3GPPSA4#122 S4-230302

20th-24th February 2023 Revision of S4-221517

**Source: Editor[[1]](#footnote-1)**

**Title: ATIAS-1: Permanent Document on ATIAS, v0.2.0**

## Agenda Item: 14.1

# Introduction

The ATIAS work item intends to specify test methods for objective characterization of terminals for 3GPP immersive services along with requirements. This Permanent Document collects candidate test methods and associated requirements that will form a pool out of which selected methods and requirements will be incorporated into TS 26.260 (Objective test methodologies for the evaluation of immersive audio systems) and, respectively, TS 26.261 (Terminal audio quality performance requirements for immersive audio services).

# Candidate test methods and requirements

## Candidate sending side test methods and requirements

[

### Send side audio performance assessment for Immersive Audio Systems [1]

The following methods have been incorporated from [1]:

#### Send frequency response of captured Ambisonics components

*Definition:*

Ratio of the sound pressure magnitude spectrum of the DUT for Ambisonics component (, ): and a reference diffuse field sound pressure spectrum ( ). Letters and respectively denote Ambisonics degree and index. This means, for each Ambisonics component a send frequency response is measured:

.

*Ideal characteristic:*

The ideal send frequency response of the captured Ambisonics components would be flat, i.e. be frequency independent (for plane wave input).

*How to formulate requirements:*

One conceivable way to measure against requirements can be taken from TS 26.260. A decorrelated pink noise test stimulus is simultaneously played over all speakers of a periphonic array. The reference diffuse field sound pressure spectrum is obtained through recordings with a diffuse-field microphone positioned at the geometric centre of the periphonic array. The sound pressure magnitude spectrum of the DUT for Ambisonics component (, ) is obtained through measuring with the DUT at the geometric centre and extracting and analysing that (B-format) component after coding, transmission and decoding as in 26.260 Figure 1.

The sensitivity/frequency response shall meet the ideal characteristics, within some tolerance.

#### Send directional response of captured Ambisonics components

##### Angular-dependent sensitivity

*Definition:*

Microphone angular-dependent sensitivity can be described by (in Volt/Pascal or digital amplitude per Pascal), e.g. referring to the output voltage or digital level generated by a microphone for a given sound pressure at the microphone location for an incident plane wave from a certain direction . Likewise, the angular-dependent sensitivity of the capture of each Ambisonics component (, ) can be described as . As this sensitivity may also be frequency dependent, frequency is a further parameter:

.

Angular-dependent sensitivity of the capture of an Ambisonics component (, ) can also be defined for multiple () incident plane waves from respective directions . This yields:

with and .

Of interest may be the cases , and . For , the two directions should be 90 degrees apart (see [1-1]). For , a reasonable setup could be where the directions would correspond to the Fliege positions of the considered Ambisonics order .

*Ideal characteristic:*

The ideal angular-dependent sensitivity for the captured Ambisonics component (, ) is obtained from the spherical harmonics equations. For simplicity, it is assumed that the spherical coordinate systems of the measurement room and the captured Ambisonics signals are aligned.

For the case , we get

with being the real valued spherical harmonics of the degree and index with .

Thus, if the reference client offers output of this Ambisonics component it will in the ideal case and valid single plane-wave assumption (e.g. a distant-enough point source to approximate plane waves at the UE) follow the above relation. The ideal frequency characteristic is flat. Care needs to be taken to apply the proper Ambisonics component ordering and normalization.

For any , we get

These cases are not discussed further in this contribution.

*How to formulate requirements:*

The measurement for the simplest case of can be done according to the principles outlined above (under 2.1) using pink noise test stimuli. Unlike 2.1, only a single speaker at a time is used.

The angular-dependent sensitivity shall meet the ideal characteristics, within some tolerance.

##### Direction of arrival

*Definition:*

Taken from [1-2]:

The FOA signals are formed with a set of pre-calculated filters that convert the signal from Eigenmike to FOA. Creation of such filters has been described, e.g., in [5]. The result is an FOA signal that contains four signal channels. This signal is formulated into FuMa-ordering (i.e., the traditional B-format ordering) as the public DirAC equations used in the reference have been defined for it. Thus, the signal is:

The next step is to calculate intensity and energy related parameter values according to equations:

The energy and the intensity are calculated per subframe, i.e., the values of the four different time slots are summed together to produce the total subframe intensity and energy. Further steps are thus performed with subframe accuracy.

It should be noted that these are not the actual physical properties but parameters that are closely related to the physical properties. Actual direction of intensity is opposite of above and scaling factors have been omitted for efficiency and SN3D scaling is assumed for the FOA signal to simplify calculations. This analysis methodology is an optimized version of the typical DirAC analysis (see [6, pages 90-91]).

After formation of these parameters, they are combined in frequency to form frequency-band based values to reduce required amount of further calculations and data. This combination is simply done by summation of all values within a frequency band as further steps do not require that absolute scale is maintained. In further processing, it is assumed that these parameters work on the desired frequency resolution.

The next step is to estimate the direction of arrival (DOA), i.e., the spatial direction parameter, and the direct-to-total energy ratio. For spatial direction, this is done with equation:

for azimuth and equation:

for elevation. Here, arctan function is assumed to be the computational variant “atan2” that solves the correct quadrant automatically.

*Ideal characteristic:*

The DOA angle calculated from the Ambisonics components from the UE capture system, is identical to the ground truth angle to the incident sound

How to measure:

The UE is connected via a wireless link to a reference client

Loudspeakers are placed on a suitable grid (reusing what is already specified, e.g. according to Annex A or B).

The loudspeakers emit sound one at a time. Their incidence angle is known. It remains to be evaluated whether an exponential sweep, pink noise or speech is the most suitable.

The Ambisonics components from the reference client are used to calculate the DOA (according to above).

The DOA is compared to the ground truth angle, potentially in several frequency bands and potentially time-averaged.

If the sending UE is properly implemented in terms of directionality, phase and scaling of Ambisonics components, the DOA metric is expected to correspond to the ground truth angle.

*How to formulate requirements:*

The DOA angle calculated from the Ambisonics components from the UE capture system shall be within some tolerances w.r.t. the ground truth angle to the incident sound.

##### Virtual microphones

*Definition:*

A virtual microphone is created by a linear combination of the ambisonics components. As an example, for FOA, a cardioid, super-cardioid, figure-of-eight etc microphone can be constructed and pointed in an arbitrary direction, by adding or subtracting portions of the four FOA ambisonics components.

It is called a *virtual* microphone because its characteristics are manipulated without affecting the microphone itself. Compare with music production; the mix engineer can *in post-production* select where to point a microphone, and even construct an infinite number of virtual microphones pointing in various directions with various beam widths. A simplified variant of FOA capture is MS stereo recording using one omni and one figure-of-eight capsule, where the stereo width is decided in post-production by turning (usually two) virtual microphones’ mutual angle and setting their pickup pattern to taste. For ATIAS, this means the test system can steer an infinite number of virtual microphones anywhere and controlling the width of the pickup beam, without interacting with the UE that is “just” supplying ambisonics components to the test system.

*Ideal characteristics*

An ideal first-order directional microphone’s directional pickup in a plane can be described as:

, where *p* =1 provides omni-directional, *p*=0.5 cardioid, *p* =0.37 super-cardioid, *p* =0.25 hyper-cardioid, *p* =0 bi-directional, etc.

This polar pattern can further be rotated, as illustrated below. The GIF illustrates adding portions of FOA components W, X and Y (“Pattern” corresponds to *p* above):

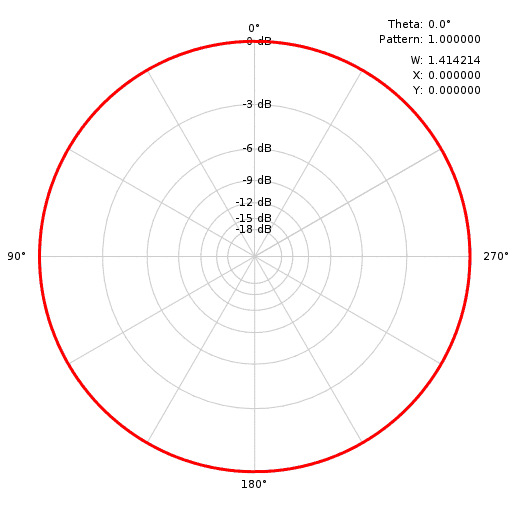


Figure 1 Illustration of beam and pattern steering (copied from [1-3])

With these simple means we are able to e.g. construct two simultaneous virtual microphones, one pointing at a sound source, the other pointing away from the sound source. This can be as simple as summing/subtracting the outputs from the reference decoder:

* Virtual mic A = W+Y
* Virtual mic B = W-Y

In the ideal case, if using two virtual back-to-back cardioids, one pointing towards a single sound source and one pointing away from it, only the first microphone shall produce an output signal.

*How to formulate requirements:*

Example: When a single sound source is placed at the positive Y direction in relation to the UE, the level of signal (W+Y) shall exceed the level of signal (W-Y) by at least X dB[[2]](#footnote-2).

This test can be expanded to cover a variety of incidence angles. Or it can be expanded to rotate the virtual microphones while keep a single source constant.

The merit of using two virtual microphones (which are preferably implemented simultaneously), instead of only one, is that the testing is robust to dynamic range processing in the UE, thus spatial aspects can be robustly assessed with this method. This is achieved by the virtual microphones “pointing” at two or several incidence angles while the UE is subjected to only one source from one angle. A merit compared to [1-1], is that we avoid two simultaneous sound sources and two separable signals.

The method is equally usable for HOA, by defining appropriate linear combinations for the ambisonics components from the decoder.

]

[

The following methods have been incorporated from [4]:

### X.1.2 Scene-based audio send spatial separation with two simultaneous acoustic sources

#### X.1.2.1 Definition

The Scene-based audio send spatial separation with two simultaneous acoustic sources is the difference in level combinations observed at the output of the reference IVAS decoder/[renderer] when the UE is subjected to two simultaneous acoustic sources at predefined directions, *(ii* *i*=1,...,L.

