ETSI ES 202 739 V1.3.1 (2009-09)

ETSI Standard

Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband VoIP terminals (handset and headset) from a QoS perspective as perceived by the user

Reference

RES/STQ-00125

Keywords

quality, telephony, terminal, VoIP

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

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).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This ETSI Standard (ES) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ).

Introduction

Traditionally, the analogue and digital telephones were interfacing switched-circuit 64 kbit/s PCM networks. With the fast growth of IP networks, wideband terminals providing higher audio-bandwidth and directly interfacing packet-switched networks (VoIP) are being rapidly introduced. Such IP network edge devices may include gateways, specifically designed IP phones, soft phones or other devices connected to the IP based networks and providing telephony service. Since the IP networks will be in many cases interworking with the traditional PSTN and private networks, many of the basic transmission requirements have to be harmonized with specifications for traditional digital terminals. However, due to the unique characteristics of the IP networks including packet loss, delay, etc. New performance specification, as well as appropriate measuring methods, will have to be developed. Terminals are getting increasingly complex, advanced signal processing is used to address the IP specific issues.

NOTE: Requirement limits are given in tables, the associated curve when provided is given for illustration.

1 Scope

The present document provides speech transmission performance requirements for 8 kHz wideband VoIP handset and headset terminals; it addresses all types of IP based terminals, including wireless and soft phones.

In contrast to other standards which define minimum performance requirements it is the intention of the present document to specify terminal equipment requirements which enable manufacturers and service providers to enable good quality end-to-end speech performance as perceived by the user.

In addition to basic testing procedures, the present document describes advanced testing procedures taking into account further quality parameters as perceived by the user.

2 References

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.
- Non-specific reference may be made only to a complete document or a part thereof and only in the following cases:
  - if it is accepted that it will be possible to use all future changes of the referenced document for the purposes of the referring document;
  - for informative references.

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

NOTE: While any hyperlinks included in this clause were valid at the time of publication ETSI cannot guarantee their long term validity.

2.1 Normative references

The following referenced documents are indispensable for the application of the present document. For dated references, only the edition cited applies. For non-specific references, the latest edition of the referenced document (including any amendments) applies.

[1] ETSI I-ETS 300 245-5: "Integrated Services Digital Network (ISDN); Technical characteristics of telephony terminals; Part 5: Wideband (7 kHz) handset telephony".

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

[3] ITU-T Recommendation G.108: "Application of the E-model: A planning guide".

[4] ITU-T Recommendation G.109: "Definition of categories of speech transmission quality".

[5] ITU-T Recommendation G.122: "Influence of national systems on stability and talker echo in international connections".

[6] ITU-T Recommendation G.711: "Pulse code modulation (PCM) of voice frequencies".

[7] ITU-T Recommendation G.722: "7 kHz audio-coding within 64 kbit/s".

[8] ITU-T Recommendation G.722.1: "Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss".

[9] ITU-T Recommendation G.729.1: "G.729 based Embedded Variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729".

[10] ITU-T Recommendation G.1020: "Performance parameter definitions for quality of speech and other voiceband applications utilizing IP networks".

[11] ITU-T Recommendation P.50: "Artificial voices".

[12] ITU-T Recommendation P.56: "Objective measurement of active speech level".

[13] ITU-T Recommendation P.57: "Artificial ears".

[14] ITU-T Recommendation P.58: "Head and torso simulator for telephonometry".

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

[16] ITU-T Recommendation P.79: "Calculation of loudness ratings for telephone sets".

[17] ITU-T Recommendation P.340: "Transmission characteristics and speech quality parameters of hands-free terminals".

[18] ITU-T Recommendation P.380: "Electro-acoustic measurements on headsets".

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

[20] ITU-T Recommendation P.502: "Objective test methods for speech communication systems using complex test signals".

[21] ITU-T Recommendation P.581: "Use of head and torso simulator (HATS) for hands-free terminal testing".

[22] ITU-T Recommendation P.862: "Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs".

[23] IEC 61260: "Electroacoustics - Octave-band and fractional-octave-band filters".

[24] ISO 3 (1973): "Preferred numbers - Series of preferred numbers".

[25] L16-256: "TIA-920, Transmission Requirements for Wideband Digital Wireline Telephones, TELECOMMUNICATIONS INDUSTRY ASSOCIATION, TIA/EIA".

2.2 Informative references

The following referenced documents are not essential to the use of the present document but they assist the user with regard to a particular subject area. For non-specific references, the latest version of the referenced document (including any amendments) applies.

[i.1] ETSI EG 201 377-1: "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".

[i.2] ETSI EG 202 396-1: "Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database".

[i.3] ETSI EG 202 425: "Speech Processing, Transmission and Quality Aspects (STQ); Definition and implementation of VoIP reference point".

[i.4] ETSI EG 202 396-3: "Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise Part 3: Background noise transmission - Objective test methods".

[i.5] ITU-T Recommendation P.800.1: "Mean Opinion Score (MOS) Terminology".

[i.6] ETSI TR 102 648-1: "Speech Processing, Transmission and Quality Aspects (STQ); Test Methodologies for ETSI Test Events and Results; Part 1: VoIP Speech Quality Testing".

[i.7] NIST net.

NOTE: Available at (http://snad.ncsl.nist.gov/itg/nistnet/).

[i.8] Netem.

NOTE: Available at (http://www.linuxfoundation.org/en/Net:Netem).

[i.9] Trace Control for Netem (TCN): "A. Keller, Trace Control for Netem, Semester Thesis SA-2006-15, ETH Zürich, 2006".

[i.10] Poschen, S., Kettler, F.; Raake, A.; Spors, S.: "Testing Wideband Terminals", DAGA 2008, March 10-13, Dresden, Proceedings.

3 Definitions and abbreviations

3.1 Definitions

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

artificial ear: device for the calibration of earphones incorporating an acoustic coupler and a calibrated microphone for the measurement of the sound pressure and having an overall acoustic impedance similar to that of the median adult human ear over a given frequency band

codec: combination of an analogue-to-digital encoder and a digital-to-analogue decoder operating in opposite directions of transmission in the same equipment

diffuse field equalization: equalization of the HATS sound pick-up, equalization of the difference, in dB, between the spectrum level of the acoustic pressure at the ear Drum Reference Point (DRP) and the spectrum level of the acoustic pressure at the HATS Reference Point (HRP) in a diffuse sound field with the HATS absent using the reverse nominal curve given in table 3 of ITU-T Recommendation P.58 [14]

ear-Drum Reference Point (DRP): point located at the end of the ear canal, corresponding to the ear-drum position

freefield reference point: point located in the free sound field, at least in 1,5 m distance from a sound source radiating in free air (in case of a head and torso simulator (HATS) in the centre of the artificial head with no artificial head present)

Head And Torso Simulator (HATS) for telephonometry: manikin extending downward from the top of the head to the waist, designed to simulate the sound pick-up characteristics and the acoustic diffraction produced by a median human adult and to reproduce the acoustic field generated by the human mouth

Mouth Reference Point (MRP): is located on axis and 25 mm in front of the lip plane of a mouth simulator

nominal setting of the volume control: when a receive volume control is provided, the setting which is closest to the nominal RLR of 2 dB

3.2 Abbreviations

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

CSS Composite Source Signal
D D-value of terminal
DRP ear Drum Reference Point
EL Echo Loss
ELR Echo Loudness Rating
HATS Head And Torso Simulator

MOS-LQOy Mean Opinion Score - Listening Quality Objective

NOTE: See ITU-T Recommendation P.800.1 [i.5].

ADC	A/D-Converter
AM-FM	Amplitude Modulation - Frequency Modulation
DAC	D/A-Converter
ERP	Ears Reference Point
HRP	HATS Reference Point
IP	Internet Protocol
LAN	Local Area Network
MOS-LQOM	Mean Opinion Score - Listening Quality, Objective, Mixed
MRP	Mouth Reference Point
N	Noise
NLP	Non Linear Processor
PCM	Pulse Code Modulation
PLC	Packet Loss Concealment
POI	Point Of Interconnect
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RLR	Receive Loudness Rating
SLR	Send Loudness Rating
Ssi(diff)	Send sensitivity, Diffuse Sound Field
Ssi(direct)	Send sensitivity, Direct Sound Field
STMR	SideTone Masking Rating
TCLw	Terminal Coupling Loss (weighted)
TCN	Trace Control for Netem
TOSQA	Telecommunication Objective Speech Quality Assessment
VAD	Voice Activity Detection
VoIP	Voice over IP

4 General considerations

4.1 Coding algorithm

The assumed coding algorithm is according to ITU-T Recommendation G.722 [7]. VoIP terminals may support other coding algorithms.

NOTE: Associated Packet Loss Concealment, e.g. as defined in ITU-T Recommendation G.722 [7], Appendixes 3 and 4 should be used.

4.2 End-to-end considerations

In order to achieve a desired end-to-end speech transmission performance (mouth-to-ear) it is recommended that the general rules of transmission planning are carried out with the E-model of ITU-T Recommendation G.107 [2] taking into account that the E-model does not yet address wideband transmission planning; this includes the a-priori determination of the desired category of speech transmission quality as defined in ITU-T Recommendation G.109 [4].

While, in general, the transmission characteristics of single circuit-oriented network elements, such as switches or terminals can be assumed to have a single input value for the planning tasks of ITU-T Recommendation G.108 [3], this approach is not applicable in packet based systems and thus there is a need for the transmission planner's specific attention.

In particular the decision as to which delay measured according to the present document should is acceptable or representative for the specific configuration is the responsibility of the individual transmission planner.

ITU-T Recommendation G.108 [3] with its amendments provides further guidance on this important issue.

The following optimum terminal parameters from a user's perspective need to be considered:

- minimized delay in send and receive direction;
- optimum loudness Rating (RLR, SLR);
- compensation for network delay variation;
- packet loss recovery performance;
- maximized terminal coupling loss.

4.3 Parameters to be investigated

4.3.1 Basic parameters

The basic parameters are based on I-ETS 300 245-5 [1].

4.3.2 Further parameters with respect to speech processing devices

For VoIP terminals that contain non-linear speech processing devices, the following parameters require additional attention in the context of the present document:

- objective evaluation of speech quality for VoIP terminals;
- doubletalk capability;
- time-variant impairments:
  - switching behaviour;
  - partial echo effects;
  - occurrence of artefacts;
  - clock accuracy;
- background noise performance of the terminal;
- etc.

The measurements of these further parameters with respect to speech processing devices which are a novelty to terminal requirement standards have been successfully used in the ETSI VoIP speech quality test events TR 102 648-1 [i.6].

5 Test equipment

5.1 IP half channel measurement adaptor

The IP half channel measurement adaptor is described in EG 202 425 [i.3].

5.2 Environmental conditions for tests

The following conditions shall apply for the testing environment:

a) ambient temperature: 15 °C to 35 °C (inclusive);

b) relative humidity: 5 % to 85 %;

c) air pressure: 86 kPa to 106 kPa (860 mbar to 1 060 mbar).

5.3 Accuracy of measurements and test signal generation

Unless specified otherwise, the accuracy of measurements made by test equipment shall be equal to or better than:

Table 1: Measurement accuracy

Item	Accuracy
Electrical signal level	±0,2 dB for levels ≥ -50 dBV ±0,4 dB for levels < -50 dBV
Sound pressure	±0,7 dB
Frequency	±0,2 %
Time	±0,2 %
Application force	±2 N

Unless specified otherwise, the accuracy of the signals generated by the test equipment shall be better than:

Table 2: Accuracy of test signal generation

Quantity	Accuracy
Sound pressure level at Mouth Reference Point (MRP)	±3 dB for frequencies from 100 Hz to 200 Hz ±1 dB for frequencies from 200 Hz to 4 000 Hz ±3 dB for frequencies from 4 000 Hz to 8 000 Hz
Electrical excitation levels	±0,4 dB across the whole frequency range.
Frequency generation	±2 % (see note)
Time	±0,2 %
Specified component values	±1 %
NOTE: This tolerance may be used to avoid measurements at critical frequencies, e.g. those due to sampling operations within the terminal under test.	

For terminal equipment which is directly powered from the mains supply, all tests shall be carried out within ±5 % of the rated voltage of that supply. If the equipment is powered by other means and those means are not supplied as part of the apparatus, all tests shall be carried out within the power supply limit declared by the supplier. If the power supply is a.c., the test shall be conducted within ±4 % of the rated frequency.

5.4 Network impairment simulation

At least one set of requirements is based on the assumption of an error free packet network, and at least one other set of requirements is based on a defined simulated malperformance of the packet network.

An appropriate network simulator has to be used, for example NIST Net [i.7] (http://snad.ncsl.nist.gov/itg/nistnet) or Netem [i.8].

Based on the positive experience, STQ have made during the ETSI Speech Quality Test Events with "NIST Net" this will be taken as a basis to express and describe the variations of packet network parameters for the appropriate tests.

Here is a brief blurb about NIST Net:

- The NIST Net network emulator is a general-purpose tool for emulating performance dynamics in IP networks. The tool is designed to allow controlled, reproducible experiments with network performance sensitive/adaptive applications and control protocols in a simple laboratory setting. By operating at the IP level, NIST Net can emulate the critical end-to-end performance characteristics imposed by various wide area network situations (e.g. congestion loss) or by various underlying subnetwork technologies (e.g. asymmetric bandwidth situations of xDSL and cable modems).

- NIST Net is implemented as a kernel module extension to the Linux operating system and an X Window System-based user interface application. In use, the tool allows an inexpensive PC-based router to emulate numerous complex performance scenarios, including: tunable packet delay distributions, congestion and background loss, bandwidth limitation, and packet reordering/duplication. The X interface allows the user to select and monitor specific traffic streams passing through the router and to apply selected performance "effects" to the IP packets of the stream. In addition to the interactive interface, NIST Net can be driven by traces produced from measurements of actual network conditions. NIST Net also provides support for user defined packet handlers to be added to the system. Examples of the use of such packet handlers include: time stamping/data collection, interception and diversion of selected flows, generation of protocol responses from emulated clients.

