ETSI TS 103 801 V1.1.1 (2020-11)

Speech and multimedia Transmission Quality (STQ); Subjective test methodologies for the evaluation of echo control systems

Reference
DTS/STQ-285

Keywords
conversation, double talk, echo, impairment,
listening quality, test

ETSI

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

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

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

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables 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.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

Foreword

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

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

In speech communication devices of all kinds, echo artefacts and double talk impairments can occur. These might dramatically degrade a conversation between users, i.e. the quality of experience in general. With an increasing usage of hands-free terminals (e.g. motor vehicle, handheld or desktop devices) and new types of devices supporting voice services (e.g. smart home devices or wearables), the cancellation of echo and providing duplex communication at the same time is still a challenging task for signal processing components.

The objective assessment of degradations caused by echo and/or poor double talk performance is already covered in several specifications, but mainly based on simple analyses in level or spectrum. The impact on the conversation as perceived by the user is typically rarely investigated.

The auditory evaluation of a conversation between two human test subjects in a laboratory may be quite cumbersome. Even though some listening test specifications already exist in several standardization bodies for these scenarios, the reproducibility of results may vary a lot due to several degrees of freedom, e.g. a randomly degraded communication channel or usage of free speech.

The present document provides a subjective test framework for the evaluation of echo artefacts and double talk impairments, based on the Third-Party Listening Test (TPLT) approach. On one hand, a conversation is simulated as close as possible to human perception, in particular including the acoustics of involved terminals as well as self-hearing and self-masking in talking phases. On the other hand, the proposed test methodology utilizes pre-recorded signals, designed with respect to best-possible reproducibility in listening labs. This approach is well known from classical subjective evaluations of speech, audio and/or video. This leads to a decreased naturalness and spontaneity compared to a real conversation between subjects. However, the compromise between these two opposite approaches provides a wider range of use cases. In addition, the signals used for subjective testing may be re-used for predictive models.

1 Scope

The present document provides a framework for auditory testing of echo artefacts and double talk impairments that may occur in telecommunication devices of all kind.

The present document assesses degradations in end-to-end scenarios as perceived by the listener at the reference-side. Only degradations caused by the terminal located at the device-side are taken into account by the framework. Since the network delay between reference-side and device-side (and vice-versa) also has an impact on the DUT's signal processing and/or the listener's quality of experience, this parameter is included in the present document as well - any other degradations (e.g. packet-loss in one of the two directions) are out of scope.

Only DCR scales are supported in the auditory test, in particular for echo artefacts and double talk disturbances, which have the most impact on conversations (more may be added in the future). ACR scales e.g. speech distortion or overall quality are not considered for auditory testing.

Any instrumental model predicting results according to the introduced listening test design is out of scope.

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 https://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] Recommendation ITU-T P.10/G.100: "Vocabulary for performance, quality of service and quality of experience".

[2] Recommendation ITU-T P.800: "Methods for subjective determination of transmission quality".

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

[4] ITU-T Handbooks: "Handbook of subjective testing practical procedures".

[5] Recommendation ITU-T P.805: "Subjective evaluation of conversational quality".

[6] Recommendation ITU-T P.700: "Calculation of loudness for speech communication".

[7] ETSI TS 103 737: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for narrowband wireless terminals (handset and headset) from a QoS perspective as perceived by the user".

[8] ETSI TS 103 738: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for narrowband wireless terminals (handsfree) from a QoS perspective as perceived by the user".

[9] ETSI TS 103 739: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband wireless terminals (handset and headset) from a QoS perspective as perceived by the user".

[10] ETSI TS 103 740: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband wireless terminals (handsfree) from a QoS perspective as perceived by the user".

[11] ETSI TS 102 924: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for Super-Wideband / Fullband handset and headset terminals from a QoS perspective as perceived by the user".

[12] ETSI TS 102 925: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for Super-Wideband / Fullband handsfree and conferencing terminals from a QoS perspective as perceived by the user".

[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] ETSI TS 126 132 "Universal Mobile Telecommunications System (UMTS); LTE; Speech and video telephony terminal acoustic test specification (3GPP TS 26.132)".

[17] Recommendation ITU-T P.501: "Test signals for use in telephony and other speech-based applications".

[18] Recommendation ITU-R BS.708: "Determination of the electro-acoustical properties of studio monitor headphones".

[19] IEC 60268-7:2010: "Sound system equipment - Part 7: Headphones and earphones".

[20] ETSI TS 103 281: "Speech and multimedia Transmission Quality (STQ); Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals".

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

[22] ETSI ES 202 737: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for narrowband VoIP terminals (handset and headset) from a QoS perspective as perceived by the user".

[23] ETSI ES 202 738: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for narrowband VoIP loudspeaking and handsfree terminals from a QoS perspective as perceived by the user".

[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 ES 202 740: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband VoIP loudspeaking and handsfree terminals from a QoS perspective as perceived by the user".

[26] Recommendation ITU-T G.191: "Software tools for speech and audio coding standardization".

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

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] F. Kettler, H.-W. Gierlich, E. Diedrich and J. Berger: "Echobeurteilung beim Abhören von Kunstkopfaufnahmen im Vergleich zum aktiven Sprechen", DAGA Conference, Hamburg, 2001.

[i.2] Recommendation ITU-T P.835: "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm".

[i.3] Recommendation ITU-T P.76: "Determination of loudness ratings; fundamental principles".

[i.4] ETSI TR 126 931: "Universal Mobile Telecommunications System (UMTS); LTE; Evaluation of Additional Acoustic Tests for Speech Telephony (3GPP TR 26.931)".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the terms given in Recommendation ITU-T P.10/G.100 [1] and the following apply:

attribute: description of a certain quality dimension of a stimulus, which is auditorily assessed by subjects in a listening test (e.g. annoyance of echo)

NOTE: Multiple attributes may be assessed for a single stimulus within one trial.

category: magnitude, which quantifies the degree of quality or degradation within an attribute

NOTE: The meaning of a certain category may be expressed by labels/descriptions, numbers or graphical alignment in the voting console to the test subject.

device-side: end-point of a telecommunication connection, which is dedicated to and operated by a device under test

NOTE: For the signal-based TPLT, a HATS is used here in order to cause double talk.

double talk: phase within a conversation (or a speech-based test signal), where the user B/ reference side as well as the user A/ DUT side are talking

double talk impairment: audible degradation in terms of quality and/or intelligibility, which is inserted by the device-side and is perceived by the listener at the reference-side

NOTE: Technically, it is typically caused by the simultaneous talker activity of both sides.

double talk source signal: signal originated from device-side and transmitted to reference-side

echo artefact: artefact generated by the signal processing in sending direction of the device-side (e.g. due to linear/non-linear coupling of signal components from receiving to sending direction of the device under test)

NOTE: It is triggered in talking phases of the reference-side.

reference-side: end-point of a telecommunication connection, which is operated by a reference device or gateway in order to capture stimuli for a TPLT

NOTE: This may be realized either electrically or acoustically with a HATS.

scale: list of categories, sorted by the degree of quality or degradation for a given attribute

signal under test: signal transmitted from device-side to reference-side

NOTE: May contain echo artefacts and/or double talk impairments caused by signal processing of DUT.

Single Talk (ST): phase within a conversation (or a speech-based test signal), where only one side/end is talking (either user B/ reference side or user A/ DUT side) is talking

source signal: signal originated from reference-side and transmitted to device-side

NOTE: May also be inserted electrically at POI to the DUT.

3.2 Symbols

For the purposes of the present document, the symbols given in Recommendation ITU-T P.10/G.100 [1] and the following apply:

a_{DT}	Attenuation (in dB) during double talk segments
g_{EL}	Factor (in dB) to obtain a certain echo loss
δ(k)	Dirac impulse (linear transmission)
ΔT	Duration of delay introduced in the echo path
dB	deciBel
dB_{SPL}	Sound Pressure Level in dB, referenced to 20 µPa
dB_{Pa}	Sound Pressure Level in dB, referenced to 1 Pa
dB_{V}	Voltage in dB, referenced to 1 Volt
$dB_{Pa/V}$	Sensitivity in receiving direction (Pascal per Volt), expressed in dB
$dB_{V/Pa}$	Sensitivity in sending direction (Volt per Pascal), expressed in dB
h(k)	Impulse response of echo path
Pa	Pascal (pressure)
T_C	Duration of concurrent talk (uplink and downlink active)
T_D	Duration of activity in downlink path
T_L	Duration of long interrupts
T_P	Duration of trailing and leading pause
T_S	Duration of short interrupts
T_U	Duration of activity in uplink path
x(k)	downlink signal sent to Device-side
x_{ST}(k)	Sidetone signal based on x(k)
y(k)	uplink signal sent by Device-side

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in Recommendation ITU-T P.10/G.100 [1] and the following apply:

5G NR	5G New Radio
ACR	Absolute Category Rating
AEC	Acoustic Echo Control
ASL	Active Speech Level
CT	Conversational Test
DCR	Degradation Category Rating
DT	Double Talk
DUT	Device Under Test
ES	Echo Suppression
FB	FullBand (20 Hz to 20 kHz)
FIR	Finite Impulse Response
GSM	Global System for Mobile Communications
INF	Infinity
IP	Internet Protocol
LTE	Long Term Evolution
NB	NarrowBand (300 Hz to 3 400 kHz)
NR	Noise Rating
NS	Noise Suppression
POI	Point of Interconnection
RCV	Receiving Direction
SLR	Sending Loudness Rating
SND	Sending Direction
SPL	Sound Pressure Level
ST	Single Talk
SWB	Super-wideband (50 Hz to 14 kHz)
TALT	Talking And Listening Test
TPLT	Third-Party Listening Test
UMTS	Universal Mobile Telecommunications System
VoIP	Voice-over-IP

WB WideBand (100 Hz to 7 kHz)

4 Fundamentals of acoustic echo control characteristics

4.1 Overview

Figure 1 depicts the simplified technical principles and components of a bidirectional end-to-end speech communication between user A (left) and user B (right). On each side, a terminal with electric and acoustic send and receive path is used. Both paths include several signal processing blocks like AEC, ES, NR, AGC and codec. The acoustic paths may range from handsets close to the ear up to recent hands-free application. The devices transmit voice signals over arbitrary and cascaded networks (e.g. VoIP access, mobile network or even satellite link).

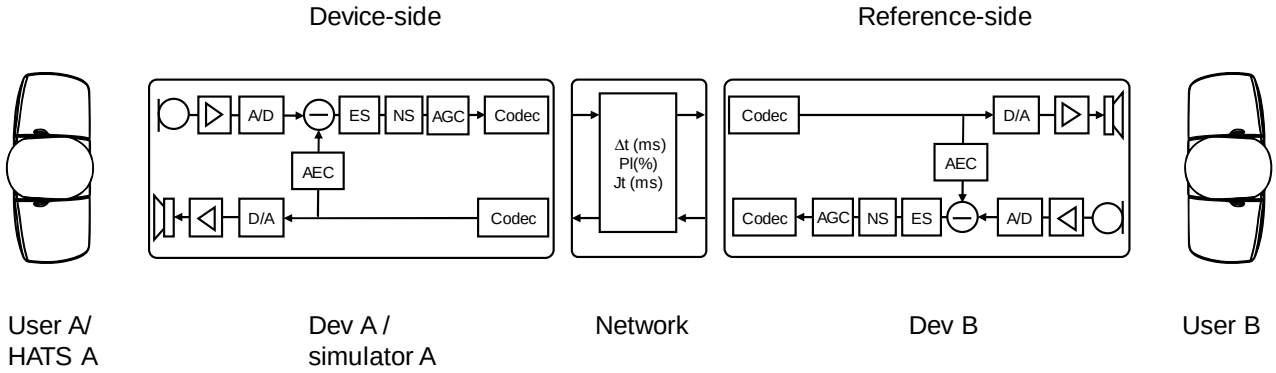

Figure 1: Technical scheme of conversation in telecommunication

NOTE: In the present document, the specific type of network is in general of minor relevance, since the degree of degradations mostly depends on the delay. However, network-specific features (e.g. coding and decoding of speech signal) should be regarded whenever possible.

4.2 Formation of Echo Artefacts

In the following, echo artefacts are described from the perspective of user B (reference side), as illustrated in Figure 3. User B starts talking and the reference device transmits the signal in sending direction via the network where delay, jitter and packet loss are possibly inserted. The signal is then played back at the device side (e.g. by loudspeaker or handset) and coupled back into the DUT's microphone. Here typically signal-processing components like an (acoustical) echo canceller and/or suppressor try to remove the echo signal. Any remaining signal is called *residual echo*, which may be even further degraded by the following signal processing units (NS, AGC, etc.).

The residual echo is transmitted back to the reference device via the network and played back to user B. In general, the resulting residual echo to be perceived by user B may be a delayed, attenuated and (linearly and/or non-linearly) distorted version of the source signal transmitted by user B. Since the roundtrip delay of the whole transmission is typically in the range of (at least) a few hundred milliseconds, user B may perceive already an echo signal while he/she is still talking. In this case, the echo signal may be partially masked by the sidetone of his/her own voice.

4.3 Formation of Double Talk Impairments

In the following, double talk impairments are described from the perspective of user B (reference side), as illustrated in Figure 1. The signal transmission paths (including network and signal-processing elements in terminals) are similar as for the formation of echo artefacts, but this time also user A is talking. A typical real-life scenario would be for example that user A is talking continuously and user B starts to interrupt him/her. The DUT and the signal-processing at the device-side has to accurately differentiate between the (wanted) near-end speech of talker A and the (unwanted) far-end signal, which may be coupled back into the sending path. A typical impact of a possibly imperfect signal processing is that the user A's speech signal is degraded by temporal clipping, which is perceived as an annoying discontinuity at the reference side by user B.

Figure 2 shows the basic methodology of double talk impairment generation, which are introduced into the uplink-transmitted signal. The second and third row of the chart provide the talking activities of users A and B at the device side. For now, a perfect transmission of user B's speech signal from the reference-side is assumed.

An imperfect signal processing in user A's terminal (AEC and ES) attenuate the uplink signal during double talk ranges (see first row and DT ranges 1, 2 and 3 in Figure 2) in order to suppress any possible echo signal into the network. In case the downlink signal is paused (between DT range 1 and 2), the Echo Suppresser of device A does not attenuate the uplink signal anymore. However, in case the signal processing does not adapt on this conversation, temporal clipping can be inserted at any time as soon as user B is talking again.

In general, such double talk impairments (delayed, strongly distorted and clipped version of the uplink signal) introduced by the DUT can severely degrade the user experience at the reference side for listener B.

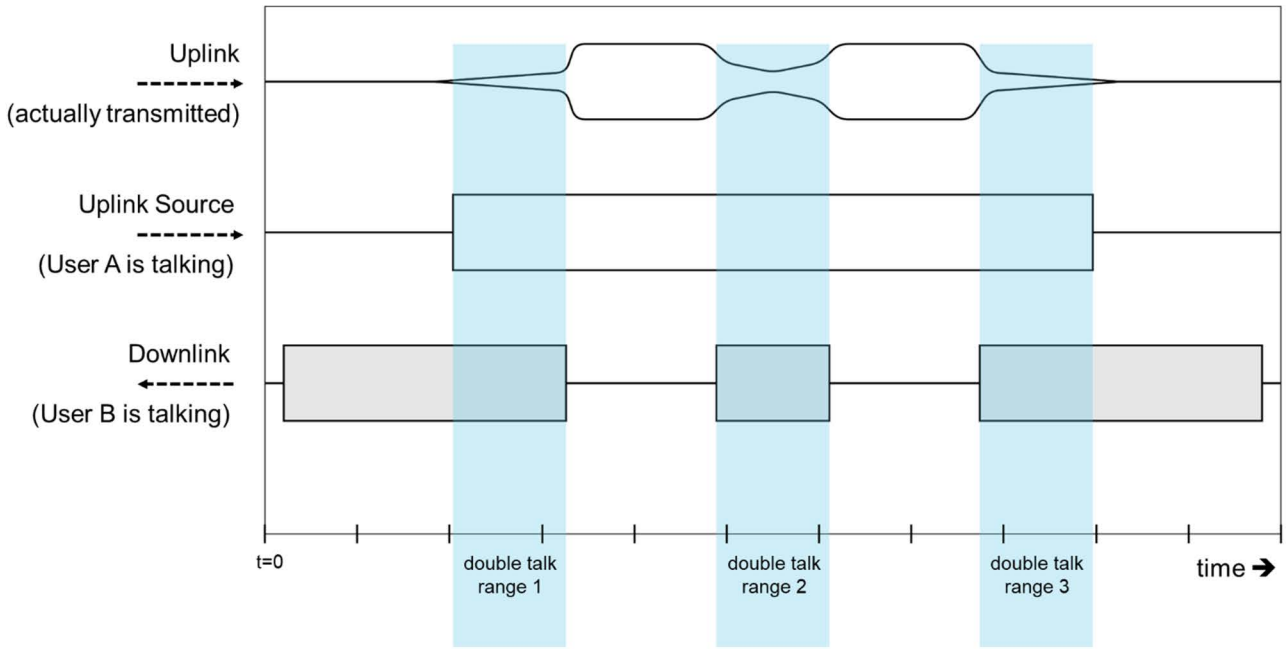

Figure 2: Types of double talk scenarios

Figure 2 also depicts three different types of double talk ranges, each having a different impact on double talk impairment perception:

- Double talk range 1: the continuously transmitted downlink activity (from the DUT point of view) is interrupted by the speech signal activity of user A (device side). The initial part of the uplink speech signal perceived by user B will be degraded.
- Double talk range 3: the continuously transmitted uplink activity (from the DUT point of view) is interrupted by the speech signal activity of the reference side's user B. At least for a certain time, the downlink activity would mute user A. This scenario presents a more obvious annoyance of the typical human listener at side B.
- Double talk range 2: combination of the two aforementioned scenarios.

