ETSI TS 103 504 V1.1.1 (2020-07)

TECHNICAL SPECIFICATION

Speech and multimedia Transmission Quality (STQ); Methods and procedures for evaluating performance of voice-controlled devices and functions: far talk voice assistant devices

Reference

DTS/STQ-260

Keywords

accessibility, assessment, performance, quality,
usability

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

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

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

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

Copyright Notification

No part may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm except as authorized by written permission of ETSI.
The content of the PDF version shall not be modified without the written authorization of ETSI.
The copyright and the foregoing restriction extend to reproduction in all media.

© ETSI 2020.
All rights reserved.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members.
3GPP™ and **LTE™** are trademarks of ETSI registered for the benefit of its Members and
of the 3GPP Organizational Partners.
oneM2M™ logo is a trademark of ETSI registered for the benefit of its Members and
of the oneM2M Partners.
GSM® and the GSM logo are trademarks registered and owned by the GSM Association.

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

Voice Assistant Devices are devices that use voice recognition, speech synthesis and Natural Language Processing (NLP) to provide services through applications. Voice Assistant Devices may include, for example, smart speakers, cars, headsets, smartphones and IoT devices.

Due to the abundance of Voice Assistant Device form factors and use cases, the present document defines test methods and performance requirements for several acoustic environments and test setups. The home-like test environment (whether real, as described in clause 4.2.2, or simulated, as described in clause 4.2.3) is the default environment for qualifying devices intended for in-home use. Additional test setups are described for vehicular use in clause 4.3 and a wider range of simulated generic acoustic environments in clause 4.4. Test labs may choose to test a device under any or all applicable acoustic environments described in the present document.

The user experience with Voice Assistant Devices is largely dependent on the background noise, reverberation and speech material used in testing. The present document describes methods and procedures for evaluating performance of voice-controlled devices and functions.

1 Scope

The present document defines test methods, performance metrics and requirements for the voice assistance functionality of devices, including devices meant to be used in far talk conditions. The methods include definition of: input speech signals, positional relations of talkers and devices; acoustic environment characterization and reproduction, including background noise and reverberation; and collection of performance characteristics and statistical analyses.

The test methods, performance metrics and requirements of devices meant to be used in close and near talk conditions are out of scope.

The voice call functionality of voice assistant devices are out of scope and covered by ETSI ES 202 738 (narrowband) [i.1], ETSI ES 202 740 (wideband) [i.2] and ETSI TS 102 925 (super-wideband) [i.3].

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

[2] ETSI TS 103 557: "Speech and multimedia Transmission Quality (STQ); Methods for reproducing reverberation for communication device measurements".

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

[4] ANSI/ASA S12.2-2019: "Criteria For Evaluating Room Noise".

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

[6] Recommendation ITU-T P.1100: "Narrowband hands-free communication in motor vehicles".

[7] Recommendation ITU-T P.1110: "Wideband hands-free communication in motor vehicles".

[8] Recommendation ITU-T P.51: "Artificial mouth".

[9] Recommendation ITU-T P.341: "Transmission characteristics for wideband digital loudspeaking and hands-free telephony terminals".

[10] IEC 61672-1:2013: "Electroacoustics - Sound level meters - Part 1: Specifications".

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 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".

[i.2] 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".

[i.3] 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".

[i.4] ETSI ETS 300 807: "Integrated Services Digital Network (ISDN); Audio characteristics of terminals designed to support conference services in the ISDN".

[i.5] Recommendation ITU-T P Suppl. 16: "Guidelines for placement of microphones and loudspeakers in telephone conference rooms and for Group Audio Terminals".

[i.6] Teemu Halkosaari, Markus Vaalgamaa: "Directivity of human and artificial speech", from Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction, 8th and 9th June 2004 - Mainz, Germany.

[i.7] Papoulis, A. and Pillai, S. U. (2002): "Probability, Random Variables and Stochastic Processes", 4th ed. McGraw-Hill, New York.

[i.8] Brown, L.D., Cai, T.T., and DasGupta, A. (2001): "Interval Estimation for a Binomial Proportion", Statistical Science, 16(2):101-133.

[i.9] Moritz N., Goetze S., Appell JE. (2011): "Ambient Voice Control for a Personal Activity and Household Assistant". In: Wichert R., Eberhardt B. (eds) Ambient Assisted Living. Springer, Berlin, Heidelberg.

[i.10] Chu, W.T. and Warnock, A.C.C. (2002): "Detailed directivity of sound fields around human talkers," Technical Report, Institute for Research in Construction (National Research Council of Canada, Ottawa ON, Canada), pp. 1-47.

[i.11] Bozzoli, F., Viktorovitch, M., Farina, A. (2005): "Balloons of Directivity of Real and Artificial Mouth Used in Determining Speech Transmission Index", Presented at the 118th AES Convention, Barcelona, Spain, 28-31 May, 2005.

[i.12] Monson, B.B. and Hunter, E.J. (2012): "Horizontal directivity of low- and high-frequency energy in speech and singing", J. Acoust. Soc. Am. 132, pp. 433-441.

[i.13] Mehta, M., Johnson, J., Rocafort., J. (1998): "Architectural Acoustics: Principles and Design", 1st ed. Prentice Hall, Upper Saddle River, NJ.

[i.14] Diaz, C., and Pedrero, A. (2005): "The reverberation time of furnished rooms in dwellings", Applied Acoustics, 66:945-956.

[i.15] ISO 3382-2: "Acoustics -- Measurement of room acoustic parameters -- Part 2: Reverberation time in ordinary rooms".

[i.16] ETSI ES 202 396-1: "Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database".

[i.17] Recommendation ITU-T P.501, Amendment 1: "Test signals for use in telephonometry, Amendment 1".

[i.18] Recommendation ITU-R BS.1770-3: "Algorithms to measure audio programme loudness and true-peak audio level".

3 Definition of terms, symbols and abbreviations

3.1 Terms

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

close talk: voice-based interaction between the voice assistant device and the user at a distance lesser than 30 cm

far talk: voice-based interaction between the voice assistant device and the user at a distance greater than 2 m

near talk: voice-based interaction between the voice assistant device and the user at a distance between 30 cm and 2 m

task: speech-triggered functionality that a voice-assistant device executes

3.2 Symbols

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

dB	decibel
dB_{FS}	dB referenced to digital full scale
dB_{Pa}	dB referenced to 1 Pascal
dB_{SPL}	dB referenced to sound pressure level (2e-5 Pascal)
dB_{SPL}(A)	dB_{SPL} A-weighted according to IEC 61672-1 [10]
dB_{SPL}(C)	dB_{SPL} C-weighted according to IEC 61672-1 [10]

3.3 Abbreviations

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

ASR	Active Speech Recognition
CUR	Correct Utterance Ratio
DRR	Direct to Reverb Ratio
DUT	Device Under Test
FAR	False Acceptance Rate
FRR	False Rejection Rate
HATS	Head And Torso Simulator
HVAC	Heating Ventilation Air Conditioning
LOS	Line-Of-Sight
LUFS	Loudness Unit relative to Full Scale
MRP	Mouth Reference Point
NC	Noise Criteria
NLP	Natural Language Processing
PTCT	Perceived Task Completion Time
PWWD	Perceived Wake Word Delay
RIR	Room Impulse Response
RMS	Root-Mean-Square
RMSE	Root-Mean-Square Error
SNR	Signal to Noise Ratio
TCR	Task Completion Rate

TCS	Task Completion Score
TCT	Task Completion Time
TV	Television
WER	Word Error Rate
WWD	Wake Word Delay

4 Acoustic environments for testing

4.1 Introduction

Voice assistant devices come in a variety of form factors and have many intended use cases. This clause defines several acoustic environments for testing, each of which may be more applicable to some devices than others. The indoor acoustic environment described in clause 4.2 (whether real, as described in clause 4.2.2, or simulated, as described in clause 4.2.3) is the default environment for qualifying devices intended for in-home use (e.g. smart speakers, soundbars, etc.). Furthermore, the vehicular acoustic environment described in clause 4.3 is applicable to the testing of in-car devices (e.g. voice-controlled navigation systems, smart dashboards, etc.). Lastly, the generic test environments presented in clause 4.4 provide a wider range of simulated adverse acoustic conditions under which a voice assistant device may be tested or developed.

Test labs may choose to test a device under any or all applicable acoustic environments presented in this clause. Devices shall be qualified using all required conditions and methods for the selected acoustic environment described in this clause and throughout the present document.

4.2 Indoor acoustic environments

4.2.1 General considerations

4.2.1.1 Introduction

The indoor acoustic environments described in this clause are applicable to devices, which are envisioned to be used inside the home (e.g. smart speakers, soundbars, etc.).

Many voice assistant devices are designed for indoor use cases. Such devices are subject to a variety of reverberation characteristics and interfering background noise conditions unique to the acoustic environment in which the device is being used. Such acoustic variables impact the performance of device algorithms including wake word detection, speech recognition, and task completion.

This clause provides information on the reverberation characteristics and interfering background noise conditions that should be considered in the testing of indoor voice assistant devices. Two test environments are described, within which the recommended acoustic environments for testing can be realized. First, a home-like test environment is defined, which emulates a realistic "open floor plan" living space with kitchen and living room characteristics. Then, methods for the lab-based simulation of acoustic environments are presented.

NOTE: The usage of electrically or digitally inserted signals instead of a complete acoustic setup may be used for offline testing of voice assistant devices. In such cases, all recordings of test material in an acoustic environment are captured with one or more measurement microphone(s) instead of a DUT. These pre-recorded signals can then be used for automated and offline testing of speech processing components.

4.2.1.2 Reverberation characteristics

Reverberation is a source of variation in the performance of ASR systems [i.9]. The reverberation characteristics of indoor spaces can vary substantially depending on the treatment of surfaces and volume of the environment.

The most common parameter used to characterize the reverberant behaviour of a room is the reverberation time T_{60}. This is the required time for an energy level decay of 60 dB. Due to the measurement noise floor, T_{60} is often estimated by measuring a decay of 20 or 30 dB and extrapolating the time for a decay of 60 dB. Methods to estimate T_{60} from T_{20} and T_{30} measurements are provided by reference in ETSI TS 103 557 [2].

In addition to T_{60}, the early to late reverberation ratio, or *clarity index*, further qualifies the reverberation characteristics of a space. This metric is motivated by the perception of multi-path sounds, which arrive within a short time interval, as a single source. It indicates how much of the indirect-path energy is perceived as coloration of the direct-path as opposed to reverberance or echo. The *clarity index* is calculated according to equation 4.2.1.2-1.

$$C_{\frac{n_e}{f_s}} = 10log_{10}\left(\frac{\sum_{n=0}^{n_e} h^2(n)}{\sum_{n=n_e+1}^{\infty} h^2(n)}\right) in \text{ dB} \qquad (4.2.1.2\text{-}1)$$

In equation 4.2.1.2-1, n_e/f_s is typically chosen as either 50 ms or 80 ms. For the purposes of the present document, the values n_e/f_s = 50 ms and f_s = 48 kHz are used.

The T_{60} and C_{50} of indoor environments vary depending on the volume of the room, the material and treatment of large surfaces, and the presence of acoustically absorbing features including carpet, drapes, furniture, etc. Annex E presents an investigation of T_{60} measurements made in several indoor environments.

4.2.1.3 Interfering background noise

Voice assistant devices are expected to function in the presence of day-to-day background noise. These interferences may impact the ability of a voice assistant device to detect wake words and/or accurately recognize a target request. Interfering background noises may be stationary or non-stationary with respect to both spectral composition and spatial positioning. Furthermore, the frequency components of an interfering background noise may be band-limited or cover the entire audible spectrum. Finally, noise scenarios may include a single discrete source or multiple directional and/or diffuse sources. It is important to assess the immunity of voice assistant devices to a variety of interfering background noise types and scenarios.

4.2.2 Home-like test environment

4.2.2.1 Introduction

This clause defines criteria for a home-like test environment that can be used to emulate the acoustic environments in which voice assistant devices are commonly used. The goal of the home-like test environment is to provide a space with reasonable reproducibility and repeatability, while offering some level of flexibility in construction and material selection. Furthermore, the test environment is designed to realize a variety of realistic reverberation characteristics and interfering background noise conditions.

4.2.2.2 Floor plan and room requirements

This clause contains general requirements for the home-like test environment, which shall be simultaneously met. The test environment emulates an "open floor plan" apartment with a kitchen and a living room area. Figure 4.2.2.2-1 presents the home-like test environment floor plan. The placements and dimensions of furniture, DUT positions, target talker locations, and noise source locations are identified. Furthermore, Table 4.2.2.2-1 provides the relative placement information for each pair of Talker and DUT positions.

- Furniture: The test space shall include furniture and decorative elements, providing a mixture of acoustically reflective and absorptive surfaces that is representative of a home environment. Furniture shall have the dimensions and be placed according to the layout in Figure 4.2.2.2-1.

- Room dimensions:

  - Internal volume: The test space shall have an internal volume equal or greater than 70 m^3 and equal or less than 90 m^3 when not occupied with furniture, subject to the following constraints:

    - Internal length: The test space shall have an internal length equal or greater than 6 m and equal or less than 9 m.

  - Internal width: The test space shall have an internal width equal or greater than 3,5 m and equal or less than 5 m.
  - Internal height: The test space shall have an internal height equal or greater than 2,2 m and equal or less than 3,2 m.

- Floor surface: The test space should have an acoustically reflective floor surface (i.e. no carpet). The floor surface should be covered by rugs to reduce the T_{60} time where necessary.
- Ceiling surface: The test space should have an acoustically absorptive ceiling with a mean absorption coefficient of 0,5 between 250 Hz and 2 kHz.
- Ambient noise floor: The home-like test environment shall have an ambient noise floor equal to or less than NC 30. This requirement applies to all locations where the DUT will be tested. Measurement of Noise Criteria is described in ANSI/ASA S12.2-2019 [4].

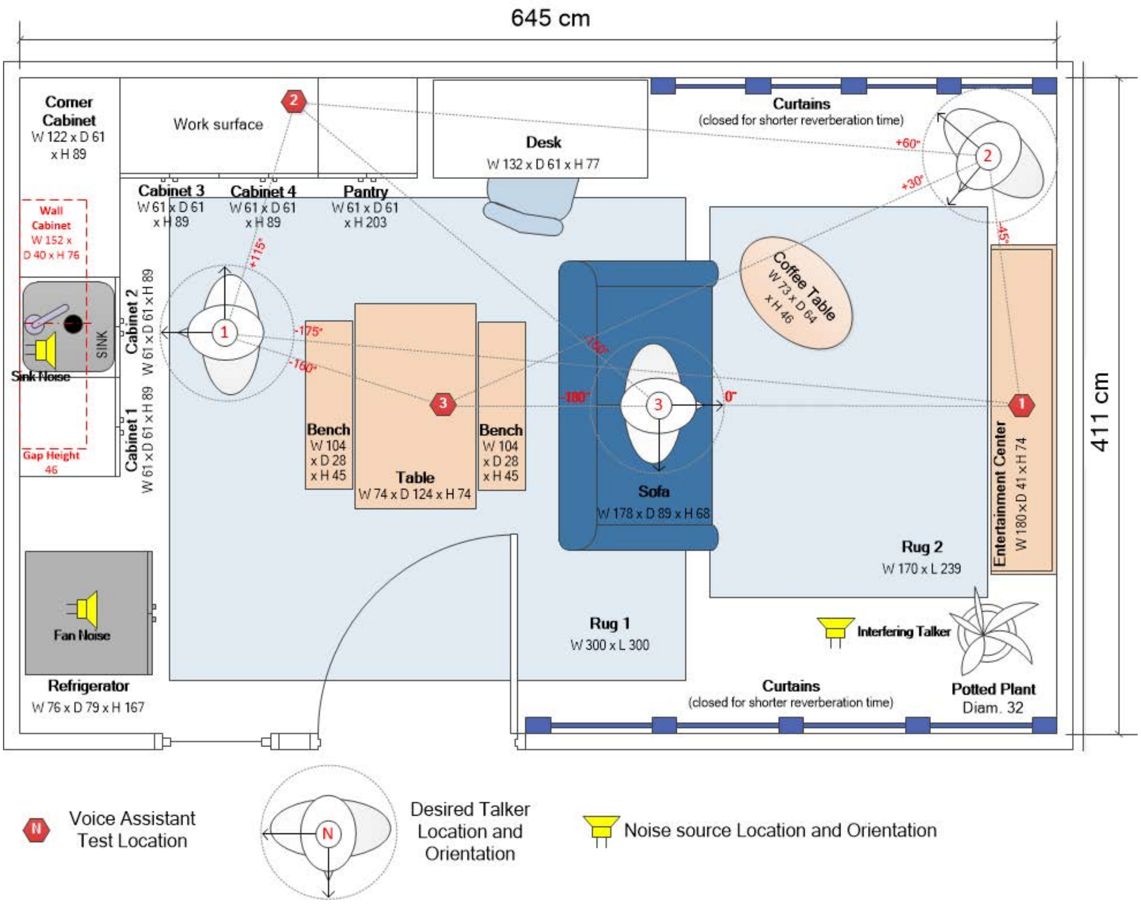

645 cm
411 cm
Corner Cabinet W 122 x D 61 x H 89
Work surface
Desk W 132 x D 61 x H 77
Curtains (closed for shorter reverberation time)
+60°
+30°
45°
Cabinet 3 W 61 x D 61 x H 89
Cabinet 4 W 61 x D 61 x H 89
Pantry W 61 x D 61 x H 203
Wall Cabinet W 152 x D 40 x H 76
+115°
Coffee Table W 73 x D 64 x H 46
Cabinet 2 W 61 x D 61 x H 89
SINK
Sink Noise
-175°
-160°
-100°
-180°
0°
Bench W 104 x D 28 x H 45
Table W 74 x D 124 x H 74
Bench W 104 x D 28 x H 45
Sofa W 178 x D 89 x H 68
Entertainment Center W 180 x D 41 x H 74
Cabinet 1 W 61 x D 61 x H 89
Gap Height 46
Rug 2 W 170 x L 239
Rug 1 W 300 x L 300
Fan Noise
Refrigerator W 76 x D 79 x H 167
Interfering Talker
Curtains (closed for shorter reverberation time)
Potted Plant Diam. 32
1
2
3
N
Voice Assistant Test Location
Desired Talker Location and Orientation
Noise source Location and Orientation

Figure 4.2.2.2-1: Home-like test environment floor plan (dimensions are in cm)

Table 4.2.2.2-1: Relative placement information for Talker and DUT positions

Talker Position	DUT Position	Talker Height (see note 1)	DUT Height (see note 2)	Distance (see note 3)	Relative Height (see note 4)	Azimuth, Elevation (see note 5)
1	1	150 cm	75 cm	520 cm	75 cm	-175°, -8°
2	1	150 cm	75 cm	180 cm	75 cm	-45°, -25°
3	1	100 cm	75 cm	250 cm	25 cm	0°, -5°
1	2	150 cm	90 cm	140 cm	60 cm	115°, -25°
2	2	150 cm	90 cm	450 cm	60 cm	60°, -8°
3	2	100 cm	90 cm	290 cm	10 cm	-150°, -2°
1	3	150 cm	75 cm	180 cm	75 cm	-160°, -25°
2	3	150 cm	75 cm	390 cm	75 cm	30°, -11°
3	3	100 cm	75 cm	130 cm	25 cm	-180°, -11°
Tolerance		±5 cm	±5 cm	±10 cm	±10 cm	±5°

NOTE 1: Talker height is measured from the floor to the HATS/mouth simulator MRP.
NOTE 2: DUT height is measured from the floor to the base of the DUT.
NOTE 3: Distance is the length of a straight vector from the HATS/mouth simulator MRP to the base surface at the DUT position.
NOTE 4: Relative height is the height difference between the HATS/mouth simulator MRP and the DUT base.
NOTE 5: Azimuth and elevation angles are measured from the MRP to the base of the DUT. Azimuth angles are positive in the clockwise direction and elevation angles are positive inclination and negative declination.

4.2.2.3 Reverberation characteristics

The home-like test environment shall allow for room adjustments to meet several acoustic conditions for testing. Figure 4.2.2.3-1 demonstrates an alternative layout where Rug 1 and Rug 2 are removed from the room and both sets of curtains are opened. This layout increases reverberation time compared to the layout presented in Figure 4.2.2.2-1.

