included in its evaluation (see clause 5);
- information on subjective reference assessment methods, which are essential to calibrate objective methods, showing what results can be obtained (see clause 6, annex D);
- information on the objective comparison measurement methods available and how they work, especially the most recent methods (see clause 7);
- overview of other assessment methods (see clauses 8 and 9).

In a second part of the present document (to be developed later), the criteria for the evaluation of such objective measurement systems will be specified, namely:

- requirements concerning the technical characteristics of speech quality measurement;
- methods to test the conformity of these methods to the subjective reference assessments; and finally
- criteria to compare and evaluate the current methods.

4.2 Related work in standardization

On all of the above mentioned topics a lot of work has already been done in the past by a number of standards bodies:

- ETSI TC STQ http://portal.etsi.org/STQ
  This Technical Committee is responsible for the "co-ordination, production (where appropriate) and maintenance of end-to-end speech quality related deliverables" (TC/STQ Terms of Reference).
- 3GPP SA
  The work done in 3GPP SA 4 concentrates on codec quality in mobile networks (in particular for Half Rate, Enhanced Full Rate and Adaptive Multi-Rate codecs) and therefore is not primarily oriented towards mouth-to-ear speech quality aspects. However, it is a very important source of information especially for the subjective rating of speech samples and for the characteristics of speech samples to be used for assessment and measurement. Note that this work done in SA 4 was previously performed by ETSI SMG 11.
- ETSI Project TIPHON http://portal.etsi.org/STQ
  According to the Terms of Reference (ToR) the ETSI Project TIPHON addresses the following topics:
  - initial focus should be on voice communications although in the future other forms of data communications could be taken into account;
  - the marketing activities should be carried out in order to affirm TIPHON's awareness;
  - the project does not have a mandate with respect to the European Commission in its ToR;
  - an activity to verify and demonstrate TIPHON specifications should be created.
- ITU-T Study Group 12
  The current work in ITU-T Study Group 12 (study period 2001 to 2004) is focused both on terminal and acoustic tests and on mouth-to-ear network aspects. Several questions are addressing mouth-to-ear speech quality issues, in particular:
  - Q4/12: Telephonometric methodologies for hands-free terminals and speech enhancement devices (including AEC and Noise Reduction);
  - Q6/12: Analysis methods using complex measurement signals;
  - Q7/12: Methods, tools and test plans for the subjective assessment of speech and audio quality;
  - Q8/12: Extension of the E-Model;

- Q9/12: Objective measurement of speech quality under conditions of non-linear and time-variant processing;
- Q11/12: Speech transmission planning for multiple interconnected networks (e.g. public, private, Internet);
- Q16/12: In-service non-intrusive assessment of voice transmission performance.

- ITU-T Study Group 2/QSDG
  The "Quality of Service Development Group" is a subgroup of ITU-T Study Group 2. Its members are network operators and manufacturers from all over the world.
  According to their Terms Of Reference, the tasks of QSDG are the following:
  - encourage participation in QoS activities;
  - identify and develop performance monitoring and evaluation;
  - improve QoS, include practices in TSS documentation;
  - disseminate information about QoS techniques and procedures;
  - encourage development of co-ordinated approach of QoS;
  - other activities to improve.

- EURESCOM
  EURESCOM is a private company owned by European network operators and doing research in the field of network operation. Among others, there is a project P603 in EURESCOM which has been finished recently, and a subsequent project is in the state of definition (see [4] and [5]).

- ETSI AT and former technical bodies such as ATA, DTA, MTA, TE4 and TE5.

5 Definition of mouth-to-ear speech quality

5.1 General definition

Mouth-to-ear speech quality (also "end-to-end speech quality") is defined as the degree of speech quality that a listener perceives at his terminal with a talker at the far end. (In some cases this definition may be too restrictive, e.g. when considering talker echo.)

This definition raises a number of questions and clarifications to be made:

- An absolute physical definition of "speech quality" does not exist; the only "baseline" we have is the subjective perception of human listeners.
- Speech quality ultimately is a psycho-acoustic phenomenon involving a complex interaction of many parameters within the process of human perception, although many of the individual parameters can be measured purely electrically.
- Mouth-to-ear in this context implies that there is a transmission of the speech signal by some kind of network; it is to be defined what that network consists of.
- In today's liberalized environment a network provider can no longer prescribe the terminal equipment being used by his customers; his reach and therefore his responsibility is limited to his network and ends at the outlet on the customer's premises.
- Speech quality is but one component of the overall quality perceived by a telecommunications user.

In the following clauses we list the required parameters and the conditions under which these have to be assessed or measured, respectively.

5.2 Human perception characteristics of speech quality

The human hearing and recognition system being highly non-linear and by far not completely understood today, we cannot analytically predict the human perception of the quality of a speech signal being transmitted through a network. However, it is clear that there are objective (physically measurable) factors as well as inter- and intra-individual aspects. Therefore, a quantitative expression of speech quality will always be a statistical mean value. The averaging is not limited to the objectively measurable factors but also includes a "mean physiological and psychological sensitivity" of human beings.

5.2.1 Physical characteristics and psychological impacts

The perceived overall speech quality is determined by a number of underlying psychological parameters. The most important ones are intelligibility, naturalness and loudness. In turn, these parameters are determined by the physical characteristics of the network under consideration, as illustrated in table 1. (The parameters are only examples.)

In the context of the present document, the main psychological characteristics are:

- *Intelligibility:* Quality of perception of the meaning or information content of what the speaker has said [20], see Bibliography, (5).
- *Naturalness:* Degree of fidelity to the speaker's voice.
- *Loudness:* Absolute loudness level at the receiver's side.

Table 1: Examples of dependence of psychological characteristics on physical characteristics

Physical characteristics	Psychological characteristics			
	Intelligibility	Naturalness	Loudness	Overall speech quality
Signal level	X	X	X	X
Noise	X			X
Frequency response	X	X	X	X
Distortion	X	X		X
Delay	X			X
Echo	X			X
Packet losses	X			X

In addition, the following parameters are important:

- (Speech) Sound quality (similar to naturalness): Perceived sound quality of telephone speech.
- Double talk capability: Ability to really interact (double talk) in a conversation.
- Quality of background noise transmission in single talk and double talk conditions.
- Speech level variations during single talk and double talk.
- Disturbances caused by switching during single talk and double talk (completeness of speech transmission).
- Disturbances caused by echoes during single talk and double talk.

5.2.2 Inter-subject differences

Auditive cognition: The sensitivity characteristics vary with each individual.

Hardness of hearing: Some people have more difficulties than others.

5.2.3 Intra-subject differences

Even the very same person does not always perceive speech the same way: According to her/his actual situation of interest, mood and expectation the momentary attention varies greatly.

The inter-subject as well as the intra-subject differences are the reason why any objective measurement result cannot be compared directly to the subjective perception of any given individual, but shall be compared with an average value of subjective opinions (*Mean Opinion Score*, MOS). To be relevant, these MOS values have to be based on sufficiently large sets of speech samples and test persons.

5.2.4 Language-dependent differences

Frequency range ↔ spectra of consonants: There are languages with very intensive use of consonants (e.g. Slavic languages) which are more easily intelligible with a larger bandwidth. Given the increasing fraction of international traffic, any network (and especially codec) around the world should accommodate equally for all languages.

5.3 Network-related issues

5.3.1 Reference configuration for mouth-to-ear measurement

Mouth-to-ear measurement in an objective/repeatable manner means measuring in a pseudo-natural environment. While the network part is the responsibility of the network provider(s), these cannot be blamed for difficult environmental conditions at the sender's or receiver's side. Hence standardized terminal equipment and standardized subscriber's environment should be assumed (and introduced into the quality assessment model with the corresponding parameters) while the network(s) in-between are actually being measured (see also clause 5.4).

For the network-related measurements, the following network configuration shall be assumed:

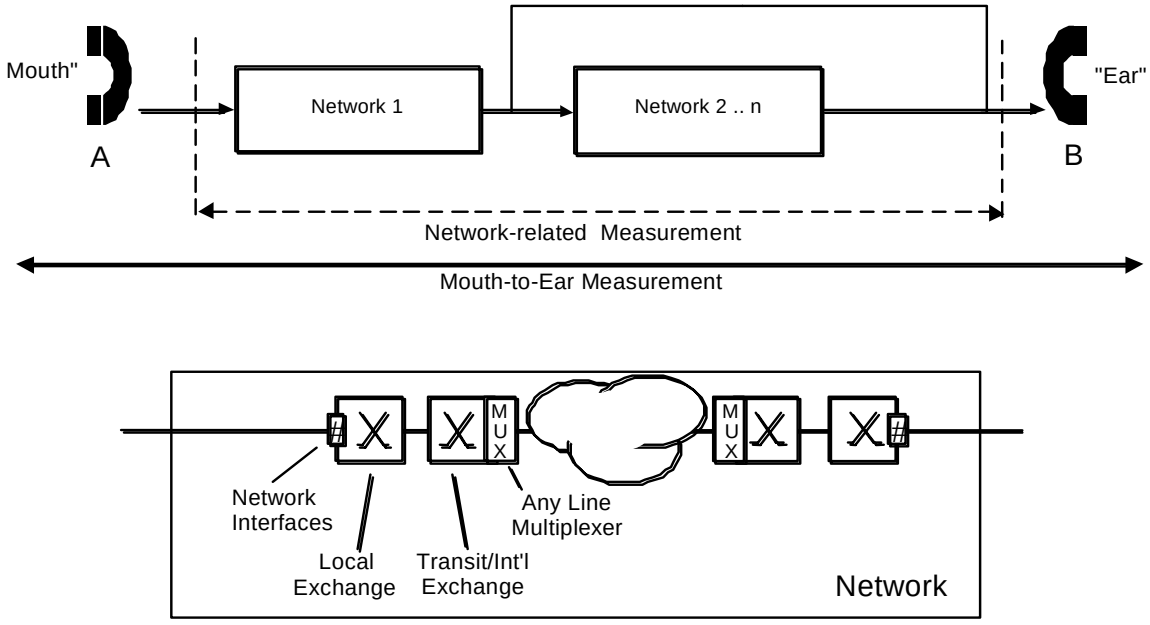


Figure 2: Reference network for mouth-to-ear speech quality (top); detail of component network (bottom)

Figure 2 shows the reference network to be used. It consists of one or more network segments belonging to possibly different network providers. These networks may be ISDN/PCM (including satellite links) or ATM or Internet. The terminal access (subscribers A and B) consists of POTS, ISDN or mobile equipment whose acoustical characteristics regarding speech quality are determined in acoustic laboratories.

Using such a reference model gives rise to several questions which are treated in more detail in the following:

1) Network related questions:

   - Is it possible to measure the quality parameters of each single network in a chain and to determine analytically the quality of a concatenation of networks?

   - Is the impact on quality of concatenated networks equal to the sum of the deteriorations of each sub-network?

   - For circuit-switched and packet-switched networks as well as for fixed and mobile networks, the parameters to be measured and the methods to evaluate their influence on speech quality will most probably be different.

2) Subscriber's environment related questions:

   - Is it correct to assume a standardized terminal equipment and a standardized subscriber's environment?

   - Should a terminal equipment be standardized per type of network (analogue, fixed ISDN, mobile (GSM), other)?

   - If yes, which network should such a standardized terminal refer to? To the network the subscriber is directly connected to, or to the network causing the (main part of) speech quality problems?

3) Is such a separation admissible at all?

   - Is all terminal equipment "immune" against all signal distortion types of all networks? If not, what are the restrictions to be made?

5.3.2 Standardization of quality parameters

There are two classes of parameters which converge to a widely agreed set of characteristics:

- the transmission parameters characterizing a network and the terminals connected to it; and
- the value (rating) expressing the speech transmission quality.

The standardization of the transmission parameters characterizing a network and the terminals connected to it is one of the merits of the work done in preparing ETR 250 (E-Model) [1], EG 201 050 [2] and the G100 Series of ITU-T Recommendations.

The speech transmission quality is most frequently expressed as a MOS value, either derived from the subjective assessment of speech quality (see clause 6) or expressed in ratings derived from the E-Model with the first class of parameters as input values. These scores are also used for the final result of objective quality measurement systems (see clause 7).

Apart from transmission parameters and speech quality values, there is a third category of characteristics, namely the parameters describing ping-pong, robot voice and similar phenomena occurring mainly in mobile and packet switched networks. This third set of parameters is not yet standardized.

5.3.3 Modelling of networks - anomalies

For the determination of the speech quality of concatenated networks it would be desirable to measure each single network separately and to derive the compound speech quality value using some analytical method. Presently, no such method is known. (The E-model used for planning purposes circumvents the problem by assuming that transmission impairments are additive on a psychological scale.)

Are the following anomalies modellizable at all, or do we have to measure them in vivo since there is no numeric method to predict them analytically? Should "modellizable" include the option to set up a physical model of a network or of multiple networks which can be interconnected for measurements?