#### X.1.2.2 Test conditions

**Free-field propagation conditions**

- The test environment 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 [200Hz].[Editor’s note: The 200Hz value is inherited from TS 26.260 while TS 26.132 specifies 275 Hz for communication. In case the latter is used, test signals should not have energy below 275Hz, since echo-free conditions are necessary for this test]

[QCOM – Andre] Keep 200Hz for: (1) (communication use case has important energy in that specrum region) and (2) consistency with existing setup/specs.

**[Test environment noise floor]**

[Editor’s note: The test environment noise floor may not have to specified in this clause. Likely, a general clause for the whole specification is sufficient.]

**Loudspeaker array**

An array of coaxial loudspeakers is is located at a set of predefined directions *(ii*, *i=1,…,6*, from the geometric center of the UE. The different locations may be realized using multiple loudspeakers or by rotation of at least two loudspeakers or by rotation of the UE.

[QCOM – Andre] Harmonize with the existing test setup (periphonic array in 26.260). It should be possible to maintain the same test method, but using positions already existing in the periphonic array by having different targets for the delta FOA signal components instead of a hard zero. Alternative angles would also be more aligned with what the device will be exposed to in actual use (a hard +/-90deg azi/ele is typically not in the field of view and not the focus of a capture). Alternatively, and less preferrable, the periphonic array can potentially be augmented with these additional 5 positions (Y2 is already present).

In case the UE has motion compensation (automatic rotation of the captured soundfield depending on pose), physical rotation of the UE shall not be used to achieve the predefined directions and it shall be ensured that there is no misalignment of X, Y and Z axes due to the motion compensation.

[QCOM – Andre] Here again, the alignment with the periphonic array would be of help since the test lab would not have to worry about the issue above and no movement of the device would be required. See comments above.

[Editor’s note: this applies for general audio case. For communication HATS playback might be considered.]

The distance from the loudspeaker front baffle to the center of the UE shall be at least 1m. [Editor’s note: check if 1m is sufficient, considering the proximity effect at 200/275 Hz, which we would like to avoid as it biases the test result. This should go (later) into a separate clause]

The loudspeakers shall be equalized with the UE absent, using a [measurement microphone and diffuse-field equalization] placed at the UE position. The microphone shall point in the positive Z direction with its membrane in the XY plane.

Table X: Location of loudspeakers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Position | i | i [deg] | i [deg] | Comment |
| X1 | 1 | 0 | 0 | 0 deg (frontal) incidence to the DUT, along the X-axis |
| X2 | 2 | 0 | 180 | opposite to X1, along the X-axis |
| Y1 | 3 | 0 | -90 | -90 deg incidence to the DUT, along the Y-axis |
| Y2 | 4 | 0 | 90 | opposite to Y1, along the Y-axis |
| Z1 | 5 | 90 | 0 | “from the ceiling” incidence to the DUT, along the Z-axis |
| Z2 | 6 | -90 | 0 | opposite to Z1, along the Z-axis |

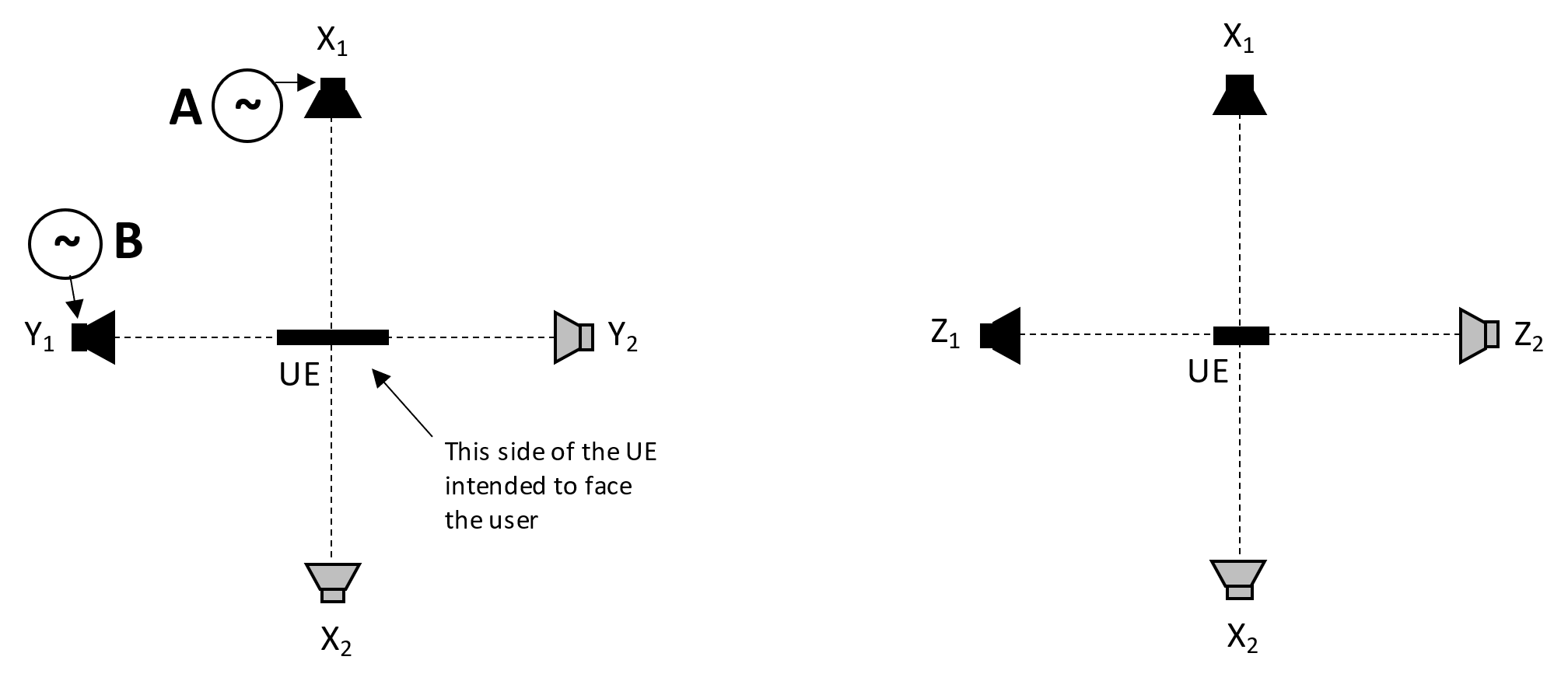


Figure X: Location of loudspeakers

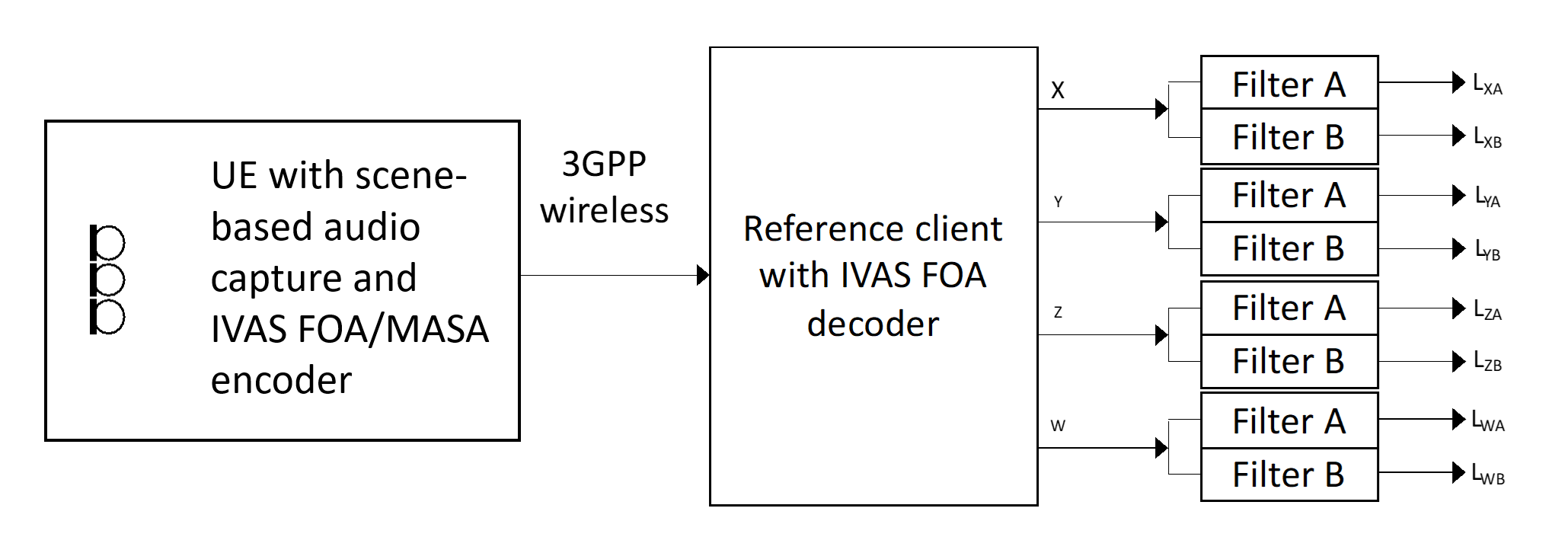


Figure X: Example using FOA; The UE under test is connected to a test system composed of a 3GPP wireless system simulator and a reference client with B-format output and frequency-domain filters for analysis. For HOA, the same setup is used and the higher order ambisonics components at the receiver are ignored. For MASA input capture, IVAS MASA encoder is utilized

[QCOM – Andre] This should be made more generic with the addition of reference IVAS renderer (either binaural renderer and then use auditory model derived metrics or a high channel count loudspeaker renderer)

#### X.1.2.3 Measurement

The following procedure shall be used:

1. The UE under test is connected to a test system composed of a 3GPP wireless system simulator and reference client with an IVAS session established with B-format output. The codec shall be operated with scene-based audio or metadata-assisted spatial audio input format at [512] kbit/s. The audio input format and bitrate shall be reported. The decoder/renderer option shall be FOA.

[QCOM – Andre] Define that maximum bit-rate supported should be used. Would need to add some note that test may not be suitable for UEs featuring noise suppression due to the multitone test signal (alternatively, the test lab may need to disable noise suppression but this is generally less preferrable). As mentioned above, this could be made more generic by using decoder/renderer to binaural or high channel count.

b) A modulated multi-tone test signal A is played over a loudspeaker at position X1. Simultanously, a modulated multi-tone test signal B is played over a loudspeaker at position Y1. See Annex X for a description of the multi-tone signals.

