# ETSI TS 126 260 V18.1.0 (2024-07)


**5G;**
**Objective test methodologies**
**for the evaluation of immersive audio systems**
**(3GPP TS 26.260 version 18.1.0 Release 18)**

Reference
RTS/TSGS-0426260vi10

Keywords
5G

***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 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

# Intellectual Property Rights

## Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are 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 Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, 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.

**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.

# Legal Notice

This Technical Specification (TS) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities. These shall be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between 3GPP and ETSI identities can be found under https://webapp.etsi.org/key/queryform.asp.

# 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.

# Contents

Intellectual Property Rights ..... 2
Legal Notice ..... 2
Modal verbs terminology ..... 2
Foreword ..... 5
Introduction ..... 5
1 Scope ..... 6
2 References ..... 6
3 Definitions, symbols and abbreviations ..... 7
3.1 Definitions ..... 7
3.2 Symbols ..... 8
3.3 Abbreviations ..... 9
4 Objective Test Methodologies for Immersive Audio Systems ..... 9
4.0 General ..... 9
4.0.1 Applicability ..... 9
4.0.2 Test equipment ..... 10
4.0.3 Test environment acoustic properties ..... 10
4.1 Objective Test Methodologies for Assessment of Immersive Audio Systems in the Sending Direction ..... 11
4.1.1 Diffuse-field Send Frequency Response for Scene-based Audio ..... 11
4.1.1.1 Introduction ..... 11
4.1.1.2 Definition ..... 11
4.1.1.3 Test method with periphonic array ..... 12
4.1.1.3.1 Test Conditions ..... 12
4.1.1.3.2 Measurement ..... 12
4.1.1.4 Test method with loudspeaker array and turn table ..... 13
4.1.1.4.1 Test Conditions ..... 13
4.1.1.4.2 Measurement ..... 14
4.1.2 Directional response measurement for scene-based audio ..... 15
4.1.2.1 Definition ..... 15
4.1.2.2 Test conditions ..... 15
4.1.2.3 Measurement ..... 15
4.2 Objective Test Methodologies for Assessment of Immersive Audio Systems in the Receiving Direction ..... 15
4.2.1 Headset Binaural Diffuse-field Receive frequency response for Scene-based audio ..... 15
4.2.1.1 Introduction ..... 15
4.2.1.2 Definition ..... 15
4.2.1.3 Test Conditions ..... 15
4.2.1.4 Measurement ..... 16
4.2.2 Nominal System Sensitivity in Receive Direction for Channel-based audio ..... 16
4.2.2.1 Introduction ..... 16
4.2.2.2 Definition ..... 16
4.2.2.3 Test Conditions ..... 16
4.2.2.4 Measurement ..... 17
4.2.3 Motion to Sound Latency in Dynamic Binaural Rendering Systems ..... 17
4.2.3.1 Introduction ..... 17
4.2.3.2 Requirements ..... 17
4.2.3.3 Calibration ..... 18
4.2.3.4 Evaluation Environment ..... 19
4.2.3.5 Data acquisition ..... 19
4.2.3.6 Data Analysis ..... 19
5 Objective Test Methodologies for IVAS-based UEs ..... 21
5.1 Overview ..... 21
5.2 Interface Definitions ..... 21
5.3 Test conditions ..... 22

5.3.1 Test environment acoustic properties ..... 22
5.3.2 System simulator and reference client ..... 22
5.3.3 Test equipment ..... 23
5.4 Test arrangement ..... 23
5.4.1 Capture modes ..... 23
5.4.2 UE types and positioning ..... 24
5.4.2.1 Overview ..... 24
5.4.2.2 Handset UE ..... 25
5.4.2.3 Headset UE ..... 25
5.4.2.4 Handheld hands-free UE ..... 27
5.4.2.5 Table-mounted UE ..... 29
5.4.2.6 Loudspeaker UE ..... 30
5.4.2.7 Electrical interface UE ..... 31
5.4.3 UE configuration ..... 32
5.5 Test signals ..... 32
5.5.1 Test signal calibration ..... 32
5.5.2 Virtual positioning ..... 33
5.6 Test methods for sending direction ..... 34
5.6.1 Delay ..... 34
5.6.2 Loudness ..... 35
5.6.3 Frequency response (single source) ..... 35
5.6.3.1 Test method ..... 35
5.6.3.2 IVAS format-specific definitions ..... 35
5.6.4 Directional information (single source) ..... 36
5.6.4.1 Test method ..... 36
5.6.4.2 IVAS format specific definitions ..... 37
5.7 Test methods for receiving direction ..... 39
5.7.1 Delay ..... 39
5.7.2 Loudness ..... 40
5.7.2.1 Test method ..... 40
5.7.2.2 IVAS format specific definitions ..... 40
5.7.3 Frequency response (single source) ..... 40
5.7.4 Interaural differences for binaural rendering ..... 41

**Annex A (normative): Order dependent directions ..... 42**

**Annex B (normative): Directions in Gaussian spherical grid ..... 47**

B.1 Definition ..... 47

B.2 Example loudspeaker array ..... 47

**Annex C (normative): Cross-correlation analysis ..... 48**

**Annex D (informative): Change history ..... 49**

History ..... 50

# Foreword

This Technical Specification has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

- x the first digit:
  - 1 presented to TSG for information;
  - 2 presented to TSG for approval;
  - 3 or greater indicates TSG approved document under change control.
- y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.
- z the third digit is incremented when editorial only changes have been incorporated in the document.

# Introduction

Audio is a key component of an immersive multimedia experience and 3GPP systems are expected to deliver immersive audio with a high Quality of Experience. However, industry agreed methods to assess the Quality of Experience for immersive audio are relatively few and the present document seeks to address this gap by providing objective test methods for the assessment of immersive audio.

# 1 Scope

The present document specifies objective test methodologies for 3GPP immersive audio systems including channel based, object based, scene-based, parametric and hybrids of these formats. The objective evaluation methods described in the present document are applicable to audio capture, coding, transmission and rendering as indicated in their corresponding clauses. They also include testing of IVAS-based UEs [26].

# 2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] J. Fliege und U. Maier: "A two-stage approach for computing cubature formulae for the sphere," Dortmund University, 1999.

[3] ISO 3745 - Annex A: "Acoustics - Determination of sound power levels and sound energy levels of noise sources using sound pressure -- Precision methods for anechoic rooms and hemi-anechoic rooms - Annex A: General procedures for qualification of anechoic and hemi-anechoic rooms".

[4] ISO/R 1996-1972: "Acoustics – Assessment of noise with respect to community response".

[5] ANSI S1.4: "Specifications for Sound Level Meters".

[6] ISO 3: "Preferred numbers – Series of preferred numbers".

[7] B. Rafaely, “Analysis and design of spherical microphone arrays,” IEEE Transactions on Speech and Audio Processing, no. 13, 2005, pp. 135 – 143

[8] M. Poletti, “Unified Description of Ambisonics Using Real and Complex Spherical Harmonics,” Ambisonics Symposium 2009, June 25-27, 2009, Graz, Austria.

[9] Recommendation ITU-T G.100.1 (06/2015): "The use of the decibel and of relative levels in speechband telecommunications".

[10] Recommendation ITU-T P.56 (12/2011): "Objective measurement of active speech level".

[11] Recommendation ITU-T P.57 (06/2021): "Artificial ears".

[12] Recommendation ITU-T P.58 (03/2023): "Head and torso simulator for telephonometry".

[13] Recommendation ITU-T P.79 (11/2007): "Calculation of loudness ratings for telephone sets".

[14] Recommendation ITU-T P.501 (05/2020): "Test signals for use in telephonometry".

[15] Recommendation ITU-T P.581 (07/2022): "Use of head and torso simulator for hands-free and handset terminal testing".

[16] Recommendation ITU-T P.340 (05/2000): "Transmission characteristics and speech quality parameters of hands-free terminals".

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

[18] Recommendation ITU-T P.380 (07/2022): "Electro-acoustic measurements on headsets".

[19] Recommendation ITU-T P.381 (03/2023): "Technical requirements and test methods for analogue wired headsets or headphones and corresponding universal interface of terminals".

[20] Recommendation ITU-T P.382 (03/2023): "Technical requirements and test methods for analogue wired multi-microphone headsets or headphones and corresponding universal interface of terminals".

[21] Recommendation ITU-T P.383 (03/2023): "Technical requirements and test methods for digital headsets or headphones and corresponding interfaces of terminals".

[22] Recommendation ITU-T P.700 (06/2021): "Calculation of loudness for speech communication".

[23] Recommendation ITU-R BS.1770-5 (11/2023): "Algorithms to measure audio programme loudness and true-peak audio level".

[24] IEC 60268-1:1985: "Sound system equipment. Part 1: General".

[25] IEC 61260-1:2014: "Electroacoustics - Octave-band and fractional-octave-band filters - Part 1: Specifications".

[26] 3GPP TS 26.131: "Speech and video telephony terminal acoustic test specification".

[27] 3GPP TS 26.132: "Speech and video telephony terminal acoustic test specification".

[28] 3GPP TS 26.250: "Codec for Immersive Voice and Audio Services - General overview".

[29] 3GPP TS 26.253: " Codec for Immersive Voice and Audio Services; Detailed Algorithmic Description incl. RTP payload format and SDP parameter definitions".

[30] 3GPP TS 26.254: "Codec for Immersive Voice and Audio Services - Rendering".

[31] 3GPP TS 26.258: "Codec for Immersive Voice and Audio Services; C code (floating-point)".

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

[33] USB Implementors' Forum: "HID Usage Tables for USB", Version 1.5.

[34] H. Wittek and G. Theile: "The recording angle – based on localisation curves", proceedings of the 112th AES convention, Munich, 2002.

# 3 Definitions, symbols and abbreviations

## 3.1 Definitions

For the purposes of the present document, the terms and definitions given in 3GPP TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in 3GPP TR 21.905 [1].

**spherical coordinates:** A coordinate system with x-axis pointing to the front, y-axis to the left and z-axis to the top, with the distance $r$ from the origin, the azimuth $\phi$ in mathematical positive orientation (counter-clockwise) and the elevation angle $\theta$ relative to the z-axis (with 0 degrees pointing to the equator and +90 degrees pointing to the North pole). See Figure 0.

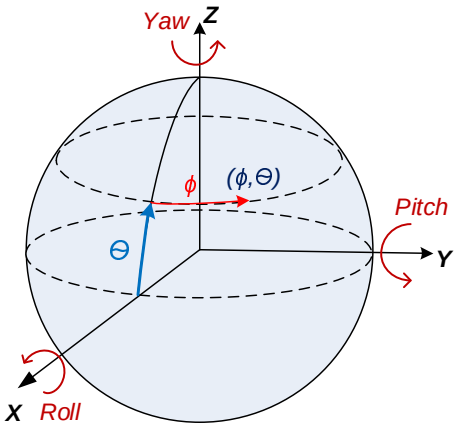


**Figure 0: Spherical coordinate system**

**Coded Formats**: The format represented in the IVAS coded frames, which is generally the input format to the encoder.

**User Capture**: A use case where the user's voice is the primary intended signal to be captured by the UE.

**Spatial Capture**: A use case where acoustic scenes that include directional and/or diffuse sound sources are intended to be captured by the UE.

**Stereo Panorama**: the spatial image of a stereo signal, in which the sound source directions lie in a range from -100% (left) to 100% (right).

## 3.2 Symbols

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

| | |
|---|---|
| (A) | A-weighting |
| dB | Decibel |
| dBFS | Digital level in dB, where 0 dBFS refers to the RMS level of a DC-free sinusoidal signal exercising the full scale of the digital interface/file. |
| dBm0 | Digital overload point in dB |
| dBov | Digital level in dB, where 0 dBov refers to the RMS level of a DC-free rectangular signal exercising the full scale of the digital interface/file. |
| dBPa | Sound pressure level in dB, referenced to 1 Pa |
| dBSPL | Sound pressure level in dB, referenced to 2E-5 Pa |
| dBV | Voltage level in dB, referenced to 1 V |
| $f$ | Frequency (in Hertz) |
| Hz | Unit of frequency (Hertz) |
| $G(f)$ | Frequency response of measured versus reference signal |
| $LA_{eq}$ | Sound level in dB equivalent to the total A-weighted sound energy measured over a stated period of time. |
| $L_{eq}$ | Sound level in dB equivalent to the unweighted sound energy measured over a stated period of time. |
| Ω | Electrical resistance in Ohm |
| $\hat{P}(f)$ | Spectral magnitude of measured signal |
| $P_{ref}(f)$ | Spectral magnitude of reference signal |
| $\hat{P}_k(f)$ | Spectral magnitude of measured signal (k-th channel) |
| $\hat{P}_l^m(f)$ | Spectral magnitude of measured Ambisonics degree *l* and index *m*. |
| Pa | Unit of pressure (Pascal) |
| ϕ | Azimuth angle (phi) |
| *r* | Distance from a point in space to the origin of the spherical coordinate system (radius of sphere) |
| RCV | Receiving (direction) |
| SND | Sending (direction) |

| | |
|---|---|
| $\theta$ | Elevation angle (theta) |
| $T_R$ | UE delay in receiving direction |
| $T_{TER}$ | Test equipment delay in receiving direction |
| $T_S$ | UE delay in sending direction |
| $T_{TES}$ | Test equipment delay in sending direction |
| V | Unit of voltage (Volt) |
| $\zeta(\phi)$ | Estimate for a single source direction in the stereo panorama that is physically positioned in direction $\phi$ relative to the capturing device |

## 3.3 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [1].

| | |
|---|---|
| EXT | External output |
| FOA | First-Order Ambisonics |
| HATS | Head-And-Torso Simulator |
| HFRP | Hands-free Reference Point |
| HID | Human Interface Device |
| HOA2 | Higher-Order Ambisonics (2nd order) |
| HOA3 | Higher-Order Ambisonics (3rd order) |
| HRP | HATS Reference Point |
| ILD | Interaural Level Difference |
| ISM | Independent Stream with Metadata |
| ITD | Interaural Time Difference |
| IVAS | Immersive Voice and Audio Services |
| ISAR | Immersive Audio for Split Rendering Scenarios |
| LKFS | Loudness, K-weighted, relative to Full Scale |
| MASA | Metadata-Assisted Spatial Audio |
| MRP | Mouth Reference Point |
| N3D | Normalization (full) 3-Dimensional (of an Ambisonics signal) |
| NR | Noise Rating |
| OMASA | Objects (ISM) with Metadata-Assisted Spatial Audio |
| OSBA | Objects (ISM) with Scene-Based Audio |
| RLL | Receive Loudness Level |
| RLR | Receive Loudness Rating |
| SBA | Scene-Based Audio (Ambisonics) |
| SLR | Send Loudness Rating |
| SN3D | Schmidt Normalization (semi) 3-Dimensional (of an Ambisonics signal) |
| SS | System Simulator |
| TC | Transport Channel (for MASA) |
| THD | Total Harmonic Distortion |
| USB | Universal Serial Bus |

# 4 Objective Test Methodologies for Immersive Audio Systems

## 4.0 General

### 4.0.1 Applicability

This clause describes general objective test methodologies for immersive audio systems. For testing IVAS-based devices, refer to clause 5.

## 4.0.2 Test equipment

Unless specified otherwise, the accuracy of electric and acoustic measurements made by test equipment shall meet the requirements defined in clause 5.3 of 3GPP TS 26.132 [27].

For tests with head tracking, HATS rotation around the vertical axis should be realized using a motorized turntable or a HATS with motorized head rotation. For motorized or manual rotations of HATS and/or UE, error in orientation (elevation and azimuth) shall not exceed ±2°.

NOTE 1: A motorized rotation of HATS and/or UE is recommended. Some UEs may not have a natural reference orientation (which, for instance, may be defined by the direction of a screen). In this case, the UE may reset the reference direction automatically after a span of time, e.g., to the current device orientation. This should be taken care of during the measurement. The measurement with the rotated HATS and/or UE should be performed quickly enough to prevent the reference direction from being spuriously readjusted. Therefore, it benefits from automation.

Head-and-torso simulators (HATS) used for acoustic testing are specified in ITU-T Recommendation P.58 [12], corresponding artificial ears for testing in receiving direction are specified in ITU-T Recommendation P.57 [11] (Type 3.3 or Type 4).

NOTE 2: For testing insert-type earphones (one of the Headset UE types) in receiving direction, a Type 4 artificial ear is recommended.

In sending direction, HATS equipped with mouth simulators (or equivalent stand-alone mouth simulator) used as a single sound source for spatial or user capture (see clause 5.4.1) in the test arrangement shall comply with ITU-T P.58 [12] and ITU-T P.581 [15]. An artificial mouth of a HATS can be replaced with an equivalent stand-alone mouth simulator, unless:

- the UE is tested in receiving direction (artificial ears are required), or
- headset UE or handset UE is tested (HATS is necessary to mount the UE).