- Concatenation of different codec algorithms?
- Concatenation of different networks?

- More generally, the joint effect of different types of impairments (further material, see ITU-T Study Group 12, Question 8).
- Roaming and handover of mobile stations, especially in the case of different network/service operators and in the case of partially analogue networks (echoes!).

5.4 Terminal equipment related issues

Today there is a set of parameters being measured in order to state the compliance of terminal equipment to national or international requirements and standards. A description of such a set of parameters as well as the methods and prerequisites to be used can be found in [3] and in:

- TBR 008 [41]: Digital telephones (ISDN);
- TBR 009 [42]: GSM telephones;
- TBR 010 [43]: DECT telephones;
- TBR 038 [44]: Analogue telephones (PSTN).

Guidance on setting up test procedures for new types of telephone terminal implementations (non-linear and/or time-variant) can be found in [3], [4]and [8].

In addition to the methods described above, a lot of effort has been put into the evaluation of more sophisticated methods, allowing the evaluation of non linear and/or time-variant equipment. The investigations have been made for terminal equipment, especially hands-free telephones (in combination with handset telephones) and network echo cancellers. It was found that almost all parameters influencing speech and conversational quality are the same for the network devices and the HFT-terminals. Thus there is high confidence that the relevant objective criteria can be used for the evaluation of terminal equipment, the network and the mouth-to-ear transmission quality including terminal equipment. Further information can be found in annex C.

Guidance on setting up test procedures for new types of telephone terminal implementations (non-linear and/or time-variant) can be found in [5]. However, these standards only specify manual measurement methods, and additional guidance is needed on the use of automatic methods.

5.5 Technical basis for measurement

(The influences of codecs and terminal acoustics are taken into account as all the other influences, too, but not investigated in particular, since these issues are treated in ITU-T Study Group *12 and 3GPP SA already.)*

5.5.1 Quantification and measurement of speech quality

As there is no commonly agreed objective definition of "speech quality", assessment of this attribute is necessarily subjective. However, in order to quantify speech quality in an objective and repeatable manner, individual aspects need to be eliminated from the assessment.

In the past, this has been achieved by *subjective measurement methods*, e.g. listening experiments. In a carefully designed experiment, averaging a sufficiently large sample of individual opinions indeed yields an accurate rating of the "true" speech quality. Listeners are presented speech samples, which they rate according to an integer-valued opinion scale. By averaging the individual opinion scores, a *Mean Opinion Score* (MOS) results, which depends less on individual preferences. Clause 6 presents a survey of subjective measurement methods.

More recently, *objective comparison measurement methods* have been introduced. These methods compute a quality value from a speech sample. Their ultimate goal is to estimate as closely as possible the MOS value that would result from a subjective measurement. Objective measurement methods are attractive because they require less effort (no listening panel) and can be automated. Nevertheless, calibration with MOS values from subjective measurements is inevitable. Clause 7 investigates objective comparison measurement methods in more detail.

In the following, the terms "subjective MOS" and "objective MOS" refer to values obtained from subjective and objective measurements, respectively.

5.5.2 Required characteristics of speech samples

At present, there is no commonly agreed set of speech samples for neither subjective measurement experiments nor objective methods. Nevertheless, the requirements below have been applied successfully in practice, and most subjective measurement experiments and objective methods are based on similar sample characteristics. And although some objective measurements are performed with artificial speech, it is still necessary to have a *reference configuration*. Moreover, the parameters of artificial speech samples can be derived from the following requirements.

Therefore, the speech samples to be used for measurements should meet the following criteria (see [11]):

1) Physical characteristics:

Frequency range:	300 Hz to 3 400 Hz (narrowband systems);
	100 Hz to 7 000 Hz (wideband systems);
Duration:	5 s to 10 s (not including header and trailer sequences);
Density:	70 % speech, 30 % pauses.

2) Phonetic characteristics:

Language(s):	English, German, French, Swedish, Italian *(other languages, e.g. slavic?);*
Distribution of phonemes:	The distribution within the sample should represent the standard distribution of phonemes in the chosen language;
Female/male speakers:	50/50 %.

3) Recording characteristics:

Sample Rate:	≥ 8 kHz (narrowband systems, at network insertion point);
	≥ 16 kHz (wideband systems, at network insertion point);
Digital resolution:	> 12 + 1 bit (original stored sample);
Recording resolution:	> 95 dB (i.e. 16 bit);
DC offset:	0.

6 Subjective measurement of speech quality

Although subjective measurements of speech quality require a substantial effort, they are indispensable as a reference for objective measurement methods. This clause gives a summary of the status of standardization in the field of subjective determination of speech transmission quality. It is based mainly on the documents ITU-T Recommendations P.800 [10] and P.830 [11]. In the context of the present document, we limit ourselves mainly to listening-opinion tests; conversation-opinion tests are very time-consuming and hard to design in such a way that the results are repeatable.

The results of subjective listening-opinion tests are influenced by a wide variety of conditions. Therefore, utmost care shall be taken to obtain reliable and reproducible results. Some of the factors to be controlled are:

- *Speech material:* Perception depends on the gender of talkers, their pronunciation, the language, length and content of samples, the recording room and equipment characteristics.
- *Experiment set-up:* Results can depend on nationality and gender of listeners, recent previous experience with listening tests, instruction of listeners about the experiment, duration of test sessions, and order of presentation of speech samples.
- *Listening conditions:* Loudness of presented speech samples and choice of equipment (headphones/telephone handsets) can influence the rating.

ITU-T Recommendations P.800 [10] and P.830 [11] contain guidelines on how to cope with these factors to obtain reliable and reproducible test results.

6.1 Subjective measurement methods

Annex D presents a basic overview on some subjective measurement methods. It is by no means intended to be complete the interested reader is referred to ITU-T Recommendation P.800 [10], which contains in-depth information about the measurement of subjective speech quality. In addition to the methods sketched in annex D, there are other test methods; for instance, the method of paired comparisons is very useful when the quality differences between the test cases are small.

6.2 Application of statistical methods

Experiments to estimate speech quality often yield a substantial amount of data. Statistical methods are a useful tool in both planning and evaluation of speech quality listening experiments:

- *Planning of experiment:* Choose size of listener panel, number of speech samples, etc.
- *Interpretation of results:* Assess reliability and accuracy of results, detect dependencies between parameters, etc.

As there are many good textbooks on statistics (see e.g. [21] and [22]), the present document contains just a very basic overview of some statistical concepts. The interested reader is referred to annex E.

7 Objective measurement methods

In order to measure speech transmission quality on a regular basis, it is necessary to avoid the complicated and expensive procedure of subjective determination, and to use objective systems instead. Today there are several systems and methods in use, some of which are in a rather experimental state, while others are commercially available products. The following non-exhaustive list shows some situations related to transmission over networks where such objective measurement methods are being applied:

- *Mobile Communications:* In mobile communication systems (e.g. GSM), speech quality measurement campaigns can unveil coverage problems, base station failures (e.g. handover problems), etc.
- *Speech compression devices:* In networks with speech compression devices (e.g. codecs, DCME), speech quality can be severely impacted due to the interaction of such devices with each other and with effects like noise, echoes, etc. Monitoring mouth-to-ear speech quality can detect such problems.
- *Voice over IP ("Internet telephony"):* The characteristics of IP-based networks are different from conventional telephony networks because IP was originally designed for data traffic. For voice over IP, the most critical parameters are delay and the degree of packet loss. It is not clear today which degree of speech quality can be achieved on such networks, and monitoring speech quality could help to improve service quality. In order to do so, however, objective methods shall be able to cope with such impairments.
- *Cascade of networks and/or analogue interfaces:* In today's liberalized environment, it is increasingly likely that a call is routed through several networks. On its way, the speech signal could be compressed and expanded repeatedly, or undergo several A/D and D/A conversions. The impact of such cascading on speech quality is virtually impossible to predict, but can be assessed by mouth-to-ear speech quality measurements.
- *Private networks (e.g. corporate networks, closed user groups) interconnected with the public ISDN/PSTN*: After the pre-installation transmission planning (e.g. according to EG 201 050 [2]), it might be highly desirable to measure and monitor the "real" speech quality in order to obtain feedback on the quality and reliability of the transmission planning process; i.e. comparing the expected influence of codecs, linear distortions, AD/DA interfaces, echoes, etc. with their real-world impact.

Objective determination of speech quality is based on two distinct methods as shown in figure 1:

- Signal-based methods, comparing speech samples before and after transmission through networks (or using only the speech sample after transmission).
- Parameter-based methods involving models.

For the signal-based comparison methods there exist a number of systems. The rest of this clause describes the building blocks that are used in more recent systems. It should be noted that these systems implement and combine the building blocks in different ways, which results in performance differences. Moreover, some systems may be designed for acceptable performance in a wide range of applications, while others may aim at accurate results in just one specific application.

The current ITU-T Recommendation for signal-based comparison is called PESQ (ITU-T Recommendation P.862 [28]). It replaces the earlier standard PSQM (withdrawn ITU-T Recommendation P.861, see bibliography) and has been selected amongst several candidates, and its performance shows a high correlation with subjective scores on a large number of databases covering a large number of conditions.

Parameter-based methods are represented mainly by the E-model, which is described extensively in ETR 250 [1].

Today it is not clear how these complementary methods fit together. In addition, the model-based approach is limited to relevant phenomena that can be modelled, which is not the case for "exotic" effects such as garage algorithms, handover/roaming, fading, packet loss, transcoding because of codec concatenation etc.

7.1 Basics of speech sample based objective measurement methods

Currently, all objective measurement systems for speech quality measurement use two signals as their input, namely an original signal (reference pattern) and the corresponding output signal after its transition through the network under test.

The signal processing within objective methods based on the comparison of speech samples can be structured into three major steps as follows (see figure 3):

- pre-processing;
- psycho-acoustic modelling;
- speech quality estimation model.

These steps are generally implemented with the same building blocks (see figure 3) in all systems for speech quality evaluation, namely:

- a signal adjustment unit (adapting signal delay, loudness differences, and useful signal duration);
- a unit that models and/or measures the environment characteristics;
- a time-frequency mapping;
- a model of psycho-acoustic sound perception;
- a method to compare the measured signal (or rather the parameters describing it) with the reference;
- a function to determine a single value describing the speech quality;
- a function to transform this result according to subjective/auditory evaluation scales.

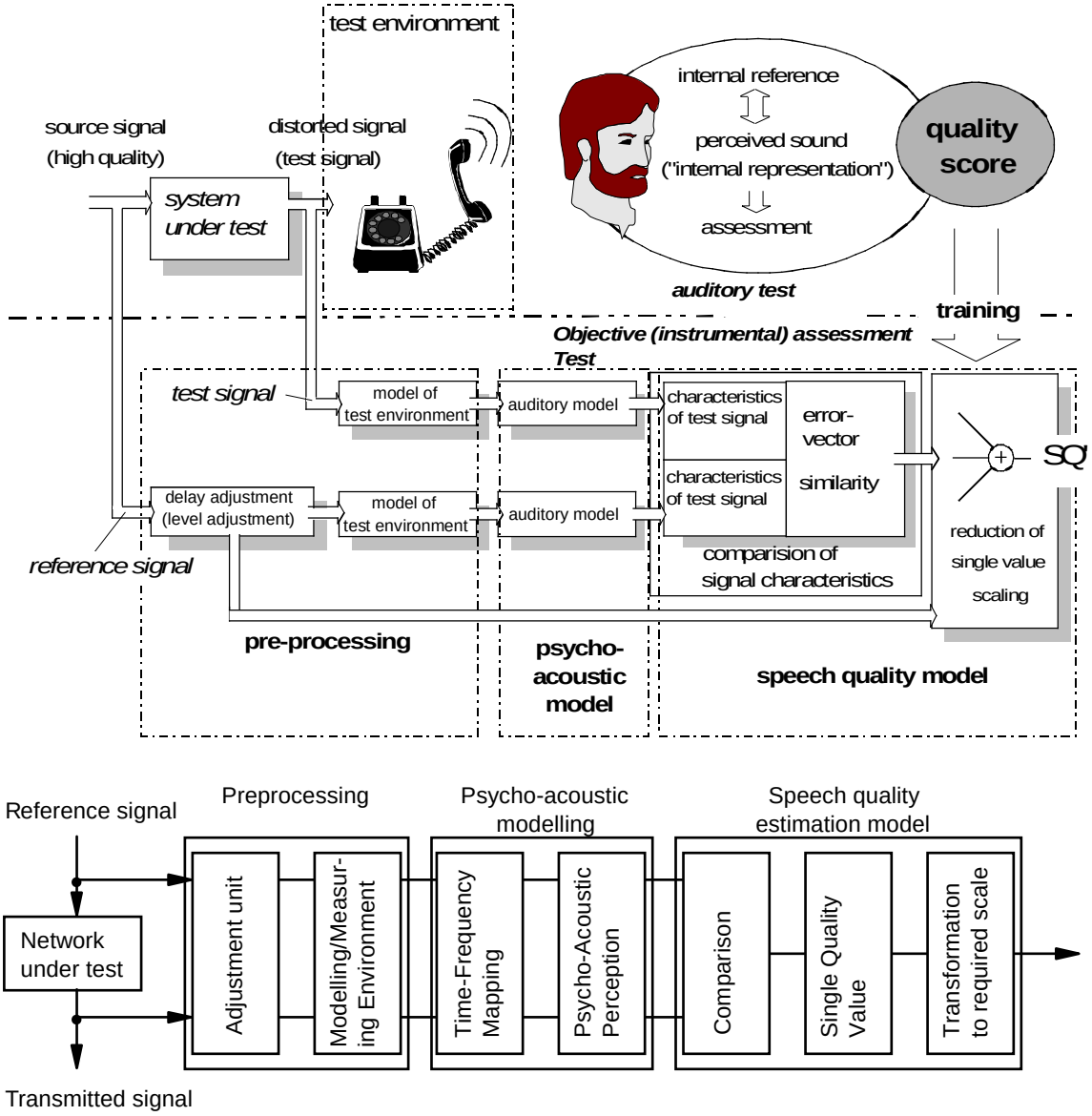


Figure 3: Block diagram of mouth-to-ear speech quality measurement systems

7.2 Pre-processing

7.2.1 Adjustment unit

All objective measurement systems in use require that reference and processed signal be adjusted properly to each other. Towards this aim, several tasks have to be performed. (Some of these tasks may be deferred to a later stage.)

