**Source: IVAS-8a Editor\***

**Title: IVAS Permanent Document IVAS-8a: Test Plan for Selection Phase – Input to editing session based on v.0.8.4**

**Version: v.0.8.5 - Draft**

**Agenda Item: 7.5**

**Document for: Discussion and Agreement**

Document History:

|  |  |  |
| --- | --- | --- |
| v.0.1.0 | 27 August 2021 | Initial Skeleton, Example Test Designs |
| v.0.1.1 | 04 November 2021 | Consolidation of references, editorial changes |
| v.0.2.0 | 19 November 2021 | Editorial corrections |
| v.0.2.1 | 8 February 2022 | Editorial corrections |
| v.0.3.0 | 22 February 2022 | Editorial corrections |
| v.0.3.1 | 5 May 2022 | Port of text from EVS-8b, conclusions from S4aA220004, inclusion of candidate laboratories from S4-211523, S4-220152 |
| v.0.4.0 | 20 May 2022 | Integration of Experiment overview template from S4-220666, editorial corrections |
| v.0.4.1 | 12 August 2022 | Integration of agreed text from S4a220005, editorial corrections |
| v.0.5.0 | 25 August 2022 | Integration of agreed text from S4-220921 |
| v.0.5.1 | 13 January 2023 | Integration in brackets agreed text from S4aA230003, S4aA230004, editorial changes |
| v.0.6.0 | 3 February 2023 | As agreed at the Jan 13 Audio SWG call |
| v.0.6.1 | 20 February 2023 | Integration of text agreed at Feb 3 Audio SWG call from S4aA230016, S4aA230018 (to put in brackets), and S4aA230019 (to put in brackets) |
| v.0.7.0 | 24 February 2023 | As agreed at the Feb 2023 SA4 #122 plenary. Integration of agreed text from s4-230245, Test Plan related remaining work from S4-230254, and edits during the meeting. |
| v.0.7.1 | 17 March 2023 | Integration of agreed text from S4aA230024, completed responsibilities, Selection Time plan skeleton, editorial changes |
| v.0.7.2 | 3 April 2023 | Integration o agreed text and proposals from S4aA230028, S4aA230030, and S4aA230032. |
| v.0.7.3 | 3 April 2023 | Integration of LL pricing indication |
| v.0.7.4 | 17 April 2023 | Integration of agreed text from S4aA230048, changes to allocation of experiments to LLs and Selection Time Plan following discussion and agreement on S4aA230036 and S4aA230052, editorial changes. |
| v.0.7.5 | 20 April 2023 | Integration of agreed text from S4-230555, S4-230557, S4-230560, S4-230561, S4-230562, S4-230596, S4-230602, S4-230605, and S4-230616, plus editorial changes. |
| v.0.8.0 | 21 April 2023 | Editorial changes |
| v.0.8.1 | 27 April 2023 | Input to editing session |
| v.0.8.2 | 28 April 2023 | Output of editing session |
| v.0.8.3 | 8 May 2023 | Integration of agreed text from S4aA230054, S4aA230057, and S4aA230060 |
| v.0.8.4 | 22 May | Integration of agreed text from S4aA230064, S4aA230066, and S4aA230070, editorial corrections. |

*Editor’s note: The following test was originally copied from S4-230254 for further reference. It was further edited to take into account progress achieved at and after the SA4 meeting #122:*

*The following remaining work was identified:*

* *Complete various open aspects wrt sound material*
  + *Complete definition of test material, incl.* 
    - *Background material*
  + *Define method for test item generation for the immersive conversation use-case using P.SUPPL800. This might include*
    - *Definition of scenes*
    - *Selection of method for scene composition*
    - *Collection of required scene data, e.g. FB stereo and spatial impulse responses – see e.g. discussion in [9]*
    - *Collection of corresponding background audio*
  + *Initiate the call for test material collection, e.g. for background and critical generic audio items; this might likely involve* 
    - *setup of a legal framework on non-disclosure of the collected material*
    - *clearance on publicly available material*
    - *set up of infrastructure to collect submitted material*
* *Add detailed list of experiments, incl.*
  + *Definition of references, anchors, CuTs*
  + *Background levels*
  + *FER*
  + *Level*
  + *Loudspeaker or headphone listening*
  + *…*
* *Annexes A – D need to be reviewed, completed and agreed. This includes:*
  + *Sample Instructions to Subjects and Data Collection*
  + *Presentation Orders*
  + *Data to be Provided by LL*
  + *Selection Testing Timeline needs to be completed*
* *In addition, many sections of the permanent document are populated but not formally agreed.*

1. Introduction

This document contains the Test Plan for the Selection Phase of the Codec (IVAS).

1. References, Conventions, and Contacts
   1. Permanent Documents

The following documents provide additional information on the IVAS codec development project.

|  |  |
| --- | --- |
| P-doc | Title |
| IVAS-1 | IVAS Codec Development Overview |
| IVAS-2 | IVAS Project Plan |
| IVAS-3 | IVAS Performance Requirements |
| IVAS-4 | IVAS Design Constraints |
| IVAS-5 | Selection Rules for Selection Phase |
| IVAS-6 | Deliverables for Selection Phase |
| IVAS-7a | Processing Plan for Selection Phase |
| IVAS-7b | Processing Plan for Characterization Phase |
| IVAS-8a | Test Plan for Selection Phase |
| IVAS-8b | Test Plan for Characterization Phase |
| IVAS-9 | IVAS Usage Scenarios |

The latest version of these documents can be found in the following link.

<https://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/IVAS_Permanent_Documents>

* 1. Reference Documents

1. S4-211523: MESAQIN.com and FORCE Technology - SenseLab expression of interest to participate in IVAS codec selection and characterization work.
2. S4-220152: Interest in participation in IVAS codec selection and characterisation phase.
3. Recommendation ITU-R BS.2051-3 (05/2022): Advanced sound system for programme production.
4. S4-211155: On IVAS example test designs, Source: Nokia Corporation.
5. S4-210848: IVAS MASA spatial speech quality evaluation, Source: Nokia Corporation.
6. S4-191167: Description of the IVAS MASA C Reference Software, Source: Nokia Corporation.
7. S4-210840: Updates to IVAS MASA C Reference Software, Source: Nokia Corporation.
8. Recommendation ITU-T P.800 (08/1996): Methods for subjective determination of transmission quality.
9. Recommendation ITU-T P.811 (01/2019): Subjective test methodology for evaluating Speech oriented stereo communication systems over headphones.
10. S4-211151: Example designs for IVAS codec tests, Source: Dolby Laboratories Inc.
11. S4-210836: On reference designs for IVAS codec tests, Source: Dolby Laboratories Inc.
12. Recommendation ITU-R BS.1770-4 (10/2015): Algorithms to measure audio programme loudness and true-peak audio level.
13. ITU-T Handbook of subjective testing practical procedures, 2011.
14. S4-200158: A Reference Audio Renderer for Qualification, Source: Dolby Laboratories Inc.
15. S4-211160: Experience of P.800 for stereo testing, Source: Ericsson LM.
16. S4-130155: EVS Permanent Document EVS-7a: Processing functions for qualification phase.
17. AFsp Programs and Routines: http://www-mmsp.ece.mcgill.ca/Documents/Software/Packages/AFsp/audio/html/AFsp.html.
18. S4aA220005: On reference designs for IVAS codec tests - Update, Source: Dolby Laboratories Inc.
19. S4aA220007 - DCR test experiments for FOA and HOA3 input in 7.0+4 and binaural listening setup.
20. F. Zotter and M. Frank, “All-Round Ambisonic Panning and Decoding,” in J. Audio Eng. Soc., Vol. 60, No. 10, 2012.
21. T22-SG12-220607-TD-GEN-0138!!MSW-E: Draft new ITU-T P.800-series – Supplement P.SUPPL800: ITU-T Rec. P.800 use case examples.
22. Recommendation ITU-R BS.1534 (10/2015): Method for the subjective assessment of intermediate quality level of audio systems.
23. 3GPP TR 26.952: Codec for Enhanced Voice Services (EVS); Performance characterization.
24. S4-030821: PSS/MMS High-Rate Audio Selection Test and Processing Plan, Version 2.2.
25. Audio File Format Specifications: WAVE, <https://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/WAVE/WAVE.html>.
26. AFsp Package <https://www-mmsp.ece.mcgill.ca/Documents/Downloads/AFsp/>.
27. Recommendation ITU-T P.191 (03/2023): Software tools for speech and audio coding standardization.
28. S4-230221: Processing updates for IVAS MASA C Reference Software.
29. IEEE Recommended Practice for Speech Quality Measurements, in IEEE Transactions on Audio and Electroacoustics, vol. 17, no. 3, pp. 225-246, September 1969, doi: 10.1109/TAU.1969.1162058.a
    1. Key Acronyms

20KPB 20-20k Hz Flat band-pass FIR filter

BIT Beijing Institute of Technology

CL Cross-check Laboratory

CuT Codec under Test

DCR Degradation Category Rating

DTX Discontinuous transmission

ESDRU Energy-based Spatial Distortion Reference Unit

EVS Enhanced Voice Services

FB Full Band

FE Frame Erasure

FOA First-Order Ambisonics

GAL Global Analysis Laboratory

HL Host Laboratory

HOA3 Higher-Order Ambisonics, 3rd order

HP50 50 Hz high-pass FIR filter

IVAS Immersive Voice and Audio Services

LKFS Loudness, K-weighted, relative to Full Scale

LL Listening Laboratory

MASA Metadata-Assisted Spatial Audio

MC Material Collection entity

MNRU Modulated Noise Reference Unit

MUSHRA Multi Stimulus test with Hidden Reference and Anchor

PC Proponent Company

SDRU Spatial Distortion Reference Unit

SNR Signal-to-Noise Ratio

SPL Sound Pressure Level

SWB Super Wide Band

WB Wide Band

1. Roles and Responsibilities
   1. Overview of the Selection Test Process

The execution of the IVAS codec Selection subjective testing is under the responsibility of the LLs participating in the Selection Phase.

The execution of the IVAS codec Selection objective testing is under the responsibility of the PC participating in the Selection Phase.

SA4 selects and ETSI will contract the LLs to perform the subjective listening tests described in this document. SA4 selects the languages used in each experiment conducted by each LL. SA4 further selects the HL, the CL, and the GAL to perform respective tasks defined in this document, and ETSI will contract the GAL.

The LLs and volunteering contributors (SA4 companies) shall provide unprocessed 48 kHz sampled clean speech, background material, model parameters, music and mixed content, and critical generic audio content samples to the MC. The format of the material is WAVE [25], 16-bit little endian format. For multi-track audio, the audio tracks are ordered according to Table 5 of IVAS Processing Plan (IVAS-7a).

The material collection entity (MC) shall control that the unprocessed raw material (both artificially created and real recorded) and the model parameters meet the requirements defined by SA4, collect a pool of model parameters and sound materials and choose the model parameters and sound materials to be used in the experiments.

The PC shall deliver a CuT executable to the HL and ETSI.

The CL shall perform cross-check of the HL processing.

The LLs shall insert the raw voting data into the workbook provided by the GAL and forward the workbook directly to the GAL. In addition, each LL must provide a report of experiments to SA4 no later than the document submission deadline for the selection meeting.

* 1. Allocation of Additional Roles

LLs: Mesaqin.com, FORCE Technology [1], HEAD acoustics GmbH/IKS [2], Macquarie University

HL: Contributors of the Public Collaboration

CL: Contributors of the Public Collaboration

MC: Contributors of the Public Collaboration

GAL: [HEAD acoustics GmbH/IKS]

* 1. Responsibilities

Many of the procedures to be followed are defined in this test plan, with further information being given in IVAS Processing Plan (IVAS-7a).

* + 1. Proponent Companies

The specific responsibilities of each PC are:

* Delivery to the HL of a preliminary CuT executable
* Delivery to the HL and ETSI of a final CuT executable
* Develop the processing scripts using the condition lists defined in this document and the processing steps defined in IVAS-7a.
* Provide the randomization playlists for 9 P.SUPPL800 subjective experiments to be described in this document. The playlists will be different for the two tests of the same experiment conducted in different languages. These playlists will be reused for all experiments. Each LL will receive the randomization playlists only for the experiments to be conducted by that LL. The playlists will be delivered in Excel spreadsheet format.
  + 1. Listening Laboratories
       1. Requirements for the Listening Laboratories
* Provide a listening environment that conforms to the requirements in [8] including:
  + Having a background noise level of less than NR-25.
* For each P.SUPPL800 listening test, use subjects that are native speakers of the tested language.
* Provide a person during the training session of each P.SUPPL800 test that is able to answer questions from the subjects in their native language.
* Provide to SA4 the P.SUPPL800 instructions for subjects in each of the languages to be tested by the LL for the Selection Testing.
  + - 1. Tasks for the Listening Laboratories
* LLs shall record or obtain, if not otherwise available, original clean mono speech material (unprocessed 48 kHz sampled speech) for the tests allocated to them and provide it to the MC.
* LLs shall record or obtain, if not otherwise available, original stereo music and mixed content material (unprocessed 48 kHz sampled signals) for the tests allocated to them and provide it to the MC.
* For any tests, LLs may record or obtain original clean mono speech or stereo/immersive material (unprocessed 48 kHz sampled signals) and provide it to the MC.
* LLs shall have the option to declare their material provided to the MC as not available for use by other LLs.
* Obtain from the HL the processed test materials for all tests to be conducted by the LL.
* Perform the listening tests in accordance with this document.
* Delivery to the GAL of all raw voting data using the data delivery file provided by the GAL for all tests to be conducted by the LL.
* Delivery of a LL report to the IVAS Selection Meeting which includes:
  + Confirmation that the LL testing environment conforms to the requirements of the Selection test for all tests conducted by the LL.
  + Provision of listening test instructions for subjects in each of the languages tested by the LL.
  + Age and gender information for the set of subjects used in each listening test, and over all listening tests in each tested language tested by the LL.
  + Discussion of any problems encountered during testing and the solution used to address the problem.
    1. Host Laboratory
       1. E.1 Included tasks

The following list defines the tasks expected to be carried out by the Host Laboratory (HL). The tasks have to be conducted and completed following the schedule for the IVAS Selection phase defined in IVAS-2.

Cross-checking activities:

* Preliminary cross-check using the preliminary CuT executables from the Public Collaboration public repository to identify potential problems.
* Interaction with the PC in order to identify and resolve potential problems.
* Interaction with the CL in order to identify and resolve potential problems.
* Final cross-check with the CL using the final CuT executables, reference executables, and the clean speech, background material, model parameters, music and mixed content, and critical generic audio content provided by the LLs and MC for each experiment.

Processing and Delivery activities:

* Get from the Public Collaboration public repository preliminary CuT executables.
* Deliver the cross-checked processing scripts to SA4.
* Get from the Public Collaboration public repository final CuT executables.
* Receive all source databases and model parameters from the LLs and MC.
* Processing and delivery of all test files per experiment for 2 x 23 experiments to the LLs in phases to meet testing schedules after completion of final cross-check.

Reporting activities:

* Delivery and Presentation of HL report. The report should include a discussion of any problems encountered during the cross-check and processing efforts. The dates for final test material delivery to the LLs should be included.
  + 1. Cross-check Laboratory
       1. Included tasks

The following list defines the tasks expected to be carried out by the Cross-check Laboratory (CL). The tasks have to be carried out following the schedule for the IVAS Selection phase defined in IVAS-2.

Cross-checking activities:

* Preliminary cross-check with the HL using the preliminary CuT executables to identify potential problems.
* Final cross-check with the HL using the final CuT executables from the HL, reference executables, and the clean speech, background material, model parameters,music and mixed content, and critical generic audio content provided by the LLs and MC for each experiment and available from the HL.

Processing and Delivery activities:

* Get processing scripts and all associated executables and parameter files for the experiments from the Public Collaboration public repository.
* Get the preliminary and final CuT executables from the Public Collaboration public repository.
* Receive all source materials from the LLs and MC.
* Processing and cross-check of all test files per experiment for 2 x 23 experiments in phases as needed for the LLs to meet testing schedules.
  + 1. Material Collection Entity (MC)
* MC shall collect the clean mono speech, background material, music and mixed content, real recorded stereo/immersive signals, and a pool of parameters for artificially created stereo/immersive sound material (e.g., impulse responses).
* MC shall control that the unprocessed material (for both artificially created and real recorded content) and parameters for artificially created stereo/immersive sound material meet the requirements defined by SA4.
* MC shall choose the parameters and sound materials to be used in the experiments.
  + 1. Global Analysis Laboratory
       1. Tasks

The following list defines the tasks expected to be carried out by the Global Analysis Laboratory (GAL). The tasks have to be carried out following the schedule for the IVAS Selection phase defined in IVAS-2.

* Provide the raw voting data delivery worksheets for the 46 subjective tests (i.e., 23 experiments, each in two LLs) to the appropriate LLs. Each LL will receive the data delivery only for the experiments to be conducted by that LL. The worksheets will be delivered in Excel spreadsheet format.
* Receive the raw voting data from the LLs in the appropriate data delivery worksheets.
* Conduct statistical Terms of Reference (ToR) tests as specified in clause 3.3.6.2. The ToR tests compare the subjective scores of the CuT against the scores for specified reference conditions. Each subjective experiment contains a number of ToR tests to be computed by the GAL.
* Prepare a GAL report to be presented at the Selection meeting as scheduled in the IVAS Project Plan IVAS-2.
  + - 1. Statistical analysis of results

The GAL report will present the results of the Terms of Reference (ToR) tests using Student's Independent Groups t-test (single-sided at 95% confidence level). Results of the Requirement ToR tests for each experiment will be presented in a table as illustrated Table 1.

In the example below for **Requirement ToR tests**:

* Requirement ToR tests that are passed, (i.e., CuT "not worse than" Requirement) are indicated by **CuT NWT Ref**.
* Requirement ToR tests that are exceeded, (i.e., CuT "better than" Requirement) are indicated by **CuT BT Ref**.
* Requirement ToR tests that are failed (i.e., CuT "worse than" Requirement) are indicated by **CuT WT Ref**.

Table : Example of Requirement ToR test results



* + 1. SA4
* SA4 defines the methods and models for artificial creation of sound material based on original (mono) sound material.
* SA4 defines the stereo/immersive scenes including, e.g., environments/rooms, relative placement of talkers to capture point, and overtalk by talkers.
* SA4 (volunteering members) shall provide the parameter sets for models/methods for artificial creation of sound material based on original (mono) sound material.
* SA4 defines the set of requirements for original sound material (e.g., sampling frequency, formats).
* SA4 (volunteering members) shall record or obtain original stereo/immersive material of Generic audio (unprocessed 48 kHz sampled signals) for BS.1534 experiments.
* SA4 (volunteering members) may record or obtain original clean mono speech material (unprocessed 48 kHz sampled speech).