Unless otherwise specified, active loudspeakers (or passive loudspeakers with an amplification system) used in this specification shall meet the following requirements:

- The spectrum of the acoustic signal produced by the loudspeaker shall be equalized under free field conditions with a measurement microphone positioned on the main loudspeaker axis at 1 m from the loudspeaker membrane. The achieved equalized spectrum in 1/3rd octave bands, when measured in the test environment, shall be within ±1 dB from 100 Hz to 200 Hz and shall be within ±0.5 dB from 200 Hz to 20 kHz.
- THD $\leq$ [0.5 %] when measured at 1 m on axis with an 85 dBSPL sinusoidal signal for frequencies $\geq$ 100 Hz
- THD $\leq$ [2%] when measured at 1 m on axis with an 85 dBSPL sinusoidal signal for frequencies $\geq$ 50 Hz and < 100 Hz.
- Maximum long-term level $\geq$ 96 dBSPL when measured at 1 m on axis with the simulated programme signal defined in clause 7 of IEC 60268-1 [24].
- A self-generated noise level $\leq$ [5 dBSPL(A)], when measured at 1 m on axis in a free-field environment.

NOTE 3: For certain tests, e.g., when measuring the self-generated noise of the UE, the loudspeakers may need to be disabled to prevent their self-generated noise from impacting the results.

## 4.0.3 Test environment acoustic properties

The test environment (anechoic chamber) shall contain a free-field volume, wherein free-field sound propagation conditions shall be observed. The free-field sound propagation conditions shall be observed down to a frequency of 200 Hz. Qualification or verification of this requirement can be conducted using the methods and limits for deviation from ideal free-field conditions as specified in ISO 3745 [3] or ITU-T P.340 [16].

For testing headset UE in receiving direction, the test environment according to clause 6.1 of ETSI TS 103 224 [32] may be used.

NOTE: The usage of the test environment according to ETSI TS 103 224 for other UE types and use cases is for further study.

The equivalent continuous sound level of the test environment in each 1/3rd octave band, $L_{eq}(f)$ shall be less than the limits of the NR15 curve and should be less than the limits of the NR10 curve, when following the noise rating (NR) determination procedures in [4].

# 4.1 Objective Test Methodologies for Assessment of Immersive Audio Systems in the Sending Direction

## 4.1.1 Diffuse-field Send Frequency Response for Scene-based Audio

### 4.1.1.1 Introduction

This test is applicable to UEs capturing scene-based audio (e.g. First and Higher Order Ambisonics).

NOTE: Currently, the test method uses a periphonic loudspeaker array for generation of a diffuse soundfield. Additional loudspeaker setups for the derivation of the diffuse sound field are under consideration.

The test environment acoustic properties shall meet the requirements in clause 4.0.3.

### 4.1.1.2 Definition

The Diffuse-field Send Frequency Response for Scene-based Audio is defined as the transfer function, $G(f)$, between:

$\hat{P}(f)$, the estimated sound pressure magnitude spectrum obtained from a diffuse-field scene-based audio capture and reference synthesis at the geometric center of a *free-field volume*; and

$P(f)$, the sound pressure magnitude spectrum obtained from a diffuse-field microphone recording the same diffuse field at the origin of a spherical coordinate system.

Figure 1 describes a typical block diagram for the scene-based audio sending direction with measurement points when using a periphonic loudspeaker array.

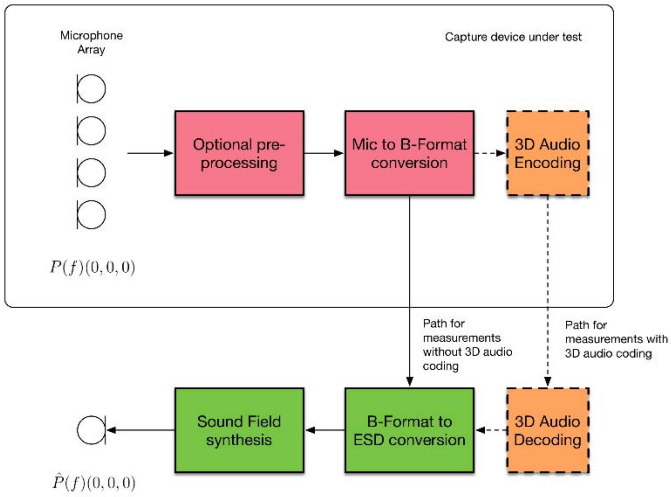


**Figure 1: Scene-based audio capture block diagram for sending direction measurements**

**Definition of Equivalent Spatial Domain**

The equivalent spatial domain representation, **w**(t)**,** of a $N^{th}$ order Ambisonics soundfield representation **c**(t) is obtained by rendering **c(**t) to $K$ virtual loudspeaker signals $w_j(t)$, $1 \le j \le K$, with $K = (N+1)^2$. The respective virtual loudspeaker positions are expressed by means of a spherical coordinate system, where each position lies on the unit sphere, i.e., a radius of 1. Hence, the positions can be equivalently expressed by order-dependent directions $\boldsymbol{\Omega}_j^{(N)}=(\theta_j^{(N)}, \phi_j^{(N)})$, $1 \le j \le K$, where $\theta_j^{(N)}$ and $\phi_j^{(N)}$ denote the inclinations and azimuths, respectively. These directions are defined according to [2] and reproduced in Annex A for convenience.

The rendering of $\boldsymbol{c}(t)$ into the equivalent spatial domain can be formulated as a matrix multiplication:

$$\mathbf{w}(t) = (\boldsymbol{\Psi}^{(N,N)})^{-1} \cdot \mathbf{c}(t),$$

where $(\cdot)^{-1}$ denotes the inversion.

The matrix $\boldsymbol{\Psi}^{(N,N)}$ of order $N$ with respect to the order-dependent directions $\boldsymbol{\Omega}_j^{(N)}$ is defined by:

$$\boldsymbol{\Psi}^{(N,N)} := [\mathbf{S}_1^{(N)} \quad \mathbf{S}_2^{(N)} \quad \dots \quad \mathbf{S}_K^{(N)}],$$

with:

$$\mathbf{S}_j^{(N)} := [S_0^0(\boldsymbol{\Omega}_j^{(N)}) \quad S_{-1}^{-1}(\boldsymbol{\Omega}_j^{(N)}) \quad S_{-1}^{0}(\boldsymbol{\Omega}_j^{(N)}) \quad S_{-1}^{1}(\boldsymbol{\Omega}_j^{(N)}) \quad S_{-1}^{1}(\boldsymbol{\Omega}_j^{(N)}) \quad \dots \quad S_N^N(\boldsymbol{\Omega}_j^{(N)})]^T,$$

where $S_n^m(\cdot)$ represents the real valued spherical harmonics of the order n and degree m as defined in [8].

The matrix $\boldsymbol{\Psi}^{(N,N)}$ is invertible so that the HOA representation $\boldsymbol{c}(t)$ can be converted back from the equivalent spatial domain by:

$$\mathbf{c}(t) = \boldsymbol{\Psi}^{(N,N)} \cdot \mathbf{w}(t)$$

### 4.1.1.3 Test method with periphonic array

#### 4.1.1.3.1 Test Conditions

**Periphonic loudspeaker array**

a) A *periphonic loudspeaker array* shall be placed within the free-field volume with the geometric center of the *periphonic loudspeaker array* coinciding with the geometric center of the free-field volume.

b) The *periphonic loudspeaker array* shall have a radius greater or equal than 1 meter.

c) The *periphonic loudspeaker array* shall be composed of $(N+1)^2$ coaxial loudspeaker elements. Each of the $(N+1)^2$ coaxial loudspeaker elements shall be equalized (if necessary) according to the requirements in Clause 4.0.2. $N$ should be equal to 4.

d) The $(N+1)^2$ coaxial loudspeaker elements shall be positioned according to the azimuth and elevation coordinates given in Annex B.

e) All coaxial loudspeaker elements shall be oriented such that their acoustic axis intersects at the geometric center of the *free field volume*.

f) The radius of each coaxial loudspeaker element shall be such that, at the geometric center of the *free-field volume*, the far field approximation for the coaxial loudspeaker axial pressure amplitude decay holds true.

g) The active loudspeakers (or passive loudspeakers with an amplification system) should meet the requirements in Clause 4.0.2

#### 4.1.1.3.2 Measurement

**Reference Spectrum measurement for periphonic loudspeaker array method**

a) A diffuse-field / random incidence, or multi-field microphone is mounted in the *free-field volume* such that the tip of the microphone corresponds to the geometric center of the *free-field volume* and the geometric center of the *periphonic loudspeaker array*.

NOTE 1: Diffuse-field / random incidence microphones, are described in [5].

b) $(N+1)^2$ decorrelated pink noise signals are played simultaneously over each of the $(N+1)^2$ coaxial loudspeakers of the *periphonic loudspeaker array*.

c) The playback level is adjusted such that the *LAeq*, measured over a 30s time window at the geometric center of the *periphonic loudspeaker array*, is equal to 78dBSPL(A) ± 0.5dB.

d) The reference sound pressure at the geometric center of the *free-field volume*, *p(t)*, is captured with the diffuse-field or multi-field microphone.

e) The magnitude spectrum of the reference sound pressure, *P(f)*, is calculated for the 1/12th octave intervals as given by the R40 series of preferred numbers in [6].

NOTE 2: For ideal (calibrated) loudspeakers, the *P(f)* spectra should have equal energy in each 1/12th octave intervals.

**Estimated Spectrum measurement**

a) The scene-based audio capture device under test is mounted in the *free-field volume* such that its geometric center coincides with the geometric center of *free-field volume* and the geometric center of the *periphonic loudspeaker array*.

b) $(N+1)^2$ decorrelated pink noise signals are played simultaneously over each of the $(N+1)^2$ coaxial loudspeakers of the *periphonic loudspeaker array*. The pink noise signals shall be identical to the signals used for the reference spectrum measurement.

c) The B-format scene-based audio format representation (compressed or uncompressed, depending on the use case being tested) is stored for offline analysis.

d) The B-format scene-based audio format representation is uncompressed (if necessary) and converted to an *equivalent spatial domain representation* of order $N_{DUT}$ (B-Format to ESD conversion in Figure 1), where $N_{DUT}$ corresponds to the Ambisonics order of the device under test.

e) $\hat{p}(t)$, the estimate of the sound field at the geometric center of *the free-field volume* and *periphonic loudspeaker array*, is synthesized using the *equivalent spatial domain representation* of order $N_{DUT}$.

NOTE 3: $\hat{p}(t)$ can be taken from the W component of the B-Format signal, as an alternative to implementing the B-Format to ESD conversion in step d).

f) The magnitude spectrum of the estimated sound pressure, $\hat{P}(f)$, is calculated for the 1/12th octave intervals as given by the R40 series of preferred numbers in [6].

**Calculation of send frequency response for scene-based audio**

The send frequency response for scene-based audio, *G(f)*, is calculated as $G(f) = \frac{\hat{P}(f)}{P(f)}$.

### 4.1.1.4 Test method with loudspeaker array and turn table

#### 4.1.1.4.1 Test Conditions

**Loudspeaker array**

a) A calibrated *loudspeaker array* shall be placed within the *free-field volume*.

b) The *loudspeaker array* shall comprise one or several semi-arcs having a radius greater or equal than 1 meter. The radius shall be reported.

c) The *loudspeaker array* shall be composed of *N+1* loudspeaker elements. The ambisonic order *N* shall be reported.

d) Each loudspeaker in the array shall be calibrated with a frequency response of [at least 100 Hz-20,000 Hz] and minimum phase response.

e) The coordinates of the loudspeaker elements are defined according to a Gaussian spherical grid [7] of order *N*. Directions shall comply with Annex B.1 and the *N+1* elevations of the spherical grid shall be reported.

**Turn table**

a) A turn table with a resolution of 0.5 degrees shall be used. The rotation axis of the turn table and the vertical axis of the semi-arcs shall be aligned The turn table shall be adjusted in height so that the device under test is positioned at the geometric center of the *loudspeaker array*.

b) For measurement, an azimuth step of 180/(*N*+1) degrees shall be used.

#### 4.1.1.4.2 Measurement

**Reference Spectrum measurement**

a) A diffuse-field / random incidence, or multi-field microphone is mounted in the *free-field volume* such that the tip of the microphone corresponds to the geometric center of the *free-field volume* and the geometric center of the *loudspeaker array*.

NOTE 1: Diffuse-field / random incidence microphones, are described in [5].

Repeat steps b-c) with an azimuth angular resolution of 180/(*N*+1) degrees:

b) An exponential sweep sine signal is played over each of the *N*+1 loudspeakers of the *loudspeaker array*.

c) The impulse response at the geometric center of the *loudspeaker array p(t)* is measured for each loudspeaker position.

d) The magnitude spectrum of the reference sound pressure, *P(f)*, is calculated for the 1/12$^{th}$ octave intervals as given by the R40 series of preferred numbers in [6].

NOTE 2: For ideal (calibrated) loudspeakers, the *P(f)* spectra should have equal energy in each 1/12$^{th}$ octave intervals.

**Estimated Spectrum measurement**

a) The scene-based audio capture device under test is mounted in the *free-field volume* such that its geometric center coincides with the geometric center of *free-field volume* and the geometric center of the *loudspeaker array*.

b) Repeat steps b-c) with an azimuth angular resolution of 180/(*N*+1) degrees:

c) An exponential sweep sine signal is played over each of the *N*+1 loudspeakers of the *loudspeaker array*. The sweep signals shall be identical to the signals used for the reference spectrum measurement.

d) The impulse response at the geometric center of the *loudspeaker array p(t)* is measured for each loudspeaker position.

e) The magnitude spectrum of the estimated sound pressure, $\hat{P}(f)$, is calculated for the 1/12$^{th}$ octave intervals as given by the R40 series of preferred numbers in [6].

**Calculation of send frequency response for scene-based audio**

The send frequency response for scene-based audio, *G(f)*, is calculated as $G(f) = \frac{\hat{P}(f)}{P(f)}$.

Due to practical constraints (e.g. reflections on turn table), measurements for specific elevations (e.g. < - degrees) may be unreliable and discarded. In this case, the above measurement procedure may be conducted in two phases by measuring only directions for one hemisphere (e.g. top hemisphere, with elevations >0) in each phase. The device under test shall be flipped upside down between the two phases, and this two-phase approach shall be reported.

### 4.1.2 Directional response measurement for scene-based audio

#### 4.1.2.1 Definition

The directional response for scene-based audio is defined as the transfer function, represented as an impulse response, $\mathbf{h}(\theta_i, \phi_i)$, between a device under test and a loudspeaker located at an equal distance *r* and L predefined directions, $(\theta_i, \phi_i)$, $i$=1,...,L.

#### 4.1.2.2 Test conditions

The test environment acoustic properties shall meet the requirements in clause 4.0.3.

**Loudspeaker array**

A loudspeaker array comprising L loudspeakers located at a set of predefined directions $(\theta_i, \phi_i)$, $i$=1,...,L, from the geometric center of the *loudspeaker array* shall be used.

#### 4.1.2.3 Measurement

For each loudspeaker position $(\theta_i, \phi_i)$, $i$=1,...,L , the following procedure shall be used:

a) An exponential sweep sine test signal is played over the loudspeaker.

NOTE: The impact of codec on the exponential sweep sine test signal needs to be verified before performing the measurements. An activation signal may be needed.

b) The impulse response $\mathbf{h}(\theta_i, \phi_i)$ at the geometric center of the *loudspeaker array* is measured.

## 4.2 Objective Test Methodologies for Assessment of Immersive Audio Systems in the Receiving Direction

### 4.2.1 Headset Binaural Diffuse-field Receive frequency response for Scene-based audio

#### 4.2.1.1 Introduction

This test is applicable to UEs rendering scene-based audio (e.g. First and Higher Order Ambisonics) over a binaural headset.

#### 4.2.1.2 Definition

The Headset Binaural Diffuse-field Receive Frequency Response for Scene-based Audio (for left and right ears) is defined as the transfer function, $G_{L,R}(f)$, between:

a) $P_{L,R}(f)$, the binaurally recorded sound pressure magnitude spectra, obtained when a diffuse field signal in the equivalent spatial domain representation, w(t), is played on the DUT; and

b) $P_{ref\,L,R}(f)$, the reference sound pressure magnitude spectra, obtained by direct convolution of the diffuse field signal in the equivalent spatial domain representation, w(t) with its corresponding set of HRTFs.