- *Delay adjustment:* Since the transmitted signal is delayed in the network, it is necessary to align it properly with the reference signal. This can be done by correlating reference and transmitted signal and searching for the correlation peak. It may be helpful to insert a specific synchronization signal in the reference.

- *Loudness adjustment:* Most measurement systems assess the quality of the transmitted sample by computing some kind of difference to the reference signal. Therefore, the loudness of both signals shall be equal, which can be achieved by making the average signal power equal (after removing any DC component). It is also possible to adjust loudness in the psycho-acoustic (perceptual) domain (see clause 7.4).

- *Useful duration adjustment:* If the transmitted sample is affected by echo, its duration is longer than that of the reference. Moreover, some measurement methods may choose to ignore parts of the sample (e.g. speech pauses). It is thus necessary to remove undesired parts of the signal, such that reference and transmitted signal are properly aligned and of equal length.

- *Filtering:* It may be known that some frequencies (e.g. outside the telephone band) cannot pass on to the listener. In this case, such frequencies may be filtered out. Since this operation could destroy valuable information, it shall be applied with caution.

7.2.2 Modelling and/or measuring transmitter and receiver environment

In the typical objective measurement system, signals are not transmitted from mouth to ear, but rather between electrical interfaces. Therefore, the impact of the environment on speech quality is missing. If desired, this influence can be included in the measurement process.

- *Transmitter side:* The speech material can be recorded in a "typical" or "average" environment, which has to be defined appropriately. By recording speech samples in several different environments, the expected range of talker conditions can be covered. In this case, objective measurements result in a speech quality distribution over different talker conditions, from which the mean and other statistical quantities can be derived.

- *Receiver side:* An artificial listener environment can be created, by emulating noise, room reverberation, and handset characteristics. As an alternative, a more realistic evaluation can be performed by including items such as background noise, room reverberations and the telephony terminal itself in the "measurement chain". In other words, the measurement between electrical interfaces is extended to a real mouth-to-ear measurement by including a realistic test room and by including acoustical devices that substitute the human mouth and ear (and body). Obviously, in both cases the result will only be an approximation to a real-world environment. It is important that both the reference and the transmitted signal are subjected to the listener environment model. Analogously to the transmitter side, a multitude of listener conditions can be applied to obtain a speech quality distribution over different listener conditions.

7.3 Psycho-acoustic sound perception

Little is known about the mechanisms that govern human perception of "speech quality". Due to this lack of knowledge it seems a natural step to eliminate from the speech signal all information that the auditory system ignores. The retained information is expected to be relevant in the sense that it contains all necessary information to decide on the quality of the transmitted signal. Therefore, most speech quality measurement schemes attempt some modelling of psycho-acoustic sound perception.

7.3.1 Time-frequency mapping

In most systems, a pre-processing unit transforms the signal into a time-frequency representation; this is motivated by psycho-acoustic considerations. In contrast to later stages, the preprocessing unit does not attempt to reduce the amount of data. Typically, the signal is segmented and transformed to the frequency domain as follows:

- *Segmentation* of the signal in overlapping blocks: segmentation models the short-term nature of human perception. Typically, segments are 15 ms to 40 ms long. Overlapping successive blocks (e.g. by 50 %) is a smoothing operation to avoid abrupt changes from block to block, but can also be viewed as modelling "memory" or "inertia" of the human listener.

- *Windowing operation:* The segment is multiplied sample-by-sample with some windowing function (see figure 4). Each windowing function has an associated trade-off between frequency resolution and suppression of spurious frequencies.

- *Fourier transform/power spectrum* for each block: The Fourier transform computes the frequency spectrum of a signal. The power spectrum is then obtained by squaring the amplitude value at each frequency. Since the outer part of the human auditory system essentially performs a spectral analysis, Fourier transformation is a first step towards modelling psycho-acoustic sound perception.

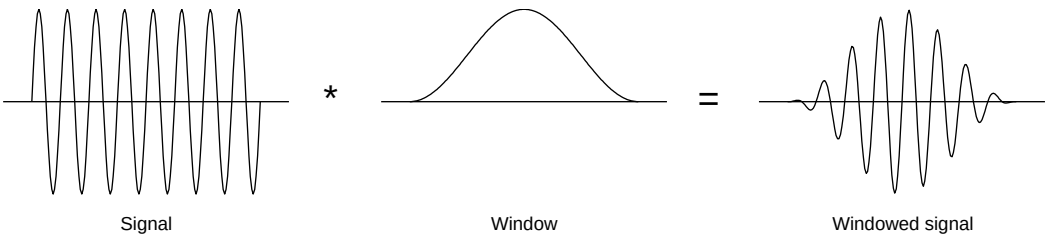


Figure 4: Windowing of speech signal

Time-frequency mapping, transformation to perceptual domain (critical bands) and frequency masking effects (see below) may be reproduced by using a perceptual filterbank as the core time-frequency transformation [23]. This has the advantage of greater temporal resolution in the signal analysis, but is more difficult to implement than alternative techniques.

7.3.2 Linear prediction coefficients

Some early measurement systems were based on Linear Prediction Coefficients (LPC). The goal of such methods is to find a linear short-term model of low order for the speech signal. In other words, these systems model the speech *source* (the vocal tract) rather than speech perception. Due to the sensitivity of LPC parameters, they were often transformed to log-area ratios (LAR), which are less sensitive. LPC/LAR coefficients are no longer used in recent systems.

7.3.3 Cepstrum

Transformation of the signal to the cepstral domain is another early attempt (see e.g. [24]) to model human sound perception. The cepstrum transformation is sketched in figure 5. In practice, the transformation is computed more directly using Linear Prediction Coefficients (LPC); as a desirable side-effect, spectral smoothing is achieved due to the limited number of LPC coefficients in the calculation.

The cepstral approach has several attractive features:

- Linear distortions of the received signal (i.e. echoes and delays) appear as additive components in the cepstral domain. Therefore, the difference between the cepstra of reference and transmitted signal is a measure for linear distortion.
- The logarithmic power level compression in the frequency domain is akin to the loudness compression in the auditory system (see below).
- The cepstrum is easily computed using LPC coefficients.

On the other hand, the cepstrum method is a rather *ad hoc* model of the human auditory system. Moreover, performance in the presence of non-linear distortions may be unsatisfactory.

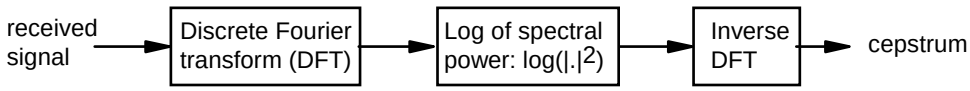


Figure 5: Transformation to cepstral domain

7.3.4 Mapping to perceptual (critical band) domain

The auditory system operates mainly in the frequency domain, with a non-linear relation between measured frequency (in Hz) and perceived frequency (in Bark). Figure 6 shows one of the approximations that are in use, namely:

$$f = 600 \times \sinh(b / 6) \qquad \text{(f: Hz, b: Bark)} \quad (1)$$

The *perceptual* or *critical band domain* represents a linear scale for the human perception of frequencies. In particular, high frequencies are compressed logarithmically; e.g. frequencies f, 2f and 4f are perceived as "equally-spaced".

Such a mapping of the speech signal to the perceptual domain is often found in objective speech quality measurement systems. The general mapping scheme is depicted in figure 7. (As mentioned above, this step can be implemented in a perceptual filterbank.)

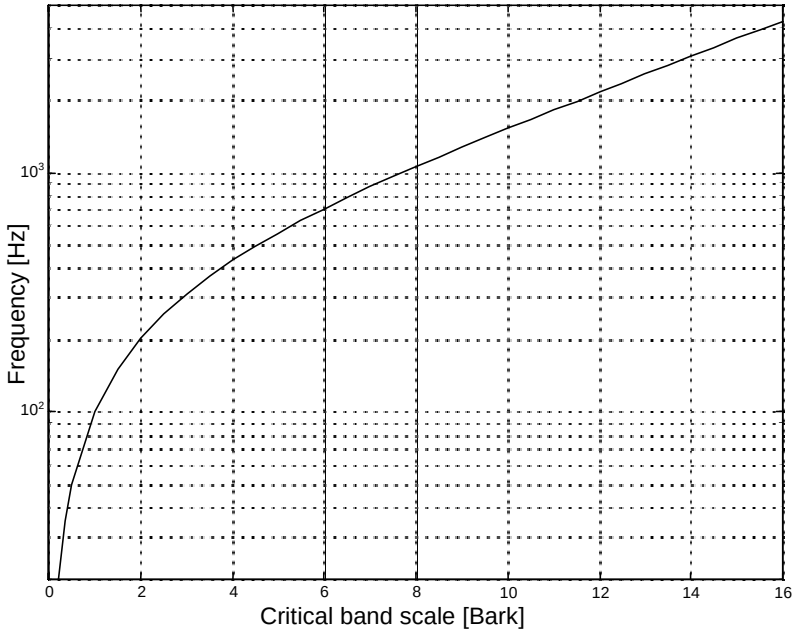


Figure 6: Relation between frequency domain and perceptual domain

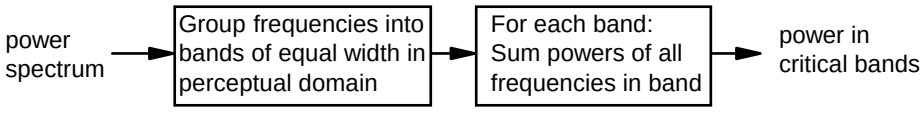


Figure 7: Mapping from frequency domain to perceptual domain

7.3.5 Frequency masking

The frequency analysis capabilities of the auditory system are not perfect. If two frequency components in a sound are sufficiently close to each other, the weaker one cannot be perceived. The extent of this frequency masking effect depends on the involved frequencies, their loudness levels, and other factors. The principle is sketched in figure 8.

Frequency masking can be approximated e.g. by a convolution in the frequency domain or in the perceptual domain. The slopes (i.e. the extent of exciting/masking adjacent frequencies) of the "masking triangle" depend on both centre frequency and intensity. (As mentioned above, this step can be implemented in a perceptual filterbank.)

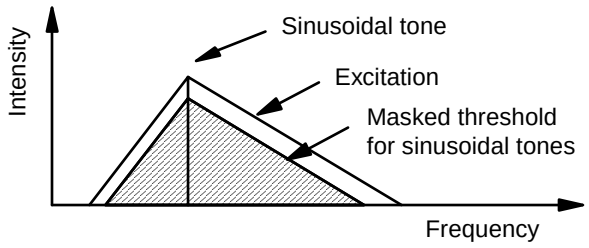


Figure 8: Frequency masking (from [16])

7.3.6 Time masking

Analogously to frequency masking, the auditory system is incapable of discriminating two short pulses separated by a small time interval, as sketched in figure 9. A coarse approximation of time-domain spreading is implicitly achieved for high frequencies by segmenting the speech signal, since time resolution is essentially determined by the time shift from one segment to the next. In general, this shift is at least 10 ms, which is sufficient for approximating time smearing at medium and high frequencies. On the other hand, low frequencies experience much larger (> 100 ms) time masking. Therefore, if this effect is to be modelled, it shall be implemented explicitly.