1. Information relevant to all Experiments
   1. General Technical Notes

Any and all deviations from the specifications contained in this document and the IVAS Processing Plan (IVAS-7a) must be documented and submitted to SA4 along with the experimental report.

* 1. General Consideration of Experiments
* IVAS Selection Test is separated into two main use case scenarios, namely speech centric Immersive conversation, and Generic immersive audio. The Immersive conversation use case targets lower bitrates and the evaluation is done by naïve listeners. The Generic immersive audio assumes higher bitrates and the evaluation is done by experienced listeners.
* Each experiment is performed twice and is tested in two different LLs. Each P.SUPPL800 experiment is run in two different languages with native listeners.
  + 1. Immersive conversation
* Source material:
  + Clean speech
  + Speech with background
  + Music and mixed content
* Input formats:
  + Stereo, including binaural
  + FOA
  + Object-based audio
  + MASA
* Lower bitrates, up to approximately the bitrate having as reference multi-mono EVS at 24.4 kbps per channel, as specified in IVAS Performance Requirements (IVAS-3).
* Including DTX conditions
* Including FE conditions
* Listening environment: headphones
* Test methodology: P.SUPPL800 [21]
  + 1. Generic immersive audio
* Source material: Generic audio
* Higher bitrates
* [No DTX conditions]
* FE conditions
* Input formats:
  + Stereo, [including binaural]
  + FOA
  + HOA3
  + Object-based audio
  + MASA
  + Channel-based audio
* Listening environment:
  + Headphones
  + 7.1 + 4 loudspeaker setup
* Test methodology: BS.1534 (MUSHRA) [22]
  1. Methodology

The following test methodologies shall be used in the IVAS Selection test: P.SUPPL800 [21] will be used in experiments designed to evaluate the Immersive conversation use case scenario, and BS.1534 [22] will be used in experiments designed to evaluate the Generic immersive audio use case scenario. High-level configuration of experiments for both methodologies is outlined below.

* + 1. P.SUPPL800
* Test duration should not exceed 2 hours per listening panel. Typical value of voting period was used for estimation of test durations, but actual voting period is not specified.
* Randomizations constructed under randomized blocks experimental design described in [13].
* 6 categories for each test. Categories are defined for each experiment separately.
* 6 samples/category (1 for each listening panel) plus 1 sample/category for preliminaries.
* 30 naïve listeners, 6 listening panels (5 listeners per panel), each panel with an independent randomization
* 180 votes for each condition.
* Total number of conditions for each experiment: 36
* Number of trials: number of test conditions x 6 talkers/categories = 216 trials.
  + 1. BS.1534
* Number of items per experiment: 12
* 14 experienced listeners
* Maximum total number of conditions: 8
* Number of anchor conditions: 2
  + Direct
  + 7 kHz low-pass anchor

Note: the exact number of anchors may vary depending on actual experiment.

* 1. Opinion Scales

Table 3 defines opinion scale used for ITU-T P.SUPPL800 DCR test. Instructions in English for the P.SUPPL800 test are provided in Annex A.

Table : Opinion scale for ITU-T P.SUPPL800 DCR test

|  |  |
| --- | --- |
| **Impairment** | **Scale** |
| No impairment | 5 |
| Small impairment | 4 |
| Moderate impairment | 3 |
| Large impairment | 2 |
| Very large impairment | 1 |

* 1. Material

All audio material shall be sampled at 48 kHz with Full Band (FB) content. The audio material is to be delivered to the HL as 16-bit little endian WAVE format files [25] following the naming convention provided in the IVAS Processing Plan (IVAS-7a). For multi-track audio, the audio tracks are ordered according to Table 5 of IVAS Processing Plan (IVAS-7a). Additionally, it should be verified that the audio material can be processed with the AFsp package tools [26].

The following categories of audio content will be used in IVAS Selection Test using P.SUPPL800:

* Clean speech: Except for experiment P800-6 (1 object), each sample contains two (or more) different talkers in conversation scenario. The talkers transition from one to another as in natural conversation, possibly with partial overlap.
* Speech with background: the details about the environment are specified in Annex E.
* Music and Mixed content – categories specified in section 4.5.3.

The following category of audio content will be used in IVAS Selection Test using BS.1534:

* Generic audio – critical generic audio items including speech with and/or without background, music, mixed.
  + 1. Speech Material for P.SUPPL800 testing

Except for the experiment P800-3 (Music and mixed content stereo experiment), P.SUPPL800 test experiments will use artificially created immersive audio. LLs shall provide clean speech mono audio samples and music and mixed content stereo samples. SA4 would provide scene descriptions and scripts to create the immersive audio.

The recording SNR should be in accordance with P.800 at least 40dB but preferably 50dB or higher, The leading and trailing inactivity portions should be shorter than [20] ms. The reverberation time RT60 should be in accordance to P.800 less than 500ms, preferably below 200ms. The length of the sentences should typically correspond to the length of traditional Harvard sentences [29].

* + 1. Background Material
* Immersive conversation use case scenario (P.SUPPL800 testing): A mix-based approach using separate background recordings will be used. The minimum lengths of noise files shall be 80 s.

The following guideline is applied to the noise types used.

Car noise is intended to test the performance of the codec under steady state background noise and should be recorded in a moving car. A constant speed between 80km/h (50mph) and 110km/h (70mph) is recommended. The make and model of the car should be reasonably common in the country of the recording. Typically, the windows of the car should be closed, and the radio turned off.

Office noise is intended to represent a typical office environment. This noise type should also contain typical office sounds, such as keyboard noise, computer fans, telephones ringing, printers, air conditioner, etc.

Street noise is intended to represent a typical street environment. It should contain unsteady traffic noise for example recorded at traffic lights where cars stop, human noise such as steps. It should not contain speech, but baby cries are allowed.

* Generic immersive audio use case scenario (BS.1534 testing): Primarily, full recordings of complete immersive scenes including background will be used. A mix-based approach might be used in addition.
  + 1. Music and Mixed Content Material for P.SUPPL800 testing

Music and mixed content samples shall contain meaningful contents and the duration of each sample shall be approximately 8 and at least 7seconds. The following categories shall be used:

* Classical music
* Modern instrumental music
* Modern vocal music
* Radio Jingle
* Movie Trailer
* Advertisement

This means that LL shall provide 7 samples per category, 6 for evaluation and 1 for preliminaries.

* + 1. Critical Generic Audio Items for BS.1534 testing
       1. Steps of Critical Test Item Selection

The following steps are based on [24]:

* Call for test material according to the generic audio signal categories described below.
* MC collects candidate material submitted in response to the call and selects a number of critical items to be used in the Selection test.
* MC selects a limited set of training items to be used in a training phase.
  + - 1. Test Material

First, a call will be sent out for test material according to a number of generic audio signal categories as specified below. All 3GPP members are invited to submit test material to MC. The submitting organization shall assign the items to the below-mentioned audio signal categories. Then, MC will identify 12 critical items per experiment, plus four items for training, which are representative for assumed typical IVAS application scenarios.

Generic audio signal categories:

Stereo – generic stereo audio signals with a focus on music categories:

* Pop, with and/or without vocals
* Classic, with and/or without vocals
* Single instruments
* a capella vocals, solo and/or choir
* Mixed speech and music
* Speech with and/or without background noise

Multi-Channel (5.1 and 7.1.4) – generic channel-based audio signals from produced content:

* Music including concerts with live audience
* Film soundtracks with and/or without speech dialogue
* Effects (e,g, nature, city/transport sounds)

Scene-Based Audio / MASA – generic immersive audio signals in the form of complex scenes, captured and/or produced content which may or may not include speech:

* Nature sounds (e.g. forest, water, wind)
* City sounds (e.g. traffic, bus, train)
* Music including concerts with live audience
* Babble-like sound (e.g. market, restaurant, conference)
* Event/Sport-like sound
* Conferencing scene with and/or without background noise/music

Object-Based Audio - Realistic immersive audio signals, e.g.:

* Scenarios comprising voice, music, background objects.
* Conversational scenarios of several talkers with or without background, with or without partial overtalk of no more than two talkers. Talkers may be moving around the scene at natural pace.

The length in time of the items will be 10 s at a maximum.

MC will further maintain and report to SA4 a list indicating the number of proposed items per submitting organization.

In case the submitted material is insufficient/inadequate to conduct the tests, MC will add the missing test items.

* + - 1. Training material

Limited material will be used in the training phase in which the subjects familiarize with the testing methodology and environment.

The training will be conducted with four sound items. These items will be identified by MC and shall not be re-used in the blind grading phase. The training phase shall be executed as a separate short MUSHRA session.

* 1. Listening Systems and Listening Environments

The IVAS Selection Test will use the following listening systems:

* Stereo headphones for binaural listening, e.g.:
  + Beyerdynamic DT 770 Pro for P.SUPPL800 experiments
  + Sennheiser HD 650 for BS1534 experiments
* Loudspeaker listening system – 7.1+4 loudspeaker setup [3].
  1. Experimental Procedure
     1. Experimental Procedure for P.SUPPL800 experiments

Initially the experimenter should provide a written copy of the experiment instructions to the listeners. When the listeners have acknowledged that they understand the instructions, they will be presented with a practice session to rate the preliminary conditions. After the practice session has been completed, the experimenter should ask if there are any questions. Only questions about the rating procedures or the meaning of the instructions should be answered. Any technical questions on matters such as the experimental methodology or details of the types of distortions they are rating must not be answered.

* 1. Results and Analysis

On completion of the experiments, the LLs must provide the raw voting data to the GAL for the purpose of performing a global analysis. The raw voting data for each experiment shall be delivered in the spreadsheet provided by the GAL for that purpose.

1. Subjective Experiments

The purpose of the 23 experiments (Experiments P800-1 – P800-9, and BS1534-1a – BS1534-7b) is to evaluate the performances of the IVAS codec candidate algorithm with respect to the performance requirements and objectives defined in (IVAS-3).

The details provided in this section and in corresponding Annexes are those that are specific to each particular experiment. Generic information can be found in Section 4. Therefore, the LLs should use the information in Section 4 in conjunction with the information given in this section and Annexes.

Table 4 shows high-level overview of P.SUPPL800 experiments. Table 5 shows high-level overview of BS.1534 experiments. Finally, Table 6 shows allocation of experiments to LLs and languages proposed by LLs for each P.SUPPL800 experiment.

Detail conditions for each subjective experiment are defined in Annex E for P.SUPPL800 experiments and in Annex F for BS.1534 experiments.

Table : High-level overview of P.SUPPL800 experiments

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Exp | Input format | Source material | Listening environment | Bitrates kbps | FER/jitter | DTX |
| P800-1 | Stereo | Clean speech | Headphones | ≤ 48 | ≤ 5% | Y |
| P800-2 | Stereo | Speech+Background | Headphones | ≤ 64 | ≤ 5% | Y |
| P800-3 | Stereo | Mixed & Music | Headphones | ≤ 64 | ≤ 5% | Y |
| P800-4 | FOA | Clean speech | Headphones | ≤ 96 | ≤ 5% | N |
| P800-5 | FOA | Speech+Background | Headphones | ≤ 96 | ≤ 5% | Y |
| P800-6 | 1 Object | Clean speech | Headphones | ≤ 64 | ≤ 5% | Y |
| P800-7 | 2 Objects | Clean speech | Headphones | ≤ 64 | ≤ 5% | Y |
| P800-8 | MASA | Clean speech | Headphones | ≤ 80 | ≤ 5% | N |
| P800-9 | MASA | Speech+Background | Headphones | ≤ 80 | ≤ 5% | Y |

Table : High-level overview of BS.1534 experiments

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Exp** | **Input format** | **Source material** | **Listening environment** | **Bitrates kbps** | **FER/jitter** | **DTX** |
| BS1534-1a | Stereo | Generic Audio | Headphones |  | ≤ 5% | N |
| BS1534-1b | Stereo | Generic Audio | Headphones |  | ≤ 5% | N |
| BS1534-2a | 5.1 | Generic Audio | 5.1 |  | ≤ 5% | N |
| BS1534-2b | 5.1 | Generic Audio | 5.1 |  | ≤ 5% | N |
| BS1534-3a | 7.1.4 | Generic Audio | 7.1 + 4 |  | ≤ 5% | N |
| BS1534-3b | 7.1.4 | Generic Audio | 7.1 + 4 |  | ≤ 5% | N |
| BS1534-4a | FOA | Generic Audio | Headphones |  | ≤ 5% | N |
| BS1534-4b | FOA | Generic Audio | Headphones |  | ≤ 5% | N |
| BS1534-5a | HOA3 | Generic Audio | Headphones |  | ≤ 5% | N |
| BS1534-5b | HOA3 | Generic Audio | 7.1 + 4 |  | ≤ 5% | N |
| BS1534-6a | Objects | Generic Audio | Headphones |  | ≤ 5% | N |
| BS1534-6b | Objects | Generic Audio | Headphones |  | ≤ 5% | N |
| BS1534-7a | MASA | Generic Audio | Headphones |  | ≤ 5% | [Y/N] |
| BS1534-7b | MASA | Generic Audio | Headphones |  | ≤ 5% | N |

Notes:

* Stereo may include binauralized samples (without head tracking).
* For inputs 7.1+4, FOA, HOA3, Objects & MASA vertical dimension is assumed in the samples.
* DTX on/off is assumed within the same experiment, where DTX on is used for relevant conditions.
* All experiments except for stereo P.SUPLL800 experiments are assumed Full Band experiments, i.e., the direct reference condition is always FB. P.SUPPL800 stereo experiments are SWB experiments.

Note: the assumption is to have at least 6 weeks for subjective testing, from receiving the processed samples to delivering the listening results, assuming a dry run could be available a week before.

Note: The databases are not assumed pristine.

Minimum requirements for P.SUPPL800 experiments: 6 talkers (3 male + 3 female) per experiment, 14 single sentences per talker.

Table 6 shows allocation of LLs so that each experiment is conducted twice, each time by a different LL. For P.SUPPL800 experiments, each experiment is run twice with different languages.

Table : Allocation of experiments to LLs and P.SUPPL800 languages

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Exp** | **Source material** | **Listening environment** | **Languages** | | | | **Pricing** |
| **Force Technology** | **Head Acoustics/ IKS** | **MQ University** | **Mesaqin** | **Euros** |
| P800-1 | Clean speech | Headphones | JAP |  |  | FR | 36000 |
| P800-2 | Speech+Background | Headphones |  | GER |  | MAN | 36000 |
| P800-3 | Mixed & Music | Headphones | DAN |  |  | MAN | 36000 |
| P800-4 | Clean speech | Headphones | JAP |  | ENG |  | 36000 |
| P800-5 | Speech+Background | Headphones | DAN | GER |  |  | 36000 |
| P800-6 | Clean speech | Headphones | JAP |  | ENG |  | 36000 |
| P800-7 | Clean speech | Headphones | DAN |  |  | MAN | 36000 |
| P800-8 | Clean speech | Headphones | DAN | GER |  |  | 36000 |
| P800-9 | Speech+Background | Headphones | JAP |  |  | FR | 36000 |
| BS1534-1a | Generic Audio | Headphones | x |  |  | x | 20000 |
| BS1534-1b | Generic Audio | Headphones |  | x |  | x | 20000 |
| BS1534-2a | Generic Audio | 5.1 | x | x |  |  | 26000 |
| BS1534-2b | Generic Audio | 5.1 | x | x |  |  | 26000 |
| BS1534-3a | Generic Audio | 7.1 + 4 | x | x |  |  | 26000 |
| BS1534-3b | Generic Audio | 7.1 + 4 | x | x |  |  | 26000 |
| BS1534-4a | Generic Audio | Headphones | x |  |  | x | 20000 |
| BS1534-4b | Generic Audio | Headphones |  | x |  | x | 20000 |
| BS1534-5a | Generic Audio | Headphones | x |  |  | x | 20000 |
| BS1534-5b | Generic Audio | 7.1+4 | x | x |  |  | 26000 |
| BS1534-6a | Generic Audio | Headphones |  | x |  | x | 20000 |
| BS1534-6b | Generic Audio | Headphones |  | x |  | x | 20000 |
| BS1534-7a | Generic Audio | Headphones |  | x |  | x | 20000 |
| BS1534-7b | Generic Audio | Headphones |  | x |  | x | 20000 |
| Total |  | | | | | | **634000** |

JAP = Japanese

FR = French

GER = German

MAN = Mandarin

DAN = Danish

ENG = English

1. Sample Instructions to Subjects and Data Collection

These instructions shall be translated properly to the LL’s language and be given to the listeners. The instructions given to the listeners shall be provided for information in the LL report.



|  |
| --- |
| **INSTRUCTIONS TO NAÏVE LISTENERS FOR P.SUPPL800 DCR TEST**  In this experiment you will be evaluating systems that might be used for future immersive telecommunication services using spatial audio. Spatial audio means that you can locate various sound sources around yourself. For example, a first talker may appear to talk from the left-hand side and a second talker from the right-hand side, a talker can be moving, etc.  In each trial, you will hear a *reference* audio sample followed by a *test* sample. The *test* sample has the same content as the *reference* sample, but it is possibly impaired after it has passed through a telecommunication system.  Your task is to evaluate the overall impairment of the second sample compared to the first sample, comprising both degradations in the sound quality (e.g., due to additional noise, roughness, clicks or other distortions), and/or differences in the spatial representation (e.g., sound source location, distance, spatial width, movement, etc.).  You should listen carefully to both samples within a trial. When they have finished, select the category that best describes your overall impression about the amount of any impairment you can perceive in the second sample relative to the first sample:  5 - No impairment  4 - Small impairment  3 - Moderate impairment  2 - Large impairment  1 - Very large impairment  Note that the level of impairments present in different *test* samples is expected to span the complete range of the rating scale during the experiment.  Please do not discuss your opinions with other listeners participating in the experiment. If you have any questions, please ask the test administrator. |

**Voting screen** (after playback of sample pair)

|  |
| --- |
| Please rate the OVERALL IMPAIRMENT of the second sample compared to the first sample:  5 - No impairment  4 - Small impairment  3 - Moderate impairment  2 - Large impairment  1 - Very large impairment |

1. P.SUPPL800 Presentation Orders

[

The GAL will provide the Presentation Order for each P.SUPPL800 experiment to the Listening Lab assigned to conduct the test. The presentation order for each experiment has been developed by the GAL using a partially-balanced randomized-blocks experimental design and sample allocation for conducting Dependent Groups Student T-tests for the specified Terms of Reference tests. Each Presentation Order includes six blocks, corresponding to six categories and includes a separate presentation sequence for each of 6 panels of subjects. The Presentation Orders will be delivered to the Listening Labs in the form of Data Delivery Excel spreadsheets described in Annex C. Presentation Orders will be cross-checked before the actual listening tests start.