#### 4.2.1.3 Test Conditions

The test environment acoustic properties shall meet the requirements in clause 4.0.3.

The set of HRTFs used by the UE shall be documented and available to the test lab.

#### 4.2.1.4 Measurement

**Reference sound pressure magnitude spectra**

The reference sound pressure magnitude spectra are derived offline. The reference sound pressure magnitude spectra for the left and right ears, $P_{ref\ L,R}(f)$ is the frequency domain representation of the convolution between the set of equivalent spatial domain signals, **w**(t), with its corresponding set head related transfer functions $\boldsymbol{h}_{L,R}(t)$, for each direction $j$ in an equivalent spatial domain of order $N_{DUT}$, i.e.:

$$P_{ref\ L,R}(f) = \mathcal{F}\left( \sum_{j=1}^{(N_{DUT}+1)^2} w_j(t) * h_{j\ L,R}(t) \right)$$

The signals $w_j(t)$, for $1 \leq j \leq (N_{DUT}+1)^2$, are uncorrelated pink noise signals of 30s length.

**Binaurally recorded sound pressure magnitude spectra**

The binaurally recorded sound pressure magnitude spectra is obtained as follows:

a) The binaural headset is placed on a HATS.

b) The DUT shall be configured such that the set of HRTFs used for binaural rendering correspond to the HATS used for testing.

c) The DUT volume control (if any) is adjusted for its nominal setting.

d) The binaural time-domain signals are recorded with HATS.

e) The binaurally recorded sound pressure magnitude spectra, $P_{L,R}(f)$ is obtained by taking the Fourier transform of the binaurally recorded time-domain signals.

**Calculation of headset binaural diffuse-field receive frequency response for scene-based audio**

The headset binaural diffuse-field frequency response for scene-based audio, $G(f)$, is calculated for each supported Ambisonics order $N_{DUT}$ as:

$$G(f) = \frac{P_{L,R}(f)}{P_{ref\ L,R}(f)}$$

### 4.2.2 Nominal System Sensitivity in Receive Direction for Channel-based audio

#### 4.2.2.1 Introduction

This test is applicable to UEs rendering channel-based audio (e.g. 7.1.4) over a binaural headset.

#### 4.2.2.2 Definition

The nominal system sensitivity in receive direction for channel-based audio is defined as the difference between the sound pressure level (in dBSPL(A)) produced by the DUT on HATS and the root mean square of the digital test signal (in dBFS).

#### 4.2.2.3 Test Conditions

The test environment acoustic properties shall meet the requirements in clause 4.0.3.

The specific HATS used for the recording shall be described in the test report. The set of HRTFs used by the UE shall be documented and available to the test lab.

#### 4.2.2.4 Measurement

For each audio channel supported by the DUT, a pink noise signal with -18 dBFS RMS level is played, with the signals played only one channel at a time.

The *LAeq* (in dBSPL(A)) is measured continuously for a period of 30 s for each of the left and right ears.

The sensitivity $G_{i\,L,R}$ is expressed as the difference of the recorded sound pressure levels at the left and right ears and the root mean square digital level of the pink noise test signal, i.e. -18 dBFS.

$$G_{i\,L,R} = LA_{eq} - 18$$

### 4.2.3 Motion to Sound Latency in Dynamic Binaural Rendering Systems

#### 4.2.3.1 Introduction

Motion to Sound latency is the time difference between the event of a change in head rotation and when the immersive audio signal is finally compensated for the head motion. The method in this specification is intended to verify that the overall motion-to-sound latency that a user experiences upon rotating their head is within acceptable limits.

The method allows full measurement of motion to sound, i.e. including both the latency of the head tracking sensor as well as the audio playback. This includes all components of a real setup and therefore contains all possible causes of additional latency that a user may experience.

The method also provides a latency value for the isolated audio processing of the binaural renderer without the aforementioned external hardware, assuming that the binaural renderer can process audio data as an audio processing plugin that can be evaluated in isolation.

NOTE: This method requires synchronized playback of two renderer instances and may not be suitable for the measurements of UEs where such synchronization is not possible.

#### 4.2.3.2 Requirements

The following will be required:

Software:

- Audio processing software to run and record output of two renderers simultaneously
- Head tracker software

Hardware:

- Host machine for audio processing
- Head tracker hardware
- Stereo audio recording interface
- Stereo audio playback interface
- Mechanical setup to rotate the head tracking sensor in a precise and reproducible way

An exemplary hardware setup can be seen below in Figure 2, the method however can also be implemented using different systems under test and accompanying equipment:

**Figure 2: Hardware Overview (Setup in Position 1 on the left, Position 2 on the right)**

The audio processing environment uses two parallel signal chains, each containing its own instance of the same binaural renderer being tested. The test is concerned only with yaw angles, so values of pitch and roll should be set to zero at the beginning of the test and can be ignored thereafter.

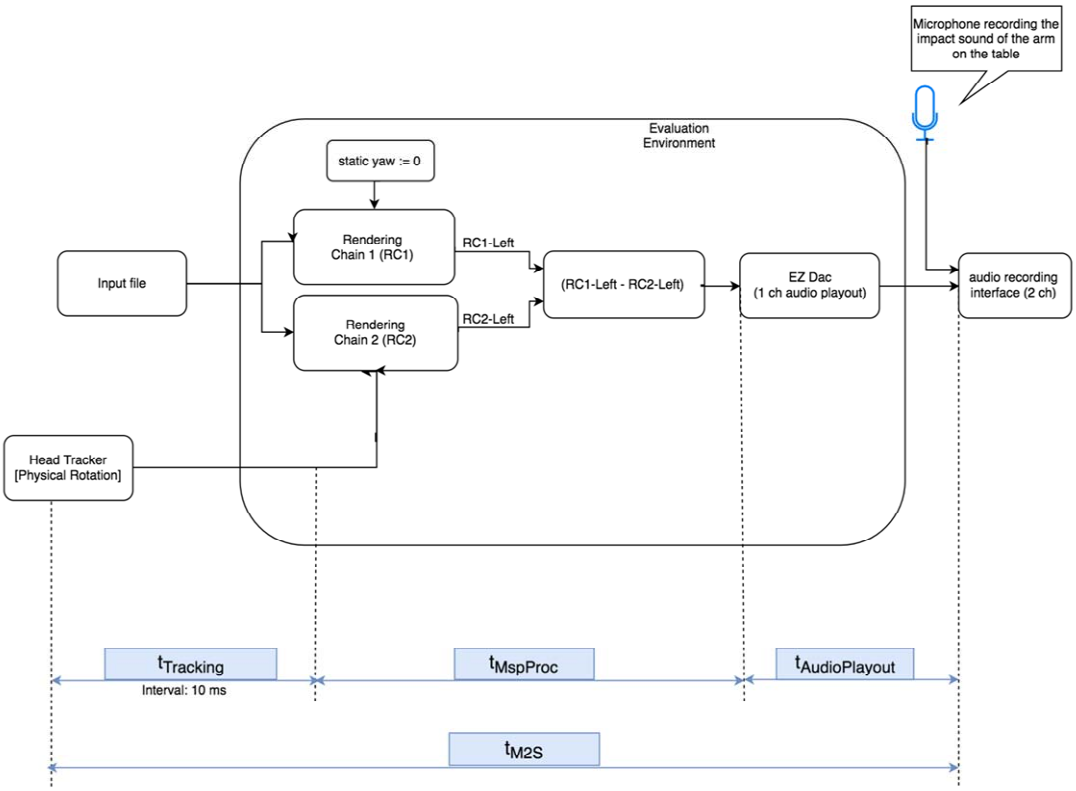


**Figure 3: Generic Audio Processing Environment**

The initial conditions are that Rendering Chain 1 (RC1) has a static yaw head rotation angle of 0 degrees and RC2 uses the physical rotation of the head tracker to get its yaw value. A white noise signal is virtually placed directly in front of the listener (0 degrees azimuth, elevation), meaning that rotation of the arm directly affects how the white noise source is rendered.

#### 4.2.3.3 Calibration

The first step is to calibrate the final position of the rotating arm (Position 2 / P2). The rotating arm is moved manually and requires only a limited range of motion - from some small rotation away from the table (Position 1 / P1), 20 to 30 degrees will be ample, through until contact with the table (P2). The arm should be placed at P2 and set up so that this position also corresponds to 0 degrees yaw.

#### 4.2.3.4 Evaluation Environment

An object within the evaluation environment, e.g. using Max/MSP, should be created to set the value of yaw to exactly 0 degrees once the real value of yaw (received from the head tracker) is <0.2 degrees.

NOTE: This tolerance value was chosen to be as small as possible while ensuring that it does not bounce (dependent on the accuracy of the tracking system)

This object should be designed to latch to zero once the actual value is under the tolerance threshold, so that any small accidental rebound of the rotating arm does not affect the yaw angle fed to the renderer - it artificially remains at exactly 0, which is important to ensure that both rendering chains have exactly the same head rotation when the arm is in its final position (P2). The output from the evaluation environment is captured by the recording audio interface, which therefore includes any latency introduced by playback.

#### 4.2.3.5 Data acquisition

A test run begins by starting to record on the recording audio interface. The rotating arm is set to Position 1, then the audio processing set running and starts feeding the input source to both renderer instances. A microphone is positioned near the contact point at the table. This mono room microphone will be recorded synchronously with the output from the evaluation environment, with its purpose being to log the point of contact of the arm with the table, which should be done with a good amount speed and vigour so that the microphone picks up a loud knock at the table. Shortly after this (one or two seconds for example), with the test run now complete, playback and recording can be stopped.

Some milliseconds after the collision, the latest yaw value detected by the tracking system will have been passed into the evaluation environment ($t_{Tracking}$). With the target yaw value now reached (latched to zero in Max), both rendering chains will have the identical values of head rotation and therefore, after some further short delay, the output of both renderer instances will be identical.

#### 4.2.3.6 Data Analysis

The overall motion-to-sound latency (tM2S) is taken as the time from the moment of collision until the point at which the two output signals are identical.

To easily visually inspect when this point occurs, one output channel of one signal chain (e.g. RC1-left) is subtracted from the same output channel of the complementary signal chain (RC2-left).

NOTE 1: This could be done manually in audio editor software after processing, but this would require recording at least three channels synchronously (one from each renderer chain, and one of the room microphone). Instead, the subtraction of signals can be done within Max/MSP, meaning only the output of this operation (one channel) and the room microphone can conveniently be recorded with a stereo audio interface. In addition to the stereo WAVE file recorded by a separate audio application, the Max/MSP application also writes to a separate mono WAVE file once it detects that it is in the final tolerated yaw position (latched on). This mono WAVE contains only the subtracted signals as described above, from which the tMspProc time can also be measured.

Evaluation is performed offline in audio editor software. The tMspProc time is measured from the start of the file until the point at which consecutive zero samples begin. This value encompasses any motion-to-sound latency caused by the tested renderer chain as well as any other latency caused by Virtual Studio Technology (VST) plugin framework buffering. The tMspProc time shall be measured from the audio frame boundary at which the latched-on yaw value is activated and applied within that audio frame.

NOTE 2: Since the yaw rotation update rate of the tracker is typically in the range of a few milliseconds, there is a framing mismatch when compared to the audio framing, but this mismatch will not be incorporated in the tMspProc value but rather only in the tM2S measurement.

An example measurement of the tMspProc is displayed in Figure 4. For tM2S this is measured by selecting the duration between the visible collision peak in the microphone channel and the point at which the other channel reduces to silence. Figure 5 shows an example measurement for the motion-to-sound latency.

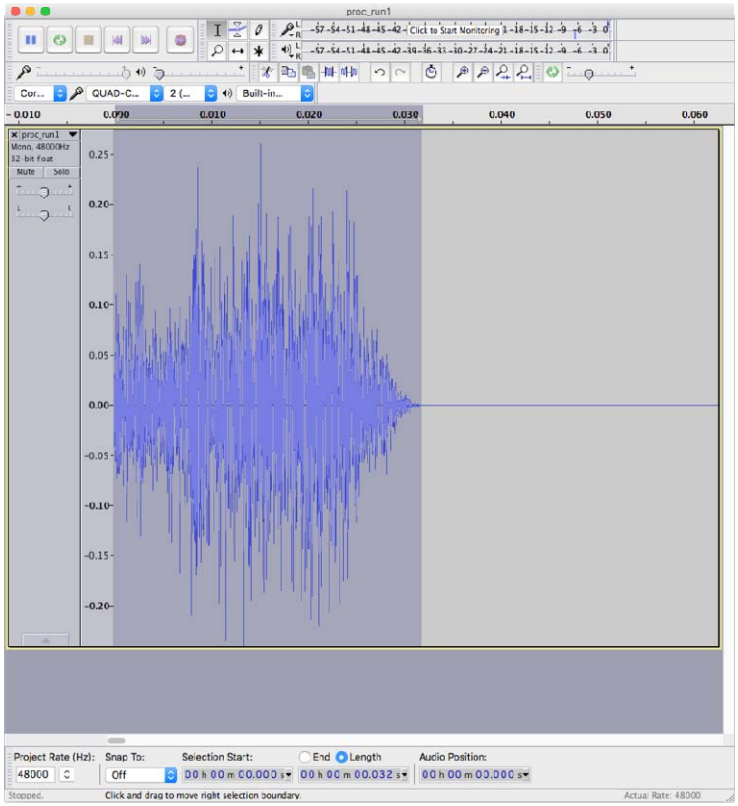


**Figure 4: tMspProc latency measurement**

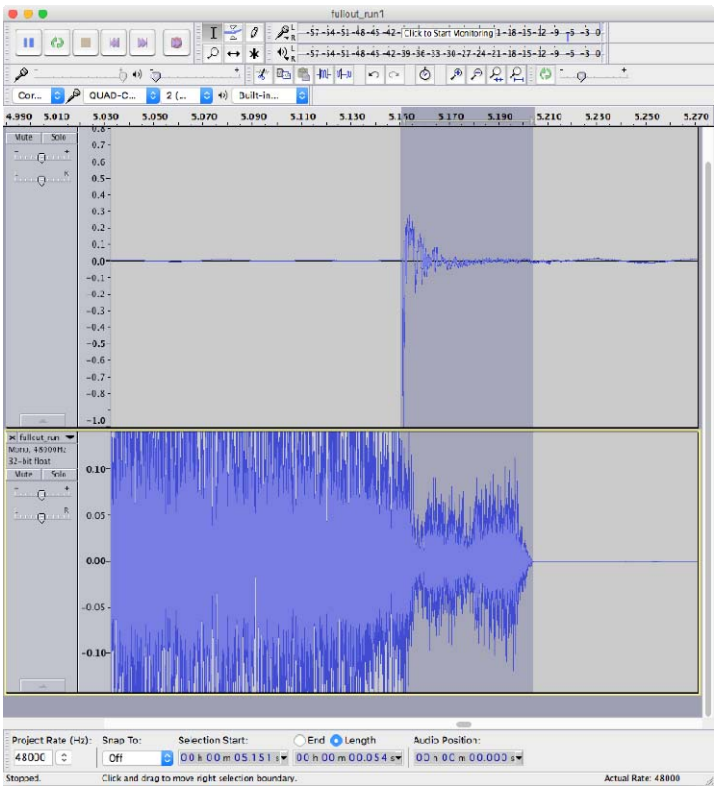


**Figure 5: Motion-to-sound (tM2S) latency measurement**

In Figure 5 the room recording is on the top, subtracted renderer output is on the bottom. Marked region is the time passed since the arm hits the table (recorded knock) and when the subtracted binaural renderer output reaches silence.

NOTE 3: Unlike the tMspProc measurement, the tM2S measurement is taken from signals recorded from hardware audio interfaces, hence it is not possible to look for continuous silence since the resulting file will always contain some noise added by the digital-to-analog and analog-to-digital converters. For this reason, it is important to ensure a high signal-to-noise ratio in the signal provided to audio interfaces, to make it easier to inspect where the cancellation occurs.

# 5 Objective Test Methodologies for IVAS-based UEs

## 5.1 Overview

This clause adds objective test methodologies specific to IVAS-based UEs. It provides for testing of UEs on either an acoustical or an electrical interface. In case of acoustical interface testing, a HATS (head and torso simulator) is used for a realistic simulation of an average user. A motorized turntable, or a HATS with motorized head rotation, is used to test head-tracking behavior of the UE under test. In contrast to speech-based mono telephony, which typically aims to suppress the acoustical background noise, IVAS-based UEs may capture and/or reproduce spatial information of the acoustic scene.