Editor’s note: The impact of codec on the test signal needs to be verified before performing the measurements.

c) The output of each ambisonics component (W, X, Y, Z) is captured. After an initial conditioning time of [5] seconds the remainder of the captured signal is converted to the frequency domain as described in Annex X. The signals are filtered by two different comb filters, filter A and filter B, with passbands corresponding to frequencies in signals A and B respectively. The filters are realized by including/excluding certain frequency bins as described in Annex X.

d) The levels after the filters, averaged over the whole duration, are calculated by summing the power of the selected bins.

e) The level metrics according to Table X are calculated.

Table X: Assessment of spatial separation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Simultaneous sources | | Requirements on the B-format outputs of the reference decoder | | |
| Source A | Source B | Signal component A | Signal component B | Motivation |
| Position X1 | Position Y1 | LXA – LYA > [N] dB,  |LWA – LXA| < [M] dB | LYB – LXB > [N] dB,  |LWB – LYB| < [M] dB | Signal component A is ideally only seen in X and W, Signal component B is ideally only seen in Y and W |
| |LWA – LWB| < [P] dB | | Signal component A in W equally strong as B in W |
| Position X1 | Position Z1 | LXA – LZA > [N] dB,  |LWA – LXA| < [M] dB | LZB – LXB > [N] dB,  |LWB – LZB| < [M] dB | Signal component A is ideally only seen in X and W, Signal component B is ideally only seen in Z and W |
| |LWA – LWB| < [P] dB | | Signal component A in W equally strong as B in W |
| Position X2 | Position Y2 | LXA – LYA > [N] dB,  |LWA – LXA| < N dB | LYB – LXB > [N] dB,  |LWB – LYB| < N dB | Signal component A is ideally only seen in X and W, Signal component B is ideally only seen in Y and W |
| |LWA – LWB| < [P] dB | | Signal component A in W equally strong as B in W |
| Position X2 | Position Z2 | LXA – LZA > [N] dB,  |LWA – LXA| < [M] dB | LZB – LXB > [N] dB,  |LWB – LZB| < [M] dB | Signal component A is ideally only seen in X and W, Signal component B is ideally only seen in Z and W |
| |LWA – LWB| < [P] dB | | Signal component A in W equally strong as B in W |
| The test is repeated where signals A B are interchanged, to avoid a potential bias. The results from the two tests are averaged. The values M (maximum of difference to omni component), N (minimum of off-axis rejection) and P (maximum unbalance of omnidirectional capture) are TBD. | | | | |

[Editor’s note: in case there will be different specifications for the test methods and the requirements, the table can be moved to the latter document, e.g. TS 26.261.]

# Annex A: Test signal definition

## A.1 Definition

The test signal shall be generated according to the ITU-P.501 [1] (subclause 7.2.4.1) and as provided in equation A.1.

A screenshot of a computer

Description automatically generated with low confidence

Figure x– Two channel test signal generation for double-talk evaluations  
based on AM-FM signals

n = 1,2,... (A.1)

where

In ITU-T P.501, the following parameters are defined in a frequency-independent manner: , and . The center frequencies for test signal are defined in the Table A.1.

The frequency-dependent modulation bandwidth is determined as follows:

Table A.1: Centre frequencies and bandwidths (1/3rd octave bands)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Center Frequency [Hz] | Talker | Freq. Start [Hz] | Freq. Stop [Hz] | [Hz] |
| 250 | A | 237 | 272 | [TBD (should essentially be Fstop – Fstart)] |
| 315 | B | 306 | 345 |  |
| 400 | A | 388 | 424 |  |
| 500 | B | 487 | 529 |  |
| 630 | A | 612 | 669 |  |
| 800 | B | 776 | 849 |  |
| 1000 | A | 974 | 1058 |  |
| 1250 | B | 1216 | 1323 |  |
| 1600 | A | 1547 | 1697 |  |
| 2000 | B | 1948 | 2117 |  |
| 2500 | A | 2432 | 2646 |  |
| 3150 | B | 3150 | 3344 |  |
| 4000 | A | 3882 | 4120 |  |
| 5000 | B | 5000 | 5144 |  |
| 6300 | A | 6300 | 6491 |  |
| 8000 | B | 8000 | 8239 |  |
| 10000 | A | 10000 | 10287 |  |
| 12500 | B | 12500 | 12859 |  |

[Editor’s note: add a column to indicate which of the above frequencies to disable in tests with MASA. The frequency spacing may still need practical verification.]

[Editor’s note: once the 200/275 Hz lower limit for freefield conditions has been decided, the lowest frequency of the test signal can be selected based on this.]

## A.2 Shaping filter

To generate [typical|speech-like] frequency characteristics, three different shaping filters can be applied:

- Low-pass at 250 Hz and 3 dB/octave roll-off characteristics

- Low-pass at 250 Hz and 5 dB/octave roll-off characteristics

- Average speech-spectrum, as per ITU-T P.50 / ITU-T P.810 [xx]

The three shaping filters are illustrated in Figure A.2.

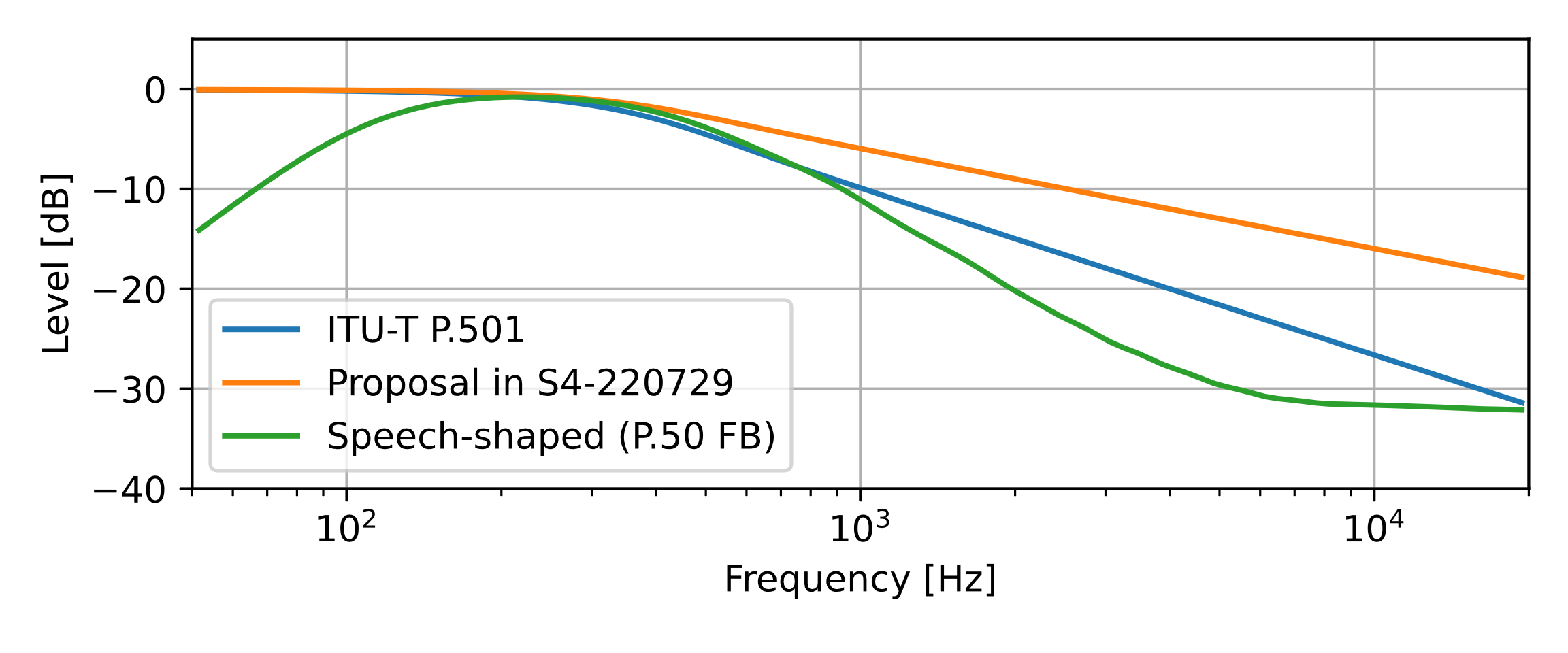


Figure A.2: Shaping filters for test signal

[Editor’s note: different shaping filters can be considered for general audio and communication scenarios]

A.3 Spectral maskThe spectral masks for the calculation of individual per-source / per-talker levels are defined as follows.

The signals are sampled at 48kHz sampling rate and transferred to the frequency domain using a [2^16] FFT, Hann window, [50%] overlap. Frequency bins are multiplied by 1 if they are within the passbands, and by 0 if they are outside.

The passbands of the masks are defined by the stimulus carrier frequencies and the frequency modulation plus a further widening by one frequency bin at each side, see Table A.1.

where

]

[

The following method(s) have been incorporated from [5]:

### Spatial perception test for stereo UE in ATIAS

1. Introduction

Spatial perception is an essential part of immersive audio. To get the immersive experience, the UE must have a correct auditory location. Otherwise, users will be confused about the discrepant spatial information.

So, this proposal aims to add a test method to determine the central and left\right direction for a hand-held hands-free stereo audio system in the sending direction. Then, discuss the threshold of mismatch between different spatial information.

The clause 3 is a test method proposed for the permanent doc S4-221517. Clause 4 is discussing the associated requirements.

1. Rationale

The stereo format is becoming common on the market because of the demand for immersive experience, though an official standardized test method hasn't appeared.

Some devices may have an excellent sound image location and quality, but it may be discrepant when it integrates with other information, like visual.

This proposal measures the ILD\ITD of UE, which has been investigated for decades to evaluate the sound image location. And discuss the central and left\right direction, the primary function of stereo audio.

1. Test setup
   1. Introduction

This test is applicable to liner system UEs capturing stereo audio.

* 1. test conditions

- The test conditions should follow the Free-field propagation conditions and test environment noise floor described in TS 26.260[1].