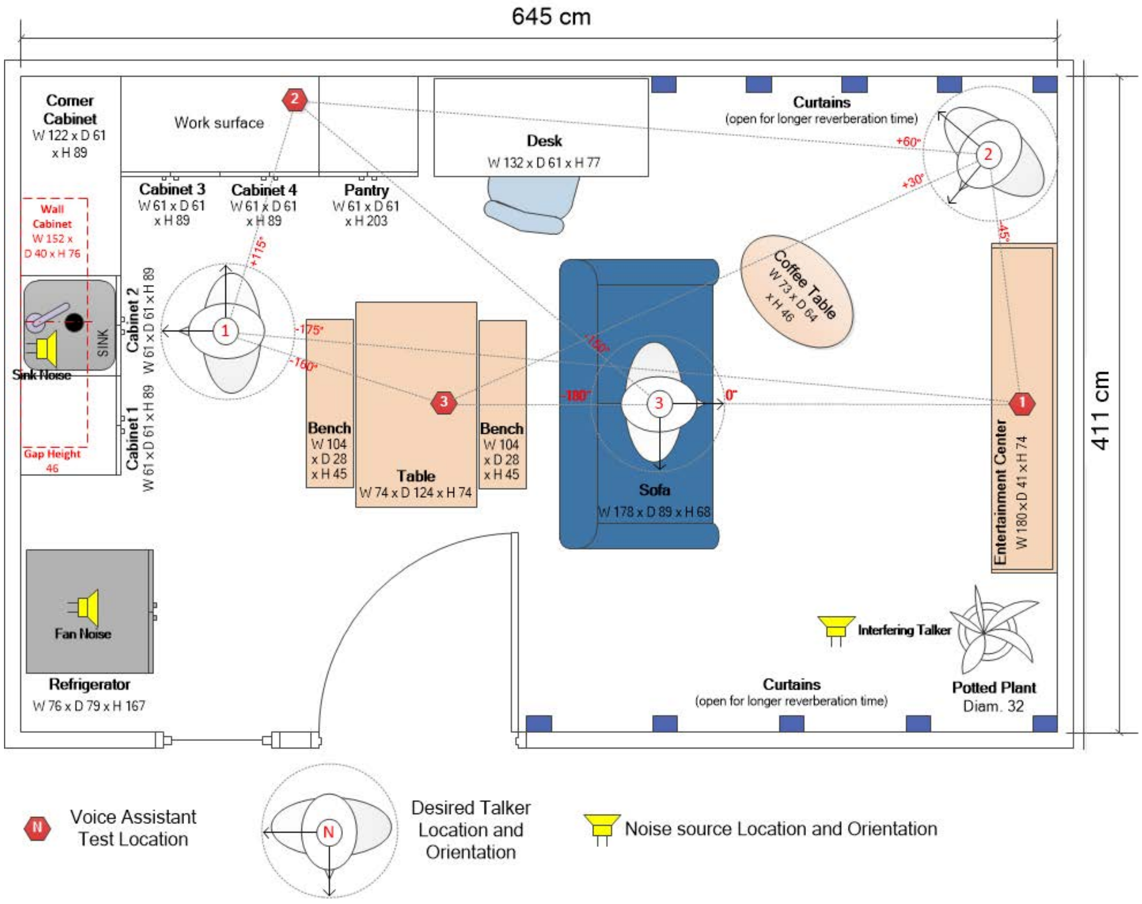

645 cm
411 cm
Corner Cabinet W 122 x D 61 x H 89
Work surface
Desk W 132 x D 61 x H 77
Curtains (open for longer reverberation time)
+60°
+30°
-45°
2
Cabinet 3 W 61 x D 61 x H 89
Cabinet 4 W 61 x D 61 x H 89
Pantry W 61 x D 61 x H 203
Wall Cabinet W 152 x D 40 x H 76
+115°
Coffee Table W 73 x D 64 x H 46
SINK
Sink Noise
Cabinet 2 W 61 x D 61 x H 89
1
-175°
-160°
-150°
3
-180°
0°
1
Cabinet 1 W 61 x D 61 x H 89
Gap Height 46
Bench W 104 x D 28 x H 45
Table W 74 x D 124 x H 74
Bench W 104 x D 28 x H 45
Sofa W 178 x D 89 x H 68
Entertainment Center W 180 x D 41 x H 74
Fan Noise
Refrigerator W 76 x D 79 x H 167
Interfering Talker
Curtains (open for longer reverberation time)
Potted Plant Diam. 32
N
Voice Assistant Test Location
N
Desired Talker Location and Orientation
Noise source Location and Orientation

Figure 4.2.2.3-1: Home-like test environment alternative layout (dimensions are in cm)

Table 4.2.2.3-1 provides a set of required acoustic conditions for testing in the home-like environment. DUTs shall be tested in each of the acoustic conditions presented in Table 4.2.2.3-1 for DUT Position 1 (entertainment center) and Talker Position 3 (couch). The acoustic conditions should be tested for other position combinations.

Table 4.2.2.3-1: Acoustic conditions for testing

Acoustic Condition	Reverberation Time (T_{60} in seconds)						Clarity Index (C_{50} in dB)
	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	
Condition 1	0,4	0,3	0,3	0,3	0,3	0,3	10,5
Condition 2	0,5	0,45	0,45	0,5	0,5	0,5	6,0
Tolerance	±0,1	±0,05	±0,05	±0,05	±0,05	±0,1	±2,5

Test labs shall demonstrate the compliance of their home-like test environment to the acoustic conditions presented in Table 4.2.2.3-1. The octave-band T_{60} and overall C_{50} shall be measured and averaged between each target talker and DUT position to determine compliance.

NOTE: Measurement of room acoustic qualities including ambient noise floor, reverberation time, and clarity index are generally made using a mono, omni-directional reference microphone. HATS, directional, or multi-channel microphones should not be used.

4.2.2.4 Interfering background noise

The home-like environment uses discrete noise sources representing typical noise sources that are found in a home environment. The discrete noise sources are positioned according to the layout in Figure 4.2.2.2-1.

For the kitchen sink and interfering talker noise sources, near-field monitor loudspeakers with a 3" to 5" woofer diameter shall be used as the sound source. The near-field monitors shall have a flat frequency response of ±2,5 dB between 80 Hz and 20 kHz.

Due to low frequency components in the refrigerator fan noise source, a near-field monitor loudspeaker with a 5" to 7" woofer diameter shall be used as the sound source. The near-field monitor shall have a flat frequency response of ±2,5 dB between 50 Hz and 20 kHz.

For all noise sources, it is recommended but not required to use a coaxial driver, to have better control of the acoustic axis of the loudspeaker.

The level of each noise source shall be calibrated at DUT Position 1, 10 cm above the entertainment center, in the absence of a DUT. Levels shall be calibrated to the values shown in Table 4.2.2.4-1 according to the acoustic condition.

Table 4.2.2.4-1: Interfering noise sources for testing of voice assistants

	Noise Condition	Description	Height	dB_{SPL}(A)		
				Acoustic Condition 1	Acoustic Condition 2	Acoustic Condition 3
1	Ambient Noise Floor	Only the ambient noise floor of the test space is present.	-	< 30	< 30	[TBD]
2	Kitchen Sink	A loudspeaker located at the position of the sink in the test space and simulating a kitchen sink running water (high frequency dominated stationary noise).	1,1 m	56 ± 1	59 ± 1	[TBD]
3	Refrigerator Fan	A loudspeaker located in the lower left corner in the test space simulating a refrigerator fan (low frequency dominated stationary noise).	1,8 m	39 ± 1	39 ± 1	[TBD]
4	Interfering Talker	A loudspeaker located at a position in the living space simulating an interfering talker (non-stationary noise).	1,5 m	57 ± 1	60 ± 1	[TBD]

	Noise Condition	**Description**	**Height**	**dB$_{\text{SPL}}$(A)**		
				Acoustic Condition 1	**Acoustic Condition 2**	**Acoustic Condition 3**
5	Kitchen Sink + Refrigerator Fan + Interfering Talker	In this scenario, the interfering talker content, sink noise and refrigerator noise are presented simultaneously (complex noise scene).	-	Calibrate each noise source individually to the levels provided above.		

DUTs shall be tested under the interfering noise conditions from Table 4.2.2.4-1 for DUT Position 1 (entertainment center), Talker Position 3 (couch), and Acoustic Condition 2. Interfering noise scenarios should be tested for other combinations of DUT position, talker position, and acoustic condition.

Generic, monophonic audio files for interfering noise playback can be retrieved from https://docbox.etsi.org/stq/Open/TS%20103%20504/Interfering%20noise%20source%20files.

4.2.3 Lab-based simulation of acoustic environments

4.2.3.1 Introduction

This clause defines criteria for lab-based simulation of acoustic environments that can be used wherever high repeatability and reproducibility in testing are desired. Lab-based simulation shall feature both reverberation and background noise simulation systems for the realization of various reverberation characteristics and interfering background noise scenarios.

When testing voice assistant devices, reverberation simulation and background noise reproduction are achieved in tandem. For accurate and coherent simulation of an acoustic environment, the multi-channel noise file used for background noise simulation and the RIR used for reverberation reproduction shall be collected with the same microphone array in the same acoustic environment and at the same location and orientation.

4.2.3.2 Room requirements

Test room requirements are described in ETSI TS 103 557 [2]. An example loudspeaker setup for a simulation arrangement for far talk devices is shown in Figure 4.2.3.2-1.

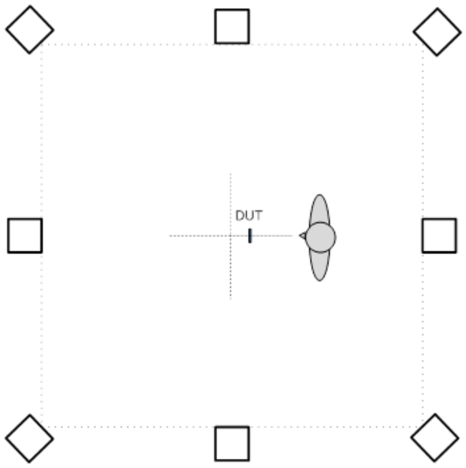


Figure 4.2.3.2-1: Example loudspeaker setup for lab-based simulation of acoustic environments

4.2.3.3 Simulation of reverberation characteristics

The loudspeaker equalization/calibration process and reverberation simulation procedure in a lab-based test environment are defined in ETSI TS 103 557 [2].

Clause A.2 of ETSI TS 103 557 [2] presents a database of multi-channel room impulse responses for the lab-based simulation of a home-like test environment constructed to the specifications in clause 4.2.2 RIRs are provided between each target talker and DUT position shown in Figure 4.2.2.2-1 in the achievable acoustic conditions described in clause 4.2.2.3.

DUTs shall be tested with simulated reverberation covering all acoustic conditions presented in Table 4.2.2.3-1 for one relative positioning of Talker and DUT. The acoustic conditions should be tested for multiple Talker-DUT relative positionings. Further information on Talker and DUT positioning in the lab-based test environment is provided in clauses 5.1.3 and 5.2.2, respectively.

4.2.3.4 Simulation of interfering background noise

The loudspeaker equalization/calibration process and background noise simulation procedure in a lab-type environment are described in ETSI TS 103 224 [1].

Clause 8.4 of ETSI TS 103 224 [1] presents a database of multi-channel recordings for the lab-based simulation of interfering background noise conditions presented in Table 4.2.2.4-1 (excluding the ambient noise floor conditions). The recordings were made in a home-like test environment constructed to the specifications in clause 4.2.2, at each of the DUT positions presented in Figure 4.2.2.2-1, and in the achievable acoustic conditions described in clause 4.2.2.3.

DUTs shall be tested under simulated background noise conditions covering all of the scenarios presented in Table 4.2.2.4-1 measured in a space complying with Acoustic Condition 2. The DUT should be tested against noise types measured in other acoustic conditions. Each background noise condition shall be tested for one relative positioning of Talker and DUT and should be tested for multiple relative positionings. Further information on Talker and DUT positioning in the lab-based test environment is provided in clauses 5.1.3 and 5.2.2, respectively.

4.3 Vehicular acoustic environments

The vehicular acoustic environments described in this clause are applicable to devices, which are envisioned to be used inside the car (e.g. voice-controlled navigation systems, smart dashboards, etc.).

Multiple setups and procedures of the present document are also applicable to vehicle-mounted voice assistant devices. Here, the context is related to more specific tasks (like e.g. navigation or phone calls).

All tests shall be carried out inside the car cabin, as shown in Figure 4.3-1. Since most of the voice assistant functionalities in car cabins are driver-related, corresponding tests shall be carried out with a HATS or mouth simulator positioned at the driver's position. Mounting instructions should follow the guidelines described in clause 7.1 of Recommendation ITU-T P.1100 [6] or Recommendation ITU-T P.1110 [7]. Any other seating (co-driver, back seats) may be tested if applicable (e.g. a dedicated voice assistant device located at a certain seat).

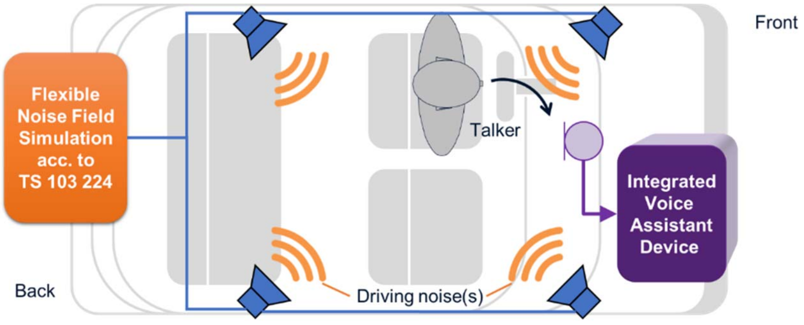

Front
Flexible Noise Field Simulation acc. to TS 103 224
Talker
Integrated Voice Assistant Device
Back
Driving noise(s)

Figure 4.3-1: Evaluation setup for vehicle-mounted voice assistant device

The recording and reproduction of driving noises shall be realized according to clause 7 of ETSI TS 103 224 [1]. At least two driving conditions with medium and high speed shall be considered. As a reference, an additional performance evaluation in silence (no driving noise) is recommended.

Variations of vehicle speed, the setting of Heating/Ventilation/Air Conditioning (HVAC) and window position may be used to generate different evaluation scenarios. All these parameters shall be reported for each noise recording. Annex D of Recommendation ITU-T P.1100 [6] and Recommendation ITU-T P.1110 [7] provides a set of user scenarios, which can be used for testing.

4.4 Generic acoustic environments

4.4.1 Introduction

The indoor acoustic environment described in clause 4.2 (either real or simulated) represents the usage of voice assistant devices in an "open floor plan" home-like environment. The multiple talker and DUT positions and orientations provide a good insight into the device performance for everyday use inside the home.

However, the acoustic conditions presented in clause 4.2 are rather moderate, especially regarding the reverberation. The simulated generic acoustic environments described in this clause may be used to test voice assistant devices in a wider range of adverse conditions. In contrast to clause 4.2, these are purely simulated acoustic conditions based on pre-recorded acoustic scenarios according to ETSI TS 103 557 [2].

4.4.2 Room requirements

Test room requirements are identical to the ones described in clause 4.2.3.2.

4.4.3 Simulation of reverberation characteristics

The loudspeaker equalization/calibration process and reverberation simulation procedure in a lab-based test environment are defined in clause A.1 of ETSI TS 103 557 [2]. These conditions use the asymmetric microphone array configuration according to ETSI TS 103 224 [1]. DUTs shall be tested with simulated reverberation covering all acoustic conditions presented in Table 4.4.3-1.

Table 4.4.3-1: Reverberation scenarios used for generic environments

Title	RT60 [ms]	Identifier in Table A.1 of ETSI TS 103 557 [2]
Bathroom	583	5
Kitchen	547	4
Livingroom	388	2
Officeroom	544	3

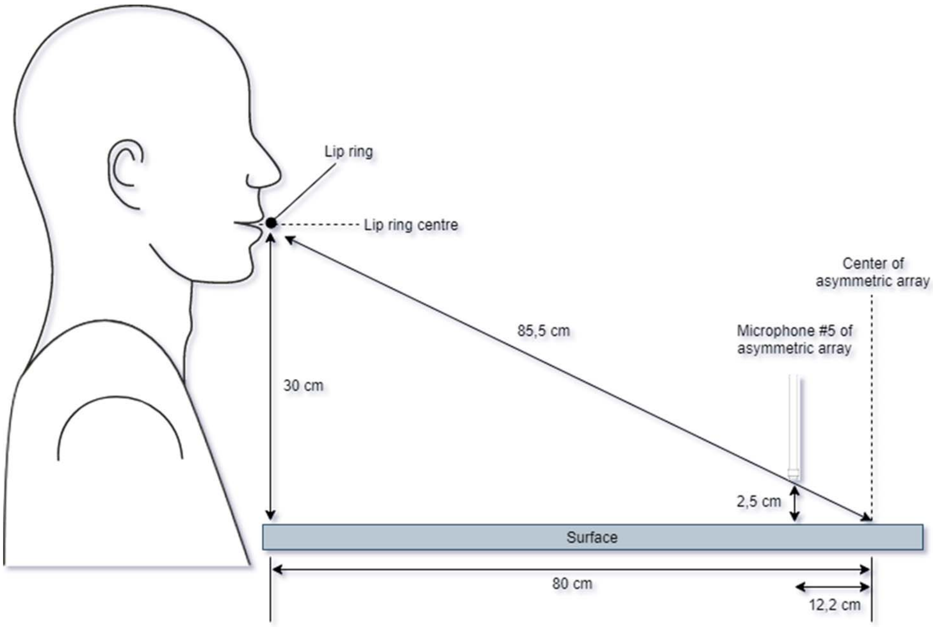


Figure 4.4.3-1: Talker to DUT orientation for generic acoustic environments

All room impulse responses refer to the setup of group audio terminals as described in clause 4.2.4 of Recommendation ITU-T P.341 [9]. For all scenarios described in Table 4.4.3-1, the setup for talker and DUT as shown in Figure 4.4.3-1 is used for the reproduction of reverberation.

DUTs shall be tested with certain combinations of simulated background noise conditions and reverberation scenarios, which are described in clause 4.4.5.

4.4.4 Simulation of interfering background noise

The loudspeaker equalization/calibration process and background noise simulation procedure for the generic test environments are included in clause 8 (Table 8.2) of ETSI TS 103 224 [1].

The noise types according to Table 4.4.4-1 shall be used for testing. These recordings were made in several acoustic environments which are identical to the reverberation scenarios of Table 4.4.3-1.

DUTs shall be tested with certain combinations of simulated background noise conditions and reverberation scenarios, which are described in clause 4.4.5.

Table 4.4.4-1: Interfering noise conditions used for generic environments

Title	Description	Average Level of Microphone #5 (dB)
Bathroom	Recording of a bathroom scenario, including shower, razor, sink, toilet flushing, hairdryer	72,5
Bathroom_withMusic	Same as "Bathroom", but with additional playback of radio broadcast	74,1
Kitchen	Recording of a kitchen scenario, including range hood, frying, tableware rattle, mixer, sink, knife on cutting board	67,3
Livingroom	Recording of a living room scenario, including vacuum cleaner, clink of drinking glass, coughing, TV, cleaning up	64,4
Officeroom	Recording of an office room scenario, including projector, writing by hand and keyboard, phone ringing, phone call, outside noise	54,9

4.4.5 Definition of generic conditions

Table 4.4.5-1 defines meaningful combinations of simulated noise conditions (see Table 4.4.4-1) and reverberation scenarios (see Table 4.4.3-1). For measurements without noise simulation, i.e. under silence conditions, no noise type is provided. For measurements without simulation of reverberation, no scenario is provided. DUTs shall be tested for all acoustic generic conditions G00 to G09 listed in Table 4.4.5-1. G00 acts as a baseline condition, in which the DUT is tested without simulated reverberation or the addition of interfering background noise.

Table 4.4.5-1: Interfering noise conditions used for generic environments

ID	Reverberation according to Table 4.4.3-1	Noise type according to Table 4.4.4-1
G00	-	-
G01	Bathroom	-
G02	Bathroom	Bathroom
G03	Bathroom	Bathroom_withMusic
G04	Kitchen	-
G05	Kitchen	Kitchen
G06	Livingroom	-
G07	Livingroom	Livingroom
G08	Officeroom	-
G09	Officeroom	Officeroom

5 Test setup and configurations

5.1 Desired talker

5.1.1 Introduction

For testing of voice assistant devices, either actual subjects, a HATS, or a standalone mouth simulator shall be used as the desired talker sound source. Voice assistant testing involves extensive repeat of wake words and/or voice commands, for which the use of pre-recorded speech materials reproduced with a HATS/mouth simulator is generally preferred over actual subjects.

Actual subjects may be used in situations where there are concerns on test result bias, e.g. due to the potential for pre-recorded speech to be used in the voice recognition model training. Actual subjects may also be desired in situations where there is a need for very accurate mouth directivity conditions. More information on speech and HATS/mouth simulator directivity is provided in annex D.

5.1.2 Mouth directivity and orientation

Reproduction directivity is an important consideration when choosing an appropriate loudspeaker to emulate humans as sound sources. Humans have various directivity characteristics that exercise the actual acoustic propagation characteristics that a device will be exposed to in practical use. For example, sound source directivity is important when attempting to:

1) replicate situations where the talker is not speaking towards the voice assistant device; and

2) properly trigger room acoustic reflections.

By design and standard, mouth simulator directivity is defined for the region in front of the lip plane [8]. Furthermore, HATS mouth directivity is defined for the region in front of the lip plane and few points in the rear plane [3].

Annex D presents a study comparing the directivity characteristics of humans, mouth simulators, HATS, and a conventional spherical loudspeaker. The HATS and mouth simulator demonstrate directivity properties more representative of a human talker than the loudspeaker. However, both HATS and mouth simulators deviate from human speech directivity, particularly in the rear plane and at high frequencies. Nevertheless, where testing with real humans is impractical, directivity properties of a HATS/mouth simulator offer a better match to humans than a conventional loudspeaker.

Depending on the number of DUT microphones and their layout, the orientation of the target talker with respect to the DUT may impact the device's performance. Example cases of talker orientation with respect to a DUT are shown in Figure 5.1.2-1. Please refer to clause 5.1.3 for required target talker orientations for testing.

