ETSI TS 103 557 V1.3.1 (2020-03)

TECHNICAL SPECIFICATION

Speech and multimedia Transmission Quality (STQ); Methods for reproducing reverberation for communication device measurements

Reference

RTS/STQ-290

Keywords

quality, reverberation, simulation, speech,
terminal, testing

ETSI

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

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

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

Important notice

Copyright Notification

Contents

Intellectual Property Rights

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).

The present document is to be used in conjunction with:

- ETSI TS 103 224 [1] series: "A sound field reproduction method for terminal testing including a background noise database".

The present document describes a sound field recording and reproduction technique which can be applied for all types of terminals but is especially suitable for modern multi-microphone terminals including array techniques. While ETSI TS 103 224 [1] focuses on background noise, the present document considers the reproduction of reverberation.

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

Many devices that employ microphones to pick up speech signals are used in a hands-free manner. Since there is usually a larger distance between the talker and the device, the microphone signals contain a significant amount of noise and reverberation.

This includes, e.g. phones in hands-free mode, group-audio terminals or smart speakers with speech recognition capabilities as well as terminals in handset or headset mode. Note that the same issues can also arise for hand-held devices depending on the acoustic conditions, see [i.1].

Testing of these devices requires a realistic reproduction of both the noise as well as the reverberation in a defined and reproducible manner. For background noise reproduction, ETSI has standardized a reproduction method (with an accompanying database of background noise signals) in ETSI TS 103 224 [1].

1 Scope

The present document describes a methodology for recording and reproducing different room characteristics and realistic reverberation under conditions that are well-defined and tailored for a calibrated setup in a lab environment. The individual aspects of the description are:

- Measurement of room impulse responses.
- Processing of test signals.
- Loudspeaker setup, calibration and equalization.

The methodology is fundamentally designed for use without access to internals of the Device Under Test (DUT), e.g. the exact positions and orientations of the device's microphones or the unprocessed microphone signals. The methodology is intended to be used for performance evaluation of all types of devices where the room characteristics may impact the performance.

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

[3] N. Xiang: "Evaluation of reverberation times using a nonlinear regression approach" in The Journal of the Acoustical Society of America Vol. 98, 1995.

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

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] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant and P. Vary: "Do We Need Dereverberation for Hand-Held Telephony?", International Congress on Acoustics (ICA), Sydney, 2010.

[i.2] Recommendation ITU-T P.341 (03/2011): "Transmission characteristics for wideband digital loudspeaking and hands-free telephony terminals".

[i.3] ISO 3382-1: "Acoustics -- Measurement of room acoustic parameters -- Part 1: Performance spaces".

[i.4] Recommendation ITU-T P.501: "Test signals for use in telephonometry".

[i.5] ETSI TS 103 738: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for narrowband wireless terminals (handsfree) from a QoS perspective as perceived by the user".

[i.6] ETSI TS 103 740: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband wireless terminals (handsfree) from a QoS perspective as perceived by the user".

3 Definition of terms, symbols and abbreviations

3.1 Terms

Void.

3.2 Symbols

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

$C_{t_{\mathrm{CO}}}$	Clarity with cut-off time t_{CO}
f_s	Sampling frequency
$g_i(k)$	Reverberant components of the impulse response to microphone i
$h_i(k)$	Impulse response to microphone i
$h_N(k)$	Impulse response to the microphone closest to the HATS
$h_{D,i}(k)$	Direct path component of the impulse response to microphone i
K	Length of the impulse response
M	Number of microphones
MM-#	Microphone number # of MM
MSA-#	Microphone number # of MSA
MU-#	Microphone number # of MU
RT_{60}	Reverberation time
$s(k)$	Source signal
t_{CO}	Cut-off time for clarity (50 ms for speech, 80 ms for music)
$\mathrm{T_{DIFF}}$	Delay difference between HATS and reverberation reproduction system

3.3 Abbreviations

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

DRR	Direct-to-Reverberant energy Ratio
DUT	Device Under Test
FFT	Fast Fourier Transform
HATS	Head And Torso Simulator
LRC	Lip Ring Centre
MLS	Maximum Length Sequence
MM	Measurement Microphone(s)
MRP	Mouth Reference Point
MSA	Microphone Sound Array
MU	Mock-Up phone

SFR	Send Frequency Response
SLR	Send Loudness Rating
SNR	Signal-to-Noise Ratio

4 Rationale

Including reverberation in a realistic manner is of paramount importance for accurate testing of, e.g. phones in hands-free mode, group-audio terminals or smart speakers with speech recognition capabilities. While it is possible to test this simply by using the device in a reverberant room, the present document introduces an alternative approach that is based on the explicit reproduction of the reverberant sound field at several microphone positions.

The present method does not require many rooms or variable acoustics setups for testing multiple acoustic conditions. The reverberant sound field can both be measured directly in a reverberant room or it can be calculated based on a database of impulse responses that is provided in combination with the present document.

NOTE 1: The room acoustics has an impact on three transmission paths that could be considered from a measurement perspective:

- from the mouth of the user to the microphone(s) of the DUT (Sending);
- from the loudspeaker(s) of the DUT to the ears of the user (Receiving);
- from the loudspeaker(s) of the DUT to the microphone(s) of the DUT (Echo-path).

NOTE 2: The methodology described in the present document is intended for the first scenario. Although it might be applicable to the other scenarios as well (possibly with modifications), this has not been verified yet and is subject to further study.

5 Room simulation in Sending

5.1 Impulse Response Measurement

5.1.1 Microphone setup

5.1.1.1 Fixed Microphone setup

A fixed microphone setup should be used for DUTs with smaller form factors (e.g. mobile phones in hands-free operation). The microphone setup shall consist of $M = 8$ microphones and conform to the description in ETSI TS 103 224 [1], clause 5. Since this setup is device-independent, no new impulse response measurements are necessary when testing a new device.

5.1.1.2 Flexible Microphone setup

For larger DUTs, a flexible microphone setup with a use-case dependent number of microphones M shall be used. The microphone setup shall conform to the description in ETSI TS 103 224 [1], clause 7. Since this setup is device-dependent, testing a new device requires new impulse response measurements.

5.1.2 Sound source

Since the testing scenario consists of a human talker in a reverberant environment, a HATS with an equalized artificial mouth according to Recommendation ITU-T P.58 [2] shall be used for sound generation. This applies to measurement of impulse responses as well as to recording of reverberant speech signals.

5.1.3 Measurement procedure

There exist different possibilities for measuring impulse responses, e.g. using Maximum Length Sequences (MLS) or using swept-sines (sweeps). The advantage of sweeps is that non-linearities can easily be observed and that the SNR in lower frequencies is higher than with MLS. Using logarithmic sweeps is therefore recommended for system identification. The sweep should cover a frequency range from 20 Hz to 20 kHz and the length of the sweep should be chosen in such a way that no significant components of the impulse response are truncated. Accordingly, a sweep length of at least 2 s should be used for typical rooms while larger rooms might need longer sweeps. While the sweep can be used directly as the measurement signal, it is recommended to construct the measurement signal from the individual sweep by repeating it at least five times. If the repetition is used, the determination of the impulse response should be based on an averaging of all but the first sweep period. The first period is discarded to avoid transient effects and the averaging should be performed in the time domain.

5.2 Loudspeaker setup for reproducing reverberation based on the fixed microphone setup

5.2.1 Introduction and System Overview

In order to correctly reproduce the sound field in a reverberant room, the setup described in ETSI TS 103 224 [1] is not sufficient. Direct sound as well as the reverberant components shall be considered. A combination of the multi-channel loudspeaker setup with the artificial mouth of a HATS is used for the reproduction.

The multi-channel loudspeaker setup shall follow the description in ETSI TS 103 224 [1]. If possible, the HATS should be positioned at the same position (distance, angle, mouth direction, etc.) for the sound field reproduction with respect to the microphone arrangement as in the original reverberant room. If this is not possible, the HATS shall be positioned at the same angle with respect to the microphone arrangement as in the original reverberant room and the change in distance shall be considered when adjusting levels and delays between the direct path and the reverberant components.

An overview of the signal processing is given in Figure 1. The two paths are visible here as well: The signal for the HATS is only subject to a level (multiplication by coefficient *a*) and delay (delay element of length T_{DIFF}) adjustment which is covered in clause 5.2.2.2. The reverberant signal components are reproduced by removing the direct path from the impulse responses (see clause 5.2.2.1) and generating the target signals for the system according to ETSI TS 103 224 [1] by convolving the source signal with the remaining parts of the impulse responses.