## 5.2 Interface Definitions

UE testing is realized by connecting a terminal to a test system composed of a system simulator and a reference client. The system simulator simulates the access network, provides core network functionalities and a point of interconnection (POI). The reference client serves as the far-end communication endpoint at the POI and provides IVAS encoder, decoder and rendering functionalities. Test sequences are both captured and fed into the reference client for sending and

receiving direction tests, as illustrated in Figure 6. Alternatively, IVAS encoding, decoding, and rendering can be part of the system simulator instead of the reference client. No further transcoding beyond linear PCM shall take place.

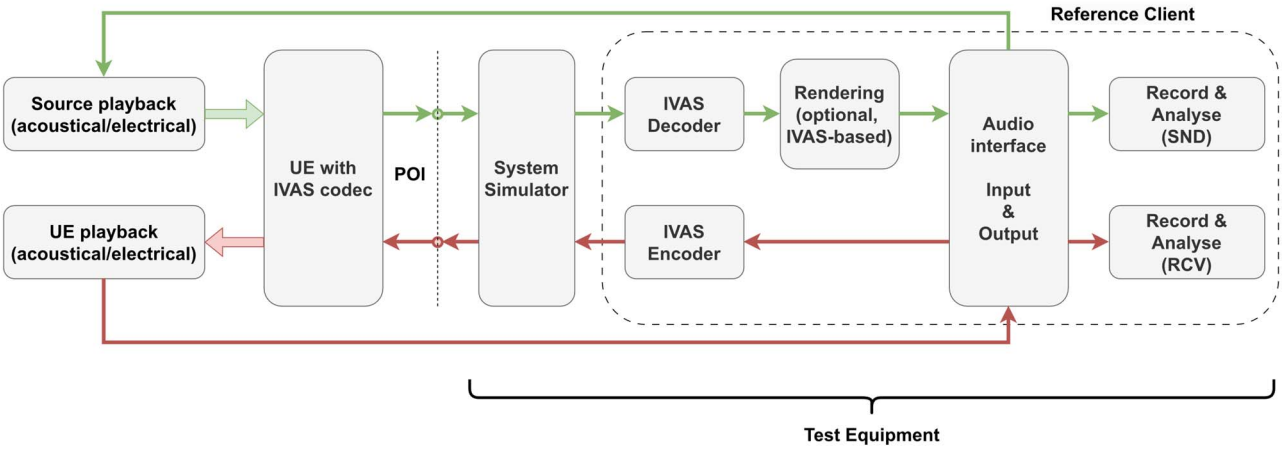


**Figure 6: General test setup with UE and reference client**

If any conversion from/to analogue signals are required by any component of the test equipment or for analysis purposes, an overload point of 3.14 dBm0 shall be used for all encoded or decoded signal channels (same definition as in 3GPP TS 26.132 [27] for mono signals). The relationship between levels measured in dBm0, dBV and dBov is described in clause 8 of ITU-T G.100.1 [9].

The details of the acoustical and electrical test setup as well as IVAS session parameters are UE-type-dependent and are given in the following clauses.

## 5.3 Test conditions

### 5.3.1 Test environment acoustic properties

The test environment acoustic properties shall meet the requirements in clause 4.0.3.

### 5.3.2 System simulator and reference client

The system simulator configuration and radio conditions on the air interface shall be as specified in clause 4.2 of 3GPP TS 26.132 [27]. Unless otherwise specified for the respective test, the system simulator shall provide an error-free radio connection to the UE under test.

The UE shall be connected to a system simulator and test equipment supporting the IVAS codec [28]. Unless specified otherwise, 48 kHz sampling rate and fullband mode shall be used.

Since UEs may provide different capabilities for sending (capture) and receiving (rendering) direction, a bidirectional communication may not be symmetric regarding IVAS Coded Formats and corresponding bitrates.

During negotiating and exchange of session parameters via Session Description Protocol (SDP), the UE advertises its supported and preferred Coded Formats in sending and receiving direction, as well as corresponding bitrate ranges.

- The reference client shall advertise via SDP the envisioned IVAS audio format for testing sending and receiving direction along with corresponding bitrate as specified in Table 1 as the preferred and only capabilities.
- For sending direction, the decoder in the reference client shall be configured for the EXT decoder output format, if not specified otherwise by any test method.
- For receiving direction, the encoder in the reference client shall be configured for the Coded Format envisioned for testing.

**Table 1: Bitrates per audio format used for testing**

| Coded Format | Subformat | Default Bitrate for testing… | Max. bitrate [kbit/s] |
|---|---|---|---|
| Stereo | Stereo | TBD | 256 |
| | Binaural | TBD | 256 |
| ISM | 1 | TBD | 128 |
| | 2 | TBD | 256 |
| | 3 | TBD | 384 |
| | 4 | TBD | 512 |
| SBA | FOA | TBD | 512 |
| | HOA2 | TBD | 512 |
| | HOA3 | TBD | 512 |
| MASA | 1 TC | TBD | 512 |
| | 2 TC | TBD | 512 |
| OSBA | | TBD | 512 |
| OMASA | | TBD | 512 |
| Multichannel | 5.1 | TBD | 512 |
| | 7.1 | TBD | 512 |
| | 5.1.2 | TBD | 512 |
| | 5.1.4 | TBD | 512 |
| | 7.1.4 | TBD | 512 |
| NOTE: The maximum bitrates are provided for information. | | | |

NOTE 1: Default bitrate configurations for testing are for further study.

NOTE 2: The maximum bitrates listed in Table 1 may not be supported by all UEs.

NOTE 3: The Coded Format ISAR (see [28]) is for further study.

The bitrate shall be provided by the network operator configuration. If the bitrate is not available from the operator, the maximum bitrate supported by the UE shall be used. The test operator shall report the bitrate used for testing.

### 5.3.3 Test equipment

The same requirements as specified in clause 4.0.2 apply for IVAS-based testing.

## 5.4 Test arrangement

### 5.4.1 Capture modes

To simulate a single sound source, either a loudspeaker or a HATS equipped with an artificial mouth (or equivalent stand-alone mouth simulator) positioned relative to the UE at a certain angle/distance is used. Two different types of use cases are distinguished:

1) User capture, where the sound source is always a HATS equipped with an artificial mouth and typically only speech (or speech-like) test signals are used.

2) Spatial capture, where the sound source is one or more loudspeakers positioned around the UE. Unless otherwise specified, the default sound source direction (for single source tests) is located at 0° azimuth and 0° elevation.

Both capture modes are illustrated in Figure 7. The applicable capture mode is part of the test arrangement selection process, which is described in clause 5.4.2.1. Some test methods may only be applicable for one or the other capture mode.

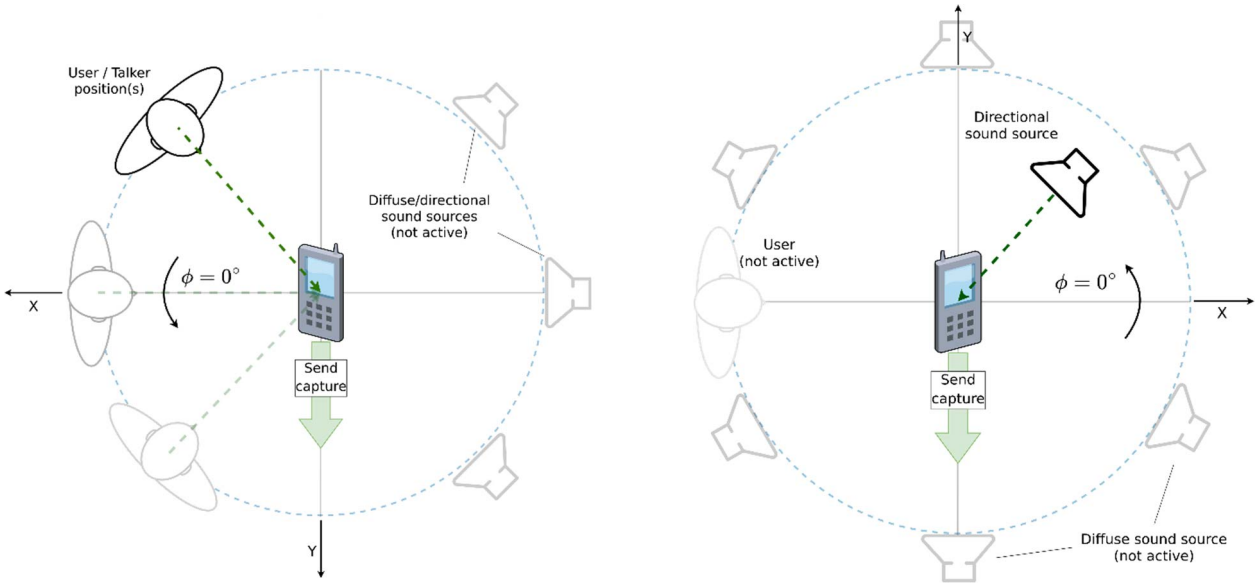


**Figure 7: Device orientation and sound sources for User Capture (left) and Spatial Capture (right)**

NOTE: Figure 7 is indicative and the illustrated example does not refer to any specific test arrangement or method. Diffuse sound fields for Spatial Capture are out of scope for IVAS-based UE testing in the present document.

## 5.4.2 UE types and positioning

### 5.4.2.1 Overview

An IVAS-enabled UE may support one or more UE types. Each UE type is composed of a *SND-UE-type* and a *RCV-UE-type*.

- The SND-UE-type is a combination of a certain audio capturing mode (which can be acoustic or electric) and a negotiated coded format (see clause 5.3.2).
- The RCV-UE-type is a combination of a negotiated coded format (see clause 5.3.2) and a certain audio playback mode (acoustic or electric).
- Each audio capturing/playback mode has a corresponding physical test arrangement.

NOTE 1: The definition of IVAS-enabled UE types is more complex than in 3GPP TS 26.132 [27]. Due to the variety of new applications that are enabled by the IVAS codec, the capture and playback Coded Formats and/or rendering configuration may not necessarily be the same in sending and receiving direction.

UEs may support multiple Coded Formats in sending and receiving direction, which are negotiated during call setup. At least one supported Coded Format shall be tested in all directions supported by the UE, which is selected according to the following priority:

1) Format specified by the manufacturer (if applicable).
2) Preference of the UE, as indicated during negotiation in SDP.
3) Test operator selects format based on form factor and envisioned use case of the UE.

The IVAS format for each tested direction shall be documented in the test report. Other available supported formats may be tested as well to ensure best-possible compatibility with other UE types.

Complementary to the well-defined IVAS formats, capturing/playback modes and corresponding interfaces are given in the following clauses along with several UE types, which may be applicable in SND and/or RCV. All UE types with acoustical interfaces assume that microphones and loudspeakers/headset of the UE are either integrated into the device

or that necessary additional off-the-shelf equipment (like e.g., headset, microphone array, loudspeaker array) is either bundled with the device or explicitly recommended by the manufacturer.

If no acoustical interface is available, the electrical interface shall be tested and the test setup according to clause 5.4.2.7 applies.

NOTE 2: If testing at the electrical interface is not possible for important technical reasons (e.g., due to a non-standardized electrical interface), an acoustic test may be carried out with suitable third-party equipment, which is, e.g., recommended by the manufacturer. This allows to perform tests at an acoustic interface to informally assess the performance of the UE (without applying requirements).

The physical test arrangement used for UE testing in sending and receiving direction is in general specified by the manufacturer by:

- Referencing one of the following clauses (recommended),
- Referencing a test arrangement from other standards (e.g., ITU-T P.340 [16] or P.341 [17]),
- Specifying an individual test arrangement.

In case no instructions on the test arrangement are provided by the manufacturer, the test operator shall select one based on the envisioned use case, form factor, etc. from either one of the following clauses or from other standards. If no suitable test arrangement can be identified for certain UEs with acoustical interface, the test operator may set up an individual arrangement or modify an existing one. In any case, the arrangement used for testing shall be documented in the test report.

If not specified otherwise or explicitly excluded in the test description, all methods for sending and receiving direction are applicable for all UE types and all Codec Formats (see Table 1).

### 5.4.2.2 Handset UE

The EVS-compatible mono mode of IVAS shall be tested according to clause 9 of TS 26.132 [27]. Requirements according to clause 7 of 3GPP TS 26.131 [26] apply.

Immersive testing for handset UE is not included in the present document due to the following reasons:

- RCV: Monaural listening typically cannot provide any directional spatial information.
- SND: The device is typically located close to the user's head, which can limit the capture of spatial information.

NOTE: Immersive audio formats may also be applicable to handset UEs, for example, to capture ambient sound at the near end. However, so far, test methods are not specified for handset UE and are for further study.

### 5.4.2.3 Headset UE

The test setup for headset UE for sending and receiving directions is shown in Figure 8. It applies to all devices that provide a head-worn acoustical frontend. The acoustical frontend may either be internal or external to the UE device. In the latter case, it may be connected via wired or wireless link (e.g., analogue jack, Bluetooth, or USB).

The head-worn acoustical frontend may provide headtracking data through a wired or wireless link (see Figure 8), which can be used for rendering binaural audio in receiving direction.

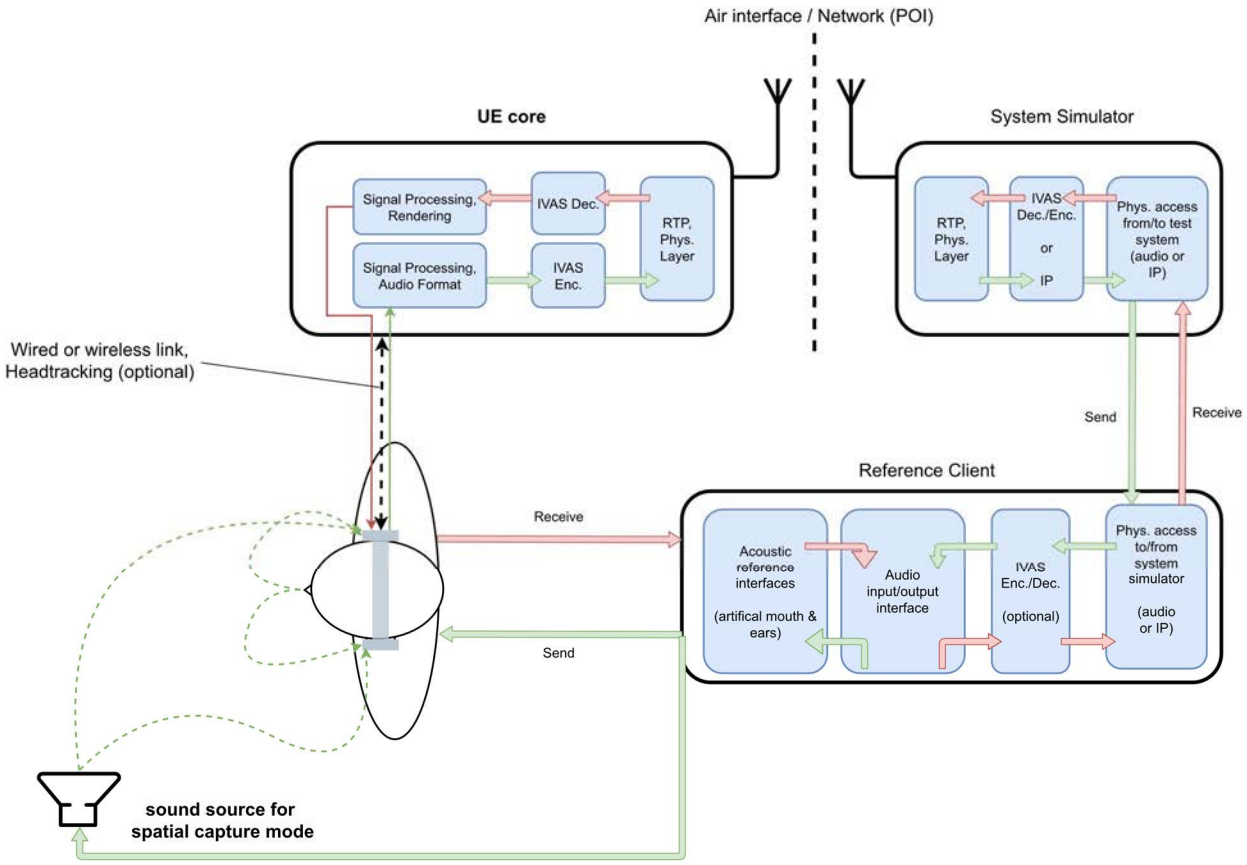


**Figure 8: Headset UE and test equipment**

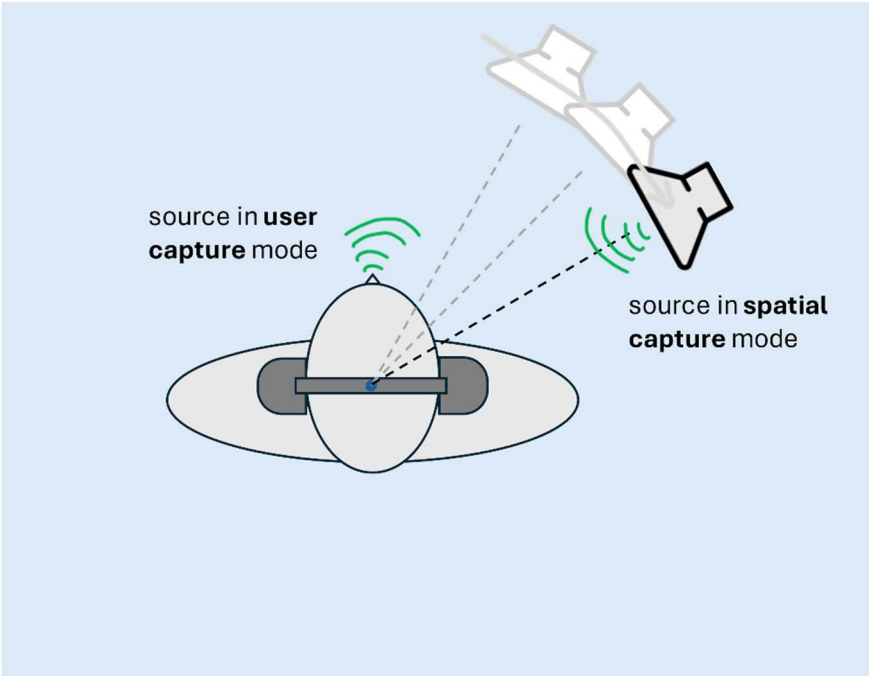


**Figure 9: Sound source positioning for Headset UE tests**

If not stated otherwise, headsets shall be placed on the HATS in their recommended wearing position. Sound source positioning for sending tests is illustrated in Figure 9. For the purposes of sound source rotation, the HRP of the HATS is considered the geometric center of the UE.