5 Auditory Assessment of Conversations

5.1 Overview

The subjective assessment described in the present document focuses on the performance evaluation of (real or simulated) devices under test, which include one or more of the aforementioned signal processing elements. In this case, user A is replaced by a HATS to reproducibly send and receive acoustic signals. User B could in general be a real person, who listens to his/her own voice (and resulting possible echo artefacts) and to the pre-recorded talker A (including possible double talk impairments). However, for sake of reproducibility, user B is also replaced by HATS, which records the aforementioned signal components. This setup has several advantages:

- Talker signals from user B can be captured electrically and inserted electrically at a later stage to the actual device to be evaluated.
- Certain network conditions (in particular additional delay) can be emulated.
- Well-defined listening situation at the reference-side may be conducted in a listening lab.

Following the principles described above, the setup can be simplified as shown in Figure 3.

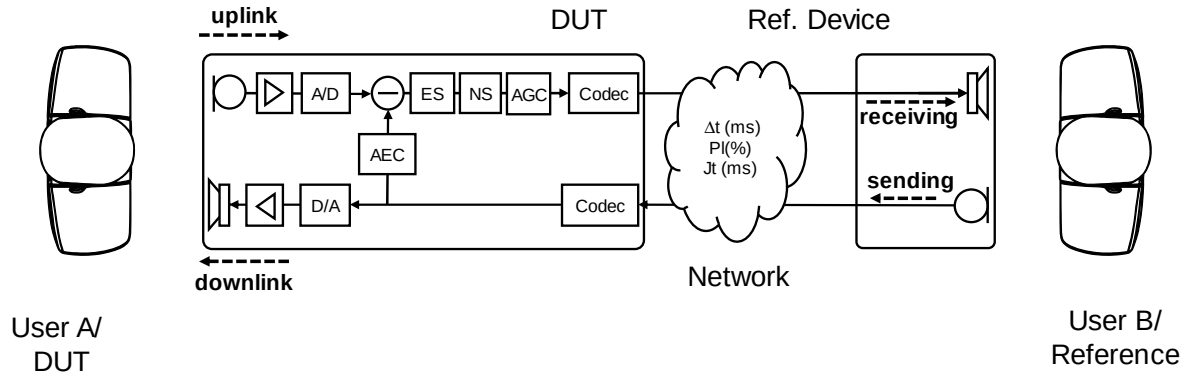

Figure 3: Simplified recording setup for generation of TPLT samples

5.2 Possible Types of Listening Test

5.2.1 Overview

Recommendations ITU-T P.831 [3] and P.805 [5] describe three suitable auditory test methods for evaluating interactive scenarios like echo artefacts or double talk impairments, i.e. Conversational Test (CT), Talking And Listening Test (TALT) and third-party listening test (TPLT). In the following clauses, a brief introduction is provided for each methodology.

In clause 4.3.1 of ETSI TR 126 931 [i.4], an additional subjective testing approach for echo artefacts and double talk impairments is introduced. In contrast to the listening test types introduced in the following, this approach did not take the sidetone of the listener's own voice into account, i.e. it was designed as a listening-only test.

5.2.2 Conversational Test

A CT, as the most complex test method involves two parties actively conversing over a live connection. Conducting a CT is very time-consuming and complex. In particular, it is a very difficult task to control all individual influences such as talking behaviour, individual speech levels, use of terminals, etc. The conversation procedure is completely interactive, the impairments are introduced online, mostly realized by use of different devices. This implies that actual user devices need to be manipulated from their intended use case in order to produce sufficiently disturbing impairments. User behaviour cannot be influenced directly, for example the double-talk ratio of the conversation cannot be forced, only encouraged by the specific user task of the conversation, e.g. by a *Kandinsky test* (see also Recommendation ITU-T P.805 [5]).

5.2.3 Talking-and-listening test

In a TALT, the test subject interface via a telephone handset with a live simulation of speech signals impaired by echo artefacts and double talk impairments. All impairments need be introduced in real-time because the source speech signal is generated by the test subject himself. Therefore, the test subject can listen to echoes of their own voice in case of echo artefacts.

TALT is more efficient to judge disturbed speech signals than in a CT. The conversation procedure is interactive on the reference-side, but the interactivity needs to be limited in order to work in conjunction with the online simulation of impaired speech signals. For the sake of comparability, the spoken text of the test subjects needs be scripted up to a certain degree. The real-time simulation of impaired speech signals is a demanding challenge especially in case of simulating both impairments simultaneously.

5.2.4 Third-Party listening tests

In a TPLT as described in Recommendation ITU-T P.831 [3], the test subjects listen to and consequently judge impaired speech signals that are part of a conversation between the two recorded parties. The impaired speech signals are generated offline, typically using artificial head recordings. The speech signals either can be recordings of (impaired) real devices or are generated by an online-simulator. Usually, both simulated and recorded signals may be used for a TPLT.

The test procedure is non-interactive and completely pre-determined. All test samples are recorded or simulated in advance. Therefore, the listening test designer can choose from a high number of types of impairments, which allows a thorough preparation of the listening test. Furthermore, numerous test conditions (e.g. terminals/signal processing components involved or network conditions) may be included.

The downside of this approach is that the presentation method of the listening samples is somewhat artificial. Nevertheless, in previous studies, the execution of TALT and TPLT under identical test conditions typically leads to comparable results, as shown in [i.1].

5.2.5 Summary

As already indicated in Recommendation ITU-T P.805 [5], each of the aforementioned listening/conversational tests have to deal with a trade-off between realistic conversations, number of test conditions, reproducibility and effort. A comparison of the specific advantages and disadvantages of the corresponding subjective test method is provided in Table 1 and Table 2.

Table 1: Advantages and disadvantages of conversational test methods

Type	Advantage	Disadvantage
CT	Close-to-reality, natural conversation situation, high subject immersion	Very time/cost inefficient, complex, low reproducibility
TALT	Close-to-reality, test subjects listens to his/her own voice	Time/cost inefficient, complex, varying listening situation for subjects, possible mediocre reproducibility
TPLT	High reproducibility, time/cost efficient, highly scalable	Listening situation is more artificial (ear witness)

These conversational test procedures differ in terms to the degree of involvement (interactive, non-interactive) and the way the disturbed and perceived speech is generated (online, offline). The peculiarities of each test method are summarized in Table 2.

Table 2: Comparison of test methods regarding procedures and impairments

Type	Conversation procedure	Impairment generation
CT	Interactive conversation on both sides, user A and user B	Online generation of impaired signals due to (imperfect) signal processing in both user devices DevA and DevB
TALT	Pre-recorded conversation on side A (using HATS recording) and interactive but scripted conversation for user B	Online generation of impaired signal, DevA is replaced by a real-time simulator
TPLT	Non-interactive, completely pre-recorded/simulated in advance	Impairment is generated/simulated in advance (only offline)

The setup introduced in Figure 1 may in general be used by all three types of auditory test methods. For CT and TALT, user A and B are two human test subjects, who follow the instructions provided by the supervisor of the test.

For the auditory evaluation of signals affected by echo artefacts and double talk impairments, degradations can be introduced either *online* or *offline*:

- *Online* describes a processing scenario, where controlled degradations are inserted in real-time into the signals presented to the test subjects. Due to the interactivity demand (test subjects are actually talking), CT and TALT require online processing.
- *Offline* describes a processing scenario, where all degraded test signals are generated in advance and played back later for the listening test subject in a controlled manner and a controlled environment. Only a TPLT can utilize offline recordings.

From the terminal testing perspective, it is in general preferred to use stimuli, which are created offline (e.g. to re-use them for prediction models). In consequence, the test methodology used in the present document is based only on the TPLT design according to Recommendation ITU-T P.831 [3].

5.3 Selection of Speech Material

This clause describes the source speech material, which shall be used for the generation of degraded listening test conditions. All sequences described in the following shall be suitable for the playback via mouth simulator of HATS, i.e. available in fullband (20 Hz - 20 kHz audio bandwidth). If not specified otherwise, speech signals shall be calibrated to -4,7 dB_{Pa} Active Speech Level (ASL) according to Recommendation ITU-T P.56 [21].

