Final draft ETSI ES 202 737 V1.5.1 (2016-10)

ETSI STANDARD

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

Reference

RES/STQ-242

Keywords

narrowband, quality, speech, 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

Important notice

Copyright Notification

Contents

Intellectual Property Rights

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

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 final draft ETSI Standard (ES) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ), and is now submitted for the ETSI standards Membership Approval Procedure.

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

Traditionally, the analogue and digital telephones were interfacing switched-circuit 64 kbit/s PCM networks. With the fast growth of IP networks, terminals 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 harmonised with specifications for traditional digital terminals. However, due to the unique characteristics of the IP networks including packet loss, delay, etc. new performance specifications, 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. Also, the VoIP terminals may use other than 64 kbit/s PCM (Recommendation ITU-T G.711 [7]) speech algorithms.

The advanced signal processing of terminals is targeted to speech signals. Therefore, wherever possible speech signals are used for testing in order to achieve mostly realistic test conditions and meaningful results.

The present document provides speech transmission performance for narrowband VoIP handset and headset terminals.

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 4 kHz narrowband 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.

It is the intention of the present document to describe terminal performance parameters in such way that the remaining variation of parameters can be assessed purely by the E-model.

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

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.

The following referenced documents are necessary for the application of the present document.

[1] ETSI EN 300 726: "Digital cellular telecommunications system (Phase 2+) (GSM); Enhanced Full Rate (EFR) speech transcoding (GSM 06.60)".

[2] ETSI TS 126 171: "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description (3GPP TS 26.171)".

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

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

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

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

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

[8] Recommendation ITU-T G.723.1: "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s".

[9] Recommendation ITU-T G.726: "40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)".

[10] Recommendation ITU-T G.729: "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)".

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

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

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

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

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

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

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

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

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

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

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

[22] IEC 61260-1: "Electroacoustics - Octave-band and fractional-octave-band filters - Part 1: Specifications".

[23] Recommendation ITU-T P.800.1: "Mean Opinion Score (MOS) terminology".

[24] ETSI ES 202 739: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband VoIP terminals (handset and headset) from a QoS perspective as perceived by the user".

[25] ETSI TS 103 224: "Speech and multimedia Transmission Quality (STQ); A sound field reproduction method for terminal testing including a background noise database".

[26] Recommendation ITU-T P.863: "Perceptual objective listening quality assessment".

[27] Recommendation ITU-T P.863.1: "Application guide for Recommendation ITU-T P.863".

[28] Recommendation ITU-T P.1010: "Fundamental voice transmission objectives for VoIP terminals and gateways".

[29] IETF RFC 3550: "RTP: A Transport Protocol for Real-Time Applications".

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

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

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] ETSI EG 201 377-1: "Speech and multimedia Transmission Quality (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 202 425: "Speech Processing, Transmission and Quality Aspects (STQ); Definition and implementation of VoIP reference point".

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

[i.4] NIST Net.

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

[i.5] Netem.

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

[i.6] Trace Control for Netem (TCN): "A. Keller, Trace Control for Netem, Semester Thesis SA-2006-15, ETH Zürich, 2006".

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

Composite Source Signal (CSS): signal composed in time by various signal elements

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 Recommendation ITU-T P.58 [14]

Ear Reference Point (ERP): virtual point for geometric reference located at the entrance to the listener's ear, traditionally used for calculating telephonometric loudness ratings

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

NOTE: 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): point 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:

AM-FM	Amplitude Modulation-Frequency Modulation
AMR	Adaptative Multi-Rate

AMR-NB	Adaptive Multi-Rate NarrowBand
CS	Composite Source
CSS	Composite Source Signal
DRP	ear Drum Reference Point
EC	Echo Canceller
EFR	Enhanced Full Rate
EL	Echo Loss
ERP	Ear Reference Point
ETH	Eidgenössische Technische Hochschule
FFT	Fast Fourrier Transform
GSM	Global System for Mobile communications
HATS	Head And Torso Simulator
IEC	International Electrotechnical Commission
IP	Internet Protocol
IPDV	IP Packet Delay Variation
ITU-T	International Telecommunication Union - Telecommunication standardization sector
MOS	Mean Opinion Score
MOS-LQOy	Mean Opinion Score - Listening Quality Objective, y being N for narrow-band, M for mixed and S for superwideband

NOTE: See Recommendation ITU-T P.800.1 [23].

MRP	Mouth Reference Point
NIST	National Institute of Standards and Technology
NLP	Non Linear Processor
PBX	Private Branch eXchange
PC	Personal Computer
PCM	Pulse Code Modulation
POLQA	Perceptual Objective Listening Quality Assessment
PLC	Packet Loss Concealment
PN	Pseudo-random Noise
POI	Point Of Interconnect
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RLR	Receive Loudness Rating
RMS	Root Mean Square
RTP	Real Time Protocol
SLR	Send Loudness Rating
STMR	SideTone Masking Rating
TCLw	Terminal Coupling Loss (weighted)
TCN	Trace Control for Netem
TDM	Time Division Multiplex
TOSQA	Telecommunication Objective Speech Quality Assessment
VAD	Voice Activity Detector

4 General considerations

4.1 Default coding algorithm

VoIP terminals shall support the coding algorithm according to Recommendation ITU-T G.711 [7] (both μ-law and A-law). VoIP terminals may support other coding algorithms.

NOTE: Associated Packet Loss Concealment (PLC) e.g. as defined in Recommendation ITU-T G.711 [7] appendix I 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 Recommendation ITU-T G.107 [3] taking into account that the E-model does not yet address headsets; this includes the a-priori determination of the desired category of speech transmission quality as defined in Recommendation ITU-T G.109 [5].

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 Recommendation ITU-T G.108 [4], 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.

Recommendation ITU-T G.108 with its amendments [4] provides further guidance on this important issue.

The following optimum terminal parameters from a users' 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.

5 Test equipment

5.1 IP half channel measurement adaptor

The IP half channel measurement adaptor is described in ETSI EG 202 425 [i.2].

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
Measured maximum frequency	20 kHz

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 14 000 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.4] (http://snad.ncsl.nist.gov/itg/nistnet/) or Netem [i.5].

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 sub network 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 nistnet, 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 (Nistnet only runs on routers).
- With an amendment of Netem, TCN (Trace Control for Netem) [i.6] 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 Nistnet distortions, generate a file out of this and playback the exact same distortions with Netem.

NOTE : NIST Net™, NETEM™, Linux™ and X Window System™ are examples of suitable products available commercially. This information is given for the convenience of users of the present document and does not constitute an endorsement by ETSI of these product(s).

5.5 Acoustic environment

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 shall 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.

Standardized measurement methods for measurements with variable echo paths are for further study.

5.6 Influence of terminal delay on measurements

As delay is introduced by the terminal, care shall be taken for all measurements where exact position of the analysis window is required. It shall be checked that the test is performed on the test signal and not on any other signal.

6 Requirements and associated measurement methodologies

6.1 Notes

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.

6.2 Test setup

6.2.1 General

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

When a coder with variable bit rate is used for testing terminal electro acoustical parameters, the bit rate recognized giving the best characteristics should be selected, e.g.:

- AMR-NB (ETSI TS 126 171 [2]): 12,2 kbit/s.
- Recommendation ITU-T G.729.1 [11]: 32 kbit/s.

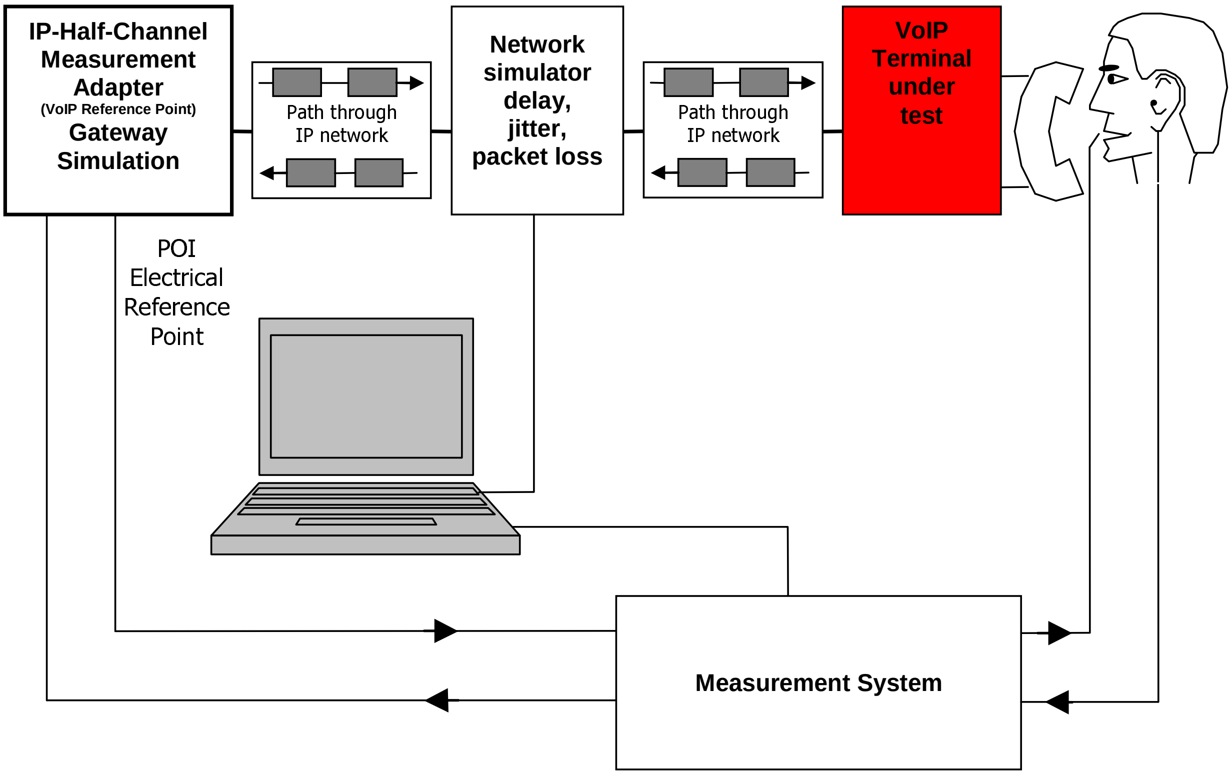


Figure 1: Half channel terminal measurement

6.2.2 Setup for handsets and headsets

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

Recommendations for positioning headsets are given in Recommendation ITU-T 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 Recommendation ITU-T 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 8 N is used for handset testing. No application force is used for headsets.