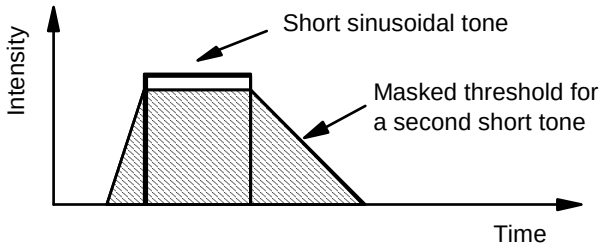


Figure 9: Time masking (from [16])

7.3.7 Psycho-acoustic loudness

The relation between physical loudness (signal power) and perceived loudness is highly non-linear. Two main effects can be distinguished:

- *Frequency-dependent sensitivity:* The hearing threshold of the auditory system is frequency-dependent, with maximum sensitivity in medium frequencies.
- *Loudness compression:* The perceived loudness is a non-linear function of the effective signal power, normalized to the hearing threshold. The mapping can be approximated by logarithmic compression, by raising the signal to a fractional power, or by using a function derived from psycho-acoustic experiments. In any case, the shape of the curve will look similar to figure 10.

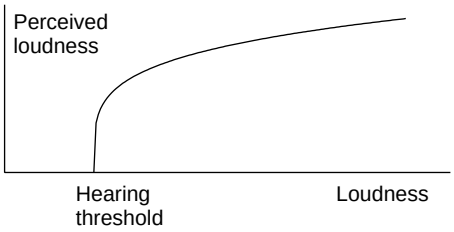


Figure 10: Loudness compression curve

7.3.8 Hair cell firing

Spectral analysis in the auditory system is performed by hair cells along the cochlea. Ultimately, these hair cells "fire" (i.e. generate all-or-none electrical spikes) based on their time-frequency analysis of the sound signal. While the exact relation between sound signal and firing patterns is not known, the firing activity can be described by probability distributions. Generally, firing activity of a hair cell increases with stimulation of that cell.

It can be argued that firing probabilities are a very close approximation of the information transmitted to the brain. On the other hand, they constitute a "low-level" representation without an intuitive interpretation. In contrast, effects such as loudness compression or frequency warping are based on human perception on a higher level, since they include the processing steps of the brain.

7.4 Comparison of reference and transmitted signal

After converting both reference and transmitted speech signal to a suitable representation, they are compared to assess the quality of the transmitted signal. The latter is usually assumed to be degraded, and thus the amount of "difference" or "non-similarity" between the two signals is taken as an indicator for the degradation of the transmitted signal. Hence, objective measurement schemes are akin to degradation category rating (see clause 6.2).

The comparison is most often performed on individual parameters or sets of parameters (e.g. all parameters from one signal segment). It is also possible to compute several comparison results for one set of parameters, using different comparison methods.

In a second step, the vector of comparison results is condensed to a single value (see clause 7.6). In this clause, the main comparison methods are discussed. For the sake of simplicity, it will be assumed that the comparison takes place between a vector ($x_1 ... x_n$) of the reference and a vector ($y_1 .. y_n$) of the transmitted signal.

7.4.1 Euclidean distance

The Euclidean distance is obtained as:

$$d_E = \sqrt{\sum_{i=1}^{n} c_i \times |y_i - x_i|^2} \tag{2}$$

It measures the length of the parameter difference vector in the usual geometrical sense. The factors c_i can be used to assign different weights to parameter differences, e.g. to compensate for different parameter ranges, or to weight the reliability of certain parameters.

7.4.2 Generalized distance

Instead of using Euclidean distance, one can compute more generally the L_p norm of a difference vector as:

$$d_p = \left(\sum_{i=1}^{n} c_i \times |y_i - x_i|^p\right)^{1/p} \tag{3}$$

For $p = 2$, the distance d_p coincides with the Euclidean distance. With increasing p, more weight is given to large differences. Again, the coefficients c_i can be used for weighting individual parameters.

7.4.3 Asymmetric differences

In some cases, the *sign* of the difference y_i - x_i is important. As an example, if a parameter represents the perceived loudness in a specific frequency band, an increased loudness in the transmitted signal could be more annoying than a reduced loudness. In such cases, it may be useful to compute asymmetric differences. As a simple example, consider:

$$d_i = \begin{cases} y_i - x_i & y_i \geq x_i \\ (x_i - y_i)/2 & y_i < x_i \end{cases} \quad (4)$$

Obviously, there are many ways of computing asymmetric differences.

7.4.4 Distance between probability functions

If the parameters to be compared are viewed as probability functions, the "distance" between these functions can be expressed in several ways. An example is the Kullback Leibler distance (relative entropy).

$$D(x \,\|\, y) = \sum_{i=1}^{n} x_i \log(x_i / y_i) \quad (5)$$

where both $x_i.x_n$ and $y_i.y_n$ are probability vectors that sum to one. The Kullback Leibler distance is also an example for asymmetric differences, since in general $D(x \,\|\, y)$ is not equal to $D(y \,\|\, x)$.

7.4.5 Multi-resolution analysis

Multi-resolution analysis is motivated by human auditory perception, which apparently works on various scales (e.g. short term/long term, broadband/narrowband). It can be used with any difference-computing scheme to obtain several difference values from one set of data.

For a multi-resolution analysis, one first measures differences between reference and transmitted signal on a coarse scale (e.g. broad frequency band or large time interval). The results are recorded, and the average difference is removed from the transmitted signal. Next, the process is repeated on a finer scale (e.g. on subbands or subintervals). This can be viewed as "zooming-in" in local signal features.

7.4.6 Compression to single number

The comparison process usually results in a vector of comparison results. In order to obtain a quality rating, the results shall be compressed to a single number. Two general methods can be distinguished.

- *Averaging:* The comparison results are averaged; different weights might be assigned to the components.
- *Classification:* The vector of comparison results is classified into one of a few quality classes. This can be achieved by means of a pattern classification method.

7.4.7 Mapping to MOS scale

The single number obtained in the compression step can be used as a measure for speech transmission quality. If compression is achieved by averaging, however, this single number is in general a non-linear function of the subjectively determined MOS value. In order to obtain an objective MOS value, it is necessary to map the result to the MOS scale. The mapping function is determined experimentally and constitutes the calibration of the method against stored subjective information.

8 Overview of INMD

The in-service, Non-Intrusive Measurement Device (INMD) described in ITU-T Recommendation P.561 [9] is intended for in-service (maintenance) application and aims at detecting network anomalies that affect the transmission performance of voice services. In contrast to the objective comparison methods described so far, INMD does not use a reference signal and thus belongs to the class of absolute estimation methods (see figure 1b). This simplifies the measurement process because measurements can be performed in-service during ordinary calls, thus eliminating the need for specific test calls and for equipment at the talker side. In turn, this allows a large number of measurements to be performed.

The INMD requires measurement of the following parameters:

Speech and noise characterization:

- active speech level: average signal level during active (non-silent) intervals of a speech connection;
- noise level (psophometric weighted): average signal level during speech pauses, weighted to account for psychoacoustic perception;
- speech activity factor: ratio of active speech time to total elapsed time.

Echo characterization:

- speech echo path delay (single or multiple reflection measurement): time delay of echo in received signal.

Additionally at least one of the following echo measurements:

- echo loss (single or multiple reflection measurement): frequency-weighted attenuation of echo path;
- echo path loss (single or multiple reflection measurement): unweighted attenuation of echo path;
- speech echo path loss (single or multiple reflection measurement): unweighted attenuation of speech echo path.

Optionally, the INMD may assess various other factors, e.g. originating/terminating address or crosstalk.

The parameters measured in the INMD are related to some parameters of the E-model (see next clause). In some sense, the INMD can be viewed as the "measurement counterpart" of the E-model, which itself is more concerned with planning issues. However, the INMD outputs raw measurement results which have a complex relationship with quality. It is possible to map these to a prediction of the conversational quality of the connection [18].

Currently, INMD is mainly considered for networks with "simple" distortions; the use of INMD in networks that include non-linear processing devices such as low-rate codecs is for further study.

9 Overview of the E-Model

The E-model (see ETR 250 [1], EG 201 050 [2] and ITU-T Recommendation G.107 [7]) is used as a tool for transmission planning in telephony networks. Its provides estimates of the expected communication quality resulting from combinations of impairments. The E-model differs in some respects from the methods described above:

- It is not a measuring tool, but a planning tool, although it can be used in association with measurements.
- It estimates two-way speech quality (i.e. communication quality) and takes into account echoes, delays, etc.

The input to the E-model consists of parameters which are available at the time of planning. (Note that planning is possible both prior to and after installation of a network.) These include factors such as noise, delays, echoes, and handset characteristics, which are subject to internationally accepted standards and recommendations, or which are known from experience or measurements. Additionally, the E-model weights the influence of modern digital equipment (low-rate codecs, multiplexers, etc.) on communication quality. In many cases, kind and number of such devices are known at the time of planning.

The E-model is based on the assumption that transmission impairments can be transformed into "psychological factors", and that these factors are additive on a "psychological scale". In other words, the subjective perception of speech quality is supposed to be equal to the sum of transmission impairments.

The E-model first computes a "base value" for quality, which is determined from network noise. Each further impairment is expressed as an impairment value, which is subsequently subtracted from the base value. This results in a predicted speech quality for a specific network. Finally, the resulting value for the speech quality can be used to estimate, what fraction of the user population would rate the quality as "good or better" and "poor or worse".

In particular, the E-model computes a *transmission rating factor R* as:

$$R = Ro - Is - Id - Ie + A \quad (6)$$

The transmission rating factor *R* consists of a base value *Ro*, of impairments *Is*, *Id*, and *Ie*, and of an expectation (or advantage) factor *A* as follows:

- *Ro* denotes the signal-to-noise ratio (SNR) of the connection. It includes noise in the network, in the environment at talker and listener side, and noise effects at the listener side. In the absence of other impairments, SNR is known to be a good indicator of speech quality.
- *Is* represents simultaneous impairments. These include excessive loudness levels, a sidetone level outside the comfortable range, and quantization distortions. The latter effect only includes "simple" distortions, such as PCM conversion. (Low-rate codecs and other non-linear devices are accounted for in *Ie*.)
- *Id* contains impairments due to delays and echoes. In particular, it includes talker echo, listener echo, and excessive delays.
- *Ie* comprises the impairments due to modern speech compression techniques (low-rate codecs, DCMEs). Since these impacts are difficult to quantify, *Ie* values for specific equipment are tabulated in [1] and [2].
- *A* allows the adjustment of quality in special situations by introducing non-technical aspects into the quality assessment, in particular customer expectations/requirements on QoS. For instance, customers rate GSM more highly than fixed connections despite their limited objective speech quality due to the advantages of mobility.

In the final step of the E-model, a non-linear mapping transforms the *R* value to an equivalent MOS value. (The value *R* is being replaced by *Itot = 94,3 - R*.)

10 Use of building blocks in some known systems

10.1 Comparison-based schemes

Table 2 illustrates the use of building blocks in some objective speech quality measurement systems. Omission from or inclusion in the table shall not be interpreted as a statement on the suitability of a system for objective speech quality measurements. Rather, the systems were chosen to cover the discussed building blocks. Note further that in some cases building blocks are implemented only partially, or implicitly as part of another block.

The overview table contains the following systems:

- *Cepstral Distance (CD):* Early scheme (1984) [24];
- *Cochlear Discrimination Information (CDI):* Model of cochlea up to hair cell firing [19];
- *Perceptual Evaluation of Speech Quality (PESQ):* ITU-T Recommendation P.862 [28];
- *Perceptual Speech Quality Measure (PSQM):* Former ITU-T Recommendation P.861 (see Bibliography);
- *Measuring Normalizing Blocks (MNB):* Contribution to ITU Study Group 12 [13];
- *PACE [*15*] and [*25*]*;

- *Telecommunication Objective Speech Quality Assessment (TOSQA) [*14*]*, see Bibliography, (4);
- *Perceptual Analysis/Measurement System (PAMS).*

Table 2: Use of building blocks in some objective speech quality measurement systems

Building block	CD	CDI	PESQ	PSQM	MNB	PACE	TOSQA	PAMS
Adjustment unit	•	•	•	•	•	•	•	•
Modelling transmission/receiver environment	---	---	•	•	---	---	•	•
Time-frequency mapping	(•)	(•)	•	•	•	•	•	•
Linear Prediction Coefficients	---	---	---	---	---	---	---	---
Cepstrum	•	---	---	---	---	---	---	---
Critical-band filtering	---	•	•	•	•	•	•	•
Frequency masking	---	•	•	•	(•)	•	•	•
Time masking	---	•	•	---	---	•	(•)	•
Perceptual loudness	(•)	•	•	•	•	•	•	•
Hair-cell firing	---	•	---	---	---	---	---	---
Euclidean distance (symmetric)	•	---	---	---	---	•	---	---
Asymmetric distance	---	---	•	•	•	---	•	•
Kullback Leibler distance	---	•	---	---	---	---	---	---
Multi-resolution differences	---	---	•	---	•	•	---	•
Compression by averaging	•	•	•	•	•	•	•	•
Compression by classification	---	---	---	---	---	---	---	---
Mapping to MOS scale	•	n/a	•	•	(•)	•	(•)	•

Legend:
- • method used
- (•) method used partially or implicitly
- --- method not used
- n/a not applicable

10.2 E-Model

In contrast to comparison-based methods, the E-model as described in clause 9 is mainly based on an input-output design. It assumes that impairments can be traced to simple electrical parameters of the network. Groups of such parameters and experimental evidence of their impact are fitted to some function in order to obtain additive impairment factors. In some sense, this implies a transformation to the perceptual domain. However, the E-model does not contain the building blocks typical for comparison-based schemes.