The key points of Netem can be summarized as follows:

- Netem is nowadays part of most Linux distributions, it only has to be switched on, when compiling a kernel. With Netem, there are the same possibilities as with NIST Net, there can be generated loss, duplication, delay and jitter (and the distribution can be chosen during runtime). Netem can be run on a Linux-PC running as a bridge or a router (NIST Net only runs on routers).

- With an amendment of Netem, Trace Control for Netem (TCN) which was developed by ETH Zurich, it is even possible, to control the behaviour of single packets via a trace file. So it is for example possible to generate a single packet loss, or a specific delay pattern. This amendment is planned to be included in new Linux kernels, nowadays it is available as a patch to a specific kernel and to the iproute2 tool (iproute2 contains Netem).

- It is not advised to define specific distortion patterns for testing in standards, because it will be easy to adapt devices to these patterns (as it is already done for test signals). But if a pattern is unknown to a manufacturer, the same pattern can be used by a test lab for different devices and gives comparable results. It is also possible to take a trace of NIST Net distortions, generate a file out of this and playback exact the same distortions with Netem.

6 Acoustic environment

In general two possible approaches need to be taken into account: EITHER room noise and background noise are an inherent part of the test environment OR room noise and background noise shall be eliminated to such an extent that their influence on the test results can be neglected.

Unless stated otherwise measurements shall be conducted under quiet and "anechoic" conditions. Depending on the distance of the transducers from mouth and ear a quiet office room may be sufficient e.g. for handsets where artificial mouth and artificial ear are located close to the acoustical transducers.

However, for some headsets or handset terminals with smaller dimension an anechoic room will be required.

In cases where real or simulated background noise is used as part of the testing environment, the original background noise must not be noticeably influenced by the acoustical properties of the room.

In all cases where the performance of acoustic echo cancellers shall be tested a realistic room which represents the typical user environment for the terminal shall be used.

7 Requirements and associated measurement methodologies

NOTE 1: In general the test methods as described in the present document apply. If alternative methods exist they may be used if they have been proven to give the same result as the method described in the present document. This will be indicated in the test report.

NOTE 2: Due to the time variant nature of IP connections delay variation may impair the measurements. In such cases the measurement has to be repeated until a valid measurement result is achieved.

7.1 Test setup

The preferred acoustical access to terminals is the most realistic simulation of the "average" subscriber. This can be made by using Head And Torso Simulator (HATS) with appropriate ear simulation and appropriate means to fix handset and headset terminals in a realistic and reproducible way to the HATS. HATS is described in ITU-T Recommendation P.58 [14], appropriate ears are described in ITU-T Recommendation P.57 [13] (type 3.3 and type 3.4 ear), a proper positioning of handsets under realistic conditions is to be found in ITU-T Recommendation P.64 [15].

The preferred way of testing a terminal is to connect it to a network simulator with exact defined settings and access points. The test sequences are fed in either electrically, using a reference codec or using the direct signal processing approach or acoustically using ITU-T specified devices.

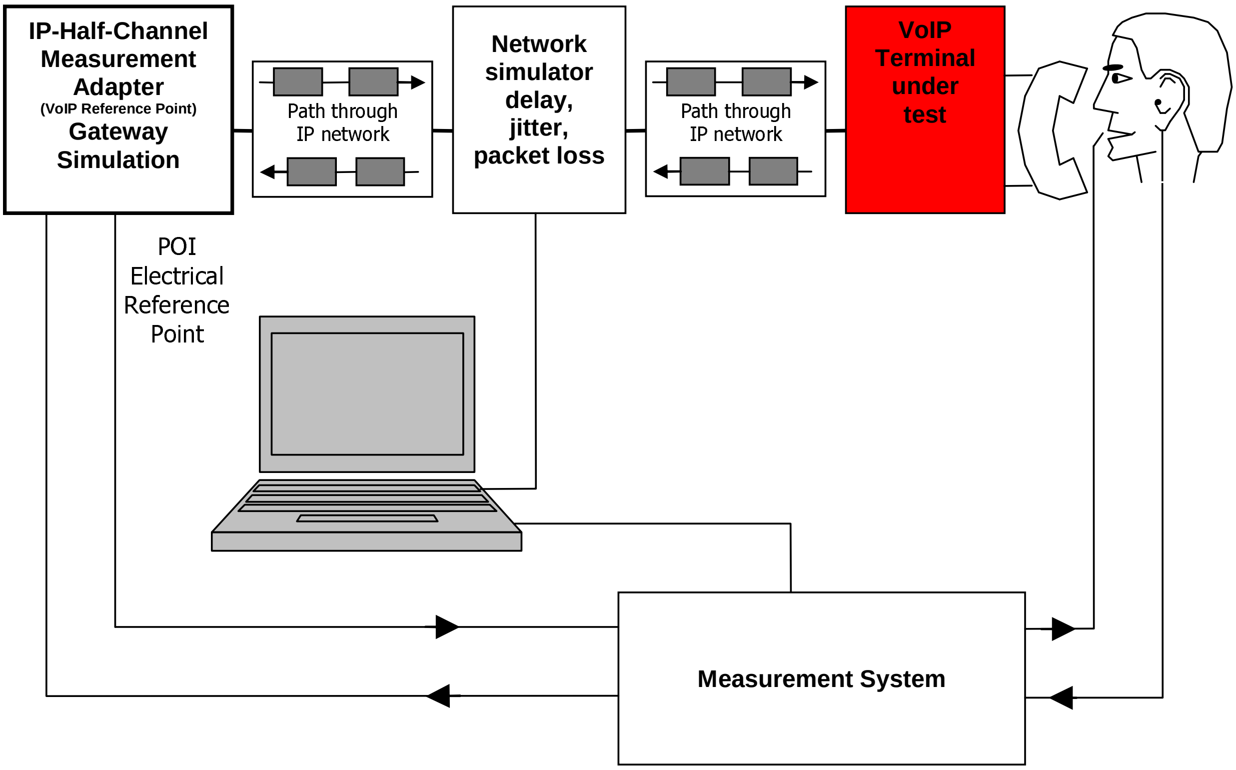


Figure 1: Half channel terminal measurement

7.1.1 Setup for handsets and headsets

When using a handset telephone the handset is placed in the HATS position as described in ITU-T Recommendation P.64 [15]. The artificial mouth shall be conform with ITU-T Recommendation P.58 [14]. The artificial ear shall be conform with ITU-T Recommendation P.57 [13], type 3.3 or type 3.4 ears shall be used.

Recommendations for positioning headsets are given in ITU-T Recommendation P.380 [18]. If not stated otherwise headsets shall be placed in their recommended wearing position. Further information about setup and the use of HATS can be found in ITU-T Recommendation P.380 [18].

Unless stated otherwise if a volume control is provided the setting is chosen such that the nominal RLR is met as close as possible.

Unless stated otherwise the application force of 8N is used for handset testing. No application force is used for headsets.

7.1.2 Position and calibration of HATS

All the send and receive characteristics shall be tested with the HATS, it shall be indicated what type of ear was used at what application force. For handsets if not stated otherwise 8 N application force shall be used.

The horizontal positioning of the HATS reference plane shall be guaranteed within ±2º.

The HATS shall be equipped with two type 3.3 or type 3.4 artificial ears. For binaural headsets two artificial ears are required. The type 3.3 or type 3.4 artificial ears as specified in ITU-T Recommendation P.57 [13] shall be used. The artificial ear shall be positioned on HATS according to ITU-T Recommendation P.58 [14].

The exact calibration and equalization can be found in ITU-T Recommendation P.581 [21]. If not stated otherwise, the HATS shall be diffuse-field equalized. The reverse nominal diffuse field curve as found in table 3 of ITU-T Recommendation P.58 [14] shall be used.

7.1.3 Test signal levels

Unless specified otherwise, the test signal level shall be -4,7 dBPa at the MRP.

Unless specified otherwise, the applied test signal level at the digital input shall be -16 dBm0.

7.1.4 Setup of background noise simulation

A setup for simulating realistic background noises in a lab-type environment is described in EG 202 396-1 [i.2].

EG 202 396-1 [i.2] contains a description of the recording arrangement for realistic background noises, a description of the setup for a loudspeaker arrangement suitable to simulate a background noise field in a lab-type environment and a database of realistic background noises, which can be used for testing the terminal performance with a variety of different background noises.

The principle loudspeaker setup for the simulation arrangement is shown in figure 2.

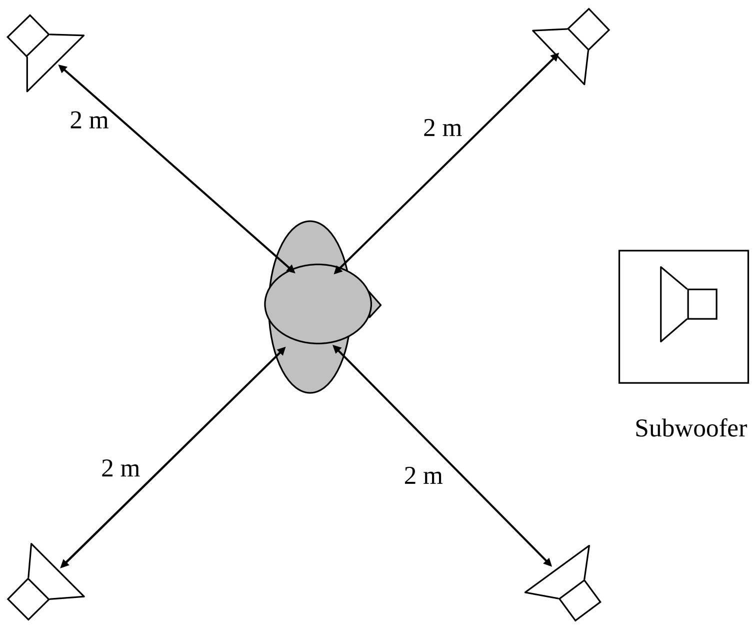


Figure 2: Loudspeaker arrangement for background noise simulation

The equalization and calibration procedure for the setup is described in detail in EG 202 396-1 [i.2].

If not stated otherwise this setup is used in all measurements where background noise simulation is required.

The following noises of EG 202 396-1 [i.2] shall be used:

Recording in pub	Pub_Noise_binaural	30 s	L: 77,8 dB(A) R: 78,9 dB(A)	binaural
Recording at sales counter	Cafeteria_Noise_binaural	30 s	L: 68,4 dB(A) R: 67,3 dB(A)	Binaural
Recording in business office	Work_Noise_Office_Callcenter_binaural	30 s	L: 56,6 dB(A) R: 57,8 dB(A)	Binaural

7.2 Coding independent parameters

7.2.1 Send Frequency response

Requirement

The send frequency response of the handset or the headset shall be within a mask as defined in table 3 and shown in figure 3. This mask shall be applicable for all types of handsets and headsets.

Table 3

Frequency	Upper Limit	Lower Limit
100 Hz	0 dB	
200 Hz	5 dB	-5 dB
5 000 Hz	5 dB	-5 dB
6 300 Hz	5 dB	-10 dB
8 000 Hz	5 dB	
NOTE: The limits for intermediate frequencies lie on a straight line drawn between the given values on a linear (dB) – logarithmic (Hz) scale.		

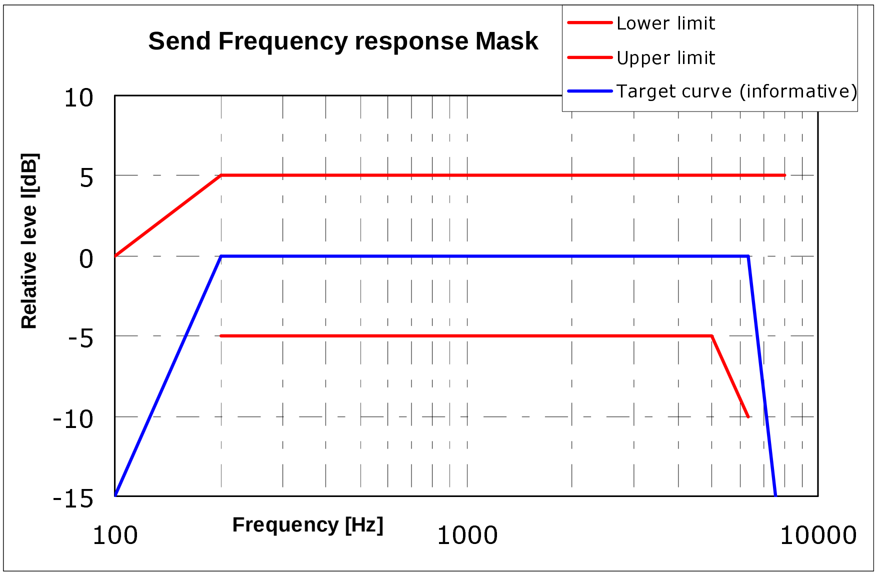


Figure 3: Send frequency response mask

NOTE 1: The basis for the target frequency responses in send and receive is the orthotelefonic reference response which is measured between 2 subjects in 1 m distance under free field conditions and is assuming an ideal receive characteristic. Under these conditions the overall frequency response shows a rising slope. In opposite to other standards the present document no longer uses the ERP as the reference point for receive but the diffusefield. With the concept of diffusefield based receive measurements, a rising slope for the overall frequency response is achieved by a flat target frequency response in send and a diffusefield based receive frequency response.

NOTE 2: A "balanced" frequency response is preferable from the perception point of view. If frequency components in the low frequency domain are attenuated in a similar way frequency components in the high frequency domain should be attenuated.

Measurement method

The test signal to be used for the measurements shall be the artificial voice according to ITU-T Recommendation P.50 [11]. If the signal to noise ratio in the high frequency domain is not sufficient Composite Source Signal (CSS) as defined in ITU-T Recommendation P.501 [19] shall be used. The spectrum of acoustic signal produced by the artificial mouth is calibrated under free field conditions at the MRP. The test signal level shall be -4,7 dBPa, duration 20 s (10 s female, 10 s male voice), measured at the MRP. The test signal level is averaged over the complete test signal sequence.

The handset terminal is setup as described in clause 7.1. The handset is mounted in the HATS position (see ITU-T Recommendation P.64 [15]). The application force used to apply the handset against the artificial ear is noted in the test report.