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

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

NOTE: The inverse average diffuse field response characteristics of HATS as found in Recommendation ITU-T P.58 [14] is used and not the specific one corresponding to the HATS used. Instead of using the individual diffuse field correction, the average correction function is used because, for handset and headset measurements, mostly the artificial ear, ear canal and ear impedance simulation are effective. The individual diffuse-field correction function of HATS includes all diffraction and reflection effects of the complete individual HATS which are not effective in the measurement and potentially would lead to bigger measurement uncertainties than using the average correction.

6.2.4 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.

6.2.5 Setup of background noise simulation

A setup for simulating realistic background noises in a lab-type environment is described in ETSI TS 103 224 [25].

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

The following noises of ETSI TS 103 224 [25] shall be used.

Table 2a

Pub Noise (Pub)	HATS and microphone array in a pub	30 seconds	1: 77,2 dB 2: 76,6 dB 3: 75,7 dB 4: 76,0 dB 5: 76,0 dB 6: 76,3 dB 7: 76,0 dB 8: 76,4 dB
Sales Counter (SalesCounter)	HATS and microphone array in a supermarket	30 seconds	1: 66,6 dB 2: 66,1 dB 3: 65,7 dB 4: 66,5 dB 5: 66,3 dB 6: 66,8 dB 7: 66,6 dB 8: 67,1 dB
Callcenter 2 (Callcenter)	HATS and microphone array in business office	30 seconds	1: 60,2 dB 2: 60,0 dB 3: 60,1 dB 4: 60,8 dB 5: 60,2 dB 6: 60,6 dB 7: 60,2 dB 8: 60,7 dB

6.3 Coding independent parameters

6.3.1 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 2. This mask shall be applicable for all types of handsets and headsets.

Table 3: Send frequency response

Frequency	Upper Limit	Lower Limit
100 Hz	-10 dB (see notes 2 and 3)	
300 Hz	5 dB	-5 dB
3 400 Hz	5 dB	-5 dB
4 000 Hz	5 dB	
NOTE 1: The limits for intermediate frequencies lie on a straight line drawn between the given values on a linear (dB) - logarithmic (Hz) scale. NOTE 2: Under conditions of high background noise a limit of -18 dB is recommended. NOTE 3: In ETSI ES 202 739 [24], the limit is 0 dB.		

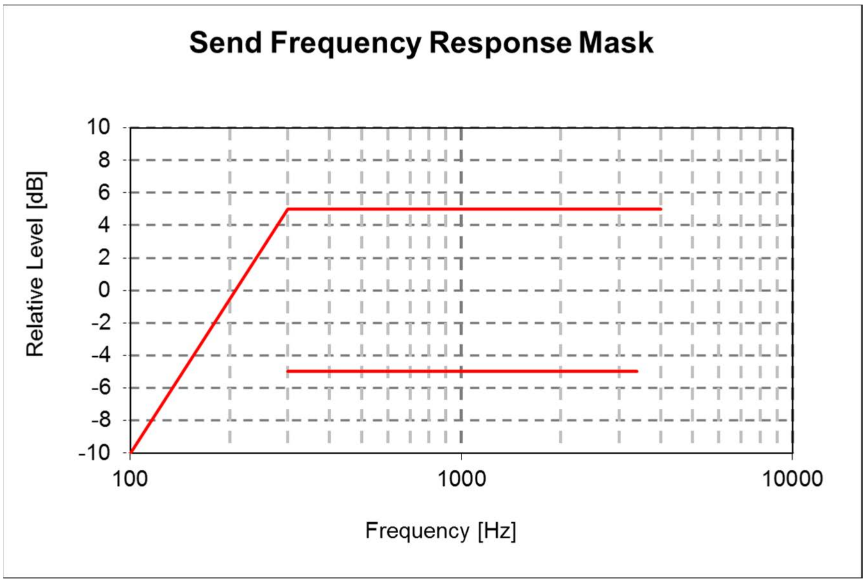


Figure 2: Send frequency response mask

NOTE: 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.

Measurement method

The test signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [19]. 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 seconds (10 seconds female, 10 seconds 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 6.2. The handset is mounted in the HATS position (see Recommendation ITU-T 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 Recommendation ITU-T P.380 [18] the results are averaged (averaged value in dB, for each frequency).

Measurements shall be made at one twelfth-octave bands as given by the IEC 61260-1 [22] for frequencies from 100 Hz to 4 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.

6.3.2 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 British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [19]. 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 6.2. The handset is mounted in the HATS position (see Recommendation ITU-T 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 Recommendation ITU-T 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 14 frequencies given in table 1 of Recommendation Recommendation ITU-T P.79 [16], bands 4 to 17. 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 Recommendation ITU-T P.79 [16], formula 5-1, over bands 4 to 17, using m = 0,175 and the send weighting factors from Recommendation ITU-T P.79 [16], table 1.

6.3.3 Mic mute

Requirement

The SLR (Send Loudness Rating) with mic mute on shall be 50dB higher than with mic mute off.

Measurement method

The test signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [19]. 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 6.2. The handset is mounted in the HATS position (see Recommendation ITU-T 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 Recommendation ITU-T 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 14 frequencies given in table 1 of Recommendation TU-T P.79 [16], bands 4 to 17. 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 Recommendation ITU-T P.79 [16], formula 5-1, over bands 4 to 17, using m = 0,175 and the send weighting factors from Recommendation ITU-T P.79 [16], table 1.

6.3.4 Linearity range for SLR

Requirement

The sensitivity determined with input sound pressure levels between -24,7 dBPa and 5,3 dBPa the limits as described in table 4 apply.

Table 4: Linearity range of SLR: ΔSLR = SLR - SLR@-4,7 dBPa

Input Level	Target ΔSLR	Upper limit	Lower limit
-24,7 dBPa	0	2,00 dB	-2 dB
-19,7 dBPa	0	2,00 dB	-2 dB
-14,7 dBPa	0	2,00 dB	-2 dB
-9,7 dBPa	0	2,00 dB	-2 dB
-4,9 dBPa	0	2,00 dB	-2 dB
-4,7 dBPa	0	0 dB	0,00 dB
-4,5 dBPa	0	2,00 dB	-2,00 dB
0,3 dBPa	0	2,00 dB	-2,00 dB
5,3 dBPa	0	4,00 dB	-4,00 dB
NOTE: The limits for intermediate levels lie on a straight line drawn between the given values on a linear (dB) - linear (dB) scale.			

NOTE: It is assumed that the variation of gain is mostly codec independent. In case codec specific requirements are needed this is found in the codec specific section.

Measurement method

The test signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [19]. 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, duration 20 seconds (10 seconds female, 10 seconds male) 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 6.2. The handset is mounted in the HATS position (see Recommendation ITU-T 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 14 frequencies given in table 1 of Recommendation Recommendation ITU-T P.79 [16], bands 4 to 17. 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 Recommendation ITU-T P.79 [16], formula 5-1, over bands 4 to 17, using m = 0,175 and the send weighting factors from Recommendation ITU-T P.79 [16], table 1.

6.3.5 Send distortion

Requirement

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
NOTE: Limits at intermediate frequencies lie on a straight line drawn between the given values on a linear (dBratio) - logarithmic (frequency) scale.	

Measurement method

The terminal will be positioned as described in clause 6.2.

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 and 1 000 Hz. The duration of the sine wave shall be less than 1 second. 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 3,15 kHz.

The female speaker signal of the short conditioning sequence described in clause 7.3.7 of Recommendation ITU-T P.501 [19] shall be used for activation. The level of this activation signal is -4,7 dBPa at the MRP.

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

6.3.6 Out-of-band signals in send direction

Requirement

With any signal above 4,6 kHz and up to 8 kHz applied at the MRP at a level of -4,7 dBPa, the level of any image frequency shall be below the level obtained for the reference signal by at least the amount (in dB) specified in table 6.