]

1. Data to be Provided by LL

[

The GAL will provide a Data Delivery spreadsheet for each experiment to the Listening Lab assigned to conduct the test. Attached to this Annex is an Excel file containing an example data delivery spreadsheet for a single experiment - a Music and Mixed content DCR test including 36 conditions (c01,c02, ... ,c36) and 6 categories (a1,a2,a3,a4,a5,a6) for each of 6 panels of subjects. For each panel, the data delivery spreadsheet includes 6 blocks of 36 trials each. The table below illustrates the data delivery spreadsheet for the 36 trials in Block-1 for Panel-1. For each trial, the table shows both the Reference file (condition c01) and the Test file (conditions c01-c36) followed by 5 data cells, one per subject, to be filled by the Listening Lab with the raw voting data provided by the subjects. In the table and in the attached file, the file naming convention is as follows:

***leee*a*y*s*zz*.cnn.wav** ***l*** = Listening Lab, ***eee*** = Experiment, **a*y*** = Category, **s*zz*** = Sample, **c*nn*** = Condition (see IVAS-7a)



]

Editor’s note: Text and the table copied from EVS Selection and need to be updated.

1. Selection Testing Timeline

Table H.1: Testing Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| **Month** | **Meeting/date** | **Task** | **Active Parties** |
| April-2023 | April 17-21 | 3GPP SA4 e-meeting #123 |  |
| May-2023 | May 19 | All payments of the Funding Agreement (FA), including the second payment, are expected to be received by ETSI at the latest by May 19. ETSI will ensure that invoices for the payments are sent out in time, in accordance with the FA. | ETSI |
| May-2023 | May 19 | Stable version (ready for signature) exists for proper legal framework among proponent companies (includes host lab, cross-check lab), listening labs, GAL to cover use of audio test material (unprocessed and processed), and test results (raw voting data). |  |
| May-2023 | May 22-26 | 3GPP SA4 meeting #124  Finalization of selection phase documents   * Selection Rules (IVAS-5) * Selection Deliverables (IVAS-6)          Selection Processing Plan (IVAS-7a)         Selection Test Plan (IVAS-8a) |  |
| June-2023 | June 2 | Proper legal framework exists (signed) among proponent companies(includes host lab, cross-check lab), listening labs, GAL to cover use of executables, source codes, audio test material (unprocessed and processed), and test results (raw voting data). |  |
| June-2023 | June 16, 17:00 CEST | Pre-release: Submission of IVAS codec candidate executable (floating-point code) for Selection testing |  |
| June-2023 | June 27, 17:00 CEST | Final release: Submission of IVAS codec candidate executable (floating-point code) for Selection testing. |  |
| June-2023 | June 27 | Start of processing audio samples to be delivered to LLs |  |
| June-2023 |  | Delivery of processed audio samples to LLs |  |
| 21-Aug |  | 3GPP SA4 meeting #124 - IVAS codec Selection meeting |  |
|  |  |  |  |

Editor’s note: Only agreed text copied from IVAS-2 in the table that needs to be completed.

1. P.SUPPL800 Experiments
   1. Experiment P800-1: Stereo Clean Speech Test

[

* + 1. Experiment setup

Tables E.1.1 to E.1.3 show conditions to be used for this experiment, list of preliminaries and full list of conditions, respectively.

Table E.1.1: Factors and conditions for Experiment P800-1

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Candidate | 1 | CuT |
| Bitrates | 5 | 13.2, 16.4, 24.4, 32, 48 |
| DTX | 2 | DTX on and off |
| Input level | 1 | -26 LKFS |
| Input frequency mask | 1 | HP50 |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, 5% [1 delay/jitter profile] |
|  |  |  |
| **Codec references** |  |  |
| Codec references  Bitrates | 1  7 | EVS dual mono  2x7.2, 2x8.0, 2x9.6, 2x13.2, 2x16.4, 2x24.4, 2x32 kbps |
| Input level  DTX | 1  2 | -26 LKFS  DTX on and off |
| Input frequency mask | 1 | HP50 |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, 5% [1 delay/jitter profile] |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | -26 LKFS |
| P.50 MNRU  ESDRU | 4  3 | Q=16, 20, 24, 28 dB  *α* = 0.1, 0.4, 0.7 |
| Input frequency mask | 1 | HP50 |
| **Common Conditions** |  |  |
| Test item generation: pre-processing incl. spatialization | 1 | Model-based relying on convolution of raw mono clean speech sentences with Room Impulse Responses respective to various talker positions relative to a capture point as described in the ITU-T Reverberation Tool [27] |
| Audio sampling frequency/bandwidth | 1 | 32 kHz/maximum available audio bandwidth up to SWB |
| Kind of samples | 1 | Sentence pair uttered by different talkers and genders (3 male and 3 female) |
| Number of categories | 6 | 6 Different environments and talker interactions |
| Number of samples | 7 | 6 + 1 (preliminaries) samples per category |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve listeners |
| Randomizations | 6 | 6 panels of 5 listeners |
| Rating Scale | 1 | P.SUPPL800 |
| Listening System | 1 | Headphones, in accordance with clause 4.6 |
| Listening Environment | 1 | No room noise, in accordance with clause 4.6 |
| Languages | 2 | Japanese, French |

Table E.1.2: Preliminaries for Experiment P800-1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Trial #** | **Label** | **Sample** | **Condition** | **Bitrate** | **FER/Profile** |
| 1 | c19 |  | EVS | 2x13.2 | [5%, profile] |
| 2 | c02 |  | MNRU Q=28 dB | - | - |
| 3 | c06 |  | ESDRU = 0.7 | - | - |
| 4 | c13 |  | EVS | 2x16.4 | No errors |
| 5 | c08 |  | ESDRU = 0.1 | - | - |
| 6 | c16 |  | EVS | 2x7.2 | [5%, profile] |
| 7 | c04 |  | MNRU Q=20 dB | - | - |
| 8 | c01 |  | Reference | - | - |
| 9 | c15 |  | EVS | 2x32 | No errors |
| 10 | c07 |  | ESDRU = 0.4 | - | - |
| 11 | c05 |  | MNRU Q=16 dB | - | - |
| 12 | c21 |  | EVS | 2x24.4 | [5%, profile] |

Table E.1.3: Test conditions for Experiment P800-1,  
clean speech under clean and impaired channel conditions [including delay/jitter profiles]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Label** | **Condition** | **Bitrate [kbps]** | **DTX** | **FER/Profile** | **ToR** |
| c01 | Reference | - | - | - |  |
| c02 | MNRU Q=28 dB | - | - | - |  |
| c03 | MNRU Q=24 dB | - | - | - |  |
| c04 | MNRU Q=20 dB | - | - | - |  |
| c05 | MNRU Q=16 dB | - | - | - |  |
| c06 | ESDRU | - | - | - |  |
| c07 | ESDRU | - | - | - |  |
| c08 | ESDRU | - | - | - |  |
| c09 | EVS | 2x7.2 | off | No errors |  |
| c10 | EVS | 2x8 | off | No errors |  |
| c11 | EVS | 2x9.6 | off | No errors |  |
| c12 | EVS | 2x13.2 | off | No errors |  |
| c13 | EVS | 2x16.4 | off | No errors |  |
| c14 | EVS | 2x24.4 | off | No errors |  |
| c15 | EVS | 2x32 | off | No errors |  |
| c16 | EVS | 2x7.2 | off | [5%, profile] |  |
| c17 | EVS | 2x8 | off | [5%, profile] |  |
| c18 | EVS | 2x9.6 | off | [5%, profile] |  |
| c19 | EVS | 2x13.2 | off | [5%, profile] |  |
| c20 | EVS | 2x16.4 | off | [5%, profile] |  |
| c21 | EVS | 2x24.4 | off | [5%, profile] |  |
| c22 | EVS | 2x32 | off | [5%, profile] |  |
| c23 | EVS | 2x13.2 | on | No errors |  |
| c24 | EVS | 2x8 | on | 5% |  |
| c25 | CuT | 13.2 | off | No errors | NWT c10 OR BT c09 |
| c26 | CuT | 16.4 | off | No errors | NWT c11 OR BT c10 |
| c27 | CuT | 24.4 | off | No errors | NWT c12 OR BT c11 |
| c28 | CuT | 32 | off | No errors | NWT c13 OR BT c12 |
| c29 | CuT | 48 | off | No errors | NWT c15 OR BT c14 |
| c30 | CuT | 13.2 | off | [5%, profile] | NWT c17 OR BT c16 |
| c31 | CuT | 16.4 | off | [5%, profile] | NWT c18 OR BT c17 |
| c32 | CuT | 24.4 | off | [5%, profile] | NWT c19 OR BT c18 |
| c33 | CuT | 32 | off | [5%, profile] | NWT c20 OR BT c19 |
| c34 | CuT | 48 | off | [5%, profile] | NWT c22 OR BT c21 |
| c35 | CuT | 24.4 | on | No errors | NWT c23 |
| c36 | CuT | 13.2 | on | 5% | NWT c24 |

* + 1. Content type categories and scene definitions

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Room | Reverb | Microphone Setup |  | Overtalk [s](2 | Bandwidth( | Talker positions(3 | Talker selection by panel |
| cat 1 | small | anechoic | A-B (100cm) |  | 1 | Max [available up to SWB] | 1-7  5-3  2-6  4-1  3-4  7-2 | P1: f1m1 P2: m2f2 P3: f3m3 P4: m1f1 P5: f2m2 P6: m3f3 |
| cat 2 | large | anechoic | A-B (150cm) |  | -1 | max [available up to SWB] | 5-11  1-6  3-7  5-8  9-7  10-9 | P1: m3f3 P2: f1m1 P3: m2f2 P4: f3m3 P5: m1f1 P6: f2m2 |
| cat 3 | small | anechoic | M-S |  | -1 | max [available up to SWB] | 1-7  5-3  2-6  4-1  3-4  7-2 | P1: f2m2 P2: m3f3 P3: f1m1 P4: m2f2 P5: f3m3 P6: m1f1 |
| cat 4 | small | echoic | A-B (100cm) |  | 1 | max [available up to SWB] | 1-7  5-3  2-6  4-1  3-4  7-2 | P1: m1f1 P2: f2m2 P3: m3f3 P4: f1m1 P5: m2f2 P6: f3m3 |
| cat 5 | large | echoic | A-B (150cm) |  | 1 | max [available up to SWB] | 2-8  9-4  6-10  11-8  10-12  12-1 | P1: f3m3 P2: m1f1 P3: f2m2 P4: m3f3 P5: f1m1 P6: m2f2 |
| cat 6 | small | echoic | Binaural |  | -1 | max [available up to SWB] | 1-7  5-3  2-6  4-1  3-4  7-2 | P1: m2f2 P2: f3m3 P3: m1f1 P4: f2m2 P5: m3f3 P6: f1m1 |

**Notes:**

**(2** Overtalk [s] means the duration in seconds by which the two sentences in the sound item uttered by different talkers are overlapping. A negative number means that there is a corresponding pause between the two sentences.

**(3** The talker positions are part of the scene definition of the different categories. They correspond to the talker positions as depicted in Figures 14.3 and 14.5 of [27] for the large and the small room, respecively.

Editor’s note: Reference conditions (following IVAS-3) missing for the CuT conditions in square brackets.

]

* 1. Experiment P800-2: Stereo Speech and Background Test

[

* + 1. Experiment setup

Tables E.2.1 to E.2.3 show conditions to be used for this experiment, list of preliminaries and full list of conditions, respectively.

Table E.2.1: Factors and conditions for Experiment P800-2

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Candidate | 1 | CuT |
| Bitrates | 6 | 13.2, 16.4, 24.4, 32, 48, 64 |
| DTX | 2 | DTX on and off |
| Input level | 1 | -26 LKFS |
| Input frequency mask | 1 | HP50 |
| Noise | 1 | 15dB |
| Error Conditions | 0 | No errors |
|  |  |  |
| **Codec references** |  |  |
| Codec references  Bitrates | 1  8 | EVS dual mono  2x7.2, 2x8.0, 2x9.6, 2x13.2, 2x16.4, 2x24.4, 2x32, 2x64 kbps |
| Input level  DTX | 1  2 | -26 LKFS  DTX off, DTX on |
| Input frequency mask | 1 | HP50 |
| Noise | 1 | 15dB |
| Error Conditions | 0 | No errors |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | -26 LKFS |
| P.50 MNRU  ESDRU | 5  4 | Q=12, 17, 22, 27, 32 dB  *α* = 0.1, 0.3, 0.5, 0.7 |
| Input frequency mask | 1 | HP50 |
| **Common Conditions** |  |  |
| Test item generation: pre-processing incl. spatialization | 1 | Model-based relying on convolution of raw mono clean speech sentences with Room Impulse Responses respective to various talker positions relative to a capture point as described in the ITU-T Reverberation Tool [27] |
| Audio sampling frequency/bandwidth | 1 | 32 kHz/maximum available audio bandwidth up to SWB |
| Kind of samples | 1 | Sentence pair uttered by different talkers and genders (3 male and 3 female) |
| Number of categories | 6 | 6 Different environments with different background types (car,office, street) and talker interactions |
| Number of samples | 7 | 6 + 1 (preliminaries) samples per category |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve listeners |
| Randomizations | 6 | 6 panels of 5 listeners |
| Rating Scale | 1 | P.SUPPL800 |
| Listening System | 1 | Headphones, in accordance with clause 4.6 |
| Listening Environment | 1 | No room noise, in accordance with clause 4.6 |
| Languages | 2 | German, Mandarin |

Table E.2.2: Preliminaries for Experiment P800-2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Trial #** | **Label** | **Sample** | **Condition** | **Bitrate** | **DTX** |
| 1 | c22 |  | EVS | 2x13.2 | on |
| 2 | c02 |  | MNRU Q=32 dB | - | - |
| 3 | c07 |  | ESDRU = 0.7 | - | - |
| 4 | c15 |  | EVS | 2x16.4 | off |
| 5 | c10 |  | ESDRU = 0.1 | - | - |
| 6 | c19 |  | EVS | 2x7.2 | on |
| 7 | c04 |  | MNRU Q=22 dB | - | - |
| 8 | c01 |  | Reference | - | - |
| 9 | c17 |  | EVS | 2x32 | off |
| 10 | c08 |  | ESDRU = 0.5 | - | - |
| 11 | c05 |  | MNRU Q=17 dB | - | - |
| 12 | c24 |  | EVS | 2x24.4 | on |

Table E.2.3: Test conditions for Experiment P800-2,  
clean speech with background noise under clean channel conditions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Condition** | **Bitrate [kbps]** | **DTX** | **ToR** |
| c01 | Reference | - | - |  |
| c02 | MNRU Q=32 dB | - | - |  |
| c03 | MNRU Q=27 dB | - | - |  |
| c04 | MNRU Q=22 dB | - | - |  |
| c05 | MNRU Q=17 dB | - | - |  |
| c06 | MNRU Q=12 dB |  |  |  |
| c07 | ESDRU *α* = 0.7 | - | - |  |
| c08 | ESDRU *α* = 0.5 | - | - |  |
| c09 | ESDRU *α* = 0.3 |  |  |  |
| c10 | ESDRU *α* = 0.1 | - | - |  |
| c11 | EVS | 2x7.2 | off |  |
| c12 | EVS | 2x8 | off |  |
| c13 | EVS | 2x9.6 | off |  |
| c14 | EVS | 2x13.2 | off |  |
| c15 | EVS | 2x16.4 | off |  |
| c16 | EVS | 2x24.4 | off |  |
| c17 | EVS | 2x32 | off |  |
| c18 | EVS | 2x48 | off |  |
| c19 | EVS | 2x7.2 | on |  |
| c20 | EVS | 2x8 | on |  |
| c21 | EVS | 2x9.6 | on |  |
| c22 | EVS | 2x13.2 | on |  |
| c23 | EVS | 2x16.4 | on |  |
| c24 | EVS | 2x24.4 | on |  |
| c25 | EVS | 2x32 | on |  |
| c26 | CuT | 13.2 | off | NWT c12 OR BT c11 |
| c27 | CuT | 16.4 | off | NWT c13 OR BT c12 |
| c28 | CuT | 24.4 | off | NWT c14 OR BT c13 |
| c29 | CuT | 32 | off | NWT c15 OR BT c14 |
| c30 | CuT | 48 | off | NWT c17 OR BT c16 |
| c31 | CuT | 64 | off | NWT c18 OR BT c17 |
| c32 | CuT | 13.2 | on | NWT c20 OR BT c19 |
| c33 | CuT | 16.4 | on | NWT c21 OR BT c20 |
| c34 | CuT | 24.4 | on | NWT c22 OR BT c21 |
| c35 | CuT | 32 | on | NWT c23 OR BT c22 |
| c36 | CuT | 48 | on | NWT c25 OR BT c24 |

* + 1. Content type categories and scene definitions

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Room | Reverb | Microphone Setup |  | Background | SNR  In dB | Overtalk [s](2 | Bandwidth | Talker positions(3 | Talker selection by panel |
| cat 1 | car | car | A-B (15cm) |  | car | 15 | -1 | Max [available up to SWB] | Driver-Pansenger  BackRight-Driver  Driver-BackCenter  BackLeft-Driver  BackRight-BackLeft  BackCenter-BackRight | P1: f1m1 P2: m2f2 P3: f3m3 P4: m1f1 P5: f2m2 P6: m3f3 |
| cat 2 | - | - | A-B (20cm) |  | street | 15 | -1 | max [available up to SWB] |  | P1: m3f3 P2: f1m1 P3: m2f2 P4: f3m3 P5: m1f1 P6: f2m2 |
| cat 3 | small | echoic | M-S |  | office | 15 | -1 | max [available up to SWB] | 1-7  5-3  2-6  4-1  3-4  7-2 | P1: f2m2 P2: m3f3 P3: f1m1 P4: m2f2 P5: f3m3 P6: m1f1 |
| cat 4 | small | echoic | A-B (100cm) |  | office | 15 | -1 | max [available up to SWB] | 1-7  5-3  2-6  4-1  3-4  7-2 | P1: m1f1 P2: f2m2 P3: m3f3 P4: f1m1 P5: m2f2 P6: f3m3 |
| cat 5 | large | echoic | A-B (150cm) |  | office | 15 | -1 | max [available up to SWB] | 2-8  9-4  6-10  11-8  10-12  12-1 | P1: f3m3 P2: m1f1 P3: f2m2 P4: m3f3 P5: f1m1 P6: m2f2 |
| cat 6 | small | echoic | Binaural |  | office | 15 | -1 | max [available up to SWB] | 1-7  5-3  2-6  4-1  3-4  7-2 | P1: m2f2 P2: f3m3 P3: m1f1 P4: f2m2 P5: m3f3 P6: f1m1 |

**Notes:**

**(2** Overtalk [s] means the duration in seconds by which the two sentences in the sound item uttered by different talkers are overlapping. A negative number means that there is a corresponding pause between the two sentences.

**(3** The talker positions are part of the scene definition of the different categories. For office noise, they correspond to the talker positions as depicted in Figures 14.3 and 14.5 of [27] for the large and the small room.

Editor’s note: The spatial room impulse response for car-noise category is tbd, based on availability. The talker positions both for car- and street-noise categories are tbd.