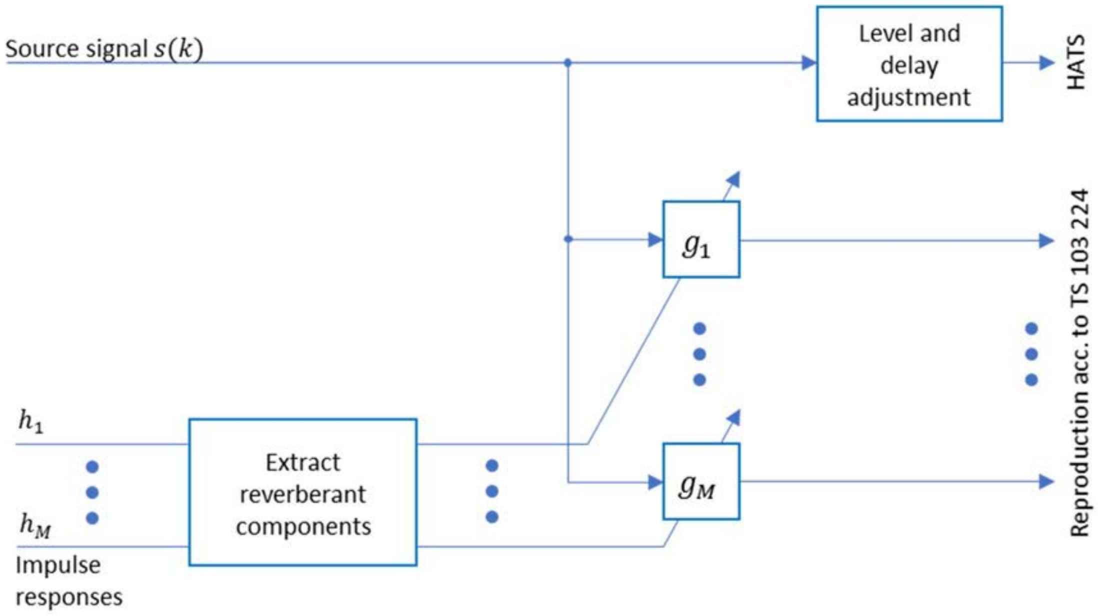
Source signal $s(k)$
Level and delay adjustment
HATS
g_1
g_M
Reproduction acc. to TS 103 224
h_1
h_M
Impulse responses
Extract reverberant components

Figure 1: Overview of the signal processing for the reproduction setup

For the fixed microphone setup, the impulse responses can be taken directly from the provided database. For the flexible microphone setup, individual measurements have to be carried out to use the reproduction system.

5.2.2 Preparations

5.2.2.1 Separation of impulse responses

The reproduction system needs two signal parts: the direct sound (single channel that is played over the HATS) and the reverberant components (eight target signals that are fed into the reproduction system according to ETSI TS 103 224 [1]). An example impulse response $h_i(k)$ is shown in Figure 2 with the direct path component $h_{D,i}(k)$ in orange and the remaining reverberant components $g_i(k)$ in blue. These two components constitute the entire impulse response according to:

$$h_i(k) = h_{D,i}(k) + g_i(k)$$

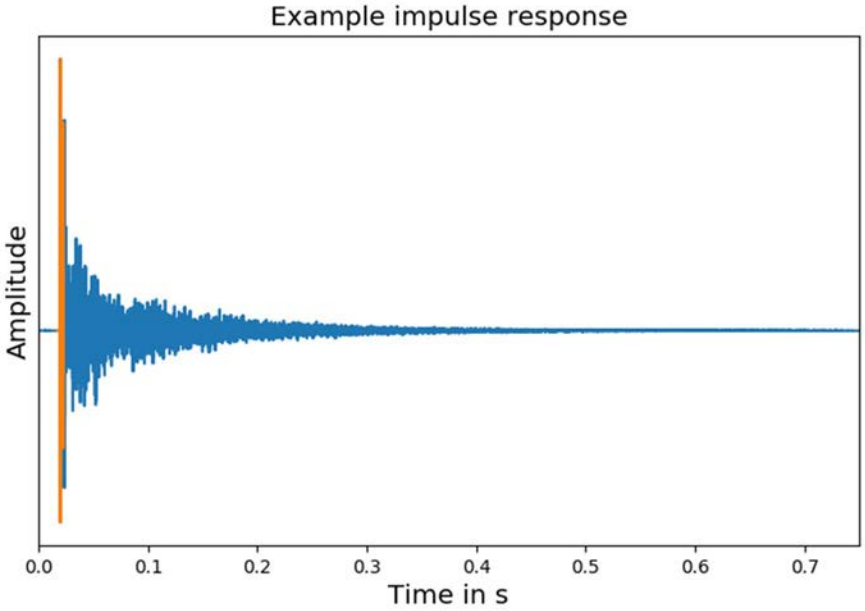

Figure 2: Time domain representation of an example impulse response

An overview of the necessary signal processing is depicted in Figure 5. From the measured impulse responses $h_1(k)$ ... $h_M(k)$, the reverberant components $g_1(k)$ ... $g_M(k)$ are extracted by removing the first few milliseconds of the impulse response, i.e. the direct path. These reverberation filters $g_1(k)$ ... $g_M(k)$ are then used to calculate the input signals for the reproduction system according to ETSI TS 103 224 [1].

The extraction of the direct path from the impulse response is done by searching for the largest absolute amplitude in the impulse response and selecting a window of ±2,5 ms around this position. To avoid signal processing artefacts, squared sinusoidal sections shall be used for fading in and out. The entire 5 ms section shall have 0,25 ms of squared sine fade-in in the beginning and 0,25 ms of squared sine fade-out in the end.

The resulting window function is depicted in Figure 3 and an enlarged view of the fade-in is presented in Figure 4. If the largest amplitude in the impulse response is closer than 2,5 ms to the beginning of the impulse response, the fade-in shall be omitted and the first part of the impulse response (from the start to 2,5 ms after the largest amplitude) shall be used for the direct path. The remainder of the impulse response contains all the reverberant components.

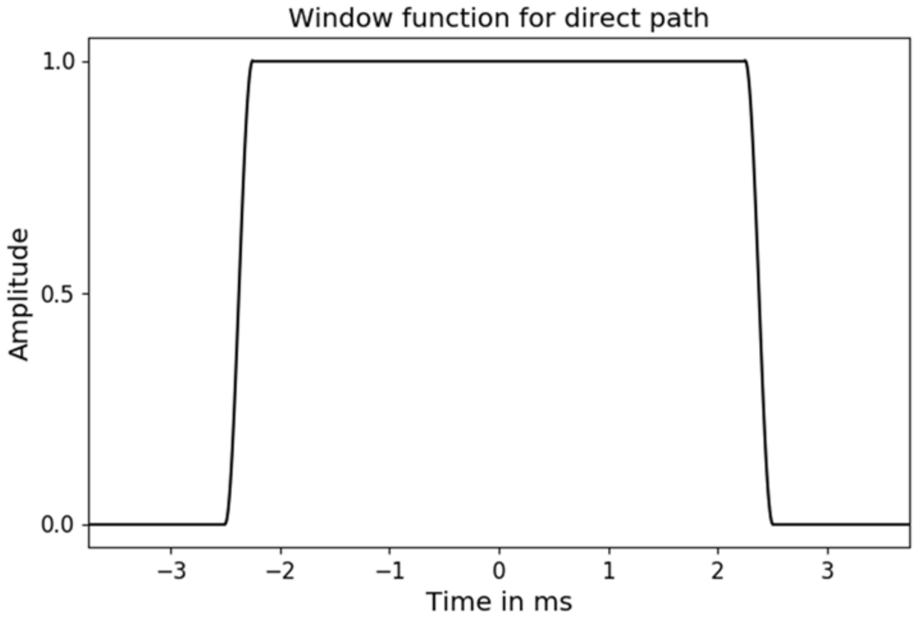

Figure 3: Window function for extracting the direct path from the impulse responses (time scale relative to maximum position of impulse response)

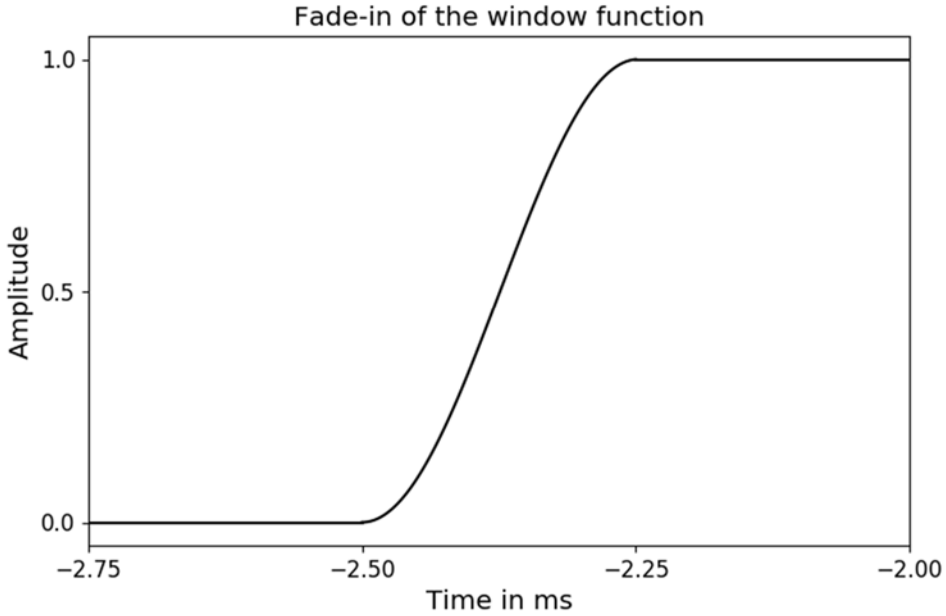

Figure 4: Fade-in of the window function

The remainder of the impulse response contains all the reverberant components.