In case of headset measurements the tests are repeated 5 times, in conformance with ITU-T Recommendation P.380 [18]. The results are averaged (averaged value in dB, for each frequency).

Measurements shall be made at one twelfth-octave intervals as given by the R.40 series of preferred numbers in ISO 3 [24] for frequencies from 100 Hz to 8 kHz inclusive. For the calculation the averaged measured level at the electrical reference point for each frequency band is referred to the averaged test signal level measured in each frequency band at the MRP.

The sensitivity is expressed in terms of dBV/Pa.

7.2.2 Send Loudness Rating (SLR)

Requirement

The nominal value of Send Loudness Rating (SLR) shall be:

- SLR(set) = 8 dB ± 3 dB

Measurement method

The test signal to be used for the measurements shall be the artificial voice according to ITU-T Recommendation P.50 [11], duration 20 s (10 s female, 10 s male voice). If the signal to noise ratio in the high frequency domain is not sufficient CSS as defined in ITU-T Recommendation P.501 [19] shall be used. The spectrum of acoustic signal produced by the artificial mouth is calibrated under free field conditions at the MRP. The test signal level shall be -4,7 dBPa, measured at the MRP. The test signal level is averaged over the complete test signal sequence.

The handset or headset terminal is setup as described in clause 7.1. The handset is mounted in the HATS position (see ITU-T Recommendation P.64 [15]). The application force used to apply the handset against the artificial ear is noted in the test report.

In case of headset measurements the tests are repeated 5 times, in conformance with ITU-T Recommendation P.380 [18]. The results are averaged (averaged value in dB, for each frequency).

The send sensitivity shall be calculated from each band of the 20 frequencies given in table 1 of ITU-T Recommendation P.79 [16], bands 1 to 20. For the calculation the averaged measured level at the electrical reference point for each frequency band is referred to the averaged test signal level measured in each frequency band at the MRP.

The sensitivity is expressed in terms of dBV/Pa and the SLR shall be calculated according to ITU-T Recommendation P.79 [16], see annex A.

7.2.3 D-Factor

Requirement

For VoIP terminals the D-factor shall be:

- D-factor (DelSM) ≥ 2 dB

NOTE: Wideband calculation is for further study, provisionally the measurement is based on narrowband.

Measurement method

The background noise simulation as described in clause 7.1 is used.

Handset or headset terminals are mounted as described in clause 7.1. Measurements are made on one-third octave bands according to IEC 61260 [23] for the 14 bands centred at 200 Hz to 4 kHz (bands 4 to 17). For each band the diffuse sound sensitivity Ssi(diff) is measured. The sensitivity shall be expressed in terms of dBV/Pa.

The direct sound field sensitivity Ssi(direct) is measured as described in clause 7.2.2 (SLR).

The D value according to ITU-T Recommendation P.79 [16], annex E, formulas E2 and E3 is calculated in bands 4 to 17. The coefficients Ki as described in table E1 are used.

The direct sound sensitivity shall be measured using the test set-up specified in clause 7.1 and a speech like test signal as defined in ITU-T Recommendation P.50 [11] or P.501 [19]. The type of test signal used shall be stated in the test report. The direct sound sensitivity is measured in one-third octave bands according to IEC 61260 [23] for the 14 bands centred at 200 Hz to 4 kHz (bands 4 to 17). For each band the direct sound sensitivity Ssi(direct) is measured. The sensitivity shall be expressed in terms of dBV/Pa.

The value of the D-factor shall be calculated according to ITU-T Recommendation P.79 [16], annex E, formulas E2 and E3, over the bands from 4 to 17, using the coefficients Ki from table E1 of ITU-T Recommendation P.79 [16].

7.2.4 Linearity range for SLR

Requirement

The sensitivity determined with input sound pressure levels between -24,7 dBPa and 5,3 dBPa shall not differ by more than ±2 dB from the sensitivity determined with an input sound pressure level of -4,7 dBPa. For the input sound pressure level of 5,3 dBPa a limit of +4 dB to -2 dB applies.

Table 4

Linearity range of SLR: ΔSLR = SLR - SLR@-4,7 dBPa			
Input Level	**Target ΔSLR**	**Upper limit**	**Lower limit**
-24,7 dBPa	0	2 dB	-2 dB
-19,7 dBPa	0	2 dB	-2 dB
-14,7 dBPa	0	2 dB	-2 dB
-9,7 dBPa	0	2 dB	-2 dB
-4,9 dBPa	0	2 dB	-2 dB
-4,7 dBPa	0	0 dB	0 dB
-4,5 dBPa	0	2 dB	-2 dB
0,3 dBPa	0	2 dB	-2 dB
5,3 dBPa	0	4 dB	-4 dB
NOTE: It is assumed that the variation of gain is mostly codec independent. In case codec specific requirements are needed, they are found in clause 7.3.			

Measurement method

The test signal to be used for the measurements shall be the artificial voice according to ITU-T Recommendation P.50 [11]. If the signal to noise ratio in the high frequency domain is not sufficient CSS as defined in ITU-T Recommendation P.501 [19] shall be used. The spectrum of acoustic signal produced by the artificial mouth is calibrated under free field conditions at the MRP. The test signal levels shall be -24,7 dBPa up to 5,3 dBPa in steps of 5 dB, measured at the MRP. The test signal level is averaged over the complete test signal sequence.

The handset terminal is setup as described in clause 7.1. The handset is mounted in the HATS position (see ITU-T Recommendation P.64 [15]). The application force used to apply the handset against the artificial ear is noted in the test report.

The send sensitivity shall be calculated from each band of the 20 frequencies given in table 1 of ITU-T Recommendation P.79 [16], bands 1 to 20. For the calculation the averaged measured level at the electrical reference point for each frequency band is referred to the averaged test signal level measured in each frequency band at the MRP.

The sensitivity is expressed in terms of dBV/Pa and the SLR shall be calculated according to ITU-T Recommendation P.79 [16], annex A.

7.2.5 Send distortion

Requirement

The terminal will be positioned as described in clause 7.1.

The ratio of signal to harmonic distortion shall be above the following mask:

Table 5

Frequency	Ratio
315 Hz	26 dB
400 Hz	30 dB
1 kHz	30 dB
2 kHz	30 dB
NOTE: Limits at intermediate frequencies lie on a straight line drawn between the given values on a linear (dB ratio) - logarithmic (frequency) scale.	

Measurement method

The terminal will be positioned as described in clause 7.1.

The signal used is an activation signal followed by a sine wave signal with a frequency at 315 Hz, 400 Hz, 500 Hz, 630 Hz, 800 Hz, 1 000 Hz and 2 000 Hz. The duration of the sine wave shall be less than 1 s. The sinusoidal signal level shall be calibrated to -4,7 dBPa at the MRP.

The signal to harmonic distortion ratio is measured selectively up to 6,3 kHz.

An artificial voice according to ITU-Recommendation P.50 [11] or a speech like test signal as described in ITU-T Recommendation P.501 [19] can be used for activation. Level of this activation signal will be -4,7 dBPa at the MRP.

NOTE: Depending on the type of codec the test signal used may need to be adapted.

7.2.6 Send noise

Requirement

The maximum noise level produced by the VoIP terminal at the POI under silent conditions in the send direction shall not exceed -68 dBm0 (A).

No peaks in the frequency domain higher than 10 dB above the average noise spectrum shall occur.

Measurement method

For the actual measurement no test signal is used. In order to reliably activate the terminal an activation signal is introduced before the actual measurement. The activation signal shall be a sequence of 4 composite source signals (CSS) as described in ITU-T Recommendation P.501 [19]. The spectrum of the acoustic signal produced by the artificial mouth is calibrated under free field conditions at the MRP. The activation signal level shall be -4,7 dBPa, measured at the MRP. The activation signal level is averaged over the complete activation signal sequence. Alternatively other speech like test signals (e.g. artificial voice) with the same signal level can be used for activation.

The handset terminal is set-up as described in clause 7.1. The handset is mounted at the HATS position (see ITU-T Recommendation P.64 [15]).

The send noise is measured at the POI in the frequency range from 100 Hz to 8 kHz. The analysis window is applied directly after stopping the activation signal but taking into account the influence of all acoustical components (reverberations). The averaging time is 1 s. The test house has to ensure (e.g. by monitoring the time signal) that during the test the terminal remains in activated condition. If the terminal is deactivated during the measurement, the measurement time has to be reduced to the period where the terminal remains in activated condition.

The noise level is measured in dBm0(C).

7.2.7 SideTone Masking Rating STMR (mouth to ear)

Requirement

The STMR shall be 16 dB ± 4 dB for nominal setting of the volume control.

For all other positions of the volume control, the STMR must not be below 8 dB.

NOTE: It is preferable to have a constant STMR independent of the volume control setting.

Measurement method

The test signal to be used for the measurements shall be the artificial voice according to ITU-T Recommendation P.50 [11]. The spectrum of the acoustic signal produced by the artificial mouth is calibrated under free field conditions at the MRP. The test signal level shall be -4,7 dBPa, measured at the MRP. The test signal level is averaged over the complete test signal sequence.

The handset or the headset terminal is setup as described in clause 7.1. The handset is mounted in the HATS position (see ITU-T Recommendation P.64 [15]) and the application force shall be 13 N on the artificial ear type 3.3 or type 3.4.

Where a user operated volume control is provided, the measurements shall be carried out the nominal setting of the volume control. In addition the measurement is repeated at the maximum volume control setting.

Measurements shall be made at one twelfth-octave intervals as given by the R.40 series of preferred numbers in ISO 3 [24] for frequencies from 100 Hz to 8 kHz inclusive. For the calculation the averaged measured level at each frequency band (ITU-T Recommendation P.79 [16], table 3, bands 1 to 20) is referred to the averaged test signal level measured in each frequency band.

The Sidetone path loss (LmeST), as expressed in dB, and the SideTone Masking Rate (STMR) (in dB) shall be calculated from the formula 5-1 of ITU-T Recommendation P.79 [16], using m = 0,225 and the weighting factors from table 3 of ITU-T Recommendation P.79 [16].

7.2.8 Sidetone delay

Requirement

The maximum sidetone-round-trip delay shall be ≤ 5 ms, measured in an echo-free setup.

Measurement method

The handset or the headset terminal is setup as described in clause 7.1. The handset is mounted in the HATS position (see ITU-T Recommendation P.64 [15]).

The test signal is a CS-signal complying with ITU-T Recommendation P.501 [19] using a pn sequence with a length of 4 096 points (for the 48 kHz sampling rate) which equals to the period T. The duration of the complete test signal is as specified in ITU-T Recommendation P.501 [19]. The level of the signal shall be -4,7 dBPa at the MRP.

The cross-correlation function $\Phi xy(\tau)$ between the input signal $S_x(t)$ generated by the test system in send direction and the output signal $S_y(t)$ measured at the artificial ear is calculated in the time domain:

$$\Phi_{xy}(\tau) = \lim_{T \to \infty} \sum_{t=-T/2}^{T/2} S_x(t) S_y(t+\tau) \qquad (1)$$

The measurement window T shall be exactly identical with the time period T of the test signal, the measurement window is positioned to the pn-sequence of the test signal.

The sidetone delay is calculated from the envelope E(τ) of the cross-correlation function Φxy(τ). The first maximum of the envelope function occurs in correspondence with the direct sound produced by the artificial mouth, the second one occurs with a possible delayed sidetone signal. The difference between the two maxima corresponds to the sidetone delay. The envelope E(τ) is calculated by the Hilbert transformation H {xy(τ)} of the cross-correlation:

$$H\{xy(\tau)\} = \sum_{-\infty}^{\infty} \frac{\Phi_{xy}(u)}{\Pi(\tau - u)} \quad (2)$$

$$E(\tau) = \sqrt{[\phi xy(\tau)]^2 + \{H[\phi xy(\tau)]\}^2} \quad (3)$$

It is assumed that the measured sidetone delay is less than T/2.

7.2.9 Terminal Coupling Loss weighted (TCLw)

Requirement

The TCLw shall be ≥ 55 dB.

With the volume control set to maximum TCLw shall be ≥ 46 dB. The volume control shall be set back to nominal after each call unless TCLw ≥ 55 dB can be maintained also with maximum volume setting.

Measurement method

The handset or headset terminal is setup as described in clause 7.1. The handset is mounted in the HATS position (see ITU-T Recommendation P.64 [15]) and the application force shall be 2 N on the artificial ear type 3.3 or type 3.4 as specified in ITU-T Recommendation P.57 [13]. The ambient noise level shall be less than -64 dBPa(A) for handset and headset terminals. The attenuation from electrical reference point input to electrical reference point output shall be measured using a speech like test signal.

Before the actual test a training sequence consisting of 10 s male artificial voice followed by 10 s female artificial voice according to ITU-T Recommendation P.50 [11] is applied. The training sequence level shall be -16 dBm0 in order not to overload the codec.

The test signal following immediately the training sequence is a PN-sequence complying with ITU-T Recommendation P.501 [19] with a length of 4 096 points (for the 48 kHz sampling rate) and a crest factor of 6 dB. The length of the complete test signal composed of at least four sequences of CSS shall be at least one second (1,0 s). The test signal level is -3 dBm0 (from 50 Hz to 7 kHz). The low crest factor is achieved by random alternation of the phase between -180° and 180°.

The TCLw is calculated according to ITU-T Recommendation G.122 [5], clause B.4 (trapezoidal rule) but using the frequency range of 300 Hz to 6 700 Hz (instead of 300 Hz to 3 400 Hz). For the calculation the averaged measured echo level at each frequency band is referred to the averaged test signal level measured in each frequency band. For the measurement a time window has to be applied adapted to the duration of the actual pn-sequence of the test signal (200 ms) choosing the pn-sequence of the third CS-Signal.

NOTE: The extension of the frequency range is for further study.

7.2.10 Stability loss

Requirement

With the handset lying on and the transducers facing a hard surface, the attenuation from the digital input to the digital output shall be at least 6 dB at all frequencies in the range of 100 Hz to 8 kHz. In case of headsets the requirement applies for the closest possible position between microphone and headset receiver.