In addition to target talker orientation, variations in DUT orientation through e.g. rotation of the device may impact performance. Please refer to clause 5.2.3 for more information.

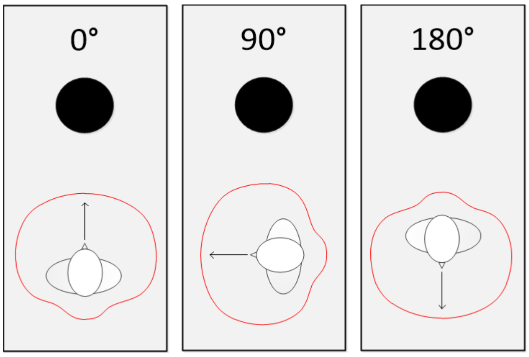


Figure 5.1.2-1: Example target talker orientations with respect to the DUT

5.1.3 HATS/mouth simulator setup

For both home-like and lab-based test environments, a HATS, according to Recommendation ITU-T P.58 [3], or a mouth simulator, according to Recommendation ITU-T P.51 [8], shall be used. Based on the improved directivity characteristics demonstrated in annex D, a HATS should be used.

Home-like test environment

- HATS/mouth simulator height: The HATS/mouth simulator shall be positioned such that the MRP is at a height of 150 cm from the floor for standing Talker Position 1 and Position 2, and 100 cm from the floor for sitting Talker Position 3. The elevation of the DUT with respect to the HATS/mouth simulator is determined by the relative height and distance between the Talker and DUT positions presented in Table 4.2.2.2-1. DUTs shall be tested under each of the three elevation conditions presented in Table 5.1.3-1.

Table 5.1.3-1: Elevation conditions based on relative height and distance between talker and DUT

Elevation Condition 1	0º
Elevation Condition 2	-10º
Elevation Condition 3	-25º
Tolerance	±5º

- HATS/mouth simulator position: HATS/mouth simulator positions in the home-like test environment are defined in Figure 4.2.2.2-1. Each of the three Talker positions shall be tested at least once.

- HATS/mouth simulator orientation: The HATS/mouth simulator shall be oriented relative to the DUT according to the azimuth angles defined in Table 5.1.3-2 for Talker and DUT positions shown in Figure 4.2.2.2-1. The 0º azimuth angle is defined for the condition where the DUT bisects the HATS/mouth simulator median plane. For smart speakers, at least six unique orientations (including 0º) from Table 5.1.3-2 shall be tested. For soundbars, the three talker orientations for DUT Position 1 in Table 5.1.3-2 shall be tested. See clause 5.2.2 for more information.

Table 5.1.3-2: Desired talker orientation with respect to the DUT in the home-like test environment

	DUT Position 1	DUT Position 2	DUT Position 3
Talker Position 1	-175º	+115º	-160º
Talker Position 2	-45º	+60º	+30º
Talker Position 3	0º	-150º	-180º

- HATS/mouth simulator equalization: The HATS/mouth simulator shall be equalized to a flat frequency response at the MRP according to the equalization procedures defined in Recommendation ITU-T P.58 [3].

Lab-based test environment

- HATS/mouth simulator height: The HATS/mouth simulator height relative to the DUT shall be adjusted based on the reverberation simulation condition according to the specifications provided in clause 5.2.1 in ETSI TS 103 557 [2]. The DUT shall be tested for each of the elevation conditions presented in Table 5.1.3-1.

- HATS/mouth simulator orientation: The HATS/mouth simulator shall be oriented relative to the DUT based on the reverberation simulation condition according to the specifications provided in clause 5.2.1 of ETSI TS 103 557 [2]. For smart speakers, at least six unique talker orientations (including 0º) from Table 5.1.3-2 shall be tested. For soundbars, the three talker orientations listed under DUT Position 1 in Table 5.1.3-2 shall be tested. See clause 5.2.2 for more information.

- HATS/mouth simulator equalization: The HATS/mouth simulator shall be equalized to a flat frequency response at the MRP according to the equalization procedures defined in Recommendation ITU-T P.58 [3].

Car cabin test environment

The HATS/mouth simulator shall be setup according to clause 4.3. Testing of the co-driver position is recommended.

Generic test environment

The HATS/mouth simulator shall be setup according to clause 4.4, i.e. in group audio terminal position as shown in Figure 4.4.3-1.

5.1.4 Speech material and level

5.1.4.1 Introduction

When testing the automatic speech recognition functionality of a voice assistant device, the speech test materials shall include different types of suitable utterances/words/sentences, depending on the DUT/type of measurement, representing typical commands that are expected to be uttered to the device. The speech material is typically composed of a wake word, followed by a short pause, followed by a question or command to the voice assistant.

EXAMPLES:

- <wake word>, <pause>, how is the weather today in Tokyo?
- <wake word>, <pause>, play Jazz music.
- <wake word>, <pause>, who is the president of Namibia?
- <wake word>, <pause>, who is the wealthiest person in the world?
- <wake word>, <pause>, what is the capital of Bolivia?
- <wake word>, <pause>, what is three times three?

Use of closed form questions with unique correct responses may enable test automation if the response can be obtained in textual format. For example, in examining the answer to the question of Bolivia's capital, the word "Sucre" would be expected to be present. In contrast, the command to play Jazz music elicit different responses from the device.

NOTE: A database of wake words and/or questions may be provided by the owner of a voice assistant service.

5.1.4.2 Speech sound pressure level

The wake word and question shall be individually normalized to a nominal active speech level of -1,7 dB_{Pa} at the MRP for quiet conditions (e.g. those conditions without interfering background noise or barge-in content). This level is equivalent to the typical speech level at the MRP for speakerphone hands-free terminals [6].

When testing with interfering background noise and/or barge-in content, the active speech level shall be increased to account for the "Lombard effect". Recommendation ITU-T P.1100 [6] provides a simple Lombard effect model based on the long-term A-weighted noise level. This model is reproduced here in equation 5.1.4.2-1:

$$I(N) = \begin{cases} 0.0 & for \quad N < 50 \\ 0.3(N-50) & for \quad 50 \leq N < 77 \\ 8.0 & for \quad N \geq 77 \end{cases} \qquad (5.1.4.2\text{-}1)$$

Where *I* is the dB increase in active speech level from the nominal level and *N* is the long-term A-weighted noise level measured at the HATS/mouth simulator head position. *N* shall be determined for each environmental noise and each given test condition (e.g. barge-in content, interfering background noise, or a combination of the two) made at the HATS/mouth simulator MRP prior to speech level calibration. Values of *I* and *N* shall be reported by test labs whenever used.

The active speech level calculator is provided in Recommendation ITU-T P.56 [5].

NOTE: Nominal speech level and intonation may change when interacting with voice assistant devices as compared to other standard communication scenarios (e.g. speakerphone hands-free terminals). The Lombard effect model described above does not take this into account.

5.1.4.3 Command pause

The command pause is a short gap in speech between the wake word and the command. The length of the command pause may be adjusted to assess the performance of the device under different pause conditions. The length of the pause shall be indicated in the test report.

5.1.4.4 Number of talkers

A minimum of 10 different talkers, representing expected device users, should be used for testing. The distribution of talkers shall be balanced between male and female talkers, spanning a wide range of talker's pitch.

5.1.4.5 Requirements on speech materials recordings

The recording of speech materials shall be conducted in a quiet and mostly anechoic environment, with T_{60} < 0,15 s from 200 Hz to 8 kHz. The recording room ambient noise, microphone self-noise and acquisition hardware self-noise shall be capable of meeting an NC 20 or lower noise rating measured according to ANSI/ASA S12.2-2019 [4]. The distance between the talker's lip plane and the microphone diaphragm shall be equal or greater than 20 cm and equal or less than 50 cm. The recordings shall be equalized to compensate for possible coloration introduced by the recording microphone frequency response.

5.1.4.6 Other considerations on speech materials

Other aspects for consideration and reporting (recommended, but not mandatory):

- The different talkers should have either no or the same dialect, different dialects should not be mixed. The dialect used for one evaluation shall be reported. Any other pertinent linguistic or talker details should be provided.
- In case of conducting the tests with multiple languages, commands and questions should contain the same or at least coherent meaning. Language-dependent peculiarities should not be violated, but in order to avoid biases due to language differences, comparable tasks should be defined.
- When recording utterances of a talker, it is recommended to capture multiple intonations of the same command/questions/task.
- Natural language understanding/processing: in case of task-related tests (i.e. targeting at completing rather a task than word/sentence recognition), it is recommended to let the talkers freely speak instead of providing an exact written sentence. Talkers should be instructed to use different wording and/or synonyms for same task.
- If tests with background noise are carried out, it is recommended to include Lombard effect of each talker. During the recording of utterances, the corresponding noise is presented via closed headphones to the talker. Binaural noises for this purpose are provided in ETSI TS 103 224 [1].
- The usage of intentionally disturbed utterances should be considered: Unnatural/longer speech pauses, parts not comprehensible or not intended for assistant device, filler words, stammering, etc.

5.2 Voice assistant device setup

5.2.1 Introduction

The setup of the voice assistant device with respect to the talker, in combination with the acoustic characteristics of the environment and noise level, determines the Direct to Reverb Ratio (DRR) and Signal to Noise Ratio (SNR) experienced by the voice assistant device. The DRR and SNR are important parameters to define the device performance and are dependent on the device distance to the desired talker, the direction of the talker with respect to the device, whether there is a direct line of sight between the DUT and the desired talker, etc.

NOTE: For vehicle-mounted assistants, not all setups of this clause are applicable.

5.2.2 Device positioning in the test environment

Home-like test environment

Placement of the DUT is an important factor to be considered in testing. In a lab environment, it is often common to setup DUTs in the centre of the room for testing. However, most voice assistant devices are used at home in a location close to a wall outlet. Since DUTs feature multi-microphone array technologies and because the placement of the DUT impacts the direction of sound incidence to the device, proximity to walls is a factor to be considered in testing. Therefore, the DUT positions in Table 5.2.2-1 shall be tested in the home-like environment. These DUT positions are marked in Figure 4.2.2.2-1.

For smart speakers, each DUT position shall be tested with at least two of the desired talker positions also marked in Figure 4.2.2.2-1. All three Talker positions should be tested for each DUT position. For soundbars, DUT Position shall be tested with all three Talker positions.

Table 5.2.2-1: List of test locations for the DUT

DUT Position	Description	Height (cm)	Smart Speaker	Soundbars
1	DUT against a wall over an entertainment center	75 ± 5	X	X
2	DUT in the corner of a room over a kitchen cabinet partially blocked by a cabinet	90 ± 5	X	N/A
3	DUT in the middle of a room over a table	75 ± 5	X	N/A

The voice assistant device shall be positioned at different distances from the desired talker or HATS/mouth simulator to assess the device robustness to different DRR and SNR conditions.

In the home-like test environment, the device under test shall be positioned and tested at the distances determined by the relation between the device and talker positions marked in Figure 4.2.2.2-1. These distances are measured from the table surface at the DUT location to the HATS/mouth simulator MRP and are listed in Table 5.2.2-2.

Table 5.2.2-2: DUT distance to HATS/mouth simulator

	DUT Position 1	DUT Position 2	DUT Position 3
Talker Position 1	520 ± 10 cm	140 ± 10 cm	180 ± 10 cm
Talker Position 2	180 ± 10 cm	450 ± 10 cm	390 ± 10 cm
Talker Position 3	250 ± 10 cm	290 ± 10 cm	130 ± 10 cm

Lab-based test environment

In the lab-based test environment, the distance between the device and target talker is captured by the reverberation simulation setup. Therefore, in the lab-based test environment, the device shall be placed at the centre of the room. When possible, the HATS/mouth simulator distance and orientation shall be realized according to the reverberation simulation RIR as described in ETSI TS 103 557 [2]. If space restrictions within the lab-based test environment do not permit such placement, the HATS/mouth simulator shall be placed at a nominal distance of 1,0 m from the DUT at the azimuth angle and orientation corresponding to the reverberation simulation condition.

Car cabin test environment

When using the car cabin test environment, the effects of distance are captured by testing the performance of the voice assistant device with different seating positions (see clause 5.1.3). The seating positions used for testing shall be indicated in the test report.

Generic test environment

The DUT is tested only in the group audio terminal position as shown in Figure 4.4.3-1.

5.2.3 Device orientation

The voice assistant device shall be tested with different device orientations, unless the device is expected to be used solely at a given orientation (e.g. a soundbar intended to be wall mounted). For devices that are not intended to be used solely at a given orientation, three orientations shall be tested as per Table 5.2.3-1.

Table 5.2.3-1: DUT orientation with respect to the talker

Orientation 1	Nominal (indicated by testing laboratory)
Orientation 2	Nominal + 120º rotation clockwise
Orientation 3	Nominal + 240º rotation clockwise

The nominal orientation used for the voice assistant device setup shall be reported by the testing laboratory.

Home-like test environment

For the home-like test environment, the different DUT orientations from Table 5.2.3-1 shall be tested for DUT Position 1 and Talker Position 3 shown in Figure 4.2.2.2-1 under Acoustic Condition 2 from Table 4.2.2.3-1.

Lab-based test environment

For the lab-based test environment, the different DUT orientations from Table 5.2.3-1 shall be tested with a reverberation simulation which complies with Acoustic Condition 2 from Table 4.2.2.3-1. Furthermore, the RIR shall simulate a Talker orientation of 0º and elevation of 0 ± 5º.

Generic test environment

For all generic test environments, DUT orientations 1 and 2 from Table 5.2.3-1 shall be tested. Orientation 3 should be tested.

5.2.4 Line-of-sight issues

The voice assistant device shall be tested both with and without direct Line-Of-Sight (LOS) to the desired talker.

Line-of-sight issues are potentially problematic when beamforming algorithms attempt to identify a dominant direction of arrival for the speech. Note that line-of-sight is not to be confused with the direction of the talker with respect to the device. In a line-of-sight situation, the talker may be facing in the direction of the device, but no direct acoustic path between talker and DUT exists due to the presence of a physical obstruction. Two line-of-sight conditions shall be tested, described as in Table 5.2.4-1.

Table 5.2.4-1: Line of sight conditions

Condition	Description
LOS Condition 1	Direct line-of-sight between device and talker
LOS Condition 2	No direct line-of-sight between device and talker

Home-like test environment

For the home-like test environment, the different line-of-sight conditions are covered by a combination of furniture placement, talker and device positioning and are defined by the DUT and Talker positioning outlined in Figure 4.2.2.2-1. For instance, the acoustic path between Talker Position 3 (couch) and DUT Position 1 (entertainment center) has a direct line-of-sight (LOS Condition 1). However, the acoustic path between Talker Position 2 (corner) and DUT Position 2 (counter) has no direct line-of-sight (LOS Condition 2).

Lab-based test environment

Simulation of line-of-sight issues in a lab-based environment with reverberation simulation is open for further study.

Car-based test environment

When using the car cabin test environment, the effects of line-of-sight are captured by testing the performance of the voice assistant device with different seating positions. The seating positions used for testing shall be indicated in the test report.

Generic test environment

Simulation of line-of-sight issues for the generic test environments with reverberation simulation is not applied for the generic test environment.

5.3 Interaction tests

5.3.1 Barge-in

5.3.1.1 Introduction

Voice controlled devices can play back self-generated sounds (e.g. music) by integrated loudspeakers. The task of recording and playing back at the same time is similar to a double talk situation in telecommunication: the device has to carry out an echo cancellation and/or suppression of the known output signal to perform the speech recognition task accurately.

At a minimum, echo scenarios are expected to cover multiple device playback levels and a variety of echo content (e.g. broadband noise, music, etc.).

Barge-in is a common interaction test where a radio/music or podcast is playing on the Voice Assistant speakers, and the user tries to wake up the assistant to activate the voice command feature. Figure 5.3.1.1-1 illustrates the interaction between DUT, echo signal e(k) and barge-in speech signal s(k).

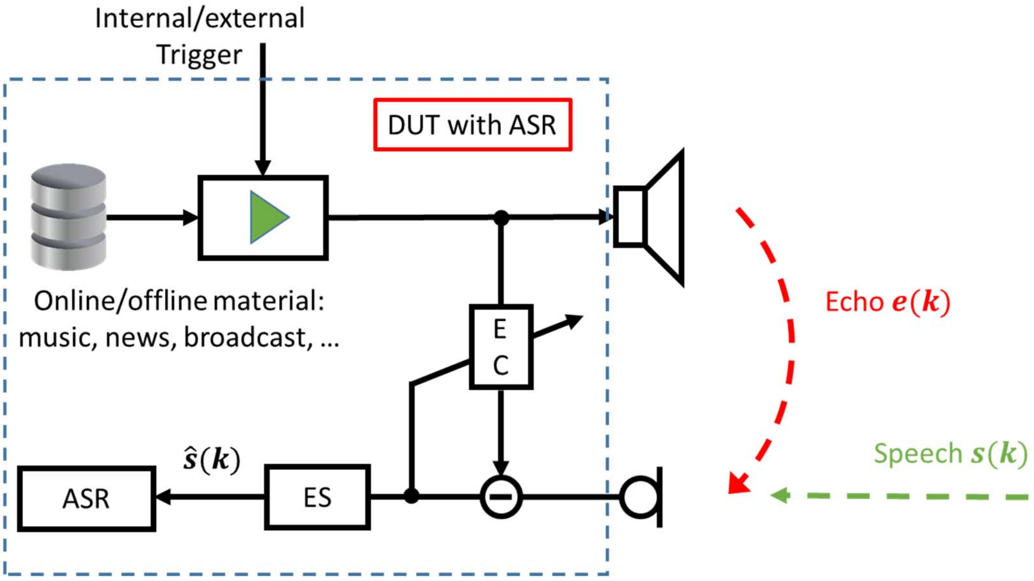

Internal/external Trigger
DUT with ASR
Online/offline material: music, news, broadcast, ...
E C
Echo $e(k)$
Speech $s(k)$
$\hat{s}(k)$
ASR
ES

Figure 5.3.1.1-1: Principle of barge-in/echo testing

5.3.1.2 Barge-in test conditions

Home-like test environment

DUT and HATS/mouth simulator position: For the home-like test environment, the DUT shall be positioned at DUT Position 1 and Talker Position 3 and should be positioned at the other location combinations.

- Speech level: The speech level for the desired talker shall be set according to clause 5.1.4.2.

- Echo signal: The echo signal to be used depends on the device rendering capability as indicated in Table 5.3.1.2-1. Devices shall be tested with the maximum channel count pink noise echo signal supported by the device as well as stereo music. The device should be tested with all supported echo signals. For example, a DUT that supports stereo rendering shall be tested with stereo decorrelated and correlated pink noise and music barge-in content.

- Playback level: The volume control of the device is adjusted at DUT Position 1 and Talker Position 3 to produce the playback levels indicated in Table 5.3.1.2-1. DUT playback levels shall be measured at the HATS/mouth simulator MRP (see following note). The volume control position is then used for any other test locations.

- Acoustic condition: Barge-in functionality shall be tested in Acoustic Condition 2 from Table 4.2.2.3-1. Other acoustic conditions should be tested.

- Background noise: In addition to barge-in without interfering background noise, barge-in should also be tested in the presence of the background noise conditions from Table 4.2.2.4-1. If used, the background noise condition shall be reported.

Table 5.3.1.2-1: Barge-in test content for different DUT rendering capabilities in home-like environment

DUT rendering capability	Barge-in test content	High playback level (dB$_{SPL}$(C))	Low playback level (dB$_{SPL}$(C))
Mono or Stereo (e.g. smart speaker)	Stereo music	67	57
	Stereo decorrelated pink noise	67	57
	Stereo correlated pink noise	67	57
5.1 (e.g. soundbar, distributed systems)	Cinematic excerpt	73	60
	Multi-channel decorrelated pink noise	67	57
	Multi-channel correlated pink noise	67	57
5.1.2 (e.g. soundbar, distributed systems)	Cinematic excerpt	73	60
	Multi-channel decorrelated pink noise	67	57
	Multi-channel correlated pink noise	67	57
7.1.4 (e.g. soundbar, distributed systems)	Cinematic excerpt	73	60
	Multi-channel decorrelated pink noise	67	57
	Multi-channel correlated pink noise	67	57
NOTE: A DUT supporting only mono rendering is tested with stereo music and pink noise content. Downmixing from stereo is a necessary feature for such devices and testing with stereo content assesses barge-in performance in this common usage scenario.			

Pink noise and stereo music barge-in content (generated according to annex F) can be retrieved from https://docbox.etsi.org/STQ/Open/TS%20103%20504/Barge-in%20test%20signals.

When barge-in testing is conducted with a cinematic excerpt, test labs shall report the content used.

NOTE: The decision to measure and calibrate barge-in content at the MRP is motivated by the need for a well-defined reference point in the test setup, which is independent of DUT form factor (e.g. single driver smart speaker, multi-channel sound bar, distributed loudspeaker system, etc.) and the speech reproduction equipment (e.g. HATS or mouth simulator). Further investigation is required to provide calibration levels at other potential reference points such as the drum reference point of a HATS.