Each of the stimuli presented in the TPLT shall not exceed an overall duration of 12,0 s. At least one utterance of user B (downlink path) with a minimum duration of $T_D > 1,0$ s shall be used. In case of double talk stimuli, user A at the device side (uplink path) shall be active for a duration T_U of at least 1,5 s. The minimum duration T_C of the source signal of the uplink path shall ensure that at least 0,5 s of concurrent talk with the downlink path is provided for each listening test sample. Both uplink and downlink sources shall include a maximum of four single utterances (e.g. four short interrupts or one longer talking phrase).

Similar to existing listening test specifications regarding speech quality, e.g. Recommendation ITU-T P.800 [2] or Recommendation ITU-T P.835 [i.2], it is recommended to use standard speech samples whenever possible. Suitable speech material in several languages can be found in Recommendation ITU-T P.501 [17] or in annex E of ETSI TS 103 281 [20].

However, in some cases it is necessary to use self-recorded speech signals for a third-party listening test, for example:

- It is desired to use more realistic speech samples regarding conversational aspects (e.g. increased interactivity, question/answer or short but meaningful interrupts).
- Standard speech material of the aforementioned sources is not available in mother tongue of the participating test subjects.

Such self-recorded test signals shall be professionally recorded real speech. The following requirements shall be met:

- Minimization of the recording noise floor (less than 30 $dB_{SPL}(A)$).
- Type, positioning and handling of the recording microphone should guarantee a linear frequency response (within ±3 dB limits) with regards to the Mouth Reference Point (MRP).

- No clipping shall occur.
- No additional noises should be inserted by the talker (e.g. breathing noise).
- At least two different talkers shall be used.
- Female and male talkers shall be represented equally.
- The speech sequence shall include at least two samples (time ranges to be cropped for the presentation in the auditory test).

For the definition of the time domain composition of the speech signals it needs to be distinguished between echo-only tests (only downlink is active) and double talk tests (up- and downlink are active):

- For echo-only tests, the time domain composition of the source signal can be kept simple. Each ending of an utterance shall not interfere with the beginning of the next one. In order to consider a maximum expected echo delay, each sample shall always provide a leading and trailing silence period T_P of at least 0,5 s (1,0 to 2,0 s are recommended).
- For double talk tests, the time domain composition of the source signal typically is more complex. For the downlink path, the same requirements regarding trailing/leading pauses and overall activity/duration as for echo-only apply here. In addition, the overlap between uplink and the downlink channel needs to be carefully designed. It is preferable to include both shorter (duration $T_S < 1,0$ s) and longer interrupts (duration $T_L > 1,0$ s).

Figure 4 illustrates an example structure for a valid source signal for both directions, including samples for echo-only and double talk testing. For echo-only testing, a single sentence ($T_D > 1,0$ s) for the downlink path is used. For double talk testing, user A is continuously talking (uplink path, $T_U > 1,5$ s) during the whole duration of the sample. Concurrent talk of user B (downlink path) is present with a short ($T_S < 1,0$ s) and a long ($T_L > 1,0$ s) interrupt. The overall duration of concurrent talk $T_C > 0,5$ s is provided as well.

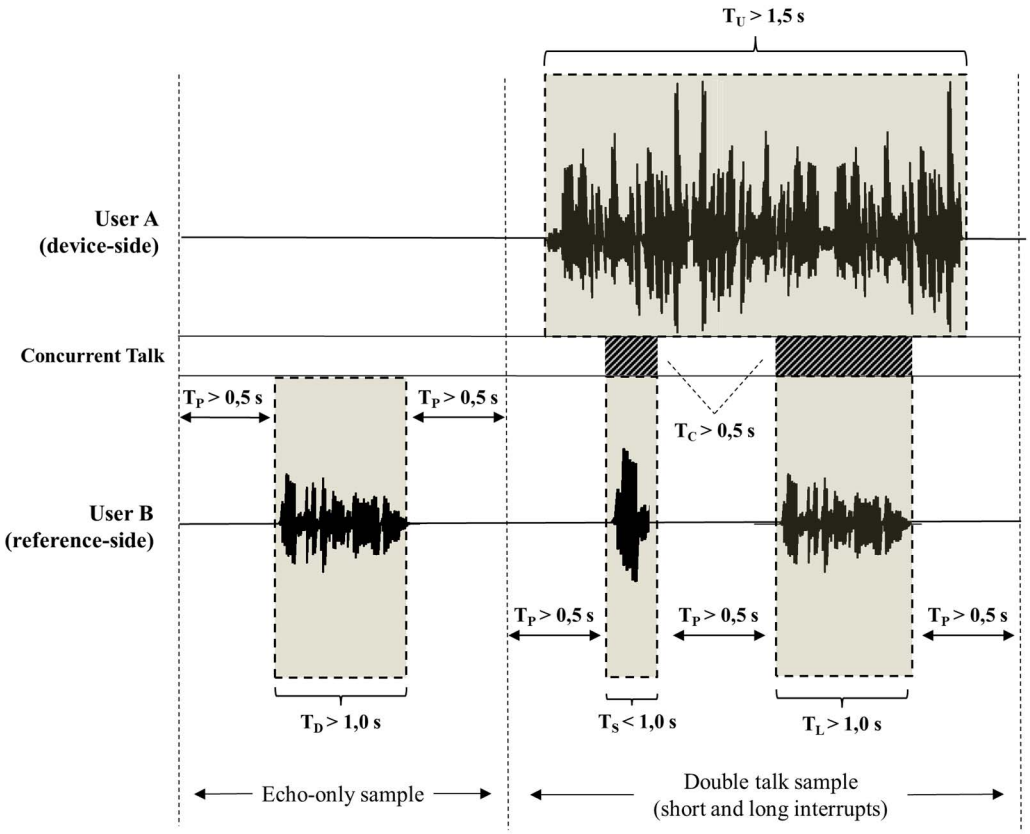

Figure 4: Example of source signal for generation of TPLT samples

NOTE 1: As indicated in Figure 4, it is possible to use a concatenated source sequence, which includes time ranges for echo-only as well as for double talk testing. After the generation of degraded speech material, this longer sequence has to be cropped accordingly.

NOTE 2: The single and double talk sequences described in clauses 7.3.5 (British English) and 7.4.1.3 (Mandarin) of Recommendation ITU-T P.501 [17] comply in general with the aforementioned specifications, but further cropping into shorter samples for presentation in the listening test is required.

NOTE 3: The single and double talk sequences (British English) described in clauses 7.3.5 of Recommendation ITU-T P.501 [17] are also used in ETSI TS 126 132 [16] for the evaluation of echo control characteristics. There the single talk signal is used for downlink and the double talk signal for uplink path.

Any listening test according to the present document shall use speech sequences in the language that corresponds to the mother tongue of the participating test subjects.

5.4 Generation of test conditions

5.4.1 Introduction

Figure 5 provides a modified version of the general communication scenario shown in Figure 1 and Figure 3, which describes the recording of binaural stimuli for the TPLT. For this purpose, user B at the reference-side is replaced by a HATS, which acts first as a talker, then in a second step as a listener.

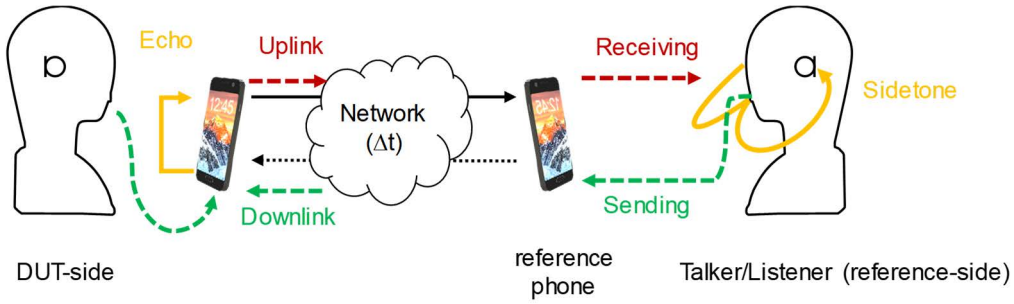

Figure 5: Components of reference-side speech signals

Stimuli for the listening test are binaural and consist of the following components:

1) Receive signal, perceived via a (real or simulated) reference phone. It may include the echo signal and/or the transmitted (and possibly degraded) speech signal of the talker at the device-side.

2) Acoustic sidetone, i.e. the talker's own voice recorded binaurally (with possibly mounted reference terminal).

For the recordings of receive signal and sidetone, diffuse-field correction according to Recommendation ITU-T P.58 [14] shall be applied on the captured ear signals.

5.4.2 Requirements on Test Equipment

For the playback and recording of speech signals, accesses to acoustic interfaces and different types of electrical POI are required. For this purpose, high quality measurement equipment shall be used for all recordings described in the following clauses. Depending on the use case, terminal type and bandwidth according to Table 3, the test equipment and recording room/environment shall comply with the corresponding specifications given in the column reference.