All channels of the impulse response shall be separated according to the given procedure. As described, the direct paths are used for determining the delay and level differences between the two system components. For calculating the reproduction targets for the reproduction system according to ETSI TS 103 224 [1], however, the reverberant components are used.

5.2.2.2 Delay and level adjustment

For the aforementioned level and delay compensation, the signal from the artificial mouth needs to be adjusted to the signal from the reproduction system. This shall be achieved by comparing the level of and the delay between two signals. A linear or logarithmic sweep signal of at least 2 s covering a frequency range from ≤ 100 Hz to ≥ 4 000 Hz shall be both played by the artificial mouth and convoluted with the direct path components to get target signals for the reproduction system.

As shown in Figure 5, this leads to two recorded signals: $y_{\mathrm{HATS}}(k)$ and $y_{\mathrm{REP}}(k)$. The delay difference between the two systems shall be estimated by calculating the cross-correlation function $\Phi y_{\mathrm{HATS}} y_{\mathrm{REP}}(\lambda)$ between the two signals and searching for the peak value of its envelope E(λ). The envelope E(λ) is calculated using the Hilbert transformation $H\{\Phi y_{\mathrm{HATS}} y_{\mathrm{REP}}\}(\lambda)$ of the cross-correlation:

$$H\left\{\Phi_{y_{\mathrm{HATS}}y_{\mathrm{REP}}}\right\}(\lambda) = \sum_{u=-\infty}^{+\infty} \frac{\Phi_{y_{\mathrm{HATS}}y_{\mathrm{REP}}}(\lambda)}{\pi \cdot (\lambda - u)}$$

$$E(\lambda) = \sqrt{\Phi^2_{y_{\mathrm{HATS}}y_{\mathrm{REP}}}(\lambda) + H^2\left\{\Phi_{y_{\mathrm{HATS}}y_{\mathrm{REP}}}\right\}(\lambda)}$$

The position on the time axis of the peak value is the delay difference T_{DIFF} between the two system paths. The signal that arrives earlier (this will usually be the signal of the artificial mouth) shall be delayed by T_{DIFF} to temporally align the two paths.

The level alignment can be carried out in similar fashion. The levels of both signals shall be calculated according to Recommendation ITU-T P.56 [4] and the attenuation or amplification *a* in the path of the HATS shall be set to compensate the level difference between the two signals.

NOTE: In case the distance between HATS and DUT is identical to the one included in the impulse responses, the result of the level adjustment is expected to be 0 dB. However, it is recommended to apply the determined level adjustment anyway, in order to compensate for small differences between simulated and real setup.

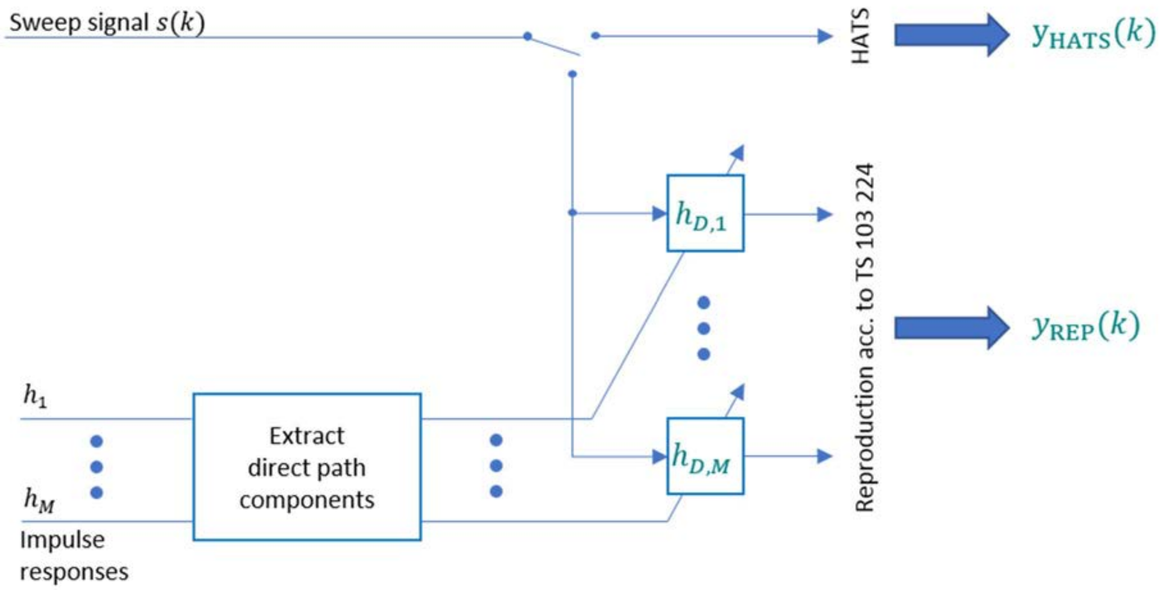

Figure 5: Overview of the signal processing for the delay compensation

5.2.3 Test room requirements

The requirements are identical to the requirements in ETSI TS 103 224 [1]. In addition, the reverberation time of the measurement chamber that is used for the reproduction imposes a lower bound on the range of rooms that can be reproduced. Thus, an anechoic room is recommended. If semi-anechoic rooms or measurement chambers are used, the effects of additional reflections of the reproduction room have to be considered.

5.2.4 Equalization and calibration

The calibration of the microphone array and the equalization of the loudspeaker system are identical to ETSI TS 103 224 [1]. The artificial mouth of the HATS shall conform to Recommendation ITU-T P.58 [2], no additional equalization is needed.

5.2.5 Accuracy of the reproduction arrangement

5.2.5.1 Evaluation parameters

The accuracy of the reproduction shall be verified by measuring the following four parameters and asserting that they are within the given tolerance range around the true value for each dataset.

The first three parameters are based on the impulse response $h(k)$ of the entire reproduction system. The impulse response of the reproduction system should be determined in the same manner as described in clause 5.1.3. Note that the measurement signal (usually a sweep) has to be fed into the entire reproduction system and not only emitted by the artificial mouth, i.e. the measurement signal is the source signal $s(k)$ in Figure 1. The impulse response and all the derived parameters shall be determined for all eight microphones in the fixed microphone setup and for all microphones in the flexible microphone setup, respectively.

5.2.5.2 Reverberation time

Probably the most widely used parameter to characterize the acoustic behaviour of a room is its reverberation time RT_{60}, the required time for a level decay of 60 dB. Several methods for estimating $RT_{60}$are available, but even the most common standard ISO 3382-1 [i.3] only provides guidelines instead of a robust and clear calculation. Thus, the reverberation time shall be determined from the impulse responses by the method described in [3], which is an extension of clause 6.2 of ISO 3382-1 [i.3] and is not prone to overestimations.

An overall reverberation time is then determined by averaging the individual reverberation times for the microphones in the considered setup. The tolerance for an accurate reproduction arrangement shall be ±50 ms compared to the real scenario.

5.2.5.3 Clarity

The clarity is a measure that is applied in room acoustics to quantify the usability of a room for a certain class of signals. It is a logarithmic ratio between the energy in the early part of the room impulse response (direct sound and early reflections) and the energy in the late reverberation up to the length K of the impulse response:

$$C_{t_{\mathrm{CO}}} = 10 \cdot \log_{10}\left(\frac{\sum_{0}^{f_s \cdot t_{CO}} h^2(k)}{\sum_{f_s \cdot t_{CO}}^{K} h^2(k)}\right)$$

For speech, the cut-off time $\mathrm{t_{CO}}$ is usually chosen to be 50 ms while for music, the value is 80 ms. Since the expected use cases will be mostly speech-focused, $\mathrm{C_{50}}$ is reported in the present document.

An overall clarity is then determined by averaging the individual clarities for the microphones in the considered setup. The tolerance for an accurate reproduction arrangement shall be ±0,5 dB compared to the real scenario.

5.2.5.4 Direct-to-Reverberant energy Ratio

The DRR is conceptually fairly similar to the aforementioned clarity; it is a ratio between the energy of the direct sound (without the early reflections) and the energy of all reverberant components. This corresponds to a value of 5 ms for the cut-off time.