Table 6: Out-of-band signal limit, send

Frequency	Minimum attenuation
4,6 kHz	30 dB
8 kHz	40 dB
NOTE: The limits for intermediate frequencies lie on a straight line drawn between the given values on a linear (dB) - logarithmic (kHz) scale.	

Measurement method

The terminal will be positioned as described in clause 6.2.

For a correct activation of the system, the female speaker signal of the short conditioning sequence described in clause 7.3.7 of Recommendation ITU-T P.501 [19] shall be used for activation. Level of this activation signal shall be -4,7 dBpa at the MRP.

For the test, an out-of-band signal shall be provided as a frequency band signal centred on 4,65 kHz, 5 kHz, 6 kHz, 6,5 kHz, 7 kHz and 7,5 kHz respectively. The level of any image frequencies at the digital interface shall be measured.

The levels of these signals shall be -4,7 dBPa at the MRP.

The complete test signal is constituted by t1 ms of in-band signal (reference signal), t2 ms of out-of-band signal and another time t1 ms of in-band signal (reference signal).

The observation of the output signal on the first and second in-band signals permits control if the set is correctly activated during the out-of-band measurement. This measurement shall be performed during t2 period.

A value of 250 ms is suggested for t1.

T2 depends on the integration time of the analyser, typically less than 150 ms.

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

6.3.7 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 -64 dBm0p.

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 the female speaker of the short conditioning sequence described in clause 7.3.7 of Recommendation ITU-T 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

The handset terminal is set-up as described in clause 6.2. The handset is mounted at the HATS position (see Recommendation ITU-T P.64 [15]).

The send noise is measured at the POI in the frequency range from 100 Hz to 4 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 second. 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(P).

Spectral peaks are measured in the frequency domain in the frequency range from 100 Hz to 3,4 kHz. The frequency spectrum of the idle channel noise is measured by a spectral analysis having a noise bandwidth of 8,79 Hz (determined using FFT 8 k samples/48 kHz sampling rate with Hanning window or equivalent). The idle channel noise spectrum is stated in dB. A smoothed average idle channel noise spectrum is calculated by a moving average (arithmetic mean) 1/3rd octave wide across the idle noise channel spectrum stated in dB (linear average in dB of all FFT bins in the range from 2^(-1/6)f to 2^(+1/6)f). Peaks in the idle channel noise spectrum are compared against a smoothed average idle channel noise spectrum.

6.3.8 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 shall 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 British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [19]. 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 headset terminal is setup as described in clause 6.2. The handset is mounted in the HATS position (see Recommendation ITU-T 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 at 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 bands as given by the IEC 61260-1 [22] for frequencies from 100 Hz to 8 kHz inclusive. For the calculation the averaged measured level at each frequency band (Recommendation ITU-T 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 Recommendation ITU-T P.79 [16], using m = 0,225 and the weighting factors of in table 3 of Recommendation ITU-T P.79 [16].

6.3.9 Sidetone delay

Requirement

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

Measurement method

The handset or headset terminal is setup as described in clause 6.2. The handset is mounted in the HATS position (see Recommendation ITU-T P.64 [15]).

The test signal is a CSS complying with Recommendation ITU-T 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 Recommendation ITU-T P.501 [19]. The level of the signal shall be -4,7 dBPa at the MRP.

The cross-correlation function Φxy(τ) 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:

$$\Phi_{xy}(\tau) = \frac{1}{T} \int_{t=\frac{-T}{2}}^{\frac{T}{2}} S_x(t) \cdot S_y(t+\tau) \qquad (1)$$

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:

$$H\{xy(\tau)\} = \sum_{u=-\infty}^{+\infty} \frac{\Phi_{xy}(u)}{\pi(\tau - u)} \qquad (2)$$

$$E(\tau) = \sqrt{\left[\Phi_{xy}(\tau)\right]^2 + \left[H\{xy(\tau)\}\right]^2} \qquad (3)$$

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

6.3.10 Terminal Coupling Loss weighted (TCLw)

Requirement

The TCLw shall be ≥ 46 dB for all settings of the volume control (if supplied).

NOTE: A TCLw ≥ 50 dB is recommended as a performance objective. Depending on the idle channel noise in the sending direction, it may not always be possible to measure an echo loss ≥ 50 dB.

Measurement method

The handset or the headset terminal is setup as described in clause 6.2. The handset is mounted in the HATS position (see Recommendation ITU-T 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 Recommendation ITU-T 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 the compressed real speech signal described in clause 7.3.3 of Recommendation ITU-T P.501 [19]. The signal level shall be -10 dBm0.

The TCLw is calculated according to Recommendation ITU-T G.122 [6], clause B.4 (trapezoidal rule). 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. The first 17,0 seconds of the test signal (6 sentences) are discarded from the analysis to allow for convergence of the acoustic echo canceller. The analysis is performed over the remaining length of the test sequence (last 6 sentences).

6.3.11 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 200 Hz to 4 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 the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [19] 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 Recommendation ITU-T 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 200 Hz to 4 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 3;

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

  1) the mouthpiece and earcap shall face towards the surface;

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

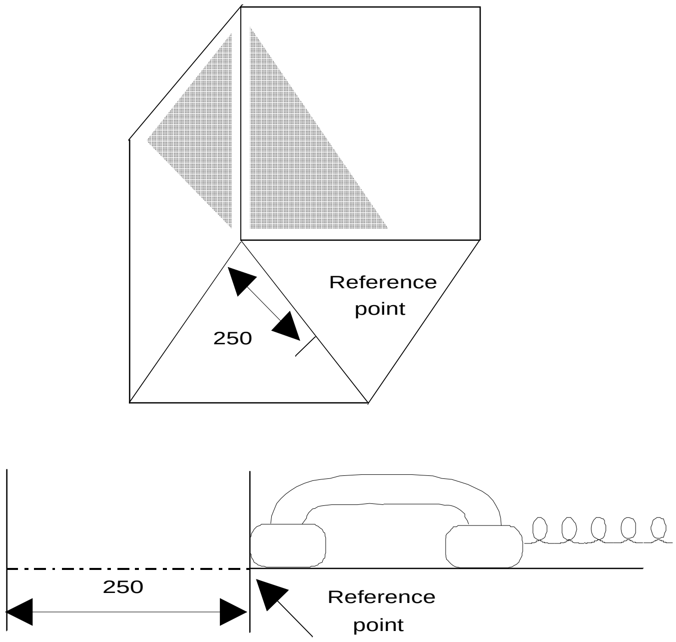
  3) the extremity of the handset shall coincide with the normal to the reference point, as shown in figure 3;

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 3;

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



NOTE: All dimensions in mm.

Figure 3

6.3.12 Receive frequency response

Requirement

The receive frequency response of the handset or the headset shall be within a mask as defined in table 7 and shown in figures 4 and 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 7: Receive frequency response mask

Frequency	Upper Limit 8 N	Lower Limit 8 N	Upper Limit 2 N and 13 N	Lower Limit 2 N and 13 N
100 Hz	4 dB		6 dB	
300 Hz	4 dB	-4 dB	6 dB	-6 dB
3 400 Hz	4 dB	-4 dB	6 dB	-6 dB
4 000 Hz	4 dB		6 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 diffuse field conditions and is assuming an ideal receive characteristic. This flat response characteristic 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 diffusefield based receive measurements a rising slope for the overall frequency response is achieved by a flat target frequency response in send and a diffuse field 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.