Lab-based test environment

- DUT and HATS/mouth simulator position: For the lab-based test environment, the DUT shall be positioned at the centre of the space with the HATS/mouth simulator 1,0 m away. Barge-in shall be tested with a reverberation simulation corresponding to a HATS/mouth simulator orientation of 0º, an elevation of 0 ± 5º, and a nominal DUT orientation. Other Talker and DUT position and orientation combinations should be tested.

- Speech level: The speech level for the desired talker shall be set according to clause 5.1.4.2 with one amendment. To achieve a Lombard gain, *N*, which is comparable to the Lombard gain used in the home-like test environment, *N* shall be measured with the interfering background noise calibrated to the level presented in Table 5.3.1.2-1, as opposed to Table 5.3.1.2-2. After the measurement of *N* and determination of the appropriate speech level with Lombard gain, the barge-in content level shall be recalibrated accordingly.

- Echo signal: The echo signal to be used depends on the device rendering capability as indicated in Table 5.3.1.2-2. Devices shall be tested with the maximum channel count pink noise echo signal supported by the device as well as stereo music. The device should be tested with all supported echo signals. For example, a DUT that supports stereo rendering shall be tested with stereo decorrelated and correlated pink noise and music barge-in content.

- Playback level: The volume control of the device is adjusted at the required position to produce playback levels indicated in Table 5.3.1.2-2. DUT playback levels shall be measured at the HATS/mouth simulator MRP (see note above). The volume control position is then used for any other DUT and HATS/mouth simulator positions.

- Acoustic condition: Barge-in functionality shall be tested with simulated Acoustic Condition 2 (medium reverberation) from Table 4.2.2.3-1. Other acoustic condition simulations should be tested.
- Background noise: In addition to barge-in without interfering background noise, barge-in should be tested under simulated background noise conditions. If used, the background noise condition shall be reported.

Table 5.3.1.2-2: Barge-in test content for different DUT rendering capabilities in lab-based environment

DUT rendering capability	Barge-in test content	High playback level (dB$_{SPL}$(C))	Low playback level (dB$_{SPL}$(C))
Mono or Stereo (e.g. smart speaker)	Stereo music	75	65
	Stereo decorrelated pink noise	75	65
	Stereo correlated pink noise	75	65
5.1 (e.g. soundbar, distributed systems)	Cinematic excerpt	81	68
	Multi-channel decorrelated pink noise	75	65
	Multi-channel correlated pink noise	75	65
5.1.2 (e.g. soundbar, distributed systems)	Cinematic excerpt	81	68
	Multi-channel decorrelated pink noise	75	65
	Multi-channel correlated pink noise	75	65
7.1.4 (e.g. soundbar, distributed systems)	Cinematic excerpt	81	68
	Multi-channel decorrelated pink noise	75	65
	Multi-channel correlated pink noise	75	65
NOTE 1: A DUT supporting only mono rendering is tested with stereo music and pink noise content. Downmixing from stereo is a necessary feature for such devices and testing with stereo content assesses barge-in performance in this common usage scenario. NOTE 2: Barge-in levels are calibrated at the target talker MRP for testing in both the simulated lab-based and real home-like environments. When located at the required calibration position, the target talker MRP is 1 m from the DUT in the lab-based environment and 2,5 m from the DUT in the home-like environment. The inverse square law is used to compensate for this difference in distance [i.13], resulting in the 8 dB difference in the calibration levels presented in Table 5.3.1.2-1 and this table.			

Pink noise and stereo music barge-in content (generated according to annex F) can be retrieved from https://docbox.etsi.org/STQ/Open/TS%20103%20504/Barge-in%20test%20signals.

When barge-in testing is conducted with a cinematic excerpt, test labs shall report the content used.

Car-based test environment

- HATS/mouth simulator location: For the car-cabin test environment, the HATS/mouth simulator shall be positioned at the driver's seat location.
- Speech level: The speech level for the desired talker shall be set according to clause 5.1.4.2.
- Echo signal: The echo signal to be used depends on the device playback capability as indicated in Table 5.3.1.2-1.

Generic test environment

- DUT and HATS/mouth simulator position: The HATS/mouth simulator shall be setup according to clause 4.4, i.e. in group audio terminal position as shown in Figure 4.4.3-1.
- Speech level: The speech level for the desired talker shall be set according to clause 5.1.4.2 with one amendment. To achieve a Lombard gain, *N*, which is comparable to the Lombard gain used in the home-like test environment, *N* shall be measured with the interfering background noise calibrated to the level presented in Table 5.3.1.2-1, as opposed to Table 5.3.1.2-3. After the measurement of *N* and determination of the appropriate speech level with Lombard gain, the barge-in content level shall be recalibrated accordingly.
- Echo signal: The echo signal to be used depends on the device rendering capability as indicated in Table 5.3.1.2-3. Devices shall be tested with the maximum channel count pink noise echo signal supported by the device as well as stereo music. The device should be tested with all supported echo signals. For example, a DUT that supports stereo rendering shall be tested with stereo decorrelated and correlated pink noise and music barge-in content.

- Playback level: The volume control of the device is adjusted at the required position to produce playback levels indicated in Table 5.3.1.2-3. DUT playback levels shall be measured at the HATS/mouth simulator MRP (see note above). The volume control position is then used for any other DUT and HATS/mouth simulator positions.

- Conditions: Barge-in functionality shall be tested with generic condition G00 and G09 from Table 4.4.5-1. Other generic condition simulations should be tested.

Table 5.3.1.2-3: Barge-in test content for different DUT rendering capabilities in generic environments

DUT rendering capability	Barge-in test content	High playback level (dB$_{SPL}$(C))	Low playback level (dB$_{SPL}$(C))
Mono or Stereo (e.g. smart speaker)	Stereo music	76	66
	Stereo decorrelated pink noise	76	66
	Stereo correlated pink noise	76	66
5.1 (e.g. soundbar, distributed systems)	Cinematic excerpt	82	69
	Multi-channel decorrelated pink noise	76	66
	Multi-channel correlated pink noise	76	66
5.1.2 (e.g. soundbar, distributed systems)	Cinematic excerpt	82	69
	Multi-channel decorrelated pink noise	76	66
	Multi-channel correlated pink noise	76	66
7.1.4 (e.g. soundbar, distributed systems)	Cinematic excerpt	82	69
	Multi-channel decorrelated pink noise	76	66
	Multi-channel correlated pink noise	76	66
NOTE 1: A DUT supporting only mono rendering is tested with stereo music and pink noise content. Downmixing from stereo is a necessary feature for such devices and testing with stereo content assesses barge-in performance in this common usage scenario. NOTE 2: Barge-in levels are calibrated at the target talker MRP for testing in both the simulated generic and real home-like environments. When located at the required calibration position, the target talker MRP is 85,5 cm from the DUT in the simulated generic environment and 2,5 m from the DUT in the home-like environment. The inverse square law is used to compensate for this difference in distance [i.13], resulting in the 9 dB difference in the calibration levels presented in Table 5.3.1.2-1 and this table.			

Pink noise and stereo music barge-in content (generated according to annex F) can be retrieved from https://docbox.etsi.org/STQ/Open/TS%20103%20504/Barge-in%20test%20signals.

When barge-in testing is conducted with a cinematic excerpt, test labs shall report the content used.

5.3.2 Conversation mode

The device is asked a question and gives an answer. The DUT should answer an ambiguous second question using information from the first question.

EXAMPLES:

- User: "How old is Tom Hanks?"
- Answer: "He is 61 years old."
- User: "Which movies did he play in?"
- Answer of DUT1: "His last movie was 'The circle'" (expected/valid answer).
- Answer of DUT2: "Which person do you mean?" (wrong answer).

5.3.3 Dialogue mode

If a device is tested for the completion of tasks, it is desired to have it completed successfully, reliably, quickly and without unnecessary additional interactions. The accuracy of each task can be expressed per talker (and possibly per varying intonations/synonyms) and is scored according to the scheme in Table 5.3.3-1.

Table 5.3.3-1: Task accuracy scoring

Score	Definition
+2	The task was understood correctly and completed, without extra commands
+1	The task was understood correctly in general, but required one or more extra command(s) for completion
0	The task was not understood and could not completed with or without extra commands
-1	The task was misunderstood, but could be cancelled
-2	The task was misunderstood, could not be cancelled, and an incorrect task was executed
NOTE: Due to potential ambiguities in the interpretation of an average task accuracy score over trials, labs should report aggregate task accuracy scores using a histogram or equivalent method.	

5.4 Wake word detection tests

5.4.1 False Rejection Rate (FRR)

False Rejection Rate (FRR) should be evaluated with at least 60 wake word trials for statistically significant results on a per condition basis. See annex C for more information on sample size recommendations for statistically significance in Bernoulli trials, for FRR and other performance metrics described in clause 6.

The FRR is calculated according to equation 6.3-1 in clause 6.3.

5.4.2 False Acceptance Rate (FAR)

While devices should detect and respond to occurrences of their wake word, they should not mistake other speech for their wake word, which is a false acceptance (sometimes also referred as 'false alarm' or 'false detect'). To measure the False Acceptance Rate (FAR), the DUT shall be exposed to 24 hours of generic audio material containing speech in the language being tested (e.g. news or talk radio). The generic audio material should not contain the wake words for the device. In the event they do, the number of occurrences shall be indicated in the test report.

The inclusion of generic audio material for FAR testing is open for further study. Test labs shall report the source material used for FAR testing.

The additional sound source is realized with a HATS/mouth simulator placed at 1 m, directly facing the DUT (orientation 0^o). The HATS/mouth simulator MRP shall be at the same height as the base of the DUT. Furthermore, the HATS/mouth simulator shall be equalized to a flat frequency response at the MRP according to the equalization procedures defined in Recommendation ITU-T P.58 [3]. The audio material shall be reproduced at -1,7 dB_{Pa} at the MRP. Other FAR material reproduction levels may be tested and reported in the test report.

In a home-like test environment, the FAR shall be tested for Acoustic Condition 3 (low reverberation) from Table 4.2.2.3-1 and should be tested for other acoustic conditions. For lab-based testing, FAR shall be tested without reverberation simulation.

The FAR is calculated according to equation 6.4-1 in clause 6.4.

6 Performance metrics

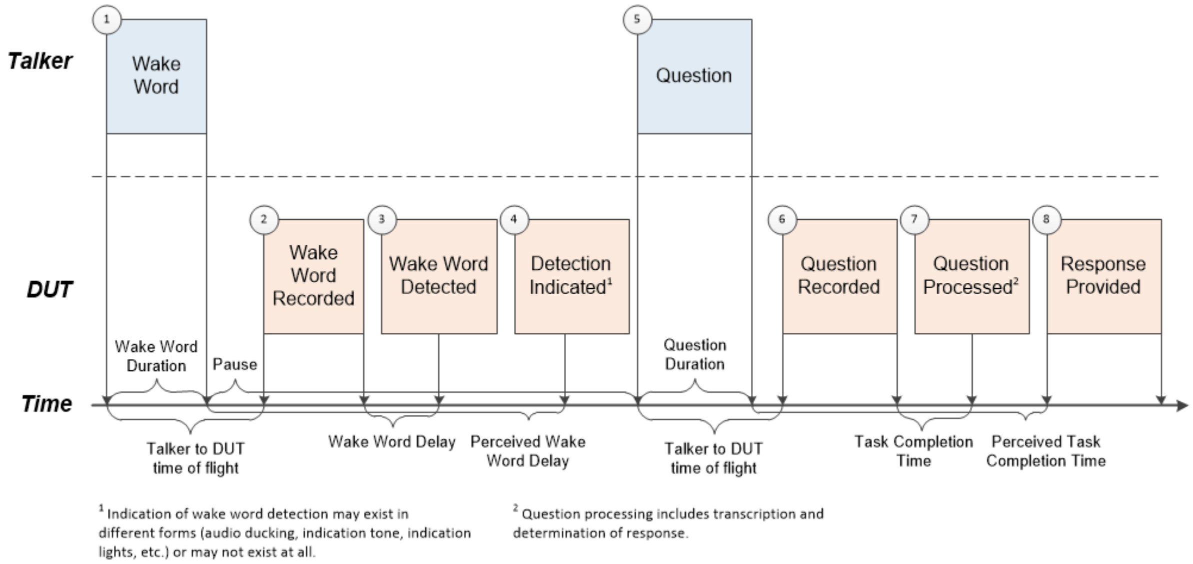
6.1 Introduction

The test procedures described in clause 5 can be carried out over a large number of variables. The metrics introduced in this clause are described per utterance/sentence, per talker and/or per condition. In this context, condition means a specific combination of the variables:

- Distance/Playback Level.
- Background noise scenario.
- Reverb scenario.

- Echo/Non-Echo.
- Device orientation.
- Line-of-sight configuration.
- Specific task.
- (other variables).

Figure 6.1-1 provides the timeline of a successful voice assistant task. The steps in Figure 6.1-1 are referenced in the following clauses to help illustrate the DUT performance metrics.



[1] Indication of wake word detection may exist in different forms (audio ducking, indication tone, indication lights, etc.) or may not exist at all.

[2] Question processing includes transcription and determination of response.

Figure 6.1-1: Example voice assistant task timeline

6.2 Word Error Rate (WER)

To evaluate this metric, the written transcript shall be available for each source utterance. It can directly be used for the calculation of the word error rate $WER_{i,j,k}$ (corresponding to the i-th utterance of the j-th talker in the k-th condition) according to equation 6.2-1.

$$WER_{i,j,k} = \frac{S_{i,j,k}+D_{i,j,k}+I_{i,j,k}}{N_{i,j,k}} \tag{6.2-1}$$

Here $S_{i,j,k}$ equals the number of substitutions, $D_{i,j,k}$ the number of deletions, $I_{i,j,k}$ the number of insertions and $N_{i,j,k}$ the number of words in the reference for i-th utterance, j-th talker, and k-th condition.. In order to obtain the average WER for a certain condition, a weighted sum according to equations 6.2-2 and 6.2-3 is used. Note that the number of words N may differ per utterance/sentence, per talker and per condition - especially when taking natural language (intonations, synonyms, unscripted speech) into account.

$$\overline{WER}_k = \sum_i (WER_{i,j,k} \cdot w_{i,j,k}) \tag{6.2-2}$$

$$w_{i,j,k} = {}^{N_{i,j,k}}\!\Big/_{\sum_i N_{i,j,k}} \tag{6.2-3}$$

Additionally, the Correct Utterance Ratio (CUR) can be determined according equation 6.2-4. The CUR specifies how many complete utterances were correctly detected in the k-th condition. This measure may be important for the analysis of several tasks, i.e. where all words within an utterance are important.

$$CUR_k = \frac{\sum_k WER_{i,j,k}==0}{(\sum_k WER_{i,j,k}==0)+(\sum_k WER_{i,j,k} > 0)} \tag{6.2-4}$$

NOTE 1: This metric can only be measured if the DUT allows access to the transcribed question from step 7 of Figure 6.1-1.

NOTE 2: See annex C for more information on sample size recommendations for statistically significance in Bernoulli trials.

6.3 False Rejection Rate (FRR)

The FRR of the j-th talker and k-th condition is calculated according to equation 6.3-1 as the ratio between the number of wake words missed, $N_{j,k}^{-}$, and the total number of wake word trials, $N_{j,k}$.

$$FRR_{j,k} = \frac{N_{j,k}^{-}}{N_{j,k}} \qquad (6.3\text{-}1)$$

NOTE: See annex C for more information on sample size recommendations for statistically significance in Bernoulli trials.

6.4 False Acceptance Rate (FAR)

The FAR is calculated according to equation 6.4-1 as the number of false alarms, $N^{\pm}$, divided by the observation time. The FAR is reported in false alarms per unit time.

$$FAR = \frac{N^{\pm}}{\Delta T} \qquad (6.4\text{-}1)$$

6.5 Task Completion Rate (TCR)

For the evaluation of dialogue systems according to clause 5.3.3, each task is evaluated for the i-th utterance (e.g. multiple modifications for the same task), the j-th talker and k-th condition according to Table 5.3.3-1, providing a single result $V_{i,j,k}$ for each trial. The Task Completion Rate (TCR) per condition is the number of completed tasks over all utterances and talkers divided by the total number of trials for this task, as specified in equation 6.5-1.

$$TCR_k = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} (V_{i,j,k} > 0) \qquad (6.5\text{-}1)$$

In addition, an average Task Completion Score (TCS) can be calculated according to equation 6.5-2.

$$\overline{TCS_k} = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} V_{i,j,k} \qquad (6.5\text{-}2)$$

NOTE: See annex C for more information on sample size recommendations for statistically significance in Bernoulli trials.

6.6 Task Completion Time (TCT)

The Task Completion Time (TCT) is presented in Figure 6.1-1 as the time elapsed between when the DUT completes recording the question (step 7), $T_{i,j,k}^{rec}$, and when the DUT finishes processing the question and generating the response (step 8), $T_{i,j,k}^{proc}$. The TCT per condition k is calculated for all utterances i and talkers j according to equation 6.6-1. The TCT is only evaluated for trials in which the task is successfully completed.

$$TCT_k = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} (T_{i,j,k}^{proc} - T_{i,j,k}^{rec}) \qquad (6.6\text{-}1)$$

NOTE: This metric can only be measured when the DUT provides access to internal timestamps for the recorded question and completion of response generation.

6.7 Perceived Task Completion Time (PTCT)

The Perceived Task Completion Time (PTCT) is presented in Figure 6.1-1 as the time elapsed between when the talker completes the question utterance (step 5), $T_{i,j,k}^{utt}$, and when the DUT begins its response (step 9), $T_{i,j,k}^{resp}$. The PTCT per condition k is calculated for all utterances i and talkers j according to equation 6.7-1. The PTCT is only evaluated for trials in which the task is successfully completed.

$$PTCT_k = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} (T_{i,j,k}^{resp} - T_{i,j,k}^{utt}) \qquad (6.7\text{-}1)$$

NOTE: Unlike TCT, this metric does not require access to internal DUT information.

6.8 Wake Word Delay (WWD)

The Wake Word Delay (WWD) is presented in Figure 6.1-1 as the time elapsed between when the DUT completes recording the wake word (step 2), $TW_{i,j,k}^{rec}$, and when the DUT flags that the wake word has been detected (step 3), $TW_{i,j,k}^{det}$. WWD per condition k is calculated for utterance i and talker j according to equation 6.8-1. Only utterances where the wake word is successfully detected are used.

$$WWD_k = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(TW_{i,j,k}^{det} - TW_{i,j,k}^{rec}\right) \qquad (6.8\text{-}1)$$

NOTE: This metric can only be measured when the DUT provides access to internal timestamps for the recorded wake word and detection flag.

6.9 Perceived Wake Word Delay (PWWD)

The Perceived Wake Word Delay (PWWD) is presented in Figure 6.1-1 as the time elapsed between when the talker completes uttering the wake word (step 1), $TW_{i,j,k}^{utt}$, and when the DUT indicates that the wake word has been detected (step 4), $TW_{i,j,k}^{ind}$. PWWD per condition k is calculated for utterance i and talker j according to equation 6.9-1. Only utterances where the wake word is successfully detected are used.

$$PWWD_k = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} (TW_{i,j,k}^{ind} - TW_{i,j,k}^{utt}) \qquad (6.9\text{-}1)$$

NOTE: Unlike WWD, this metric does not require access to internal DUT information. However, this metric requires that the DUT indicates wake word detection (e.g. through audio ducking, indication tone, etc.).

7 Performance requirements

The definition of performance requirements for the previously described metrics is open for further study.

Annex A (normative): Room acoustics and electro acoustic equipment positioning

The positioning of transducers in the acoustic environment can strongly influence their effective performances and suitable installation criteria should be followed in order to maximize the signal-to-noise and signal-to-reverberation ratios.

In particular the main parameters to be taken into account when installing teleconference/videoconference systems are:

- Room acoustics (e.g. reverberation).
- Background noise.
- Sound insulation (privacy), mainly for individual use.

Additional parameters to be taken into account are at least:

- A room suitable for a normal face-to-face conference shall be selected.
- Maximum talker to microphone distance shall be determined taking into account both the noise and reverberation dependencies.
- The microphones and loudspeakers shall be positioned in accordance with both these distances.
- The microphone type should be chosen according to the room environment.

More detailed information is available in ETSI ETS 300 807 [i.4] and Recommendation ITU-T Supplement P 16 [i.5].

Annex B (informative): Example test plan

B.1 Introduction

This example test plan covers the required test conditions defined in the present document for both a smart speaker with single-channel playback and a soundbar with surround playback capabilities. The test plan can be carried out in either a home-like or lab-based test environment.

B.2 False Acceptance Rate test

The FAR of the smart speaker and soundbar form factors are tested according to clause 5.4.2. In the home-like test environment, the room acoustics are set according to Acoustic Condition 2 from Table 4.2.2.3-1. In the lab-based test environment, no reverberation simulation is used.

B.3 Task completion tests

B.3.1 Test setup

In the home-like test environment, realization of the acoustic condition, talker position, DUT position, and interfering noise condition is defined in clause 4.2.2. Barge-in conditions are defined in clause 5.3.1.2.

For lab-based testing, the acoustic condition, DUT position, and talker position are realized through reverberation simulation according to clause 4.2.3.3. The interfering noise condition is simulated according to clause 4.2.3.4 and barge-in conditions are covered in clause 5.3.1.2.

The speech reproduction system (e.g. HATS or mouth simulator) is set up for each testing environment and trial condition according to clause 5.1.3.