Table 3: References to requirements on test equipment

Terminal Type	Bandwidth	Use case	Reference
VoIP Desktop Terminals	NB	Handset/Headset	ETSI ES 202 737 [22]
		Hands-free	ETSI ES 202 738 [23]
	WB	Handset/Headset	ETSI ES 202 739 [24]
		Hands-free	ETSI ES 202 740 [25]
	SWB/FB	Handset/Headset	ETSI TS 102 924 [11]
		Hands-free	ETSI TS 102 925 [12]
Mobile Phones (GSM, UMTS, LTE and 5G NR networks)	NB/WB/SWB/FB	Handset/Headset/Hands-Free	ETSI TS 126 132 [16]
Wireless Terminals	NB	Handset/Headset	ETSI TS 103 737 [7]
		Hands-free	ETSI TS 103 738 [8]
	WB	Handset/Headset	ETSI TS 103 739 [9]
		Hands-free	ETSI TS 103 740 [10]
	SWB/FB	Handset/Headset	ETSI TS 102 924 [11]
		Hands-free	ETSI TS 102 925 [12]

5.4.3 Recordings on Reference-side

5.4.3.1 Sending Direction

In order to transmit the speech signals originated from the reference-side with a realistic absolute level range, the following insertion methods in sending direction are possible:

1) Playback over the artificial mouth, acoustic insertion into a reference device (e.g. a known phone providing high quality). This device shall be handset- or headset-like, i.e. providing a short distance between MRP and input microphone. Positioning of this device shall comply with Recommendation ITU-T P.64 [15]. The sending frequency response, loudness rating and possible introduced delay shall comply with the corresponding requirements of ETSI TS 103 737 [7], ETSI TS 103 739 [9] or ETSI TS 102 924 [11]. Any delay in introduced by the reference device shall be compensated. Results of these measurements shall be enclosed to the listening test results.

2) Acoustic insertion is simulated by filtering the source speech with a sending frequency response of a known reference device. Again, sending frequency response, loudness rating and possible introduced delay shall comply with the corresponding requirements of ETSI TS 103 737 [7], ETSI TS 103 739 [9] or ETSI TS 102 924 [11]. The delay in introduced by the reference device shall be compensated to the source signal. Results of these measurements shall be enclosed to the listening test results.

3) Acoustic insertion is simulated by filtering the source speech with a default frequency response (flat transfer function in the bandwidth-specific frequency range), as shown in annex B. The simulated send sensitivity depends on the bandwidth: for NB a constant sensitivity of -13,55 dB(V/Pa) and for WB/SWB/FB, -14,09 dB(V/Pa) shall be applied. See annex B for more details.

The HATS on the device-side does not act as a listener, but may also talk simultaneously to the incoming speech signal from the reference-side (double talk).

5.4.3.2 Sidetone

In order to simulate the binaural self-hearing of talking perceived by the talker/listener at the reference-side (sidetone) for the listening test stimuli, the following recording or simulation methods are possible:

1) Recording with active reference device mounted to HATS. In this case, the sidetone has to be recorded separately for each source signal at the reference-side.

2) Simulation by filtering the speech source with a sidetone transfer function (MRP to DRP) of a known reference device.

3) Default sidetone: Simulation by filtering the speech source with transfer function MRP-to-DRP without any device mounted.

NOTE: The usage of L_{MEHS} from Recommendation ITU-T P.76 [i.3] (human sidetone path MRP to ERP) as an additional weighting/filtering or replacement in this clause for joint air/bone conduction is currently under study, which is not yet included in the aforementioned recordings.

5.4.4 Recording of degraded signals

Listening test speech samples simulated with the subjective test procedure of the present document should reflect impairments, which may occur in real-life communication situations as well. However, the number of conditions may grow rapidly, when trying to combine multiple source files, terminals and/or other degradations.

Table 4 provides an overview about possible impairments on the speech signal, which can be caused by components included in the terminal.

Table 4: Possible impairments caused by terminal components

Signal Processing Unit	Possible Impairments
Microphone/ A/D-converter	Room Noise, electrical noise, clipping
Codec	Codec noise, distortion, unnatural sounding speech
ES, AEC	Double talk attenuation, double talk coloration, delay, insufficient echo loss causing residual echo
AGC	Noise amplification, waveform modulation
NS	Musical tones, distortion of speech parts, missing speech parts
Network	Delay, packet loss (in case of IP-network)
Loudspeaker/ D/A-converter	Electrical noise, clipping, harmonic distortion

For the generation of degraded speech signals with real terminals, the simulated or measured signals as determined in clause 5.4.3.1 (reference-side) are inserted electrically at the POI into the downlink path of the DUT (see Figure 5). For double talk scenarios, the playback via mouth simulator shall be synchronized with the downlink signal, i.e. the speech signal to be transmitted from the device-side shall be delayed by the receiving delay of the DUT. Finally, the signal of the uplink path of the DUT is captured again at the POI.

Degraded signals may also be generated by full or partly offline simulations (e.g. simulation of terminal and room acoustics versus acoustic capture with a mock-up device). The output of any simulated terminal shall correspond to the POI of the uplink path.

NOTE: For recordings including double talk, i.e. near-end speech from the device-side (user A is talking), the transmitted signal in uplink should have a certain minimum quality. Test subjects are instructed to pay attention only to signal degradations, which are introduced by concurrent talk (see clause 5.7.1). In case the uplink is already severely degraded even without echo artefacts or double talk impairments, a low transmission quality (to be determined with measurements given in e.g. ETSI TS 103 737 [7] to ETSI TS 102 925 [12]) may impact the assessment of subjective results.

Between 12 and 54 test conditions (recorded or processed) shall be created using the speech material described in clause 5.3. It is recommended to include 48 conditions (excluding reference conditions, see clause 5.5) in the evaluation in a TPLT.

5.4.5 Calibration of test signals

For the generation of stimuli, the sending signals from device-side are captured electrically at the POI. In order to provide correct level and suitable frequency shaping, the acoustic capture on the reference-side can be realized similar as in clause 5.4.3.1:

1) Electric insertion into a reference device (e.g. a known phone providing high quality) and acoustic capture via HATS. Diffuse-field correction according to Recommendation ITU-T P.58 [14] shall be applied on the ear signals.
   This device shall be handset- or headset-like, i.e. providing a short distance between MRP and input microphone. Positioning of this device shall comply with Recommendation ITU-T P.64 [15].
   The receiving frequency response, loudness rating and possible introduced delay of the reference device shall comply with the corresponding requirements of ETSI TS 103 737 [7], ETSI TS 103 739 [9] or ETSI TS 102 924 [11]. Results of these measurements shall be enclosed to the listening test results. Any delay introduced by the reference device in receiving direction shall be compensated for the stimuli.

2) Acoustic capture via HATS is simulated by filtering the signal under test with a receiving frequency response (including diffuse-field correction according to Recommendation ITU-T P.58 [14]) of a known reference device.
Again, the receiving frequency response, loudness rating and possible introduced delay (may be compensated) shall comply with the corresponding requirements of ETSI TS 103 737 [7], ETSI TS 103 739 [9] or ETSI TS 102 924 [11]. Results of these measurements shall be enclosed to the listening test results.

3) Acoustic insertion is simulated by filtering the signal under test with a default SWB frequency response and no additional delay, as shown in Figure B.1 in Annex B.2 (assuming that a diffuse-field correction according to Recommendation ITU-T P.58 [14] is included). In order to obtain realistic receiving levels, the following parameters shall be met and reported enclosed to the listening test results: The sensitivity shall be 4,83 dB(Pa/V) ± 2 dB (referring to WB/SWB/FB mode of Recommendation ITU-T P.700 [6]).

NOTE: When using a (real or simulated) reference device, it is recommended to use the identical device as for the reference sending recordings, as described in clause 5.4.3.1.

Applying the calibration procedure described above on the signals under test, absolute levels of (virtually or actually) of received signals (residual echo artefacts, double talk impairments) plus sidetone generation (see clause 5.4.3.2) are then within a realistic range.

5.5 Reference conditions

As common practise for other standardized speech quality tests, reference conditions shall be used in TPLT complying with the present document.

Reference conditions are a well-established method to conduct meaningful comparisons of auditory test results from different laboratories or from the same laboratory at different times. These conditions always include a best possible condition (also often denoted as *clean* or *direct*), as well as conditions where known, controlled degradations have been added to the speech materials. This reference system also provides specific anchor points. The direct condition represents the very best condition that is attainable in the experiment.

Below, two different processing sets for echo-only and double talk TPLT are described, which shall be applied to the speech material used also for the test conditions.