An overall ratio is then determined by averaging the individual ratios for the microphones in the considered setup. The tolerance for an accurate reproduction arrangement shall be ±1,5 dB compared to the real scenario.

5.2.5.5 Coherence

In contrast to the other three parameters (which can all be calculated individually for each channel), the coherence describes the relation between original source (S) and reproduced (R) sound pressure at a given microphone channel i. It is usually given in the form of the magnitude-squared coherence, which is calculated as the ratio between the squared magnitude of the cross-power spectral density and the product of the two auto-power spectral densities:

$$C_{R_iS_i}(f) = \frac{\left|P_{R_iS_i}(f)\right|^2}{P_{R_iR_i}(f) \cdot P_{S_iS_i}(f)} \text{ ,} \forall\, i \in [1..M]$$

This parameter quantifies the performance limit of the reproduction system over the entire frequency range.

This analysis versus frequency produces numerous result curves and an aggregation to a meaningful single value is not obvious. Thus, a requirement on coherence for an accurate reproduction arrangement as well as the analysis parameters (FFT size, overlap, etc.) are for further study.

5.3 Loudspeaker setup for reproducing reverberation based on the flexible microphone setup

The constraints to the number and frequency range of the loudspeakers that are given in ETSI TS 103 224 [1] apply in the present document as well.

5.4 Impulse response database and signal generation

An impulse response database is provided in Annex A. The corresponding files are available separately from the present document.

5.5 Validation and examples

5.5.1 Reproduction of room acoustical parameters

Room impulse responses were recorded in six different rooms according to Table 1.

Table 1: Investigated rooms

Label	Room	Approximate volume
Room 1	Meeting room with additional damping panels	120 m^3
Room 2	Listening laboratory	107 m^3
Room 3	Office room	54 m^3
Room 4	Meeting room	120 m^3
Room 5	Small workshop	172 m^3
Room 6	Large workshop	1 000 m^3

In all measurements, the microphone setup consisted of 14 microphones as depicted in Figure 6. Eight of the microphones are needed for recording the sound field as described in ETSI TS 103 224 [1] (labelled MSA-1 to MSA-8). Four microphones are situated in a phone mock-up which is used as the device under test for this investigation (MU-1 to MU-4) and the final two microphones are additional measurement microphones that were installed at the far left and far right of the entire group of microphones (MM-1 and MM-2).

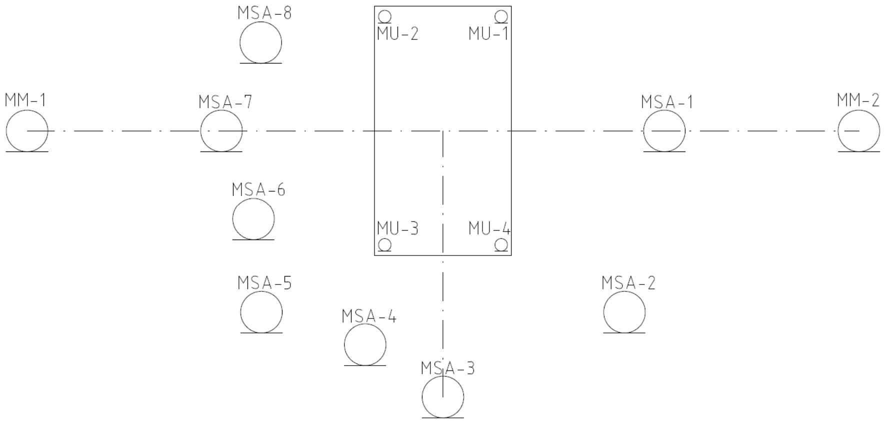

Figure 6: Microphone arrangement used for the measurements

The recordings in the different rooms were done both in the measurement configuration for a group-audio terminal according to Recommendation ITU-T P.341 [i.2] (denoted *Std.* in the figures) and a room-specific alternative position (denoted *Alt.* in the figures), which was further away from the DUT (see Figure 7 and Figure 8 for the two setups in the meeting room).

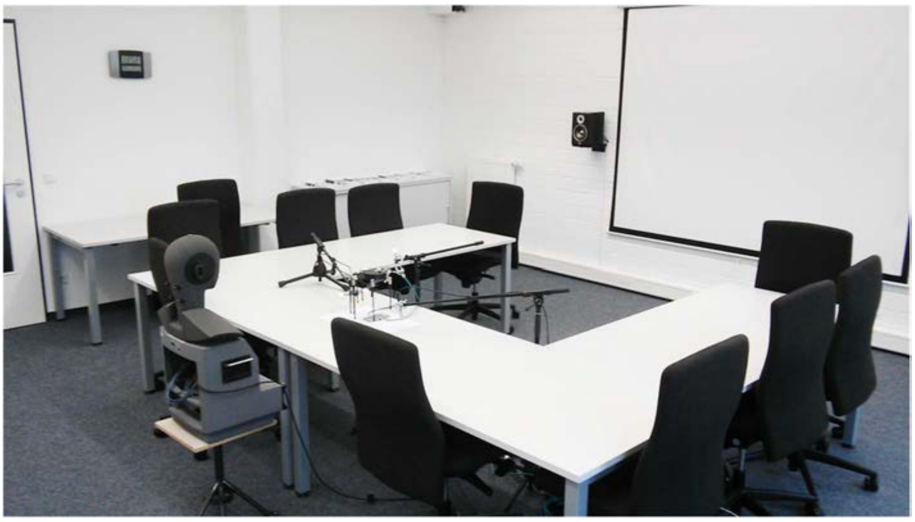

Figure 7: Measurement setup according to [i.2] in the meeting room

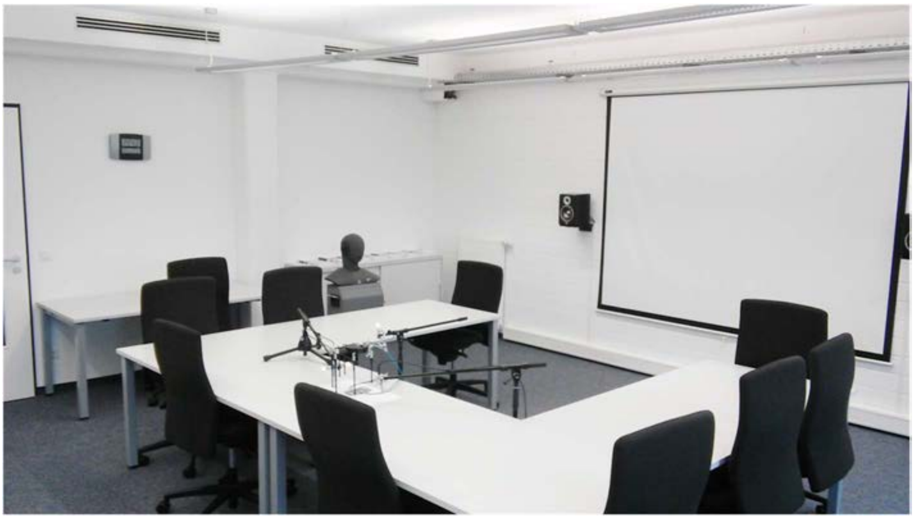

Figure 8: Alternative measurement position in the meeting room

The different room acoustic parameters were measured both in the real rooms and in the reproduced reverberant environment, a comparative overview between the results is given in this clause. The results are always averaged over all microphones unless otherwise noted.

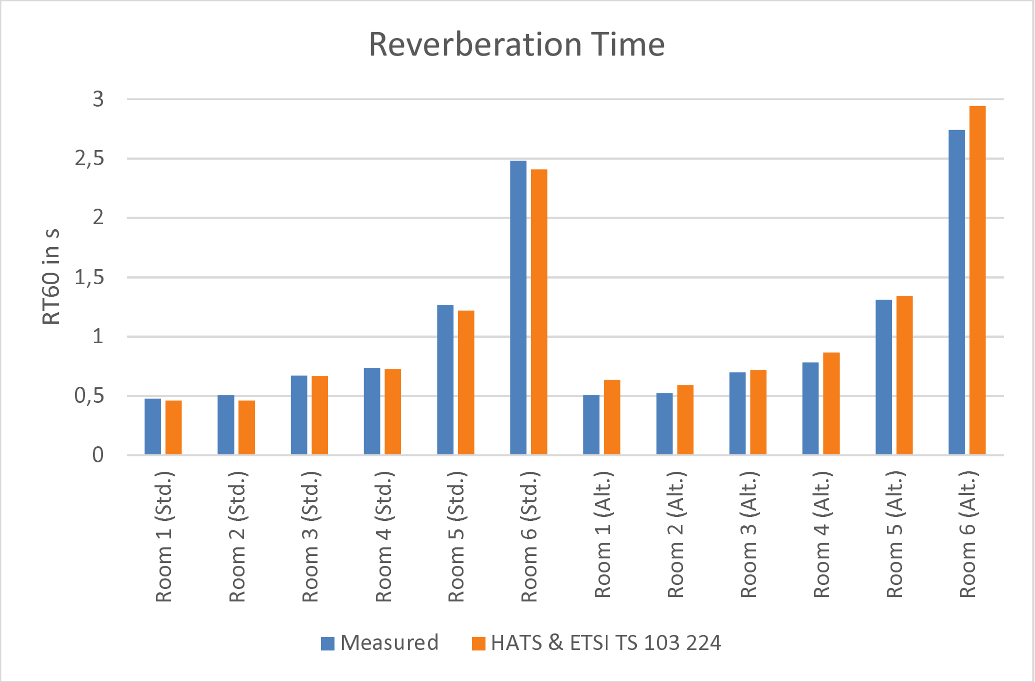

Figure 9: Comparison between measured and reproduced reverberation times

The estimated reverberation times are very similar. While there is a very slight decrease for the six measurements in the group-audio terminal position, the changes are not practically relevant - the largest difference is a change of 3,6 % for room 1. For the alternative positions, the estimated reverberation times are practically identical.