B.3.2 Stimuli

Speech stimuli are generated, and level normalized according to clause 5.1.4. 10 total talkers are used (5 male and 5 female) with 2 utterances per talker for a total of 20 utterances per condition. The language, dialect, and recording information for the utterances are provided. A pause of 0,5 seconds is used between each wake word and question.

NOTE: 20 utterances per condition provides poor statistical significance on a per condition basis (see annex C). Performance should therefore be evaluated on the aggregate of all conditions. For statistically significant results on a per condition basis, it is recommended that at least 60 utterances are used per condition.

B.3.3 Smart speaker test conditions

Table B.3.3-1 presents the 18 required test conditions for a smart speaker with single-channel playback capabilities. With 20 utterances per condition, there are a total of 360 trials. If each utterance is approximately 30 seconds, the entire test will require roughly 3 hours.

Table B.3.3-1: Required test conditions for a smart speaker

	Acoustic Condition	DUT Position	Talker Position	Noise Condition	Barge-In Content	DUT Orientation
1	1 (low reverb)	1 (entertainment center)	3 (Couch)	Ambient	None	Nominal
2	3 (high reverb)	1 (entertainment center)	3 (Couch)	Ambient	None	Nominal
3	2 (medium reverb)	1 (entertainment center)	3 (Couch)	Ambient	None	Nominal
4	2 (medium reverb)	1 (entertainment center)	3 (Couch)	Ambient	None	120º

	Acoustic Condition	DUT Position	Talker Position	Noise Condition	Barge-In Content	DUT Orientation
5	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	None	240º
6	2 (medium reverb)	1 (entertainment center)	3 (couch)	Sink	None	Nominal
7	2 (medium reverb)	1 (entertainment center)	3 (couch)	Refrigerator	None	Nominal
8	2 (medium reverb)	1 (entertainment center)	3 (couch)	Talker	None	Nominal
9	2 (medium reverb)	1 (entertainment center)	3 (couch)	All	None	Nominal
10	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	Stereo correlated pink noise low	Nominal
11	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	Stereo correlated pink noise high	Nominal
12	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	Stereo decorrelated pink noise low	Nominal
13	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	Stereo decorrelated pink noise high	Nominal
14	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	Stereo music low	Nominal
15	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	Stereo music high	Nominal
16	2 (medium reverb)	1 (entertainment center)	1 (sink)	Ambient	None	Nominal
17	2 (medium reverb)	2 (counter)	2 (corner)	Ambient	None	Nominal
18	2 (medium reverb)	2 (counter)	3 (couch)	Ambient	None	Nominal
19	2 (medium reverb)	3 (table)	1 (sink)	Ambient	None	Nominal
20	2 (medium reverb)	3 (table)	2 (corner)	Ambient	None	Nominal

B.3.4 Soundbar test conditions

Table B.3.4-1 presents the required test conditions for a 5.1 enabled soundbar. There are a total of 13 required test conditions for a surround soundbar. With 20 utterances per condition, there are a total of 260 trials. If each utterance is approximately 30 seconds, the entire test will require roughly 2 hours and 15 minutes.

Table B.3.4-1: Required test conditions for a 5.1 enabled soundbar

	Acoustic Condition	DUT Position	Talker Position	Interfering Noise	Barge-In Content
1	1 (low reverb)	1 (entertainment center)	3 (couch)	Ambient	None
2	3 (high reverb)	1 (entertainment center)	3 (couch)	Ambient	None
3	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	None
4	2 (medium reverb)	1 (entertainment center)	3 (couch)	Kitchen Sink	None
5	2 (medium reverb)	1 (entertainment center)	3 (couch)	Refrigerator	None
6	2 (medium reverb)	1 (entertainment center)	3 (couch)	Talker	None
7	2 (medium reverb)	1 (entertainment center)	3 (couch)	All	None
8	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	5.1 correlated pink noise low
9	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	5.1 correlated pink noise high
10	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	5.1 decorrelated pink noise low
11	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	5.1 decorrelated pink noise high
12	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	Stereo music low
13	2 (medium reverb)	1 (entertainment center)	3 (couch)	Ambient	Stereo music high
14	2 (medium reverb)	1 (entertainment center)	1 (sink)	Ambient	None
15	2 (medium reverb)	1 (entertainment center)	2 (corner)	Ambient	None

B.4 Results reporting

Results for an FAR test are reported as demonstrated in Table B.4-1.

Table B.4-1: Example False Acceptance Rate table of results

Device	False alarms per 24 hours
Smart speaker	
Soundbar	

Results for a task completion test with *N* total conditions are reported as demonstrated in Table B.4-2. All metrics are calculated according to clause 6.

Table B.4-2: Example task completion test table of results

Condition	False Rejection Rate	Word Error Rate	Task Completion Rate	Task Completion Time	Wake Word Delay
1					
2					
…	…	…	…	…	…
N					

Annex C (informative): Statistical considerations when selecting sample size

C.1 Introduction

Many of the performance metrics defined in clause 6, such as False Rejection Rate (FRR, clause 6.3) and Task Completion Rate (TCR, clause 6.5) are derived from trials in which there are only two possible outcomes, success or failure. Experiments with exactly two outcomes are statistically described as Bernoulli trials [i.7]. Assuming that each trial is independent with constant underlying probability of success, the statistics of the outcome of a sequence of Bernoulli trials is described by the binomial distribution [i.7]. The statistical properties of a binomial random variable differ from the more commonly encountered normal (or Gaussian) random variable, so attention to the differences is warranted for the binomially-distributed performance metrics defined in clause 6.

Sophisticated voice assistant systems can be adaptive and may incorporate continuous machine learning techniques, so that the underlying probabilities may not be fixed. However, for purposes of practical measurement considered in the present document, it is assumed that the underlying probabilities are effectively constant throughout the duration of a test, so that binomial statistics should be used to estimate statistical confidence intervals of the measured results.

C.2 Recommendation

Clause 5.1.4.4, Number of talkers, recommends a minimum of 10 different talkers, while clause 5.4.1, False Rejection Rate, recommends using at least 60 trials (utterances) per talker, so the resulting recommended sample size is 600 trials. For this total sample size, the statistical confidence interval or "error bar" on the measured results is expected to be less than 0,04 (4 %) for any measured value of FRR or TCR. However, in some cases, due to limited experimental time, only fewer trials may be possible. It is important to understand the implications of selecting a smaller test sample size on the reliability of the resulting measurements.

While there are well-known closed form expressions for the probability mass function, cumulative distribution function, and descriptive statistics for the binomial distribution [i.7], most statistics texts provide a formula for the confidence interval that makes several assumptions, mainly that the number of trials is large (Wald method). A recent study [i.8] provides a review of the issues with the underlying assumptions and a survey of various alternative methods for computing the confidence interval for binomially-distributed random variables. In [i.8], the assumptions underlying the commonly cited Wald method are listed and shortcomings identified. Several alternative methods are compared with respect to the assumptions required for the Wald method. Three methods are found to be superior with respect to these limitations: the interval estimates of Wilson, Agresti-Coull, and Jeffreys. For tests number of trials $N \geq 40$, all three methods are comparable in technical terms, and given practical consideration, the authors of [i.8] recommend the Agresti-Coull method.

It is important to understand that the confidence interval of a measured estimate of a binomially-distributed performance measure depends not only on the sample size but also on the underlying true value of the measured metric. Figure C.2-1 illustrates this dependence, plotting the lower confidence interval for the range of possible values of TCR, for different values of sample size N (50, 100, 200, 500) and confidence level (95 %, 99 %), using the Agresti-Coull method.

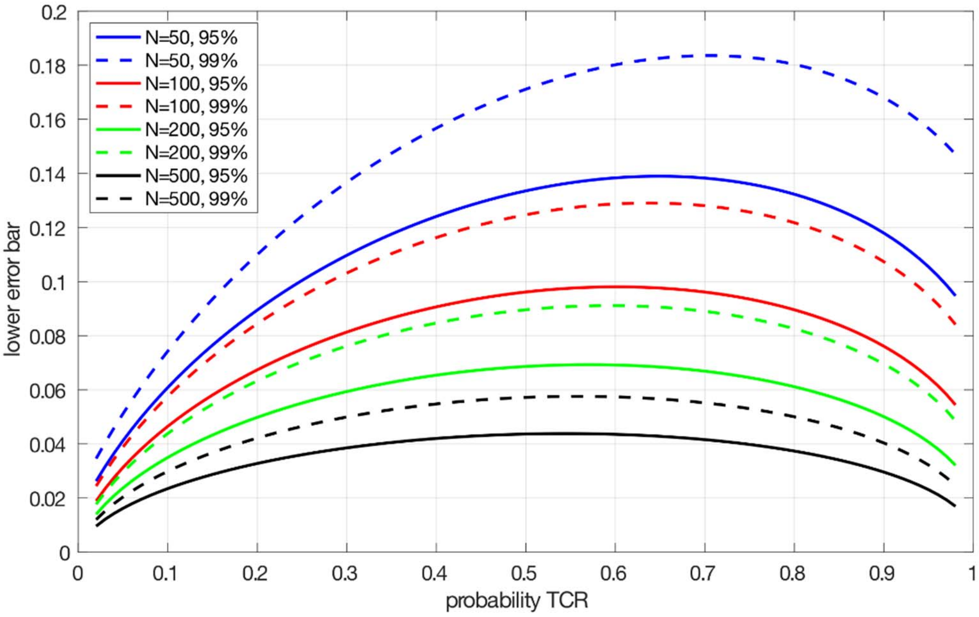


Figure C.2-1: Lower confidence interval versus TCR, for several values of Sample Size N

It is recommended to use the Agresti-Coull method to compute the expected lower (or upper) confidence interval for the desired sample size, and confirm that the resulting statistical reliability is acceptable in the specific application of concern.

Using Figure C.2-1 as an example, if the sample size of 50 is desired (e.g. 5 utterances for each of 10 talkers), then the Agresti-Coull estimate of the lower confidence interval at the 99 % confidence interval is showing by the dashed blue line. This shows that the lower confidence interval reaches a maximum of about 0,18 for TCR of about 0,7. This implies that when the measured TCR is about 0,7, the experimental uncertainty is 0,18 (at the 99 % level), so the true value could be as low as (0,7 - 0,18) 0,52. If this degree of uncertainty is acceptable for the application, then this relatively small sample size could be used.

Annex D (informative): Investigation of HATS and mouth simulator directivity

D.1 Introduction

For the accurate assessment of voice assistant devices, it is important to consider the effects of talker orientation on device performance [i.9]. Although true human subjects provide the desired speech directivity characteristics, it is often impractical to use human subjects for the tests described within the present document. Due to the extensive test time and need for reproducibility, artificial mouth simulators are a viable alternative, provided their directivity characteristics adequately reproduce the average directivity characteristics of human subjects.

The broadband normalised free-field responses of standalone artificial mouth simulators and HATS are defined and standardized in Recommendation ITU-T P.51 [8] and Recommendation ITU-T P.58 [3], respectively. Table 3c of [8] specifies the directivity requirements of a standalone mouth simulator at seven points in the frontal hemisphere. Table 7d of [3] extends the specifications in Recommendation ITU-T P.51 [8] for the mouth simulator in a HATS with one-third octave band directivity requirements at seven points in the frontal hemisphere and five points in the rear hemisphere. Although [8] and [3] are motivated by human subjects' speech directivity, the specifications are band limited to 8 kHz and provide sparse spatial sampling.

Prior research demonstrates inconsistencies between HATS and human subjects' speech directivity in both the near field [i.6] and far field [i.10], [i.11]. Chu and Warnock (2002) measured the speech directivity of 40 male and female subjects and one HATS [i.10]. Recordings were made at 15º increments along the horizontal axis, with elevations ranging from 90º (overhead) to -50º. Human speech was shown to be less directional than the HATS reproduction, with the most pronounced differences occurring behind the lip plane [i.10]. Bozzoli et al. (2005) measured the speech directivity of 10 male subjects and one HATS [i.11]. Their findings illustrated differences of up to 5 dB in rear-plane directivity between human speakers and the HATS. Contrary to the results in [i.10], Bozzoli et al. found human speech to be more directional than the HATS reproduction [i.11].

Although no comparison between humans and artificial mouth simulators is provided, Monson and Hunter (2012) collected a database of human speech directivity in the horizontal plane from 15 subjects [i.12]. This database extends previous directivity measurements to 16 kHz and validates horizontal plane data from [i.10].

In what follows, the far field speech directivity data from human subjects presented in [i.10] and [i.12] are compared to the measured directivity of two HATS, one standalone mouth simulator, and one single-driver loudspeaker.

D.2 Measurement setup and procedure

Figure D.2-1 shows the measurement setup used for far field directivity measurements in the present document. All measurements were conducted in an anechoic chamber. An array of three free-field reference microphones was used for acoustic capture. The DUT was laser aligned such that the DUT driver directly faced the microphone array at the height of the middle microphone with the turntable set to 0º.

DUT playback was calibrated to 55 dB_{SPL}(A) at 1 kHz measured at the 0º microphone. Measurements were taken at 15º increments from 0º to 180º along the horizontal axis. At each position, a two second logarithmic swept sine impulse response from 100 Hz to 18 kHz was captured at the three microphones. The resulting frequency responses were integrated to octave bands and normalized such that the DUT response was flat at 0º azimuth, 0º elevation. For visualization purposes, the results were interpolated along the horizontal axis by a factor of three. The measurement procedure was repeated for two HATS, one standalone mouth simulator, and a 4.25" spherical single-driver loudspeaker. Normalized octave band frequency response data can be found in clause D.7.

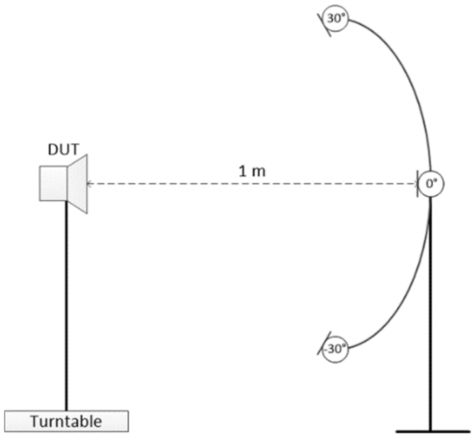


Figure D.2-1: Directivity measurement setup

D.3 Horizontal plane directivity

D.3.1 Comparing DUTs with Chu & Warnock data

Figure D.3.1-1 presents the horizontal plane directivity of the four DUTs compared to the human directivity data reported by Chu & Warnock [i.10]. Human speech is significantly less directional than the single-driver loudspeaker, particularly at frequencies above 1 kHz. Both HATS are slightly more directive than human speech at high frequencies. Finally, the mouth simulator demonstrates a decrease in directivity compared to human speech at frequencies above 1 kHz due to a higher energy lobe behind the lip plane. The mouth simulator lateral directivity around 90⁰ is comparable to that of the two HATS.

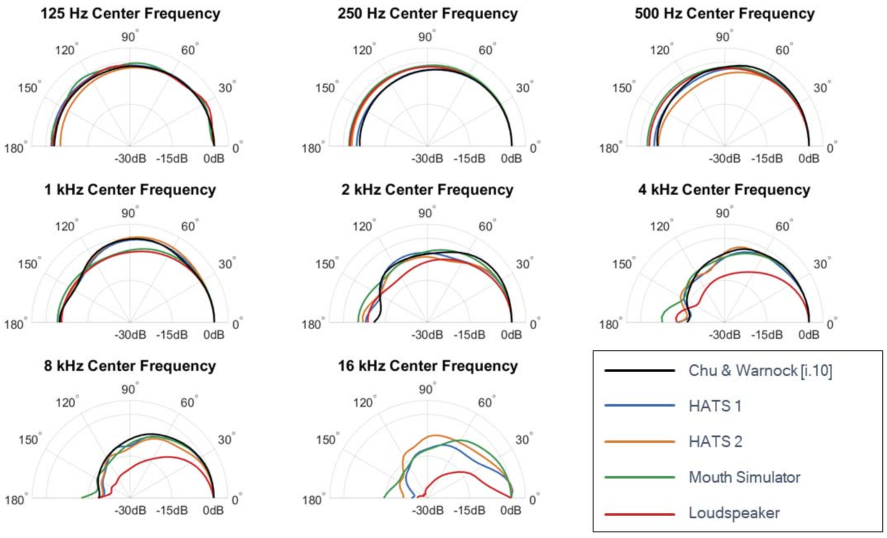


Figure D.3.1-1: Horizontal plane (0⁰ elevation) directivity of DUTs and Chu & Warnock data [i.10]

Figure D.3.1-2 presents the Root Mean Squared Error (RMSE) between each DUT and the Chu & Warnock data [i.10] calculated by azimuth angle of incidence, octave frequency band, and overall. The single driver speaker has a maximum overall RMSE of 3,93 dB, with the most significant errors occurring around the 90⁰ angle of incidence and at frequencies above 1 kHz. Although the standalone mouth simulator demonstrates a higher overall RMSE than either of the two HATS, significant differences are only seen in the rear plane (e.g. greater than 150⁰ angle of incidence). Finally, the RMSE of the two HATS generally increases with frequency and angle of incidence.

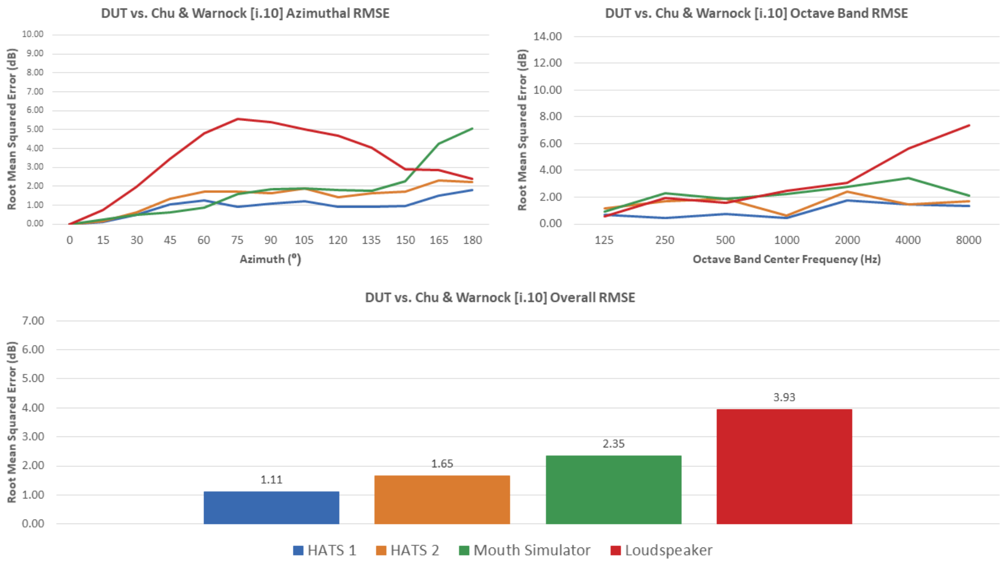


Figure D.3.1-2: Horizontal plane directivity RMSE between DUTs and Chu & Warnock data [i.10]

D.3.2 Comparing DUTs with Monson & Hunter data

Figure D.3.2-1 presents horizontal plane directivity of the four DUTs compared to the human data reported by Monson & Hunter [i.12]. As in clause D.3.1, the loudspeaker is more directional than human speech. However, the HATS directivity patterns are, on average, less directive than human speech due to the deeper rear hemisphere notch in the subjective data from [i.12]. Furthermore, the rear hemisphere difference in directivity between the mouth simulator and [i.12] data is more pronounced than in clause D.3.1.

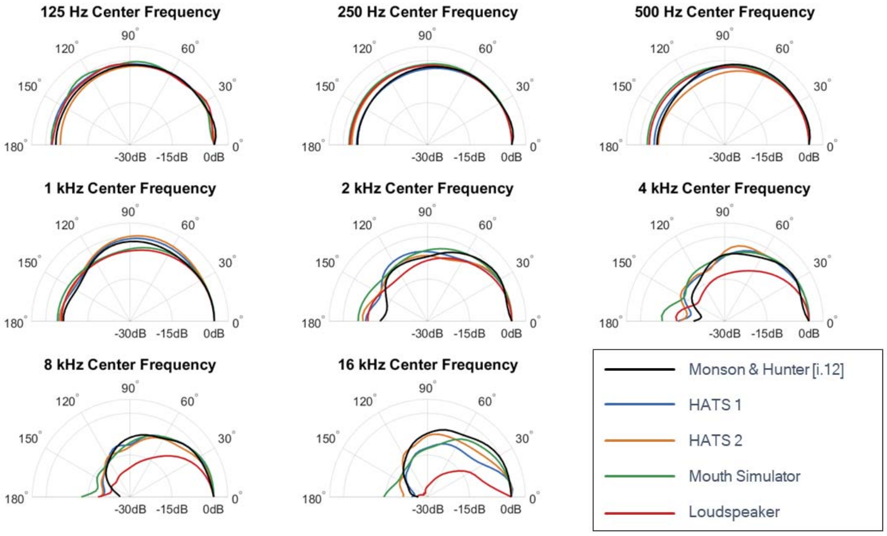


Figure D.3.2-1: Horizontal plane directivity of DUTs and Monson & Hunter data [i.12]

Figure D.3.2-2 presents the RMSE between each DUT and the [i.12] data calculated by azimuth angle of incidence, octave frequency band, and overall. The trends in RMSE match those presented in clause D.3.1. However, with the inclusion of the 16 kHz octave band, a general increase in overall RMSE is seen for all DUTs. Furthermore, the rear hemisphere RMSE of the standalone mouth simulator is exacerbated due to the deeper antinode in human speech directivity measured in [i.12].