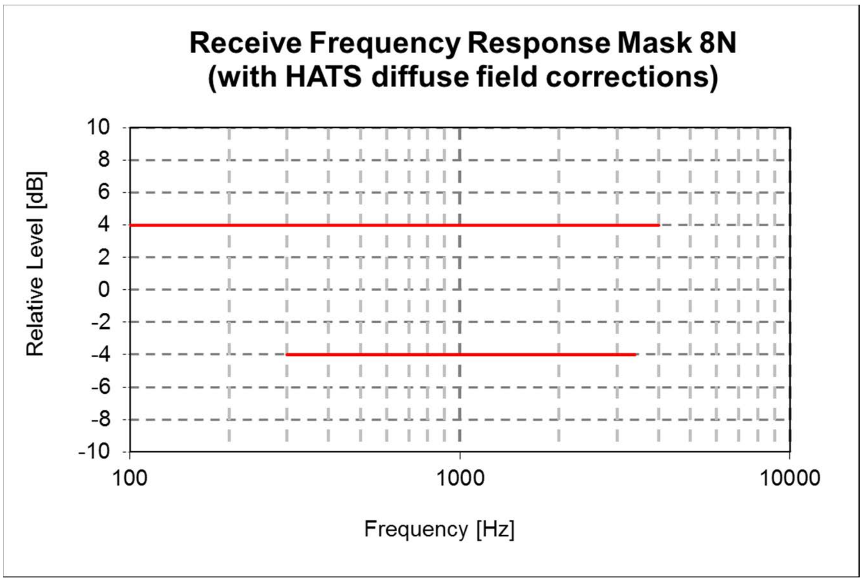


Figure 4: Receive frequency response mask for 8 N application force

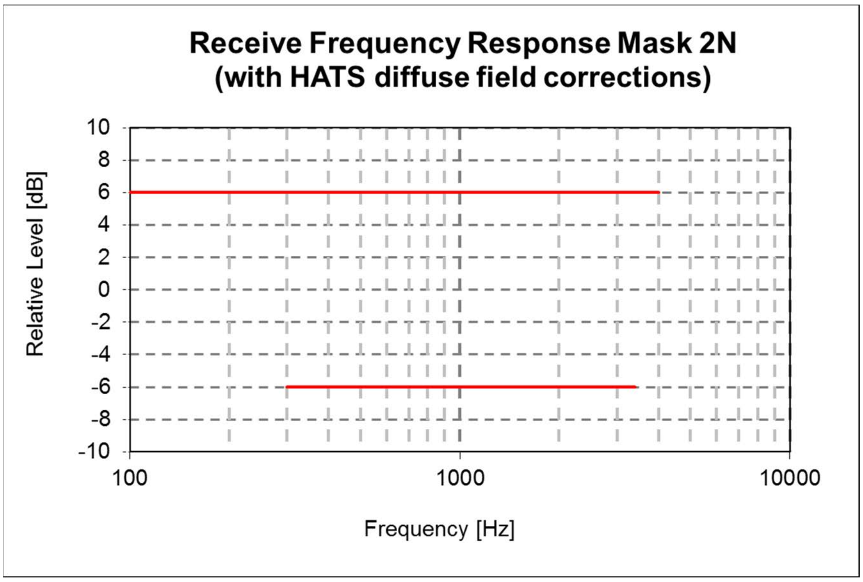


Figure 5a: Receive frequency response mask for 2 N application force

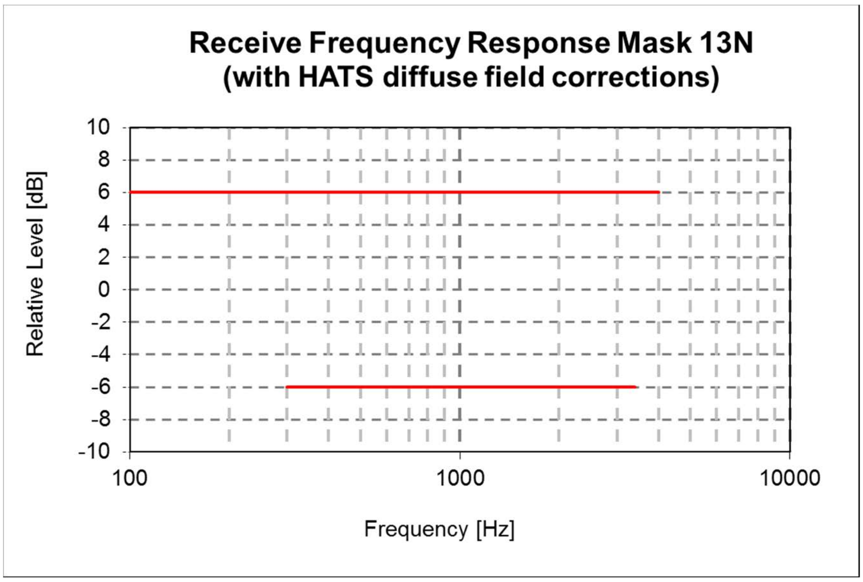


Figure 5b: Receive frequency response mask for 13 N application force

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 diffusefield correction.

pe_{ff} DRP Sound pressure measured by ear simulator Measurement data is 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 British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [19]. The test signal level shall be -16 dBm0, measured according to Recommendation ITU-T 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 6.2. The handset is mounted in the HATS position (see Recommendation ITU-T 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 Recommendation ITU-T P.380 [18] the results are averaged (averaged value in dB, for each frequency).

The HATS is diffusefield equalized as described in Recommendation ITU-T P.581 [21]. 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 bands as given by the IEC 61260-1 [22] for frequencies from 100 Hz to 4 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.

6.3.13 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 British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [19]. 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 or headset terminal is setup as described in clause 6.2. The handset is mounted in the HATS position (see Recommendation ITU-T 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** diffusefield equalized as described in Recommendation ITU-T P.581 [21]. The DRP-ERP correction as defined in Recommendation ITU-T 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 Recommendation ITU-T 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 14 frequencies given in table 1 of Recommendation ITU-T P.79 [16], bands 4 to 17. 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 Recommendation ITU-T P.79 [16], formula 5-1, over bands 4 to 17, using m = 0,175 and the receive weighting factors from table 1 of Recommendation ITU-T P.79 [16]. No leakage correction shall be applied for the measurement.

NOTE: Currently the Loudness Ratings Calculation is still based on the ERP. Therefore no diffusefield correction is applied and still the DRP-ERP correction is used.

6.3.14 Receive distortion

Requirement

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

Table 8

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
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 6.2.

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 and 1 000 Hz used for the measurements.

The signal level shall be -16 dBm0.

The female speaker signal of the short conditioning sequence described in clause 7.3.7 of Recommendation ITU-T P.501 [19] shall be used for activation.

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

The signal to harmonic distortion ratio is measured selectively up to 3,15 kHz.

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

6.3.15 Out-of-band signals in receive direction

Requirement

Any spurious out-of-band image signals in the frequency range from 4,6 kHz to 8 kHz measured selectively shall be lower than the in-band level measured with a reference signal. The minimum level difference between the reference signal level and the out-of-band image signal level shall be as given in table 9.

Table 9: Out of band signal limits, receive

Frequency	Minimum attenuation
4,6 kHz	35 dB
8 kHz	45 dB
NOTE: The limits for intermediate frequencies lie on a straight line drawn between the given values on a linear (dB) - logarithmic (kHz) scale.	

Measurement method

The handset terminal or the headset terminal is setup as described in clause 6.2.

The signal used is an activation signal followed by a sine wave signal. For input signals at the frequencies 500 Hz, 1 000 Hz, 2 000 Hz and 3 150 Hz applied at the level of -16 dBm0, the level of spurious out-of-band image signals at frequencies up to 8 kHz is measured selectively at measurement point.

The female speaker signal of the short conditioning sequence described in clause 7.3.7 of Recommendation ITU-T P.501 [19] shall be used for activation. The level of this activation signal is -16 dBm0. The out of band signal shall be measured at 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.

6.3.16 Minimum activation level and sensitivity in receive direction

For further study.

6.3.17 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.

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

Measurement method

The handset terminal or the headset terminal is setup as described in clause 6.2.

The A-weighted noise level shall be measured at DRP of the artificial ear with the diffusefield equalization active. The noise level is measured until 10 kHz.

The female speaker signal of the short conditioning sequence described in clause 7.3.7 of Recommendation ITU-T P.501 [19] shall be used for activation. The activation signal level shall be -16 dBm0.

Spectral peaks are measured in the frequency domain in the frequency range from 100 Hz to 3,4 kHz. The frequency spectrum of the idle channel noise is measured by a spectral analysis having a noise bandwidth of 8,79 Hz (determined using FFT 8 k samples/48 kHz sampling rate with Hanning window or equivalent). The idle channel noise spectrum is stated in dB. A smoothed average idle channel noise spectrum is calculated by a moving average (arithmetic mean) 1/3rd octave wide across the idle noise channel spectrum stated in dB (linear average in dB of all FFT bins in the range from 2^(-1/6)f to 2^(+1/6)f). Peaks in the idle channel noise spectrum are compared against a smoothed average idle channel noise spectrum.

6.3.18 Automatic level control in receive

For further study.

6.3.19 Double talk performance

6.3.19.1 General

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 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 Recommendations ITU-T 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.

6.3.19.2 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 10.

Table 10

Category (according to Recommendation ITU-T 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 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 the double talk speech sequence as defined in clause 7.3.5 of Recommendation ITU-T P.501 [19] as shown in figure 6. The competing speaker is always inserted as the double talk sequence sdt(t) either in send or receive and is used for analysis. The signal level in send direction shall be -4,7 dBPa, the signal level in receive direction shall be -16 dBm0.

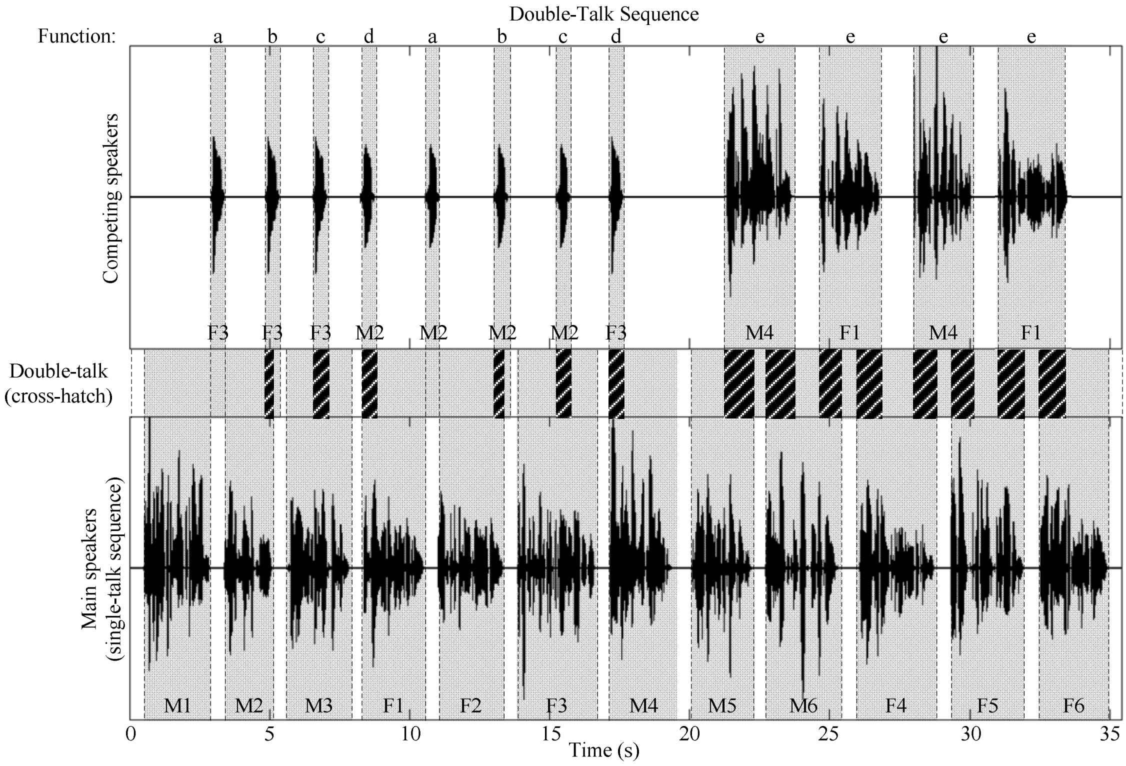


Figure 6: Double talk test sequence with overlapping speech sequences in send and receive direction