]

* 1. Experiment P800-3: Stereo Mixed and Music Test

[

The purpose of this experiment is to evaluate the performances of the IVAS candidate algorithm with respect to EVS reference conditions in FB mixed content and music for clean and impaired channel conditions [including delay/jitter].

* + 1. Experiment setup

Tables E.3.1 to E.3.3 show conditions to be used for this experiment, list of preliminaries and full list of conditions, respectively.

Table E.3.1: Factors and conditions for Experiment P800-3

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Candidate | 1 | CuT |
| Bitrates | 6 | 13.2, 16.4, 24.4, 32, 48, 64 |
| DTX | 2 | DTX on and off |
| Input level | 1 | -26 LKFS |
| Input frequency mask | 0 | 20KBP |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, [5% random frame erasures / 1 delay/jitter profile] |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 1 | EVS dual mono |
| Bitrates | 8 | 2x7.2, 2x8.0, 2x9.6, 2x13.2, 2x16.4, 2x24.4, 2x32, 2x64 kbps |
| Input level  DTX | 1  0 | -26 LKFS  DTX off |
| Input frequency mask | 0 | 20KBP |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, [5% random frame erasures / 1 delay/jitter profile] |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | -26 LKFS |
| P.50 MNRU  ESDRU | 4  3 | Q = 12, 17, 22, 27 dB  = 0.1, 0.4, 0.7 |
| Input frequency mask | 0 | 20KBP |
| **Common Conditions** |  |  |
| Audio sampling frequency/bandwidth | 1 | 48 kHz/maximum available audio bandwidth up to FB |
| Number of categories | 6 | 3 mixed content and 3 music (See clause 4.5.3) |
| Number of samples | 7 | 6 + 1 (preliminaries) samples per category |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve listeners |
| Randomizations | 6 | 6 panels of 5 listeners |
| Rating Scale | 1 | P.SUPPL800 |
| Listening System | 1 | Headphones, in accordance with clause 4.6 |
| Listening Environment | 1 | No room noise, in accordance with clause 4.6 |
| Languages | 2 | Danish, Mandarin |

Table E.3.2: Preliminaries for Experiment P800-3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Trial #** | **Label** | **Sample** | **Condition** | **Bitrate** | **FER/Profile** |
| 1 | C20 |  | EVS | 2x13.2 | [5%, profile] |
| 2 | c02 |  | MNRU Q=28 dB | - | - |
| 3 | c06 |  | ESDRU = 0.7 | - | - |
| 4 | c13 |  | EVS | 2x16.4 | No errors |
| 5 | c08 |  | ESDRU = 0.1 | - | - |
| 6 | c17 |  | EVS | 2x7.2 | [5%, profile] |
| 7 | c04 |  | MNRU Q=23 dB | - | - |
| 8 | c01 |  | Reference | - | - |
| 9 | c15 |  | EVS | 2x32 | No errors |
| 10 | c07 |  | ESDRU = 0.4 | - | - |
| 11 | c05 |  | MNRU Q=12 dB | - | - |
| 12 | c22 |  | EVS | 2x24.4 | [5%, profile] |

Table E.3.3: Test conditions for Experiment P800-3,  
mixed contents and music under clean and impaired channel conditions [including delay/jitter profiles]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Label** | **Condition** | **Bitrate [kbps]** | **DTX** | **FER/Profile** | **ToR** |
| c01 | Reference | - | - | - |  |
| c02 | MNRU Q=27 dB | - | - | - |  |
| c03 | MNRU Q=22 dB | - | - | - |  |
| c04 | MNRU Q=17 dB | - | - | - |  |
| c05 | MNRU Q=12 dB | - | - | - |  |
| c06 | ESDRU | - | - | - |  |
| c07 | ESDRU | - | - | - |  |
| c08 | ESDRU 1 | - | - | - |  |
| c09 | EVS | 2x7.2 | off | No errors |  |
| c10 | EVS | 2x8 | off | No errors |  |
| c11 | EVS | 2x9.6 | off | No errors |  |
| c12 | EVS | 2x13.2 | off | No errors |  |
| c13 | EVS | 2x16.4 | off | No errors |  |
| c14 | EVS | 2x24.4 | off | No errors |  |
| c15 | EVS | 2x32 | off | No errors |  |
| c16 | EVS | 2x48 | off | No errors |  |
| c17 | EVS | 2x7.2 | off | [5%, profile] |  |
| c18 | EVS | 2x8 | off | [5%, profile] |  |
| c19 | EVS | 2x9.6 | off | [5%, profile] |  |
| c20 | EVS | 2x13.2 | off | [5%, profile] |  |
| c21 | EVS | 2x16.4 | off | [5%, profile] |  |
| c22 | EVS | 2x24.4 | off | [5%, profile] |  |
| c23 | EVS | 2x32 | off | [5%, profile] |  |
| c24 | CuT | 13.2 | off | No errors | NWT c10 OR BT c09 |
| c25 | CuT | 16.4 | off | No errors | NWT c11 OR BT c10 |
| c26 | CuT | 24.4 | off | No errors | NWT c12 OR BT c11 |
| c27 | CuT | 32 | off | No errors | NWT c13 OR BT c12 |
| c28 | CuT | 48 | off | No errors | NWT c15 OR BT c14 |
| c29 | CuT | 64 | off | No errors | NWT c16 OR BT c15 |
| c30 | CuT | 13.2 | off | [5%, profile] | NWT c18 OR BT c17 |
| c31 | CuT | 16.4 | off | [5%, profile] | NWT c19 OR BT c18 |
| c32 | CuT | 24.4 | off | [5%, profile] | NWT c20 OR BT c19 |
| c33 | CuT | 32 | off | [5%, profile] | NWT c21 OR BT c20 |
| c34 | CuT | 48 | off | [5%, profile] | NWT c23 OR BT c22 |
| c35 | CuT | 24.4 | on | No errors | NWT c12 OR BT c11 |
| c36 | CuT | 13.2 | on | 5% | NWT c18 OR BT c17 |

]

* 1. Experiment P800-4: FOA Clean Speech Test

[

* + 1. Experiment setup

Tables E.4.1 to E.4.3 show conditions to be used for this experiment, list of preliminaries and full list of conditions, respectively.

Table E.4.1: Factors and conditions for Experiment P800-4 (Clean Speech)

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Candidate | 1 | CuT |
| Bitrates | 7 | Option1: IVAS candidate operated at 16.4, 24.4, 32, 48, 64, 80 and 96 kbps |
| DTX | 0 | DTX off |
| Input level | 1 | -26 LKFS |
| Input frequency mask | 1 | HP50 |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, 5% [1 delay/jitter profile] |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 8 | Multi-mono EVS |
| Bitrates | 8 | 3x7.2, 4x7.2, 4x8, 4x9.6, 4x13.2, 4x16.4, 4x24.4, 4x32 kbps |
| Input level  DTX | 1  0 | -26 LKFS  DTX off |
| Input frequency mask | 1 | HP50 |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, 5% [1 delay/jitter profile] |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | -26 LKFS |
| P.50 MNRU  ESDRU | 4  3 | Q=17, 22, 27, 32 dB  α = 0.55, 0.675, 0.8 |
| Input frequency mask | 1 | HP50 |
| **Common Conditions** |  |  |
| Test item generation: pre-processing incl. spatialization | 1 | Model-based relying on convolution of raw mono clean speech sentences convolved with (FOA) Spatial Room Impulse Responses respective to various talker positions relative to a capture point and spatial (FOA) ambient noise mixing. |
| Binaural renderer | 1 | FOA to binaural (external) rendering |
| Audio sampling frequency/bandwidth | 1 | 48 kHz/maximum available audio bandwidth |
| Kind of samples | 1 | Sentence pair uttered by different talkers and genders (3 male and 3 female) |
| Number of categories | 6 | 6 Different environments and talker interactions |
| Number of samples | 7 | 6 + 1 (preliminaries) samples per category |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve listeners |
| Randomizations | 6 | 6 panels of 5 listeners |
| Rating Scale | 1 | P.SUPPL800 |
| Listening System | 1 | Headphones, in accordance with clause 4.6 |
| Listening Environment | 1 | No room noise, in accordance with clause 4.6 |
| Languages | 2 | Japanese, English |

Table E.4.2: Preliminaries for Experiment P800-4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Trial #** | **Label** | **Sample** | **Condition** | **Bitrate** | **FER/Profile** |
| 1 | c20 |  | EVS | 4x13.2 | [5%, profile] |
| 2 | c02 |  | MNRU Q=32 dB | - | - |
| 3 | c06 |  | ESDRU = 0.8 | - | - |
| 4 | c14 |  | EVS | 4x16.4 | No errors |
| 5 | c08 |  | ESDRU = 0.55 | - | - |
| 6 | c17 |  | EVS | 4x7.2 | [5%, profile] |
| 7 | c04 |  | MNRU Q=22 dB | - | - |
| 8 | c01 |  | Reference | - | - |
| 9 | c16 |  | EVS | 4x32 | No errors |
| 10 | c07 |  | ESDRU = 0.675 | - | - |
| 11 | c05 |  | MNRU Q=17 dB | - | - |
| 12 | c22 |  | EVS | 4x24.4 | [5%, profile] |

Table E.4.3: Test conditions for Experiment P800-4,  
clean speech under clean and impaired channel conditions [including delay/jitter profiles]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Condition** | **Bitrate [kbps]** | **FER/Profile** | **ToR** |
| c01 | Reference | - | - |  |
| c02 | MNRU Q=32 dB | - | - |  |
| c03 | MNRU Q=27 dB | - | - |  |
| c04 | MNRU Q=22 dB | - | - |  |
| c05 | MNRU Q=17 dB | - | - |  |
| c06 | ESDRU | - | - |  |
| c07 | ESDRU | - | - |  |
| c08 | ESDRU | - | - |  |
| c09 | EVS | 3x7.2 | No errors |  |
| c10 | EVS | 4x7.2 | No errors |  |
| c11 | EVS | 4x8 | No errors |  |
| c12 | EVS | 4x9.6 | No errors |  |
| c13 | EVS | 4x13.2 | No errors |  |
| c14 | EVS | 4x16.4 | No errors |  |
| c15 | EVS | 4x24.4 | No errors |  |
| c16 | EVS | 4x32 | No errors |  |
| c17 | EVS | 4x7.2 | [5%, profile] |  |
| c18 | EVS | 4x8 | [5%, profile] |  |
| c19 | EVS | 4x9.6 | [5%, profile] |  |
| c20 | EVS | 4x13.2 | [5%, profile] |  |
| c21 | EVS | 4x16.4 | [5%, profile] |  |
| c22 | EVS | 4x24.4 | [5%, profile] |  |
| c23 | EVS | 4x32 | [5%, profile] |  |
| c24 | CuT | 16.4 | No errors | NWT c09 |
| c25 | CuT | 24.4 | No errors | NWT c10 |
| c26 | CuT | 32 | No errors | NWT c12 or BT c11 |
| c27 | CuT | 48 | No errors | NWT c14 or BT c13 |
| c28 | CuT | 64 | No errors | NWT c15 or BT c14 |
| c29 | CuT | 80 | No errors | NWT c15 or BT c14 |
| c30 | CuT | 96 | No errors | NWT c16 or BT c15 |
|  |  |  |  |  |
| c31 | CuT | 24.4 | [5%, profile] | NWT c17 |
| c32 | CuT | 32 | [5%, profile] | NWT c19 or BT c18 |
| c33 | CuT | 48 | [5%, profile] | NWT c21 or BT c20 |
| c34 | CuT | 64 | [5%, profile] | NWT c22 or BT c21 |
| c35 | CuT | 80 | [5%, profile] | NWT c22 or BT c21 |
| c36 | CuT | 96 | [5%, profile] | NWT c23 or BT c22 |

* + 1. Content type categories and scene definitions

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Category*** | ***Environment(1*** |  |  | ***Background(2*** | ***SNR [dB]*** | ***Overtalk [s](3*** | ***Bandwidth*** | ***Talker positions(4*** | ***Talker selection by panel(5*** |
| *cat 1* | *room\_1\_FOA* |  |  | *room\_1\_cleanbg\_FOA* | *[50]* | *1* | *Max* |  | *P1: f1m1 P2: m2f2 P3: f3m3 P4: m1f1 P5: f2m2 P6: m3f3* |
| *cat 2* | *room\_[1/4]\_FOA* |  |  | *room\_[1/4]\_cleanbg\_FOA* | *[50]* | *-1* | *Max* |  | *P1: m3f3 P2: f1m1 P3: m2f2 P4: f3m3 P5: m1f1 P6: f2m2* |
| *cat 3* | *room\_2\_FOA* |  |  | *room\_2\_cleanbg\_FOA* | *[50]* | *1* | *Max* |  | *P1: f2m2 P2: m3f3 P3: f1m1 P4: m2f2 P5: f3m3 P6: m1f1* |
| *cat 4* | *room\_[2/5]\_FOA* |  |  | *room\_[2/5]\_cleanbg\_FOA* | *[50]* | *-1* | *Max* |  | *P1: m1f1 P2: f2m2 P3: m3f3 P4: f1m1 P5: m2f2 P6: f3m3* |
| *cat 5* | *room\_3\_FOA* |  |  | *room\_3\_cleanbg\_FOA* | *[50]* | *1* | *Max* |  | *P1: f3m3 P2: m1f1 P3: f2m2 P4: m3f3 P5: f1m1 P6: m2f2* |
| *cat 6* | *room\_[3/6]\_FOA* |  |  | *room\_[3/6]\_cleanbg\_FOA* | *[50]* | *-1* | *Max* |  | *P1: m2f2 P2: f3m3 P3: m1f1 P4: f2m2 P5: m3f3 P6: f1m1* |

**Notes:**

**(1** Editor’s note: The specific room/environment characteristic and resulting reverb characteristic will be defined by the choice of the specific Spatial Room Impulse Responses used in the convolution process with the raw mono sentences, according to the pertaining stipulations of the test plan IVAS-8a.

**(2** Editor’s note: Background is defined by the chosen background noise file according to the pertaining stipulations of the test plan IVAS-8a. Backround name ‘clean\_bg\_[X]\_FOA’ indicates a low-noise background corresponding to environment [X], e.g., with low air-conditioning/fan noise.

**(3** Overtalk [s] means the duration in seconds by which the two sentences in the sound item uttered by different talkers are overlapping. A negative number means that there is a corresponding pause between the two sentences.

**(4** Editor’s note: The talker positions are part of the scene definition of the different categories. They should be chosen in a way from the available set of SRIRs for the used room making sure that there is a good coverage of different possible positions. Different selections should be made for the different listener panels.

**(5** All sentences by the 6 talkers shall be unique.

]

* 1. Experiment P800-5: FOA Speech+background Test

[

* + 1. Experiment setup

Tables E.5.1 to E.5.3 show conditions to be used for this experiment, list of preliminaries and full list of conditions, respectively.

Table E.5.1: Factors and conditions for Experiment P800-5

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Candidate | 1 | CuT |
| bitrates | 7 | 16.4, 24.4, 32, 48, 64, 80, 96 kbps |
| DTX | 2 | DTX off, DTX on |
| Input level | 1 | -26 LKFS |
| Input frequency mask | 1 | HP50 |
| Noise | 1 | 15dB |
| Error Conditions | 0 | No errors |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 1 | Multi-mono EVS |
| Bitrates | 8 | 3x7.2, 4x7.2, 4x8, 4x9.6, 4x13.2, 4x16.4, 4x24.4, 4x32 kbps |
| Input level  DTX | 1  2 | -26 LKFS  DTX off, DTX on |
| Input frequency mask | 1 | HP50 |
| Noise | 1 | 15dB |
| Error Conditions | 0 | No errors |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | -26 LKFS |
| P.50 MNRU  ESDRU | 4  3 | Q= 16, 20, 24, 28 dB  *α* = 0.1, 0.4, 0.7 |
| Input frequency mask | 1 | HP50 |
| **Common Conditions** |  |  |
| Test item generation: pre-processing incl. spatialization | 1 | Model-based relying on convolution of raw mono clean speech sentences convolved with (FOA) Spatial Room Impulse Responses respective to various talker positions relative to a capture point and spatial (FOA) ambient noise mixing |
| Binaural renderer | 1 | FOA to binaural (external) rendering |
| Audio sampling frequency/bandwidth | 1 | 48 kHz/maximum available audio bandwidth |
| Kind of samples | 1 | Sentence pair uttered by different talkers and genders (3 male and 3 female) |
| Number of categories | 6 | 6 Different environments with different background types (car,office, street) and talker interactions |
| Number of samples | 7 | 6 + 1 (preliminaries) samples per category |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve listeners |
| Randomizations | 6 | 6 panels of 5 listeners |
| Rating Scale | 1 | P.SUPPL800 |
| Listening System | 1 | Headphones, in accordance with clause 4.6 |
| Listening Environment | 1 | No room noise, in accordance with clause 4.6 |
| Languages | 2 | Danish, German |

Table E.5.2: Preliminaries for Experiment P800-5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Trial #** | **Label** | **Sample** | **Condition** | **Bitrate** | **DTX** |
| 1 | c21 |  | EVS | 4x13.2 | on |
| 2 | c02 |  | MNRU Q=28 dB | - | - |
| 3 | c06 |  | ESDRU = 0.7 | - | - |
| 4 | c14 |  | EVS | 4x16.4 | off |
| 5 | c08 |  | ESDRU = 0.1 | - | - |
| 6 | c18 |  | EVS | 4x7.2 | on |
| 7 | c04 |  | MNRU Q=20 dB | - | - |
| 8 | c01 |  | Reference | - | - |
| 9 | c16 |  | EVS | 4x32 | off |
| 10 | c07 |  | ESDRU = 0.4 | - | - |
| 11 | c05 |  | MNRU Q=16 dB | - | - |
| 12 | c23 |  | EVS | 4x24.4 | on |