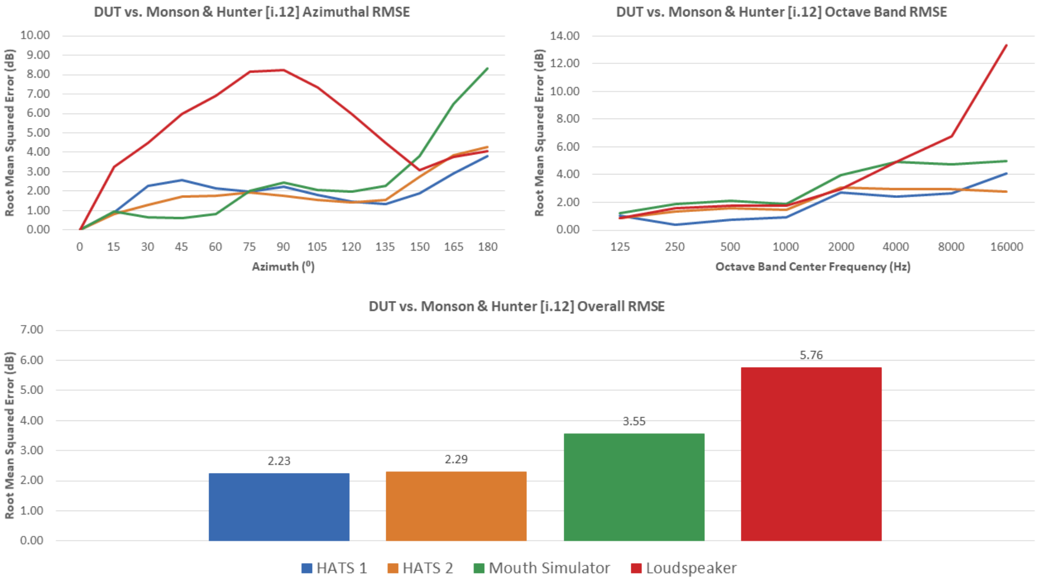


Figure D.3.2-2: Horizontal plane directivity RMSE between DUTs and Monson & Hunter data [i.12]

D.4 Upper hemisphere directivity

Figure D.4-1 presents the upper hemisphere (+30º elevation) directivity of the four DUTs compared to the data reported by Chu & Warnock [i.10]. No elevation data was provided in [i.12]. All DUTs tend to be more directive than the human speech, particularly at high frequencies.

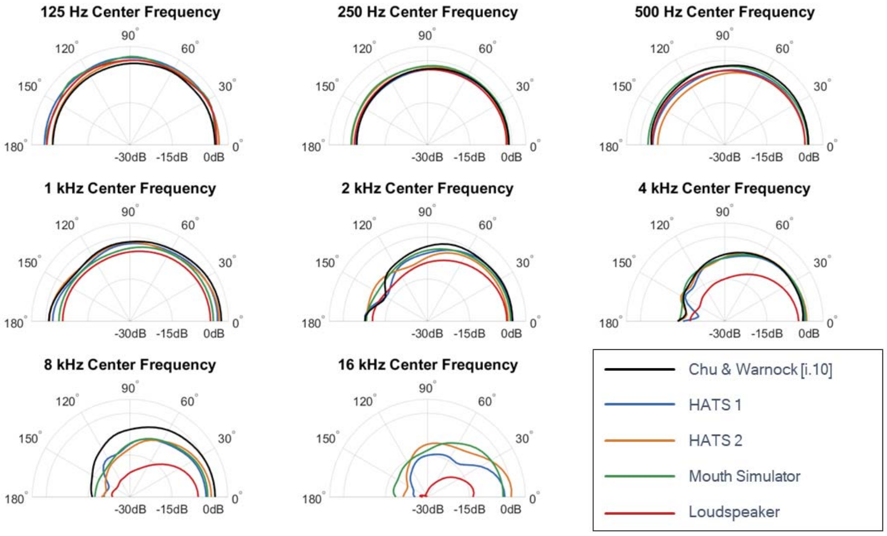


Figure D.4-1: Upper hemisphere (30º elevation) directivity of DUTs and Chu & Warnock data [i.10]

Figure D.4-2 presents the upper hemisphere RMSE between each DUT and the [i.10] data calculated by azimuth angle of incidence, octave frequency band, and overall. There is little to no difference between the RMSE of either HATS or the standalone mouth simulator. As before, the loudspeaker demonstrates the highest levels of RMSE, particularly at lateral angles of incidence and frequencies above 1 kHz.

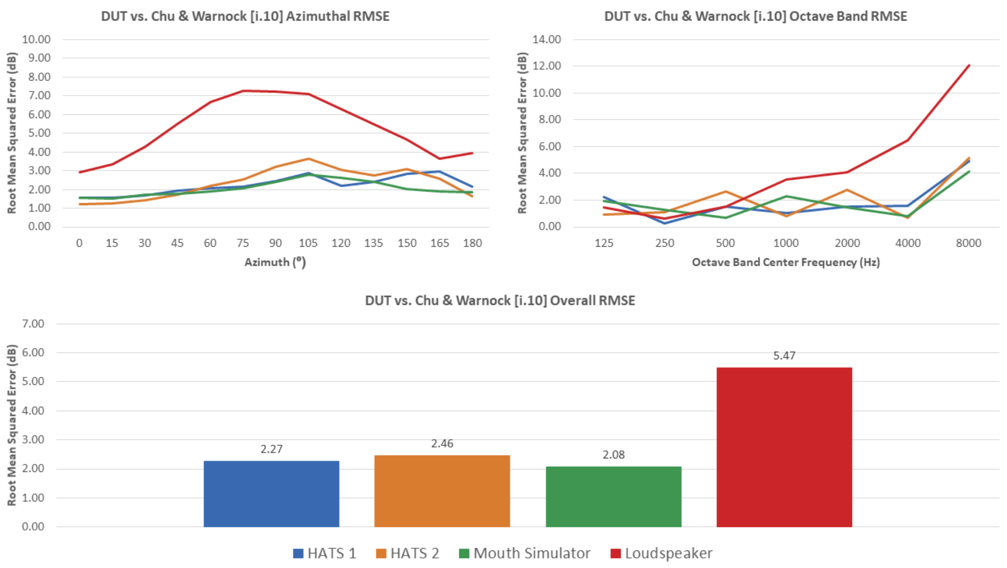


Figure D.4-2: Upper hemisphere directivity RMSE between DUTs and Chu & Warnock data [i.10]

D.5 Lower hemisphere directivity

Figure D.5-1 presents the lower hemisphere (-30º elevation) directivity of the four DUTs compared to the data reported by Chu & Warnock [i.10]. Again, the loudspeaker is more directional than [i.10] data at frequencies above 1 kHz. Furthermore, both HATS are slightly more directional than the human speech despite an increase in rear hemisphere energy at frequencies around 1 kHz. Finally, the mouth simulator is slightly less directional than the human speech on average.

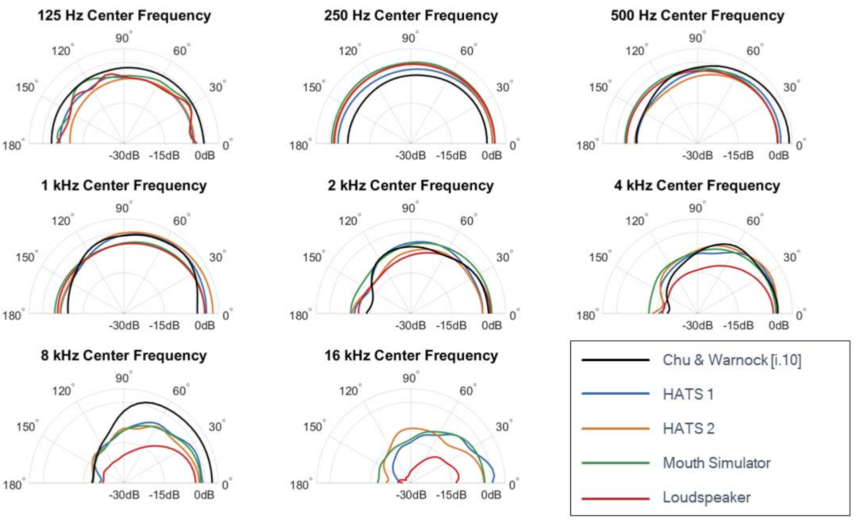


Figure D.5-1: Lower hemisphere (-30º elevation) directivity of DUTs and Chu & Warnock data [i.10]

Figure D.5-2 presents the lower hemisphere RMSE between each DUT and the [i.10] data calculated by azimuth angle of incidence, octave frequency band, and overall. As in clause D.4, the two HATS and standalone mouth simulator demonstrate similar RMSE. However, there is a general increase RMSE in the lower hemisphere as compared to the upper hemisphere. Furthermore, the loudspeaker demonstrates higher RMSE at lateral angles of incidence and frequencies above 2 kHz.

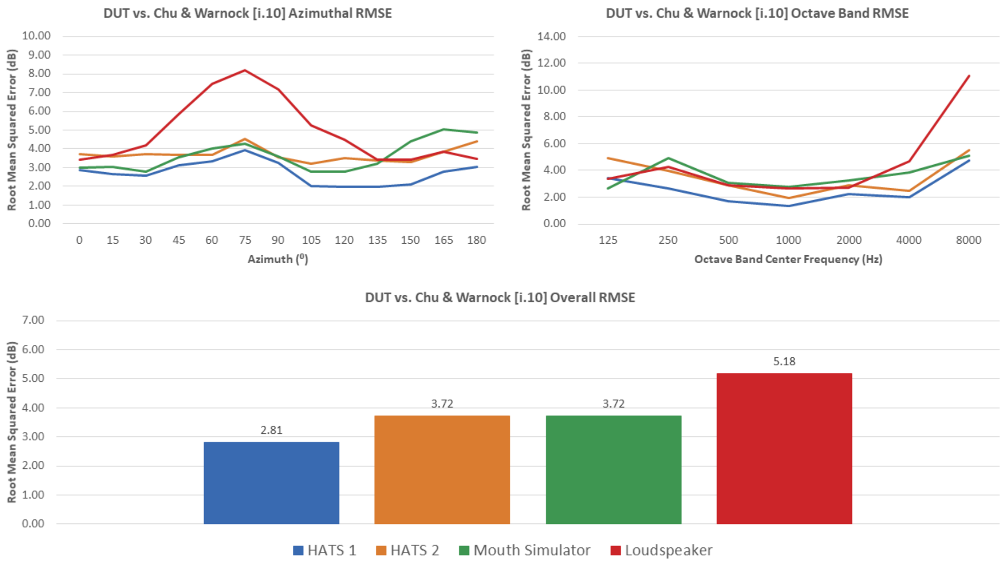


Figure D.5-2: Lower hemisphere directivity RMSE between DUTs and Chu & Warnock data [i.10]

D.6 Discussion

In the previous clauses, the reproduction directivities of two HATS, a standalone mouth simulator, and a single-driver loudspeaker were compared to human speech directivity data from [i.10] and [i.12].

In-line with the findings in Chu & Warnock [i.10] and Bozolli et al. [i.11], both HATS differ from the measured directivity of human speech, particularly in the rear hemisphere, at low elevations, and at high frequencies. However, the HATS reproduction is far more indicative of human speech directivity than the conventional loudspeaker. Furthermore, despite increased deviation in the horizontal plane, the standalone mouth simulator replicates upper and lower hemisphere directivity with as much accuracy as the HATS.

Furthermore, the results demonstrate a significant divergence in directivity between human speech and a 4,25" spherical single-driver loudspeaker. This is particularly apparent at frequencies above 1 kHz and with angles of incidence around 90º. These findings were consistent in the upper and lower hemispheres as well as on the horizontal plane.

D.7 Raw data

Table D.7-1: HATS 1 directivity data

Azimuth	Elevation	Octave band response (dB)							
(º)	(º)	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	16 kHz
0	0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
15	0	-0,01	0,00	0,13	-0,07	-0,06	-0,07	-0,28	-1,32
30	0	-0,05	-0,21	0,05	-0,47	-0,84	-0,95	-1,37	-6,82
45	0	-0,25	-0,63	-0,18	-0,46	-2,61	-2,01	-3,01	-9,23
60	0	-0,58	-1,34	-0,76	-0,18	-4,82	-2,44	-5,22	-9,50
75	0	-0,89	-2,10	-1,57	-0,12	-5,48	-4,07	-7,44	-10,28
90	0	-1,18	-2,91	-2,56	-0,61	-5,19	-6,86	-11,30	-12,39
105	0	-1,01	-3,65	-3,62	-1,83	-4,99	-10,29	-11,20	-13,90
120	0	-1,63	-4,29	-4,52	-4,05	-5,54	-11,52	-14,33	-17,13
135	0	-1,78	-4,64	-4,96	-6,40	-7,50	-11,90	-17,31	-19,06
150	0	-1,85	-4,80	-5,04	-6,96	-10,83	-13,24	-18,72	-22,13
165	0	-1,96	-4,82	-4,94	-5,66	-9,79	-17,46	-20,53	-24,98
180	0	-1,99	-4,77	-4,90	-4,82	-7,94	-13,36	-19,47	-24,23
0	30	1,74	-1,31	-1,49	1,13	-0,34	-1,67	-2,72	-2,32
15	30	1,75	-1,29	-1,39	1,13	-0,29	-1,73	-3,09	-3,75
30	30	1,75	-1,42	-1,46	0,56	-0,42	-2,14	-4,04	-8,07
45	30	1,64	-1,69	-1,66	-0,37	-0,88	-2,82	-5,80	-13,49
60	30	1,42	-2,20	-2,12	-1,32	-2,05	-4,10	-7,23	-13,76
75	30	1,21	-2,70	-2,73	-1,89	-3,79	-5,95	-8,51	-14,17
90	30	1,03	-3,24	-3,46	-2,34	-5,70	-7,52	-10,99	-15,09
105	30	1,07	-3,71	-4,17	-3,17	-6,99	-8,64	-13,84	-16,31
120	30	0,69	-4,15	-4,74	-4,53	-8,54	-11,27	-14,09	-18,94
135	30	0,56	-4,42	-4,88	-5,20	-11,60	-13,68	-16,49	-21,31
150	30	0,50	-4,56	-4,80	-4,41	-12,77	-13,76	-21,09	-24,06
165	30	0,44	-4,62	-4,58	-3,10	-9,47	-19,77	-21,38	-24,76
180	30	0,45	-4,62	-4,42	-2,49	-7,81	-15,14	-20,73	-25,58
0	-30	-4,37	0,08	0,67	0,44	-3,54	-0,30	-0,71	0,27
15	-30	-4,36	0,07	0,74	0,70	-3,55	-0,33	-0,96	-0,52
30	-30	-4,42	-0,16	0,45	1,06	-3,74	-0,55	-1,70	-4,67
45	-30	-4,64	-0,59	-0,08	1,31	-3,33	-1,13	-4,57	-5,26
60	-30	-5,02	-1,29	-0,99	1,01	-2,72	-3,98	-4,62	-9,54
75	-30	-5,36	-1,97	-2,12	0,22	-2,86	-7,07	-7,24	-11,59
90	-30	-5,69	-2,62	-3,43	-1,00	-3,78	-7,80	-10,45	-14,80
105	-30	-4,94	-3,08	-4,84	-2,69	-5,08	-9,33	-13,40	-16,26
120	-30	-6,06	-3,37	-6,26	-4,90	-7,08	-11,03	-14,45	-18,71
135	-30	-6,12	-3,41	-7,16	-6,73	-10,09	-12,33	-15,52	-20,61
150	-30	-6,15	-3,30	-7,60	-7,32	-12,48	-14,19	-17,60	-22,66
165	-30	-6,31	-3,17	-7,62	-6,99	-10,87	-17,29	-20,83	-24,79
180	-30	-6,38	-3,08	-7,59	-6,68	-9,22	-15,69	-20,85	-25,08

Table D.7-2: HATS 2 directivity data

Azimuth (º)	Elevation (º)	Octave band response (dB)							
		125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	16 kHz
0	0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
15	0	0,00	-0,05	-0,04	-0,14	-0,03	-0,14	-0,55	-0,85
30	0	-0,16	-0,22	-0,32	-0,23	-0,67	-0,48	-2,12	-4,03
45	0	-0,42	-0,47	-0,87	0,16	-2,43	-1,35	-4,70	-6,17
60	0	-0,89	-0,79	-1,71	0,71	-5,39	-2,46	-5,98	-7,00
75	0	-1,45	-1,16	-2,85	0,85	-7,06	-2,38	-8,42	-7,13
90	0	-2,03	-1,54	-4,21	0,24	-6,60	-5,07	-12,03	-8,03
105	0	-2,72	-1,93	-5,56	-1,23	-6,80	-10,65	-13,43	-11,13
120	0	-3,36	-2,30	-6,56	-3,49	-7,20	-10,82	-13,81	-12,81
135	0	-4,06	-2,62	-6,91	-5,71	-8,92	-11,36	-17,02	-17,94
150	0	-4,58	-2,87	-6,78	-6,35	-10,21	-11,86	-18,26	-18,63
165	0	-4,97	-3,02	-6,51	-5,57	-8,04	-15,77	-19,88	-20,34
180	0	-5,24	-3,06	-6,37	-4,99	-7,02	-14,00	-19,42	-21,30
0	30	1,65	-1,46	-1,45	1,70	-1,34	-0,79	-0,87	-0,01
15	30	1,67	-1,47	-1,49	1,47	-1,48	-1,08	-1,31	-0,61
30	30	1,52	-1,54	-1,68	0,48	-1,86	-1,69	-2,57	-3,93
45	30	1,31	-1,66	-2,06	-0,94	-2,42	-2,33	-4,47	-7,83
60	30	0,97	-1,81	-2,68	-1,91	-3,30	-3,34	-6,87	-9,63
75	30	0,53	-1,96	-3,55	-1,92	-4,80	-4,73	-9,20	-10,25
90	30	0,07	-2,14	-4,61	-1,87	-7,24	-7,19	-12,62	-11,08
105	30	-0,49	-2,34	-5,65	-2,53	-9,41	-9,29	-14,98	-13,24
120	30	-1,00	-2,54	-6,38	-3,66	-9,08	-10,33	-16,55	-16,71
135	30	-1,52	-2,73	-6,64	-4,02	-7,89	-12,75	-18,29	-19,34
150	30	-1,93	-2,89	-6,53	-3,03	-7,77	-12,01	-19,86	-21,74
165	30	-2,20	-3,02	-6,31	-1,84	-8,19	-15,63	-21,13	-21,98
180	30	-2,41	-3,07	-6,13	-1,30	-8,07	-13,40	-19,79	-21,28
0	-30	-4,02	0,23	-0,32	2,86	-3,59	-0,33	-2,05	-2,80
15	-30	-4,00	0,17	-0,42	2,82	-3,92	-0,07	-2,87	-3,84
30	-30	-4,12	0,06	-0,82	2,64	-4,42	0,33	-4,62	-6,75
45	-30	-4,29	-0,11	-1,52	2,19	-4,32	-1,12	-5,71	-10,26
60	-30	-4,87	-0,31	-2,48	1,52	-4,47	-2,53	-5,75	-9,78
75	-30	-5,54	-0,53	-3,69	0,73	-5,46	-4,03	-9,20	-9,61
90	-30	-6,13	-0,77	-5,11	-0,18	-6,91	-6,57	-9,70	-9,69
105	-30	-6,91	-1,03	-6,56	-1,50	-8,46	-8,52	-12,92	-10,99
120	-30	-7,69	-1,29	-7,63	-3,45	-10,74	-9,66	-15,31	-14,74
135	-30	-8,64	-1,55	-7,94	-5,69	-12,26	-11,27	-14,91	-16,72
150	-30	-9,24	-1,75	-7,63	-6,93	-12,24	-14,32	-16,53	-18,38
165	-30	-9,77	-1,89	-7,14	-6,75	-10,16	-16,83	-18,07	-21,03
180	-30	-10,16	-1,95	-6,87	-6,42	-8,34	-13,29	-18,58	-20,72