The attenuation range during double talk is determined as described in Appendix III of Recommendation ITU-T P.502 [20]. The double talk performance is analysed for each word and sentence produced by the competing speaker. The requirement has to be met for each word and sentence produced by the competing speaker.

6.3.19.3 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 11.

Table 11

Category (according to Recommendation ITU-T P.340 [17])	1	2a	2b	2c	3
	Full Duplex Capability	*Partial Duplex Capability*			*No Duplex Capability*
$A_{H,R,dt}$ *[dB]*	*≤ 3*	*≤ 5*	*≤ 8*	*≤ 10*	*> 10*

In general table 11 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 speech signal 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 signal level in send direction shall be -4,7 dBPa, the signal level in receive direction shall be -16 dBm0.

The test arrangement is according to clause 6.2.

The attenuation range during double talk is determined as described in Appendix III of Recommendation ITU-T P.502 [20]. The double talk performance is analysed for each word and sentence produced by the competing speaker. The requirement has to be met for each word and sentence produced by the competing speaker.

6.3.19.4 Detection of echo components during double talk

Requirement

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

NOTE 1: 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 Recommendation ITU-T P.340 [17]).

Table 12

Category (according to Recommendation ITU-T P.340 [17])	1	2a	2b	2c	3
	Full Duplex Capability	*Partial Duplex Capability*			*No Duplex Capability*
Echo Loss [dB]	*≥ 27*	*≥ 23*	*≥ 17*	*≥ 11*	*< 11*

Measurement method

The test arrangement is according to clause 6.2.

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 Recommendation ITU-T 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).

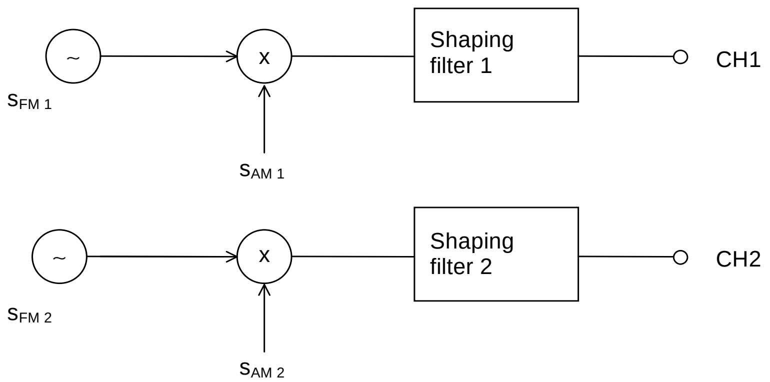


Figure 7: Measurement signals

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

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

NOTE 2: A is determined by the required test signal level as found in the individual test cases.

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

Send Direction				Receive Direction		
F_m(Hz)	$f_{mod(fm)}$ (Hz)	F_{am}(Hz)		F_m(Hz)	$f_{mod(fm)}$(Hz)	F_{am}(Hz)
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
NOTE: Parameters of the Shaping Filter: Low Pass Filter, 5 dB/oct.						

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 Recommendation ITU-T 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 11. The echo attenuation is to be achieved for **each individual frequency band** according to the different categories.

6.3.19.5 Minimum activation level and sensitivity of double talk detection

For further study.

6.3.20 Switching characteristics

6.3.20.1 Note

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

6.3.20.2 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).

Requirements

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 test signal is the short words for activation sequence described in clause 7.3.4 of Recommendation ITU-T P.501 [19] with increasing level for each single word.

The settings of the test signal are as follows.

Table 14

	Single word Duration/ Pause Duration	Level of the first single word (at the MRP)	Level Difference between two Periods of the Test Signal
Single word to determine switching characteristic **in send direction**	~600 ms/ ~400 ms	-24 dBPa (see note)	1 dB
NOTE: The signal level is determined for each utterance individually according to Recommendation ITU-T P.56 [12].			

It is assumed that the pause length of about 400 ms is longer than the hang-over time so that the test object is back to idle mode after each single word.

The test arrangement is described in clause 6.2.

The level of the transmitted signal is measured at the electrical reference point. The test signal is filtered by the transfer function of the test object. The measured signal level is referred to the filtered 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 single word which indicates the first activation of the test object. The time between the beginning of the single word and the complete activation of the test object is measured.

6.3.20.3 Silence suppression and comfort noise generation

For further study.

6.3.21 Background noise performance

6.3.21.1 Performance in send direction in the presence of background noise

Requirement

The level of comfort noise, if implemented, shall be within in a range of +2 dB and -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
NOTE: All sensitivity values are expressed in dB on an arbitrary scale.		

Measurement method

The background noise simulation as described in clause 6.2.5 is used.

The handset terminal is set-up as described in clause 6.2. The handset is mounted at the HATS position (see Recommendation ITU-T P.64 [15]).

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

In a second step a test signal is applied in receive direction consisting of an initial pause of 10 seconds and a periodical repetition of the Composite Source Signal in receive direction (duration 10 seconds) 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 seconds and 20 seconds 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 seconds and 20 seconds. Level and spectral differences between both power density spectra are analysed and compared to the requirements.

6.3.21.2 Speech quality in the presence of background noise

Requirement

Speech Quality for narrowband systems can be tested based on ETSI EG 202 396-3 [i.3]. 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-LQOn: Transmission quality of the background noise.
- S-MOS-LQOn: Transmission quality of the speech.
- G-MOS-LQOn: Overall transmission quality.

For the background noises defined in clause 6.2 the following requirements apply:

- N-MOS-LQOn ≥ 3,5.
- S-MOS-LQOn ≥ 3,5.
- G-MOS-LQOn ≥ 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 clause 6.2.

Measurement method

The background noise simulation as described in clause 6.2 is used. The handset terminal is set-up as described in clause 6.2.2. The handset is mounted at the HATS position (see Recommendation ITU-T P.64 [15]).

The background noise should be applied for at least 5 seconds 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 Recommendation ITU-T P.501 [19]. The preferred language is English since the objective method was validated with English language in narrowband. 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 ETSI EG 202 396-3 [i.3]).

2) The speech plus undisturbed background noise signal is recorded at the terminal's microphone position using an omni directional measurement microphone with a linear frequency response between 50 Hz and 6 kHz.

3) The send signal is recorded at the electrical reference point.

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

6.3.21.3 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 6.2.

NOTE: The intention of this measurement is to detect impairments (modulations, switching and others) influencing the background noise transmitted from the terminal under test when a signal from the distant end (receiving side of the terminal under test) is present. Under these test conditions no modulation of the transmitted signal should occur. Modulation, switching or other type of impairments might be caused by an improper behaviour of a nonlinear processor working in conjunction with the echo canceller and erroneously switching or modulating the transmitted background noise.

Measurement method

The test arrangement is according to clause 6.2.

The background noises are generated as described in clause 6.2.

First the measurement is conducted without inserting the signal at the far end. At least 10 seconds 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 CSS at the far end. The exactly identical background noise signal is applied. The background noise signal shall 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 seconds in order to allow adaptation of the noise reduction algorithms. After at least 5 seconds a Composite Source Signal according to Recommendation ITU-T 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 CSS 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.

6.3.22 Quality of echo cancellation

6.3.22.1 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 echo attenuation measured.

Measurement method

The test arrangement is according to clause 6.2.

The test signal consists of periodically repeated Composite Source Signal according to Recommendation ITU-T 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 seconds which represents 8 periods of the CSS. 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.

The difference between the maximum attenuation and the minimum attenuation is measured.

NOTE 1: In addition tests with more speech like signals should be made, e.g. Recommendation ITU-T P.501 [19] 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 (35 ms) of the level analysis taking into account the exponential character of the integration time in any tolerance scheme.

NOTE 3: Care should be taken not to confuse noise or comfort noise with residual echo. In cases of doubt the measured echo signal should be compared to the residual noise signal measured under the same conditions without inserting the receive signal. If the level vs. time analysis leads to the identical result it can be assumed that no echo but just comfort noise is present.