NOTE: Depending on the type of headset it may be necessary to repeat the measurement in different positions.

Measurement method

Before the actual test a training sequence consisting of 10 s male artificial voice followed by 10 s female artificial voice according to ITU-T Recommendation P.50 [11] is applied. The training sequence level shall be -16 dBm0 in order not to overload the codec.

The test signal is a PN sequence complying with ITU-T Recommendation P.501 [19] with a length of 4 096 points (for the 48 kHz sampling rate) and a crest factor of 6 dB. The duration of the test signal is 250 ms. With an input signal of -3 dBm0, the attenuation from digital input to digital output shall be measured for frequencies from 100 Hz to 8 kHz under the following conditions:

a) the handset or the headset, with the transmission circuit fully active, shall be positioned on one inside surface that is of three perpendicular plane, smooth, hard surfaces forming a corner. Each surface shall extend 0,5 m from the apex of the corner. One surface shall be marked with a diagonal line, extending from the corner formed by the three surfaces, and a reference position 250 mm from the corner, as shown in figure 4;

b1) the handset, with the transmission circuit fully active, shall be positioned on the defined surface as follows:

1) the mouthpiece and ear cup shall face towards the surface;

2) the handset shall be placed centrally, the diagonal line with the ear cup nearer to the apex of the corner;

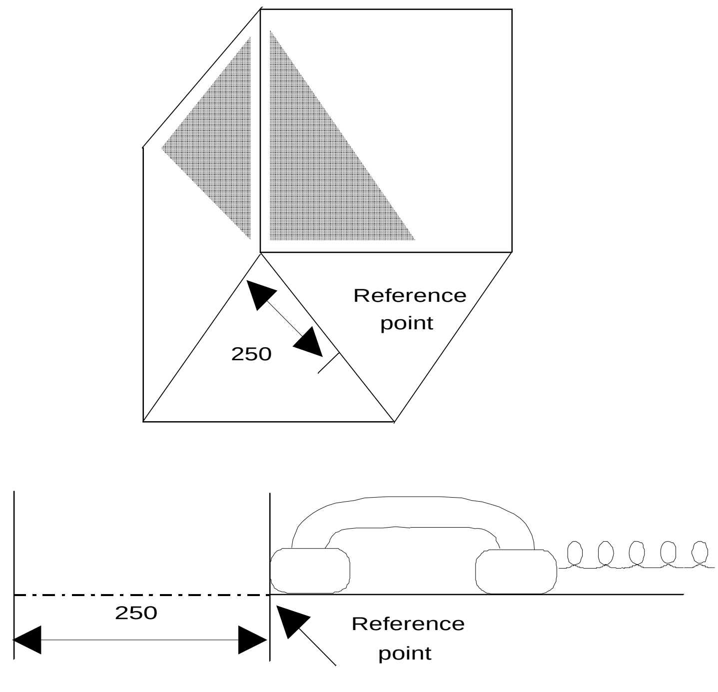
3) the extremity of the handset shall coincide with the normal to the reference point, as shown in figure 4;

b2) the headset, with the transmission circuit fully active, shall be positioned on the defined surface as follows:

1) the microphone and the receiver shall face towards the surface;

2) the headset receiver shall be placed centrally at the reference point as shown in figure 4;

3) the headset microphone is positioned as close as possible to the receiver.



NOTE: All dimensions in mm.

Figure 4

7.2.11 Receive frequency response

Requirement

The receive frequency response of the handset or the headset shall be within a mask as defined in table 6 and shown in figure 5. The application force for handsets is 2 N, 8 N and 13 N. This mask defined for 8 N application force shall be applicable for all types of headsets.

Table 6: Receive frequency response mask

Frequency	Upper limit 8 N	Lower limit 8 N	Upper limit 2 N	Lower limit 2 N	Upper limit 13 N	Lower limit 13 N
100 Hz	3 dB		3 dB		6 dB	
120 Hz	3 dB	-5 dB	3 dB	-10 dB	6 dB	-5 dB
200 Hz	3 dB	-5 dB	3 dB	-8 dB	6 dB	-5 dB
400 Hz	3 dB	-5 dB	3 dB	-8 dB	6 dB	-5 dB
1 010 Hz	See note 1	-5 dB	See note 1	-8 dB	6 dB	-5 dB
1 200 Hz	See note 1	-8 dB	See note 1	-8 dB	6 dB	-8 dB
1 500 Hz	See note 1	-8 dB	See note 1	-8 dB	See note 1	-8 dB
2 000 Hz	9 dB	-3 dB	9 dB	-3 dB	9 dB	-3 dB
3 200 Hz	9 dB	-3 dB	9 dB	-3 dB	9 dB	-3 dB
7 000 Hz	9 dB	-13 dB	9 dB	-13 dB	9 dB	-13 dB
8 000 Hz	9 dB		9 dB		9 dB	

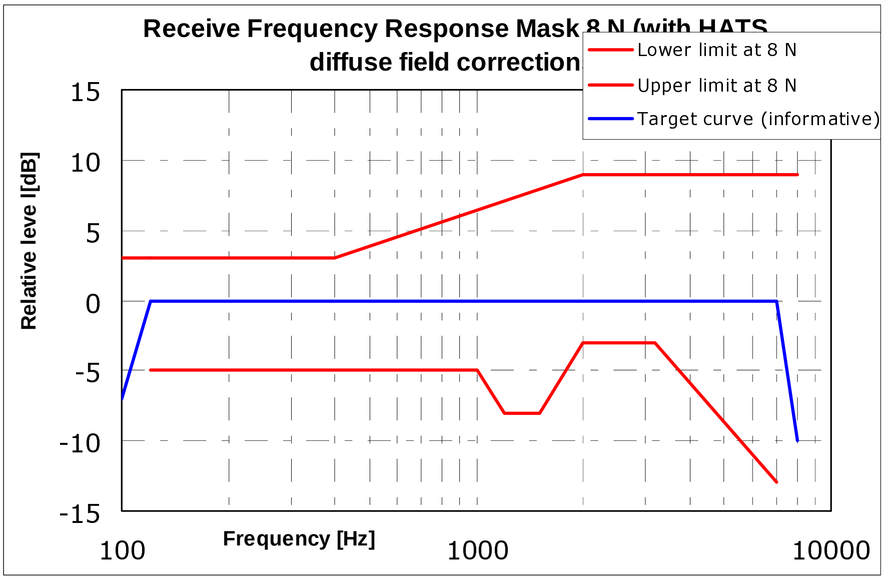
NOTE 1: The limit curves shall be determined by straight lines joining successive co-ordinates given in the table, where frequency response is plotted on a linear dB scale against frequency on a logarithmic scale. The mask is a floating or "best fit" mask.

NOTE 2: The basis for the target frequency responses in send and receive is the orthotelefonic reference response which is measured between 2 subjects in 1 m distance under free field conditions and is assuming an ideal receive characteristic. This flat response characteristics is shown as the target curve. Under these conditions the overall frequency response shows a rising slope. In opposite to other standards the present document no longer uses the ERP as the reference point for receive but the diffuse field. With the concept of diffuse field based receive measurements a rising slope for the overall frequency response is achieved by a flat target frequency response in send and a diffusefield based receive frequency response.

NOTE 3: With current technology it may be difficult or even not possible to achieve the desired frequency response characteristics for handsets with 2 N application force.

NOTE 4: With current technology it may be difficult or even not possible to achieve the desired frequency response characteristics for headsets below 250 Hz.

NOTE 5: The basis for the frequency response mask requirements is a subjective experiment which is described in Annex C. It may be difficult to be compliant with both this frequency response mask and the current frequency response mask as defined in TIA-920 [25].



NOTE: A "balanced" frequency response is preferable from the perception point of view. If frequency components in the low frequency domain are attenuated in a similar way frequency components in the high frequency domain should be attenuated.

Figure 5: Receive frequency response mask for 8 N application force

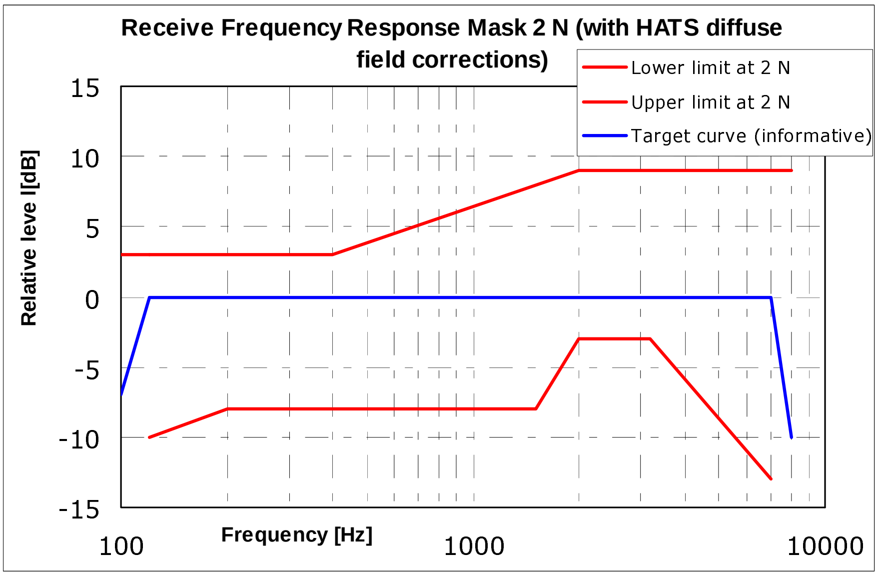


Figure 5a: Receive frequency response mask for 2 N application force

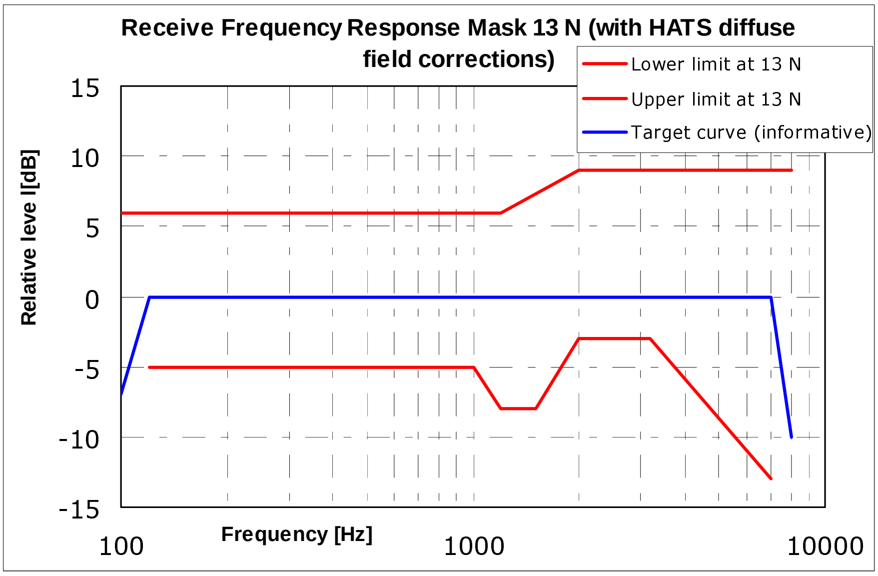


Figure 5b: Receive frequency response mask for 13 N application force

Measurement method

Receive frequency response is the ratio of the measured sound pressure and the input level (dB relative Pa/V).

$$S_{Jeff} = 20 \log (pe_{ff} / v_{RCV}) \text{ dB rel 1 Pa / V} \quad (4)$$

S_{Jeff} Receive Sensitivity; Junction to HATS Ear with diffuse field correction.

pe_{ff} DRP Sound pressure measured by ear simulator Measurement data are converted from the Drum Reference Point to diffusefield.

v_{RCV} Equivalent RMS input voltage.

The test signal to be used for the measurements shall be the artificial voice according to ITU-T Recommendation P.50 [11], duration 20 s (10 s female, 10 s male voice). If the signal to noise ratio in the high frequency domain is not sufficient CSS as defined in ITU-T Recommendation P.501 [19] shall be used. The test signal level shall be -16 dBm0, measured according to ITU-T Recommendation P.56 [12] at the digital reference point or the equivalent analogue point.

The handset terminal or the headset terminal is setup as described in clause 7.1. The handset is mounted in the HATS position (see ITU-T Recommendation P.64 [15]). The application forces used to apply the handset against the artificial ear is 2 N, 8 N and 13 N.

In case of headset measurements the tests are repeated 5 times, in conformance with ITU-T Recommendation P.380 [18]. The results are averaged (averaged value in dB, for each frequency).

The HATS is diffuse field equalized as described in ITU-T Recommendation P.581 [21]. The diffuse field correction as defined in ITU-T Recommendation P.58 [14] is applied. The equalized output signal is power-averaged on the total time of analysis. The 1/12 octave band data are considered as the input signal to be used for calculations or measurements.

Measurements shall be made at one twelfth-octave intervals as given by the R.40 series of preferred numbers in ISO 3 [24] for frequencies from 100 Hz to 8 kHz inclusive. For the calculation the averaged measured level at each frequency band is referred to the averaged test signal level measured in each frequency band.

The sensitivity is expressed in terms of dBPa/V.

7.2.12 Receive Loudness Rating (RLR)

Requirement

The nominal value of Receive Loudness Rating (RLR) shall be:

- RLR (set) = 2 dB ± 3 dB
- RLR (binaural headset) = 8 dB ± 3 dB for each earphone

Measurement method

The test signal to be used for the measurements shall be the artificial voice according to ITU-T Recommendation P.50 [11], duration 20 s (10 s female, 10 s male voice). If the signal to noise ratio in the high frequency domain is not sufficient CSS as defined in ITU-T Recommendation P.501 [19] shall be used. The test signal level shall be -16 dBm0, measured at the digital reference point or the equivalent analogue point. The test signal level is averaged over the complete test signal sequence.

The handset terminal or the headset terminal is setup as described in clause 7.1. The handset is mounted in the HATS position (see ITU-T Recommendation P.64 [15]). The application force used to apply the handset against the artificial ear is noted in the test report. The HATS is *NOT* diffuse field equalized as described in ITU-T Recommendation P.581 [21]. The DRP-ERP correction as defined in ITU-T Recommendation P.57 [13] is applied. The application force used to apply the handset against the artificial ear is noted in the test report. By default, 8 N will be used.