Echo-only:

The annoyance of perceived echo signals mainly depends on the (perceived) level and delay of the listener's own voice. The reference system according to Table 5 shall be used for echo-only tests, which defines several anchor conditions based on a simulated echo loss for the listener on the reference-side. Clause A.1 describes the generation process for these reference conditions in detail.

Table 5: Reference conditions for echo-only tests

ID	Echo Loss [dB]	Resulting Echo Level (ASL) [dB_{SPL}]	Echo Delay (ΔT) [ms]	Comment
R01	+inf.	-inf.	0	Direct/no degradation
R02	50	30,6	100	Second-best anchors (with different delays)
R03	55	25,6	200	
R04	55	25,6	400	
R05	45	35,6	400	
R06	45	35,6	600	
R07	30	50,6	400	
R08	20	60,6	800	Worst degradation

Double talk:

The annoyance of perceived double talk impairments mainly depends on the attenuation of the send path of the DUT. Since double talk scenarios may also still include echo impairments, some of the echo-only conditions apply here as well. The reference system according to Table 6 shall be used for double talk tests, which defines several anchor conditions based on simulated attenuation during double talk and the aforementioned echo loss. Clause A.2 describes the generation process for these reference conditions in detail.

Table 6: Reference conditions for double talk tests

ID	DT Attenuation (a_{DT}) [dB]	Echo Loss [dB]	Resulting Echo Level [dB_{SPL}]	Echo Delay (ΔT) [ms]	Comment
R01	0	+inf.	-inf.	0	Direct/no degradation
R02	0	50	30,6	100	Best echo-only anchor
R03	0	45	35,6	400	
R04	0	30	50,6	400	
R05	0	20	60,6	0	Worst echo-only anchor
R06	3	+inf.	-inf.	0	Best double talk-only anchor
R07	9	+inf.	-inf.	0	
R08	16	+inf.	-inf.	0	
R09	50	+inf.	-inf.	0	Worst double talk-only anchor
R10	4	50	30,6	100	Fair combined echo & DT anchor
R11	10	30	50,6	300	Medium combined echo & DT anchor
R12	25	20	60,6	600	Worst combined echo & DT anchor

NOTE: In clause 4.3.1 of ETSI TR 126 931 [i.4], anchoring of test subjects was conducted in a similar way. However, the controlled double talk impairments used in this listening test design are not applicable for the current TPLT design. The degradations introduced in the time-frequency domain need to be aligned with the self-speech signal, i.e. double talk impairments should only occur during or shortly after the sidetone is active.

5.6 Headphone playback for presentation

Headphones used for presentation of the test material to the listening panel shall be calibrated and equalized using a HATS conforming to Recommendation ITU-T P.58 [14] equipped with an artificial ear type 3.3 according to Recommendation ITU-T P.57 [13]. The HATS is diffuse field equalized.

The resulting frequency response characteristics (in one-third octave bands) of the headphones used in the subjective experiments should be within the mask given in annex 1 of Recommendation ITU-R BS.708 [18] (see Figure 1).

Alternatively, equalization can be made using a subjective method as in IEC 60268-7:2010 [19], ensuring that all frequencies for full-band listening are satisfactorily reproduced.

Since the sidetone at the reference-side is always binaural, the presentation of the test and reference conditions to listeners shall also always be binaural - independent of the type of chosen reference listening device (monaural handset/headset or binaural headset).

5.7 Listening Test Design

5.7.1 Listening Test Instructions

Since the task in a TPLT is more challenging for naïve test subjects than traditional speech quality tests, the test has to be prepared with care. A clear, simple and direct introduction shall be presented to the subjects in advance of the test. This includes a written and/or oral description in the mother tongue of the subjects of the possible degradations to be expected (echo artefacts, double-talk impairments). In particular for double talk impairments, test subjects shall be instructed that:

- Echo artefacts may be superimposed with double talk impairments (and should be addressed to the corresponding attribute, see clauses 6 and 7).
- Double talk impairments only occur during or shortly after talking (self-masking speech).

- Only the degradation during and shortly after concurrent talk should be assessed, not the absolute quality of the transmitted speech signal.

After the instruction and prior to the actual voting procedure, it is recommended to present some demonstration samples to the test subjects (without voting). These should illustrate the two perceptual dimension (echo and double talk) with representative listening examples.

The instructions of the test shall be identical for all subjects, i.e. shall not be adapted for different groups of participants (e.g. for experienced or expert listeners).

5.7.2 Choice of Listening Test Subjects

In general, the choice of test subjects should follow the guidelines of Recommendation ITU-T P.800 [2] and the Handbook of subjective testing [4] as close as possible. The subject pool should be representative of the telecommunication user pool. Groups should be heterogeneous and balanced concerning gender, age and (if possible) professional background. The native language of the test subjects shall be identical to the language used for the generation of test conditions (see clause 5.4.4) and reference conditions (see clause 5.5).

The number of different test subjects shall equal at least the number of test conditions (excluding reference conditions) divided by two.

EXAMPLE: For the recommended number of 48 test conditions, at least 24 different subjects are required. In overall, the number of votes per condition shall be at least 48.

Each sample in the TPLT shall be evaluated by at least eight subjects, i.e. leading to a minimum of eight votes per sample. Depending on the number of conditions and samples per condition, the number of test subjects shall be increased in order to meet the aforementioned requirements.

5.7.3 Test Procedure

In accordance to [4], a training phase of at least five samples shall be presented in advance of the real test. These samples shall be of same type as the actual test conditions and may originate from the test corpus itself. However, it shall be ensured that the training samples are not presented in the first session of the test. The results of the training samples are excluded from the evaluation of the test.

More specific description of the test procedures can be found in clause 6 (echo-only testing) and clause 7 (double talk testing).

5.7.4 Test Sample Presentation

The samples of test and reference conditions shall be structured into sessions. In each session, one randomly selected sample per condition is included (*balanced blocks* design, see [4]). All listening test samples within a single session shall be presented in a (pseudo-)random order to the subjects.

5.8 Requirements on the listening laboratory

Listening laboratory facilities shall comply with the requirements and follow the guidelines provided in Recommendation ITU-T P.800 [2] and the ITU-T handbook on practical procedures for subjective testing [4].

6 Assessment of Echo Artefacts

For echo-only listening tests, the common requirements of clause 5 shall be met. In addition, the following items shall be considered:

- During the recording of test conditions, no talker shall be active at the sending side of the DUT, i.e. no uplink signal shall be transmitted by the DUT (except the echo signal). See also clause 5.4.4.
- The auditory scale and questionnaire according to Table 7 shall be used. In order to allow test subjects to use intermediate category steps, Table 8 may be used instead alternatively.

- Questionnaire and labels of the categories shall be translated into the native language of the test subjects (which is the same as the language used for the test samples, see clause 5.7.2).
- Besides the assessment of echo artefacts, no other attribute shall be evaluated during the test.

Table 7: DCR scale for attribute "Echo Artefacts", according to Recommendation ITU-T P.800 [2]

Echo disturbance is...	Value
Inaudible	5
Audible but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

Table 8: DCR scale for attribute "Echo Artefacts" according to Recommendation ITU-T P.800 [2], with additional intermediate steps

Echo impairment is...	Value
Inaudible	5,0
-	4,5
Audible but not annoying	4,0
--	3,5
Slightly annoying	3,0
---	2,5
Annoying	2,0
----	1,5
Very annoying	1,0

7 Assessment of Double Talk Impairments

For double talk listening tests, the common requirements of clause 5 shall be met. In addition, the following items shall be considered:

- During the recording of test conditions, a talker shall always be active at the sending side of the DUT, i.e. no echo-only recordings as per clause 6 shall be evaluated in a double talk test. See also clause 5.4.4.
- The auditory scale and questionnaire according to Table 9 shall be used. In order to allow test subjects to use intermediate category steps, Table 10 may be used instead alternatively.
- Questionnaire and labels of the categories shall be translated into the native language of the test subjects (which is the same as the language used for the test samples, see clause 5.7.2).
- The device-specific and perceptual relevant quality degradations in an interactive communication scenario are echo artefacts and double talk impairments, as outlined in clause 4. Therefore, the two attributes shall always be assessed for each listening sample, i.e. test subjects shall vote each sample on the scale according to Table 9/Table 10 as well as on the scale for echo artefacts according to clause 6 (see Table 7/Table 8).
- Besides the assessment of double talk impairments and echo artefacts, no other attribute shall be evaluated during the test at the same time.
- It is recommended to switch the order of attributes at half of the test duration, i.e. after a break (similar to the procedure in Recommendation ITU-T P.835 [i.2]).
- It is strongly recommended to assess both attributes simultaneously during the listening test. In this case, each sample shall be presented at least twice (similarly to the procedure in Recommendation ITU-T P.835 [i.2]).
- However, as already described in clause 5.7.1, the design of a TPLT provides a high level of complexity, especially for naïve subjects, which may cause a high cognitive load. In order to avoid overwhelming, the attribute for echo artefacts as per clause 6 can be run in a subsequent and separate session with the same samples.