6.3.22.2 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	-20 dB
200 Hz	-30 dB
300 Hz	-38 dB
800 Hz	-34 dB
1 500 Hz	-33 dB
2 600 Hz	-24 dB
4 000 Hz	-24 dB
NOTE 1: All sensitivity values are expressed in dB on an arbitrary scale. NOTE 2: 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 6.2.

Before the actual measurement a training sequence consisting of the compressed real speech signal described in clause 7.3.3 of Recommendation ITU-T P.501 [19]. The level of the training sequence is -16 dBm0.

The test signal is the compressed real speech signal described in clause 7.3.3 of Recommendation ITU-T P.501 [19]. The measurement is carried out under steady-state conditions. The average test signal level is -16 dBm0, averaged over the complete test signal. 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.

6.3.22.3 Occurrence of artefacts

For further study.

6.3.22.4 Variable echo path

Requirement

This test is intended to verify that the system will maintain sufficient echo attenuation during single talk with dynamically changing echo paths. The measured echo level over time during single talk should not be more than 10 dB above the minimum noise level during the measurement.

Measurement method

The handset is positioned d = 3 cm above a horizontal hard surface, facing the surface with speaker and microphone. The surface shall be at least 35 x 35 cm. The handset is fixed like a pendulum with a non-elastic cord 3 cm above the centre of the horizontal surface, see figure 8. The pivot is 55 ± 1 cm above the hard plate.

Test setup for headsets: for further study.

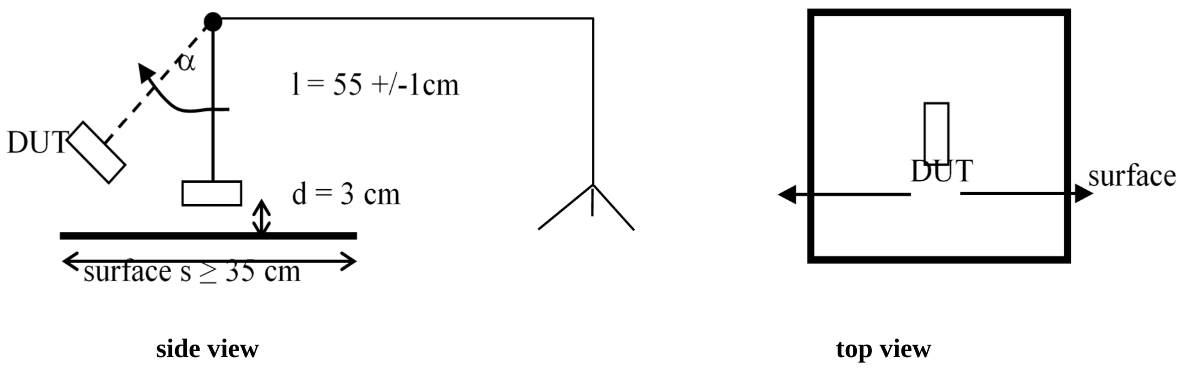


Figure 8: Positioning of handset under test

The "handset-pendulum" is displaced at least to the edge of the hard surface. The test signal playback shall start with the release of the displaced handset under test.

As test signal the compressed real speech signal described in clause 7.3.3 of Recommendation ITU-T P.501 [19] is used. The signal level shall be -10 dBm0. The terminal volume control is set to nominal RLR. The first 4 sentences of the test signal are used to allow full convergence of the echo canceller. The next 4 sentences (from 10,75 s to 22,5 s) are used for the analysis. The echo signal level is analysed over time. The echo signal level is analysed for 11,75 s, using a time constant of 35 ms.

The measurement result is displayed as echo level versus time.

No level peak should be more than 10 dB above the minimum noise level during the measurement.

6.3.23 Variant impairments; network dependant

6.3.23.1 Clock accuracy send

Requirement

The clock accuracy in send direction between the VoIP-Terminal and the IP reference interface shall be less than 150 ppm under ideal network conditions.

NOTE: The clock accuracy does not cover all possible network configurations. Especially it is not sufficient for data transmission or distributed TDM PBX where synchronization is required.

Measurement method

A sequence of CS signals (active signal length = 250 ms) is repeated for 120 s in order to analyse clock accuracy and any other time-variant delay. The pause length between two CS bursts is 100 ms and 1,2 s after every fourth burst in order to simulate a speech pause, which may lead to buffer adjustments. The test signal level shall be -4,7 dBPa at the MRP.

A cross correlation analysis versus time is carried out over the whole 120 s sequence between the received and the original test signal. The duration of the measurement (120 s) is indicated on the x-axis, the result of the cross correlation analysis (delay) is plotted on the y-axis.

The resulting clock accuracy within an analysis time range of at least 60 s is calculated as follows:

$$clockaccuracy[ppm] = \frac{delaychange[s]}{analysisduration[s]} \cdot 1 \cdot 10^6 \tag{7}$$

6.3.23.2 Clock accuracy receive

Requirement

The clock accuracy in receive direction between the IP reference interface and the VoIP-Terminal shall be less than 150 ppm under ideal network conditions.

Measurement method

A sequence of CS signals (active signal length = 250 ms) is repeated for 120 s in order to analyse clock accuracy and any other time-variant delay. The pause length between two CS bursts is 100 ms and 1,2 s after every fourth burst in order to simulate a speech pause, which may lead to buffer adjustments. The test signal level at the IP reference interface shall be -16 dBm0.

A cross correlation analysis versus time is carried out over the whole 120 s sequence between the received and the original test signal. The duration of the measurement (120 s) is indicated on the x-axis, the result of the cross correlation analysis (delay) is plotted on the y-axis.

The resulting clock accuracy within an analysis time range of at least 60 s is calculated as follows:

$$clockaccuracy[ppm] = \frac{delaychange[s]}{analysisduration[s]} \cdot 1 \cdot 10^6 \tag{8}$$

6.3.23.3 Send delay variation

Requirement

The measured maximum delay variation in send direction of the VoIP-terminal under test should be less than 1 ms.

NOTE: Any delay variation introduced in send direction will lead to potentially increased delay due to increased de-jitter buffer at the far end terminal.

Measurement method

The RTP data stream in send direction should be monitored with a tap or a switch providing a monitoring port, positioned at the location of the network impairment simulator (see clause 6.2). The test arrangement is according to clause 6.2.

The monitoring time should be 60 s. A signal like the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [19] s played back in send direction using a nominal level of -4,7 dBPa at the MRP. This speech signal is only necessary to make sure, RTP is played out, even in the case VAD is active.

The delay variation for each packet D(i) is evaluated according to IETF RFC 3550 [29]:

$$d(i) = \Delta t_{eff(i)} - \Delta t_{exp(i)}$$

$$D(i) = (15 * D(i-1) + | d(i) |) / 16 \tag{9}$$

With:

- $\Delta t_{exp(i)}$ = the expected time between packet i and packet i-1; and
- $\Delta t_{eff(i)}$ = the effective time between packet i and packet i-1.

Maximum delay variation = MAX(D(i)).

6.3.24 Send and receive delay - round trip delay

The roundtrip delay of a VoIP-terminal is defined as the sum of send and receive delays. In the following clauses the calculation of the requirements for send and receive delay are explained. For a telecommunication connection, only the roundtrip delay can be experienced. For this reason, also the requirement for VoIP-terminals is given only for the roundtrip delay. As long as the measured roundtrip delay fulfils the requirements, send or receive delays may be above the theoretical requirements.

Requirement

It is recognized that the end to end delay should be as small as possible in order to ensure high quality of the communication.

The roundtrip delay of the VoIP-terminal T_{rtd} (sum of receive and send delay) shall be less than 100 ms. (category B in Recommendation ITU-T P-1010 [28]). From the users perspective, a value less than 50 ms (category A in Recommendation ITU-T P-1010 [28]) is preferred.

NOTE 1: The limit for the roundtrip delay T_{rtd} of the VoIP-terminal is derived from the sum of the send and receive delay limits.

NOTE 2: This requirement is 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

- **Send direction**

The delay in send direction is measured from the MRP to POI. The delay measured in send direction is:

$$Ts + t_{System}$$

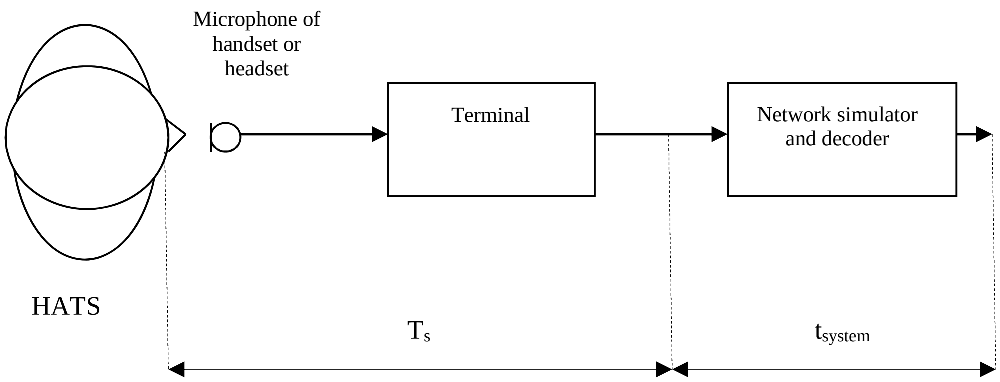


Figure 9: Different blocks contributing to the delay in send direction

The system delay t_{System} is depending on the transmission method used and the network simulator. The delay t_{system} shall be known:

1) For the measurements a Composite Source Signal (CSS) according to Recommendation ITU-T P.501 [19] is used. The pseudo random noise (pn)-part of the CSS has to be longer than the maximum expected delay. It is recommended to use a pn sequence of 16 k samples (with 48 kHz sampling rate). The test signal level is -4,7 dBPa at the MRP.