In case of headset measurements the tests are repeated 5 times, in conformance with ITU-T Recommendation P.380 [18]. The results are averaged (averaged value in dB, for each frequency).

The receive sensitivity shall be calculated from each band of the 20 frequencies given in table 1 of ITU-T Recommendation P.79 [16], bands 1 to 20. For the calculation the averaged measured level at each frequency band is referred to the averaged test signal level measured in each frequency band.

The sensitivity is expressed in terms of dBPa/V and the RLR shall be calculated according to ITU-T Recommendation P.79 [16], annex A. No leakage correction shall be applied for the measurement.

7.2.13 Receive distortion

Requirement

The ratio of signal to harmonic distortion shall be above the following mask:

Table 7

Frequency	Signal to distortion ratio limit, receive
315 Hz	26 dB
400 Hz	30 dB
500 Hz	30 dB
800 Hz	30 dB
1 kHz	30 dB
2 kHz	30 dB
NOTE: Limits at intermediate frequencies lie on a straight line drawn between the given values on a linear (dB ratio) - logarithmic (frequency) scale.	

Measurement method

The handset terminal or the headset terminal is positioned as described in clause 7.1.

The signal used is an activation signal followed by a sine wave signal with a frequency at 315 Hz, 400 Hz, 500 Hz, 630 Hz, 800 Hz, 1 000 Hz and 2 000 Hz.

An artificial voice according to ITU-Recommendation P.50 [11] or a speech like test signal as described in ITU-T Recommendation P.501 [19] can be used for activation.

The signal level shall be -16 dBm0.

Measurement are made at 315 Hz, 400 Hz, 500 Hz, 630 Hz, 800 Hz, 1 000 Hz and 2 000 Hz.

The signal to harmonic distortion ratio is measured selectively up to 6,3 kHz.

The ratio of signal to harmonic distortion shall be measured at the DRP of the artificial ear with the diffusefield equalization active.

NOTE: Depending on the type of codec the test signal used may need to be adapted.

7.2.14 Minimum activation level and sensitivity in receive direction

For further study.

7.2.15 Receive noise

Requirement

Telephone sets with adjustable receive levels shall be adjusted so that the RLR is as close as possible to the nominal RLR.

The receive noise shall be less than -57 dBPa(A).

Where a volume control is provided, the measured noise shall not be greater than -54 dBPa(A) at the maximum setting of the volume control.

Measurement method

The handset terminal or the headset terminal is setup as described in clause 7.1.

An artificial voice according to ITU-Recommendation P.50 [11] or a speech like test signal as described in ITU-T Recommendation P.501 [19] can be used for activation. The activation signal level shall be -16 dBm0.

The A-weighted noise level shall be measured at DRP of the artificial ear with the diffusefield equalization active.

7.2.16 Automatic gain control in receive

For further study.

7.2.17 Double talk Performance

During double talk the speech is mainly determined by 2 parameters: impairment caused by echo during double talk and level variation between single and double talk (attenuation range).

In order to guarantee sufficient quality under double talk conditions the talker Echo Loudness Rating (ELR) should be high and the attenuation inserted should be as low as possible. Terminals which do not allow double talk in any case should provide a good echo attenuation which is realized by a high attenuation range in this case.

The most important parameters determining the speech quality during double talk are (see ITU-T Recommendations P.340 [17] and P.502 [20]):

- attenuation range in send direction during double talk $A_{H,S,dt}$;
- attenuation range in receive direction during double talk $A_{H,R,dt}$;
- echo attenuation during double talk.

7.2.17.1 Attenuation range in send direction during double talk $A_{H,S,dt}$

Requirement

Based on the level variation in send direction during double talk $A_{H,S,dt}$ the behaviour of the terminal can be classified according to table 8.

Table 8

Category (according to ITU-T Rec. P.340 [17])	**1**	**2a**	**2b**	**2c**	**3**
	Full Duplex Capability	Partial Duplex Capability			No Duplex Capability
$A_{H,S,dt}$ **[dB]**	≤ 3	≤ 6	≤ 9	≤ 12	> 12

In general table 8 provides a quality classification of terminals regarding double talk performance. However, this does not mean that a terminal which is category 1 based on the double talk performance is of high quality concerning the overall quality as well.

Measurement method

The test signal to determine the attenuation range during double talk is shown in figure 6. A sequence of uncorrelated CS signals is used which is inserted in parallel in send and receive direction.

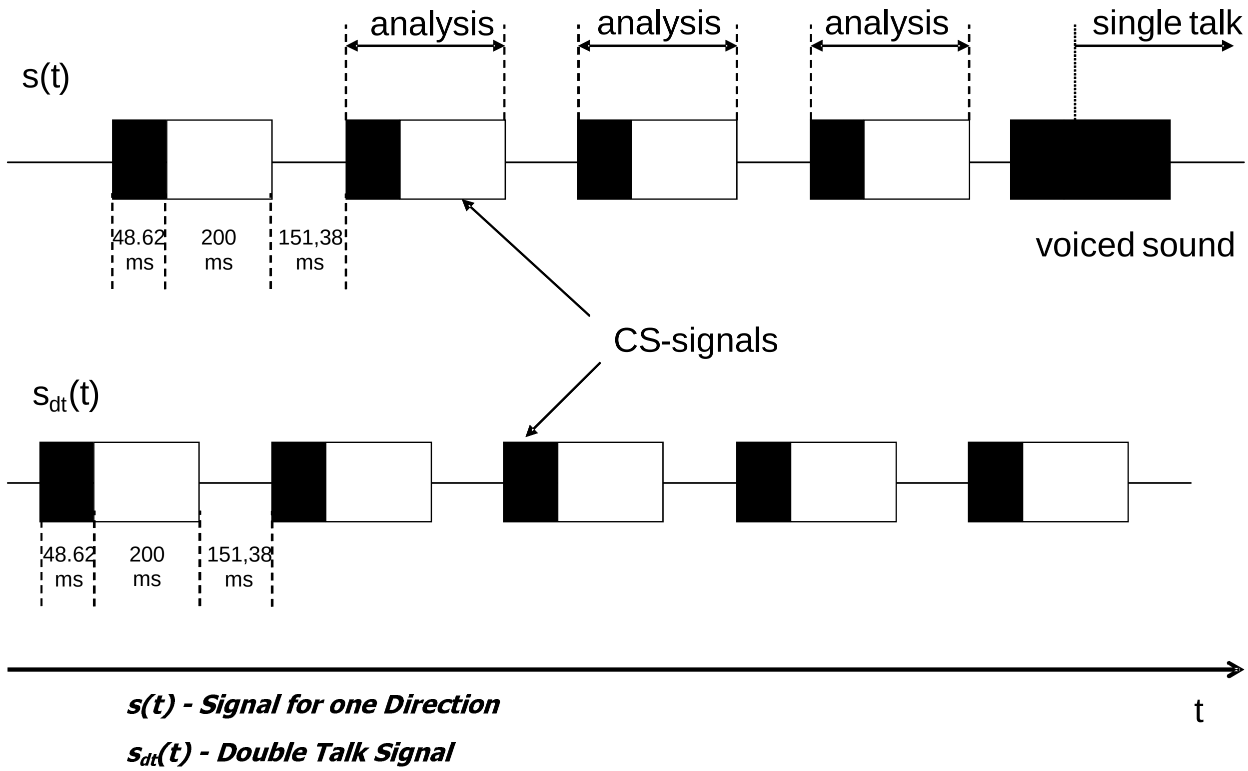


Figure 6: Double talk test sequence with overlapping CS signals in send and receive direction

Figure 6 indicates that the sequences overlap partially. The beginning of the CS sequence (voiced sound, black) is overlapped by the end of the pn-sequence (white) of the opposite direction. During the active signal parts of one signal the analysis can be conducted in send and receive direction. The analysis times are shown in figure 6 as well. The test signals are synchronized in time at the acoustical interface. The delay of the test arrangement should be constant during the measurement.

NOTE: The length of voiced sound of the double talk signal is achieved by repeating one period of the voiced sound for double talk according to ITU-T Recommendation P.501 [19] 10 times and cutting off the initial 3,3 ms of the period of the first voiced sound.

The settings for the test signals are as follows:

Table 9

	Receive direction (sdt(t))	Send direction (s(t))
Pause length between two signal bursts	151,38 ms	151,38 ms
Average signal level (assuming an original pause length of 101,38 ms)	-16 dBm0	-4,7 dBPa
Active signal parts	-14,7 dBm0	-3 dBPa

The test arrangement is according to clause 7.

When determining the attenuation range in send direction the signal measured at the electrical reference point is referred to the test signal inserted.

The level is determined as level versus time from the time domain. The integration time of the level analysis is 5 ms. The attenuation is determined from the level difference measured at the beginning of the double talk always with the beginning of the CS-signal in send direction until its complete activation (during the pause in the receive channel). The analysis is performed over the complete signal starting with the second CS-signal. The first CS-signal is not used for the analysis.

7.2.17.2 Attenuation range in receive direction during double talk $A_{H,R,dt}$

Requirement

Based on the level variation in receive direction during double talk $A_{H,R,dt}$ the behaviour of the terminal can be classified according to table 10.

Table 10

Category (according to ITU-T Rec. P.340 [17])	**1**	**2a**	**2b**	**2c**	**3**
	Full duplex capability	Partial duplex capability			Full duplex capability
$A_{H,R,dt}$ [dB]	≤ 3	≤ 5	≤ 8	≤ 10	> 10

In general table 10 provides a quality classification of terminals regarding double talk performance. However, this does not mean that a terminal which is category 1 based on the double talk performance is of high quality concerning the overall quality as well.

Measurement method

The test signal to determine the attenuation range during double talk is shown in figure 6. A sequence of uncorrelated CS signals is used which is inserted in parallel in send and receive direction. The test signals are synchronized in time at the acoustical interface. The delay of the test arrangement should be constant during the measurement.

The settings for the test signals are as follows:

Table 11

	Receive direction (s(t))	Send direction (sdt(t))
Pause length between two signal bursts	151,38 ms	151,38 ms
Average signal level (assuming an original pause length of 101,38 ms)	-16 dBm0	-4,7 dBPa
Active signal parts	-14,7 dBm0	-3 dBPa

The test arrangement is according to clause 7.

When determining the attenuation range in receive direction the signal measured at the artificial ear referred to the test signal inserted.

The level is determined as level versus time from the time domain. The integration time of the level analysis is 5 ms. The attenuation is determined from the level difference measured at the beginning of the double talk always with the beginning of the CS-signal in receive direction until its complete activation (during the pause in the send channel). The analysis is performed over the complete signal starting with the second CS-signal. The first CS-signal is not used for the analysis.

7.2.17.3 Detection of echo components during double talk

Requirement

Echo Loss during double talk is the echo suppression provided by the terminal during double talk measured at the electrical reference point.

NOTE: The echo attenuation during double talk is based on the parameter Talker Echo Loudness Rating (TELRdt). It is assumed that the terminal at the opposite end of the connection provides nominal Loudness Rating (SLR + RLR = 10 dB).

Under these conditions the requirements given in table 12 are applicable (more information can be found in annex A of the ITU-T Recommendation P.340 [17]).

Table 12

Category (according to ITU-T Rec. P.340 [17])	**1**	**2a**	**2b**	**2c**	**3**
	Full Duplex Capability	Partial Duplex Capability			No Duplex Capability
Echo Loss [dB]	≥ 27	≥ 3	≥ 17	≥ 11	< 11

Measurement method

The test arrangement is according to clause 7.1.

The double talk signal consists of a sequence of orthogonal signals which are realized by voice-like modulated sine waves spectrally shaped similar to speech. The measurement signals used are shown in figure 7. A detailed description can be found in ITU-T Recommendation P.501 [19].

The signals are fed simultaneously in send and receive direction. The level in send direction is -4,7 dBPa at the MRP (nominal level), the level in receive direction is -16 dBm0 at the electrical reference point (nominal level).

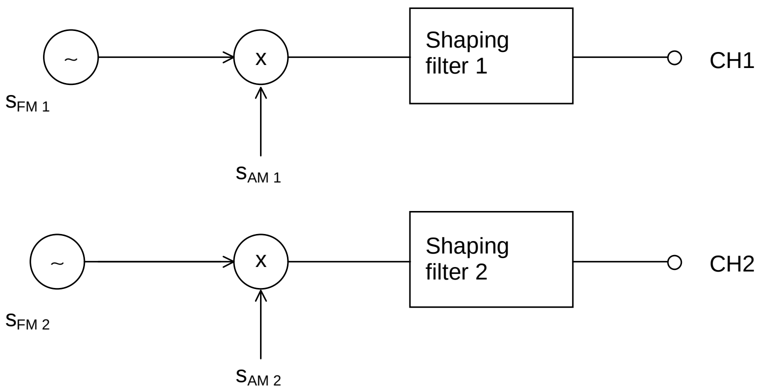


Figure 7: Measurement signals

$$s_{FM1,2}(t) = \sum A_{FM1,2} * \cos(2\pi t\, n * F_{01,2})\ ;\ n = 1,2,.... \quad (5)$$

$$s_{AM1,2}(t) = A_{AM1,2} * \cos(2\pi t F_{AM1,2}); \quad (6)$$

The settings for the signals are as follows:

Table 13: Parameters of the two test signals for double talk measurement based on AM-FM modulated sine waves

Receive direction				Send direction		
f_m(Hz)	$f_{mod(fm)}$(Hz)	F_{am}(Hz)		f_m(Hz)	$f_{mod(fm)}$(Hz)	F_{am}(Hz)
125	±2,5	3		150	±2,5	3
250	±5	3		270	±5	3
500	±10	3		540	±10	3
750	±15	3		810	±15	3
1 000	±20	3		1 080	±20	3
1 250	±25	3		1 350	±25	3
1 500	±30	3		1 620	±30	3
1 750	±35	3		1 890	±35	3
2 000	±40	3		2 160	±35	3
2 250	±40	3		2 400	±35	3
2 500	±40	3		2 650	±35	3
2 750	±40	3		2 900	±35	3
3 000	±40	3		3 150	±35	3
3 250	±40	3		3 400	±35	3
3 500	±40	3		3 650	±35	3
3 750	±40	3		3 900	±35	3
4 000	±40	3		4 150	±35	3
4 250	±40	3		4 400	±35	3
4 500	±40	3		4 650	±35	3
4 750	±40	3		4 900	±35	3
5 000	±40	3		5 150	±35	3
5 250	±40	3		5 400	±35	3
5 500	±40	3		5 650	±35	3
5 750	±40	3		5 900	±35	3
6 000	±40	3		6 150	±35	3
6 250	±40	3		6 400	±35	3
6 500	±40	3		6 650	±35	3
6 750	±40	3		6 900	±35	3
7 000	±40	3				
NOTE: Parameters of the shaping filter: F ≥ 250 Hz: Low Pass Filter, 5 dB/oct; f < 250 Hz,: High Pass Filter.						