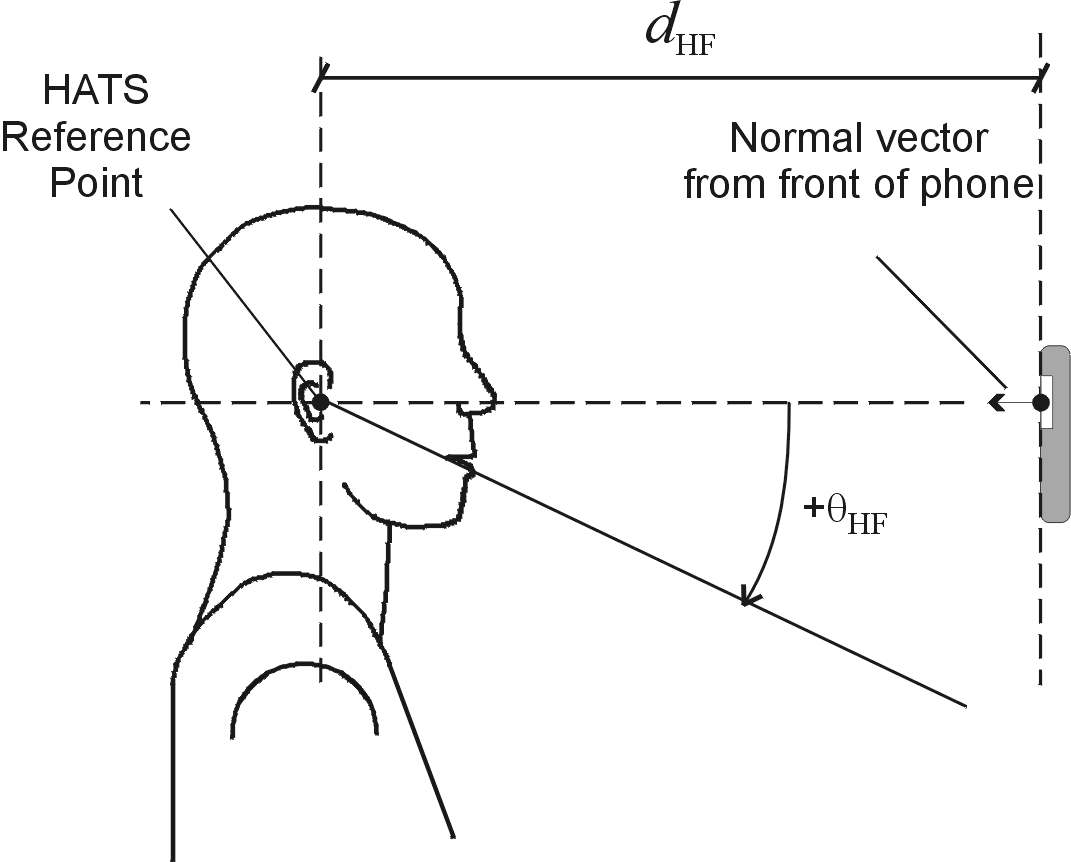
* 1. Setup for terminals

The setup is referred to TS 26.260 and TS 26.132[2]. Including the POI, reference point, etc.

Where the manufacturer gives conditions of use, these will apply for testing. If the manufacturer gives no other requirement, the DUT will be positioned according the reference usage of hand-held hands-free UE in TS 26.132 describing in the following block:

##########################################################################################

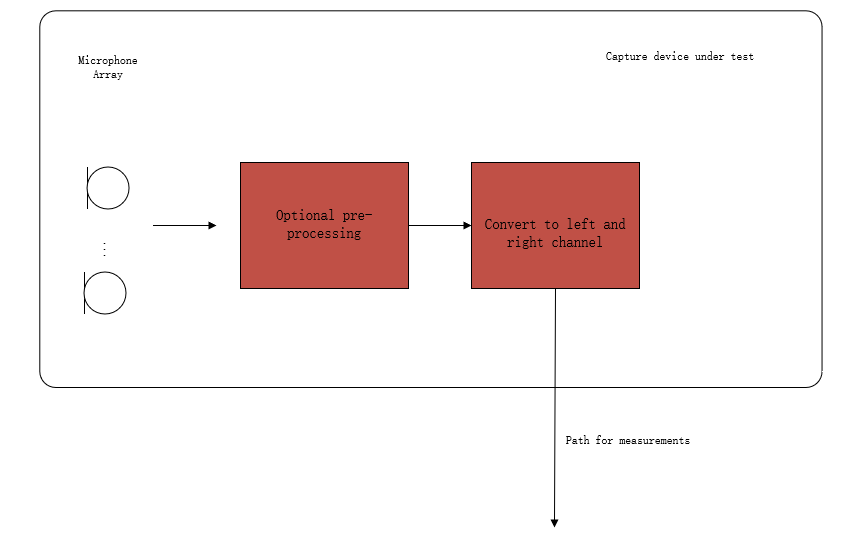
If HATS measurement equipment is used, it shall be configured to the hand-held hands-free UE according to figure 4. The HATS should be positioned so that the HATS Reference Point is at a distance dHF from the centre point of the visual display of the Mobile Station. The distance dHF is specified by the manufacturer. A vertical angle HF may be specified by the manufacturer. Where it is not specified, the nominal distance dHF shall be 42 cm and HF shall be 0º.



*Figure 4: Configuration of hand-held hands-free UE relative to the HATS*

################################################################################################

Measurement points:



**Figure 1: Audio capture block diagram for sending direction measurements**

* 1. Definition

**interchannel level difference**

Theinterchannel level difference is the sound level of the left channel minus the right channel.

**interchannel time difference**

Theinterchannel time difference is the times-of-arrival of the sounds of the left channel minus the right channel.

**NOTE:** If calculate other parameters like subband signal, SNR, etc is TBD. Since the actual performance of stereo UE hasn’t been confirmed.

**Central direction:**

To create a central direction, the left and right channels usually have the same or similar signals.

The central direction range is TBD

The requirement of a central direction is TBD

**Left\Right direction:**

The left and right channel should have sufficient difference to make sound images located on the left or right. If the sound source comes from the left direction, the interchannel time difference<0 and\or interchannel level difference >0 in general and vice versa.

The left and right range is TBD

The requirement of left and right direction is TBD

**NOTE**: Some specific situation like ILD>0 ITD>0 is TBD.

**Auditory central location**

when the left and right channel have the most similar signal the sound source is located on the central audio location

**Visual central location**

For audio device with accompanying picture, when the reference point of source is appeared on the median line of screen, the sound source is located on the visual central location

* 1. Measurement method

1. The UE device under test is mounted in the free-field volume such that its reference point is on the axis of the sound source.

Repeat steps b-c) with an azimuth angular resolution of N degrees for every possible usage range (at least cover the visual range):

1. The sound source pointed directly toward the reference point of the DUT, measuring the impulse response of DUT on the α degree from the reference line (minus for left).
2. Change the angle between sound source and DUT.

**Test signal：**

Swept-sine signal [2]

**Sound source:**

HAT and coaxial loudspeaker.

**NOTE**: Since the UE is most used for speech service, and avoid phase different cause by x-way loudspeaker.

**Calculate ITD and ILD:**

**Discrepancy between audio-visual spatial location:**

The angular offset of auditory central audio location is

The angular offset of visual central location is

Discrepancy between audio-visual spatial location:

1. Discuss on requirement

To create an immersive experience correctly, the UE should have a basic requirement that the audio spatial perception has right direction (left, central or right), at least for Audio-visual spatial perception.

NOTE: The prior research is most based on standard HDTV sound system. But the performance on UE is different greatly, the requirement should consider the limit of UE device.

References

1. 3GPP TS 26.260: " Objective test methodologies for the evaluation of immersive audio systems."
2. 3GPP TS 26.132: " Speech and video telephony terminal acoustic test specification."
3. A. Farina: Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique, Audio Engineering Society Convention Paper 5093, (2000 February)

]

[

The following method(s) have been incorporated from [6]:

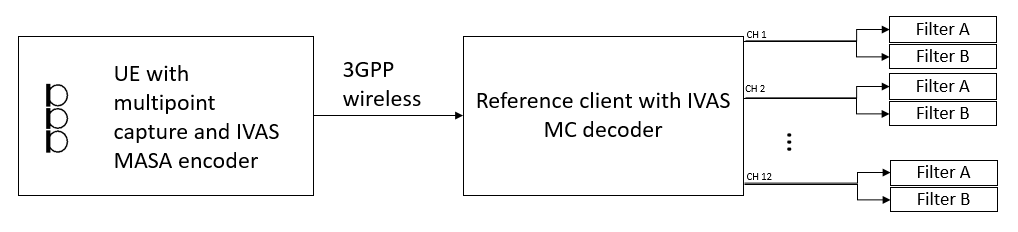
### Spatial separation for multiple acoustic sources based on multichannel output

**2. Test method**

**2.1 Test setup**

Test should be performed in an anechoic room. For the measurement, two simultaneous identical acoustic sources A and B are used. The distance from the loudspeaker front baffle to the center of the UE shall be at least 1m. To separate the sound sources at the analysis phase, multitone signals with non-overlapping frequency responses are utilized. The test signal is further defined in the section 2.2.

The UE is put into MASA capture mode at the highest available bitrate. The encoded stream is transmitted to the test system where a reference client decodes the signal and outputs the multichannel output signal with 7.1+4 loudspeaker configuration for analysis. Decoded output signal is converted to the frequency domain and filtered by two different comb filters, filter A and filter B, with passbands corresponding to frequencies in signals A and B respectively. The filters are realized by including/excluding certain frequency bins by utilizing frequency masks.



The configuration of 7.1+4 multichannel loudspeaker setup can be found in Table 1. Different combinations of two loudspeaker positions will be utilized for testing simultaneous multisource capture at different directions. In practice, the setup can be realized by changing positions and angles of two loudspeakers.