Table 9: DCR scale for attribute "Double Talk Impairments", according to Recommendation ITU-T P.800 [2]

The perceived degradation in speech quality of your conversational partner during double talk is...	Value
Inaudible	5
Audible but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

Table 10: DCR scale for attribute "Double Talk Impairments", according to Recommendation ITU-T P.800 [2], with additional intermediate steps

The perceived degradation in speech quality of your conversational partner during double talk is...	Value
Inaudible	5,0
-	4,5
Audible but not annoying	4,0
--	3,5
Slightly annoying	3,0
---	2,5
Anoying	2,0
----	1,5
Very annoying	1,0

Annex A (normative): Generation of Reference Conditions

A.1 Reference Conditions for Echo-only Listening Tests

The reference conditions for the echo-only listening test are based on the typical setup also used for the recordings of real conditions, as described in clause 5.4. The principle for the generation is shown in Figure A.1.

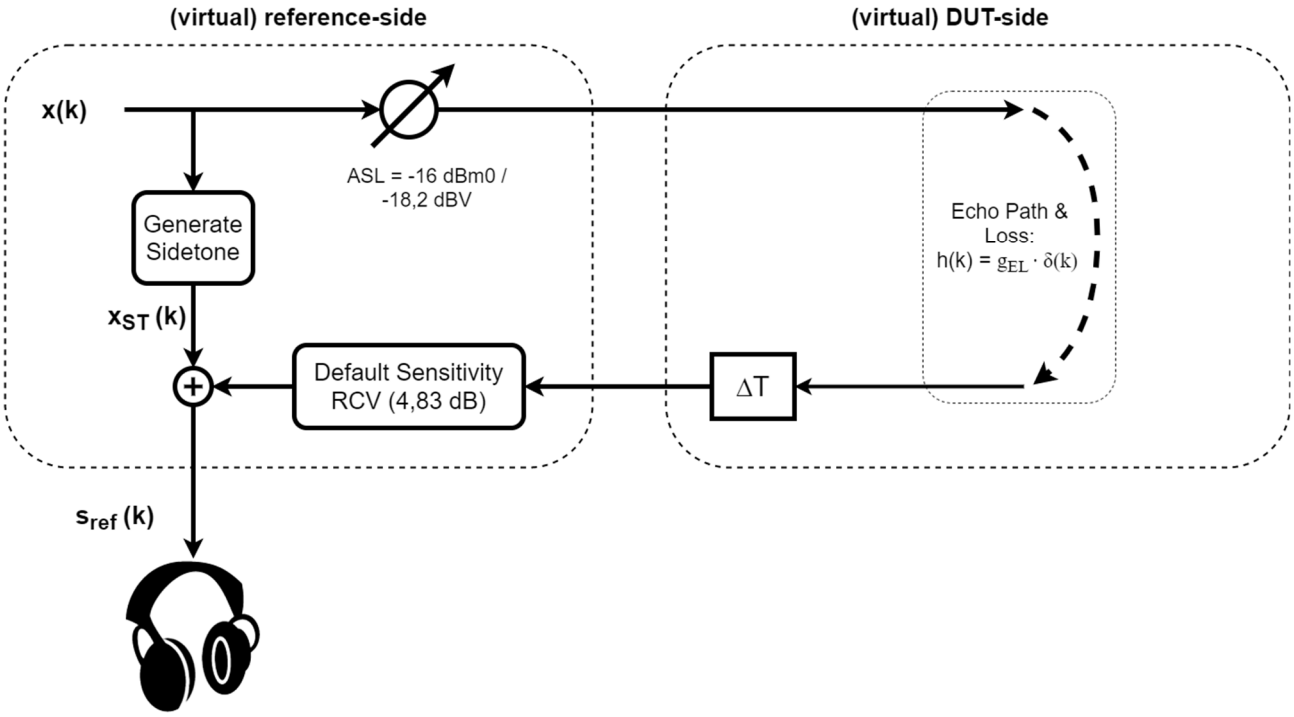

Figure A.1: Generation of reference conditions for echo-only listening test

The source signal x(k) shall be the same as used for the test conditions (see clause 5.3 and Figure 4), i.e. the downlink signal for the DUT. For the generation of reference conditions, it is assumed that x(k) is sent from the (virtual) reference side with a default network level of -16 dBm0 and without any degradation to the (virtual) DUT-side. Here a simple echo path h(k) is implemented in order to create simple and deterministic echo impairments. The signal x(k) is simply attenuated by a given echo loss g_{EL} (typically specified in dB).

This linear echo signal is then delayed by a given value ΔT and transmitted back to the (virtual) reference-side. Here a default receiving sensitivity of 4,83 $dB_{Pa/V}$ (see clause 5.4.5) is applied in order to transform the echo signal to a (pseudo-)acoustical domain with assumed DF-equalization.

Finally, the sidetone (self-masking) according to clause 5.4.3.2 is added to the acoustic echo signal to generate the stimulus for the listening test. The sidetone signal shall be generated from the source signal x(k) in the same way as for the test conditions.

NOTE: A more realistic, but also easy to reproduce echo path h(k) is under study at the time of publication of the present document (e.g. a typical transfer function between loudspeaker and microphone of a device). Nevertheless, since the two most important parameters of echo artefacts are level/attenuation and delay, the implementation of the reference conditions provide adequate anchoring of test subjects.

A.2 Reference Conditions for Double-Talk Listening Tests

The reference conditions for the double talk listening test are based on a similar processing schema as introduced in clause A.1. The extended principle for the generation is shown in Figure A.2.

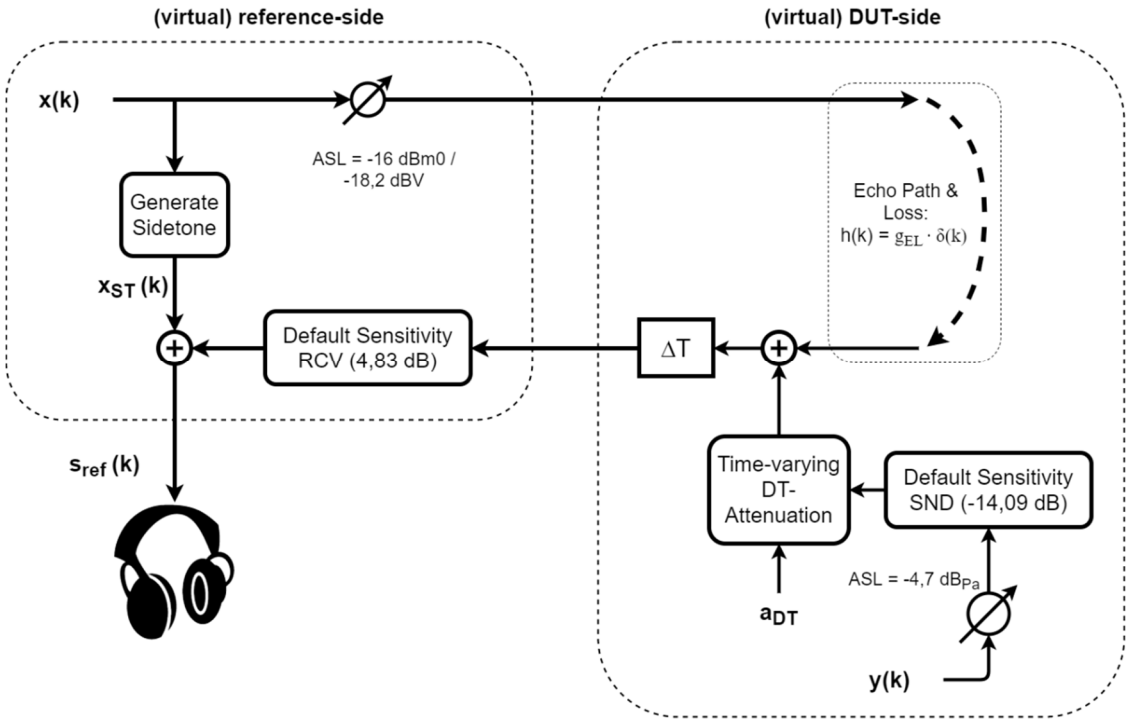

Figure A.2: Generation of reference conditions for double talk listening test

The source signals shall be the same as used for the test conditions (see clause 5.2), i.e. the downlink signal x(k) and the uplink signal y(k) are used (see Figure 4). For the generation of reference conditions, it is assumed that x(k) is sent from the (virtual) reference side with a default network level of -16 dBm0 and without any degradation to the (virtual) DUT-side. Here a simple echo path h(k) is implemented in order to create simple and deterministic echo impairments. The signal x(k) is simply attenuated by a given echo loss g_{EL} (typically specified in dB).