The test signal is measured at the electrical reference point (send direction). The measured signal consists of the double talk signal which was fed in by the artificial mouth and the echo signal. The echo signal is filtered by comb filter using mid-frequencies and bandwidth according to the signal components of the signal in receive direction (see ITU-T Recommendation P.501 [19]). The filter will suppress frequency components of the double talk signal.

In each frequency band which is used in receive direction the echo attenuation can be measured separately. The requirement for category 1 is fulfilled if in any frequency band the echo signal is either below the signal noise or below the required limit. If echo components are detectable, the classification is based on table 13. The echo attenuation is to be achieved for **each individual frequency band** according to the different categories.

7.2.17.4 Minimum activation level and sensitivity of double talk detection

For further study.

7.2.18 Switching characteristics

NOTE: Additional requirements may be needed in order to further investigate the effect of NLP implementations on the users' perception of speech quality.

7.2.18.1 Activation in send direction

The activation in send direction is mainly determined by the built-up time $T_{r,S,min}$ and the minimum activation level ($L_{S,min}$). The minimum activation level is the level required to remove the inserted attenuation in send direction during idle mode. The built-up time is determined for the test signal burst which is applied with the minimum activation level.

The activation level described in the following is always referred to the test signal level at the Mouth Reference Point (MRP).

Requirement

The minimum activation level $L_{S,min}$ shall be ≤ -20 dBPa.

The built-up time $T_{r,S,min}$ (measured with minimum activation level) should be ≤ 15 ms.

Measurement method

The structure of the test signal is shown in figure 8. The test signal consists of CSS components according to ITU-T Recommendation P.501 [19] with increasing level for each CSS burst.

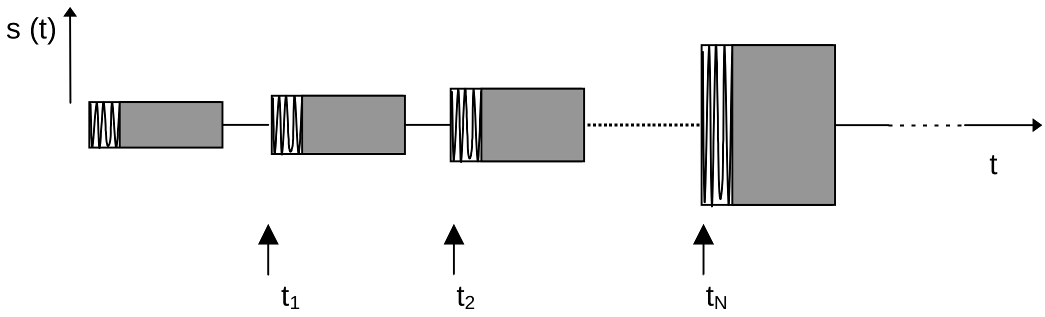


Figure 8: Test signal to determine the minimum activation level and the built-up time

The settings of the test signal are as follows:

Table 14

	CSS duration / pause duration	**Level of the first CS Signal (active signal part at the MRP)**	**Level difference between two periods of the test signal**
CSS to determine switching charachteristic **in send direction**	~250 ms / ~450 ms	-23 dBPa (see note)	1 dB
NOTE: The level of the active signal part corresponds to an average level of -24,7 dBPa at the MRP for the CSS according to ITU-T Recommendation P.501 [19] assuming a pause of about 100 ms.			

It is assumed that the pause length of about 450 ms is longer than the hang-over time so that the test object is back to idle mode after each CSS burst.

The test arrangement is described in clause 7.1.

The level of the transmitted signal is measured at the electrical reference point. The measured signal level is referred to the test signal level and displayed versus time. The levels are calculated from the time domain using an integration time of 5 ms.

The minimum activation level is determined from the CSS burst which indicates the first activation of the test object. The time between the beginning of the CSS burst and the complete activation of the test object is measured.

NOTE: If the measurement using the CS-Signal does not allow to clearly identify the minimum activation level, the measurement may be repeated by using a one syllable word instead of the CS-Signal. The word used should be of similar duration, the average level of the word should be adapted to the CS-signal level of the according CS-burst.

7.2.18.2 Silence Suppression and Comfort Noise Generation

For further study.

7.2.19 Background noise performance

7.2.19.1 Performance in send in the presence of background noise

Requirement

The level of comfort noise shall be within in a range of +2 dB to -5 dB compared to the original (transmitted) background noise. The noise level is calculated with psophometric weighting.

NOTE 1: It is advisable that the comfort noise matches the original signal as good as possible (from a perceptional point of view).

NOTE 2: Input for further specification necessary (e.g. on temporal matching).

The spectral difference between comfort noise and original (transmitted) background noise shall be within the mask given through straight lines between the breaking points on a logarithmic (frequency) - linear (dB sensitivity) scale as given in table 15.

Table 15: Requirements for spectral adjustment of comfort noise (mask)

Frequency	Upper limit	Lower limit
200 Hz	12 dB	-12 dB
800 Hz	12 dB	-12 dB
800 Hz	10 dB	-10 dB
2 000 Hz	10 dB	-10 dB
2 000 Hz	6 dB	-6 dB
4 000 Hz	6 dB	-6 dB
8 000 Hz	6 dB	-6 dB
NOTE: All sensitivity values are expressed in dB on an arbitrary scale.		

Measurement method

The background noise simulation as described in clause 7.1 is used.

The handset terminal is set-up as described in clause 7.1. The handset is mounted at the HATS position (see ITU-T Recommendation P.64 [15]).

First the background noise transmitted in send is recorded at the POI for a period of at least 20 s.

In a second step a test signal is applied in receive direction consisting of an initial pause of 10 s and a periodical repetition of the Composite Source Signal in receive direction (duration 10 s) with nominal level to enable comfort noise injection simultaneously with the background noise. For the measurement the background noise sequence has to be started at the same point as it was started in the previous measurement. Alternatively other speech like test signals (e.g. artificial voice) with the same signal level can be used.

The transmitted signal is recorded in send direction at the POI.

The power density spectra measured in send direction without far end speech simulation averaged between 10 s and 20 s is referred to the power density spectrum measured in send direction determined during the period with far end speech simulation in receive direction averaged between 10 s and 20 s. Level and spectral differences between both power density spectra are analysed and compared to the requirements.

7.2.19.2 Speech quality in the presence of background noise

Requirement

Speech Quality for wideband systems can be tested based on EG 202 396-3 [i.4]. The test method is applicable for narrowband (100 Hz to 4 kHz) and wideband (100 Hz to 8 kHz) transmission systems. LQOn is used for narrowband and LQOw is used for wideband systems. The test method described leads to three MOS-LQO quality numbers:

- N-MOS-LQOw: Transmission quality of the background noise.
- S-MOS-LQOw: Transmission quality of the speech.
- G-MOS-LQOw: Overall transmission quality.

For the background noises defined in clause 7.1 the following requirements apply:

- N-MOS-LQOw ≥ 3,5.
- S-MOS-LQOw ≥ 3,5.
- G-MOS-LQOw ≥ 3,5.

NOTE: It is recommended to test the terminal performance with other types of background noises if the terminal is likely to be exposed to other noises than specified in 7.1.

Measurement method

The background noise simulation as described in clause 7.1 is used. The handset terminal is set-up as described in clause 7.1. The handset is mounted at the HATS position (see ITU-T Recommendation P.64 [15]).

The background noise should be applied for at least 5 s in order to adapt noise reduction algorithms in advance the test.

The near end speech signal consists of 8 sentences of speech (2 male and 2 female talkers, 2 sentences each). Appropriate speech samples can be found in ITU-T Recommendation P.501 [19]. The preferred language is French since the objective method was validated with French language. The test signal level is -4,7 dBPa at the MRP.

Three signals are required for the tests:

1) The clean speech signal is used as the undisturbed reference (see EG 202 396-3 [i.4]).
2) The speech plus undisturbed background noise signal is recorded at the terminal's microphone position using an omnidirectional measurement microphone with a linear frequency response between 50 Hz and 12 kHz.
3) The send signal is recorded at the electrical reference point.

N-MOS-LQOw, S-MOS LQOw and G-MOS LQOw are calculated as described in EG 202 396-3 [i.4].

7.2.19.3 Quality of background noise transmission (with far end speech)

Requirement

The test is carried out applying the Composite Source Signal in receive direction. During and after the end of Composite Source Signal bursts (representing the end of far end speech simulation) the signal level in send direction should not vary more than 10 dB (during transition to transmission of background noise without far end speech). The measurement is conducted for all types of background noise as defined in clause 7.1.

Measurement method

The test arrangement is according to clause 7.1.

The background noises are generated as described in clause 7.1.

First the measurement is conducted without inserting the signal at the far end. At least 10 s of noise is analysed. The background signal level versus time is calculated using a time constant of 35 ms. This is the reference signal.

In a second step the same measurement is conducted but with inserting the CS-signal at the far end. The exactly identical background noise signal is applied. The background noise signal must start at the same point in time which was used for the measurement without far end signal. The background noise should be applied for at least 5 s in order to allow adaptation of the noise reduction algorithms. After at least 5 s a Composite Source Signal according to ITU-T Recommendation P.501 [19] is applied in receive direction with a duration of ≥ 2 CSS periods. The test signal level is -16 dBm0 at the electrical reference point.

The send signal is recorded at the electrical reference point. The test signal level versus time is calculated using a time constant of 35 ms.

The level variation in send direction is determined during the time interval when the CS-signal is applied and after it stops. The level difference is determined from the difference of the recorded signal levels versus time between reference signal and the signal measured with far end signal.

7.2.19.4 Quality of background noise transmission (with near end speech)

Requirement

The test is carried out applying a simulated speech signal in send direction. During and after the end of the simulated speech signal (Composite Source Signal bursts) the signal level in send direction should not vary more than 10 dB.

Measurement method

The test arrangement is according to clause 7.1.

The background noises are generated as described in clause 7.1. The background noise should be applied for at least 5 s in order to allow adaptation of the noise reduction algorithms.

The near end speech is simulated using the Composite Source Signal according to ITU-T Recommendation P.501 [19] with a duration of ≥ 2 CSS periods. The test signal level is -4,7 dBPa at the MRP.

The send signal is recorded at the electrical reference point. The test signal level versus time is calculated using a time constant of 35 ms.

First the measurement is conducted without inserting the signal at the near end. The signal level is analysed versus time. In a second step the same measurement is conducted but with inserting the CS-signal at the near end. The level variation is determined by the difference between the background noise signal level without inserting the CS-signal and the maximum level of the noise signal during and after the CS-bursts in send direction.

7.2.20 Quality of echo cancellation

7.2.20.1 Temporal echo effects

Requirement

This test is intended to verify that the system will maintain sufficient echo attenuation during single talk. The measured echo attenuation during single talk should not decrease by more than 6 dB from the maximum measured during the TCLw test.

Measurement method

The test arrangement is according to clause 7.1.

The test signal consists of periodically repeated Composite Source Signal according to ITU-T Recommendation P.501 [19] with an average level of -5 dBm0 as well as an average level of -25 dBm0. The echo signal is analysed during a period of at least 2,8 s which represents 8 periods of the CS signal. The integration time for the level analysis shall be 35 ms, the analysis is referred to the level analysis of the reference signal.

The measurement result is displayed as attenuation versus. time. The exact synchronization between input and output signal has to be guaranteed.

NOTE 1: In addition tests with more speech like signals should be made, e.g. ITU-T Recommendation P.50 [11] to see time variant behaviour of EC. However, for such tests the simple broadband attenuation based test principle as described above cannot be applied due to the time varying spectral content of the speech like signals.

NOTE 2: The analysis is conducted only during the active signal part, the pauses between the Composite Source Signals are not analysed. The analysis time is reduced by the integration time of the level analysis (35 ms).

7.2.20.2 Spectral Echo Attenuation

Requirement

The echo attenuation versus frequency shall be below the tolerance mask given in table 16.

Table 16: Echo attenuation limits

Frequency	Limit
100 Hz	-41 dB
1 300 Hz	-41 dB
3 450 Hz	-46 dB
5 200 Hz	-46 dB
7 500 Hz	-37 dB
8 000 Hz	-37 dB
NOTE: The limit at intermediate frequencies lies on a straight line drawn between the given values on a log (frequency) - linear (dB) scale.	

During the measurement it should be ensured that the measured signal is really the echo signal and not the Comfort Noise which possibly may be inserted in send direction in order to mask the echo signal.

Measurement method

The test arrangement is according to clause 7.1.

Before the actual measurement a training sequence is fed in consisting of 10 s CS signal according to ITU-T Recommendation P.501 [19]. The level of the training sequence is -16 dBm0.

The test signal consists of a periodically repeated Composite Source Signal. The measurement is carried out under steady-state conditions. The average test signal level is -16 dBm0, averaged over the complete test signal. 4 CS signals including the pauses are used for the measurement which results in a test sequence length of 1,4 s. The power density spectrum of the measured echo signal is referred to the power density spectrum of the original test signal. The analysis is conducted using FFT analysis with 8 k points (48 kHz sampling rate, Hanning window).

The spectral echo attenuation is analysed in the frequency domain in dB.

7.2.20.3 Occurrence of Artefacts

For further study.