Table D.7-3: Mouth simulator directivity data

Azimuth (º)	Elevation (º)	Octave band response (dB)							
		125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	16 kHz
0	0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
15	0	-0,70	0,08	0,02	-0,11	-0,18	-0,11	-0,10	-0,56
30	0	0,49	0,01	-0,19	-0,48	-0,65	-0,51	-1,02	-1,93
45	0	-0,39	-0,28	-0,50	-1,13	-1,36	-1,35	-3,14	-3,87
60	0	-0,78	-0,51	-0,88	-1,97	-2,28	-2,74	-5,16	-6,05
75	0	-0,18	-0,72	-1,24	-2,98	-3,36	-4,57	-7,49	-10,16
90	0	-0,52	-1,23	-1,80	-4,27	-4,67	-6,73	-10,50	-12,62
105	0	-1,37	-1,62	-2,16	-5,27	-6,02	-8,48	-13,32	-14,26
120	0	-0,59	-1,66	-2,21	-5,50	-7,28	-9,44	-15,24	-16,55
135	0	-0,89	-1,76	-2,27	-5,25	-8,06	-11,09	-16,15	-17,60
150	0	-2,67	-2,02	-2,36	-4,80	-7,24	-13,22	-16,39	-18,08
165	0	-2,11	-2,09	-2,44	-4,37	-5,92	-9,65	-18,03	-16,69
180	0	-2,42	-2,05	-2,42	-4,16	-5,32	-7,77	-12,61	-14,33
0	30	0,76	-0,94	-0,49	-0,36	-0,44	-1,20	-2,18	-2,92
15	30	0,91	-0,84	-0,48	-0,43	-0,59	-1,32	-2,39	-3,03
30	30	2,15	-0,84	-0,65	-0,73	-1,01	-1,70	-3,42	-3,76
45	30	0,93	-1,15	-0,97	-1,27	-1,70	-2,48	-5,28	-5,25
60	30	0,52	-1,35	-1,28	-2,00	-2,50	-3,77	-6,77	-7,83
75	30	1,04	-1,51	-1,58	-2,86	-3,48	-5,45	-8,39	-10,33
90	30	1,45	-1,86	-2,09	-3,98	-4,74	-7,29	-10,80	-13,29
105	30	0,53	-2,29	-2,42	-4,94	-6,21	-8,84	-13,60	-14,73
120	30	0,50	-2,32	-2,46	-5,29	-7,70	-9,69	-15,83	-15,77
135	30	1,05	-2,32	-2,53	-5,32	-9,02	-10,80	-17,04	-16,76
150	30	-0,60	-2,71	-2,69	-5,13	-9,37	-12,82	-17,29	-17,69
165	30	-0,46	-2,83	-2,72	-4,87	-8,73	-14,16	-17,48	-17,54
180	30	-0,44	-2,80	-2,71	-4,71	-8,23	-13,70	-17,37	-18,63
0	-30	-3,30	0,96	-0,59	-0,19	-0,21	-1,11	-1,48	-2,49
15	-30	-4,74	1,03	-0,58	-0,30	-0,35	-1,30	-1,64	-2,87
30	-30	-1,73	0,97	-0,78	-0,64	-0,73	-1,77	-2,72	-3,48
45	-30	-3,20	0,75	-1,12	-1,21	-1,35	-2,60	-4,79	-5,28
60	-30	-3,38	0,59	-1,48	-1,94	-2,20	-3,78	-6,18	-8,18
75	-30	-4,42	0,40	-1,87	-2,83	-3,18	-5,34	-8,07	-10,51
90	-30	-4,94	-0,06	-2,39	-3,95	-4,49	-7,13	-10,45	-13,72
105	-30	-4,81	-0,37	-2,76	-4,88	-5,96	-8,46	-13,10	-15,13
120	-30	-3,02	-0,36	-2,83	-5,19	-7,33	-9,11	-15,48	-15,89
135	-30	-3,84	-0,47	-2,93	-5,17	-8,59	-9,89	-15,93	-16,76
150	-30	-7,31	-0,71	-3,10	-4,94	-8,86	-11,26	-17,26	-16,97
165	-30	-5,91	-0,81	-3,17	-4,70	-8,19	-11,96	-18,20	-17,75
180	-30	-6,34	-0,75	-3,19	-4,51	-7,85	-12,03	-18,06	-17,63

Table D.7-4: Loudspeaker directivity data

Azimuth (º)	Elevation (º)	Octave band response (dB)							
		125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	16 kHz
0	0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
15	0	0,20	-0,05	-0,13	-0,26	-0,51	-1,05	-1,81	-9,18
30	0	1,21	-0,28	-0,40	-0,87	-1,69	-3,36	-4,79	-12,38
45	0	-0,74	-0,73	-0,85	-1,75	-3,33	-6,35	-9,19	-17,11
60	0	-1,10	-1,12	-1,36	-2,83	-5,06	-9,33	-13,77	-21,14
75	0	-1,36	-1,47	-1,86	-3,91	-6,71	-11,64	-17,00	-24,55
90	0	-1,41	-1,73	-2,30	-4,95	-8,33	-13,35	-19,50	-26,19
105	0	-1,06	-1,95	-2,66	-5,55	-9,97	-14,76	-20,81	-27,28
120	0	-1,94	-2,15	-2,88	-5,99	-11,31	-16,49	-21,80	-28,52
135	0	-2,22	-2,31	-3,05	-6,09	-11,61	-17,94	-22,80	-28,06
150	0	-2,50	-2,44	-3,11	-5,91	-10,49	-18,07	-22,43	-28,58
165	0	-2,73	-2,49	-3,15	-5,70	-9,16	-14,23	-23,03	-27,25
180	0	-2,25	-2,39	-3,16	-5,65	-8,77	-12,99	-18,50	-26,54
0	30	0,87	-1,92	-1,40	-1,34	-1,75	-3,68	-5,55	-13,70
15	30	1,32	-1,88	-1,50	-1,53	-2,16	-4,56	-6,70	-13,99
30	30	1,76	-2,14	-1,81	-1,97	-3,16	-6,22	-9,53	-16,84
45	30	0,70	-2,42	-2,24	-2,66	-4,49	-8,39	-13,41	-20,09
60	30	0,36	-2,74	-2,74	-3,46	-6,00	-10,78	-17,11	-23,88
75	30	0,34	-3,04	-3,20	-4,41	-7,55	-13,10	-19,75	-26,31
90	30	0,14	-3,21	-3,59	-5,30	-9,07	-14,95	-21,27	-27,93
105	30	0,01	-3,45	-3,87	-6,05	-10,56	-16,15	-22,63	-28,55
120	30	0,07	-3,51	-4,04	-6,49	-11,69	-17,51	-23,26	-28,95
135	30	-0,45	-3,76	-4,16	-6,52	-12,27	-18,58	-24,10	-28,94
150	30	-0,52	-3,95	-4,24	-6,30	-11,84	-19,66	-24,17	-29,71
165	30	-0,61	-4,04	-4,25	-6,13	-10,83	-18,62	-23,54	-27,71
180	30	-0,36	-3,98	-4,22	-6,03	-10,41	-17,72	-23,75	-28,33
0	-30	-3,12	1,01	-0,39	-0,14	-0,72	-2,03	-3,34	-13,25
15	-30	-4,59	0,90	-0,54	-0,38	-1,10	-2,84	-4,25	-12,45
30	-30	-2,11	0,72	-0,89	-0,93	-2,12	-4,70	-6,77	-14,87
45	-30	-4,62	0,41	-1,42	-1,68	-3,54	-7,16	-10,66	-16,35
60	-30	-5,40	0,01	-2,05	-2,56	-5,11	-9,70	-14,55	-20,31
75	-30	-5,40	-0,32	-2,65	-3,45	-6,74	-12,08	-17,70	-23,83
90	-30	-5,29	-0,64	-3,13	-4,26	-8,45	-13,90	-19,68	-25,57
105	-30	-3,83	-0,91	-3,43	-5,01	-10,10	-14,76	-21,08	-27,03
120	-30	-6,62	-1,20	-3,67	-5,57	-11,50	-15,65	-21,51	-27,38
135	-30	-4,81	-1,32	-3,75	-5,78	-12,37	-16,44	-21,86	-27,39
150	-30	-7,18	-1,56	-3,81	-5,67	-11,91	-17,53	-21,64	-27,94
165	-30	-7,71	-1,62	-3,81	-5,59	-11,00	-17,54	-22,22	-25,80
180	-30	-5,29	-1,54	-3,81	-5,56	-10,70	-17,00	-22,11	-27,08

Annex E (informative): Reverberation time data and characteristics

E.1 Introduction

Reverberation characteristics are an important feature of real-life acoustic environments. Increased reverberation time may reduce speech intelligibility [i.13] and, in turn, impact the performance of voice assistant device functionalities [i.9]. It is therefore important to test voice assistant devices under a variety of reverberation conditions, which emulate common acoustic environments.

This annex summarizes published research on the reverberation time (T_{60}) of ordinary indoor spaces presented by Diaz and Pedrero [i.14]. Further reverberation time values corresponding to room impulse responses from ETSI TS 103 557 [2] are reported. Lastly, T_{60} data from an example home-like test environment built to the specifications in clause 4.2.2 are provided.

E.2 Published data

Diaz and Pedrero collected a corpus of reverberation measurements taken in a variety of furnished and unfurnished rooms in ordinary dwellings [i.14]. Octave-band reverberation time T_{60} data are provided for 11 687 domestic rooms: 8 246 furnished bedrooms, 2 111 furnished living rooms, and 230 unfurnished rooms. Diaz and Pedrero determine that T_{60} is generally proportional to the volume of the room and can be stratified into nine distinct classes [i.14]. Figure E.2-1 reproduces the nine T_{60} classes for convenience.

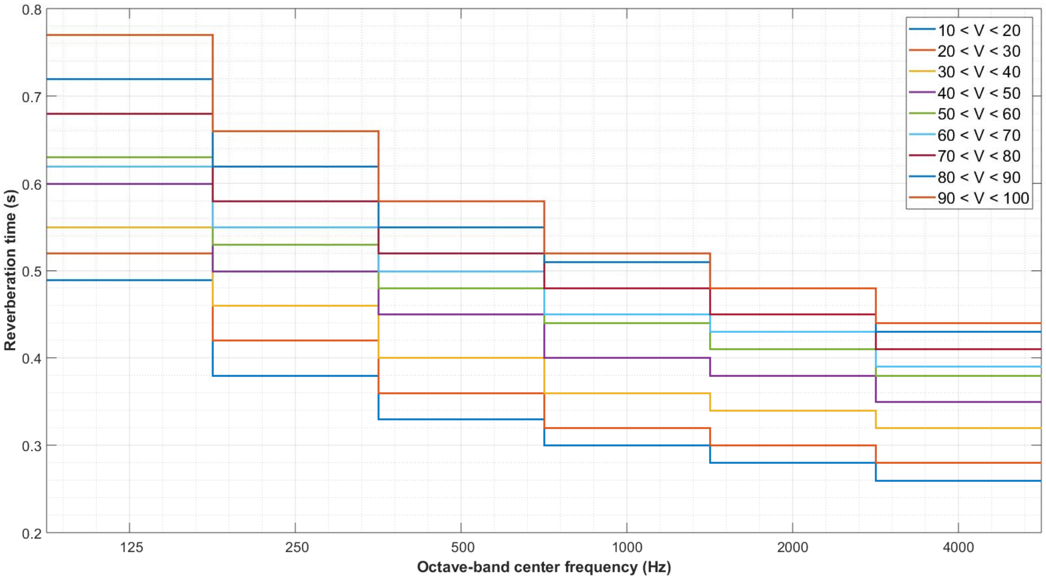
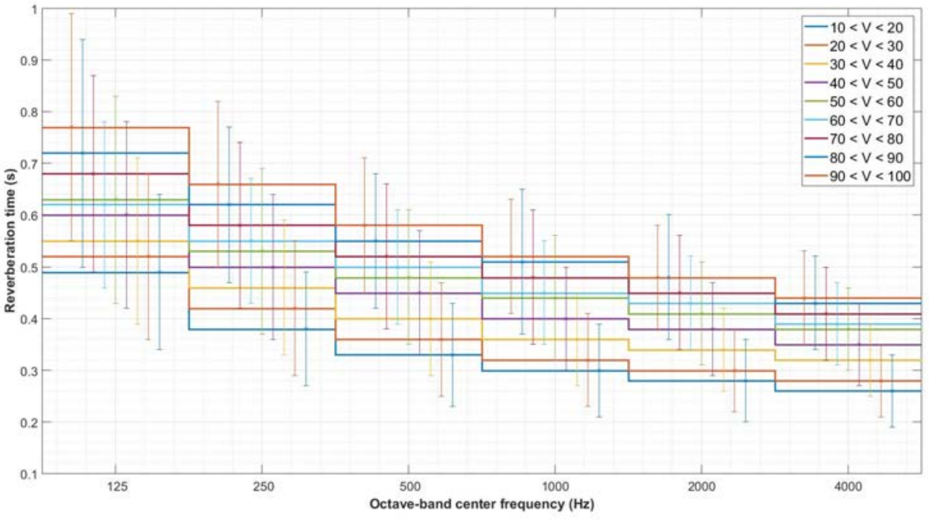


Figure E.2-1: Diaz and Pedrero [i.14] octave-band T_{60} classes based on volume in m^3

In addition to providing the mean T_{60} for each room class, Diaz and Pedrero report the octave-band standard deviation to indicate the range of reverberation characteristics per class [i.14]. Figure E.2-2 reproduces the octave-band T_{60} mean and standard deviation for convenience. The standard deviation is inversely proportional to frequency for all room classes. This indicates a wider range of low frequency reverberation characteristics for rooms within the same class.



NOTE: Diaz and Pedrero [i.14] report T_{60} for octave-bands ranging from 125 Hz to 4 kHz. Results in the remainder of this annex are reported for octave-bands ranging from 250 Hz to 8 kHz.

Figure E.2-2: Diaz and Pedrero [i.14] T_{60} classes with standard deviation

E.3 ETSI room impulse response data

ETSI TS 103 557 [2] provides a database of room impulse responses for the simulation of real-life reverberation conditions. Table E.3-1 and Figure E.3-1 present the octave-band T_{60} characteristics of four domestic spaces from ETSI TS 103 557 [2].

Table E.3-1: Octave-band T_{60} measurements in ETSI TS 103 557 [2] domestic spaces

Domestic Space	Reverberation time (T_{60} in seconds)					
	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz
Living Room	0,51	0,41	0,44	0,42	0,39	0,31
Office	0,84	0,64	0,61	0,57	0,49	0,35
Kitchen	0,39	0,44	0,55	0,61	0,57	0,46
Bathroom	0,51	0,62	0,60	0,63	0,60	0,49

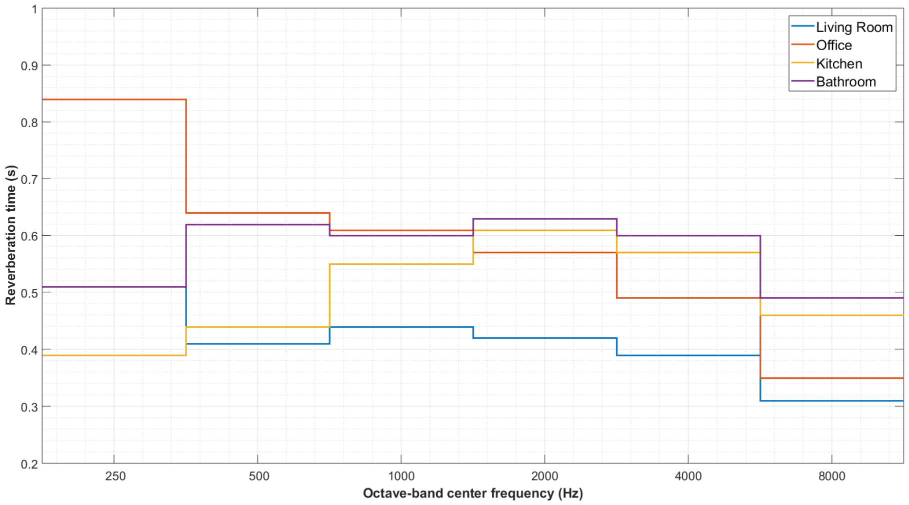


Figure E.3-1: Octave-band T_{60} measurements in ETSI TS 103 557 [2] domestic spaces

Two general trends exist in the provided T_{60} data. The first, demonstrated in the Pedrero and Diaz reverberation classes [i.14] and in the Office and Living Room data from ETSI TS 103 557 [2], is characterized by a steadily decreasing T_{60} with maximum reverberation in the low frequencies. The second T_{60} trend, demonstrated by the kitchen and bathroom environments from ETSI TS 103 557 [2], is concave as opposed to monotonically decreasing. The octave-band T_{60} increases from 250 Hz to 2 kHz and decreases in the higher frequencies above 2 kHz.

E.4 Home-like test environment T_{60} data

Reverberation time measurements were taken in an example home-like test environment built to the specifications in clause 4.2.2. Measurements were conducted in both the damped layout presented in Figure 4.2.2.2-1 (Room Layout 1) as well as the more reverberant layout shown in Figure 4.2.2.3-1 (Room Layout 2).

For each layout, the octave-band reverberation time was measured from each of the three talker locations to each of the three DUT locations shown in Figure 4.2.2.2-1 and Figure 4.2.2.3-1, respectively. The overall T_{60} for each octave-band and room layout was determined as the arithmetic mean between each of the nine talker and DUT position combinations.

The sound source used for excitation signal reproduction was omnidirectional. Measurements were made with a reference microphone placed 10 cm above the working surface at each DUT position. The T_{60} was estimated through extrapolation of T_{30} measurements made according to the ISO 3382-2 interrupted noise technique [i.15].

Tables E.4-1 and E.4-2 present the T_{60} measurements made in the home-like test environment for acoustic conditions 1 and 2, respectively.

Table E.4-1: Reverberation time measured in home-like test environment (Room Layout 1)

Talker Position	DUT Position	Reverberation time (T_{60} in seconds)					
		250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz
1 (sink)	1 (entertainment center)	0,37	0,32	0,31	0,36	0,35	0,33
1 (sink)	2 (cabinet)	0,47	0,30	0,31	0,33	0,33	0,32
1 (sink)	3 (center table)	0,39	0,31	0,31	0,34	0,34	0,34
2 (corner)	1 (entertainment center)	0,44	0,36	0,32	0,31	0,32	0,31
2 (corner)	2 (cabinet)	0,50	0,31	0,34	0,34	0,36	0,34
2 (corner)	3 (center table)	0,43	0,33	0,32	0,32	0,34	0,33
3 (couch)	1 (entertainment center)	0,46	0,31	0,32	0,36	0,36	0,34
3 (couch)	2 (cabinet)	0,52	0,38	0,31	0,34	0,35	0,32
3 (couch)	3 (center table)	0,41	0,32	0,32	0,32	0,33	0,31
Mean		0,44	0,33	0,32	0,34	0,34	0,33
Standard Deviation		0,05	0,03	0,01	0,02	0,01	0,01

Table E.4-2: Reverberation time measured in home-like test environment (Room Layout 2)

Talker Position	DUT Position	Reverberation time (T_{60} in seconds)					
		250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz
1 (sink)	1 (entertainment center)	0,49	0,41	0,40	0,46	0,52	0,49
1 (sink)	2 (cabinet)	0,56	0,39	0,41	0,45	0,49	0,46
1 (sink)	3 (center table)	0,46	0,42	0,43	0,48	0,50	0,48
2 (corner)	1 (entertainment center)	0,46	0,41	0,41	0,45	0,51	0,50
2 (corner)	2 (cabinet)	0,47	0,41	0,40	0,48	0,51	0,48
2 (corner)	3 (center table)	0,54	0,38	0,41	0,45	0,52	0,50
3 (couch)	1 (entertainment center)	0,46	0,43	0,41	0,45	0,51	0,48
3 (couch)	2 (cabinet)	0,57	0,41	0,41	0,44	0,51	0,49
3 (couch)	3 (center table)	0,48	0,40	0,41	0,46	0,51	0,48
Mean		0,50	0,41	0,41	0,46	0,51	0,48
Standard Deviation		0,05	0,02	0,01	0,01	0,01	0,01

Figure E.4-1 presents the mean and standard deviation T_{60} values measured in the example home-like test environment.

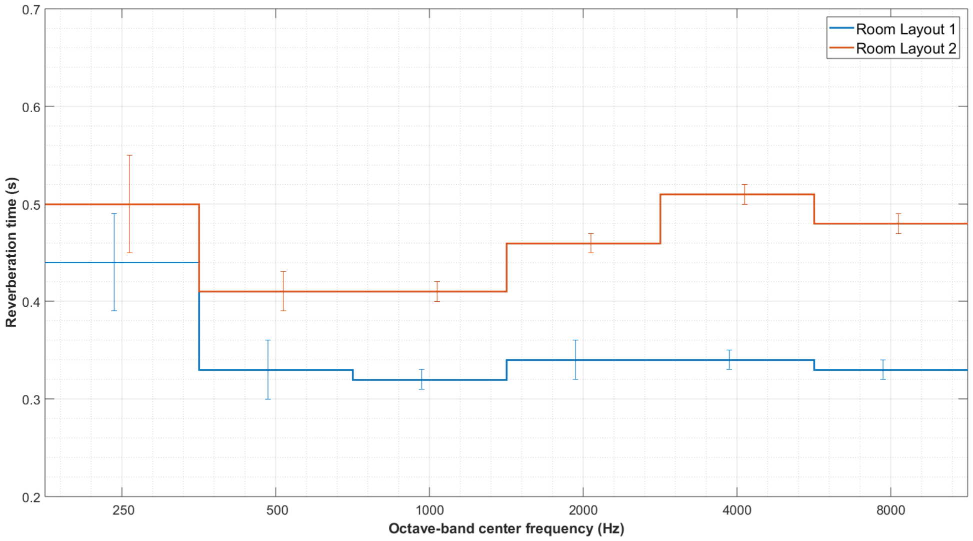


Figure E.4-1: Octave-band T_{60} measured in home-like environment

Similar to the data presented in Diaz and Pedrero [i.14], the T_{60} standard deviation measured in the home-like environment is larger at low frequencies. However, in clause E.4 the standard deviation is measured between different points in the same room whereas Diaz and Pedrero [i.14] report the standard deviation between many different rooms.