For the source signal y(k), it is assumed that it is played back over a mouth simulator with a default level of -4,7 dB_{Pa} and is acoustically inserted into the (virtual) DUT. Here the default sending sensitivity of -14,09 $dB_{V/Pa}$ (see clause 5.4.3.1) is applied. This (pseudo-)electrical signal is then processed by a simple double talk attenuation a_{DT} (typically expressed in dB). The attenuation is applied to y(k) in all signal segments, which are overlapping with the downlink signal x(k) (see Figure 2). In order to avoid too harsh level changes, a linear fade-in and fade-out of 100 ms shall be used for these segments.

This processed sending signal is then added to the linear echo signal. The superposition of both components is then delayed by a given value ΔT and transmitted back to the (virtual) reference-side. Here a default receiving sensitivity of 4,83 $dB_{Pa/V}$ (see clause 5.4.5) is applied in order to transform the echo signal to a (pseudo-)acoustical domain with assumed DF-equalization.

Finally, the sidetone (self-masking) according to clause 5.4.3.2 is added to the acoustic echo signal to generate the stimulus for the listening test. The sidetone signal shall be generated from the source signal x(k) in the same way as for the test conditions.

NOTE: A more realistic, but also easy to reproduce method for the attenuation during double talk is under study at the time of publication of the present document (e.g. signal manipulations in the time-frequency domain). Nevertheless, since the most important parameter for double talk impairments is the overall attenuation, the implementation of the reference conditions provide adequate anchoring of test subjects.

Annex B (normative): Simulation of reference sending terminal

B.1 Introduction

In order to simulate a high-quality reference terminal in sending direction, a suitable bandwidth/band-pass filter and level adjustment has to be defined. The following clauses of this Annex provide guidelines for implementation of this processing step.

B.2 Band-pass filters

To prepare source signals for NB, WB, SWB and FB applications, several FIR filters defined in Recommendation ITU-T G.191 [26] shall be used. Table B.1 show the short names of filters to use for each bandwidth. In case more than one short name is provided, the given filters have to be serially applied. Figure B.1 illustrates the corresponding transfer functions.

Table B.1: Band-pass filters of Recommendation ITU-T G.191 [26]

Bandwidth	Filter name(s)
NB	LP35 + MSIN
WB	P341
SWB	14KBP
FB	20KBP

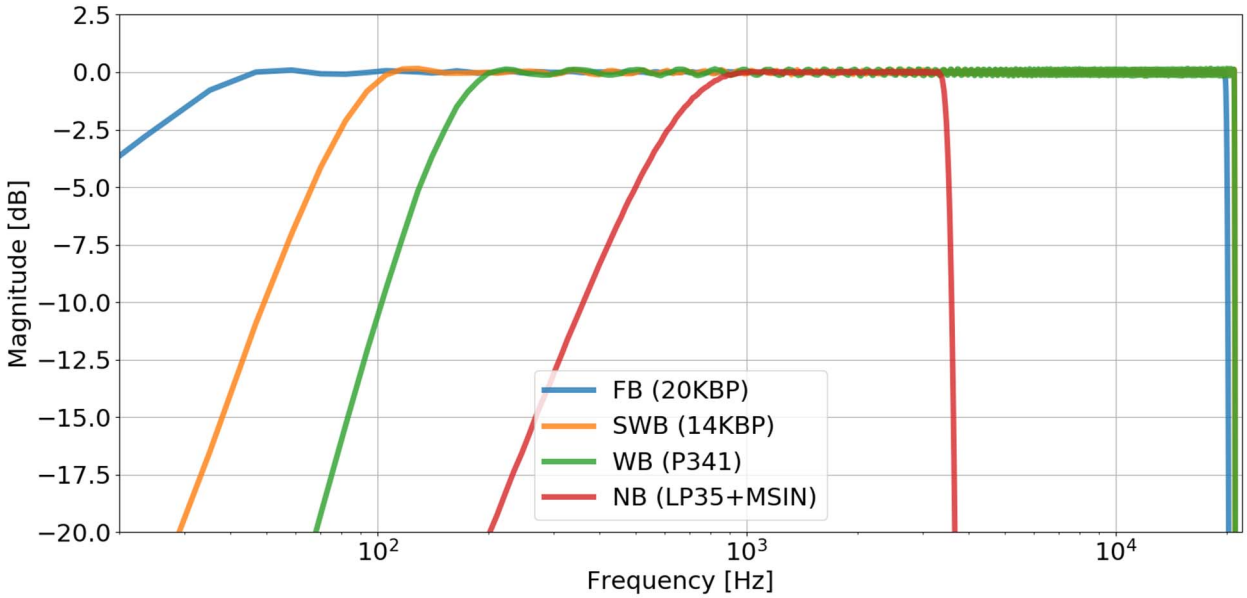

Figure B.1: Transfer functions of Recommendation ITU-T G.191 [26] filters

B.3 Sensitivity

Most measurement specifications for acoustic terminal testing define a Sending Loudness Rating (SLR) of 8 dB according to Recommendation ITU-T P.79 [27] as an optimum target. A transmission via a simulated reference system should thus also target at this SLR value. In consequence, the filters introduced in clause B.2 have additionally to be scaled by a constant sensitivity. This constant value is slightly different between NB and WB due to different weighting parameters. In case of SWB or FB transmissions, the SLR method and sensitivity value for WB shall be applied. The constant sensitivities for simulation of reference terminal in sending direction are:

- NB: -13,55 dB V/Pa for all applicable frequencies. Table B.2 provides the calculation with the corresponding weighting factors.
- WB, SWB, FB: -14,09 dB V/Pa for all applicable frequencies. Table B.3 provides the calculation with the corresponding weighting factors.

EXAMPLE: For a speech signal calibrated to an ASL_{MRP} = -4,7 dB_{Pa} according to Recommendation ITU-T P.56 [21], the simulated electrical signal for WB/SWB/FB would correspond to ASL_{POI} = -4,7 + (-14,09) = 18,79 dB_V. For NB, ASL_{POI} = -4,7 + (-13,55) = -18,25 dB_V. Both values are close to the commonly used reference level of -16 dBm0/-18,2 dB_V.

Table B.2: SLR calculation for NB with constant sensitivity

Band No.	Mid frequency [Hz]	Sensitivity [dB V/Pa]	W_R from Table 1/Recommendation ITU-T P.79 [27]	$x_i = 10^{0,1 \cdot 0,175 \cdot (S_i - W_i)}$
4	200	-13,55	76,9	0,026129133
5	250	-13,55	62,6	0,046491654
6	315	-13,55	62,0	0,047629388
7	400	-13,55	44,7	0,095636792
8	500	-13,55	53,1	0,068174977
9	630	-13,55	48,5	0,082058769
10	800	-13,55	47,6	0,085089310
11	1 000	-13,55	50,1	0,076935185
12	1 250	-13,55	59,1	0,053533421
13	1 600	-13,55	56,7	0,058969168
14	2 000	-13,55	72,2	0,031577300
15	2 500	-13,55	72,6	0,031072414
16	3 150	-13,55	89,2	0,015917505
17	4 000	-13,55	117,0	0,005192482
$SLR = -\frac{10}{0,175} \cdot \log_{10}\left(\sum_{i=4}^{17} x_i\right) = 8\ dB$				

Table B.3: SLR calculation for WB, SWB and FB with constant sensitivity

Band No.	Mid frequency [Hz]	Sensitivity [dB V/Pa]	W_R from Table A.2/Recommendation ITU-T P.79 [27]	$x_i = 10^{0,1 \cdot 0,175 \cdot (S_i - W_i)}$
1	100	-14,09	154,5	0,001121179
2	125	-14,09	115,4	0,005419073
3	160	-14,09	89,0	0,015700916
4	200	-14,09	77,2	0,025259515
5	250	-14,09	62,9	0,044944339
6	315	-14,09	62,3	0,046044207
7	400	-14,09	45,0	0,09245385
8	500	-14,09	53,4	0,065906007
9	630	-14,09	48,8	0,079327725
10	800	-14,09	47,9	0,082257404
11	1 000	-14,09	50,4	0,074374661
12	1 250	-14,09	59,4	0,051751745
13	1 600	-14,09	57	0,057006582
14	2 000	-14,09	72,5	0,030526358
15	2 500	-14,09	72,9	0,030038276
16	3 150	-14,09	89,5	0,015387746
17	4 000	-14,09	117,3	0,005019668
18	5 000	-14,09	157,3	0,001001555
19	6 300	-14,09	172,2	0,000549446
20	8 000	-14,09	181,7	0,000374692
$SLR = -\frac{10}{0,175} \cdot \log_{10}\left(\sum_{i=1}^{20} x_i\right) = 8\ dB$				

History

Document history		
V1.1.1	November 2020	Publication