**User capture**

The sound source for User Capture is the artificial mouth of the HATS wearing the headset. The playback level at MRP shall be calibrated to -4.7 dB Pa.

**Spatial capture**

If not specified otherwise, the single sound source arrangement for Spatial Capture is a loudspeaker positioned at 0° azimuth and elevation and at 1 m relative to the HRP of the HATS wearing the headset device. The playback levels at HRP (in the absence of HATS) shall be calibrated to 75 dBSPL according to clause 5.5.1.

### 5.4.2.4 Handheld hands-free UE

The test setup for handheld hands-free UE for sending and receiving directions is shown in Figure 10. It applies to all devices that can be held in front of the user.

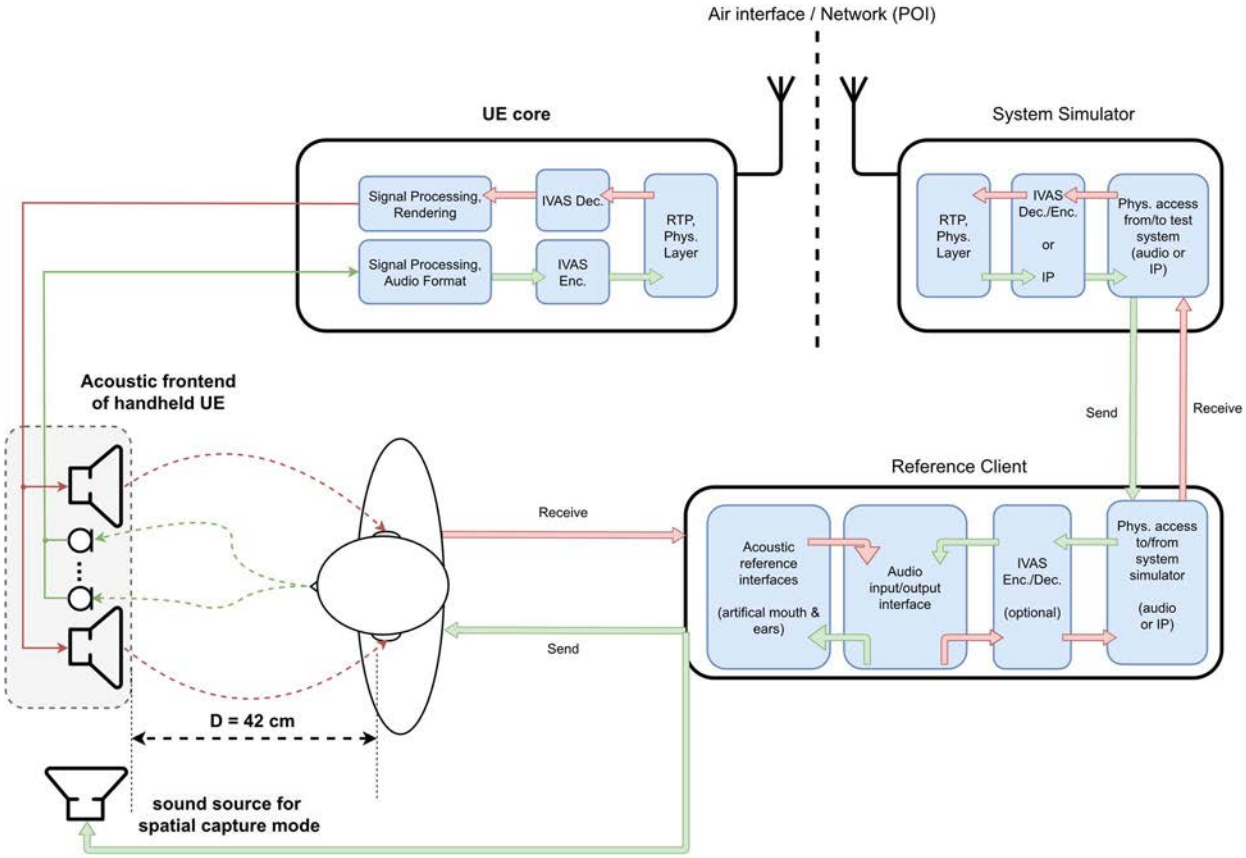


**Figure 10: Handheld hands-free UE and test equipment**

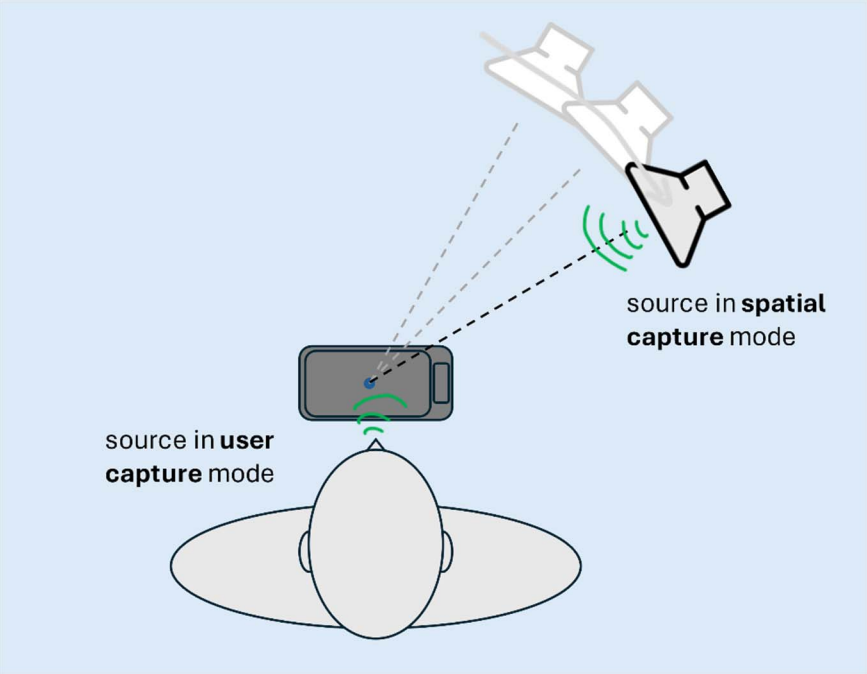


**Figure 11: Sound source positioning for Handheld UE tests**

The UE orientation (landscape/portrait and top/bottom position, front/back side facing the user) used for testing is specified by the manufacturer. If such information not available, the orientation depends on the capture mode.

The test fixture (e.g., microphone stand) used to mount the handheld hands-free UE for the measurements should be as acoustically transparent as possible and should not obstruct any of the input microphones.

If not specified otherwise, the arrangement for receiving direction is a HATS positioned at 42 cm from the center point of the visual display of the UE (same setup as in TS 26.132 [27] for handheld hands-free). Sound source positioning for sending tests is illustrated in Figure 11.

**User capture**

The sound source for User Capture is the artificial mouth of the HATS. The HATS is positioned the same way as in receiving direction. If applicable, different geometries of this setup are considered in corresponding test methods (for e.g., multi-talker scenarios or speech from certain angles). The playback levels at MRP shall be calibrated to -1.7 dB Pa.

If no manufacturer-defined orientations are defined, the UE shall be positioned in portrait mode, whereas the front side is facing the user.

**Spatial capture**

If not specified otherwise, the single sound source for Spatial Capture is a loudspeaker positioned at 0° azimuth and 0° elevation and at 1 m relative to the geometric center of the handheld hands-free UE. The playback levels at the UE shall be calibrated to 75 dBSPL according to the clause 5.5.1. If applicable, different geometries of this setup are considered in corresponding test methods (for e.g., multiple sources from different angles).

If no manufacturer-defined orientations are defined, the UE shall be positioned in landscape mode (top edge of the device turned to the left), whereas the front side is facing the user.

NOTE: Spatial capture typically targets at acoustic scenes opposite to the user of the device. Therefore, the simulated user/HATS is not considered as an integral part of the Spatial capture setup, as the acoustic impact of the simulated user/HATS in such a setup is for further study.

### 5.4.2.5 Table-mounted UE

The test setup for table-mounted hands-free UE for sending and receiving direction is shown in Figure 12. It applies to all hands-free devices that are intended for usage on tables (like e.g., conference devices). In contrast to handheld hands-free UE, the reflections of the table are explicitly included in the test setup.

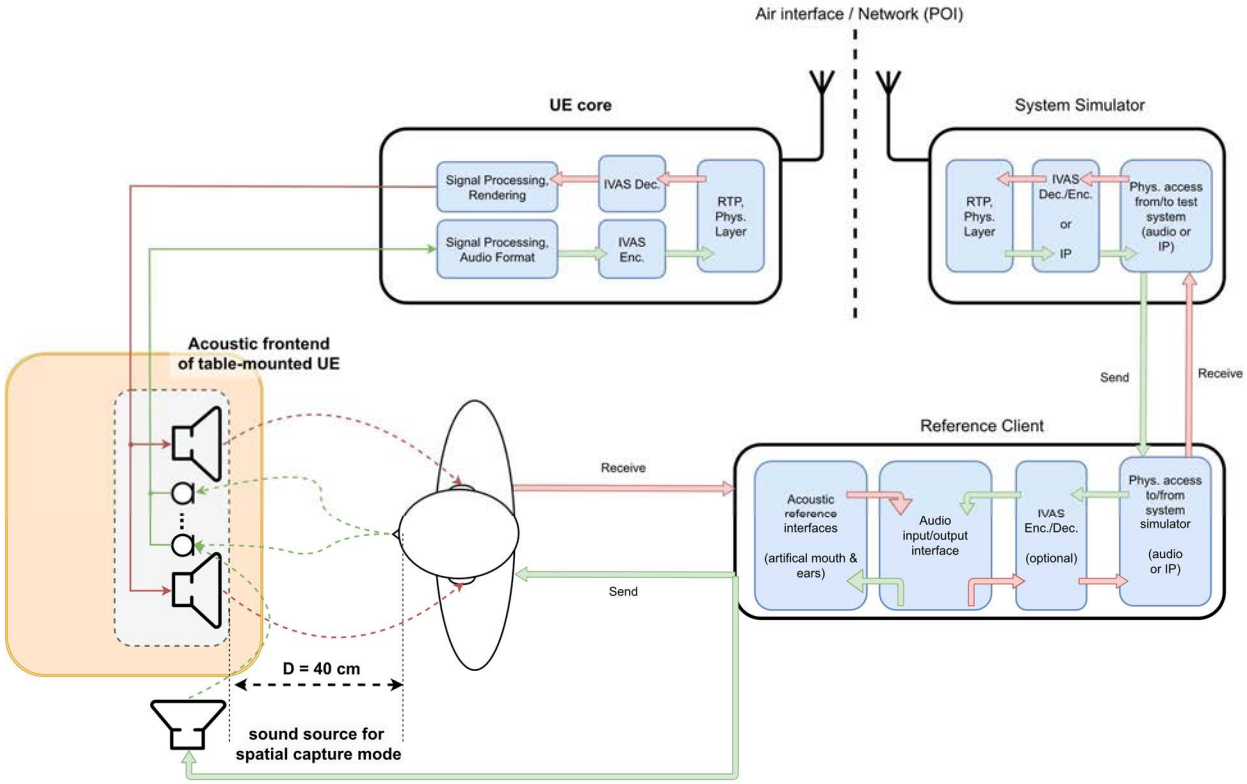


**Figure 12: Table-mounted hands-free UE and test equipment**

Figure 12 shows an example with a distance of D = 40 cm between front of the UE and lip reference plane of the user, which corresponds to the desktop hands-free setup as specified in Recommendation ITU-T P.341 [17], which is also referenced in 3GPP TS 26.132 (width W = 40 cm, height H = 30 cm). In general, multiple sub-setups may be considered for this UE type, like e.g., the "group audio terminal" position (see clause 4.2.4 of P.341 [17]) or the softphone/laptop-based setups 3GPP TS 26.132 [27].

NOTE: The term "table-mounted hands-free" is suggested here instead of "desktop hands-free", as used in e.g., 3GPP TS 26.132. The intention for this is to explicitly address also different/larger setups like e.g., conferencing scenarios with multiple microphones and loudspeaker arrays.

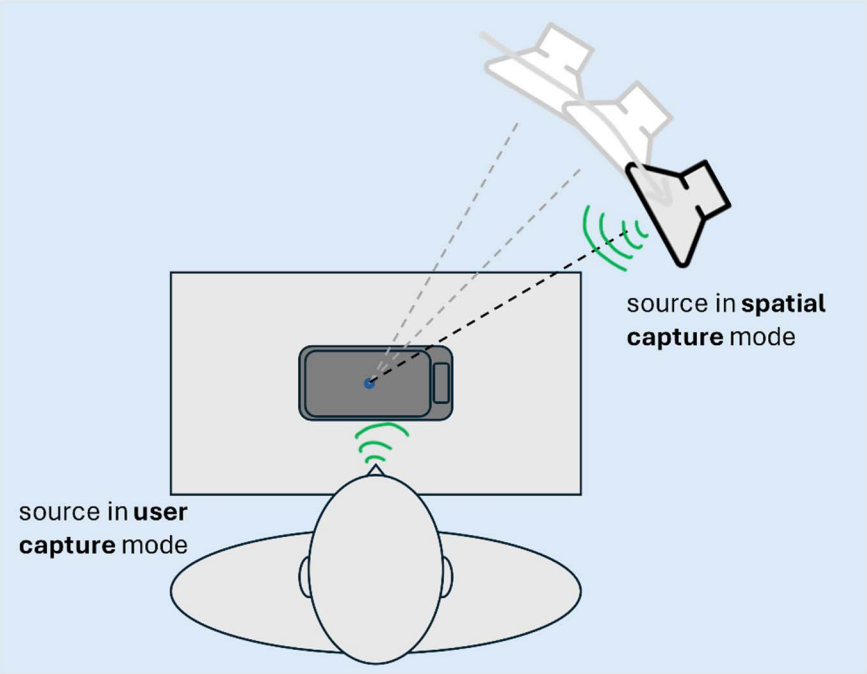


**Figure 13: Sound source positioning for Table-mounted UE tests**

If not specified otherwise, the arrangement for receiving direction is a HATS positioned at a distance of 40 cm between UE and lip reference plane and a height of 40 cm. Sound source positioning for sending tests is illustrated in Figure 13.

**User capture**

The sound source for User Capture is the artificial mouth of the HATS. The HATS is positioned the same way as in receiving direction. The playback levels at MRP shall be calibrated to -1.7 dB Pa.

**Spatial capture**

If not specified otherwise, the single sound source for Spatial Capture is a HATS or loudspeaker positioned at a height H = 40 cm above the table and at a width of 80 cm in front of the geometric center of the table-mounted UE. In spherical coordinate system, this corresponds to 0° azimuth, 26.6° elevation and 89.4 cm. The playback levels at the HFRP shall be calibrated to -24.7 dBPa.