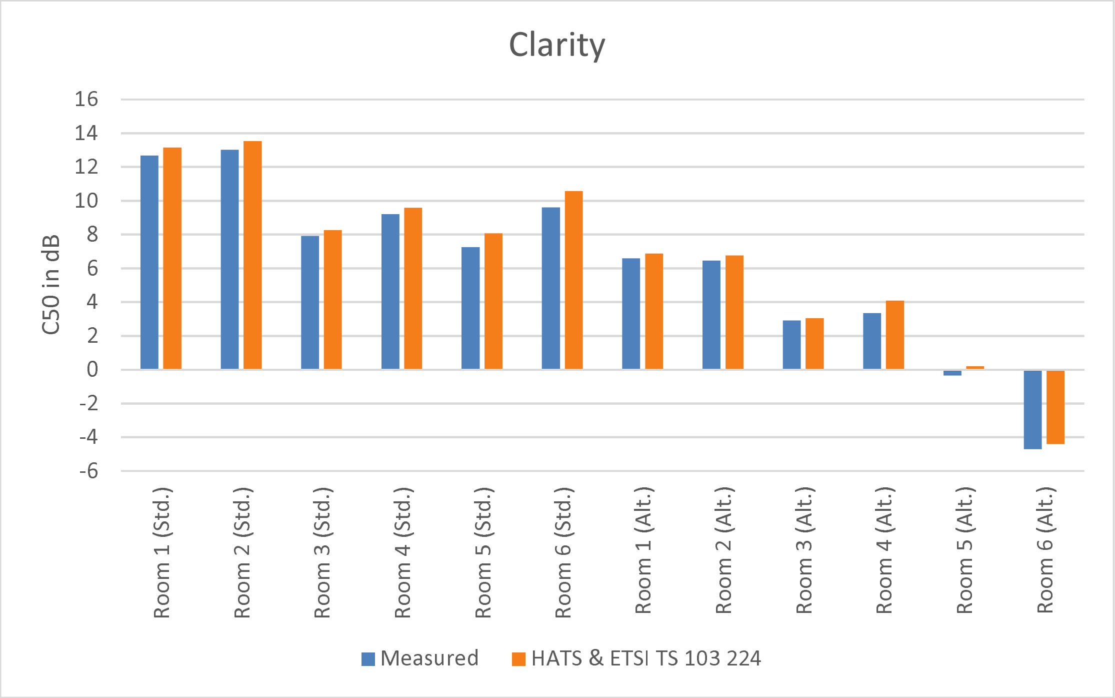

Figure 10: Comparison between measured and reproduced clarities

The same observation can be made for the clarity; there are no large deviations between the results for the measured and the reproduced sound field.

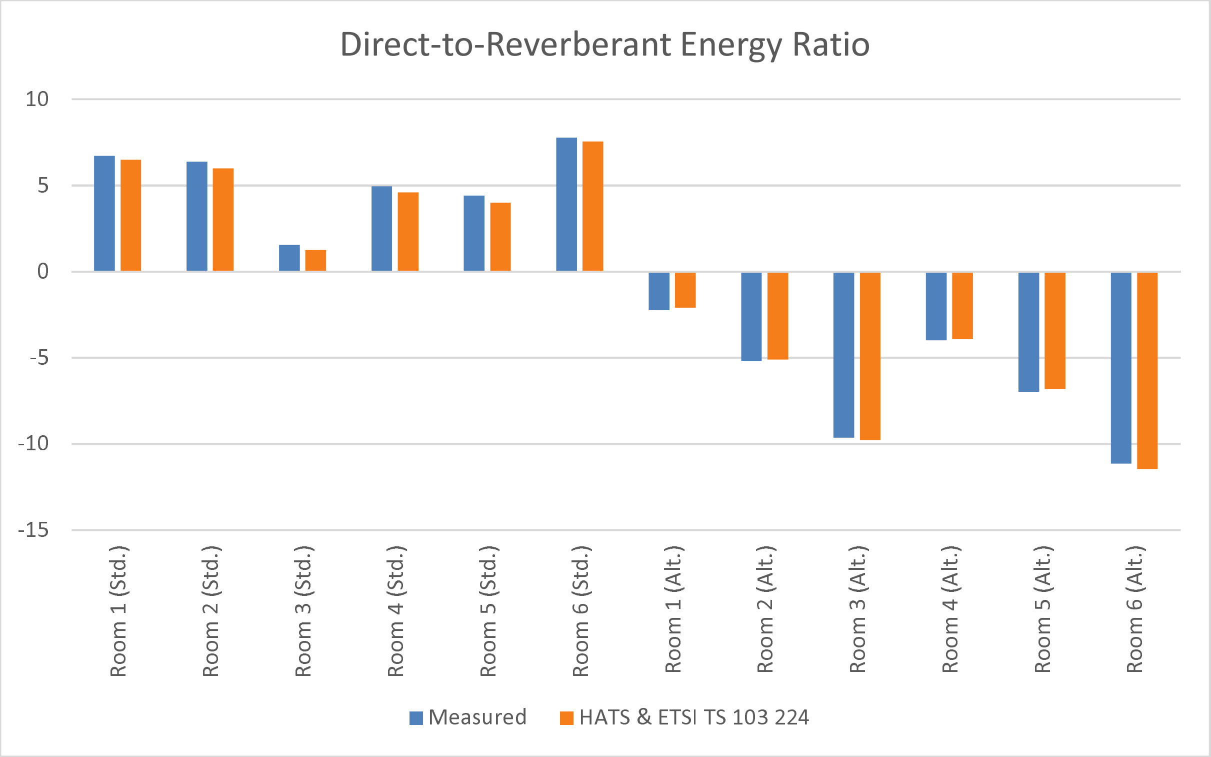

Figure 11: Comparison between measured and reproduced direct-to-reverberant energy ratios

The DRR results are again very similar between measurement and reproduction; there are no systematic and relevant changes.

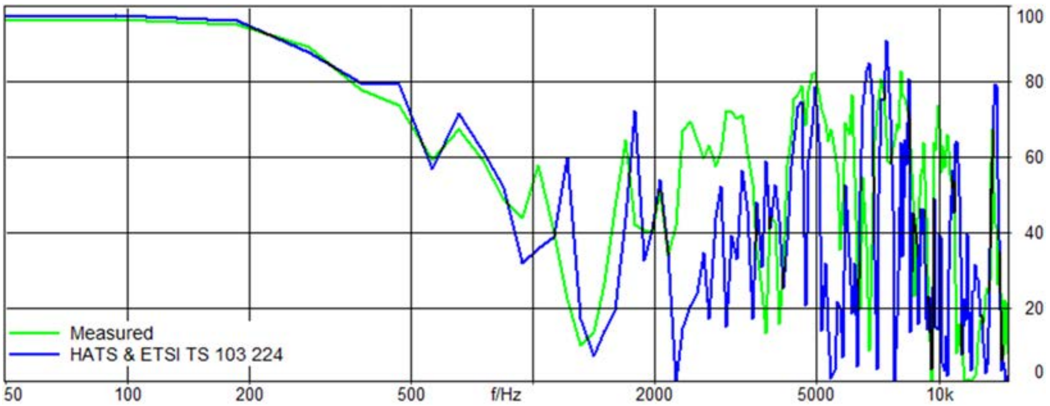

Figure 12: Comparison between measured and reproduced magnitude squared coherences

The magnitude-squared coherence between the channels MU-1 and MU-3 is depicted in Figure 5 for room 4 (the meeting room depicted in Figure 7 and Figure 8) in the group-audio terminal position. The coherences in the recording and the reproduction are similar up a frequency of approximately 2,5 kHz. Above this frequency, differences are to be expected as the distances between the microphones in the recording setup impose constraints on the sound field reproduction (see clause 5.1.1 in ETSI TS 103 224 [1]).