   The reference signal is the original signal (test signal).

   The setup of the handset/headset terminal is in correspondence to clause 6.2.

2) The delay is determined by cross-correlation analysis between the measured signal at the electrical access point and the original signal. The measurement is corrected by delays which are caused by the test equipment.

3) The delay is measured in ms and the maximum of the cross-correlation function is used for the determination.

- **Receive direction**

The delay in receive direction is measured from POI to the Drum Reference Point (DRP). The delay measured in receive direction is:

$$Tr + t_{System}$$

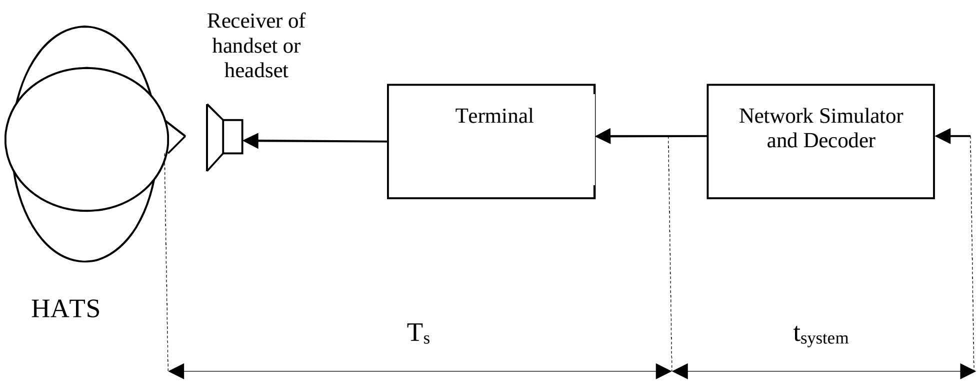


Figure 10: Different blocks contributing to the delay in receive direction

The system delay t_{System} is depending on the transmission system and on the network simulator used. The delay t_{System} shall be known:

1) For the measurements a Composite Source Signal (CSS) according to Recommendation ITU-T P.501 [19] is used. The pseudo random noise (pn)-part of the CSS has to be longer than the maximum expected delay. It is recommended to use a pn sequence of 16 k samples (with 48 kHz sampling rate). The test signal level is -16 dBm0 at the electrical interface (POI).

The reference signal is the original signal (test signal).

2) The test arrangement is according to clause 6.2.

3) The delay is determined by cross-correlation analysis between the measured signal at the DRP and the original signal. The measurement is corrected by delays which are caused by the test equipment.

4) The delay is measured in ms and the maximum of the cross-correlation function is used for the determination.

6.4 Codec specific requirements

6.4.1 Objective listening speech quality MOS-LQO 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 17

Speech coder	MOS-LQON (P.863 or TOSQA 2001)	MOS-LQOM (TOSQA 2001)
Recommendation ITU-T G.711 [7]	> 4,2	> 3,6
Recommendation ITU-T G.729 [10]	> 3,8	> 3,1
Recommendation ITU-T G.723.1 [8]	> 3,5	> 2,9
Recommendation ITU-T G.726 @ 32 kbit/s [9]	> 3,9	> 3,3
GSM EFR [1] and AMR @ 12,2 kbit/s [2]	> 4,0	> 3,4
Recommendation ITU-T G.729.1 @ 8 kbit/s [11]	> 3,8	> 3,1

NOTE 1: In narrowband acoustics, Recommendation ITU-T P.863 [26] is recommending using the superwideband mode with a narrowband reference signal, resulting in a prediction on the narrowband scale.

NOTE 2: The use of the codecs G.723.1 [8], G.729 [10] and G.729.1 [11] is not recommended due to low quality,

Measurement method

Objective listening speech quality is measured using Recommendation ITU-T P.863 [26] in superwideband mode.

The test signal to be used for the measurements shall be 4 sentence pairs (male/female) fulfilling the requirements of Recommendation ITU-T P.863.1 [27]. The 4 sentence pairs are taken from Recommendation ITU-T P.501 [19]. It shall be stated, which sentence pairs were used. The test signal level is averaged over all sentence pairs (4 sentence pairs). The measurement is done 4 times, every time using another pair of the speech sentences. The result of the measurement is the averaged value of all 4 measurements.

NOTE 3: An alternative test method is TOSQA 2001 (ETSI EG 201 377-1 [i.1]). With TOSQA, terminals used in narrowband mode only should be measured based on MOS-LQON. Terminals used in narrowband and wideband mode should be measured based on MOS-LQOM.

6.4.2 Objective listening quality MOS-LQO 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.

Requirements

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

Table 18

Speech coder	MOS-LQON (P.863 or TOSQA 2001)	MOS-LQOM (TOSQA 2001)
Recommendation ITU-T G.711 [7]	> 4,0	> 3,5
Recommendation ITU-T G.729 [10]	> 3,4	> 3,0
Recommendation ITU-T G.723.1 [8]	> 3,3	> 2,7
Recommendation ITU-T G.726 @ 32 kbit/s [9]	> 3,7	> 3,1
GSM EFR [1] and AMR @ 12,2 kbit/s [2]	> 3,8	> 3,2
Recommendation ITU-T G.729.1 @ 8 kbit/s [11]	> 3,6	> 2,9

NOTE 1: The MOS-LQO 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.

NOTE 2: In narrowband acoustics, Recommendation ITU-T P.863 [26] is recommending using the superwideband mode with a narrowband reference signal, resulting in a prediction on the narrowband scale.

NOTE 3: The use of the codecs G.723.1 [8], G.729 [10] and G.729.1 [11] is not recommended due to low quality.

Measurement method

Objective listening speech quality is measured using Recommendation ITU-T P.863 [26] in superwideband mode.

The test signal to be used for the measurements shall be 4 sentence pairs (male/female) fulfilling the requirements of Recommendation ITU-T P.863.1 [27]. The 4 sentence pairs are taken from Recommendation ITU-T P.501 [19]. It shall be stated, which sentence pairs were used. The test signal level is averaged over all sentence pairs (4 sentence pairs). The measurement is done 4 times, every time using another pair of the speech sentences. The result of the measurement is the averaged value of all 4 measurements.

NOTE 4: An alternative test method is TOSQA 2001 (ETSI EG 201 377-1 [i.1]). With TOSQA, terminals used in narrowband mode only should be measured based on MOS-LQON. Terminals used in narrowband and wideband mode should be measured based on MOS-LQOM.

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

Table 19: Network conditions for electrical-acoustical measurements (speech samples)

<table>
<tr><th>Condition</th><th>Packet Loss (Equal)</th><th>Delay Variation</th></tr>
<tr><td>0 (see note 2)
(VAD)</td><td>0</td><td>No</td></tr>
<tr><td>1</td><td>0</td><td>No</td></tr>
<tr><td>2</td><td>0</td><td>20 ms (see note 1)</td></tr>
<tr><td>3</td><td>1 %</td><td>No</td></tr>
<tr><td>4</td><td>1 %</td><td>20 ms (see note 1)</td></tr>
<tr><td>5</td><td>3 %</td><td>No</td></tr>
<tr><td colspan="3">NOTE 1: Delay variation produced with a Pareto-Distribution and r = 0,5.
NOTE 2: VAD on, all other conditions (1-5) 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.
NOTE 4: The settings are derived from the ones used in the ETSI Plugtest VoIP speech quality test events.</td></tr>
</table>

Table 20: Requirements for G.711 speech codecs

Condition	MOS-LQON (P.863 or TOSQA 2001)	MOS-LQOM (TOSQA 2001)	Delay
0	> 3,9	> 3,4	< 31 ms
1	> 4,0	> 3,5	< 31 ms
2	> 3,8	> 3,3	< 51 ms
3	> 3,8	> 3,3	< 31 ms
4	> 3,8	> 3,3	< 51 ms
5	> 3,8	> 3,3	< 31 ms

Table 21: Requirements for G.729 speech codecs

Condition	MOS-LQON (P.863 or TOSQA 2001)	MOS-LQOM (TOSQA 2001)	Delay
1	> 3,4	> 3,0	< 30 ms
2	> 3,3	> 2,9	< 50 ms
3	> 3,3	> 2,9	< 30 ms
4	> 3,3	> 2,9	< 50 ms
5	> 2,9	> 2,4	< 30 ms

Table 22: Requirements for G.723.1 speech codecs

Condition	MOS-LQON (P.863 or TOSQA 2001)	MOS-LQOM (TOSQA 2001)	Delay
1	> 3,3	> 2,7	< 50 ms
2	> 3,0	> 2,5	< 70 ms
3	> 3,0	> 2,5	< 50 ms
4	> 3,0	> 2,5	< 70 ms
5	> 2,9	> 2,4	< 50 ms

NOTE 5: In narrowband acoustics, Recommendation ITU-T P.863 [26] is recommending using the superwideband mode with a narrowband reference signal, resulting in a prediction on the narrowband scale.