Table 1 Output loudspeaker configuration

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Ch** | **1** | **2** | **3** | **4/LFE** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| **Azi** | 30° | -30° | 0° | 0° | 135° | -135° | 90° | -90° | 30° | -30° | 135° | -135° |
| **Ele** | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 35° | 35° | 35° | 35° |

In order to test the performance in horizontal plane, channel combinations of channels 1,2,3,5,7 and 8 are tested. The tested channel combinations in horizontal plane are:

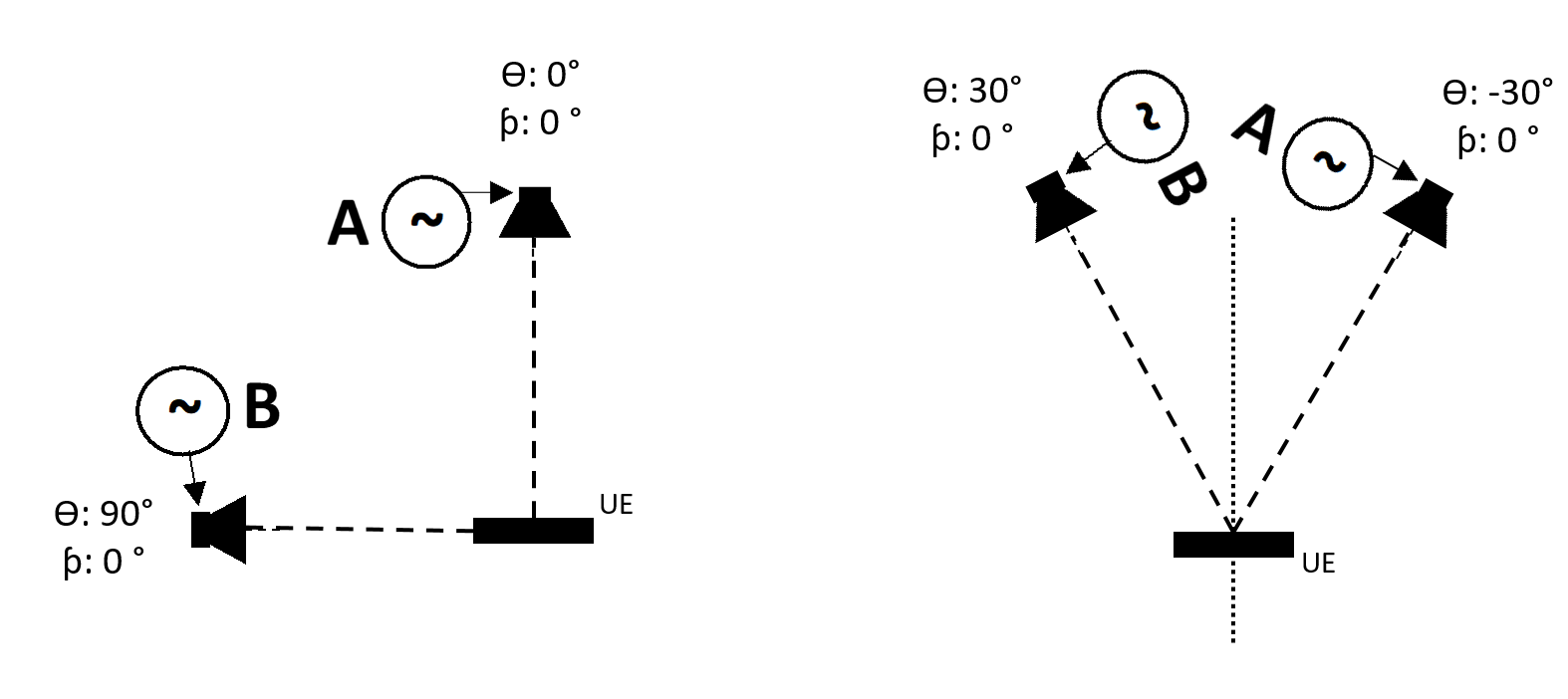
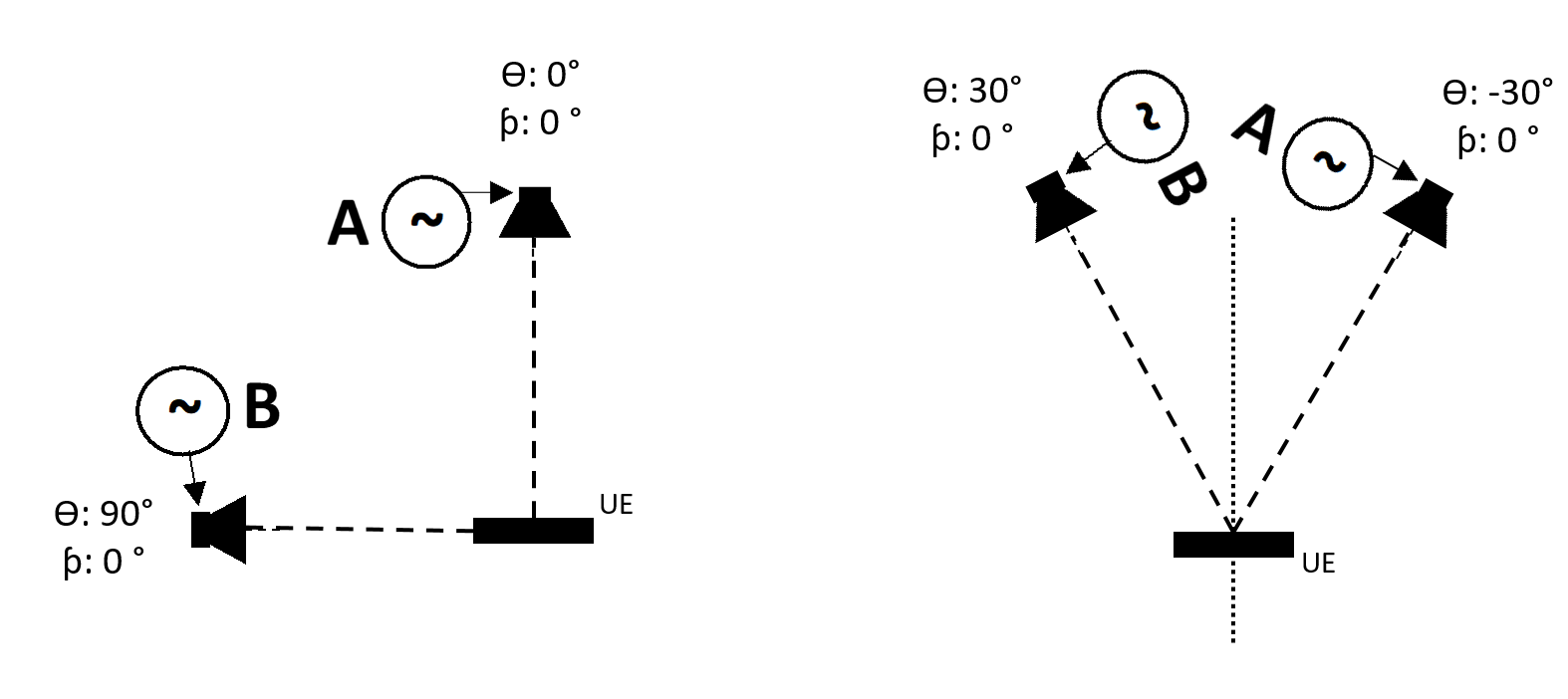
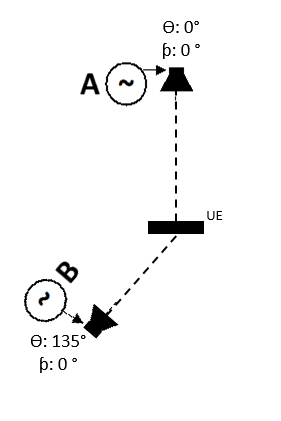
1. Ch 3 and Ch 7 ( & )
2. Ch 2 and Ch 1 ( & )
3. Ch 8 and Ch 7 ( & )
4. Ch 3 and Ch 5 ( & )

To validate the vertical performance, the sum of elevated loudspeaker channels is analysed. If the MASA capture works properly, all the sound directly above the UE should be decoded only into the channels 9-12. The sum of the elevated channels can be assessed with horizontal channels, e.g., the center channel 3 (directly in front of the UE). The tested channel combination in vertical plane is then:

1. Ch 3 and sum of Ch 9-12 ( & )

Number of the channel combinations can be increased or decreased, depending on the desired test accuracy, i.e., only combinations of channels 1,2,3,7,8 could be utilized to verify that the device can capture two non-elevated simultaneous sound sources from the front of the device. This can be extended to include elevated sources by adding the channels 9-12 to the analysis and further to include full 360-degree accuracy by adding channels 7 and 8 to the analysis.

All proposed channel/direction combinations are illustrated in the figures below:

**a) b) c) **

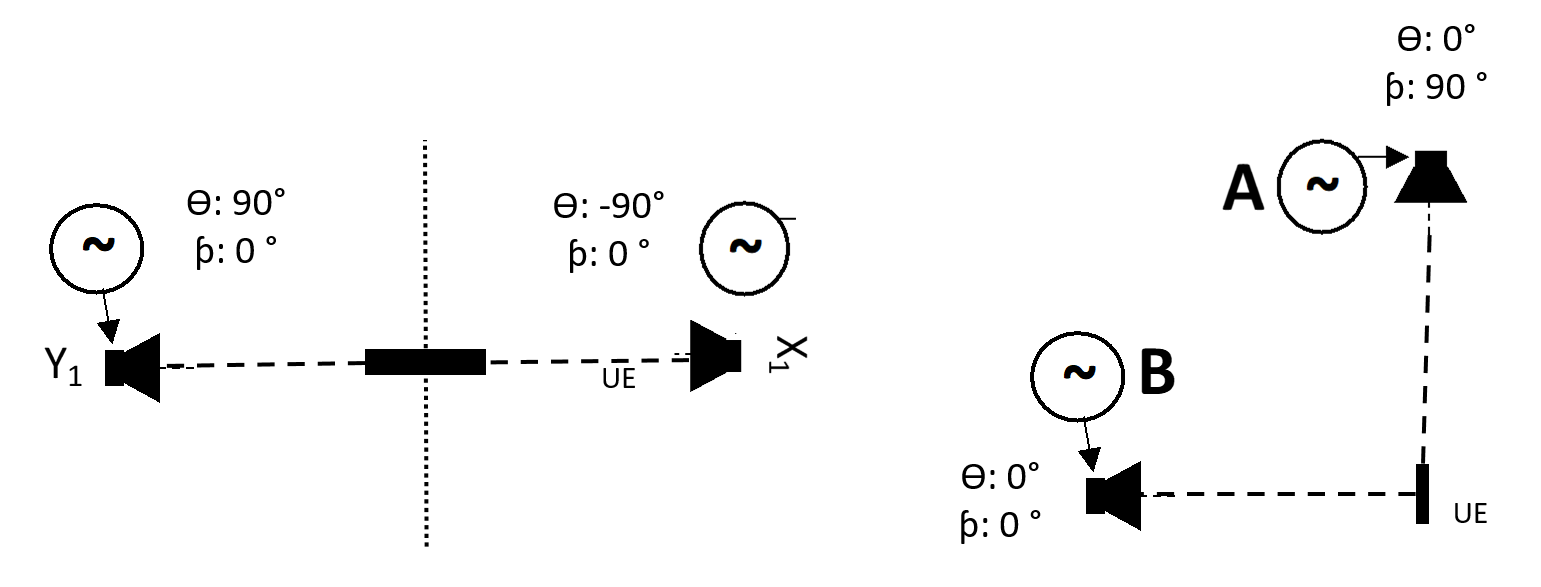
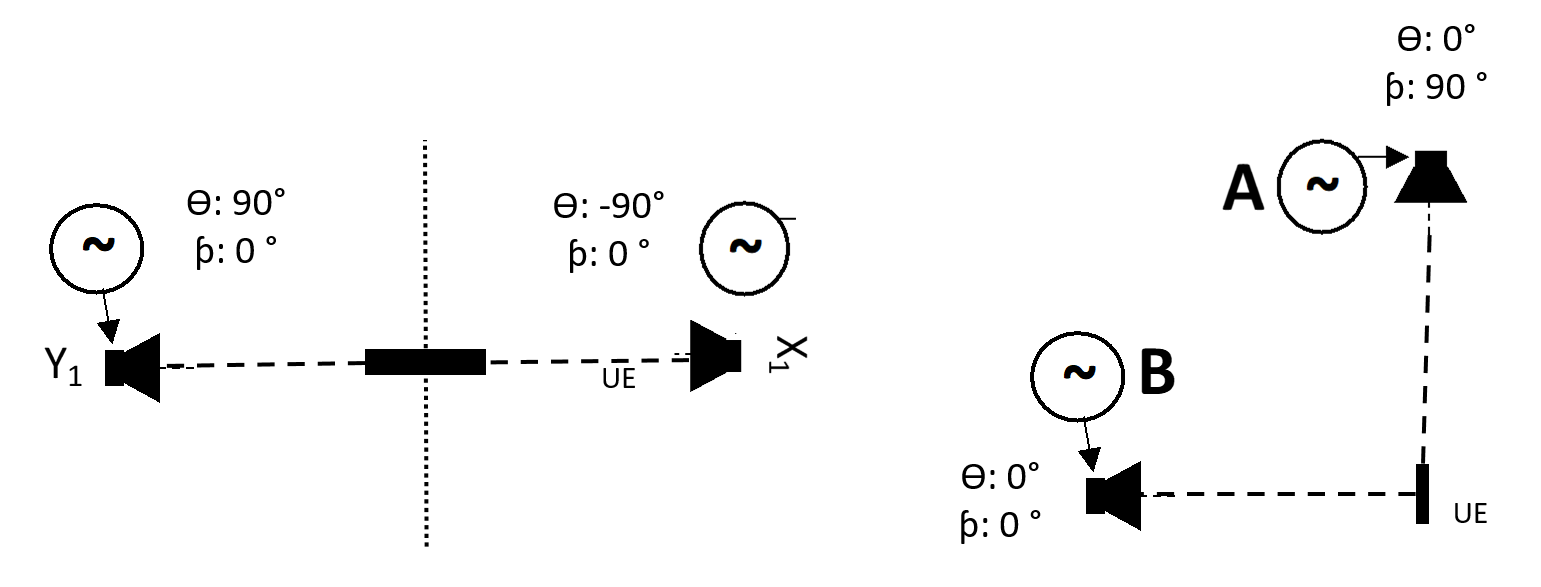
**d) e) **

Figure 1 Illustrations of proposed channel combinations

**2.2 Test signal**

Test signal shall be generated according to the ITU-P.501 [2] subclause 7.2.4.1

A screenshot of a computer

Description automatically generated with low confidence

Figure 2 AMFM modulated test signal generation for multisource capture evaluations according to the ITU-P.501 [2].

n = 1,2,...

where

In ITU-T P.501, the following parameters are defined in a frequency-independent manner: , and . The center frequencies are defined in the Table 2. The frequency dependent modulation bandwidth is determined as follows [2]:

Both shaping filters are identical: Lowpass with a cutoff frequency of 250Hz and slope of 3dB/5dB or filter with the response of average speech spectrum from ITU-T P.50 [3].

Table 2 Properties of a specific multisource test signal, modulated in amplitude

| (Hz) | (Hz) |
| --- | --- |
| 250 | 600 |
| 1000 | 1400 |
| 1800 | 2200 |
| 2600 | 3000 |
| 3400 | 3800 |
| 4200 | 4600 |
| 5000 | 5400 |
| 5800 | 6200 |
| 6600 | 7000 |
| 7400 | 7800 |
| 9000 | 11000 |
| 14000 | 18000 |

The power spectrums of the test signals are illustrated in the Figure 2. Every second frequency component is in the A signal and every other in the B signal. A and B signals are output from different sound sources at the same time.

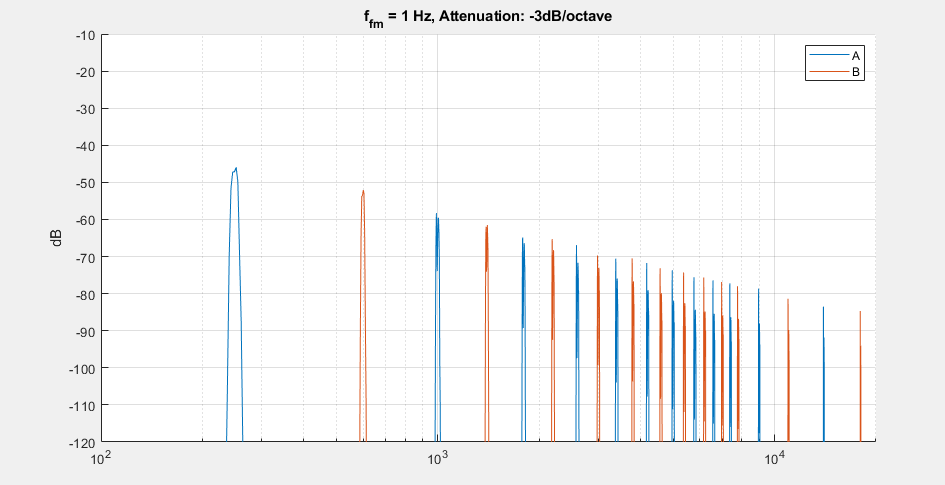


Figure 3 Power spectrum of AM/FM modulated test signal.

**2.3 Level calculations**

The levels are calculated as suggested in document [1]. Signals A and B are separated after the reference client decoder. Separation is done by applying masks on the FFT bins. 16K FFT with Hann window and 75% overlap is applied for the analysis. The width of the masks are set slightly wider (+-3 Hz) than the signal modulation. Channels levels are calculated by summing the power of the selected bins. The test is repeated with interchanged signals A and B, and the results of the two measurements are averaged.

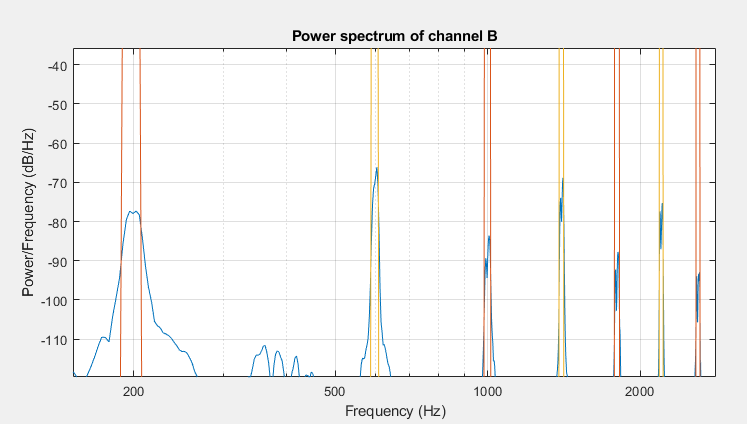
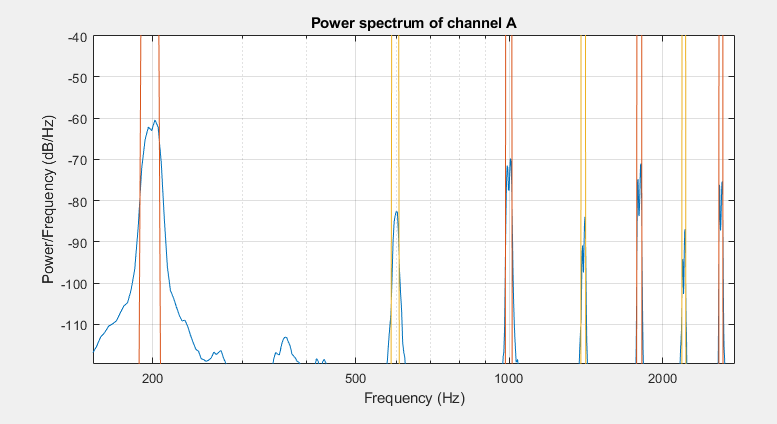


Figure 4 At the analysis stage, the corresponding frequency bins are separated by applying frequency masks. Level difference of different frequency components can be observed from the figures.

Finally, the levels are assessed according to Table 1.