The E-model has primarily been used for planning. Using it for speech quality measurements requires measurement and/or estimation of its input parameters, some of which are difficult to assess.

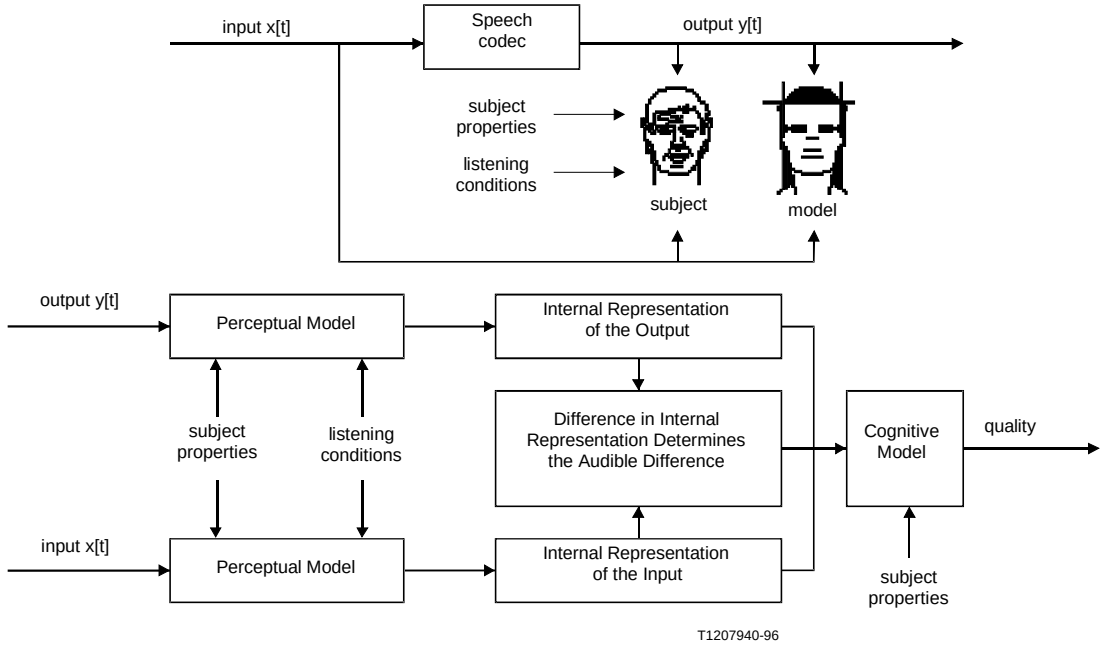
Annex A:
Void

Annex B (informative): Examples of specific systems

B.1 Perceptual Speech Quality Measure (PSQM)

PSQM was a part of ITU-T Recommendation P.861 (see bibliography), which has been withdrawn and replaced by the current ITU-T Recommendation for signal-based comparison P.862 [28].

Perceptual Speech Quality Measure (PSQM) speech quality measurement is performed according to the following schematic (see [4] and ITU-T Recommendation P.861, see bibliography):



NOTE: See ITU-T Recommendation P.861, see bibliography.

Figure B.1: Block diagram of PSQM; general view (top) and block diagram of model part (bottom)

PSQM is a method for measuring the quality of narrowband (300 Hz to 3 400 Hz) speech codecs. The idea behind PSQM is to mimic sound perception of subjects as in real-life situations. Experiments are simulated in which subjects judge the quality of speech codecs on their transparency when compared to a reference signal, mostly the input signal.

A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the output of the speech codec with the input.

Within PSQM the physical signals constituting the input and output of the speech codec are mapped onto psychophysical representations that match as closely as possible the internal representations of the speech signals (representations inside our heads). The internal representations make use of the psychophysical equivalents of frequency (Bark) and intensity (Compressed Sone). Masking is modelled in a simple way, only when two time-frequency components coincide in both the time and frequency domain is masking taken into account.

Within the PSQM approach the quality of the speech codec is judged on the basis of differences in the internal representation. This difference is used for the calculation of the noise disturbance as a function of time and frequency.

The average noise disturbance is directly related to the quality of the speech codec.

The transformation from the physical, external domain, to the psychophysical, internal domain, is performed by three operations:

- time-frequency mapping;
- frequency warping;
- intensity warping (compression).

Besides perceptual modelling the PSQM method also uses cognitive modelling in order to obtain a high correlation between subjective and objective measurements.

The output value of PSQM is an objective measure of the degradation introduced by the speech codec.

It is often desirable to express voice quality in MOS-values, but it is difficult to determine a unique function which transforms the PSQM value to the estimated MOS value (see [4] and ITU-T Recommendation P.861, see Bibliography).

PSQM is an off-line evaluation system. The original version was developed for the evaluation of waveform and CELP-type codec signals. A more recent version PSQM+ was extended to treat GSM signals and particularly GSM bit errors [5] and [12].

B.2 Measuring Normalizing Blocks (MNB)

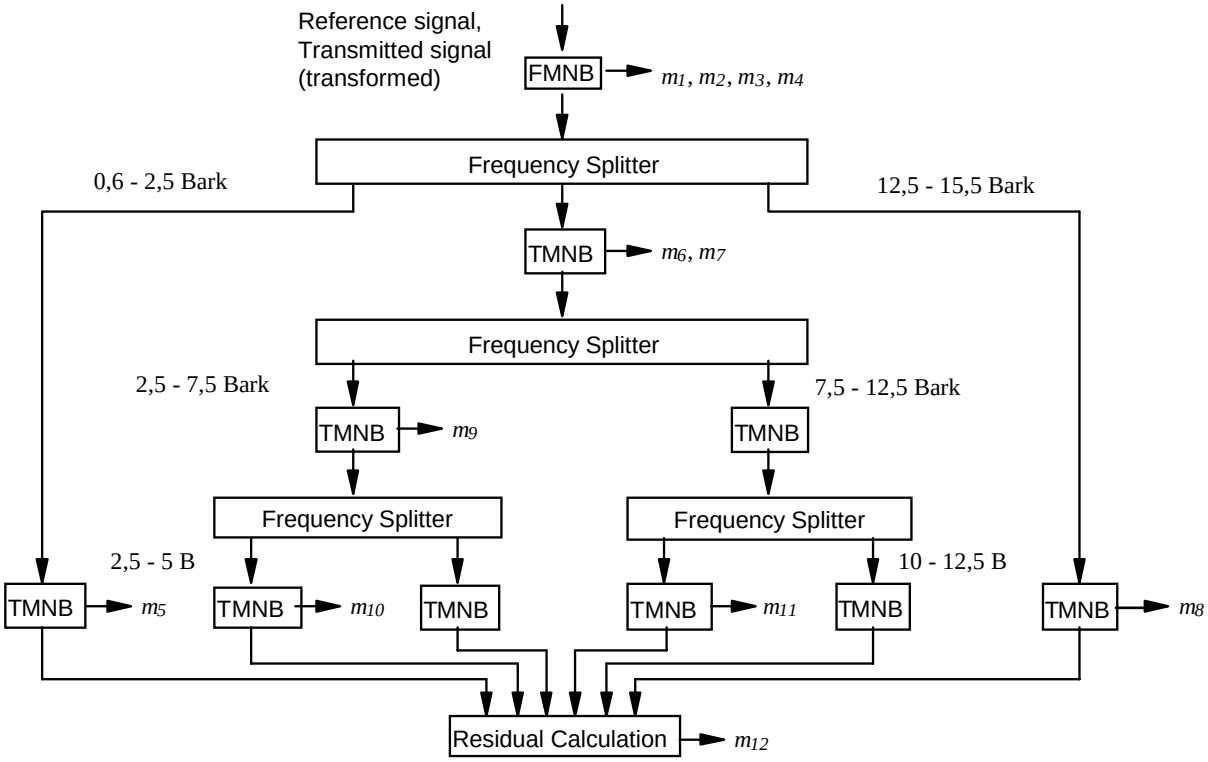


Figure B.2: MNB hierarchical structure (adapted from [13])

MNB was a part of ITU-T Recommendation P.861 (see bibliography), which has been withdrawn and replaced by the current ITU-T Recommendation P.862 [28].

The MNB system (*measuring normalizing blocks*) [13] follows the general structure in figure 3. In a first step, reference and transmitted signal are synchronized, DC components are eliminated, and the mean power of the two samples is normalized to the same level. In a second step, both signals are transformed to the frequency domain by an FFT, employing a Hamming window of 128 samples (16 ms) and a frame overlap of 50 %.

From the resulting sequence of frames, all frames are eliminated where the energy of either the reference frame or the transmitted signal frame falls below a certain threshold. Additionally, all frames with one or more zero-power frequency components are removed.

The retained frames are logarithmically transformed to approximate perceived loudness.

The subsequent comparison is based on two building blocks, the Time Measuring Normalizing Block (TMNB) and the frequency measuring normalizing block (FMNB). The TMNB integrates over some frequency scale, then measures differences and normalizes the transmitted signal frames for each time interval. Finally, the positive and negative portions of the measurements are integrated separately over time. In the FMNB, the role of time and frequency is exchanged.

The comparison process is sketched in figure B.2. Essentially, it performs a multi-resolution analysis in the critical-band domain using TMNBs. After an initial FMNB pass, differences in the highest, lowest, and medium frequencies are measured and recorded. After removing the differences in the medium frequency band, it is split into narrower bands, and the process is repeated for the individual subbands; subsequently, those bands are split and measured a third time. Finally, a residual difference between reference and transmitted signal is calculated.

The desired objective MOS value is then obtained by a linear combination of the comparison results.

B.3 PACE

The PACE algorithm [15], [25] was originally developed for a mouth-to-ear speech quality measurement system for mobile communication systems and is outlined in figure B.3.

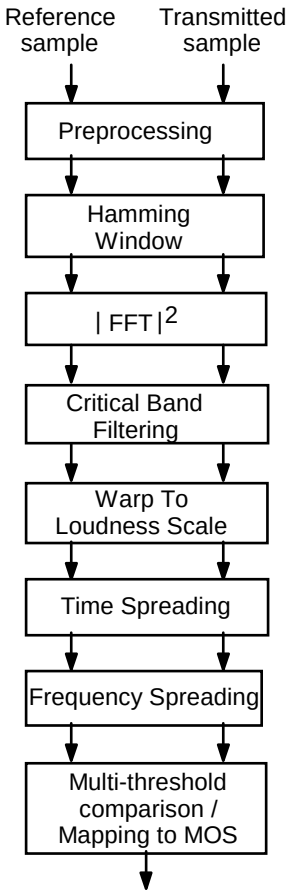


Figure B.3: Block diagram of PACE

Feature extraction consists of four steps (see figure B.3): First, the speech sample is split into 286 time slices of 18 ms duration each by using a Hamming window of length 256 and overlap of 128. The next step is to transform the signal into the frequency domain. The resulting values are divided into 18 *Bark* bands, and each frequency band is modelled separately. As a last step, a model of the psycho-physiological behaviour of the ear in response to the time variant signals with many spectral components is applied. All in all, the processing of one time slice yields 18 real values.

For comparing the reference and the transmitted signal, the similarity between the two signals is computed as a correlation coefficient, using only time slices that exceed a certain energy threshold (Correlation coefficients are equivalent to Euclidean distances). The computation is performed four times with different thresholds, thus yielding a multi-resolution analysis with respect to signal energy. Next, the resulting set of coefficients is reduced to a single value; in this calculation, coefficients corresponding to higher energy thresholds are given more importance. Finally, the result is transformed to a MOS value using a simple non-linear function.