NOTE 6: An alternative test method is TOSQA 2001 (ETSI EG 201 377-1 [i.1]). With TOSQA, terminals used in narrowband mode only should be measured based on MOS-LQON. Terminals used in narrowband and wideband mode should be measured based on MOS-LQOM.

6.4.3 Quality of jitter buffer adjustment

Requirements

The speech quality during and after inserted IP delay variation shall be as follows:

Table 23: Requirements for variant network impairments

Codec	MOS-LQON
G.711	> 3,9
G.729	> 3,4
G.723.1	> 3,1

The delay measured 20 seconds after ending of the IP delay variation shall be max. 10 ms higher than the delay measured before the IP delay variation.

Measurement method

The test signal consists of a CSS-signal, followed by 5 times the same speech sentence, fulfilling the requirements of Recommendation ITU-T P.863.1 [27], then again a CSS signal (20 seconds after the IP delay variation stops). The speech signal level is averaged over all used (original) sentences (8 sentences).

NOTE 1: The 8 used sentences consist of the 8 single sentences taken from the 4 sentence pairs used in clauses 6.4.1 and 6.4.2.

NOTE 2: For every new measurement a new call has to be setup to start with an initial delay. Depending on the algorithm used in the variable jitter buffer (e.g. jitter buffer starting with a high fill size), it may be necessary to let some time pass under clean conditions until the measurement is started.

The first CSS signal is used to measure the delay prior to the IP impairment (in clean network conditions). The second CSS signal is used to measure the delay 20 seconds after the IP impairment stops. The difference of the two delays is the measurement result for the variation of the jitter buffer per measurement. The overall result is the average of all 8 measurements.

The first sentence (during which IPDV of 50 ms is applied) is used to measure the speech quality during jitter buffer adaption (low to high). MOS-LQON of the first sentence is measured using Recommendation ITU-T P.863 [26] in superwideband mode. The overall result is the average MOS-LQOS of the 8 measurements.

The second to the fifth sentence (every 5 seconds a sentence) are used to measure the speech quality during jitter buffer adaption (high to low). MOS-LQON is measured using Recommendation ITU-T P.863 [26] in superwideband mode for each of these four sentences. The minimum MOS-LQON of these four sentences is used for the averaging over all 8 measurements. The overall result for the speech quality during jitter buffer adaption (high to low) is the average of the minimum MOS-LQON-value of the 8 measurements.

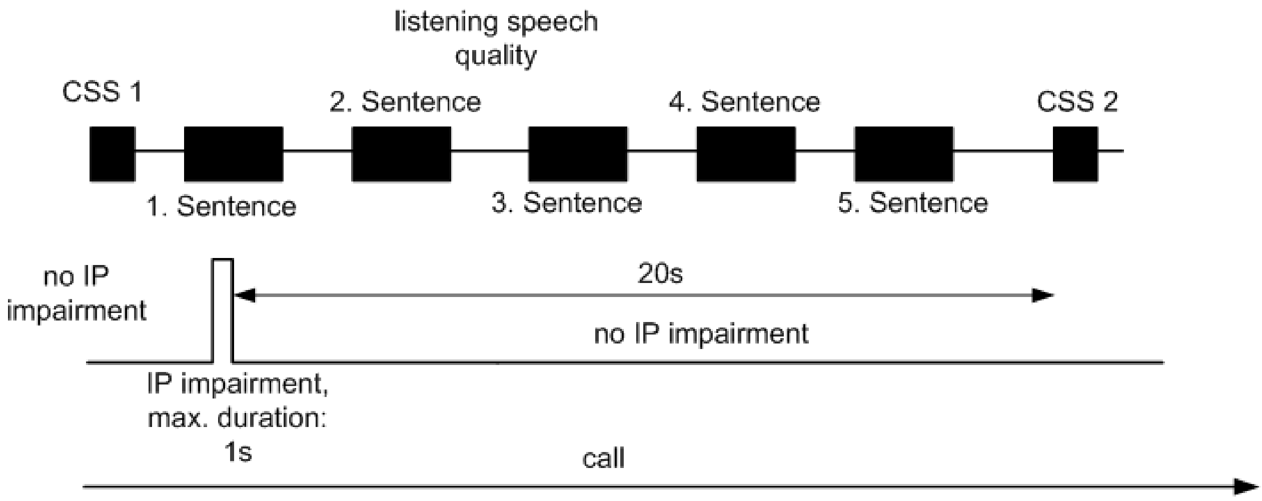


Figure 11: Test sequence to measure quality of jitter buffer adjustment (with 1 of 8 sentences)

The IP impairment consists of additional packet delay (IPDV) up to 50 ms, during maximum 1 second. The impairment can be in form of jitter, but also with only some single packets delayed. An example for the impairment can be found in annex B.

NOTE 3: Care should be given, that no packet reordering occurs (this could happen if e.g. one packet is delayed by 50 ms and the next one is not delayed, they will change order, which will not happen in real networks except in a failover situation or with bad implementations of load balancing).

Annex A (informative): Processing delays in VoIP terminals

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

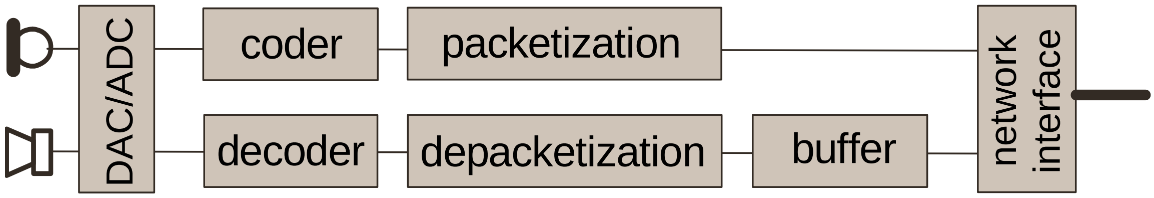


Figure A.1: Synoptic of the different functions implemented in a VoIP terminal

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

- The analogue-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-analogue 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 no 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.

Analogue to digital conversion consists in transforming speech signal from analogue to digital format. The processing delay associated to the conversion is considered as no significant.

Digital to analogue conversion consists in transforming speech signal from digital to analogue format. As analogue to digital conversion, the processing delay associated to digital to analogue 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 no 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 this table, 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. no 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): Example IP delay variation

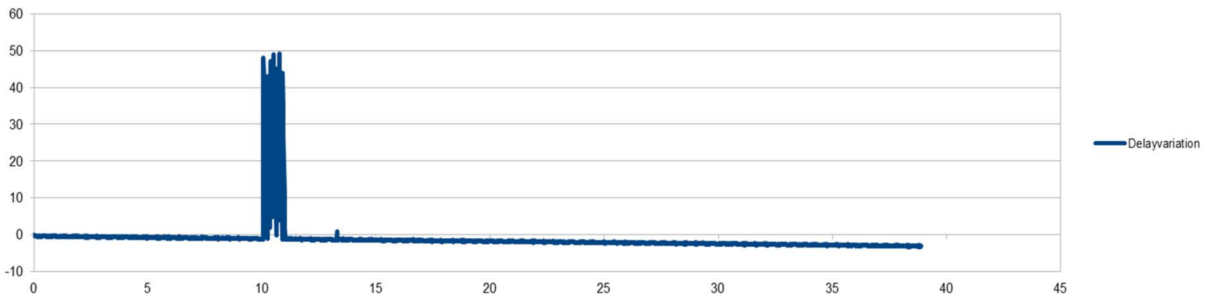


Figure B.1: Example IP delay variation (IPDV) introduced for the measurements shown later in the present document (IPDV in ms vs. time in seconds)

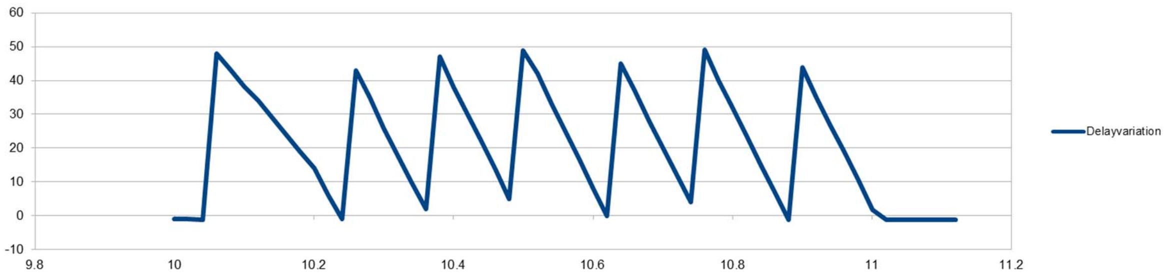


Figure B.2: Zoom in into the IP delay variation part (IPDV in ms vs. time in seconds)

Annex C (informative): Bibliography

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

History

Document history		
V1.2.1	October 2007	Publication
V1.3.1	September 2009	Publication
V1.3.2	September 2010	Publication
V1.4.1	March 2015	Publication
V1.5.1	October 2016	Membership Approval Procedure MV 20161226: 2016-10-27 to 2016-12-26