5.5.2 Application examples

As an example, for the use and the performance of the reproduction system, different hands-free telephones were positioned on a table in room 1 from the impulse response database at the group-audio terminal position. Speech utterances from a HATS were recorded. The speech signals used were a sequence of six male and six female speakers taken from Recommendation ITU-T P.501 [i.4]. The three devices have different form factors and spatial characteristics. Device 1 is designed for wideband telephony, fits easily into the space that is enclosed by the microphone array and uses a single directional microphone. Device 2 is not strictly bandlimited but appears to be optimized for narrowband telephony, has approximately the size of the microphone array and an omnidirectional microphone system. Device 3 is also focusing on narrowband telephony, is significantly larger than the microphone array and has a strongly directional receiving pattern.

The situation was reproduced in a measurement chamber where the same speech utterances were played over the delay and level compensated combination of an artificial mouth for the direct path and a system conforming to ETSI TS 103 224 [1] for the reverberant components.

Two quantities are compared in the analysis: the Send Loudness Rating (SLR) and the Send Frequency Response (SFR) according to ETSI TS 103 738 [i.5] or ETSI TS 103 740 [i.6], depending on the device.

For device 1, the SFR is depicted in Figure 13 using 1/12th octave bands. It can be seen that the shape of the curves is similar and the reproduction manages to achieve a device behaviour that is approximately equivalent to the behaviour in the real reverberant environment. The measured SLR values differ by 0,5 dB.

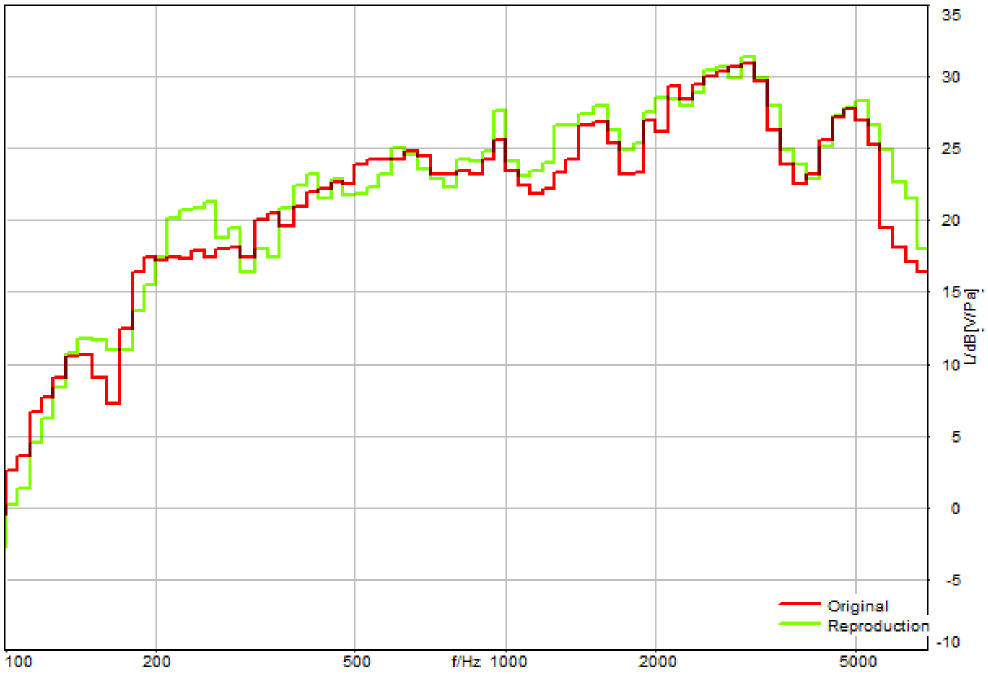

Figure 13: Send frequency responses of device 1 in the original room and the reproduced room

The SFR for device 2 is given in Figure 14. The two curves are very similar, there are some differences in single frequency bands but the amplitude of these difference is quite small - in particular for the narrowband frequency range which is the apparent design target for this device given the significant drop-off towards the lower frequencies. The difference between the SLR values for this device is 0,45 dB.

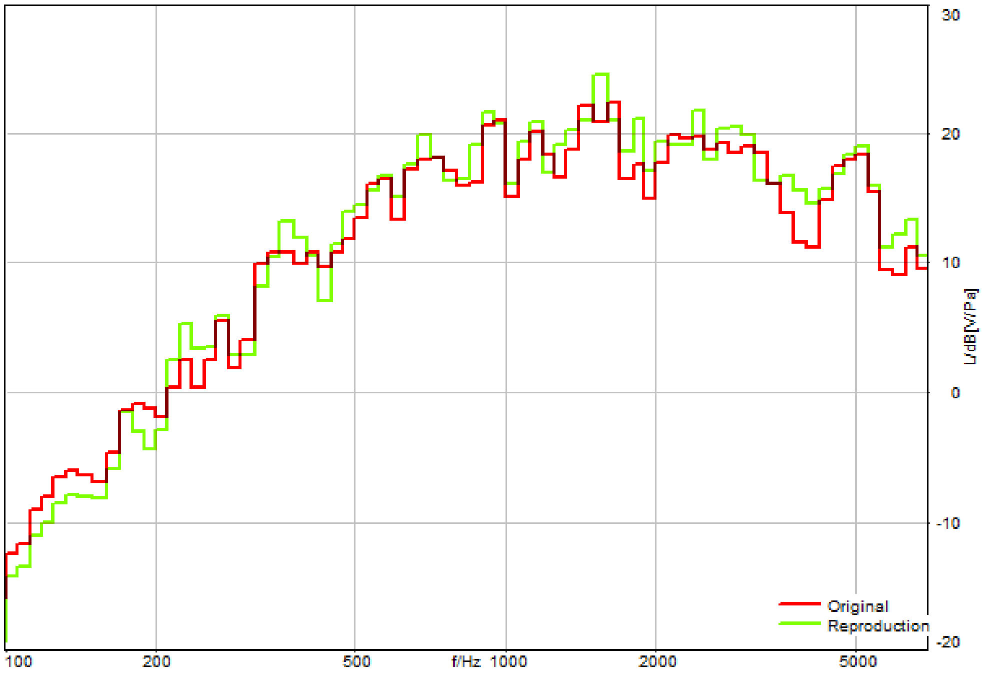

Figure 14: Send frequency responses of device 2 in the original room and the reproduced room

Finally, Figure 15 shows the SFR for device 3. This device is significantly wider than the dimensions of the microphone array. Thus, it is not surprising that the deviations between the curves are larger than for the other two devices. The SLR difference is also the largest of the three with 0,74 dB. The reproduction for this device would benefit from using the flexible version.

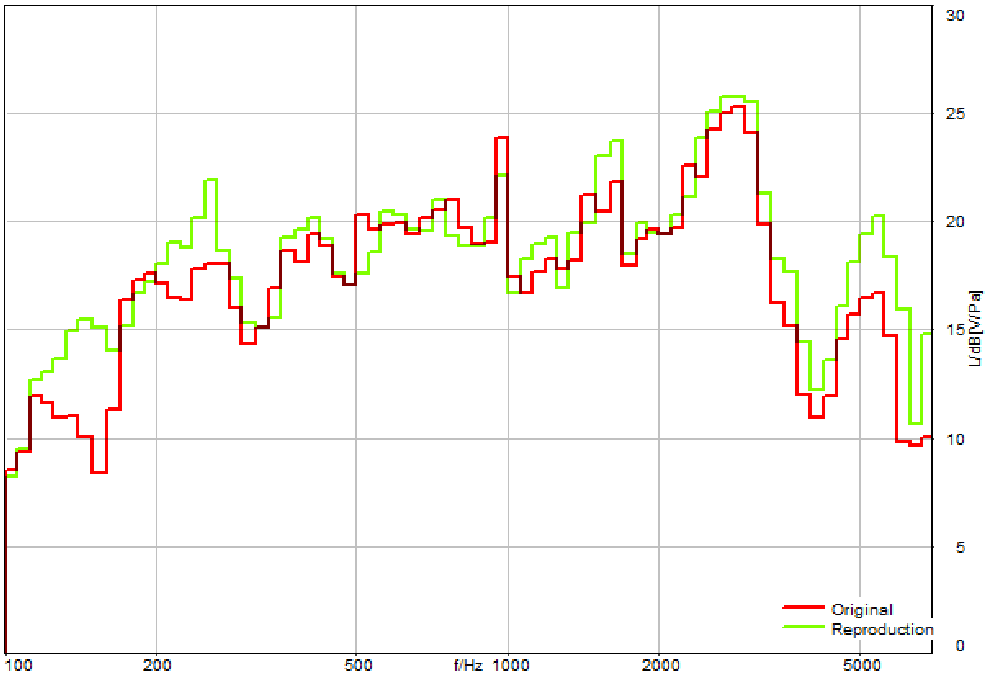

Figure 15: Send frequency responses of device 3 in the original room and the reproduced room

Annex A (normative): Impulse Response Database

A.1 Fixed microphone setup

The following multi-channel impulse responses were obtained using the fixed eight-microphone setup according to ETSI TS 103 224 [1], providing examples for the simulation of typical, reverberant rooms. In addition, details about positioning and distances of the array are specified per scenario, which are necessary for the physical test setup. If not specified otherwise, microphone 5 of the array is the closest one to the HATS and is aligned to the vertical plane according to Recommendation ITU-T P.58 [2].