B.4 Telecommunication Objective Speech Quality Assessment (TOSQA)

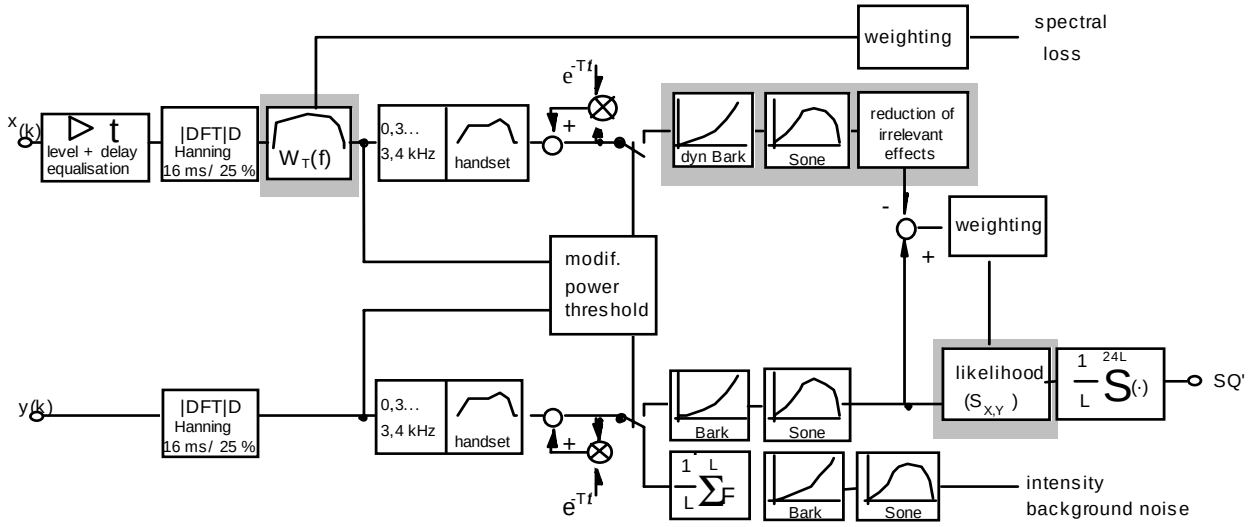


Figure B.4: Structure of TOSQA (new features and extended components are shadowed)

The basic structure of Telecommunication Objective Speech Quality Assessment (TOSQA) follows also the common structure and components as shown in figure 3. The signal processing is carried out off-line in several steps, as outlined in figure B.4. First the total delay will be estimated by a cross-correlation function. The second step computes the power levels in the telephone-band of both signals and their mean spectral envelope using an estimation of spectral power density. An auxiliary value similar to the coherence function will also be calculated to obtain an impression of the linearity of the system under test. The results from the first part of computation are used as control parameters for the main part of the program: the estimation of values describing speech quality impairments.

Along with the level and delay equalization for the input signal, the spectral distortions are also eliminated by using the estimated frequency response from step one. Both signals, the pre-processed input *x(k)* and the output signal *y(k)*, are windowed in segments of 16 ms and transformed in the frequency domain using DFT. A limitation to the relevant frequency range in the telephone band is the next part, as well as a weighting in the frequency domain with the receiving function of an ordinary telephone handset.

With the help of a power threshold the segments will be classified either to speech or non-speech parts. The weighted short time spectrums of the speech segments are the basis for frequency warping in critical bands (bandwidth: 1 Bark). TOSQA is strongly based on the ISO standard for the Zwicker model of calculation of specific loudness. Spectral masking and the compression is taken from the Zwicker model [6] and [27]. The classification in critical bands is realized by a dynamic frequency warping.

For calculating the similarity between the input and output characteristics, modified short time loudness spectrums are used. In this step also some differences between these loudness patterns that have low influence in the auditory test will be reduced. Part of this step is also an asymmetry weighting. The computed similarity is the main result of measuring the speech quality in TOSQA. The mean value of all short time results yield SQ. The internal value SQ will be transformed into a predicted MOS scale by using a non-linear monotonous mapping function. This program is also able to calculate additional values for describing quality effects (e.g. background noise, frequency response or non-linearity of the system under test).

(Text adapted from [14].)

B.5 Perceptual Analysis/Measurement System (PAMS)

PAMS is a speech quality assessment algorithm designed for mouth-to-ear assessment of telephone-bandwidth voice transmission systems. [23] presents an introduction to the processing in the perceptual layer described in figure B.5.

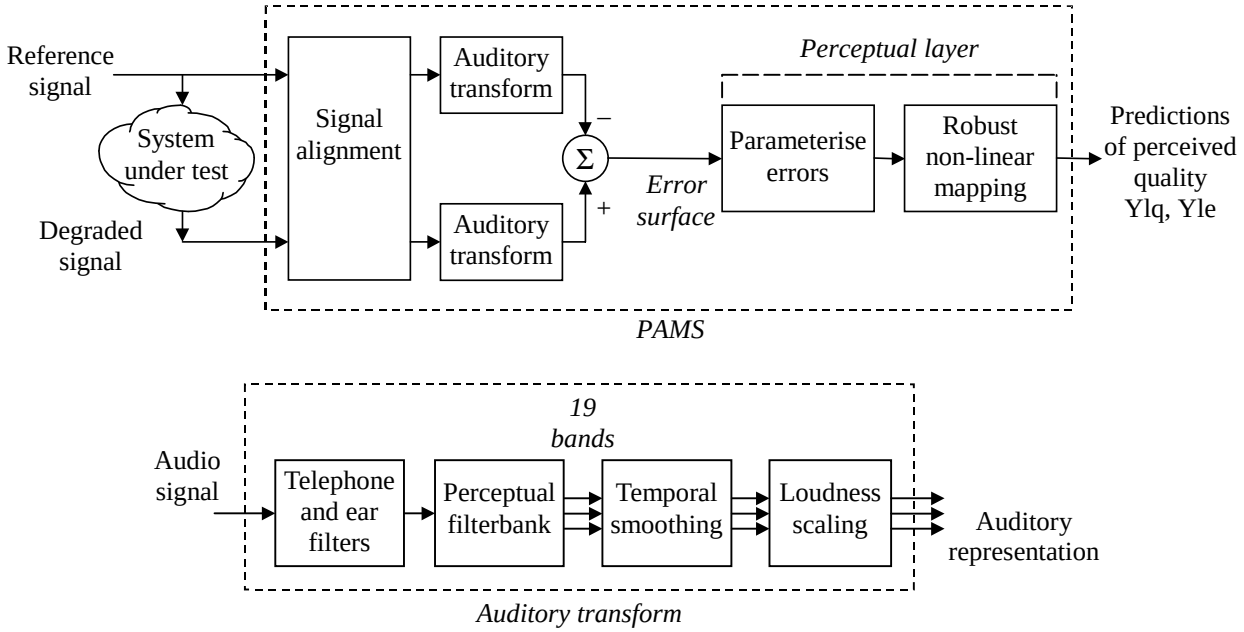


Figure B.5: Structure of PAMS

Diagrams indicating the structure of PAMS are shown in figure B.5. PAMS is given an original (reference) signal and the degraded version measured at the output of the network or system under test. The signals should be speech recordings or artificial speech-like test signals.

PAMS predicts subjective quality on two different opinion scales, Yle (listening effort) and Ylq (listening quality). These correspond to the absolute category rating scales defined in ITU-T Recommendation P.800 [10]. Listening quality is the most commonly used scale in network assessment, and is recommended for general use. The listening effort scale is also provided as it may be more appropriate for use with high levels of degradation. Certain other objective measures are produced for diagnostic purposes.

The emphasis in the design of PAMS has been to produce reliable predictions of speech quality even with unknown networks. For this reason, it includes the following components:

- A signal alignment stage to account for level and delay offsets in the network. This has been tested on a wide range of network connections, including low bit-rate codecs and packet networks. Recently this stage has been extended to make PAMS suitable for assessing VoIP connections.
- A detailed psychoacoustic model (the auditory transform) is used to estimate the audible errors introduced by the network. This transform uses a perceptual filterbank to model the time/frequency analysis, frequency perception and simultaneous masking performed by the human auditory system. Other stages include models of the acoustic response of the telephone and ear, temporal smoothing to approximate time masking, and mapping to a perceptual loudness scale.
- Error parametrization takes into account a variety of error classes, including background noise.
- Constrained non-linear mapping to perceived quality ensures that increasing errors always produce a decrease in quality, enhancing model robustness.

B.6 Perceptual Evaluation of Speech Quality (PESQ)

PESQ is the current ITU-T Recommendation for signal-based comparison (P.862 [28]), in replacement of ITU-T Recommendation P.861 (see bibliography), which has been withdrawn.

PESQ compares an original signal X(t) with a degraded signal Y(t) that is the result of passing X(t) through a communications system. The output of PESQ is a prediction of the perceived quality that would be given to Y(t) by subjects in a subjective listening test.

In the first step of PESQ a series of delays between original input and degraded output are computed, one for each time interval for which the delay is significantly different from the previous time interval. For each of these intervals a corresponding start and stop point is calculated. The alignment algorithm is based on the principle of comparing the confidence of having two delays in a certain time interval with the confidence of having a single delay for that interval. The algorithm can handle delay changes both during silences and during active speech parts.

Based on the set of delays that are found PESQ compares the original (input) signal with the aligned degraded output of the device under test using a perceptual model, as illustrated in figure B.6. The key to this process is transformation of both the original and degraded signals to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system, taking account of perceptual frequency (Bark) and loudness (Sone). This is achieved in several stages: time alignment, level alignment to a calibrated listening level, time-frequency mapping, frequency warping, and compressive loudness scaling.

The internal representation is processed to take account of effects such as local gain variations and linear filtering that may - if they are not too severe - have little perceptual significance. This is achieved by limiting the amount of compensation and making the compensation lag behind the effect. Thus minor, steady-state differences between original and degraded are compensated. More severe effects, or rapid variations, are only partially compensated so that a residual effect remains and contributes to the overall perceptual disturbance. This allows a small number of quality indicators to be used to model all subjective effects. In PESQ, two error parameters are computed in the cognitive model; these are combined to give a prediction of MOS.

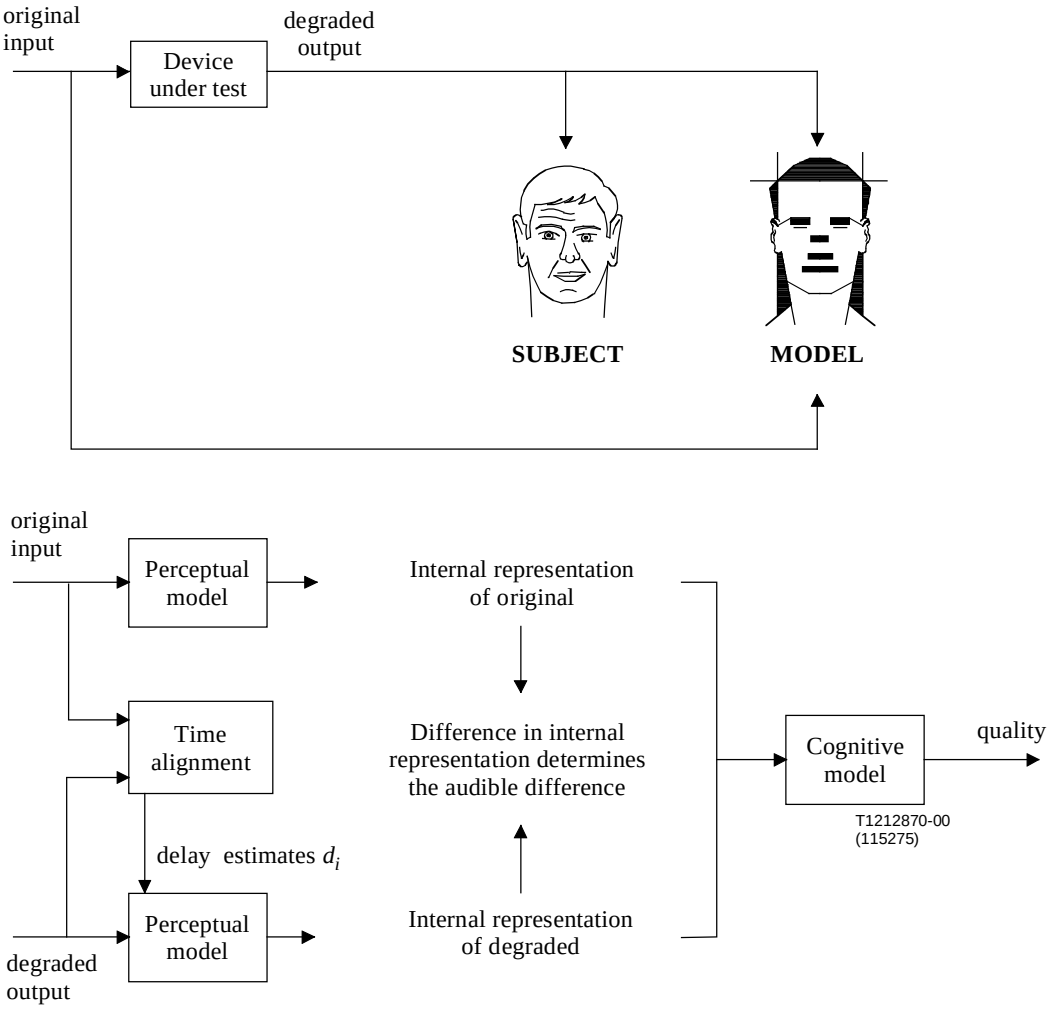


Figure B.6: Overview of the basic philosophy used in PESQ

A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the output of the device under test with the input, using alignment information as derived from the time signals in the time alignment module.