Table E.5.3: Test conditions for Experiment P800-5,  
clean speech with background noise under clean channel conditions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Condition** | **Bitrate [kbps]** | **DTX** | **ToR** |
| c01 | Reference | - | - |  |
| c02 | MNRU Q=28 dB | - | - |  |
| c03 | MNRU Q=24 dB | - | - |  |
| c04 | MNRU Q=20 dB | - | - |  |
| c05 | MNRU Q=16 dB | - | - |  |
| c06 | ESDRU | - | - |  |
| c07 | ESDRU | - | - |  |
| c08 | ESDRU | - | - |  |
| c09 | EVS | 3x7.2 | off |  |
| c10 | EVS | 4x7.2 | off |  |
| c11 | EVS | 4x8 | off |  |
| c12 | EVS | 4x9.6 | off |  |
| c13 | EVS | 4x13.2 | off |  |
| c14 | EVS | 4x16.4 | off |  |
| c15 | EVS | 4x24.4 | off |  |
| c16 | EVS | 4x32 | off |  |
| c17 | EVS | 3x7.2 | on |  |
| c18 | EVS | 4x7.2 | on |  |
| c19 | EVS | 4x8 | on |  |
| c20 | EVS | 4x9.6 | on |  |
| c21 | EVS | 4x13.2 | on |  |
| c22 | EVS | 4x16.4 | on |  |
| c23 | EVS | 4x24.4 | on |  |
|  |  |  |  |  |
| c24 | CuT | 16.4 | off | NWT c09 |
| c25 | CuT | 24.4 | off | NWT c10 |
| c26 | CuT | 32 | off | NWT c12 or BT c11 |
| c27 | CuT | 48 | off | NWT c14 or BT c13 |
| c28 | CuT | 64 | off | NWT c15 or BT c14 |
| c29 | CuT | 80 | off | NWT c15 or BT c14 |
| c30 | CuT | 96 | off | NWT c16 or BT c15 |
| c31 | CuT | 16.4 | on | NWT c17 |
| c32 | CuT | 24.4 | on | NWT c18 |
| c33 | CuT | 32 | on | NWT c20 or BT c19 |
| c34 | CuT | 48 | on | NWT c22 or BT c21 |
| c35 | CuT | 64 | on | NWT c23 or BT c22 |
| c36 | CuT | 80 | on | NWT c23 or BT c22 |

* + 1. Content type categories and scene definitions

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Category*** | ***Environment*** |  |  | ***Background(2*** | ***SNR[dB]*** | ***Overtalk [s](3*** | ***Bandwidth(3*** | ***Talker positions(4*** | ***Talker selection by panel(5*** |
| *cat 1* | *car\_1\_FOA* |  |  | *car\_1\_bg\_FOA* | *10* | *-1* | *Max* |  | *P1: f1m1 P2: m2f2 P3: f3m3 P4: m1f1 P5: f2m2 P6: m3f3* |
| *cat 2* | *car\_[1/2]\_FOA* |  |  | *car\_[1/2]\_bg\_FOA* | *10* | *-1* | *Max* |  | *P1: m3f3 P2: f1m1 P3: m2f2 P4: f3m3 P5: m1f1 P6: f2m2* |
| *cat 3* | *out\_1\_FOA* |  |  | *street\_1\_bg\_FOA* | *10-15* | *-1* | *Max* |  | *P1: f2m2 P2: m3f3 P3: f1m1 P4: m2f2 P5: f3m3 P6: m1f1* |
| *cat 4* | *out\_[1/2]\_FOA* |  |  | *[park\_1\_bg\_FOA / nature\_1\_bg\_FOA / event\_1\_bg\_FOA / street\_[1/2]\_bg\_FOA]* | *10-15* | *-1* | *Max* |  | *P1: m1f1 P2: f2m2 P3: m3f3 P4: f1m1 P5: m2f2 P6: f3m3* |
| *cat 5* | *room\_[X]\_FOA* |  |  | *office\_1\_bg\_FOA* | *10-15* | *-1* | *Max* |  | *P1: f3m3 P2: m1f1 P3: f2m2 P4: m3f3 P5: f1m1 P6: m2f2* |
| *cat 6* | *room\_[X]\_FOA* |  |  | *[cafeteria\_1\_bg\_FOA / mall\_1\_bg\_FOA/ office[1/2]\_bg\_FOA]* | *10-15* | *-1* | *Max* |  | *P1: m2f2 P2: f3m3 P3: m1f1 P4: f2m2 P5: m3f3 P6: f1m1* |

**Notes:**

**(1** Editor’s note: The specific room/environment characteristic and resulting reverb characteristic will be defined by the choice of the specific Spatial Room Impulse Responses used in the convolution process with the raw mono sentences, according to the pertaining stipulations of the test plan IVAS-8a.

**(2** Editor’s note: Background is defined by the chosen background noise file according to the pertaining stipulations of the test plan IVAS-8a.

**(3** Overtalk [s] means the duration in seconds by which the two sentences in the sound item uttered by different talkers are overlapping. A negative number means that there is a corresponding pause between the two sentences.

**(**4 Editor’s note: The talker positions are part of the scene definition of the different categories. They should be chosen in a way from the available set of SRIRs for the used room making sure that there is a good coverage of different possible positions. Different selections should be made for the different listener panels.

**(5** All sentences by the 6 talkers shall be unique.

Editor’s note:Types of background need to be aligned with Section 4.5.2.

]

* 1. Experiment P800-6: 1-Object Clean Speech Test

[

The purpose of this experiment is to evaluate the performances of the IVAS candidate for encoding, decoding, and rendering of a single object with metadata. The metadata determines the position of the object around the listener in a 3-dimensional space as described in the Scene description section below. The audio is rendered binaurally via headphones.



























* + 1. Experiment setup

Tables E.6.1 to E.6.3 show conditions to be used for this experiment, list of preliminaries and full list of conditions, respectively.

The test Categories correspond to different talkers.

Table E.6.1: Factors and conditions for Experiment P800-6

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Candidate | 1 | CuT |
| Bitrates | 6 | 13.2, 16.4, 24.4, 32, 48, 64 kbps |
| DTX | 2 | DTX ON and OFF |
| Input frequency mask | 1 | HP50 |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, 5% FERs |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 1 | EVS |
| Bitrates | 6 | 13.2, 16.4, 24.4, 32, 48, 64 kbps |
| DTX | 2 | DTX ON and OFF |
| Input frequency mask | 1 | HP50 |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, 5% FERs |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | -26 LKFS |
| MNRU | 5 | Q= 15, 23, 31, 39, 47 dB |
| ESDRU | 4 | 0.1, 0.3, 0.5, 0.7 |
| Input frequency mask | 1 | HP50 |
|  |  |  |
| **Common Conditions** |  |  |
| Audio sampling frequency/bandwidth | 1 | 48 kHz/maximum available audio bandwidth up to FB |
| Number of talkers (categories) | 6 | 3 male and 3 female |
| Number of speech samples | 7 | 6 + 1 (preliminaries) sentence pairs per talker |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve Listeners |
| Randomizations (panels) | 6 | 6 panels of 5 listeners |
| Rating Scale | 1 | P.SUPPL800 |
| Listening System | 1 | Headphones, in accordance with clause 4.6 |
| Languages | 2 | Japanese, English |

Table E.6.2: Preliminaries for Experiment P800-6

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Trial** | **Label** | **Sample** | **Condition** | **Bitrate [kb/s]** | **DTX** | **FER [%]** |
| **1** | C19 |  | EVS | 24.4 | Off | 5 |
| **2** | C04 |  | MNRU 31 | - | - | - |
| **3** | C21 |  | EVS | 13.2 | On | 0 |
| **4** | C09 |  | ESDRU 0.5 | - | - | - |
| **5** | C14 |  | EVS | 32 | Off | 0 |
| **6** | C02 |  | MNRU 15 | - | - | - |
| **7** | C12 |  | EVS | 16.4 | Off | 0 |
| **8** | C01 |  | Direct | - | - | - |
| **9** | C17 |  | EVS | 13.2 | Off | 5 |
| **10** | C06 |  | MNRU 47 | - | - | - |
| **11** | C23 |  | EVS | 24.4 | On | 0 |
| **12** | C07 |  | ESDRU 0.1 | - | - | - |

Table E.6.3 Test conditions for Experiment P800-6:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Label** | **Condition** | **Bitrate [kb/s]** | **DTX** | **FER [%]** | **Ref condition** | **ToR** |
| **C01** | Direct |  | - | - | - | - |
| **C02** | MNRU 15 |  | - | - | - | - |
| **C03** | MNRU 23 |  | - | - | - | - |
| **C04** | MNRU 31 |  | - | - | - | - |
| **C05** | MNRU 39 |  | - | - | - | - |
| **C06** | MNRU 47 |  | - | - | - | - |
| **C07** | ESDRU 0.1 |  | - | - | - | - |
| **C08** | ESDRU 0.3 |  | - | - | - | - |
| **C09** | ESDRU 0.5 |  | - | - | - | - |
| **C10** | ESDRU 0.7 |  | - | - | - | - |
| **C11** | EVS | 13.2 | Off | 0 | - | - |
| **C12** | EVS | 16.4 | Off | 0 | - | - |
| **C13** | EVS | 24.4 | Off | 0 | - | - |
| **C14** | EVS | 32 | Off | 0 | - | - |
| **C15** | EVS | 48 | Off | 0 | - | - |
| **C16** | EVS | 64 | Off | 0 |  |  |
| **C17** | EVS | 13.2 | Off | 5 | - | - |
| **C18** | EVS | 16.4 | Off | 5 | - | - |
| **C19** | EVS | 24.4 | Off | 5 | - | - |
| **C20** | EVS | 32 | Off | 5 |  |  |
| **C21** | EVS | 13.2 | On | 0 | - | - |
| **C22** | EVS | 16.4 | On | 0 | - | - |
| **C23** | EVS | 24.4 | On | 0 | - | - |
| **C24** | CuT | 13.2 | Off | 0 | C11 | NWT |
| **C25** | CuT | 16.4 | Off | 0 | C12 | NWT |
| **C26** | CuT | 24.4 | Off | 0 | C13 | NWT |
| **C27** | CuT | 32 | Off | 0 | C14 | NWT |
| **C28** | CuT | 48 | Off | 0 | C15 | NWT |
| **C29** | CuT | 64 | Off | 0 | C16 | NWT |
| **C30** | CuT | 13.2 | Off | 5 | C17 | NWT |
| **C31** | CuT | 16.4 | Off | 5 | C18 | NWT |
| **C32** | CuT | 24.4 | Off | 5 | C19 | NWT |
| **C33** | CuT | 32 | Off | 5 | C20 | NWT |
| **C34** | CuT | 13.2 | On | 0 | C21 | NWT |
| **C35** | CuT | 16.4 | On | 0 | C22 | NWT |
| **C36** | CuT | 24.4 | On | 0 | C23 | NWT |

* + 1. Content type categories and scene definitions

It is assumed that sentence pairs include leading and trailing silence of minimum of 0.5 s. The metadata corresponds to the whole duration of the samples. This means that for moving objects, only a part of the trajectory corresponds to active speech.

1. Talker sitting at a table (elevation 0°), at different azimuths:

|  |  |  |
| --- | --- | --- |
| Panel | Talker | Azimuth |
| 1 | G1 | 0° |
| 2 | G6 | 60° |
| 3 | G5 | 120° |
| 4 | G4 | 180° |
| 5 | G3 | 240° |
| 6 | G2 | 300° |

1. Standing talker (elevation 35°), at different azimuths:

|  |  |  |
| --- | --- | --- |
| Panel | Talker | Azimuth |
| 1 | G2 | 120° |
| 2 | G1 | 180° |
| 3 | G6 | 240° |
| 4 | G5 | 300° |
| 5 | G4 | 0° |
| 6 | G3 | 60° |

1. Smaller talker (child) walking around a table in the positive sense (counter clockwise), elevation 0°. Azimuth varies continuously for the sentence pair to cover the whole circle starting at:

|  |  |  |
| --- | --- | --- |
| Panel | Talker | Starting Azimuth |
| 1 | G3 | 0° |
| 2 | G2 | 60° |
| 3 | G1 | 120° |
| 4 | G6 | 180° |
| 5 | G5 | 240° |
| 6 | G3 | 300° |

1. Adult talker walking around a table in the negative sense (clockwise), elevation 35°. Azimuth varies continuously for the sentence pair to cover the whole circle starting at:

|  |  |  |
| --- | --- | --- |
| Panel | Talker | Starting Azimuth |
| 1 | G4 | 0° |
| 2 | G3 | 60° |
| 3 | G2 | 120° |
| 4 | G1 | 180° |
| 5 | G6 | 240° |
| 6 | G5 | 300° |

1. Elevation displacement: Elevation varies continuously for the sentence pair to cover the interval of -90° to 90°. Azimuth is constant for a sentence pair, but different for different for each sentence pair:

|  |  |  |
| --- | --- | --- |
| Panel | Talker | Azimuth |
| 1 | G5 | 240° |
| 2 | G4 | 300° |
| 3 | G3 | 0° |
| 4 | G2 | 60° |
| 5 | G1 | 120° |
| 6 | G6 | 180° |

1. Azimuth and elevation displacement. Azimuth and elevation vary continuously. Azimuth varies in positive sense (counter clockwise) to cover the range of 180°, with different starting azimuth for each sentence pair. Elevation varies in negative sense, from 35° to -35°:

|  |  |  |
| --- | --- | --- |
| Panel | Talker | Starting Azimuth |
| 1 | G6 | 60° |
| 2 | G5 | 120° |
| 3 | G4 | 180° |
| 4 | G3 | 240° |
| 5 | G2 | 300° |
| 6 | G1 | 0° |

**Allocation of scenes to talkers to each listening panel**

Each of the sentences uttered by a certain talker is encoded using different scene. To balance the test, in addition to listeners of each panel listening to all talkers (Categories), and all scenes are also covered in each panel. Allocation of sentences to each panel is given in the Table below:

Table E.6.4: Allocation of scenes to talkers to each listening panel

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Panel** | **G1 (M1)** | **G2 (F1)** | **G3 (M2)** | **G4 (F2)** | **G5 (M3)** | **G6 (F3)** |
| **1** | S1 (Sc a) | S1 (Sc b) | S1 (Sc c) | S1 (Sc d) | S1 (Sc e) | S1 (Sc f) |
| **2** | S2 (Sc b) | S2 (Sc c) | S2 (Sc d) | S2 (Sc e) | S2 (Sc f) | S2 (Sc a) |
| **3** | S3 (Sc c) | S3 (Sc d) | S3 (Sc e) | S3 (Sc f) | S3 (Sc a) | S3 (Sc b) |
| **4** | S4 (Sc d) | S4 (Sc e) | S4 (Sc f) | S4 (Sc a) | S4 (Sc b) | S4 (Sc c) |
| **5** | S5 (Sc e) | S5 (Sc f) | S5 (Sc a) | S5 (Sc b) | S5 (Sc c) | S5 (Sc d) |
| **6** | S6 (Sc f) | S6 (Sc a) | S6 (Sc b) | S6 (Sc c) | S6 (Sc d) | S6 (Sc e) |

]

* 1. Experiment P800-7: 2-Objects Clean Speech Test

[

The purpose of this experiment is to evaluate the performances of the IVAS candidate for encoding, decoding, and rendering of two simultaneous objects with metadata. The metadata determines the position of the objects around the listener in a 3-dimensional space as described in the Scene description section below. The audio is rendered binaurally via headphones.

* The length of the created spatial audio samples is 6 s.
* Level adjusted as specified in IVAS-7a (Processing Plan) [3].



























* + 1. Experiment setup

Tables E.7.1 to E.7.3 show conditions to be used for this experiment, list of preliminaries and full list of conditions, respectively.

The test Categories correspond to different talkers.

Table E.7.1: Factors and conditions for Experiment P800-7

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Candidate | 1 | CuT |
| Bitrates | 5 | 16.4, 24.4, 32, 48, 64 kbps |
| DTX | 2 | DTX ON and OFF |
| Input frequency mask | 1 | HP50 |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, 5% FERs |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 1 | EVS |
| Bitrates | 5 | 2x8.0, 2x13.2, 2x16.4, 2x24.4, 2x32 kbps |
| DTX | 2 | DTX ON and OFF |
| Input frequency mask | 1 | HP50 |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, 5% FERs |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | -26 LKFS |
| MNRU | 5 | Q= 15, 23, 31, 39, 47 dB |
| ESDRU | 4 | 0.1, 0.3, 0.5, 0.7 |
| Input frequency mask | 1 | HP50 |
|  |  |  |
| **Common Conditions** |  |  |
| Audio sampling frequency/bandwidth | 1 | 48 kHz/maximum available audio bandwidth up to FB |
| Number of talkers (categories) | 6 | 3 male and 3 female |
| Number of speech samples | 7 | 6 + 1 (preliminaries) sentence pairs per talker. |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve Listeners |
| Randomizations (panels) | 6 | 6 panels of 5 listeners |
| Rating Scale | 1 | P.SUPPL800 |
| Listening System | 1 | Headphones, in accordance with clause 4.6 |
| Languages | 2 | Danish, Mandarin |

Table E.7.2: Preliminaries for Experiment P800-7

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Trial** | **Label** | **Sample** | **Condition** | **Bitrate [kb/s]** | **DTX** | **FER [%]** |
| **1** | C18 |  | EVS | 2 x 16.4 | Off | 5 |
| **2** | C04 |  | MNRU 31 | - | - | - |
| **3** | C20 |  | EVS | 2 x 8.0 | On | 0 |
| **4** | C09 |  | ESDRU 0.5 | - | - | - |
| **5** | C14 |  | EVS | 2 x 24.4 | Off | 0 |
| **6** | C02 |  | MNRU 15 | - | - | - |
| **7** | C12 |  | EVS | 2 x 13.2 | Off | 0 |
| **8** | C01 |  | Direct | - | - | - |
| **9** | C16 |  | EVS | 2 x 8.0 | Off | 5 |
| **10** | C06 |  | MNRU 47 | - | - | - |
| **11** | C22 |  | EVS | 2 x 16.4 | On | 0 |
| **12** | C07 |  | ESDRU 0.1 | - | - | - |

Table E.7.3 Test conditions for Experiment P800-7:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Label** | **Condition** | **Bitrate [kb/s]** | **DTX** | **FER [%]** | **Ref condition** | **ToR** |
| **C01** | Direct | - | - | - | - | - |
| **C02** | MNRU 15 | - | - | - | - | - |
| **C03** | MNRU 23 | - | - | - | - | - |
| **C04** | MNRU 31 | - | - | - | - | - |
| **C05** | MNRU 39 | - | - | - | - | - |
| **C06** | MNRU 47 | - | - | - | - | - |
| **C07** | ESDRU 0.1 | - | - | - | - | - |
| **C08** | ESDRU 0.3 | - | - | - | - | - |
| **C09** | ESDRU 0.5 | - | - | - | - | - |
| **C10** | ESDRU 0.7 | - | - | - | - | - |
| **C11** | EVS | 2 x 8.0 | Off | 0 | - | - |
| **C12** | EVS | 2 x 13.2 | Off | 0 | - | - |
| **C13** | EVS | 2 x 16.4 | Off | 0 | - | - |
| **C14** | EVS | 2 x 24.4 | Off | 0 | - | - |
| **C15** | EVS | 2 x 32 | Off | 0 | - | - |
| **C16** | EVS | 2 x 8.0 | Off | 5 | - | - |
| **C17** | EVS | 2 x 13.2 | Off | 5 | - | - |
| **C18** | EVS | 2 x 16.4 | Off | 5 | - | - |
| **C19** | EVS | 2 x 24.4 | Off | 5 | - | - |
| **C20** | EVS | 2 x 8.0 | On | 0 | - | - |
| **C21** | EVS | 2 x 13.2 | On | 0 | - | - |
| **C22** | EVS | 2 x 16.4 | On | 0 | - | - |
| **C23** | EVS | 2 x 24.4 | On | 0 |  |  |
| **C24** | CuT | 16.4 | Off | 0 | C11 | NWT |
| **C25** | CuT | 24.4 | Off | 0 | C12 | NWT |
| **C26** | CuT | 32 | Off | 0 | C13 | NWT |
| **C27** | CuT | 48 | Off | 0 | C14 | NWT |
| **C28** | CuT | 64 | Off | 0 | C15 | NWT |
| **C29** | CuT | 16.4 | Off | 5 | C16 | NWT |
| **C30** | CuT | 24.4 | Off | 5 | C17 | NWT |
| **C31** | CuT | 32 | Off | 5 | C18 | NWT |
| **C32** | CuT | 48 | Off | 5 | C19 | NWT |
| **C33** | CuT | 16.4 | On | 0 | C20 | NWT |
| **C34** | CuT | 24.4 | On | 0 | C21 | NWT |
| **C35** | CuT | 32 | On | 0 | C22 | NWT |
| **C36** | CuT | 48 | On | 0 | C23 | NWT |

* + 1. Content type categories and scene definitions

The listening database consists of artificially created spatial audio samples from monophonic clean speech recordings where always 1 female and 1 male talker are combined in conversation-like scenarios following the Scene descriptions below.