### 5.4.2.6 Loudspeaker UE

The test setup for loudspeaker hands-free UE for receiving direction is shown in Figure 14. It applies to multichannel loudspeaker systems and speaker arrays, e.g., soundbars or automotive infotainment systems.

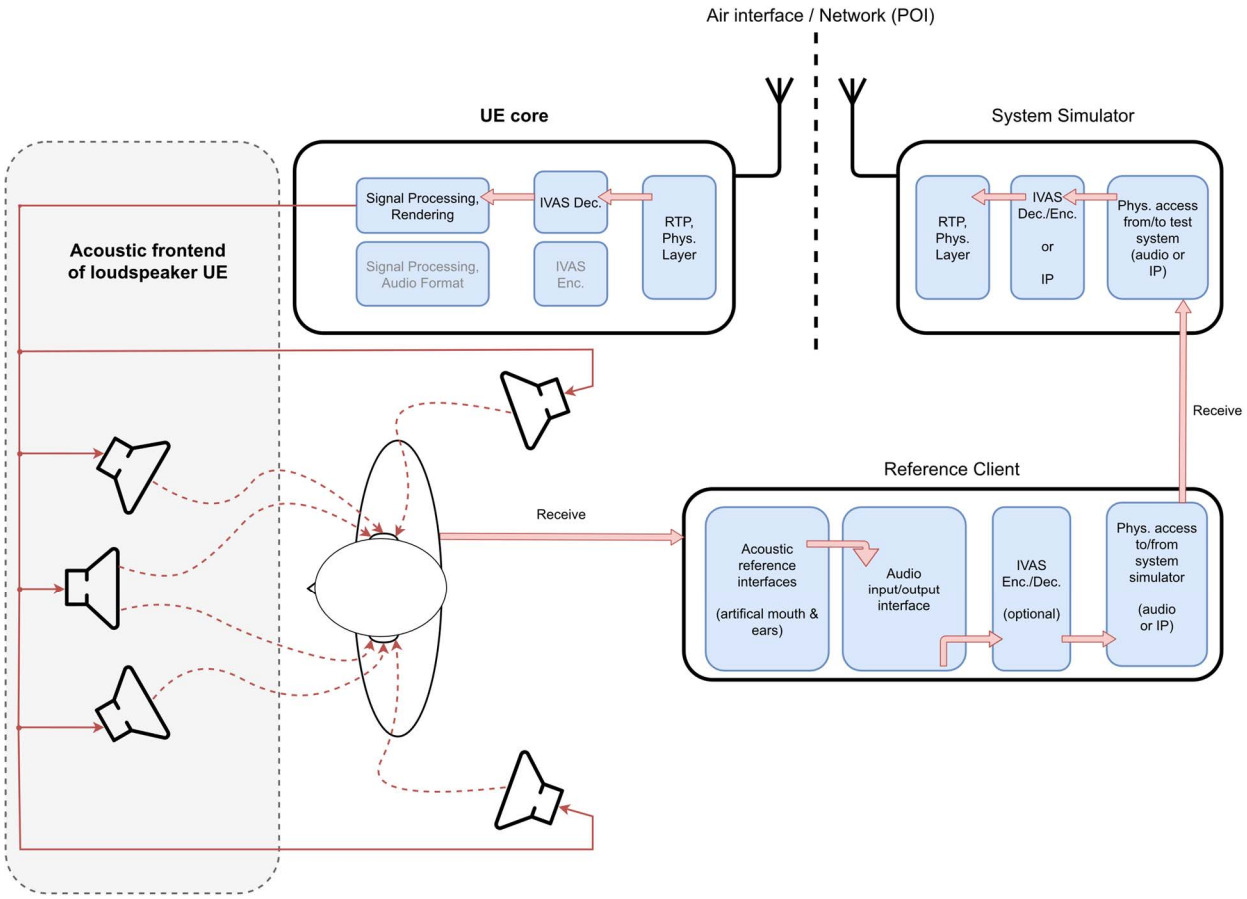


**Figure 14: Loudspeaker hands-free UE and test equipment**

### 5.4.2.7 Electrical interface UE

The test setup for electrical interface UE for sending and receiving directions is shown in Figure 15. It applies to all devices that do not provide integrated or associated equipment for capturing and/or reproduction of immersive audio. Wired or wireless digital audio interfaces according to Recommendation ITU-T P.383 [21] (e.g., Bluetooth or USB) are commonly used. Note that the interface may also be realized via an analogue jack plug, which provides up to two channels in receiving and sending direction (see Recommendations ITU-T P.381 [19] and ITU-T P.382 [20]).

The device may provide an additional input for head tracking data that can be used for rendering the receiving direction (e.g., Bluetooth or USB with HID profile [33]).

Different equipment may be connected to the electrical interface UE such that the combination of UE and additional equipment will behave like one of previous UE types (e.g., headset or loudspeaker). Test methods apply according to the envisioned use-case. The default test signal for electrical insertion of a single sound source shall correspond to the envisioned use case.

EXAMPLE: If the electrical interface of an UE is envisioned to connect a third-party immersive headset (e.g., with headtracking functionalities), the default test signal contains a virtual single sound source at the default source position specified for headset UE (see clause 5.4.2.3). Only test methods for headset UE are applicable in this case.

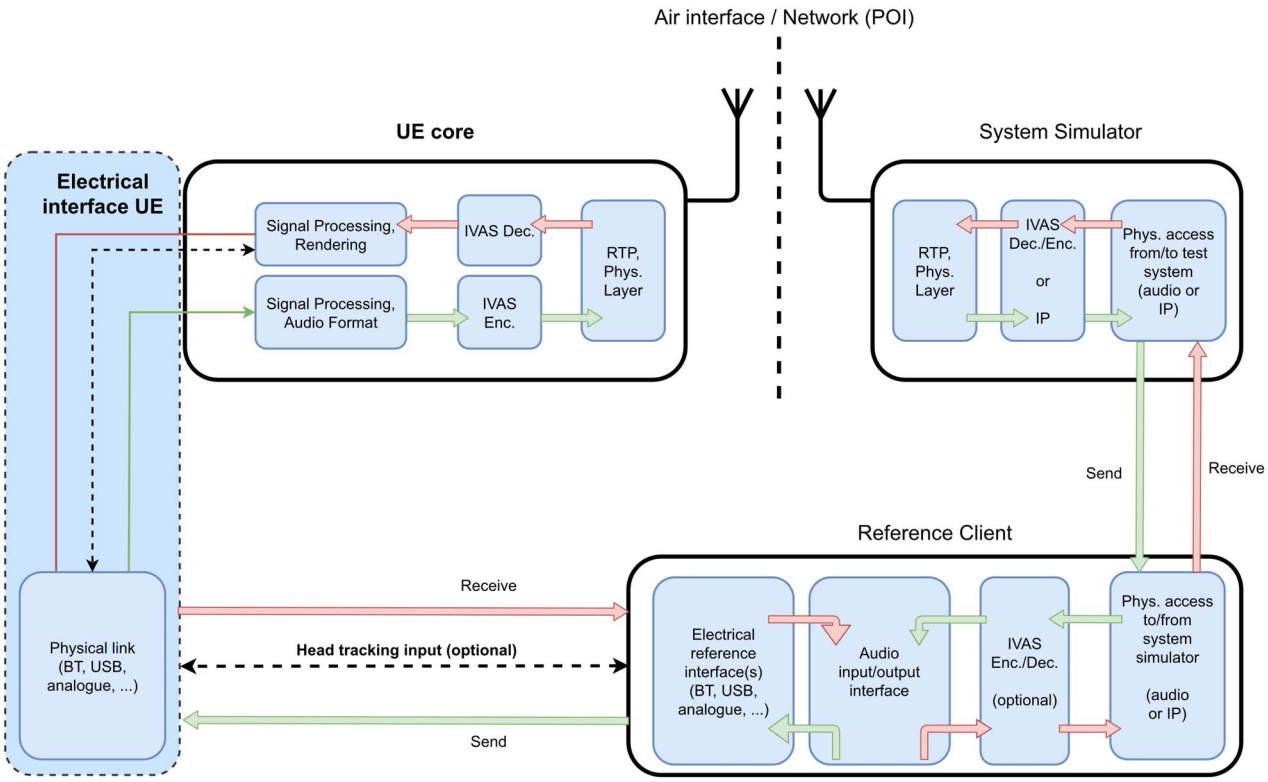


**Figure 15: Electrical interface UE and test equipment**

### 5.4.3 UE configuration

For testing, the UE shall be configured for the relevant and/or envisioned use cases as described in clause 5.4.2.

UEs shall be tested *as is*, i.e., without the test operator enabling/disabling signal enhancement features or functionalities after the UE initialization.

NOTE: If the test operator observes corrupted results and/or low UE performance for certain measurements, the tests may be repeated with such modifications (enabling/disabling signal enhancement features or functionalities) for the purposes of tracking and documenting possible root causes of the problem.

Unless stated otherwise, if a volume control is provided in receiving direction, the setting is chosen such that the nominal receiving loudness is met as closely as possible.

## 5.5 Test signals

### 5.5.1 Test signal calibration

The input signal levels for testing receiving direction or testing with an electrical interface in sending direction shall be calibrated by means of a calibration factor, which is determined according to the iterative procedure illustrated in Figure 16. The rendering of the input signal to the target format shall be done via external IVAS reference renderer, which is defined in clause 6 of 3GPP TS 26.254 [30]. The rendered output format shall be set to 7.1+4 output, except if the format of the input signal is already multichannel or stereo. The level is then calculated according to ITU-R BS.1770 [23] for these intermediate signals. If the difference between measured and target level exceeds 0.5 LKFS the procedure is repeated for up to ten iterations.

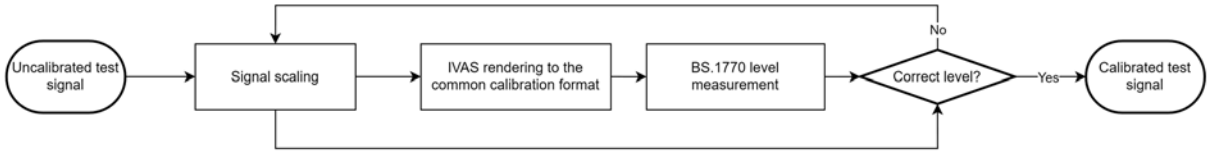


**Figure 16: Test signal input level adjustment procedure for receiving tests**

For mono real speech signals, acoustical and electrical calibration of test signals shall be carried out with active speech level according to ITU-T P.56 [10], calculated over the complete test sequence.

## 5.5.2 Virtual positioning

For testing of a single, directional sound source in receiving direction and sending direction with an electrical interface, the test signal shall be positioned virtually to represent the tested source direction.

**Object-based audio**

Unless specified otherwise, the source position of the test signal shall be set by format specific metadata as defined in Table 2. No further metadata fields shall be set.

**Table 2: Object-based audio format specific metadata**

| Azimuth | Elevation | Spread | Gain |
|---|---|---|---|
| According to the tested azimuth | According to the tested elevation | 0 | 1 |

**Scene-based audio**

Unless specified otherwise, the source position of the test signal shall be set by a multi-component Ambisonics (ACN/SN3D) signal that represents a source from the particular incidence angle. To generate the test signal, first the virtually positioned object-based format is created. Then the IVAS external renderer according to 3GPP TS 26.254 [30] is used to obtain the desired signal in scene-based audio output format from the object-based input.

**Metadata-assisted spatial audio**

Unless specified otherwise, the source position of the test signal shall be set by the format specific metadata. The applied descriptive metadata for every frame and the applied spatial metadata for every time-frequency tile shall be set as defined in Table 3 and Table 4. The fields of the MASA metadata format are specified in Annex A of 3GPP TS 26.258 [31]. Unless specified otherwise, the source signal shall be applied to each transport channel used for testing.

**Table 3: Applied descriptive metadata**

| MASA format descriptive common metadata parameters | Assigned values for every metadata frame |
|---|---|
| Format descriptor | Default |
| Number of directions | 1 (bit value 0) |
| Number of channels | 1 or 2 (bit value 0 or 1), depending on the number of applied transport channels |
| Source format | Bit values 00 (Default/unknown) |
| Variable description | 12 bit zero-padding (Default/unknown) |

**Table 4: Applied spatial metadata**

| MASA format spatial metadata parameters | Assigned values for every time-frequency tile in all MASA metadata sub-frames |
|---|---|
| Direction Index | According to the tested azimuth and elevation |
| Direct-to-total energy ratio | 1.0 |
| Spread coherence | 0.0 |
| Diffuse-to-total energy ratio | 0.0 |
| Surround coherence | 0.0 |
| Remainder-to-total energy ratio | 0.0 |

**Multichannel audio**

To generate the test signal, first the virtually positioned object-based format is created. Then the IVAS external renderer according to 3GPP TS 26.254 [30] is used to obtain the desired signal in multichannel audio output format from the object-based input.

# 5.6 Test methods for sending direction

## 5.6.1 Delay

The UE delay testing in sending direction, $T_S$, is illustrated in Figure 17. The reference client decodes and renders the input bitstream to a common analysis-dependent format. For the rendering step, the IVAS renderer [30] shall be used.

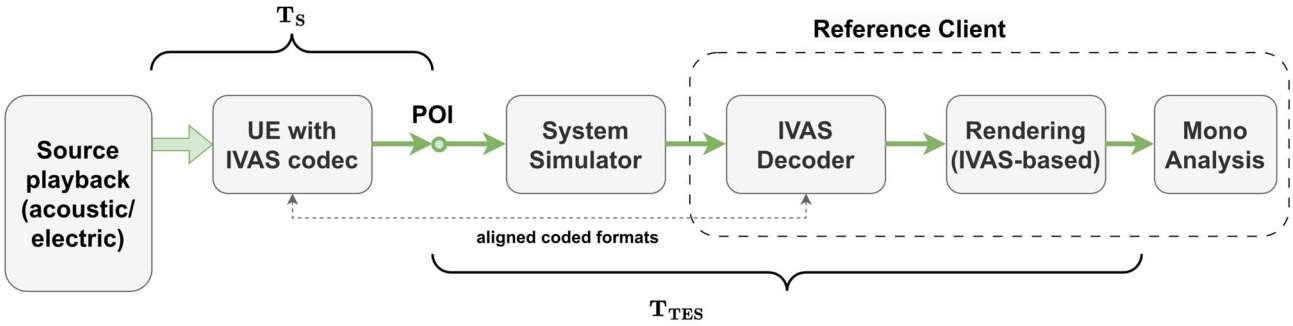


**Figure 17: Test setup for sending loudness and delay**

**Test method**

1) The default arrangement for a single sound source as defined in clause 5.4.2 is set up according to the UE type. The UE under test and the reference client are connected and configured as described in the clause 5.3.2.

2) The test signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [14], which is calibrated as specified in clause 5.4.2 for the corresponding UE type and the tested capture mode (user or spatial).

3) For acoustic measurements, the UE is mounted as described in the clause 5.4.2. The sound source corresponding to the tested capture mode (user or spatial) is positioned accordingly. The test signal is played back via the acoustic source.
For measurements with electrical interface UE, the test signal virtually positioned as described in clause 5.4.2.7. The generated source signal is played back via insertion into the electrical interface.

4) The renderer in the reference client shall be configured to mono output.

5) The UE delay in sending direction $T_S$ is obtained between the acoustical sound source and the electrical POI of the test equipment. The source signal is used as a reference for the cross-correlation analysis described in Annex C, which is used to determine $T_S$.

6) The measured delay shall be compensated by the delay $T_{TES}$ introduced by the test equipment (including possible contributions of the rendering to mono).

## 5.6.2 Loudness

The UE delay testing in sending direction is illustrated in Figure 17. The reference client decodes and renders the input bitstream to a common analysis-dependent format. For the rendering step, the IVAS renderer [30] shall be used.

**Test method**

1) The default arrangement for a single sound source as defined in clause 5.4.2 is set up according to the UE type. The UE under test and the reference client are connected and configured as described in the clause 5.3.2.

2) The test signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [14], which is calibrated as specified in clause 5.4.2 for the corresponding UE type and the tested capture mode (user or spatial).

3) For acoustic measurements, the UE is mounted as described in clause 5.4.2. The sound source corresponding to the tested capture mode (user or spatial) is positioned accordingly. The test signal is played back via the acoustic source.
For measurements with electrical interface UE, the test signal virtually positioned as described in clause 5.4.2.7. The generated source signal is played back via insertion into the electrical interface.

4) The renderer in the reference client shall be configured to mono output.