Table Requirements on spatial separation of decoded multichannel output channels. L denotes a signal level measured at the reference decoder output by summing power in selected FFT bins

|  |  |  |  |
| --- | --- | --- | --- |
| Simultaneous sources | | Requirements on the 7.1+4 Multichannel outputs of the reference decoder | |
| Source A | Source B | Signal component A | Signal component B |
| Azi: 0°  Ele: 0° | Azi: 90°  Ele: 0° | L3A – L7A > [N] dB | L7B – L3B > [N] dB |
| Azi: -30°  Ele: 0° | Azi: 30°  Ele: 0° | L2A – L1A > [N] dB | L2B – L1B > [N] dB |
| Azi: -90°  Ele: 0° | Azi: 90°  Ele: 0° | L8A – L7A > [N] dB | L7B – L8B > [N] dB |
| Azi: 0°  Ele: 0° | Azi: 135°  Ele 0° | L3A – L5A > [N] dB | L5B – L3B > [N] dB |
| Azi: 0°  Ele: 0° | Azi: 0°  Ele: 90° | L3A – Ʃ L9…12A > [N] dB | Ʃ L9…12B - L3B > [N] dB |

**3. Experiments**

**3.1 Measurement setup**

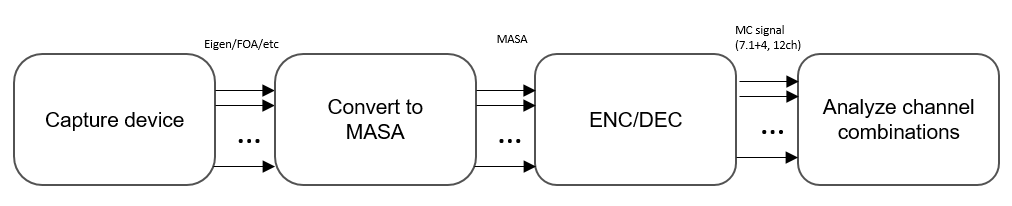


Figure 5 Processing flow from capture to analysis

Measurements were conducted in order to evaluate the applicability of the proposed test method. The processing flow from capture to analysis is shown in the Figure 3. Eigenmike and Rode NT-SF1 A-format microphones were used as capture devices. Captured signals were converted into MASA input format [3] via IVAS MASA C Reference Software [4]. MASA input signal was encoded and decoded to the 7.1+4 multichannel output format with IVAS candidate technology utilizing bitrate of 512 kbit/s. Decoded 7.1+4 Multichannel signal was analysed according to Table 2.

Two small identical loudspeakers were utilized as sound sources. The distance between loudspeakers and the DUT was 1.3m. Five different speaker position combinations were measured in horizontal plane:

1. 0° & 30°
2. 0° & 90°
3. 0° & 135°
4. +/- 30°
5. +/- 90°

**3.2 Results**

Results of Eigenmike capture are presented in Table 4 and results of Rode NT-SF1 capture in Table 5. The presented results are averaged results of two measurements with interchanged signals.

In overall, the analysed results indicate mostly proper spatial capture performance as expected. Analysed level differences are higher in average with Eigenmike compared to Rode captures. This behaviour is also expected, as the number of microphones is higher in Eigenmike, thus spatial capture accuracy is assumed to be higher.

With Rode capture, the level difference of channel combination 1 and 3 (Azi: 0° and Azi:30°) suggests inaccurate spatial separation, as the level difference of B component is negative. This behaviour is probably due to too small separation of the loudspeakers for an accurate analysis from FOA components.

Table 4, Results of Eigenmike measurements

|  |  |  |  |
| --- | --- | --- | --- |
| Eigenmike measurement results | | | |
| Source A | Source B | Signal component A | Signal component B |
| Azi: 0° | Azi: 30° | L3A – L1A = 11.54 dB | L1B – L3B = 13.5 dB |
| Azi: 0° | Azi: 90° | L3A – L7A = 18.87dB | L7B – L3B = 18.12 dB |
| Azi: 0° | Azi: 135° | L3A – L5A = 17.84 dB | L5B – L3B = 17.09 dB |
| Azi: -30° | Azi: +30° | L2A – L1A = 19.34 dB | L1B – L2B = 20.23 dB |
| Azi: -90° | Azi: +90° | L8A – L7A = 16.93 dB | L7B – L8B = 17.13 dB |
| Average | | 16.90 dB | 17.21 dB |

Table 5, Results of Rode NT-SF1 measurements

|  |  |  |  |
| --- | --- | --- | --- |
| Rode NT-SF1 measurement results | | | |
| Source A | Source B | Signal component A | Signal component B |
| Azi: 0° | Azi: 30° | L3A – L1A = 7.35 dB | L1B – L3B = -1.68 dB |
| Azi: 0° | Azi: 90° | L3A – L7A = 11.61 dB | L7B – L3B = 14.04 dB |
| Azi: 0° | Azi: 135° | L3A – L5A = 11.13 dB | L5B – L3B = 13.95 dB |
| Azi: -30° | Azi: +30° | L2A – L1A = 12.13 dB | L1B – L2B = 8.9 dB |
| Azi: -90° | Azi: +90° | L8A – L7A = 12.63 dB | L7B – L8B = 14.27 dB |
| Average | | 10.97 dB | 9.90 dB |

**4. Summary**

In this document the source proposes an end-to-end test method for assessing spatial capture accuracy of two simultaneous sound sources. In addition, results of two practical experiments are presented. The results indicate that the proposed test method is applicable. Furthermore, the results indicate that the accuracy of decoding channels of multichannel output decreases when the number of microphones (spatial capture accuracy) decreases, which is the expected behaviour.

**References**

[1] Tdoc S4-220729: On ATIAS acoustic performance testing for FOA audio, Dolby Laboratories, Inc.

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

[3] Tdoc S4-221297: IVAS Design Constraints (IVAS-4)

[4] Tdoc S4-210840: Updates to IVAS MASA C Reference Software, Nokia Corporation

]

[

The following method(s) have been incorporated from [7]:

### Direction-of-arrival estimation for MASA input format

**1. Introduction**

The ATIAS work item develops test specifications for objective characterization of terminals for 3GPP immersive services, including conversational services and non-conversational services. A previous input [1] proposed multiple acoustic performance tests for ambisonics capture, including estimation of direction-of-arrival (DOA) from First-order Ambisonics (FOA) audio. However, for Metadata-assisted spatial audio (MASA) [2] the DOA is more convenient and accurate to calculate directly from the analyzed metadata. In this contribution the source proposes a candidate method for estimating DOA for MASA input capture directly from the metadata.

**2. Test method**

*Definition:*

The estimate for DOA can be formed by mapping the direct-to-total ratio times energy weighted azimuth and elevation angles into Cartesian coordinate vectors , and over all subframes and frequency bands.

Where is the index of the frequency bands and is the index of the subframes.

For spatial direction, the angles of direction of arrival in degrees are estimated with equation:

for azimuth and equation:

for elevation. Here, arctan function is assumed to be the computational variant “atan2” that solves the correct quadrant automatically.

*Ideal characteristic:*

The DOA angle calculated from the MASA metadata from the UE capture system, is identical to the ground truth angle to the incident sound

*How to measure:*

The UE is connected via a wireless link to a reference client.

Loudspeakers are placed on a suitable grid (reusing what is already specified, e.g. according to Annex A or B).

The loudspeakers emit sound one at a time. Their incidence angle is known. It remains to be evaluated whether an exponential sweep, pink noise, or speech is the most suitable signal.

The MASA metadata from the reference client are used to calculate the DOA (according to above).

The DOA is compared to the ground truth angle, potentially in several frequency bands and potentially time averaged.

If the sending UE is properly implemented in terms of directionality and the energy ratio analysis for the MASA metadata, the DOA metric is expected to correspond to the ground truth angle.

*How to formulate requirements:*

The DOA angle calculated from the MASA metadata from the UE capture system shall be within some tolerances w.r.t. the ground truth angle to the incident sound.

**3. Experiments**

**3.1 Measurement setup**

Experiments were conducted, to estimate the impact of different DOA estimation methods for MASA capture. Performance of four different test signals were evaluated. Test signals under evaluations were Logarithmic Sine Sweep, Pink Noise, Male speech, and Female speech. Length of each test signal was 5 seconds.

8 different azimuth angles and 6 different elevation angles were evaluated (in total 41 different angles). Azimuth angles were: 0°, 15°, 30°, 45°, 60°, 90°, 135° and 180°, and elevation angles were -60°, -30°, 0°, 30°, 60° and 90°. For each azimuth angle, each elevation angle was measured, except for 90° which was measured only with azimuth angle of 0°.

All measurements were done in an anechoic chamber. Measurement device was an Eigenmike microphone. Measured signals were processed with the IVAS MASA C Reference Software [3] to obtain stereo transport signal and MASA metadata. The MASA input was encoded and decoded with IVAS candidate codec utilizing bitrate of 512 kbit/s. DOA estimations were done from decoded FOA output and from decoded MASA EXT output. Processing flows are illustrated in the figures below.

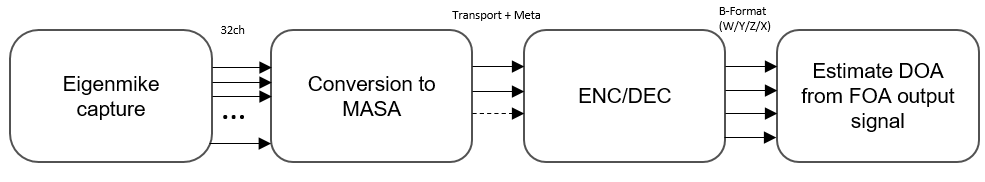


Figure 1 Processing flow for DOA estimation from FOA.

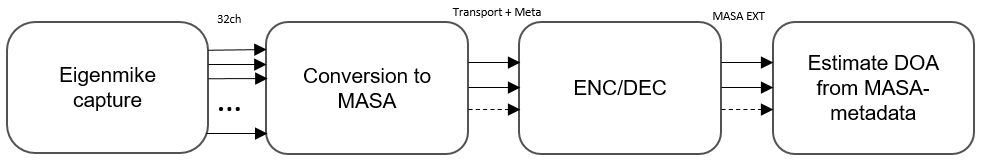


Figure 2 Processing flow for DOA estimation from MASA-metadata.

**3.2 Results**

The mean absolute error of DOA estimates over all measurement angles were calculated for different test signals. Mean absolute error was calculated as follows:

,

where is the estimated DOA and is the ground truth angle at the azimuth and elevation .