A leading and trailing silence is assumed for each artificially created spatial audio sample. The metadata corresponds to the whole duration of the sample. This means that for moving objects, only a part of the trajectory corresponds to active speech.

In one half of the samples, the 2nd talker’s utterance follows the 1st talker’s utterance simulating natural conversation. The gap between the utterances should be approximately 1 s. In the other half of the samples, the situation is similar, but the utterances partially overlap. The targeted overlap should be also approximately 1 s. Non-overlapping sentence pairs are used for Scenes a., c., and e. as described below. Overlapping sentence pairs are used for Scenes b., d., and f.

1. Two talkers sitting at a table (elevation 0°), at different azimuths. To increase positional variation, both the absolute azimuths and the difference of the azimuths of both talkers vary for each sentence pair:

|  |  |
| --- | --- |
| Azimuth of 1st talker | Azimuth of 2nd talker |
| 0° | 50° |
| 10° | 110° |
| 20° | 170° |
| 30° | 230° |
| 40° | 290° |
| 50° | 350° |

1. Two standing talkers (elevation 35°), at different azimuths. To increase positional variation, both the absolute azimuths and the difference of the azimuths of both talkers vary for each sentence pair:

|  |  |
| --- | --- |
| Azimuth of 1st talker | Azimuth of 2nd talker |
| 0° | 50° |
| 10° | 110° |
| 20° | 170° |
| 30° | 230° |
| 40° | 290° |
| 50° | 350° |

1. One talker sitting at a table (elevation 0°), second talker standing beside the table (elevation 45°). Non-overlapping utterances:

|  |  |
| --- | --- |
| Azimuth of 1st talker | Azimuth of 2nd talker |
| 1st talker | 2nd talker |
| 0° | 50° |
| 10° | 110° |
| 20° | 170° |
| 30° | 230° |
| 40° | 290° |

1. One talker sitting at a table (elevation 0°), second talker walking around the table (elevation 45°). The azimuth of the 2nd talker varies continually, positive sense is counter clockwise:

|  |  |
| --- | --- |
| Azimuth of 1st talker | Azimuth of 2nd talker |
| 50° | 180° : +1° : 120° |
| 100° | 130° : +1° : 70° |
| 150° | 80° : +1° : 20° |
| 200° | 30° : -1° : -270° (90°) |
| 250° | -20° : -1° : -320° (40°) |
| 300° | -70° : -1° : -10° |

1. Two talkers walking side-by-side around the table (elevation 45°). The azimuth is the same for both talkers and varies continually:

|  |
| --- |
| **Azimuth of the talkers** |
| 180° : +1° : 120° |
| 130° : +1° : 70° |
| 80° : +1° : 20° |
| 30° : -1° : -270° (90°) |
| -20° : -1° : -320° (40°) |
| -70° : -1° : -10° |

1. Two talkers walking around the table in opposite directions (elevation 30°), starting at the same position. Azimuths of both talkers vary continually:

|  |  |
| --- | --- |
| **Azimuth of 1st talker** | **Azimuth of 2nd talker** |
| 180° : +1° : 120° | 180° : -1° : -120° |
| 240° (-120°) : +1° : 180° | 240° (-120°) : -1° : -60° |
| 300° (-60°) : +1° : 240° | 300° (-60°) : -1° : 0° |
| 0° +1° : 300° | 0° : -1° : 60° |
| 60° : +1° : 0° | 60° : -1° : 120° |
| 120° : +1° : 60° | 120° : -1° : 180° |

**Allocation of scenes to talkers to each listening panel**The following table assumes that test Categories correspond to different talker pairs. Each of the sentence pairs uttered by a certain talker pair is encoded using a different scene. To balance the test, listeners of each panel listen to all talker pairs, and all scenes are covered in each panel.

Table E.7.4: Allocation of scenes and talkers to each listening panel

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Panel / Categories** | **G1 (M1+F1)** | **G2 (M2+F2)** | **G3 (M3+F3)** | **G4 (F1+M2)** | **G5 F2+M3()** | **G6 (F3+M1)** |
| **1** | S1 (Sc a) | S1 (Sc b) | S1 (Sc c) | S1 (Sc d) | S1 (Sc e) | S1 (Sc f) |
| **2** | S2 (Sc b) | S2 (Sc c) | S2 (Sc d) | S2 (Sc e) | S2 (Sc f) | S2 (Sc a) |
| **3** | S3 (Sc c) | S3 (Sc d) | S3 (Sc e) | S3 (Sc f) | S3 (Sc a) | S3 (Sc b) |
| **4** | S4 (Sc d) | S4 (Sc e) | S4 (Sc f) | S4 (Sc a) | S4 (Sc b) | S4 (Sc c) |
| **5** | S5 (Sc e) | S5 (Sc f) | S5 (Sc a) | S5 (Sc b) | S5 (Sc c) | S5 (Sc d) |
| **6** | S6 (Sc f) | S6 (Sc a) | S6 (Sc b) | S6 (Sc c) | S6 (Sc d) | S6 (Sc e) |

]

* 1. Experiment P800-8: MASA Clean Speech Test

[

* + 1. Experiment setup

Stereo-MASA inputs are tested in IVAS selection phase. The stereo-MASA inputs can have 1-direction spatial metadata (based, e.g., on MASA analysis [28] of original FOA content) or 2-direction spatial metadata (based, e.g., on MASA analysis [28] of original HOA2 content).

Tables E.8.1 to E.8.3 show conditions to be used for this experiment, list of preliminaries and full list of conditions, respectively.

Table E.8.1: Factors and conditions for Experiment P800-8 (Clean Speech)

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Candidate | 1 | CuT |
| Bitrates | 7 | 13.2, 16.4, 24.4, 32, 48, 64, 80 kbps |
| DTX | 0 | DTX off |
| Input level | 1 | -26 LKFS |
| Input frequency mask | 1 | HP50 |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, 5% [1 delay/jitter profile] |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 2 | Multi-mono EVS, Dual-mono EVS with unquantized metadata |
| Bitrates | 6 | 3x7.2, 4x7.2, 4x8, 4x9.6, 4x16.4, 4x24.4 kbps |
|  | 5 | 2x7.2, 2x8, 2x9.6, 2x16.4, 2x24.4 kbps |
| Input level  DTX | 1  0 | -26 LKFS  DTX off |
| Input frequency mask | 1 | HP50 |
| Noise | 0 | No noise |
| Error Conditions | 2 | 0%, 5% [1 delay/jitter profile] |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | -26 LKFS |
| P.50 MNRU  ESDRU | 4  3 | Q = 32, 27, 22, 17 dB  *α* = 0.45, 0.65, 0.85 |
| Input frequency mask | 1 | HP50 |
| **Common Conditions** |  |  |
| Test item generation: pre-processing incl. spatialization | 1 | Model-based generation according to convolution of raw mono clean speech sentences with FOA Spatial Room Impulse Responses corresponding to the talker positions relative to a capture point and spatial FOA background. MASA format generation from FOA according to MASA analysis [28]. |
| Binaural renderer | 1 | IVAS MASA C Reference Software binaural rendering [28] |
| Audio sampling frequency/bandwidth | 1 | 48 kHz/maximum available audio bandwidth up to FB |
| Kind of samples | 1 | Sentence pair uttered by different talkers and genders (3 male and 3 female) |
| Number of categories | 6 | 6 Different environments and talker interactions |
| Number of samples | 7 | 6 + 1 (preliminaries) samples per category |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve listeners |
| Randomizations | 6 | 6 panels of 5 listeners |
| Rating Scale | 1 | P.SUPPL800 |
| Listening System | 1 | Headphones, in accordance with clause 4.6 |
| Listening Environment | 1 | No room noise, in accordance with clause 4.6 |
| Languages | 2 | Danish, German |

Table E.8.2: Preliminaries for Experiment P800-8

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Trial #** | **Label** | **Sample** | **Condition** | **Bitrate** | **FER/Profile** |
| 1 | c23 |  | EVS | 4x9.6 | [5%, profile] |
| 2 | c02 |  | MNRU Q=32 dB | - | - |
| 3 | c06 |  | ESDRU = 0.85 | - | - |
| 4 | c13 |  | EVS | 4x16.4 | No errors |
| 5 | c08 |  | ESDRU = 0.45 | - | - |
| 6 | c21 |  | EVS | 4x7.2 | [5%, profile] |
| 7 | c04 |  | MNRU Q=22 dB | - | - |
| 8 | c01 |  | Reference | - | - |
| 9 | c14 |  | EVS | 4x24.4 | No errors |
| 10 | c07 |  | ESDRU = 0.65 | - | - |
| 11 | c05 |  | MNRU Q=17 dB | - | - |
| 12 | c22 |  | EVS | 4x8 | [5%, profile] |

Table E.8.3: Test conditions for Experiment P800-8,  
clean speech under clean and impaired channel conditions [including delay/jitter profiles]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Condition** | **Bitrate [kbps]** | **FER/Profile** | **ToR** |
| c01 | Reference | - | - |  |
| c02 | MNRU Q=32 dB | - | - |  |
| c03 | MNRU Q=27 dB | - | - |  |
| c04 | MNRU Q=22 dB | - | - |  |
| c05 | MNRU Q=17 dB | - | - |  |
| c06 | ESDRU | - | - |  |
| c07 | ESDRU | - | - |  |
| c08 | ESDRU | - | - |  |
| c09 | EVS | 3x7.2 | No errors |  |
| c10 | EVS | 4x7.2 | No errors |  |
| c11 | EVS | 4x8 | No errors |  |
| c12 | EVS | 4x9.6 | No errors |  |
|  |  |  |  |  |
| c13 | EVS | 4x16.4 | No errors |  |
| c14 | EVS | 4x24.4 | No errors |  |
| c15 | EVS(+unquantized metadata) | 2x7.2 | No errors |  |
| c16 | EVS(+unquantized metadata) | 2x8 | No errors |  |
| c17 | EVS(+unquantized metadata) | 2x9.6 | No errors |  |
| c18 | EVS(+unquantized metadata) | 2x16.4 | No errors |  |
| c19 | EVS(+unquantized metadata) | 2x24.4 | No errors |  |
| c20 | EVS | 3x7.2 | [5%, profile] |  |
| c21 | EVS | 4x7.2 | [5%, profile] |  |
| c22 | EVS | 4x8 | [5%, profile] |  |
|  |  |  |  |  |
|  |  |  |  |  |
| c23 | EVS | 4x16.4 | [5%, profile] |  |
| c24 | EVS | 4x24.4 | [5%, profile] |  |
| c25 | CuT | 13.2 | No errors | NWT c09 OR NWT c15 |
| c26 | CuT | 16.4 | No errors | NWT c10 OR NWT c15 |
| c27 | CuT | 24.4 | No errors | NWT c11 OR NWT c16 |
| c28 | CuT | 32 | No errors | NWT c12 OR NWT c17 |
| c29 | CuT | 48 | No errors | NWT c13 OR NWT c18 |
| c30 | CuT | 64 | No errors | NWT c14 OR NWT c19 |
| c31 | CuT | 80 | No errors | NWT c14 OR NWT c19 |
| c32 | CuT | 13.2 | [5%, profile] | NWT c20 |
| c33 | CuT | 16.4 | [5%, profile] | NWT c21 |
| c34 | CuT | 24.4 | [5%, profile] | NWT c22 |
|  |  |  |  |  |
| c35 | CuT | 48 | [5%, profile] | NWT c23 |
| C36 | CuT | 64 | [5%, profile] | NWT c24 |

* + 1. Content type categories and scene definitions (Exp P800-8: Clean speech)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Category*** | ***Environment(1*** | ***Level [dB]*** | ***Background(2*** | ***SNR [dB]*** | ***Overtalk [s](3*** | ***Bandwidth*** | ***Talker positions(4*** | ***Talker selection by panel(5*** |
| *cat 1* | *room\_1\_MASA* | *-26* | *room\_1\_cleanbg\_MASA* | *[50]* | *1* | *Max* | *tbd* | *P1: f1m1 P2: m2f2 P3: f3m3 P4: m1f1 P5: f2m2 P6: m3f3* |
| *cat 2* | *room\_[1/4]\_MASA* | *-26* | *room\_[1/4]\_cleanbg\_MASA* | *[50]* | *-1* | *Max* | *tbd* | *P1: m3f3 P2: f1m1 P3: m2f2 P4: f3m3 P5: m1f1 P6: f2m2* |
| *cat 3* | *room\_2\_MASA* | *-26* | *room\_2\_cleanbg\_MASA* | *[50]* | *1* | *Max* | *tbd* | *P1: f2m2 P2: m3f3 P3: f1m1 P4: m2f2 P5: f3m3 P6: m1f1* |
| *cat 4* | *room\_[2/5]\_MASA* | *-26* | *room\_[2/5]\_cleanbg\_MASA* | *[50]* | *-1* | *Max* | *tbd* | *P1: m1f1 P2: f2m2 P3: m3f3 P4: f1m1 P5: m2f2 P6: f3m3* |
| *cat 5* | *room\_3\_MASA* | *-26* | *room\_3\_cleanbg\_MASA* | *[50]* | *1* | *Max* | *tbd* | *P1: f3m3 P2: m1f1 P3: f2m2 P4: m3f3 P5: f1m1 P6: m2f2* |
| *cat 6* | *room\_[3/6]\_MASA* | *-26* | *room\_[3/6]\_cleanbg\_MASA* | *[50]* | *-1* | *Max* | *tbd* | *P1: m2f2 P2: f3m3 P3: m1f1 P4: f2m2 P5: m3f3 P6: f1m1* |

**Notes:**

**(1** Editor’s note: The specific room/environment characteristic and resulting reverb characteristic will be defined by the choice of the specific Spatial Room Impulse Responses used in the convolution process with the raw mono sentences, according to the pertaining stipulations of the test plan IVAS-8a.

(2 Editor’s note: Background is defined by the chosen background noise file according to the pertaining stipulations of the test plan IVAS-8a. Backround name ‘clean\_bg\_[X]\_MASA’ indicates a low-noise background corresponding to environment [X], e.g., with low air-conditioning/fan noise.

**(3** Overtalk [s] means the duration in seconds by which the two sentences in the sound item uttered by different talkers are overlapping. A negative number means that there is a corresponding pause between the two sentences.

**(4** Editor’s note: The talker positions are part of the scene definition of the different categories. They should be chosen in a way from the available set of SRIRs for the used room making sure that there is a good coverage of different possible positions. Different selections should be made for the different listener panels.

**(5** All sentences by the 6 talkers shall be unique.

]

* 1. Experiment P800-9: MASA Speech+Background Test

[

* + 1. Experiment setup

Stereo-MASA inputs are tested in IVAS selection phase. The stereo-MASA inputs can have 1-direction spatial metadata (based, e.g., on MASA analysis [28] of original FOA content) or 2-direction spatial metadata (based, e.g., on MASA analysis [28] of original HOA2 content).

Tables E.9.1 to E.9.3 show conditions to be used for this experiment, list of preliminaries and full list of conditions, respectively.