Table A.2 shows the validation metrics according to clause 5.2.5, which shall be achieved by the reproduction arrangement (tolerances are specified in clause 5.2.5 as well).

The multi-channel impulse responses can be downloaded here:
https://docbox.etsi.org/stq/Open/TS 103 557/Annex_A_ImpulseResponseDatabase/Fixed/.

Table A.1: Impulse responses for fixed microphone setup

ID	Filename	Distances			Description
		LRC to Pos. 5	LRC to centre	Lip plane [2] to centre	
1	Room1_RIR.wav	73 cm	85,5 cm	80 cm	Office room; group audio terminal setup including reflecting table, according to Recommendation ITU-T P.341 [i.2], Figure 6. Microphone array is located 2,5 cm above table surface (30 cm between LRC and table surface).
2	Livingroom	73 cm	85,5 cm	80 cm	L-shaped room with couch, coffee table and dining area; one large window front; brick walls; wooden ceiling; tiled floor.
3	Officeroom	73 cm	85,5 cm	80 cm	officeroom with three desks; walls and ceiling made of concrete, plasterboard and bricks; three windows; carpet.
4	Kitchen	73 cm	85,5 cm	80 cm	fully equipped kitchen; cuboid shape; one window; brick walls; open on one side to a living and dining area (approximately 70 m²); tiled floor; wooden ceiling.
5	Bathroom	73 cm	85,5 cm	80 cm	bathroom with shower, toilet, sideboard and sink; square base area; ceiling slope on one side; one window; tiled walls; wooden floor and ceiling.
6	medium Hall	73 cm	85,5 cm	80 cm	hall with three desks; rectangular main room with adjoining corridor; walls and ceiling mainly made of concrete; two brick walls; one big window; carpet.
7	big Hall	73 cm	85,5 cm	80 cm	large almost empty hall; rectangular main room with adjoining corridor; walls, floor and ceiling made of concrete; small window area.
8	Staircase	73 cm	85,5 cm	80 cm	rectangular open staircase with four floors; concrete walls; glass doors; window front on one side; tiled floors.

Table A.2: Parameters for validation of reproduction

ID	RT60 [ms]	Clarity [dB]	DRR [dB]
1	680	8,75	2,5
2	388	13,4	4,3
3	544	10,7	2,0
4	547	7,8	0,0
5	583	9,3	2,9
6	1 228	8,5	5,3
7	1 847	12,3	10,4
8	2 277	4,6	0,5

A.2 Impulse responses in a home-like test environment

The following multi-channel impulse responses were obtained using the fixed eight-microphone setup according to ETSI TS 103 224 [1]. All impulse responses were captured in a home-like test environment described in Annex B. The acoustic conditions, microphone array (DUT) positions, and target talker locations are presented in Annex B. Details about positioning and distances of the array are specified per scenario, which are necessary for the physical test setup. Azimuth offsets define the HATS orientation with respect to the microphone array in reference to the vertical plane according to Recommendation ITU-T P.58 [2]. Microphone alignment describes which microphone in the array is pointing towards the center of the HATS.

Table A.4 shows the validation metrics according to clause 5.2.5, which shall be achieved by the reproduction arrangement (tolerances are specified in clause 5.2.5).

The multi-channel room impulse responses can be downloaded here:
https://docbox.etsi.org/stq/Open/TS 103 557/Annex_A_ImpulseResponseDatabase/HomeLike/

Table A.3: Impulse responses for voice assistant device testing

ID	Filename	MRP [2] to centre	Description
VA1	AC1_Talker1_DUT1.wav	5,1 m	Acoustic Condition 1 (low reverberation); Talker Position 1 (sink); DUT Position 1 (entertainment center); LRC 72 cm above mic 5; Azimuth offset of 185º; Aligned between mics 5 and 6.
VA2	AC1_Talker2_DUT1.wav	1,76 m	Acoustic Condition 1 (low reverberation); Talker Position 2 (corner); DUT Position 1 (entertainment center); LRC 72 cm above mic 5; Azimuth offset of 315º; Aligned to mic 8.
VA3	AC1_Talker3_DUT1.wav	2,5 m	Acoustic Condition 1 (low reverberation); Talker Position 3 (couch); DUT Position 1 (entertainment center); LRC 22 cm above mic 5; Azimuth offset of 0º; Aligned to mic 5.
VA4	AC1_Talker1_DUT2.wav	1,41 m	Acoustic Condition 1 (low reverberation); Talker Position 1 (sink); DUT Position 2 (counter); LRC 52 cm above mic 5; Azimuth offset of 115º; Aligned to mic 6.
VA5	AC1_Talker2_DUT2.wav	4,42 m	Acoustic Condition 1 (low reverberation); Talker Position 2 (corner); DUT Position 2 (counter); LRC 52 cm above mic 5; Azimuth offset of 60º; Aligned between mics 2 and 3.
VA6	AC1_Talker3_DUT2.wav	2,93 m	Acoustic Condition 1 (low reverberation); Talker Position 3 (couch); DUT Position 2 (counter); LRC 2 cm above mic 5; Azimuth offset of 210º; Aligned between mics 3 and 4.
VA7	AC1_Talker1_DUT3.wav	1,81 m	Acoustic Condition 1 (low reverberation); Talker Position 1 (sink); DUT Position 3 (table); LRC 71 cm above mic 5; Azimuth offset of 200º; Aligned to mic 2.
VA8	AC1_Talker2_DUT3.wav	3,83 m	Acoustic Condition 1 (low reverberation); Talker Position 2 (corner); DUT Position 3 (table); LRC 71 cm above mic 5; Azimuth offset of 30º; Aligned to mic 6.
VA9	AC1_Talker3_DUT3.wav	1,28 m	Acoustic Condition 1 (low reverberation); Talker Position 3 (couch); DUT Position 3 (table); LRC 21 cm above mic 5; Azimuth offset of 180º; Aligned to mic 7.
VA10	AC2_Talker1_DUT1.wav	5,1 m	Acoustic Condition 2 (medium reverberation); Talker Position 1 (sink); DUT Position 1 (entertainment center); LRC 72 cm above mic 5; Azimuth offset of 185º; Aligned between mics 5 and 6.
VA11	AC2_Talker2_DUT1.wav	1,76 m	Acoustic Condition 2 (medium reverberation); Talker Position 2 (corner); DUT Position 1 (entertainment center); LRC 72 cm above mic 5; Azimuth offset of 315º; Aligned to mic 8.
VA12	AC2_Talker3_DUT1.wav	2,5 m	Acoustic Condition 2 (medium reverberation); Talker Position 3 (couch); DUT Position 1 (entertainment center); LRC 22 cm above mic 5; Azimuth offset of 0º; Aligned to mic 5.
VA13	AC2_Talker1_DUT2.wav	1,41 m	Acoustic Condition 2 (medium reverberation); Talker Position 1 (sink); DUT Position 2 (counter); LRC 52 cm above mic 5; Azimuth offset of 115º; Aligned to mic 6.
VA14	AC2_Talker2_DUT2.wav	4,42 m	Acoustic Condition 2 (medium reverberation); Talker Position 2 (corner); DUT Position 2 (counter); LRC 52 cm above mic 5; Azimuth offset of 60º; Aligned between mics 2 and 3.
VA15	AC2_Talker3_DUT2.wav	2,93 m	Acoustic Condition 2 (medium reverberation); Talker Position 3 (couch); DUT Position 2 (counter); LRC 2 cm above mic 5; Azimuth offset of 210º; Aligned between mics 3 and 4.
VA16	AC2_Talker1_DUT3.wav	1,81 m	Acoustic Condition 2 (medium reverberation); Talker Position 1 (sink); DUT Position 3 (table); LRC 71 cm above mic 5; Azimuth offset of 200º; Aligned to mic 2.
VA17	AC2_Talker2_DUT3.wav	3,83 m	Acoustic Condition 2 (medium reverberation); Talker Position 2 (corner); DUT Position 3 (table); LRC 71 cm above mic 5; Azimuth offset of 30º; Aligned to mic 6.
VA18	AC2_Talker3_DUT3.wav	1,28 m	Acoustic Condition 2 (medium reverberation); Talker Position 3 (couch); DUT Position 3 (table); LRC 21 cm above mic 5; Azimuth offset of 180º; Aligned to mic 7.