Annex C (informative): Terminal equipment related issues

C.1 Overview

In the overall transmission quality evaluation terminals determine widely the quality of a connection. The transfer functions and loudness ratings of a connection are mainly determined by the terminals, the background noise and the background noise transmission are highly influenced by the terminal and the acoustical environment the terminal is exposed to. The conversational properties which are the most important ones in a conversation are mainly determined by the terminal as well: double talk capability, switching characteristics and delay are dominant impairments often introduced by the terminal.

In order to find the determining factors a set of subjective test procedures have been developed allowing to extract the dominant quality aspects: Conversational test, talking and listening tests, double talk tests and listening only tests as described in [30], [31], [32] are the basis of the parameter extraction procedure.

An overview of the methodologies is given in figure C.1.

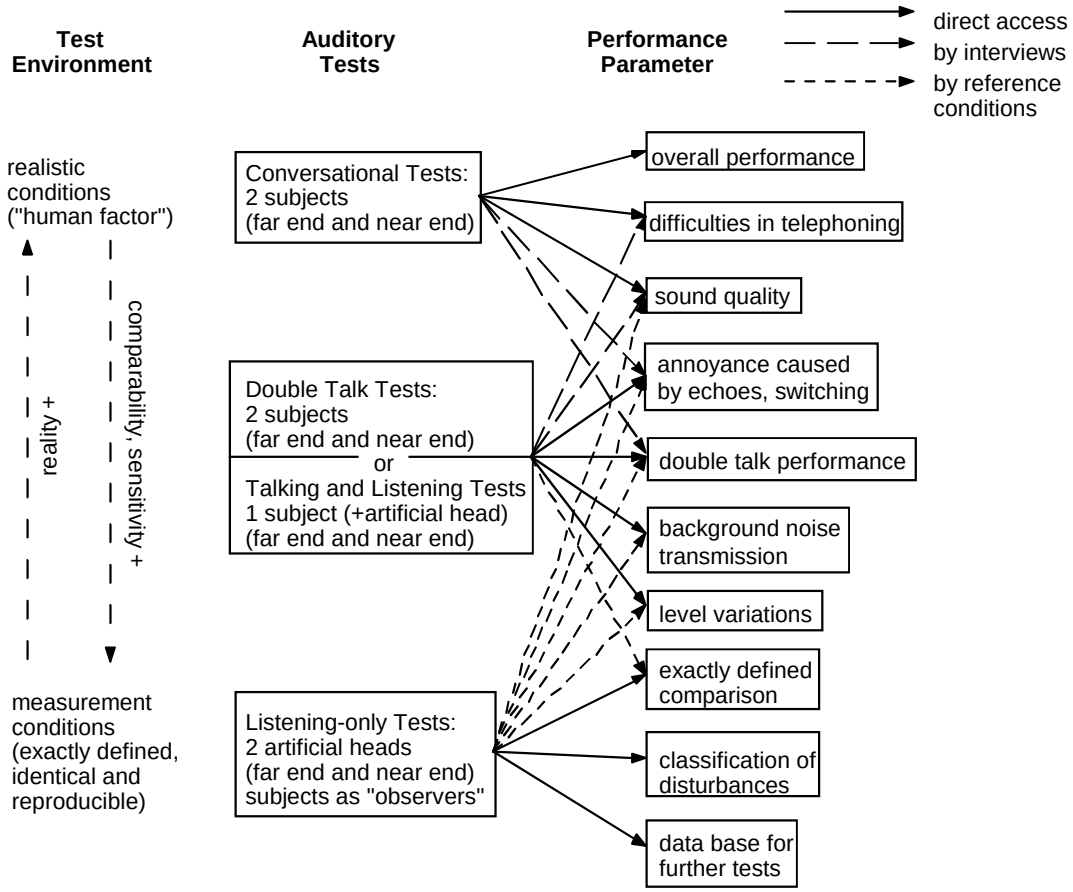


Figure C.1: Overview of test methods used for subjective evaluation

The subjectively relevant parameters determining "speech transmission quality" are detailed below. Overall quality is determined by:

- sound quality;
- quality of background noise transmission at idle, in single talk (and double talk conditions);
- speech level variations during single talk and double talk;
- disturbances caused by switching during single talk and double talk (completeness of speech transmission);
- disturbances caused by echoes during single talk and double talk.

Consequently, the objective evaluation needs to be divided into single talk measurements and double talk evaluations. In addition, evaluations are required during periods of silence where only background noise is present.

Since the typical test set-up should include all components involved in the mouth to ear transmission, a test arrangement should include the terminals "attached" to a realistic substitution of a user and his typical environment. Figure C.2 illustrates how a test set-up from end to end should look like.

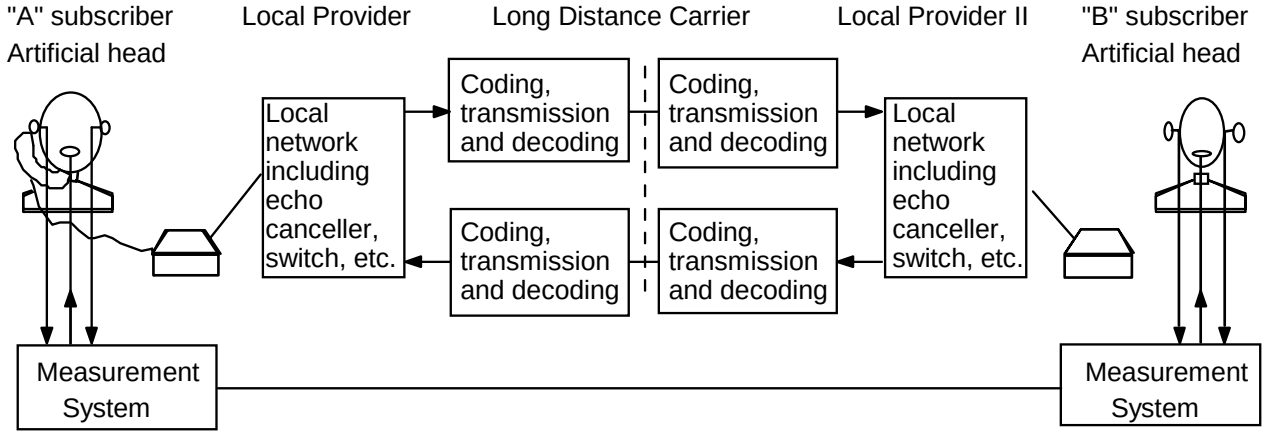


Figure C.2: Typical test set-up for determining the speech transmission quality from mouth to ear by objective parameters (example for handset/hands-free communication)

The methods to be used for the individual parameters are as follows:

- *Sound quality:* Sound quality is defined from mouth to ear in the orthotelephonic reference position. Thus the measurements need to include mouth and ear. In order to achieve realistic measurements, all measurements should be carried out using a realistic human head reproduction HATS, a realistic human ear reproduction (type 3.3 or 3.4) with realistic pressure forces [33], [34] and [35]. The sound quality is determined by frequency response, loudness rating and distortion. Those parameters can be measured using the appropriate procedures and speech-like test signals, which can be found in [8] and [36].

- *Quality of background noise transmission:* The quality of background noise transmission in single talk condition and during periods of silence (without signal or speech present) can be measured by applying the appropriate background noise to the system under test. The background noise transmission should be as low as possible. This can be verified by determining DelSM and the D-factor. A time and frequency dependent evaluation based e.g. on spectrogram evaluations derived from Fourier transformations or other kind of transformations. The variation of background noise should be low in time and frequency; in particular, no switching should occur. More detailed information can be found in [36].
  In the double talk situation, in general the same requirements are applicable although they may be more relaxed. The analysis of background noise is carried out during pauses of the double talk signal.

- *Speech level variations:* Speech level variations are typically evaluated by referring the measured output signal to the input signal. In order to measure under realistic conditions, speech or speech like test signals [8] need to be used. Also a realistic test set-up providing the appropriate conditions for the terminals (HATS equipped with type 3.3 or 3.4 artificial ear, realistic room conditions) should be chosen. Speech level variations should not be measurable. In case of any detectable level variations during single talk, they should be less than ±3 dB. In particular, no switching or interruptions should be measured. In case of switching, the quality-relevant numbers for switching time, switching levels, hang over times etc. can be found in [36].
  In the double talk situation, in general the same requirements are applicable although they may be more relaxed. The double talk measurements are carried out either by sequential combined test signals (CSS [8]) or by using orthogonal, speech-like test stimuli. More information can be found in [8], [32], [28] and [36].

- *Disturbances caused by switching:* Disturbances caused by switching from the instrumental point of view can be seen as a specific kind of speech level variation. Thus the same objective procedures as for speech level variations apply.

- *Disturbances caused by echoes:* Disturbances caused by echoes during single talk are defined by a variety of instrumental procedures: convergence behaviour of speech echo-cancellers (if present) under various conditions such as background noise or initial double talk, ERL provided by the terminal itself, adaptation speed for time-varying echo paths. All the above mentioned properties can be measured by time and frequency variant measurement procedures (e.g. spectral echo attenuation). For the measurement, speech like signals as described in [8] and [37] should be used. Limits based again on subjective evaluations can be found in [29] and [36]. All measurements should be carried out using the appropriate measurement set-up (as described above) and using the proper environmental conditions: realistic rooms, background noise and time varying echo paths. It should be noted that echo control in almost all components (network as well as terminals) is backed up by centre clippers or special speech level dependent level variations or switching.
  In the double talk situation, a more sophisticated analysis technique is required. Specific orthogonal double talk sequences either sequentially combined in time or in frequency should be used (see e.g. [37]). Detailed information for limits derived from subjective experiments are currently evaluated and will be available in a new version of [36].

Annex D (informative): Subjective measurement methods

D.1 Absolute Category Rating (ACR)

In Absolute Category Rating (ACR) tests, listeners are instructed to rate the "absolute" quality of speech samples, i.e. without comparison to a reference. The rating scale to be used depends on the parameter to assess. In the context of the present document, the most relevant scale is listening quality. ITU-T Recommendation P.800 [10] recommends the scale shown in table D.1.

Table D.1: Listening quality scale for absolute category rating

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

Experiment results are reported as numerical means in the form of mean opinion scores (MOS). The relevance of MOS results can be determined by statistical analysis, e.g. by significance tests and confidence intervals.

D.2 Degradation Category Rating (DCR)

When speech samples of good quality are evaluated, ACR tends to be insensitive, to the effect that small differences in quality are not detected. In such cases, *Degradation Category Rating* (DCR) can be applied. In this variation, listeners are presented pairs (A-B) or repeated pairs (A-B-A-B) of speech samples. The sample A is a quality reference, the quality of which depends on the application. Sample B is a degraded version of A. Listeners are instructed to rate the samples B according to a degradation category scale. ITU-T Recommendation P.800 [10] recommends the scale shown in table D.2.

Table D.2: Opinion scale for degradation category rating [10]

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

As for ACR, results are reported as numerical means, but in the form of *Degradation Mean Opinion Scores* (DMOS). In contrast to MOS, DMOS values depend on the quality reference. Thus, different experiments can only be compared if they share the same reference (or a reference of equal quality). Again, it is advisable to assess the validity of results by statistical tests.

D.3 Comparison Category Rating (CCR)

Comparison Category Rating (CCR) is similar to DCR. However, instead of presenting pairs in the order "reference sample - processed sample", the order is chosen at random. For each test condition, half of the sample pairs are presented in the order "processed sample - reference sample". Thus, the second sample may be better or worse than the first one. The listeners are asked to rate the second sample in comparison to the first one according to the scale in table D.3. A potential application for CCR is the assessment of speech enhancement systems, where the processed sample may actually be better than the reference.

Table D.3: Opinion scale for comparison category rating [10]

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

Results are expressed as *Comparison Mean Opinion Scores* (CMOS).

D.4 Interview and survey test

If the large effort is justified, speech transmission quality can be determined by "service observations", i.e. by interviewing customers. This is covered in more detail in ITU-T Recommendation P.82 [39]. A related, but more automated method called SIBYL is described in [38]. In this scheme, customers agree to have a small proportion of their calls artificially distorted according to a test programme. The volunteers are asked to vote on such calls by dialling a digit on their telephone equipment.

D.5 Conversational tests

A conversation test involves two parties conversing over a connection, and depending on the purpose of the test, either experienced or untrained subjects can be used. Such tests can be useful to both manufacturers and customers, and are an important assessment tool because they provide the closest simulation of real telephone interactions between two subscribers. Untrained subjects are used when it is important to get an indication of how the general telephone-using population would rate the overall quality and difficulty. Experienced subjects are used in situations where it is necessary to obtain information about the subjective effects of individual degradations.

The benefit of conversational testing is that it is the only way of realistically assessing the combined subjective effect of all the parameters affecting conversational quality. In particular, effects such as level variations, echo and double-talk can have a marked effect on conversational performance.