Table E.9.1: Factors and conditions for Experiment P800- 9 (Speech+Background)

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Candidate | 1 | CuT |
| Bitrates | 7 | 13.2, 16.4, 24.4, 32, 48, 64, 80 |
| DTX | 2 | DTX ON and OFF |
| Input level | 1 | -26 LKFS |
| Input frequency mask | 1 | HP50 |
| Noise | 1 | 15dB |
| Error Conditions | 0 | No errors |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 26+3 | Multi-mono EVS + Dual-mono with unquantized metadata |
| Bitrates | 6 | 3x7.2, 4x7.2, 4x8, 4x9.6, 4x16.4, 4x24.4 kbps |
|  | 3 | 2x7.2, 2x9.6, 2x16.4 kbps |
| Input level  DTX | 1  2 | -26 LKFS  DTX On and OFF |
| Input frequency mask | 1 | HP50 |
| Noise | 1 | 15dB |
| Error Conditions | 0 | No errors |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | -26 LKFS |
| P.50 MNRU  ESDRU | 4  3 | Q= 17, 22, 27, 32] dB  *α* = 0.45, 0.65, 0.85 |
| Input frequency mask | 1 | HP50 |
| **Common Conditions** |  |  |
| Test item generation: pre-processing incl. spatialization | 1 | Model-based generation according to convolution of raw mono clean speech sentences with FOA Spatial Room Impulse Responses corresponding to the talker positions relative to a capture point and spatial FOA background. MASA format generation from FOA according to MASA analysis [28]. |
| Binaural renderer | 1 | IVAS MASA C Reference Software binaural rendering [28] |
| Audio sampling frequency/bandwidth | 1 | 48 kHz/maximum available audio bandwidth up to FB |
| Kind of samples | 1 | Sentence pair uttered by different talkers and genders (3 male and 3 female) |
| Number of categories | 6 | 6 Different environments and talker interactions |
| Number of samples | 7 | 6 + 1 (preliminaries) samples per category |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve listeners |
| Randomizations | 6 | 6 panels of 5 listeners |
| Rating Scale | 1 | P.SUPPL800 |
| Listening System | 1 | Headphones, in accordance with clause 4.6 |
| Listening Environment | 1 | No room noise, in accordance with clause 4.6 |
| Languages | 2 | Japanese, French |

Table E.9.2: Preliminaries for Experiment P800-8

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Trial #** | **Label** | **Sample** | **Condition** | **Bitrate** | **DTX** |
| 1 | c21 |  | EVS | 4x9.6 | on |
| 2 | c02 |  | MNRU Q=32 dB | - | - |
| 3 | c06 |  | ESDRU = 0.85 | - | - |
| 4 | c13 |  | EVS | 4x16.4 | off |
| 5 | c08 |  | ESDRU = 0.45 | - | - |
| 6 | c19 |  | EVS | 4x7.2 | on |
| 7 | c04 |  | MNRU Q=22 dB | - | - |
| 8 | c01 |  | Reference | - | - |
| 9 | c14 |  | EVS | 4x24.4 | off |
| 10 | c07 |  | ESDRU = 0.65 | - | - |
| 11 | c05 |  | MNRU Q=17 dB | - | - |
| 12 | c20 |  | EVS | 4x8 | on |

Table E.9.3: Test conditions for Experiment P800-9,  
speech and background under clean channel conditions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Condition** | **Bitrate [kbps]** | **DTX** | **ToR** |
| c01 | Reference | - | - |  |
| c02 | MNRU Q=32 dB | - | - |  |
| c03 | MNRU Q=27 dB | - | - |  |
| c04 | MNRU Q=22 dB | - | - |  |
| c05 | MNRU Q=17 dB | - | - |  |
| c06 | ESDRU 5 | - | - |  |
| c07 | ESDRU | - | - |  |
| c08 | ESDRU | - | - |  |
| c09 | EVS | 3x7.2 | off |  |
| c10 | EVS | 4x7.2 | off |  |
| c11 | EVS | 4x8 | off |  |
| c12 | EVS | 4x9.6 | off |  |
| c13 | EVS | 4x16.4 | off |  |
| c14 | EVS | 4x24.4 | off |  |
| c15 | EVS(+unquantized metadata) | 2x7.2 | on |  |
| c16 | EVS(+unquantized metadata) | 2x9.6 | on |  |
| c17 | EVS(+unquantized metadata) | 2x16.4 | on |  |
| c18 | EVS | 3x7.2 | on |  |
| c19 | EVS | 4x7.2 | on |  |
| c20 | EVS | 4x8 | on |  |
| c21 | EVS | 4x9.6 | on |  |
| c22 | EVS | 4x16.4 | on |  |
| c23 | EVS | 4x24.4 | on |  |
| c24 | CuT | 13.2 | off | NWT c09 | |
| c25 | CuT | 16.4 | off | NWT c10 | |
| c26 | CuT | 24.4 | off | NWT c11 | |
| c27 | CuT | 32 | off | NWT c12 | |
| c28 | CuT | 48 | off | NWT c13 | |
| c29 | CuT | 64 | off | NWT c14 | |
| c30 | CuT | 80 | off | NWT c14 | |
| c31 | CuT | 13.2 | on | NWT c18 or NWT c15 | |
| c32 | CuT | 16.4 | on | NWT c19 or NWT c15 | |
| c33 | CuT | 24.4 | on | NWT c20 | |
| c34 | CuT | 32 | on | NWT c21 or NWT c16 | |
| c35 | CuT | 48 | on | NWT c22 or NWT c17 | |
| c36 | CuT | 64 | on | NWT c23 | |

* + 1. Content type categories and scene definitions (Exp P800-9: Speech+Background)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Category*** | ***Environment(1*** | ***Level [dB]*** | ***Background(2*** | ***SNR [dB]*** | ***Overtalk [s](3*** | ***Bandwidth*** | ***Talker positions(4*** | ***Talker selection by panel(5*** |
| *cat 1* | *car\_1\_MASA* | *-26* | *car\_1\_bg\_MASA* | *10* | *-1* | *Max* | *tbd* | *P1: f1m1 P2: m2f2 P3: f3m3 P4: m1f1 P5: f2m2 P6: m3f3* |
| *cat 2* | *car\_[1/2]\_MASA* | *-26* | *car\_[1/2]\_bg\_MASA* | *10* | *-1* | *Max* | *tbd* | *P1: m3f3 P2: f1m1 P3: m2f2 P4: f3m3 P5: m1f1 P6: f2m2* |
| *cat 3* | *out\_1\_MASA* | *-26* | *street\_1\_bg\_MASA* | *[15]* | *-1* | *Max* | *tbd* | *P1: f2m2 P2: m3f3 P3: f1m1 P4: m2f2 P5: f3m3 P6: m1f1* |
| *cat 4* | *out\_[1/2]\_MASA* | *-26* | *[park\_1\_bg\_MASA / nature\_1\_bg\_MASA / event\_1\_bg\_MASA / street\_[1/2]\_bg\_MASA]* | *[15]* | *-1* | *Max* | *tbd* | *P1: m1f1 P2: f2m2 P3: m3f3 P4: f1m1 P5: m2f2 P6: f3m3* |
| *cat 5* | *room\_[X]\_MASA* | *-26* | *office\_1\_bg\_MASA* | *[15]* | *-1* | *Max* | *tbd* | *P1: f3m3 P2: m1f1 P3: f2m2 P4: m3f3 P5: f1m1 P6: m2f2* |
| *cat 6* | *room\_[X]\_MASA* | *-26* | *[cafeteria\_1\_bg\_MASA / mall\_1\_bg\_MASA/ office[1/2]\_bg\_MASA]* | *[15]* | *-1* | *Max* | *tbd* | *P1: m2f2 P2: f3m3 P3: m1f1 P4: f2m2 P5: m3f3 P6: f1m1* |

**Notes:**

**(1** Editor’s note: The specific room/environment characteristic and resulting reverb characteristic will be defined by the choice of the specific Spatial Room Impulse Responses used in the convolution process with the raw mono sentences, according to the pertaining stipulations of the test plan IVAS-8a.

**(2** Editor’s note: Background is defined by the chosen background noise file according to the pertaining stipulations of the test plan IVAS-8a.

**(3** Overtalk [s] means the duration in seconds by which the two sentences in the sound item uttered by different talkers are overlapping. A negative number means that there is a corresponding pause between the two sentences.

**(**4 Editor’s note: The talker positions are part of the scene definition of the different categories. They should be chosen in a way from the available set of SRIRs for the used room making sure that there is a good coverage of different possible positions. Different selections should be made for the different listener panels.

**(5** All sentences by the 6 talkers shall be unique.

]

1. BS.1534 Experiments
   1. Experiment BS1534-1a: Stereo

[

Factors and conditions (BS1534-1a Generic Audio)

|  |  |  |
| --- | --- | --- |
| ***Main Codec Conditions*** |  |  |
| *Codec under Test (CuT)* | *2* | *IVAS candidate operated with stereo audio input at*  *48 and 64 kbps DTX off at 0% FER* |
|  |  |  |
| ***Codec references*** |  |  |
| *Codec references* | *3* | *Dual-mono EVS*  *2\*24.4 kbps, 2\*32 kbps, 2\*48 kbps DTX off at 0% FER* |
|  |  |  |
| ***Other references*** |  |  |
| *Reference* | *1* | *Direct signal, Nominal input level* |
| *Hidden Reference* | *1* | *Direct signal, Nominal input level* |
| *LP7 anchor* | *1* | *7 kHz lowpass filtered signal, nominal level* |
|  |  |  |
| ***Common Conditions*** |  |  |
| *Test item generation* | *1* | *According to material collection procedure for IVAS selection MUSHRA tests.* |
| *Audio sampling frequency/bandwidth* | *1* | *48 kHz/FB* |
| *Input frequency mask* | *1* | *20KBP* |
| *Nominal output loudness* | *1* | *-26 LKFS [12]* |
| *Listening Level* | *1* | *Adjusted by listener* |
| *Listeners* | 14 | *Experienced Listeners* |
| *Randomizations* | 14 | *Individual per listeners* |
| *Rating Scale* | *1* | *Continuous Mushra scale from 0-100 [22]* |
| *Listening System* | *1* | *High-quality headphone for diotic presentation, in accordance with clause 4.6* |
| *Listening Environment* | *1* | *No room noise* |

]

* 1. Experiment BS1534-1b: Stereo

[

Factors and conditions (BS1534-1b Generic Audio)

|  |  |  |
| --- | --- | --- |
| ***Main Codec Conditions*** |  |  |
| *Codec under Test (CuT)* | *2* | *IVAS candidate operated with stereo audio input at*  *96 and 128 kbps DTX off at 5% FER* |
|  |  |  |
| ***Codec references*** |  |  |
| *Codec references* | *3* | *Dual-mono EVS*  *2\*48 kbps, 2\*64 kbps, 2\*96 kbps DTX off at 5% FER* |
|  |  |  |
| ***Other references*** |  |  |
| *Reference* | *1* | *Direct signal, Nominal input level* |
| *Hidden Reference* | *1* | *Direct signal, Nominal input level* |
| *LP7 anchor* | *1* | *7 kHz lowpass filtered signal, nominal level* |
|  |  |  |
| ***Common Conditions*** |  |  |
| *Test item generation* | *1* | *According to material collection procedure for IVAS selection MUSHRA tests.* |
| *Audio sampling frequency/bandwidth* | *1* | *48 kHz/FB* |
| *Input frequency mask* | *1* | *20KBP* |
| *Nominal output loudness* | *1* | *-26 LKFS [12]* |
| *Listening Level* | *1* | *Adjusted by listener* |
| *Listeners* | 14 | *Experienced Listeners* |
| *Randomizations* | 14 | *Individual per listeners* |
| *Rating Scale* | *1* | *Continuous MUSHRA scale from 0-100 [22]* |
| *Listening System* | *1* | *High-quality headphone for diotic presentation, in accordance with clause 4.6* |
| *Listening Environment* | *1* | *No room noise* |

]

* 1. Experiment BS1534-2a: Multi-channel 5.1

[

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 2 | IVAS candidate operated at 64, 96 kbps at 0% FER |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 3 | Multi-mono EVS operated at 5\*13.2, 5\*16.4, 5\*24.4 kbps with the LFE channel processed using EVS operated at 9.6 kbps NB (for all Codec references) at 0% FER |
|  |  |  |
| **Other references** |  |  |
| Reference / Direct | 1 | Direct 5.1 signal, nominal level |
| Hidden Reference | 1 | Direct 5.1 signal, nominal level |
| LP7 anchor | 1 | 7 kHz lowpass filtered direct 5.1 signal, nominal level |
|  |  |  |
| **Common Conditions** |  |  |
| Test item generation | 1 | According to material collection procedure for IVAS selection MUSHRA tests. |
| Loudspeaker Rendering | 1 | 5.1 channels direct playback |
| Audio sampling frequency / bandwidth | 2 | 48 kHz / maximum available audio bandwidth (SWB, FB) |
| Input frequency mask | 1 | *20KBP* |
| Nominal output loudness | 1 | -26 LKFS [12] |
| Listening Level | 1 | Adjusted by listener |
| Listeners | 14 | Experienced Listeners |
| Randomizations | 14 | Individual per listener |
| Rating Scale | 1 | Continuous MUSHRA scale: 0-100 [22] |
| Listening System | 1 | 5.1 high-quality loudspeaker setup |
| Listening Environment | 1 | No room noise |

]

* 1. Experiment BS1534-2b: Multi-channel 5.1

[

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 2 | Option 1: IVAS candidate operated at 128, 160 kbps at 0% FER  Option 2: IVAS candidate operated at 64, 96 kbps at 3% FER |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 3 | Option 1: Multi-mono EVS operated at 4\*24.4, 5\*32, 5\*48 kbps with the LFE channel processed using EVS operated at 9.6 kbps NB (for all Codec references) at 0% FER  Option 2: Multi-mono EVS operated at 5\*13.2, 5\*16.4, 5\*24.4 kbps with the LFE channel processed using EVS operated at 9.6 kbps NB (for all Codec references) at 3% FER |
|  |  |  |
| **Other references** |  |  |
| Reference | 1 | Direct 5.1 signal, nominal input level |
| Hidden Reference | 1 | Direct 5.1 signal, nominal input level |
| LP7 anchor | 1 | 7 kHz lowpass filtered direct 5.1 signal, nominal level |
|  |  |  |
| **Common Conditions** |  | *(see BS1534-2a for full list)* |
| Audio sampling frequency / bandwidth | 1 or 2 | Option 1: 48 kHz / FB  Option 2: 48 kHz / maximum available audio bandwidth (SWB, FB) |

]

* 1. Experiment BS1534-3a: Multi-channel 7.1+4

[

Factors and conditions (BS1534-3a Generic Audio)

|  |  |  |
| --- | --- | --- |
| ***Main Codec Conditions*** |  |  |
| *Codec under Test (CuT)* | *2* | *IVAS candidate operated with multi-channel 7.1+4 input at*  *128 and 160 kbps DTX off at 0% FER* |
|  |  |  |
| ***Codec references*** |  |  |
| *Codec references* | *3* | *Multi-mono EVS coding multi-channel 7.1+4 input at*  *11\*9.6 kbps, 11\*13.2 kbps, 11\*16.4 kbps DTX off at 0% FER*  *LFE coded with 9.6kbps NB (IVAS-3)* |
|  |  |  |
| ***Other references*** |  |  |
| *Reference* | *1* | *Direct signal, Nominal input level* |
| *Hidden Reference* | *1* | *Direct signal, Nominal input level* |
| *LP7 anchor* | *1* | *7 kHz lowpass filtered signal, nominal level* |
|  |  |  |
| ***Common Conditions*** |  |  |
| *Test item generation* | *1* | *According to material collection procedure for IVAS selection MUSHRA tests.* |
| *Audio sampling frequency/bandwidth* | *1* | *48 kHz/FB* |
| *Input frequency mask* | *1* | *20KBP* |
| *Nominal output loudness* | *1* | *-26 LKFS [12]* |
| *Listening Level* | *1* | *Adjusted by listener* |
| *Listeners* | 14 | *Experienced Listeners* |
| *Randomizations* | 14 | *Individual per listeners* |
| *Rating Scale* | *1* | *Continuous MUSHRA scale from 0-100 [22]* |
| *Listening System* | *1* | *High-quality loudspeaker: 7.1+4 overhead speaker setup with the CICP19 configuration* |
| *Listening Environment* | *1* | *No room noise* |

]

* 1. Experiment BS1534-3b: Multi-channel 7.1+4

[

Factors and conditions (BS1534-3b Generic Audio)

|  |  |  |
| --- | --- | --- |
| ***Main Codec Conditions*** |  |  |
| *Codec under Test (CuT)* | *2* | *IVAS candidate operated with multi-channel 7.1+4 audio input at*  *384 and 512 kbps DTX off at 0% FER* |
|  |  |  |
| ***Codec references*** |  |  |
| *Codec references* | *3* | *Multi-mono EVS coding multi-channel 7.1+4 input at*  *11\*32 kbps, 11\*48 kbps, 11\*64 kbps DTX off at 0% FER*  *LFE coded with 9.6kbps NB (IVAS-3)* |
|  |  |  |
| ***Other references*** |  |  |
| *Reference* | *1* | *Direct signal, Nominal input level* |
| *Hidden Reference* | *1* | *Direct signal, Nominal input level* |
| *LP7 anchor* | *1* | *7 kHz lowpass filtered signal, nominal level* |
|  |  |  |
| ***Common Conditions*** |  |  |
| *Test item generation* | *1* | *According to material collection procedure for IVAS selection MUSHRA tests.* |
| *Audio sampling frequency/bandwidth* | *2* | *48 kHz/FB* |
| *Input frequency mask* | *1* | *20KBP* |
| *Nominal output loudness* | *1* | *-26 LKFS [12]* |
| *Listening Level* | *1* | *Adjusted by listener* |
| *Listeners* | 14 | *Experienced Listeners* |
| *Randomizations* | 14 | *Individual per listeners* |
| *Rating Scale* | *1* | *Continuous MUSHRA scale from 0-100 [22]* |
| *Listening System* | *1* | *High-quality loudspeaker: 7.1+4 overhead speaker setup with the CICP19 configuration* |
| *Listening Environment* | *1* | *No room noise* |

]

* 1. Experiment BS1534-4a: FOA

[

Factors and conditions (BS1534-4a, Generic Audio)

|  |  |  |
| --- | --- | --- |
| ***Main Codec Conditions*** |  |  |
| *Codec under Test (CuT)* | *3* | *IVAS candidate operated with audio input truncated to FOA at*  *96 kbps, 128 kbps, 160 kbps* |
|  |  |  |
| ***Codec references*** |  |  |
| *Codec references* | *3* | *Multi-mono EVS operated with audio input truncated to FOA at  4\*24.4 kbps, 4\*32 kbps, 4\*48 kbps* |
|  |  |  |
| ***Other references*** |  |  |
| *Reference* | *1* | *Direct rendering of HOA3 signal, Nominal input level* |
| *Hidden Reference* | *1* | *Direct rendering of HOA3 signal, Nominal input level* |
| *LP7 anchor* | *1* | *7 kHz lowpass filtered direct rendered HOA3 signal: nominal level* |
|  |  |  |
| ***Common Conditions*** |  |  |
| *Test item generation* | *1* | *According to material collection procedure for IVAS selection Mushra tests.* |
| ***Binaural renderer*** | *1* | *Ambisonics to binaural (external) rendering* |
| *Audio sampling frequency/bandwidth* | *2* | *48 kHz/FB* |
| *Input frequency mask* | *1* | *20KBP* |
| *Nominal output loudness* | *1* | *-26 LKFS [12]* |
| *Listening Level* | *1* | *Adjusted by listener* |
| *Listeners* | *14* | *Experienced Listeners* |
| *Randomizations* | *14* | *Individual per listeners* |
| *Rating Scale* | *1* | *Continuous Mushra scale from 0-100 [22]* |
| *Listening System* | *1* | *High-quality headphone for diotic presentation, in accordance with clause 4.6* |
| *Listening Environment* | *1* | *No room noise* |

]

* 1. Experiment BS1534-4b: HOA2

[

Factors and conditions (BS1534-4b, Generic Audio)

|  |  |  |
| --- | --- | --- |
| ***Main Codec Conditions*** |  |  |
| *Codec under Test (CuT)* | *2* | *IVAS candidate operated with audio input truncated to HOA2 at*  *192 kbps, 256 kbps.* |
|  |  |  |
| ***Codec references*** |  |  |
| *Codec references* | *3* | *Multi-mono EVS operated with audio input truncated to FOA at  4\*48 kbps, 4\*64 kbps, 4\*96 kbps.* |
|  |  |  |
| ***Other references*** |  |  |
| *Reference* | *1* | *Direct rendering of HOA3 signal, Nominal input level* |
| *Hidden Reference* | *1* | *Direct rendering of HOA3 signal, Nominal input level* |
| *LP7 anchor* | *1* | *7 kHz lowpass filtered direct rendered HOA3 signal: nominal level* |
|  |  |  |
| ***Common Conditions*** |  |  |
| *Test item generation* | *1* | *According to material collection procedure for IVAS selection Mushra tests.* |
| ***Binaural renderer*** | *1* | *Ambisonics to binaural (external) rendering* |
| *Audio sampling frequency/bandwidth* | *2* | *48 kHz/FB* |
| *Input frequency mask* | *1* | *20KBP* |
| *Nominal output loudness* | *1* | *-26 LKFS [12]* |
| *Listening Level* | *1* | *Adjusted by listener* |
| *Listeners* | *14* | *Experienced Listeners* |
| *Randomizations* | *14* | *Individual per listeners* |
| *Rating Scale* | *1* | *Continuous Mushra scale from 0-100 [22]* |
| *Listening System* | *1* | *High-quality headphone for diotic presentation, in accordance with clause 4.6* |
| *Listening Environment* | *1* | *No room noise* |