Table A.4: Parameters for validation of reproduction

ID	RT60 [ms]	Clarity [dB]	DRR [dB]
VA1	348	9,7	-14,9
VA2	326	13,3	1,4
VA3	332	12,9	0,4
VA4	353	12,4	-3,8
VA5	370	8,3	-18,3
VA6	371	5,9	-14,2
VA7	340	9,9	-8,0
VA8	349	11,0	-4,1
VA9	352	8,4	-4,9
VA10	476	5,3	-17,2
VA11	483	8,9	0,0
VA12	480	8,7	-2,7
VA13	477	8,4	-4,9
VA14	495	4,3	-19,1
VA15	490	3,1	-16,0
VA16	481	6,0	-9,0
VA17	488	6,6	-5,5
VA18	485	4,9	-5,0

Annex B (informative): Home-like test environment

This annex provides information on the layout of a home-like test environment used for the capture of room impulse responses presented in clause A.2. Impulse responses were captured in two acoustic conditions. The reverberation time and clarity index requirements and tolerances for the two acoustic conditions are presented in Table B.1.

Table B.1: Acoustic conditions

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 (low)	0,4	0,3	0,3	0,3	0,3	0,3	10,5
Condition 2 (medium)	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

Figures B.1 and B.2 show the home-like test environment layouts which satisfy acoustic conditions 1 and 2, respectively.

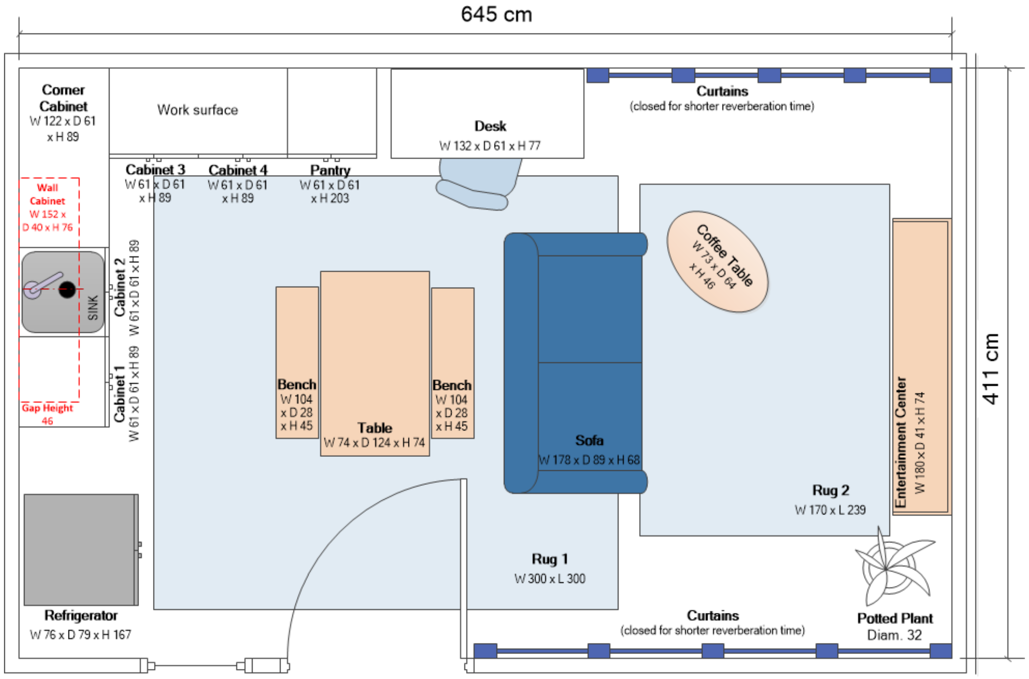

Figure B.1: Acoustic condition 1 home-like test environment layout (dimensions are in cm)

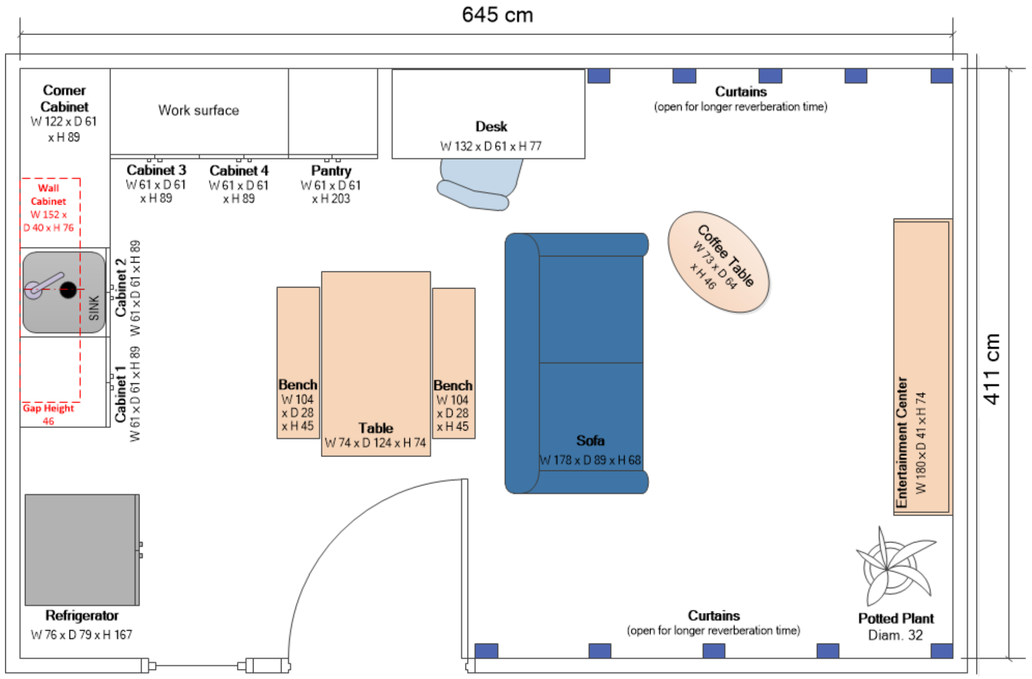

Figure B.2: Acoustic condition 2 home-like test environment layout (dimensions are in cm)

Figure B.3 presents the three microphone array (DUT) positions along with the three target talker locations. The azimuth offsets and microphone alignments are indicated as well.

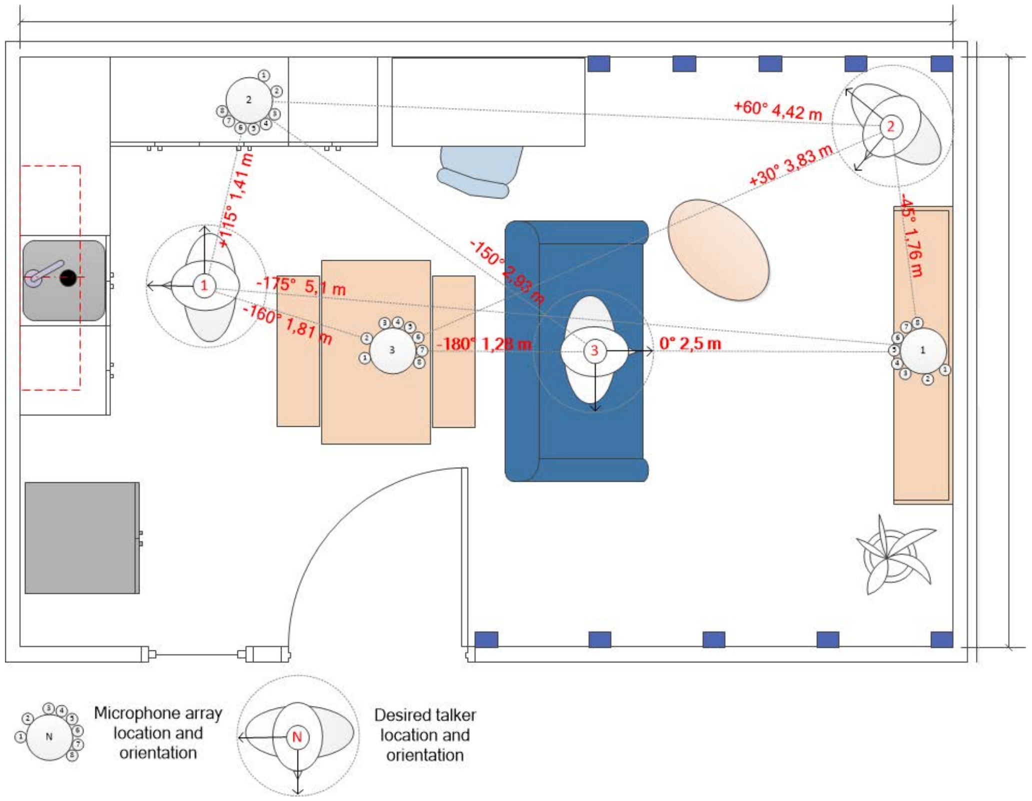

Figure B.3: Desired talker and microphone array (DUT) positions and orientations

History

Document history		
V1.1.1	December 2018	Publication
V1.2.1	August 2019	Publication
V1.3.1	March 2020	Publication