The results of estimated DOAs from decoded FOA output are presented in Table 1, and results from decoded MASA EXT output are presented in Table 2. The direct estimations from the MASA metadata are more accurate compared to the estimations made from the FOA output. In addition to more accurate results, the impact of different test signals for the results is mitigated, as the variations of the results between different test signals seems to be lower.

Table 1 Results of DOA analysis from decoded FOA output.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Eigen to MASA to FOA – Mean absolute error of DOA estimate** | | | | | |
|  | **Sweep** | **Pink Noise** | **Male speech** | **Female speech** | **Total** |
|  | 3.37° | 4.16° | 3.69° | 2.75° | 3.49° |
|  | 1.58° | 4.16° | 6.71° | 4.67° | 4.28° |

Table 2 Results of DOA analysis from decoded MASA EXT output.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Eigen to MASA to MASAEXT– Mean absolute error of DOA estimate** | | | | | |
|  | **Sweep** | **Pink Noise** | **Male speech** | **Female speech** | **Total** |
|  | 3.47° | 3.91° | 3.26° | 3.23° | 3.47° |
|  | 1.46° | 1.62° | 1.47° | 1.66° | 1.55° |

**4. Summary**

In this document we present a test method for evaluating DOA directly from MASA metadata. For validating the method, results of DOA estimation from IVAS candidate technology decoded MASA metadata and decoded FOA signal are compared. The results indicate that the proposed method is more accurate DOA estimation method for MASA captures.

**References**

[1] Tdoc S4-221449: On send side audio performance assessment for Immersive Audio Systems –additional metrics

[2] Tdoc S4-221297: IVAS Design Constraints (IVAS-4)

[3] Tdoc S4-210840: Updates to IVAS MASA C Reference Software, Nokia Corporation

]

### Send side audio performance assessment for Immersive Audio Systems in wind noise

The following methods have been incorporated from [2]:

#### Introduction

This test is applicable to UEs capturing immersive audio, including scene-based (e.g. First and Higher Order Ambisonics), binaural, channel-based (e.g. 7.1.4, 5.1, stereo), and object-based audio.

##### test conditions

- The test conditions should follow the Free-field propagation conditions and test environment noise floor described in TS 26.260[2-1].

- wind speed should be 0m/s.

- The size of free-field volume should be large enough to avoid influencing the wind.

**Wind-generator:**

ETSI TS 103 640[2-2] Annex A lists several turbulent wind generation considerations. Some most important requirements are listed here.

- The acoustic noise should be [TBD]dB less than the wind noise at effective frequency band.

- The airflow wide enough to cover the acoustic test equipment and DUT

- The device must keep the target wind speed stable during the test.

NOTE: this test method is used to measure the overload, the acoustic noise requirement of the device used to generate wind needn’t be so strictly as the requirement in ETSI TS 103 640[2-2] and IEC 60268-4[2-6].

##### Setup for terminals

The setup is referred to TS 26.260[2-1] and TS 26.132[2-3]. including the POI, reference point, etc.

Reference point:

Scene-based: geometric centre. [2-1]

Binaural: centre of the acoustic test equipment EEP-to-EEP axis.[2-2]

Object-based: geometric centre of all transducers.

Multichannel: geometric centre of all transducers.

Position:

When using handset UE, headset or hand-free terminal, the terminal should be placed on HATS, according to the corresponding standard or recommended position.

Handsets are given in ITU-T Recommendation P.64 Annex E.[2-5]

Headsets are given in ITU-T Recommendation P.340[2-4]

Measurement points [2-1]:

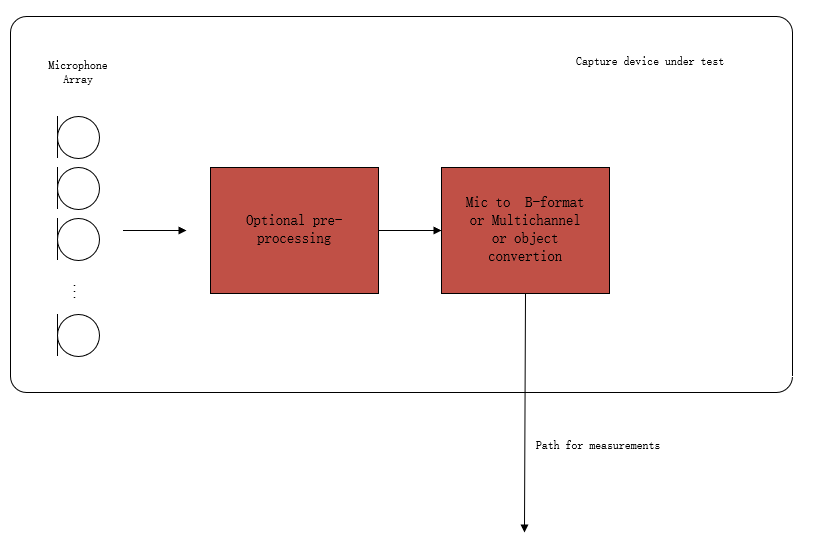


Figure 1: Audio capture block diagram for sending direction measurements

NOTE: The overload point wind speed is a limiting characteristic like the overload sound pressure. All the channels won't affect each other Some processing may cause overload at some special condition, it has damage to communication and is inevitable in windy sensorics, hence the overload caused by processing is included in the result, so select the standard audio signal to measure, and each channel should be measured independently.

##### Definition

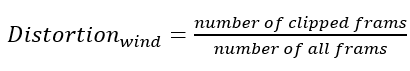
**Overload point wind speed:**

The maximum wind speed at which the distortion of the terminal does not exceed a specified limit(the value of the limit is TBD) for any possible direction of wind incidence and any channel the device outputted.

**Distortion rate:**

Since the clipping appears more frequently at higher wind speed, the probability of clipping appearing in the test signal can indicate distortion rate caused by wind.

So, the source suggests using the rate of clipped frames in all test frames as the distortion rate.



**clipped frames:**

The clipped frames will have any of the following characterises:

1. reach the up limit of signal level (the value needs to test for the DUT)
2. frequency range in high frequency is different from wind noise without clipped.

##### Wind noise measurement method with wind generator for send direction.

1. The UE device under test is mounted in the free-field volume such that its reference point is on the axis of the wind generator exit port and 30 cm from the exit port.

Repeat steps b-c) with an azimuth angular resolution of N degrees for every possible wind direction:

NOTE 2: Since limiting the wind direction in real usage scenarios is not suitable, the test should be implemented in every possible wind direction.

1. The wind generator is the target wind speed on the DUT, and the airflow should cover the DUT.
2. The output of the UE device is stored for offline analysis. The signal should be stored before the wind start and its duration time should larger than 60s.

Increasing the wind speed and repeating the test until the output signal is overloaded or reaches the expected wind speed.

NOTE 3: the wind speed should be selected carefully to avoid the overload damage caused by wind influence on the later test

**Calculation of wind-resistant ability**

The wind-resistant ability represents with the wind speed overload point, which means the terminal can work stable in all directions, and all channels with the wind speed don't exceed the overload point.

The terminal, with several audio channels output, should be calculated by every channel.

The following conclusion has been incorporated from [3]:

#### Recommendations for wind noise simulations for terminal testing

* Wind noise simulations for terminal testing have to be carefully defined under the following constraints:
  + A minimum degree of laminar flow should be ensured by means of e.g., spatial wind speed accuracy, measured at multiple points.
  + A certain degree of reproducibility should be ensured across labs and/or different test equipment solutions.
  + The noise produced by the ventilator/generator should not exceed a certain threshold to minimize the impact on the actual measurements.
  + For employment in typical measurement rooms, a manageable generator size is required – which might limit the aforementioned constraints even further.
* There is currently no specification available or known to the group that uses or defines such a wind noise simulation.
* Possible test methods and performance requirements for ATIAS should be limited to certain form factors/types of terminals. Wind noise simulation for smaller devices is most likely more feasible and reproducible than for larger ones.
* If applicable, specification of a wind noise simulation, test methods and performance requirements should be verified by round robin tests.

## Candidate receiving side test methods and requirements

[

TBD.

]

# References

[1] S4-221449: On send side audio performance assessment for Immersive Audio Systems –additional metrics, Dolby Laboratories Inc.

[1-1] S4-220482: On ATIAS acoustic performance testing for FOA audio, Dolby Laboratories Inc.

[1-2] S4-191167: Description of the IVAS MASA C Reference Software, Nokia Corporation

[1-3] GIF image: <https://commons.wikimedia.org/wiki/Category:Microphone_polar_patterns>

[2] S4-221353: Proposal of wind noise test in ATIAS, Beijing Xiaomi Mobile Software

[2-1] 3GPP TS 26.260: " Objective test methodologies for the evaluation of immersive audio systems."

[2-2] ETSI TS 103 640 V1.2.1-Test Methods and Performance Requirements for Active Noise Cancellation Headsets and other Earphones

[2-3] 3GPP TS 26.132: " Speech and video telephony terminal acoustic test specification."

[2-4] ITU-T Recommendation P.340 (05/2000): "Transmission characteristics and speech quality parameters of hands-free terminals".

[2-5] ITU-T Recommendation P.64 (06/2019): "Determination of sensitivity/frequency characteristics of local telephone systems".

[2-6] IEC 60268-4: Sound system equipment - Part 4: Microphones

[3] S4-230035: Wind noise generation for terminals, HEAD acoustics

[4] S4-230259: Spatial audio capture – spatial separation for multiple acoustic sources based on FOA components, Dolby Laboratories Inc., Nokia Corporation, HEAD acoustics

[5] S4-230189: Add the spatial perception test for stereo UE in ATIAS, Xiaomi

[6] S4-230231: On spatial separation for multiple acoustic sources based on multichannel output, Nokia Corporation

[7] S4-230232: On direction-of-arrival estimation for MASA input format, Nokia Corporation

# Revision history

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** | **Meeting** | **Subject/Comment** | **Old** | **New** |
| 2021-11-17 | SA4#121 | Initial version incorporating S4-221449 and S4-221353 | N/A | 0.1.0 |

1. Stefan Bruhn, e-mail: stefan.bruhn@dolby.com [↑](#footnote-ref-1)
2. Depending on the defined format for the ambisonics, scaling factors may be applied to the X and Y components in this example [↑](#footnote-ref-2)