Clause 4.2.2 requires the home-like test environment to have an internal room volume between 70 and 90 m^3. The T_{60} values measured in the home-like environment are lower than the mean values presented for the appropriate room classes from Diaz and Pedrero [i.14].

The Room Layout 1 T_{60} curve is comparable to the Living Room curve from ETSI TS 103 557 [2]. However, the Room Layout 2 T_{60} curve presents a new trend. It is flat to ±0,05 s around a nominal reverberation time of 0,46 s.

E.5 Recommendations

Based on the T_{60} data presented in this annex E, several recommendations for the reproduction of common reverberation characteristics for voice assistant device testing are provided:

- T_{60} requirements should cover octave-bands ranging from 250 Hz to at least 8 kHz.
- Multiple trends in T_{60} should be tested (e.g. monotonically decreasing T_{60}, concave T_{60}, flat T_{60} curves).
- Tolerances for octave-band T_{60} requirements should reflect increased standard deviation in low frequencies.
- The lower bound of T_{60} requirements should be aligned with the example home-like test environment data.
- Reverberation time characterization of the home-like test environment should be averaged over multiple talker and DUT positions.

Annex F (informative): Generation of test signals for barge-in

F.1 Introduction

Clause 5.3.1 of the present document presents several barge-in conditions for testing voice assistant devices. Tables 5.3.1.2-1, 5.3.1.2-2 and 5.3.1.2-3 detail the barge-in test content for a variety of acoustic conditions and device rendering capabilities. Annex F describes the generation of stereo music, correlated pink noise, and decorrelated pink noise barge-in test signals. While pink noise signals are mathematically well-defined and commonly understood, some details are presented so that the publicly available test files align with the proposed stereo music signal.

Test signals for evaluation of barge-in performance should have both adequate technical properties for effective testing and should be generally available for users of the test specification. Music poses a particular challenge due to licensing issues. To eliminate issues with distribution, the music signal generated in annex F has been developed based on existing and readily available ETSI and ITU-T signals. In the database associated with ETSI ES 202 396-1 [i.16], there is a music-like signal available in stereo (binaural) format. In the database associated with Recommendation ITU-T P.501 [i.17], there is a speech signal. These were selected based on availability and appropriateness, then combined and modified to meet desirable technical properties.

F.2 Signal generation

F.2.1 Technical considerations

Barge-in testing is intended to exercise the acoustic echo control of a voice assistant device. Toward that end, the test signal should have content across the full range of frequencies of interest. The spectral content should also be distributed well in time, to provide an adequate challenge over the entire temporal and spectral range. Finally, for face-validity, the test signal should include both instrumental and vocal content.

F.2.2 Original music signal

The music-like signal in [i.16], *RockMusic01m48k_ETSI_3m18s_1sRamp.wav*, meets some of these properties, but is somewhat lacking in low-frequency energy, is lacking in vocals/speech, and is somewhat sparse in spectral-temporal terms. Figure F.2.2-1 plots the power spectrum, and Figure F.2.2-2 plots the time and spectrogram of a 10-sec segment. In Figure F.2.2-2, there are regions (example marked with white box) where there is relatively little energy.

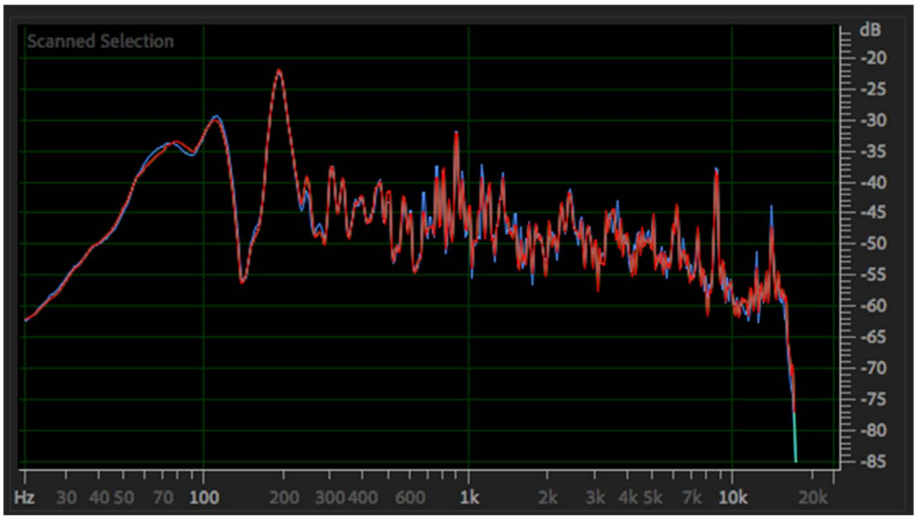


Figure F.2.2-1: Power spectrum of RockMusic from [i.16]

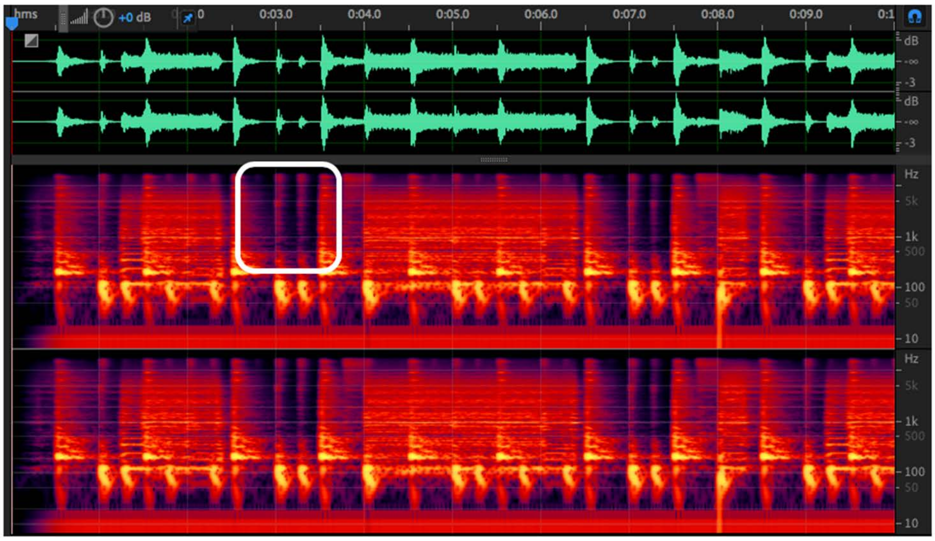


Figure F.2.2-2: Time domain (upper) and spectrogram (lower) for RockMusic from [i.16]

F.2.3 Improved music signal

To address both the lack of vocal content and the spectral-temporal sparsity, the speech signal from clause 7.3 of [i.17], *FB_male_female_single-talk_seq.wav* was mixed in. As the music content is stereo, the speech content was mixed equally in both channels. For face-validity as musical content, the cadence and timings of the speech utterances were individually adjusted to match the tempo and timing of the music. To further improve the low-frequency content, a bass track was added, consistent with the key and tempo of the source music. To enhance the audibility of the speech content, some frequency equalization was applied. Mild dynamic range compression was applied to both the speech and bass tracks, as well as to the overall mix, to improve the overall level and enhance the signal's face-validity as musical content.

Figure F.2.3-1 compares the power spectrum of the improved signal to the original. Figure F.2.3-2 shows the time-domain and spectrogram of an equivalent 10-second interval to that shown in Figure F.2.2-2. The addition of the speech and bass-track components improves the content at lower frequencies (Figure F.2.3-1) and fills in energetically sparse spectral-temporal intervals (Figure F.2.3-2).

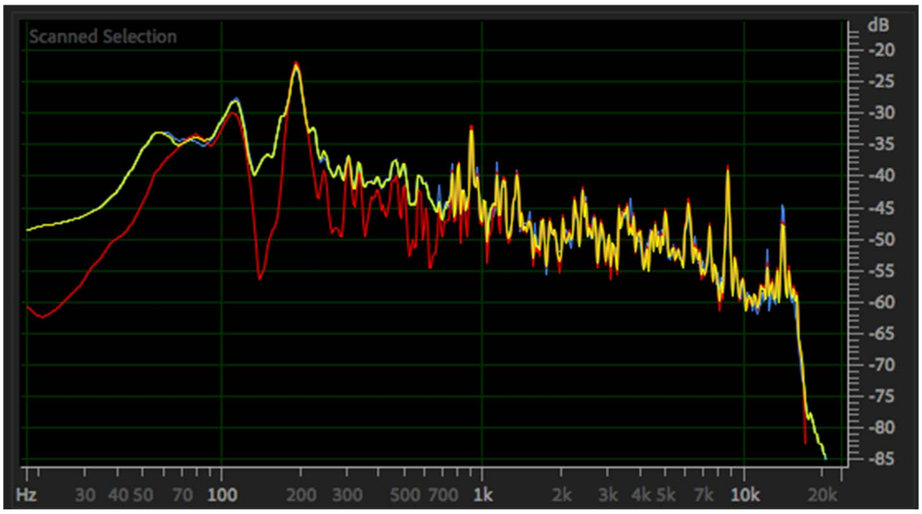


Figure F.2.3-1: Power spectrum of original (red) and improved (yellow) music signals

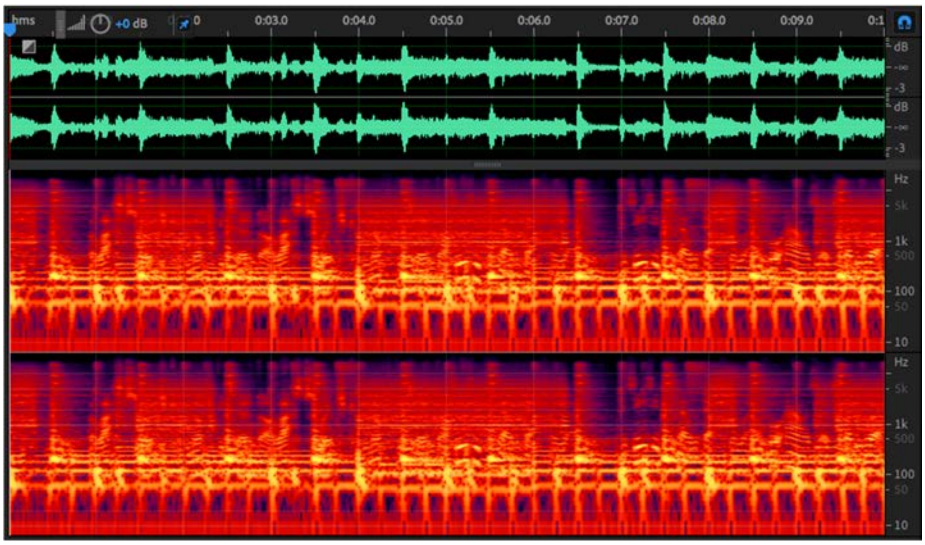


Figure F.2.3-2: Time domain (upper) and spectrogram (lower) for Improved music signal

The addition of speech content results in a small increase in overall level, while meeting the same peak values as the original signal, as shown in Table F.2.3-1.

Table F.2.3-1: Signal level measurements for original and improved music signals

	Original		Improved	
	Left	**Right**	**Left**	**Right**
Peak	-0,22	0,00	0,00	0,00
RMS	-19,67	-19,74	-18,88	-18,92
BS.1770-3, LUFS [i.18]	-16,41		-15,95	

Adding the speech content equally in each channel somewhat affects the inter-channel coherence. A typical measure for audio content is using the broadband inter-channel phase analysis. Figure F.2.3-3 shows the phase analysis over a 0,34-sec interval (16,384 samples) for the original (left) and improved (right).

The plot for the Improved music signal is slightly less open, due to the added speech content that is in common in both channels. This does not significantly change the inter-channel coherence.

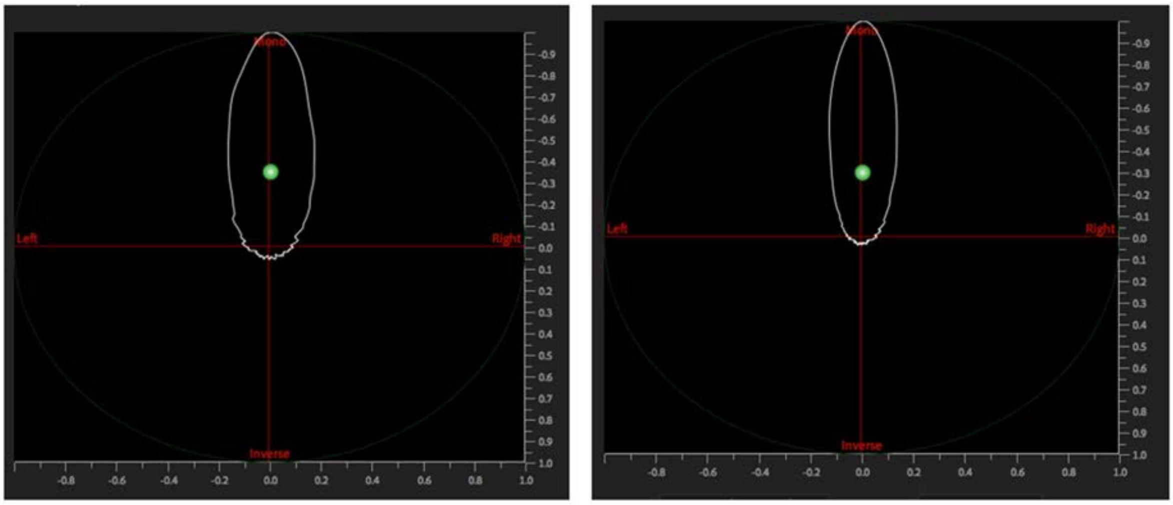


Figure F.2.3-3: Phase analysis for Original (left) and Improved (right) music signals

F.2.4 Pink noise signal

Pink noise test signals for the relevant test cases in clause 5.3.1 are constructed according to the following guidelines:

1) Level: The levels in LUFS as defined in Recommendation ITU-R BS.1770-3 [i.18] of the pink noise test signals are consistent with the level of the music signal from Table F.2.3-1.

2) Power spectrum: The pink noise test signals are filtered at 20 Hz to 20 kHz to match the common reproduction of full-range speakers. It is expected that devices that render less than this full range will apply appropriate internal filtering.

3) Temporal smoothing: The source signal for the music from [i.16] has smoothing (rise/fall) applied to the start and the end, to prevent click/pop artefact when the signal is rendered in a continuous loop. The test signals based on pink noise will have the same smoothing (rise/fall of 20 ms).

4) Coherence: Two pink noise signals are generated for each channel configuration to meet the definitions in Tables 5.3.1.2-1, 5.3.1.2-2 and 5.3.1.2-3. First, a signal where all channels are coherent (correlated), and second, where all channels are mutually incoherent (uncorrelated).

5) Multichannel assignment: For test signals with more than 2 channels, the intended channel assignment to speaker location are defined in a document included with the file, as some media containers may use different mappings (e.g. for 5.1: Ch 1: Left Front, Ch 2: Right Front, Ch 3: Centre, Ch 4: Low-frequency effects, Ch 5: Left Surround, Ch 6: Right Surround).

6) Low-frequency effects channel: For test cases with a low-frequency effects channel ("dot 1"), the test signals have no content. This is because this low-frequency energy typically lies below that of the frequencies of interest in voice commands, and because there are varying approaches to how these channels are rendered in consumer devices.

F.2.5 Stereo test signal comparison

For the stereo pink noise test signals, a high-pass filter and level is adjustment are applied to align the pink noise signals with the music signal. Figure F.2.5-1 shows the power spectra of the music signal (green trace), the correlated (red) and uncorrelated (magenta) filtered pink noise signals (note that these last two essentially overlie each other).

The levels of both channels for each pink noise sample is -16,0 LUFS, quite close to the level of -15,95 LUFS of the music signal (from Table F.2.3-1).

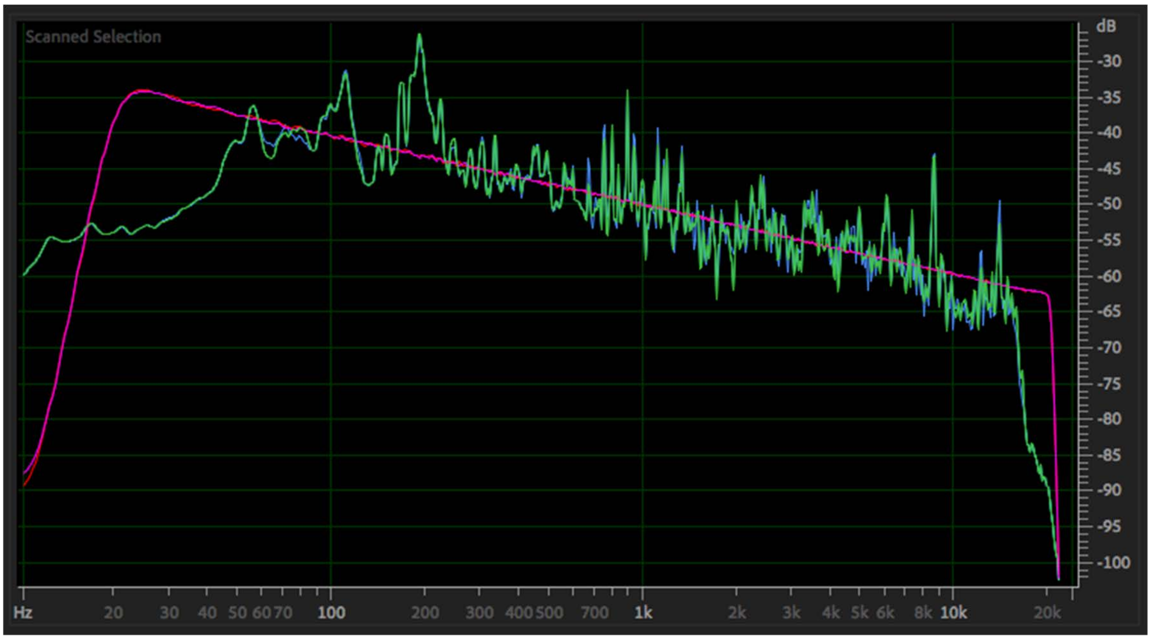


Figure F.2.5-1: Power spectra of stereo music (green/blue), correlated pink noise(red) and uncorrelated pink noise (magenta)

Figure F.2.5-2 shows the phase analysis for the correlated (left panel) and uncorrelated (right panel) pink noise signal. This figure can be compared to Figure F.2.5-2 (right panel) for the improved music test signal. As expected, the uncorrelated signal has phase analysis that is nearly circular, representing the incoherence of the phase between the two channels.

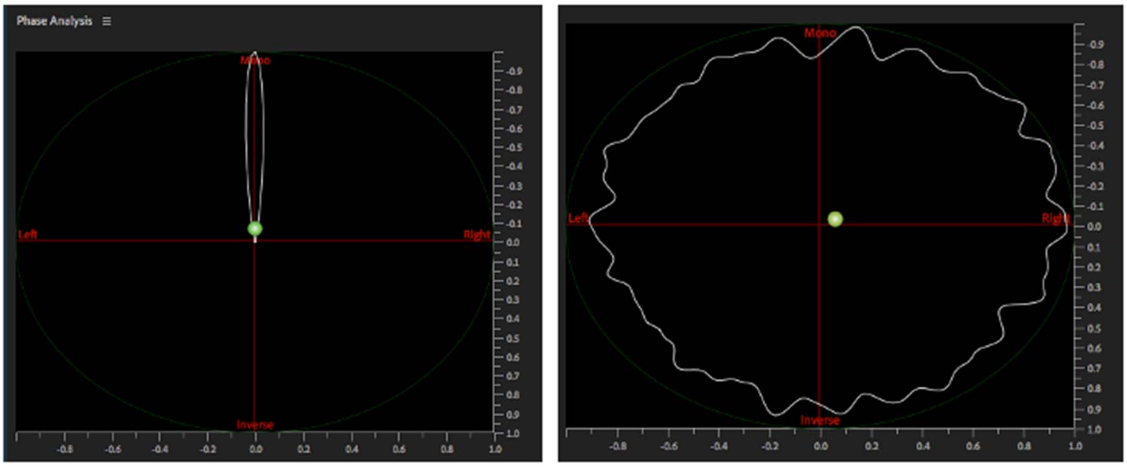


Figure F.2.5-2: Phase analysis of two-channel pink noise, correlated (left) and uncorrelated (right)

Annex G (informative): Bibliography

- International standard ISO 532 B (1975): "Method for calculating loudness".

- Zwicker E. and Fastl H. (1999): "Psychoacoustics: Facts and models", 2nd Edition, Springer-Verlag, Berlin.

- Glasberg, B. R., and Moore, B. C. J. (2002): "A model of loudness application to time-varying sounds", J. Audio Eng. Soc, Vol. 50, n°5, 331-342.

- ETSI TS 103 334: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wearable wireless terminals from a QoS perspective as perceived by the user".

- ETSI TS 103 607: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wearable wireless wideband terminals from a QoS perspective as perceived by the user".

- Recommendation ITU-T P.340: "Ammendment 2: Annex B: Objective test methods for multi-talker scenarios".

History

Document history		
V1.1.1	July 2020	Publication