7.2.21 Variant Impairments; Network Dependant

For further study.

7.2.21.1 Delay versus Time Send

For further study.

7.2.21.2 Delay versus Time Receive

For further study.

7.2.21.3 Quality of jitter buffer adjustment

For further study.

7.3 Codec Specific Requirements

7.3.1 Send Delay

For a VoIP terminal, send delay is defined as the one-way delay from the acoustical input (mouthpiece) of this VoIP terminal to its interface to the packet based network. The total send delay is the upper bound on the mean delay and takes into account the delay contributions of all of the elements shown in figures 2 and A.1 in ITU-T Recommendation G.1020 [10], respectively.

The send delay T(s) is defined as follows:

$$T(s) = T(ps) + T(la) + T(rif) + T(asp) \qquad (7)$$

Where:

T(ps) = packet size = N * T(fs)

N = number of frames per packet

T(fs) = frame size of encoder

T(la) = look-ahead of encoder

T(aif) = air interface framing

T(asp) = allowance for signal processing

The additional delay required for IP packet assembly and presentation to the underlying link layer will depend on the link layer. When the link layer is a LAN (e.g. Ethernet), this additional time will usually be quite small. For the purposes of the present document it is assumed that in the test setup this delay can be neglected.

NOTE 1: The size of T(aif) is for further study.

Requirement

The allowance for signal processing shall be T(asp) < 10 ms

NOTE 2: With the knowledge of the codec specific values for T(fs) and T(la) the requirements for send delay for any type of coder and any frame size T(fs) can easily be calculated by equation 7. Table 17 provides requirements calculated accordingly for frequently used codecs and packet sizes.

Table 17

Codec	N	T(fs) in ms	T(ps) in ms	T(la) in ms	T(aif) in ms	T(asp) in ms	T(s) in ms	T(s) requirement in ms
G.722 [7]	80	0,0625	10	0	0	10	20,0625	< 21
	160	0,0625	20	0	0	10	30,0625	< 31
G.722.1 [8]	1	20	10	5	0	10	To be completed	To be completed
	2	20	20	5	0	10	To be completed	To be completed
L16-256 [25]	160	0,0625	10	0	0	10	20,0625	< 21

Further information about the different sources of delay for different codecs can be found in annex A.

Measurement method

The test signal to be used for the measurements shall be a Composite Source Signal (CSS) as described in ITU-T Recommendation P.501 [19]. The test signal consists of the voiced part as described in ITU-T Recommendation P.501 [19] followed by a pseudo random noise sequence with a periodicity of minimum 500 ms. The spectrum of acoustic signal produced by the artificial mouth is calibrated under free field conditions at the MRP. The test signal level shall be -4,7 dBPa, measured at the MRP. The test signal level is averaged over the complete test signal sequence.

NOTE 3: If the expected delay is higher than 500 ms a pseudo random sequence with a higher periodicity should be used.

The handset terminal is setup as described in clause 7.1. The handset is mounted in the HATS position (see ITU-T Recommendation P.64 [15]). The application force used to apply the handset against the artificial ear shall be stated in the test report.

The delay is calculated using the cross correlation function between the signal at the electrical test point and the signal at the MRP. The cross correlation analysis has to be chosen in such a way that the maximum delay of 500 ms can be analysed. The measurement is corrected by the delay introduced by the test equipment.

The delay is expressed in ms, determined from the maximum of the cross correlation function.

NOTE 4: Delay may be time variant. Therefore constant monitoring of the actual delay may be required when evaluating the range of delay which can be observed in a given connection. The test setup should take into account either real network conditions or the tools needed to simulate typical causes for time variant delay (e.g. packet loss) during the measurement period. Other methods like running cross correlation or delay estimation procedures e.g. used in ITU-T Recommendation P.862 [22] may be used.

7.3.2 Receive delay

For a VoIP terminal, receive delay is defined as the one-way delay from the interface to the packet based network of this VoIP terminal to its acoustical output (earpiece). The total receive delay is the upper bound on the mean delay and takes into account the delay contributions of all of the elements shown in figures 3, A.1 and A.2 of ITU-T Recommendation G.1020 [10], respectively.

The receive delay T(r) is defined as follows:

$$T(r) = T(fs) + T(fi) + T(aif) + T(jb) + T(plc) + T(asp) \quad (8)$$

Where:

- T(fs) = frame size of encoder
- T(fi) = filter processing delay
- T(aif) = air interface framing
- T(jb) = jitter buffer size
- T(plc) = PLC buffer size
- T(asp) = allowance for signal processing

The additional delay required for IP packet dis-assembly and presentation from the underlying link layer will depend on the link layer. When the link layer is a LAN (e.g. Ethernet), this additional time will usually be quite small. For the purposes of the present document it is assumed that in the test setup this delay can be neglected.

NOTE 1: The size of T(aif) is for further study.

Requirements

The allowance for signal processing shall be T(asp) < 10 ms.

The additional delay introduced by the jitter buffer shall be T(jb) ≤ 10 ms.

For coders without integrated PLC the additional PLC buffer size shall be T(plc) < 10 ms.

For coders with integrated PLC the additional PLC buffer size shall be T(plc) = 0 ms.

NOTE 2: With the knowledge of the codec specific values for T(fs) and T(la) the requirements for receive delay for any type of coder and any frame size T(fs) can easily be calculated by equation 8. Table 18 provides requirements calculated accordingly for some frequently used codecs and packet sizes as an example.

Table 18

Codec	N	T(fs) in ms	T(fi) in ms	T(aif) in ms	T(jb) in ms	T(plc) in ms	T(asp) in ms	T(r) in ms	T(r) Requirement in ms
G.722 [7]	80	0,0625	0	0	10	10	10	30,0625	< 31
G.722 [7]	160	0,0625	0	0	10	10	10	30,0625	< 31
G.722.1 [8]	1	20	0	0	10	0	10	40	< 40
G.722.1 [8]	2	20	0	0	10	0	10	40	< 40
L16-256 [25]	160	0,0625	0	0	10	10	10	30,0625	< 31
NOTE 1: T(ps) = packet size = N * T(fs). NOTE 2: N = number of frames per packet.									

NOTE 3: These requirements are based on the lowest possible delay values which can be expected under ideal network conditions. Caution should be exercised to ensure that the terminal is operated under optimum conditions in order to avoid adverse effects, e.g. network conditions, settings and memory effects of the terminal jitter buffer.

Measurement method

The test signal to be used for the measurements shall be a Composite Source Signal (CSS) as described in ITU-T Recommendation P.501 [19]. The test signal consists of the voiced part as described in ITU-T Recommendation P.501 [19] followed by a pseudo random noise sequence with a periodicity of minimum 500 ms. The test signal level shall be -16 dBm0, measured at the electrical test point. The test signal level is averaged over the complete test signal sequence.

The handset terminal is setup as described in clause 7.1. The handset is mounted in the HATS position (see ITU-T Recommendation P.64 [15]). The application force used to apply the handset against the artificial ear shall be stated in the test report.

The delay is calculated using the cross correlation function between the signal at the electrical test point and the signal at the DRP. The cross correlation analysis has to be chosen in such a way that the maximum delay of 500 ms can be analysed. The measurement is corrected by the delay introduced by the test equipment.

The delay is expressed in ms, determined from the maximum of the cross correlation function.

NOTE 4: Delay may be time variant. Therefore constant monitoring of the actual delay may be required when evaluating the range of delay which can be observed in a given connection. The test setup should take into account either real network conditions or the tools needed to simulate typical causes for time variant delay (e.g. packet loss) during the measurement period. Other methods like running cross correlation or delay estimation procedures e.g. used in ITU-T Recommendation P.862 [22] may be used.

7.3.3 Objective listening speech quality MOS-LQOM in send direction

The listening speech quality tests are conducted under clean network conditions.

Requirements

The requirements for the listening speech quality are as follows:

Table 19

Speech coder	MOS-LQOM
G.722 [7] @64 kbit/s	> 4,0
G.729.1 [9] @ 32 kbit/s	> 4,3
G.722.1 [8] @ 12,65 kbit/s	> 4,0
L16-256 [25]	> 4,3

NOTE: Currently no test method is available for terminals, TOSQA 2001 is one method (EG 201 377-1 [i.1]), which may be used in half-channel scenarios.

Measurement method

For further study.

7.3.4 Objective listening quality MOS-LQOM in receive direction

The listening speech quality tests are conducted under clean network conditions as well as with network impairments simulated. In addition to the listening speech quality tests the delay is measured.

Requirement

The requirement for the listening speech quality and the delay under clean network conditions are as follows:

Table 20

Speech coder	MOS-LQOM (see note 1) (with ideal terminal characteristics)	MOS-LQOM
G.722 [7] @ 64 kbit/s	(> 4,0)	> 3,6
G.722.1 [8] @ 12,65 kbit/s	(> 4,0)	> 3,6
G.729.1 [9] @32 kbit/s	(> 4,2)	> 3,6
L16-256 [25]	(> 4,2)	> 3,6
NOTE 1: Informative. NOTE2: The MOS-LQOM requirements in receive are lower than the requirements set in send. This takes into account that in receive the impairment introduced by a non ideal frequency response characteristics in receive in addition to the impairment introduced by the codec impairment is more dominant then in send.		

Test method

For further study.

For the performance tests with network impairments the following settings are used.

Table 21: Network conditions for electrical-acoustical measurements (speech samples)

Condition	Packet loss (equal)	Delay variation
0c (see note 2) (VAD)	0	No
1	0	No
2	0	20 ms (see note 1)
3	1 %	No
4	1 %	20 ms (see note 1)
5	3 %	No
NOTE 1: Delay variation produced with a Pareto-Distribution and r = 0,5. NOTE 2: VAD on, all other conditions (1 to 5) tested with VAD off. NOTE 3: For some network emulation tools, it is necessary to introduce a constant delay to offer the possibility to generate a delay variation distribution. This delay has to be subtracted from the measured delay before interpreting the results.		

Table 22: Requirements for ITU-T Recommendation G.722 speech codecs

Condition	MOS-LQOM	Delay
1	> 3,6	< 31 ms
2	> 3,4	< 51 ms
3	> 3,4	< 31 ms
4	> 3,4	< 51 ms
5	> 3,2	< 31 ms
NOTE: The settings are derived from the ones used in the ETSI Plugtest VoIP speech quality test events.		

Table 23: Requirements for G.722.1 speech codecs

Condition	MOS-LQOM	Delay
1	> 3,6	< 40 ms
2	> 3,4	< 60 ms
3	> 3,4	< 40 ms
4	> 3,4	< 60 ms
5	> 3,4	< 40 ms

7.3.4.1 Efficiency of Packet Loss Concealment (PLC)

For further study.

7.3.4.2 Efficiency of delay variation removal

For further study.

Annex A (informative):
Processing delays in VoIP terminals

This annex gives some elements about delays generated in VoIP terminals. At first, we consider only wired terminals. These terminals could be schematized as shown in figure A.1.

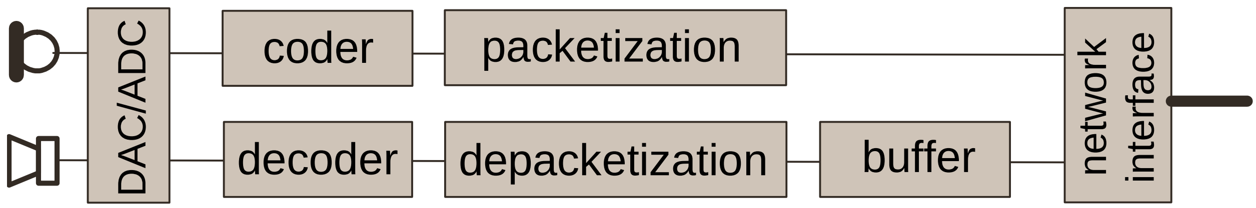


Figure A.1: Synoptic of the different functions implemented in a VoIP terminal

The implemented functions in the send part of the terminal are:

- the analog-digital conversion;
- the encoding;
- the packetization;
- the interfacing with the network.

The implemented functions in the receive part of the terminal are:

- the interfacing with the network;
- the depacketization;
- the buffering;
- the decoding;
- the digital-analog conversion.

Let us examine each function's contribution to the processing delay characterizing VoIP terminals.

On the send part of the terminal, the **network interface** operates the transfer of digital data from IP stack to IP network. At the reception, the network interface operates the transfer of digital data from IP network to IP stack. The network interface has a low contribution to the delay. The contribution is estimated at less than 2 ms per transmission way (send and receive direction).

The **packetization** represents the transfer of the audio frames through the IP stack, from the telephony applicative part of the terminal to the transmission network. The packetization consists in adding specific headers (associated to different protocols) to audio frames. The delay associated to the packetization is considered as not significant and included into encoding time.

Encoding corresponds to the compression of the speech signal. The delay associated to the encoding process depends on the implemented codec and the payload's length (number of audio frames) inserted into each IP packet. On the send part of the terminal, encoding is the main contribution to the processing delay. The delay can strongly change according to the codec and the payload's length.

Analog to digital conversion consists in transforming speech signal from analog to digital format. The processing delay associated to the conversion is considered as not significant.

Digital to analog conversion consists in transforming speech signal from digital to analog format. As analog to digital conversion, the processing delay associated to digital to analog conversion is considered as not significant.

The **depacketization** represents the transfer of the audio frames through the IP stack, from transmission network to the telephony applicative part of the terminal. The depacketization consists in tacking off the headers associated to protocols to get back audio frames after transmission. The delay associated to the depacketization is considered as not significant and included into the decoding processing time.

The first role of the **jitter buffer** is to ensure synchronization between send and receive terminals. This synchronization is carried out by buffering the audio frames received from the IP stack before send them to the decoder. The second role of the jitter buffer is to smooth a possible variation of the transmission time. If synchronization of send and receive terminals requires a minimum size of buffer, smoothing transmission delay variation requires a buffer size depending on jitter produced by the network. High variations of transmission time involve an important size of the buffer to smooth jitter. Jitter buffers can be implemented either as buffer with static size(s) (several sizes are possible) or as dynamic buffer. In the last case, size management is carried out according to QoS present on the network interface. Jitter buffer is the main contribution to the processing time on the reception part of VoIP terminal.