Typically, the following quality parameter can be assessed:

- overall quality;
- (speech) sound quality;
- difficulty in talking or listening;
- dialogue capability;
- echo performance;
- quality of the background noise transmission.

Detailed information can be found in [10], [27], [28] and [29].

D.6 Double talk tests

Comparable to the conversational test the double talk test involves two parties and, depending on the purpose of the test, either experienced or untrained subjects can be used. These double talk tests are an important evaluation tool because they assess the transmission quality during periods of double talk in detail. Conversational tests clearly pointed out that the double talk performance highly influences the naturalness of a conversation. Typically in such tests one subject is talking continuously while the other is interrupting. This gives the possibility to ask different parameters for both subjects during the test. The following table highlights the parameters that typically determine double talk performance and were therefore used and successfully evaluated.

Detailed information can be found in [31] and [32].

Table D.4: Parameters determining double talk performance

"Talking continuously"	"Interrupting"
Double talk capability	Double talk capability
Completeness of speech transmission	Completeness of speech transmission
Loudness during double talk	Loudness during double talk
	Loudness variation single talk/double talk
Echo	Echo
Echo variation single talk/double talk	
Sound quality	Sound quality single talk/double talk
Transmission of background noise	Transmission of background noise

D.7 Talking and listening tests

The subjective performance under single talk conditions can be investigated in a more efficient way if the lack of the complete conversation can be tolerated. All aspects which influence the transmission quality for subscribers while they are currently talking or while they are listening (without having a conversational partner on the other end of the connection) are covered by this test. The procedure is specially adapted for the evaluation of talking-related disturbances, such as disturbances caused by echoes, disturbances caused by switching, or disturbances related to the transmission of background noise. Typically, initial convergence situations as well as steady-state conditions are evaluated.

Detailed information can be found in [30].

D.8 Listening-only test procedure

The principle of a listening-only test procedure is to record specially designed speech material in advance to play it back to the subjects during the test session for evaluation. This test procedure is designed to evaluate and compare the individual performance parameters of different terminals, different algorithm implementations, or different measurement conditions in one test.

Subjects judge the quality either by using handsets or if conversational recordings have been made between a pair of correctly equalized HATS reproduced by correctly equalized headphones. Typically, test subjects listen to pre-recorded speech samples. In specific cases, test subjects act as third-party listeners. Third-party listener means that the subjects are observers of a conversation, standing in the position beside the near end speaker. The test is applicable to situations where the recording procedure needs to reproduce the listening situation as realistically as possible.

Reference conditions can easily be included, because they can be generated off-line. These listening examples can be presented with the real recordings during the test. Reference conditions allow results from different labs to be compared, and may include test set-ups under well defined conditions.

The listening-only test is well suited to the evaluation of specific parameters to give a very detailed and precise description of the achieved transmission quality of the candidates under test because subjects can *concentrate better on these parameters*.

Detailed information can be found in [10], [27], [28] and [29].

Annex E (informative): Application of statistical methods

E.1 Statistical relevance of results

Subjective MOS values are obtained by averaging the opinion scores of all listeners in an experiment; i.e.:

$$MOS = (x_1 + x_2 + ... + x_N) / N \quad (7)$$

where N is the number of participants and x_i denotes an individual opinion score.

The resulting MOS value is only a statistical *estimate* of the "true" MOS because there is only a limited number of listeners. Moreover, listeners use a coarse (integer-valued) opinion scale, which results in a kind of quantization noise.

It is therefore important to assess the statistical accuracy of the computed MOS values. This can be accomplished by determining a *confidence interval* in addition to a MOS value. For example, a 90 % confidence interval [a,b] indicates that the "true" MOS value lies in the interval [a,b] with a probability of 90 %.

The length of the confidence interval depends on several factors:

- *Confidence level:* A higher confidence level (e.g. 99 % instead of 90 %) results in a larger confidence interval.
- *Agreement among listeners:* A higher variance of the individual opinion scores (i.e. less agreement among listeners) yields a larger confidence interval.
- *Size of listener panel:* Increasing the number of listeners (without changing other parameters) decreases the length of the confidence interval.

If an a priori estimate for the standard deviation σ_x of individual opinion scores is available, a rule-of-thumb estimate (assuming a Gaussian distribution of individual scores) of the required number N of listeners for a specific confidence level is given by:

$$N = \left(\frac{a\sigma_x}{L/2}\right)^2 \quad (8)$$

where L is the desired interval length, σ_x is the standard deviation of individual opinion scores, and a is a constant that depends on the desired confidence level (e.g. a = 1 for 68 % and a = 2 for 95 %). As an example, a 95 % confidence that the computed MOS value is accurate to ±0,1 MOS points, given a standard deviation of 0,5 MOS points for individual scores, requires a panel of approximately $(2 \times 0{,}5/0{,}1)^2 = 100$ listeners.

A related confidence problem arises in objective measurement systems, which output a single MOS estimate at a time. Typically, one wants to know whether a certain quality is achieved or, equivalently, whether the "true" MOS value is in some given interval. This question can be answered if the probability distribution for the measured MOS value is known, conditioned on the hypothesis that the "true" MOS is in some quality interval. The necessary information may come from experience, or from repeating the measurement under comparable conditions. However, such knowledge is often more difficult to acquire than in subjective opinion tests.

Nevertheless, if statistical information is available, the relevance of a measurement can be assessed. The principle is illustrated in figure E.1. Assume first that the MOS value at point A was measured. In this case, one can deduce with high confidence that the actual speech quality is fair. On the other hand, if point B was measured, the actual speech quality may be fair or bad it is impossible to distinguish these cases from the single MOS measurement at point B.

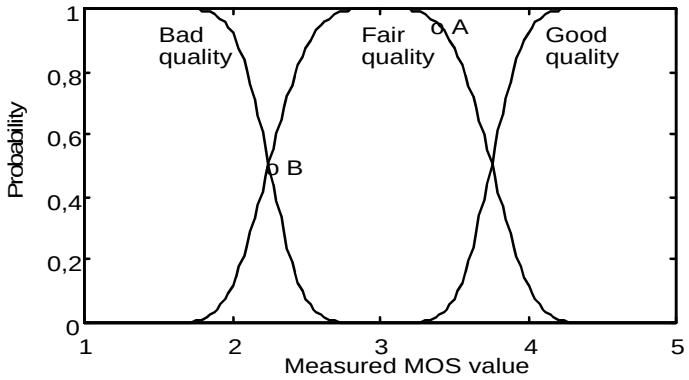


Figure E.1: Probability of measuring a certain MOS value if true quality is bad, fair or good (example)

E.2 Estimation of confidence intervals

A simple method to obtain a confidence interval is based on the estimated *standard deviation* σ_{MOS} of the MOS value:

$$\sigma_{MOS} = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i - MOS)^2} \qquad (9)$$

Under the (wrong) assumption that computed MOS values are drawn from a Gaussian probability distribution, the interval $[MOS - \sigma_{MOS}, MOS + \sigma_{MOS}]$ is a 68 % confidence interval for *MOS*. Obviously, the Gaussian assumption is problematic for small sample sizes and for cases where the underlying distribution of individual scores diverges strongly from a normal distribution.

More accurate confidence intervals are obtained with *bootstrap* techniques [40], which do not invoke a Gaussian assumption. The basic bootstrap algorithm for calculating a confidence interval for a mean is shown in table E.1.

In practice the (wrong) Gaussian calculus used here is replaced by the more correct calculus based on the Student or t-distribution. This yields better results for both the standard deviation method and the bootstrap method.

Table E.1: Bootstrap principle for calculating a confidence interval for the mean (adapted from [40])

Step 0	*Experiment.* Conduct the experiment. Suppose the number of samples is *N* and the sample mean is μ.
Step 1	*Resampling.* Using a pseudo-random number generator, draw a random sample of *N* values, with replacement, from the set of samples. This is called a bootstrap resample. Note that some of the original sample values may appear more than once, and others not at all.
Step 2	*Calculation of the bootstrap estimate.* Calculate the mean μ_k of all values in the bootstrap resample.
Step 3	*Repetition.* Repeat steps 1 and 2 a large number of times to obtain a total of *K* bootstrap estimates μ_k.
Step 4	*Approximation of the distribution of* μ_k. Sort the bootstrap estimates μ_k in increasing order.
Step 5	*Confidence interval.* The desired $(1 - \alpha)100$ % bootstrap confidence interval is given by the μ_k at position $K\alpha/2$ (lower end) and at position $K(1 - \alpha/2)$ (upper end).

E.3 ANOVA

In the context of subjective speech quality measurement, analysis of variances (ANOVA) can be used to assess the influence of different test parameters (such as gender, talkers, or listeners speech level) on the results.

To illustrate the principle, consider figure E.2. The samples are collected in disjoint groups A to D; for example, a listening panel could be presented the same speech sample with four different loudness levels, or the scores of one sample could be distributed to four groups according to the listeners age. On the left side of figure E.2, there is no evidence that the groups behave differently: the variance *within* a group masks any differences in the mean value. On the right side of figure E.2, however, the variance *between* the groups exceeds the variance within a group; apparently, the groups have different means.

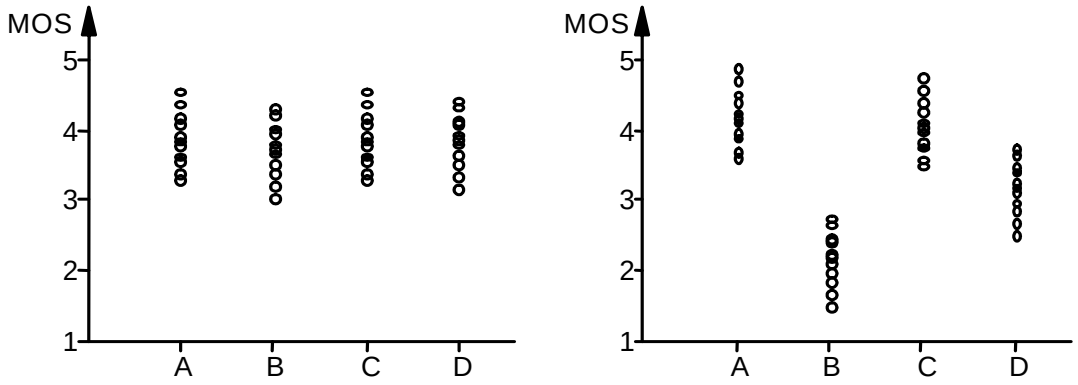


Figure E.2: Listening experiment, results for different groups; a) probably equal means, b) different means

The basic ANOVA algorithm shown in table E.2 allows to detect with a certain confidence level the case that at least one group has a different mean. The basic technique can be extended and generalized in several ways; the interested reader is referred to the statistical literature.

It should be noted that ANOVA is based on the assumption that each sample follows a Gaussian distribution with identical variance. Thus, results may be misleading if the sample size is too small.

Table E.2: ANOVA for testing whether disjoint groups of samples have the same mean

Step 0	*Experiment*	Conduct the experiment. Suppose the total number of samples is N and the samples are collected in K disjoint groups, where the k-th group contains n_k samples.
Step 1	*Group statistics*	For each group k, compute the sample mean μ_k and the sample variance σ_k^2.
Step 2	*"Within variance"*	Compute the "within variance" $\sigma_W^2 = \frac{1}{N-K}\sum_{k=1}^{K}(n_k-1)\sigma_k^2$.
Step 3	*"Between variance"*	Compute the "between variance" $\sigma_B^2 = \frac{1}{K-1}\sum_{k=1}^{K}(\mu_k-\mu)^2$, where μ is the mean of the μ_k.
Step 4	*Threshold test*	Test whether $\sigma_B^2/\sigma_W^2 > F_{\alpha,\ K-1,\ N-K}$. If this is the case, not all K groups have the same mean, with confidence level $(1 - \alpha)100$ %. (The F-function is tabulated in the statistical literature).

Annex F (informative): Bibliography

1) ITU-T Recommendation P.861: "Objective quality measurement of telephone-band (300 - 3 400 Hz) speech codecs".

2) ITU-T (1992): "Handbook on Telephonometry".

3) ITU-T Recommendation P.50: "Artificial voices; Appendix: Test Signals".

4) Berger, J. (1998): "Instrumentelle Verfahren zur Sprachqualitätsschätzung-Modelle auditiver Tests (Instrumental approaches for speech quality estimation-models of auditory tests)", Ph.D. thesis, Christian-Albrechts-University of Kiel, Shaker-Verlag. ISBN 3-8265-4092-3

5) Klaus H., Berger J. (1997): "Die Bestimmung der Telefon-Sprachqualität für die Übertragungskette vom Mund zum Ohr - Herausforderungen und ausgewählte Verfahren. Deutsche Telekom".

History

Document history		
V1.1.1	April 1999	Publication
V1.2.1	October 2002	Membership Approval Procedure MV 20021213: 2002-10-15 to 2002-12-13
V1.2.1	December 2002	Publication