5) The UE loudness in sending direction is obtained by Send Loudness Rating (SLR) according to ITU-T P.79 [13] in wideband mode. The calculation is carried out as specified in clause 8.2.3.1 of 3GPP TS 26.132 [27], except that the active speech level of the reference signal shall be calibrated to -1.7 dBPa.

## 5.6.3 Frequency response (single source)

### 5.6.3.1 Test method

The following procedure shall be used:

1) The default arrangement for a single sound source as defined in clause 5.4.2 is set up according to the UE type. The UE under test and the reference client are connected and configured as described in the clause 5.3.2.

2) The test signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [14], which is calibrated as specified in clause 5.4.2 for the corresponding UE type and the tested capture mode (user or spatial).

3) For acoustic measurements, the UE is mounted as described in clause 5.4.2. The sound source corresponding to the tested capture mode (user or spatial) is positioned accordingly. The test signal is played back via the acoustic source.
For measurements with electrical interface UE, the test signal virtually positioned as described in clause 5.4.2.7. The generated source signal is played back via insertion into the electrical interface.

4) The frequency response of the decoded output signal shall be calculated according to the format-specific definitions of the following clause. For all formats, 1/12th octave intervals as given by the R40 series of preferred numbers in ISO 3 [6], for frequencies from 100 Hz to 12 kHz (inclusive) apply. The reference magnitude spectrum $P_{ref}(f)$ is the 1/12$^{th}$ octave spectrum of the test signal.

### 5.6.3.2 IVAS format-specific definitions

**Stereo & Object-based audio**

The frequency response of a stereo or object-based audio signal is defined as a ratio of the magnitude spectrum $\hat{P}_k(f)$ of each audio channel *k* and the reference magnitude spectrum $P_{ref}(f)$. Thus, for each stereo audio signal channel *k*, the frequency response is determined by:

$$G_k(f) = \frac{\hat{P}_k(f)}{P_{ref}(f)}$$

**Scene-based audio**

The magnitude spectrum $\hat{P}(f)$ of a scene-based audio signal is determined by evaluating the root mean square of the magnitude spectrum of the $(N+1)^2$ Ambisonics coefficients $\hat{P}_l^m(f)$:

$$\hat{P}(f) = \sqrt{\frac{1}{(N+1)^2}\sum_{l=0}^{N}\sum_{m=-l}^{l}(2l+1)\left(\hat{P}_l^m(f)\right)^2}$$

Letters $l$ and $m$ respectively denote Ambisonics degree and index. The factor $\sqrt{2l+1}$ renormalizes the SN3D-normalized coefficients to N3D. The N3D-normalized root-mean-square of the coefficient-domain magnitude spectrum is equivalent of determining the root-mean-square in the spatial domain, evaluated on a uniform sampling grid (Parseval's theorem).

The frequency response is defined as the ratio of $\hat{P}(f)$ and the reference magnitude spectrum $P_{ref}(f)$. The frequency response is determined by:

$$G(f) = \frac{P(f)}{P_{ref}(f)}$$

**Metadata-assisted spatial audio**

The frequency response of a metadata-assisted spatial audio signal is defined as a ratio of the magnitude spectrum $\hat{P}_k(f)$ of each transport channel *k* nd the reference magnitude spectrum $P_{ref}(f)$. Thus, for each MASA transport channel *k*, the frequency response is determined by:

$$G_k(f) = \frac{\hat{P}_k(f)}{P_{ref}(f)}$$

## 5.6.4 Directional information (single source)

### 5.6.4.1 Test method

This test method applies to the spatial capture mode.

The default arrangement for a single sound source is used, as defined in clause 5.4.2 for each UE type. The decoded and rendered output format shall be the same as the IVAS audio format used by the UE. $L = 7$ directions shall be evaluated, as indicated in Table 5. In case the setup is realized with a turntable, the device shall be rotated around the vertical center of the UE.

**Table 5: Additional source directions**

| *i* | $\phi_i$ [°] | $\theta_i$ [°] |
|---|---|---|
| 1 | -90 | 0 |
| 2 | -60 | 0 |
| 3 | -30 | 0 |
| 4 | 0 | 0 |
| 5 | 30 | 0 |
| 6 | 60 | 0 |
| 7 | 90 | 0 |

For each sound source direction ($\phi_i$, $\theta_i$) and $i$ = 1, ..., $L$, the following procedure shall be used:

1) The test signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [14], calibrated to an active speech level according to clause 5.5.1.

2) The UE under test and the reference client are connected and configured as described in the clause 5.3.2.

3) Acoustical interface: The UE is mounted as described in clause 5.4.2 and the acoustic source is positioned such that the source direction under test is met. The test signal is played via the acoustic source.

   Electrical interface: The test signal is generated by virtually placing the acoustic source such that the source direction under test is met as described in clause 5.4.2.7.

4) The output format-dependent directional metric calculations shall be done as defined in the following clause for the tested format.

### 5.6.4.2 IVAS format specific definitions

**Stereo**

The estimated source directions in the stereo panorama $\zeta(\phi)$ shall be calculated as follows:

1) The left and right channel signals ($s_L(k)$, $s_R(k)$) are recorded by the test equipment.

2) The inter-channel time difference ($\mathrm{ICTD}(\phi_i) = \Delta_\tau$) is determined as described in Annex C between the signals $s_L(k)$ and $s_R(k)$.

3) The inter-channel level difference (ICLD) is determined as follows:

   a) The active speech level according to Recommendation P.56 [10] is calculated separately for $s_l(k)$ and $s_r(k)$, resulting in $L_L$ and $L_R$.

   b) The ICLD for source direction $\phi_i$ is calculated as:

$$\mathrm{ICLD}(\phi_i) = L_L - L_R$$

4) The equivalent level difference $\Delta(\phi_i)$ for source direction $\phi_i$ is calculated as:

$$\Delta(\phi_i) = \mathrm{ICLD}(\phi_i) + \mathrm{ICTD}(\phi_i) \cdot 17.3 \frac{\mathrm{dB}}{\mathrm{ms}} \quad [\mathrm{dB}]$$

5) The estimated source directions in the stereo panorama $\zeta(\phi_i)$ is calculated according to [34] as:

$$\zeta(\phi_i) = 100 \cdot \begin{cases} \frac{\Delta(\phi_i)}{13.5} & \text{for} \quad |\Delta| \le 6.75 \text{ dB} \\ \left(-32\left|\frac{\Delta(\phi_i)}{\mathrm{dB}}\right|^3 + 288\left|\frac{\Delta(\phi_i)}{\mathrm{dB}}\right|^2 + 20736\left|\frac{\Delta(\phi_i)}{\mathrm{dB}}\right| - 6561\right) \cdot \frac{\operatorname{sign}\Delta(\phi_i)}{273375} & \text{for} \quad 6.75 \text{ dB} < |\Delta| < 18 \text{ dB} \\ \operatorname{sign}\Delta(\phi_i) & \text{for} \quad |\Delta| \ge 18 \text{ dB} \end{cases} \quad [\%]$$

**Scene-based audio**

Direction of arrival analysis with scene-based audio shall be conducted as follows:

1) From the SBA signal decoded by the reference client, only the first four Ambisonics component channels (FOA) in time-domain B-format representation are used, i.e., higher order Ambisonics components are discarded. These four FOA components are referred to as $W$, $Y$, $Z$, $X$ in ACN notation.

2) Each channel is segmented into time-domain frames of 20 ms length. The frames are referred to by frame index $m$, leading to a sequence of frames $w_m$, $y_m$, $z_m$, $x_m$.

3) Intensity parameters $i_x$, $i_y$, $i_z$ for the three catesian directions are calculated as follows:

$$\begin{pmatrix} i_x \\ i_y \\ i_z \end{pmatrix} = -\sum_m \mathrm{g}(w_m) \begin{pmatrix} x'_m \\ y'_m \\ z'_m \end{pmatrix} w_m.$$

   The gating function $\mathrm{g}(w_m)$ equals 1, if the input frame $w_m$ to the function is larger than -48 dBov; otherwise, it equals 0.

4) The direction of arrival estimation $(\hat{\phi}_i, \hat{\theta}_i)$ is calculated based on the intensity parameter using the equations:

$$\hat{\phi}_i = \text{atan2}(i_y, i_x) \cdot \frac{180^\circ}{\pi},$$

$$\hat{\theta}_i = \text{atan2}(i_z, \sqrt{i_x^2 + i_y^2}) \cdot \frac{180^\circ}{\pi}$$

The traditional arctangent function (arctan(x/y)) is not in itself sufficient to uniquely resolve the correct quadrant, thus the two-argument version "atan2" is used:

$$\text{atan2}(y, x) = \begin{cases} \arctan\left(\frac{y}{x}\right) & \text{if } x > 0, \\ \frac{\pi}{2} - \arctan\left(\frac{x}{y}\right) & \text{if } y > 0, \\ -\frac{\pi}{2} - \arctan\left(\frac{x}{y}\right) & \text{if } y < 0, \\ \arctan\left(\frac{y}{x}\right) \pm \pi & \text{if } x < 0, \\ \text{undefined} & \text{if } x = 0 \text{ and } y = 0. \end{cases}$$

5) The estimated sound source direction of arrival $(\hat{\phi}_i, \hat{\theta}_i)$ is compared to the direction of the sound source ($\phi_i$, $\theta_i$).

**Metadata-assisted spatial audio**

Direction of arrival analysis with metadata-assisted spatial audio shall be done as follows:

1) The decoded MASA signal is captured by the test equipment. The decoded MASA metadata includes estimated direction indices and direct-to-total energy ratio quantities $r_{dir}(k, n, m)$ for each subframe $n$ and MASA spatial metadata frequency band $k$ in every metadata frame $m$. Direction indices are further converted to the direction angular values $(\phi'(k, n, m), \theta'(k, n, m))$ as described in clause 5.5.2.2 of 3GPP TS 26.253 [29]. The energy $E(k, n, m)$ of the transport channel(s) is estimed for each subframe and MASA spatial metadata frequency band in every frame. The MASA format time-frequency resolution and metadata parameters are further described in 3GPP TS 26.258 Annex A [31].

2) Estimated direction angular values (in radians) are mapped into Cartesian coordinate vectors $X$, $Y$ and $Z$ over all subframes $N = 4$ in every frame $M$ and MASA spatial metadata frequency bands $K = 20$ as:

$$X = \sum_m^M \sum_n^4 \sum_k^{20} (\cos(\phi'(k, n, m)) \cdot \cos(\theta'(k, n, m)) \cdot r_{dir}(k, n, m) \cdot E(k, n, m))$$

$$Y = \sum_m^M \sum_n^4 \sum_k^{20} \left(\sin(\phi'(k, n, m)) \cdot \cos(\theta'(k, n, m)) \cdot r_{dir}(k, n, m) \cdot E(k, n, m)\right),$$

$$Z = \sum_m^M \sum_n^4 \sum_k^{20} (\sin(\theta'(k, n, m)) \cdot r_{dir}(k, n, m) \cdot E(k, n, m))),$$

where $k$ is the index of the MASA spatial frequency band, $n$ is the index of the subframe and $m$ is the index of the frame.

3) The sound source direction estimation $(\hat{\phi}_i, \hat{\theta}_i)$ in degrees is calculated based on the mapped Cartesian coordinate vectors using the equations:

$$\hat{\phi}_i = \arctan(Y, X) \cdot \frac{180^\circ}{\pi},$$

$$\hat{\theta}_i = \arctan(Z, \sqrt{X^2 + Y^2}) \cdot \frac{180^\circ}{\pi}$$

See previous definition of the function "atan2".

4) The estimated sound source direction $(\hat{\phi}_i, \hat{\theta}_i)$ is compared to the direction of the sound source $(\theta_i, \varphi_i)$.

# 5.7 Test methods for receiving direction

## 5.7.1 Delay

**Test method**

The assessment of receiving UE delay is illustrated in Figure 18. The default arrangement (acoustical or electrical interface) in receiving direction as defined in clause 5.4.2 for each UE type is used.

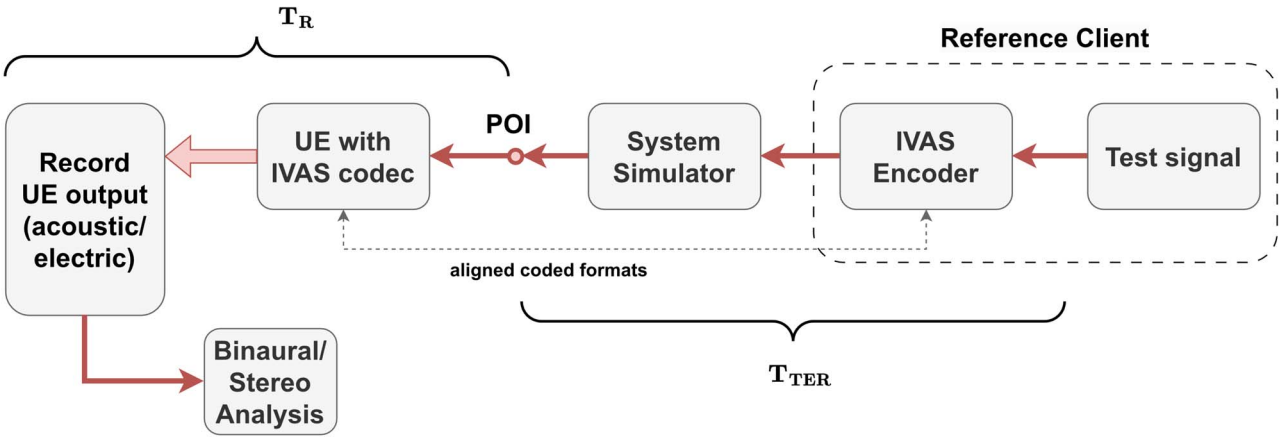


**Figure 18: Test setup for receiving delay**

1) The source signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [14].

2) Virtual source positioning according to clause 5.5.2 is used to generate the test signal from the source signal.

3) The test signal is calibrated to a loudness of -26 LKFS as defined in clause 5.5.1.

4) The UE is setup according to clauses 5.4.2 and 5.3.2. The UE and the reference client connection is setup according to clause 5.4.2, the source signal is encoded by the reference client, and inserted at the POI to the UE.

5) The capture of the UE output is carried out via:

   a) acoustical interface (headphones or loudspeakers): recording via diffuse-field equalized HATS.

   b) electrical interface: recording via corresponding reference interface. If the captured audio format is not stereo or binaural, the default IVAS binaural renderer is used to generate a binaural signal. To calibrate the digital signal into the (pseudo-)acoustical domain, it is assumed that a level of -26 dBov corresponds to an acoustical level of 73 dBSPL (-21 dBPa).

6) The UE delay in receiving direction is obtained between the electrical POI of the test equipment and the recorded signals at both ears. The cross-correlation analysis described in Annex C is carried out for both ear signals, using the mono source signal as reference. To obtain the overall delay $T_R$, the results for left and right ears are averaged.

7) The measured delay shall be compensated by the delay $T_{TER}$ introduced by the test equipment. If applicable, also the binaural renderer for the electrical interface is compensated.

## 5.7.2 Loudness

### 5.7.2.1 Test method

1) The source signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [14].

2) Virtual source positioning according to clause 5.5.2 is used to generate the test signal from the source signal.

3) The test signal is calibrated to a loudness of -26 LKFS as defined in clause 5.5.1.

4) The UE is setup according to the clause 5.4.2. The UE and the reference client connection is setup according to clause 5.3.2, the source signal is encoded by the reference client, and inserted at the POI to the UE.

5) The capture of the UE output is carried out via …

   a) acoustical interface (headphones or loudspeakers): recording via diffuse-field equalized HATS.

   b) electrical interface: recording via corresponding reference interface.
   If the captured audio format is not stereo or binaural, the default IVAS binaural renderer is used to generate a binaural signal. To calibrate the digital signal into the (pseudo-)acoustical domain, it is assumed that a level of -26 dBov corresponds to an acoustical level of 73 dBSPL (-21 dBPa).

6) The LLR in phon is calculated according to clause 8.3.3 of Recommendation ITU-T P.700 [22] with the captured or rendered binaural signal.

7) The same binaural signal should be used to calculate Receiving Loudness Rating (RLR) according to Recommendation ITU-T P.79 [13] for comparison to 3GPP TS 26.131 [26].

   a) The inverse diffuse-field correction according to Annex A of Recommendation ITU-T P.58 [12] is applied on left and right channel of the recording to obtain the signal at DRP. Then DRP-to-ERP correction is applied.

   b) The reference signal used for the RLR calculation is the original test signal specified in step 1), calibrated to -16 dBm0.

   c) The RLR is calculated according to clause 8.2.3.2 of 3GPP TS 26.132 [27].