Decoding corresponds to the rebuilding of speech signal from receive audio frames. The delay associated to decoding depends on the codec implemented. Decoding contributes in a significant way to the processing time on the reception part of VoIP terminal.

Table A.1 presents the processing times of VoIP terminals for different codecs and IP packet payload's lengths.

In table A 1 x1, x2, x3, x4, y5, x6 and x7 represent the encoding delays according to selected codec. In the same way, y1, y2, y3, y4, y5, y6 and y7 represent the decoding delays according to selected codec.

According to selected codec and payload's length, columns 5 and 6 show overall encoding and decoding delays respectively. Overall encoding time takes into account algorithm, encoding and packetization delays. Overall decoding time takes into account algorithm, decoding and depacketization delays.

Column 7 shows for each codec and payload's length the real time condition. It stands for the maximum duration to encode and decode at the same time. IP terminals have to meet this requirement.

Column 10 shows the minimum delay induced by the jitter buffer. To ensure a correct running of the VoIP terminal, the minimal size of jitter buffer has to correspond to the IP packet payload's length. Furthermore, a double buffering operation induces 10 additional ms in the overall jitter buffer processing.

Column 12 shows the minimum end-to-end delay induced by two terminals connected to a "perfect" network (i.e. with no jitter, no packet loss and with a null transmission delay), with real time condition at the lower limit (i.e. not significant encoding and decoding times).

Column 13 shows the minimum end-to-end delay induced by two terminals connected to a "perfect" network (i.e. with no jitter, no packet loss and with a null transmission delay), with real time condition at the upper limit (i.e. encoding + decoding times very close to the payload size).

Table A.1

Codec	Frame	Lookahead	Payload	Sending procesing delay = Algorithm delay + coding and packetization delay	Receiving procesing delay = Algorithm delay + coding and packetization delay	Real time condition	Network interface and ADC delay	Network interface and DAC delay	Minimum delay of the jitter buffer	Maximum delay of the jitter buffer	Minimum End to End delay with the lower jitter buffer processing time when real time condition is minimum (x+y=0)	Minimum End to End delay with the lower jitter buffer processing time when real time condition is maximum (x+y=upper limit)	Maximum End to End delay with the higher jitter buffer processing time when real time condition is minimum (x+y=0)	Minimum End to End delay with the higher jitter buffer processing time when real time condition is maximum (x+y=upper limit)
G.711	1	0	10	10+x1	y1	x1+y1<10 ms	2	2	20	400	34	44	414	424
	1	0	20	2*(10+x1)	2*y1	2*(x1+y1)<20 ms	2	2	30	400	54	74	424	444
	1	0	30	3*(10+x1)	3*y1	3*(x1+y1)<30 ms	2	2	40	400	74	104	434	464
	1	0	40	4*(10+x1)	4*y1	4*(x1+y1)<40 ms	2	2	50	400	94	134	444	484
	1	0	50	5*(10+x1)	5*y1	5*(x1+y1)<50 ms	2	2	60	400	114	164	454	504
	1	0	60	6*(10+x1)	6*y1	6*(x1+y1)<60 ms	2	2	70	400	134	194	464	524
G.729	10	5	10	(10+x2)+5	y2	x2+y2<10 ms	2	2	20	400	39	49	419	429
	10	5	20	(2*(10+x2))+5	2*y2	2*(x2+y2)<20 ms	2	2	30	400	59	79	429	449
	10	5	30	(3*(10+x2))+5	3*y2	3*(x2+y2)<30 ms	2	2	40	400	79	109	439	469
	10	5	40	(4*(10+x2))+5	4*y2	4*(x2+y2)<40 ms	2	2	50	400	99	139	449	489
	10	5	50	(5*(10+x2))+5	5*y2	5*(x2+y2)<50 ms	2	2	60	400	119	169	459	509
	10	5	60	(6*(10+x2))+5	6*y2	6*(x2+y2)<60 ms	2	2	70	400	139	199	469	529
G.723.1	30	7,5	30	(30+x3)+7,5	y3	x3+y3<30 ms	2	2	40	400	81,5	111,5	441,5	471,5
	30	7,5	60	(2*(30+x3))+7,5	2*y3	2*(x3+y3)<60 ms	2	2	70	400	141,5	201,5	471,5	531,5
NB-AMR	20	5	20	(20+x4)+5	y4	x4+y4<20 ms	2	2	30	400	59	79	429	449
	20	5	40	(2*(20+x4))+5	2*y4	2*(x4+y4)<40 ms	2	2	50	400	99	139	449	489
	20	5	60	(3*(20+x4))+5	3*y4	3*(x4+y4)<60 ms	2	2	70	400	139	199	469	529
G.722	10	1,5	10	(10+x5)+1,5	y5	x5+y5<10 ms	2	2	20	400	35,5	45,5	415,5	425,5
	10	1,5	20	(2*(10+x5))+1,5	2*y5	2*(x5+y5)<20 ms	2	2	30	400	55,5	75,5	425,5	445,5
	10	1,5	30	(3*(10+x5))+1,5	3*y5	3*(x5+y5)<30 ms	2	2	40	400	75,5	105,5	435,5	465,5
	10	1,5	40	(4*(10+x5))+1,5	4*y5	4*(x5+y5)<40 ms	2	2	50	400	95,5	135,5	445,5	485,5
	10	1,5	50	(5*(10+x5))+1,5	5*y5	5*(x5+y5)<50 ms	2	2	60	400	115,5	165,5	455,5	505,5
	10	1,5	60	(6*(10+x5))+1,5	6*y5	6*(x5+y5)<60 ms	2	2	70	400	135,5	195,5	465,5	525,5
WB-AMR	20	5	20	(20+x6)+5	y6+0,94	x6+y6<20 ms	2	2	30	400	59,94	79,94	429,94	449,94
	20	5	40	(2*(20+x6))+5	2*y6+0,94	2*(x6+y6)<40 ms	2	2	50	400	99,94	139,94	449,94	489,94
	20	5	60	(3*(20+x6))+5	3*y6+0,94	3*(x6+y6)<60 ms	2	2	70	400	139,94	199,94	469,94	529,94
G.729.1	20	25	20	(20+x7)+25+1,97	y7+1,97	x7+y7<20 ms	2	2	30	400	82,94	102,94	452,94	472,94
	20	25	40	(2*(20+x7))+25+1,97	2*y7+1,97	2*(x7+y7)<40 ms	2	2	50	400	122,94	162,94	472,94	512,94
	20	25	60	(3*(20+x7))+25+1,97	3*y7+1,97	3*(x7+y7)<60 ms	2	2	70	400	162,94	222,94	492,94	552,94

Annex B (informative): Bibliography

ITU-T Recommendation P.51: "Artificial mouth".

Annex C (informative): Optimum Frequency Responses for Wideband Transmission in Receive Direction - Underlying Subjective Experiments

For the derivation of the optimum frequency response characteristics first investigations were started with expert listeners who were instructed in the first test to adjust their personally preferred speech sound quality for several wideband devices using a software equalizer. The adjustment of the frequency response characteristics was performed in 1/3 octaves between 100 Hz and 8 000 Hz. Two main points were observed:

- the settings chosen by the experts were significantly different for all tested phones;
- in subsequent interviews the experts stated that it was very difficult to adjust a preferred speech sound without having a "reference sound" or a comparison to another device.

Initiated by these results a second expert test was conducted. The experts now had to rank the speech samples of each phone separately. These were recorded using an artificial head and the equalizer settings adjusted in the first test. The results of these tests clearly indicate a "winner" response characteristics for each phone. Furthermore the "winner" frequency response characteristics of different phones look similar.

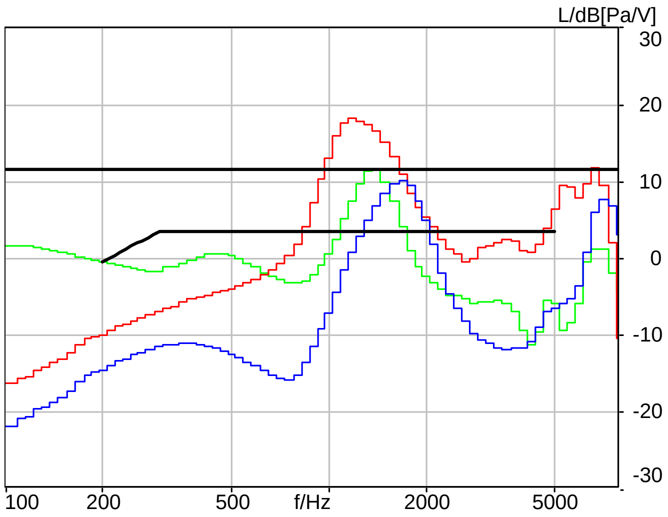


Figure C.1: Receive frequency response of 3 wideband terminals and the tolerance scheme of ES 202 739 (V1.1.1, 2007) measured in handset mode at an artificial head, 3.4 artificial ear, free-field equalization and with 8 N application force

Using these "winner" frequency responses a formal listening test with naïve test persons was conducted. Furthermore equalizer settings providing a flat frequency response measured with DRP to ERP correction, free-field and diffuse-field equalization of the artificial head were used for several phones. Additionally, a recording of the ortho-telephonic reference position was inserted (measurement with two artificial heads in 1m distance to each other). The speech sounds were assessed by 24 listeners on the 5-point MOS-scale in terms of their "overall quality". The speech material (two sentences of two male and two female speakers each) was presented diotically.

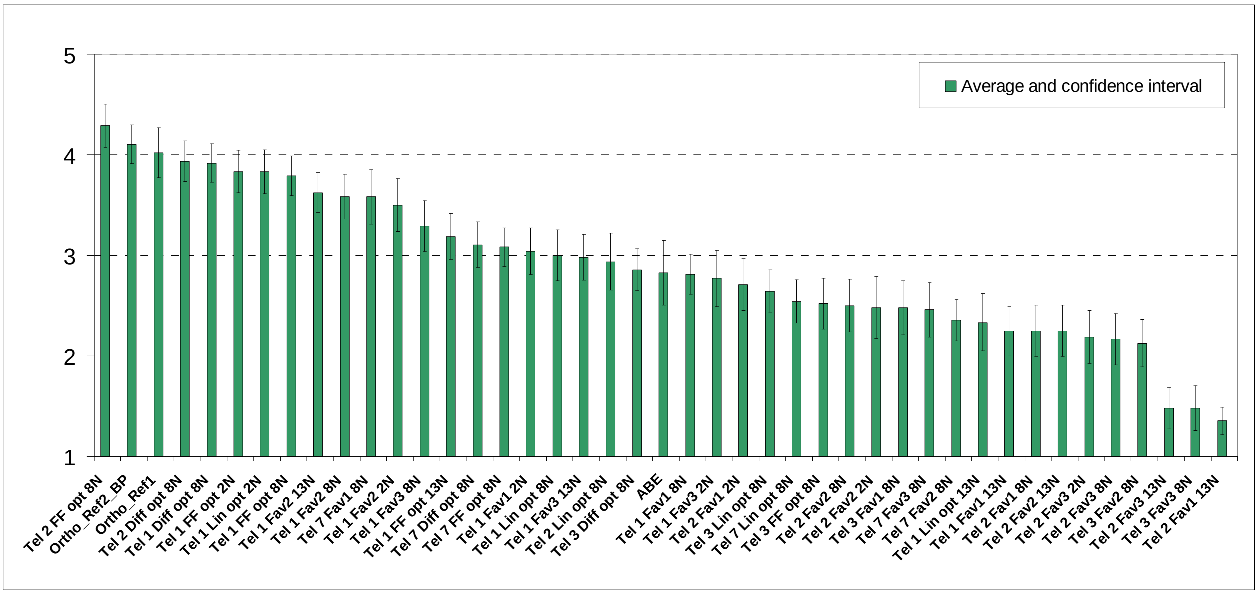


Figure C2: MOS results and confidence intervals of formal listening test in receive direction

The results - shown in **figure C 2** (mean and confidence interval) **-** indicate that the whole quality range was covered by this listening test. As expected, the ortho-telephonic reference condition was one of the best rated samples (see magenta circle). In order to derive a new tolerance scheme all responses which lead to a MOS score of at least 3.6 were extracted and plotted in one diagram (see **figure C 3**). Based on this plot, a new tolerance scheme (thick black lines in **figure C 3**) and a modified measurement setup was defined using a diffuse-field equalized artificial head.

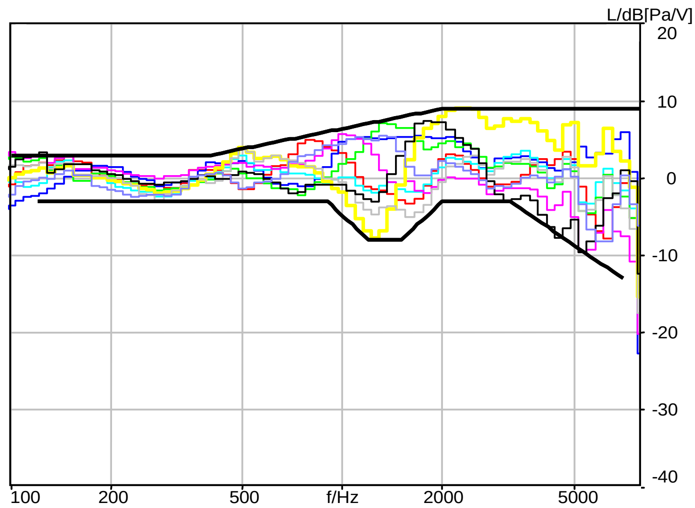


Figure C 1: Frequency responses leading to an MOS of ≥ 3.6 and proposed new tolerance scheme (thick black lines) to be used for diffuse-field corrected measurements.

This work was conducted by Deutsche Telekom Laboratories and HEAD acoustics GmbH. Further information can be found in [i.10].

History

Document history		
V1.2.1	October 2007	Publication
V1.3.1	July 2009	Membership Approval Procedure MV 20090904: 2009-07-07 to 2009-09-04
V1.3.1	September 2009	Publication