### 5.7.2.2 IVAS format specific definitions

**Stereo**

The test signal shall be used for both stereo channels.

## 5.7.3 Frequency response (single source)

**Table 6: Source directions for assessing for sensitivity/frequency characteristics**

| Source azimuth [°] | Source elevation [°] |
|---|---|
| 0 | 0 |
| 180 | 0 |
| 0 | 90 |
| 90 | 0 |
| -90 (270) | 0 |

For each tested sound source direction listed in Table 6, the following procedure shall be used:

1) The source signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [14].

2) Virtual source positioning according to clause 5.5.2 is used to generate the test signal from the source signal.

3) The test signal is calibrated to a loudness of -26 LKFS as defined in clause 5.5.1.

4) The UE and the reference client are setup according to clauses 5.4.2 and 5.3.2, the source signal is encoded by the reference client, and inserted at the POI to the UE.

5) The sensitivity/frequency characteristics are measured as described in TS 26.132 [27] and are reported for the left and the right ears/channels.

The sensitivity/frequency characteristics may in addition be measured and reported for other positions.

## 5.7.4 Interaural differences for binaural rendering

The test method only applies to headset UE or electrical interface UE (if intended for headset usage).

The test method evaluates binaural cues for test signals from different source directions as given in Table 8. If the UE supports head tracking, the test is performed for the reference direction with no HATS rotation and additionally for rotated HATS orientations. To apply requirements independently of the head rotation, the source directions in Table 8 are stated relative to the head orientation. Hence, the overall source direction is composed of the stated direction plus the head orientation. For the electrical interface test, the orientation information shall be passed to the electrical interface. For the acoustical test, HATS rotation shall be realized as described in clause 5.3.3.

**Table 7: Source directions for assessing interaural time and level differences (relative to head orientation)**

| *i* | Source azimuth $\phi_i$ [°] | Source elevation $\theta_i$ [°] |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 180 | 0 |
| 3 | 0 | 90 |
| 4 | 90 | 0 |
| 5 | -90 (270) | 0 |

For each tested sound source direction, the following procedure shall be used:

1) The source signal to be used for the measurements shall be the British-English single talk sequence described in clause 7.3.2 of Recommendation ITU-T P.501 [14].

2) If the UE supports head-tracking, the horizontal plane HATS orientations $\tilde{\varphi}_j \in \{0°, -30°, +30°\}$ are tested. Otherwise, only the default HATS orientation ($\tilde{\varphi}_j = 0°$) is tested.

3) For each simulated source direction ($\phi_i$, $\theta_i$) as listed in Table 7 and each HATS orientation $\tilde{\varphi}_j$ under test, the following procedure is repeated:

   a) The HATS is oriented to $\tilde{\varphi}_j$.

   b) Virtual source positioning according to clause 5.5.2 is used to generate the test signal from the source signal. The source direction ($\phi_i + \tilde{\varphi}_j$, $\theta_i$) is applied.

   c) The test signal is calibrated to a loudness of -26 LKFS as defined in clause 5.5.1.

   d) The UE and the reference client are setup according to clauses 5.4.2 and 5.3.2, the test signal is encoded by the reference client, and inserted at the POI to the UE.

   e) The left and right signals from the UE are captured electrically or acoustically.

   f) The ILD is defined as the difference between right and left levels (in dB) and is calculated for octave bands with center frequencies 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz and 8000 Hz. Octave bandpass filters according to IEC 61260-1 [25] shall be used.

   g) The ITD is defined as the difference in delay between right and left ear. For the calculation, the signal is prefiltered with a highpass filter with cut-off frequency 200 Hz and a lowpass filter with cut-off frequency 2000 Hz. Both filters shall use a filter order of four, which corresponds to a roll-off of 24 dB per octave. The ITD is calculated over the whole signal length using the cross-correlation method described in Annex C.

# Annex A (normative): Order dependent directions

The following tables order-dependent directions $\boldsymbol{\Omega}_j^{(N)} = (\phi_j^{(N)}, \theta_j^{(N)}),\ 1 \leq j \leq K$, where $\theta_j^{(N)}$ and $\phi_j^{(N)}$ denote the elevations and azimuths in radians, respectively.

| Index *j* | $\theta_j^{(N=1)}$ | $\phi_j^{(N=1)}$ |
|---|---|---|
| 1 | 1.570796 | 0 |
| 2 | -0.339837 | 0 |
| 3 | -0.339837 | 2.094395 |
| 4 | -0.339837 | -2.0944 |

| Index *j* | $\theta_j^{(N=2)}$ | $\phi_j^{(N=2)}$ |
|---|---|---|
| 1 | 1.570796 | 0 |
| 2 | -0.790277 | 0 |
| 3 | 0.363207 | -1.95668 |
| 4 | 0.363207 | 1.956682 |
| 5 | -0.844382 | -1.95668 |
| 6 | 0.009757 | -3.14159 |
| 7 | -0.844382 | 1.956681 |
| 8 | 0.245128 | 0.687124 |
| 9 | 0.245129 | -0.68712 |

| Index *j* | $\theta_j^{(N=3)}$ | $\phi_j^{(N=3)}$ |
|---|---|---|
| 1 | 1.570796 | 0 |
| 2 | 0.716698 | 0 |
| 3 | -0.461173 | 1.119907 |
| 4 | -1.034310 | -0.25283 |
| 5 | 0.492174 | 1.155586 |
| 6 | -0.165812 | 2.040481 |
| 7 | -0.461172 | -1.38118 |
| 8 | -0.165813 | 0.270692 |
| 9 | 0.001916 | -2.20417 |
| 10 | 0.653709 | 2.297267 |
| 11 | 0.653709 | -2.80293 |
| 12 | -0.192680 | 3.010956 |
| 13 | -1.079056 | 2.154919 |
| 14 | 0.001915 | -0.63529 |
| 15 | 0.616834 | -1.41973 |
| 16 | -0.887326 | -2.46809 |

| Index $j$ | $\theta_j^{(N=4)}$ | $\phi_j^{(N=4)}$ |
|---|---|---|
| 1 | 1.570796 | 0 |
| 2 | 0.747578 | 0 |
| 3 | -0.168324 | -2.00759 |
| 4 | 0.846499 | 1.927637 |
| 5 | 0.234515 | -1.41208 |
| 6 | 0.699165 | -2.10001 |
| 7 | 0.307091 | 2.512927 |
| 8 | 0.130649 | 1.667633 |
| 9 | -0.677517 | 1.442383 |
| 10 | 0.136843 | -0.60062 |
| 11 | -1.317269 | 0.329968 |
| 12 | -0.433118 | -1.18621 |
| 13 | -0.231864 | 2.983332 |
| 14 | 0.174242 | -2.69222 |
| 15 | -0.599985 | 0.507602 |
| 16 | -0.382009 | 2.208977 |
| 17 | -0.009394 | 0.952319 |
| 18 | -1.013813 | -1.71565 |
| 19 | 0.696199 | 0.934402 |
| 20 | -0.602139 | -0.38654 |
| 21 | -1.041921 | 2.675958 |
| 22 | -0.623111 | -2.62842 |
| 23 | 0.054056 | 0.165012 |
| 24 | 0.855489 | -1.02504 |
| 25 | 0.808243 | -3.13121 |

| Index $j$ | $\theta_j^{(N=5)}$ | $\phi_j^{(N=5)}$ |
|---|---|---|
| 1 | 1.570796 | 0 |
| 2 | -0.454100 | 0 |
| 3 | 0.323739 | -1.19666 |
| 4 | -1.175381 | 0.184066 |
| 5 | 0.947221 | 0.124282 |
| 6 | -0.193698 | -2.84022 |
| 7 | 0.500281 | -1.84701 |
| 8 | -0.663529 | 0.698758 |
| 9 | -0.613332 | 2.280239 |
| 10 | -0.588043 | -2.28482 |
| 11 | 0.946645 | -2.37569 |
| 12 | 0.333311 | 2.883411 |
| 13 | 0.967374 | -1.18504 |
| 14 | 0.436854 | -2.76846 |
| 15 | 0.510141 | 0.763488 |
| 16 | -0.063811 | -0.46491 |
| 17 | 0.048266 | -2.27504 |
| 18 | -0.148392 | 1.762138 |
| 19 | 0.945735 | 2.804486 |
| 20 | -0.125777 | -1.69175 |
| 21 | -0.241518 | -1.0321 |
| 22 | -0.063824 | 0.509415 |
| 23 | -1.240392 | -1.95737 |
| 24 | 0.542172 | -0.567 |
| 25 | 0.043647 | 2.319619 |
| 26 | -0.291045 | 2.853233 |
| 27 | -0.841101 | -3.07101 |
| 28 | -1.213891 | 2.113132 |
| 29 | -0.706626 | -1.50877 |
| 30 | -0.774625 | -0.65404 |
| 31 | -0.707445 | 1.464227 |
| 32 | 0.990842 | 1.373127 |
| 33 | -0.122664 | 1.112751 |
| 34 | 0.598614 | 2.113949 |
| 35 | 0.306690 | 0.057137 |
| 36 | 0.381934 | 1.457925 |

| Index $j$ | $\theta_j^{(N=6)}$ | $\phi_j^{(N=6)}$ |
|---|---|---|
| 1 | 1.570796 | 0 |
| 2 | 0.720144 | 0 |
| 3 | -0.308365 | 3.024454 |
| 4 | 0.068431 | 2.080642 |
| 5 | -0.495677 | -2.21373 |
| 6 | -0.018779 | -2.03598 |
| 7 | 0.426043 | 1.678014 |
| 8 | -0.259742 | 0.964363 |
| 9 | 0.179320 | -3.03552 |
| 10 | -0.249618 | -2.70206 |
| 11 | 1.074183 | 0.581055 |
| 12 | -0.781172 | -2.80103 |
| 13 | 0.457849 | 0.550136 |
| 14 | 0.523951 | -1.98436 |
| 15 | -0.006246 | -0.51212 |
| 16 | -0.788507 | -1.1411 |
| 17 | 0.228181 | -2.48765 |
| 18 | -0.418110 | -1.62282 |
| 19 | -0.512688 | -0.57506 |
| 20 | 0.572140 | 2.286204 |
| 21 | -0.867576 | -0.08741 |
| 22 | -0.624799 | 0.547028 |
| 23 | -0.446687 | 1.878965 |
| 24 | -0.789667 | 2.746717 |
| 25 | 1.047763 | -0.76025 |
| 26 | 0.247192 | -1.01978 |
| 27 | 0.720143 | 1.162107 |
| 28 | -0.081819 | 1.507148 |
| 29 | 0.226040 | 1.062706 |
| 30 | 0.709088 | -2.68135 |
| 31 | -0.249096 | -1.08377 |
| 32 | 0.573959 | 2.91352 |
| 33 | 1.069121 | 2.939099 |
| 34 | 0.135381 | -1.53966 |
| 35 | -0.057504 | 0.473238 |
| 36 | -0.975369 | -1.95522 |
| 37 | -0.666036 | 1.294994 |
| 38 | -1.146922 | 0.887936 |
| 39 | -0.357070 | 2.427548 |
| 40 | 0.200642 | -0.01608 |
| 41 | -0.965084 | 1.97199 |
| 42 | 0.681666 | -1.35341 |
| 43 | 0.112434 | 2.651183 |
| 44 | 0.528475 | -0.57647 |
| 45 | 1.003627 | 1.857517 |
| 46 | -1.275974 | -0.77916 |
| 47 | 1.051102 | -2.01121 |
| 48 | -1.315079 | 3.087768 |
| 49 | -0.326694 | -0.00446 |

# Annex B (normative): Directions in Gaussian spherical grid

## B.1 Definition

A Gaussian grid of order N consists of $2(N+1)^2$ points associated to directions $\Omega_{i,j}^{(N)} = \left(\theta_i^{(N)}, \phi_j^{(N)}\right)$, $0 \le i \le N$ and $0 \le j < 2(N+1)$, where $\theta_i^{(N)}$ and $\phi_j^{(N)}$ denote the elevation and azimuth, respectively. These directions are defined as follows [7]: The elevations $\theta_i^{(N)}$ are computed as the zeros of the (N+1)-th degree Legendre polynomial $P_{N+1}\left(\cos(\theta_i^{(N)})\right) = 0$, while the azimuths are given by $\phi_j^{(N)} = \frac{2\pi j}{2(N+1)}$ (in radians) or j.180/(*N*+1) (in degrees), $0 \le j < 2(N+1)$.

Directions in test setup shall comply with the theoretical values $\Omega_{i,j}^{(N)}$ with an accuracy of +/-0.5 degree for all azimuths and +/-0.5 degree for elevations in the range [-80,+80] degrees. For elevation >80 degrees and <-80 degrees, the accuracy shall be respectively +4/-0.5 degrees and +0.5/-4 degrees.

## B.2 Example loudspeaker array

An example implementation with an ambisonic order N = 29 is described below:

- A turn table with constant step size of 6 degrees and starting at 0 degree (to obtain 60 positions in azimuth).
- Two fixed semi-arcs of radius 2.5 meters separated in azimuth by 90 degrees with 15 loudspeakers on each semi-arc; the elevations of loudspeakers are given (in degrees) by -85, -80, -74, -68, -62, -56, -50, -44, -38, -32, -27, -21, -15, -9, -3, 3, 9, 15, 21, 27, 32, 38, 44, 50, 56, 62, 68, 74, 80, 85, where succesive values are alternatively allocated to each semi-arc.

NOTE: In practice, the elevation of -85 degrees may be replaced by a nearby value (e.g. -82 degrees) to leave room for the mounting structure at the bottom of the loudspeaker array.

# Annex C (normative): Cross-correlation analysis

The following analysis method is used to determine the time difference (delay) between two time-discrete signals $x(k)$ and $y(k)$ by applying segmental cross-correlation with period T (in samples) and overlap L (in percent).

If not specified otherwise, a sampling rate of 48 kHz is assumed.

The envelope $E(i,\tau)$ of the segmental cross-correlation function $\Phi_{xy}(i,\tau)$ between $x(k)$ and $y(k)$ is calculated by means of the Hilbert transformation:

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

$$H\left\{\Phi_{xy}(i,\tau)\right\} = \sum_{u=-\frac{T}{2}}^{+\frac{T}{2}} \frac{\Phi_{xy}(i,u)}{\pi(\tau-u)}$$

$$\Phi_{xy}(i,\tau) = \frac{1}{T}\sum_{k=\frac{-T}{2}}^{\frac{T}{2}} x(i,k)\cdot \mathrm{y}(i,k+\tau)$$

Each segment $i$ has a duration of T samples, using an overlap of L percent. The time difference $\Delta_\tau$ is then determined by the time lag $\tau$ that provides the maximum value.

$$\bar{E}(\tau) = \frac{1}{N}\sum_{i=1}^{N} E(i,\tau)$$

$$\Delta_\tau = \underset{\tau}{\operatorname{argmax}}\, \bar{E}(\tau)$$

For expected shorter time differences (up to ~85 ms), T = 8192 and L = 50% are recommended.

For expected longer time differences (up to ~1.4 s), T = 131072 and L = 50% are recommended.

If an aggregate value over the whole signal sequence is to be determined, T shall comprise the whole signal length (no overlap).

# Annex D (informative): Change history

| Change history | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Date** | **Meeting** | **TDoc** | **CR** | **Rev** | **Cat** | **Subject/Comment** | **New version** |
| 2018-09 | SA-81 | SP-180644 | | | | Presented to TSG SA#81 for approval | 1.0.0 |
| 2018-09 | SA-81 | | | | | Approved at TSG SA#81 | 15.0.0 |
| 2018-12 | SA-82 | SP-180969 | 0001 | 2 | | Corrections to test method with loudspeaker array and turn table | 15.1.0 |
| 2020-03 | SA-87-e | SP-200105 | 002 | 1 | F | Correction of sensitivity calculation for immersive audio playback | 15.2.0 |
| 2020-07 | - | - | - | - | - | Update to Rel-16 version (MCC) | 16.0.0 |
| 2022-04 | - | - | - | - | - | Update to Rel-17 version (MCC) | 17.0.0 |
| 2022-12 | SA-98-e | SP-221058 | 0004 | 2 | D | Corrections to TS 26.260 Clause 4 | 18.0.0 |
| 2024-06 | SA#104 | SP-240853 | 0006 | 4 | B | Objective Test Methodologies for IVAS-based UEs | 18.1.0 |

# History

| Document history | | |
|---|---|---|
| V18.0.0 | May 2024 | Publication |
| V18.1.0 | July 2024 | Publication |
| | | |
| | | |
| | | |