]

* 1. Experiment BS1534-5a: HOA3

[

Factors and conditions (BS1534-5a, Generic Audio)

|  |  |  |
| --- | --- | --- |
| ***Main Codec Conditions*** |  |  |
| *Codec under Test (CuT)* | *2* | *IVAS candidate operated with HOA3 input at*  *192 kbps, 256 kbps.* |
|  |  |  |
| ***Codec references*** |  |  |
| *Codec references* | *3* | *Multi-mono EVS operated with audio input truncated to FOA at  4\*48 kbps, 4\*64, 4\*96 kbps.* |
|  |  |  |
| ***Other references*** |  |  |
| *Reference* | *1* | *Direct rendering of HOA3 signal, Nominal input level* |
| *Hidden Reference* | *1* | *Direct rendering of HOA3 signal, Nominal input level* |
| *LP7 anchor* | *1* | *7 kHz lowpass filtered direct rendered HOA3 signal: nominal level* |
|  |  |  |
| ***Common Conditions*** |  |  |
| *Test item generation* | *1* | *According to material collection procedure for IVAS selection Mushra tests.* |
| ***Binaural renderer*** | *1* | *Ambisonics to binaural (external) rendering* |
| *Audio sampling frequency/bandwidth* | *2* | *48 kHz/FB* |
| *Input frequency mask* | *1* | *20KBP* |
| *Nominal output loudness* | *1* | *-26 LKFS [12]* |
| *Listening Level* | *1* | *Adjusted by listener* |
| *Listeners* | *14* | *Experienced Listeners* |
| *Randomizations* | *14* | *Individual per listeners* |
| *Rating Scale* | *1* | *Continuous Mushra scale from 0-100 [22]* |
|  |  |  |
| *Listening System* | *1* | *High-quality headphone for diotic presentation, in accordance with clause 4.6.* |
| *Listening Environment* | *1* | *No room noise* |

]

* 1. Experiment BS1534-5b: HOA3

[

Factors and conditions (BS1534-5b, Generic Audio)

|  |  |  |
| --- | --- | --- |
| ***Main Codec Conditions*** |  |  |
| *Codec under Test (CuT)* | *2* | *IVAS candidate operated with HOA3 input at*  *384 kbps, 512 kbps.* |
|  |  |  |
| ***Codec references*** |  |  |
| *Codec references* | *2* | *Multi-mono EVS operated with audio input truncated to FOA at  4\*96 kbps, 4\*128 kbps.* |
|  |  |  |
| ***Other references*** |  |  |
| *Reference* | *1* | *Direct rendering of HOA3 signal, Nominal input level* |
| *Hidden Reference* | *1* | *Direct rendering of HOA3 signal, Nominal input level* |
| *LP7 anchor* | *1* | *7 kHz lowpass filtered direct rendered HOA3 signal: nominal level* |
|  |  |  |
| ***Common Conditions*** |  |  |
| *Test item generation* | *1* | *According to material collection procedure for IVAS selection Mushra tests.* |
| ***Binaural renderer*** | *1* | *Ambisonics to binaural (external) rendering* |
| *Audio sampling frequency/bandwidth* | *2* | *48 kHz/FB* |
| *Input frequency mask* | *1* | *20KBP* |
| *Nominal output loudness* | *1* | *-26 LKFS [12]* |
| *Listening Level* | *1* | *Adjusted by listener* |
| *Listeners* | *14* | *Experienced Listeners* |
| *Randomizations* | *14* | *Individual per listeners* |
| *Rating Scale* | *1* | *Continuous Mushra scale from 0-100 [22]* |
| *Listening System* | *1* | *Calibrated and conformant 7.1.4 listening room with single listener in center.* |
| *Listening Environment* | *1* | *No room noise* |

]

* 1. Experiment BS1534-6a: Objects

[

It is proposed to test with 3 Objects in BS1534-6a at mid bitrates.

Note: It is assumed that the rendered immersive audio scenes reflect reasonable use case scenarios. E.g., if objects correspond to several talkers, a partial overtalk of two talkers is expected, as well as some talkers moving around the scene at natural pace. However, it is not expected that all talkers are active all the time, with unnaturally rapid displacements.

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 3 | IVAS candidate operated at 48, 64, 96 kbps |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 3 | Multi-mono EVS operated at 3\*16.4, 3\*24.4, 3\*32 kbps |
|  |  |  |
| **Other references** |  |  |
| Reference / Direct | 1 | Direct Object signal(s), nominal input level |
| Hidden Reference | 1 | Direct Object signal(s), nominal input level |
| LP7 anchor | 1 | 7 kHz lowpass filtered direct Object signal(s), nominal level |
|  |  |  |
| **Common Conditions** |  |  |
| Test item generation | 1 | According to material collection procedure for IVAS selection MUSHRA tests. |
| Binaural rendering | 1 | Objects to binaural (external) rendering |
| Audio sampling frequency / bandwidth | 1 | 48 kHz / FB |
| Input frequency mask | 1 | 20KBP |
| Nominal output loudness | 1 | -26 LKFS [12] |
| Listening Level | 1 | Adjusted by listener |
| Listeners | 14 | Experienced Listeners |
| Randomizations | 14 | Individual per listener |
| Rating Scale | 1 | Continuous MUSHRA scale: 0-100 [22] |
| Listening System | 1 | High-quality headphones, diotic presentation, in accordance with clause 4.6 |
| Listening Environment | 1 | No noise |

]

* 1. Experiment BS1534-6b: Objects

[

It is proposed to test with 4 Objects in BS1534-6b at mid/high bitrates. Alternatively, one experiment can focus on clean channel performance while the other focuses on DTX or FER performance.

Note: It is assumed that the rendered immersive audio scenes reflect reasonable use case scenarios. E.g., if objects correspond to several talkers, a partial overtalk of two talkers is expected, as well as some talkers moving around the scene at natural pace. However, it is not expected that all talkers are active all the time, with unnaturally rapid displacements.

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 3 or 1 | Option 1: IVAS candidate operated at 96, 160, 256 kbps with DTX off at 0% FER  Option 2 and Option 3: IVAS candidate operated at 96 kbps with DTX off at 0% FER |
| 0 or 1 | Option 2: IVAS candidate operated at 96 kbps with DTX on at 0% FER  Option 3: IVAS candidate operated at 96 kbps with DTX off at 3% FER |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 3 or 1 | Option 1: Multi-mono EVS operated at 4\*24.4, 4\*48, 4\*64 kbps with DTX off at 0% FER  Option 2 and Option 3: Multi-mono EVS operated at 4\*24.4 kbps with DTX off at 0% FER |
| 0 or 1 | Option 2: Multi-mono EVS operated at 4\*24.4 kbps with DTX on at 0% FER  Option 3: Multi-mono EVS operated at 4\*24.4 kbps with DTX off at 3% FER |
|  |  |  |
| **Other references** |  |  |
| Reference / Direct | 1 | Direct Object signal(s), nominal input level |
| Hidden Reference | 1 | Direct Object signal(s), nominal input level |
| LP7 anchor | 1 | 7 kHz lowpass filtered direct Object signal(s), nominal level |
|  |  |  |
| **Common Conditions** |  | *(see BS1534-6a for full list)* |

]

* 1. Experiment BS1534-7a: MASA

[

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 2 | IVAS candidate operated at 64, 80 kbps |
|  | 1 | IVAS candidate operated at 48 kbps with DTX off at 3% or 5% FER |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 1 | Multi-mono EVS operated at 4\*24.4 kbps with DTX off at 0% FER |
| 1 | Multi-mono EVS with unquantized metadata operated at 2\*24.4 kbps with DTX off at 0% FER |
| 1 | Multi-mono EVS operated at 4\*16.4 kbps with DTX off at 3% or 5% FER |
|  |  |  |
| **Other references** |  |  |
| Reference / Direct | 1 | Direct MASA signal(s), nominal input level |
| Hidden Reference | 1 | Direct MASA signal(s), nominal input level |
| LP7 anchor | 1 | 7 kHz lowpass filtered direct MASA signal(s), nominal level |
|  |  |  |
| **Common Conditions** |  |  |
| Test item generation | 1 | According to material collection procedure for IVAS selection MUSHRA tests |
| Binaural rendering | 1 | MASA to binaural (external) rendering |
| Audio sampling frequency / bandwidth | 1 | 48 kHz / FB |
| Input frequency mask | 1 | 20KBP |
| Nominal output loudness | 1 | -26 LKFS ([ITU-R BS.1770-4]) |
| Listening Level | 1 | Adjusted by listener |
| Listeners | 14 | Experienced Listeners |
| Randomizations | 14 | Individual per listener |
| Rating Scale | 1 | Continuous MUSHRA scale: 0-100 [ITU-R BS.1534] |
| Listening System | 1 | High-quality headphones, diotic presentation, in accordance with clause 4.6 |
| Listening Environment | 1 | No noise |

]

* 1. Experiment BS1534-7b: MASA

[

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 2 | IVAS candidate operated at 160, 256 kbps with DTX off at 0% FER |
| 1 | IVAS candidate operated at 96 kbps with DTX off at 3% FER |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 2 | Multi-mono EVS operated at 4\*48, 4\*96 kbps with DTX off at 0% FER |
| 2 | Multi-mono EVS with unquantized metadata operated at 2\*48, 2\*96 kbps with DTX off at 0% FER |
| 1 | Multi-mono EVS operated at 4\*32 kbps with DTX off at 3% FER |
|  |  |  |
| **Other references** |  |  |
| Reference / Direct | 1 | Direct MASA signal(s), nominal input level |
| Hidden Reference | 1 | Direct MASA signal(s), nominal input level |
| LP7 anchor | 1 | 7 kHz lowpass filtered direct MASA signal(s), nominal level |
|  |  |  |
| **Common Conditions** |  | *(see BS1534-7a for full list)* |

]

Examples of test designs potentially relevant for IVAS codec testing

Introduction

This Appendix contains a collection of experimental designs that are deemed potentially relevant for IVAS codec testing. When creating the IVAS codec selection and characterizations test plans SA4 may decide to resort to concepts of these designs.

Example 1: Modified P.800 DCR test of parametric spatial speech [4], [5]

Test purpose

The main purposes for the experiment were: to evaluate the updated IVAS MASA C Reference Software package [6], [7]; to study the suitability of modified ITU-T P.800 [8] DCR and P.811 [9] methodologies for experiments using real spatial speech recordings; to evaluate quality of potential reference conditions for MASA format with degradation anchors spanning both signal and spatial quality dimensions.

Test outline

The listening test experiment was designed for evaluation of potential reference conditions for the parametric metadata-assisted spatial audio (MASA) format with degradation anchors spanning both signal and spatial quality dimensions.

Content types and material generation:

* Realistic spatial speech items in real environments and controlled environments where background was generated using loudspeakers
* The audio capture use cases can be described as “realistic spatial audio communications and user-generated content capture scenarios”
* Audio was recorded in various indoor and outdoor environments using Eigenmike, Eigenmike + external microphone pair, Ambisonic + external cardioid pair, and (for a single category) a multi-microphone smartphone mockup
* Majority of the captured signals were analyzed with the updated IVAS MASA C Reference Software [S4-210840] with the sole exception of the smartphone mockup samples that were analyzed using an in-house parametric analysis method
* Binaural rendering was performed with IVAS MASA C Reference Software [6], [7] package for all conditions.

Evaluation and listening system/environment:

* Modified P.800 DCR test method using real spatial speech recordings with parametric representation
* Anchor conditions based on P.50 MNRU and P.811 ESDRU
* Binaural listening was conducted using Sennheiser HD650 headphones in quiet booths

Detailed test description

* Following provides detailed description of the test:
* 16 test subjects
* Eight sample categories
* Four randomizations for each 4-listener set
* Four samples per category (one for each listening panel)
* 128 votes casted for each condition
* Total of 24 conditions: 7 Reference conditions, 8 coded reference 2xEVS conditions (with unquantized (UQ) spatial metadata), 9 CuTs
* 5-scale DCR test methodology with updated instructions and revised voting scale
* Degradation references: P.50 MNRU and ESDRU
* P.50 MNRU Q values of 30, 24, and 18 dB were used
* ESDRU values of 0.85, 0.70, and 0.55 were used
* Average trial duration: 20 s
* 8 s reference sample + 0.5 s silence + 8 s test sample + 3.5 s voting period
* Test duration: ~1.8 h per listening panel including instructions, preliminaries, and rest breaks

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 9 | Nokia-internal IVAS MASA coding system |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 8 | Dual-mono EVS (2xEVS) with unquantized MASA metadata operated at 2\*8(WB), 2\*9.6, 2\*13.2, 2\*16.4, 2\*24.4, 2\*32, 2\*48, 2\*64 kbps.  Rendering with IVAS MASA C Reference binaural renderer [6], [7]. |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | Analysed with the updated IVAS MASA C Reference software [S4-210840]. No transport stream nor MASA spatial metadata compression.  Rendering done with IVAS MASA C Reference binaural renderer [6], [7]. |
| P.50 MNRU (applied to MASA transport streams) | 3 | Q = 18, 24, 30 dB (output loudness set to nominal level) |
| ESDRU (applied to binaural rendering) | 3 | α = 0.55, 0.7, 0.85 (output loudness set to nominal level) |
|  |  |  |
| **Common Conditions** |  |  |
| Test item generation | 4 | Multi-channel recordings in real environments analysed with the updated IVAS MASA C Reference Software [7] in various configurations or (for single category) using an in-house system. |
| Binaural rendering | 1 | Rendering done with IVAS MASA C Reference renderer [6], [7]. |
| Audio sampling frequency / bandwidth | 2 | 48 kHz/SWB except for reference condition 2xEVS@2\*8kbps which used 48 kHz/WB |
| Rating Scale | 1 | DCR with modified instructions and scale considered more suitable for binaural/spatial telephony (see “Instructions to listeners”) |
| Languages | 1 | Finnish |
| Listening System | 1 | Sennheiser HD650 headphones for binaural presentation |
| Listening Environment | 1 | No room noise |

Instructions to listeners

The following set of instructions were given to all listeners as printouts. Note that the instructions were in Finnish, and they are here translated into English to aid the reader.

|  |
| --- |
| Listening instructions:  You will hear through stereo headphones pairs of binaural speech samples. Binaural means that you can locate various sound sources around yourself while listening with headphones. For example, a first talker may appear to talk from the left-hand side and a second talker from the right-hand side. This may also be called spatial audio. In traditional mono audio you cannot hear the direction of the talkers like in spatial audio. Instead, both talkers appear to talk from the same position inside your head.  The samples you are about to hear were recorded in real environments and may contain in addition to main talkers’ speech various ambient noises, music, and distant chatter by other people.  The first speech sample of each pair is the original. Right after the first sample you will hear the sample again. For the second sample there may have been used some future mobile phone technology. Your task is to evaluate the second speech sample compared to the first speech sample. Your task is to evaluate both the voice quality and the spatial representation of the second speech sample compared to the first speech sample. We can call this combination of voice quality and the spatial quality the Overall quality of the sample.  The Overall quality degradation of the second speech sample compared to the first speech sample is evaluated using the following scale:  5 Degradation is inaudible  4 Degradation is barely audible  3 Degradation is audible but not annoying  2 Degradation is slightly annoying  1 Degradation is annoying  ----------------------  Do not take refreshments with you to the booth (you can have refreshments during the breaks)  Leave your mobile phone on the table outside the listening booths  Do not discuss about the speech samples with other people during the comfort breaks |

Compared to standard P.800 instructions, the listeners are guided to consider the overall quality, including any degradation of the speech or other sound, and any change in the spatial presentation quality before casting their vote. For degradation scale, a more sensitive wording is used. Instead of “1 Degradation is very annoying” we use here “1 Degradation is annoying” for lowest quality and an additional step is inserted between original scores of 4 and 5. This score is “4 Degradation is barely audible”. This sensitivity adjustment of the scale can reduce the effect of quality saturation at the upper end of the voting scale when conditions are close to transparency. This modification also increases usage of the lowest score of 1, particularly in case of relatively high-quality samples thus providing additional separation between conditions.

In addition to the textual instructions, verbal instructions were given prior to listening to all listeners. Before the listening test, several introductory samples were played back covering the full range of degradations appearing in the actual test.

Example 2: Example P.800 DCR test of spatial (FOA) speech [10]

Introduction

Below is a P.800 DCR [8] test design example for subjective testing of spatial (FOA) speech quality. The example has been imported from Tdoc S4-210836 [11]. Results obtained from the test execution are not provided here but are available in the original documents [11] for Experiment 1 and [18] for Experiment 2.

Test Purpose

Build an opinion about suitability of modified P.800 DCR test methodology for quality assessments of immersive conversational speech.

Test Outline

* 2 Experiments
* Exp1: use case ‘immersive conferencing’ with Ambisonics (FOA) spatial speech, 6 content type categories constructed as follows:
* Model-based relying on convolution of raw mono clean speech sentences convolved with (FOA) Spatial Room Impulse Responses respective various talker positions relative to a capture point. The Spatial Room Impulse Responses were recorded in the respective conference rooms.
* Spatialized sentences are combined to sentence pairs and mixed with spatial (FOA) ambient noise.
* 2 relatively low background noise levels (30, 40 dB SNR, based on level normalization according to ITU-R BS.1770-4 [12])
* Reverberance typical for 2 conference rooms (large and small)
* 2 talker interactions types: sentence pairs with and without ‘overtalking’ (1s overtalk)
* Language: Polish
* Lab: Dolby Wroclaw (Poland)
* Exp2: Immersive telephony while on the move (outside) with Ambisonics (FOA) spatial speech, 6 content type categories constructed as follows:
* Model-based relying on convolution of raw mono clean speech sentences convolved with (FOA) Spatial Room Impulse Responses respective various talker positions relative to a capture point. The Spatial Room Impulse Responses were recorded in the respective test environments (car) or a low-echoic room approximating the other environments.
* Spatialized sentences are combined to sentence pairs and mixed with spatial (FOA) ambient noise.
* Moderate to high background noise levels (15, 20, 25dB SNR, based on level normalization according to ITU-R BS.1770-4 [12])
* Various environments: street, car, public indoor (shopping mall, subway station)
* No talker interactions (no ‘overtalking’): sentence pairs without ‘overtalking’ (1s gap)
* Language: American English
* Lab: Dolby San Francisco (USA)/remote (home environment)

General Consideration of Experiments

* Six categories of content types.
* 30 subjects, five listening panels (six subjects per panel), each panel with an independent randomization.
* Five samples per category (one for each listening panel).
* Randomizations constructed under “partially-balanced/randomized blocks” experimental design described in “Practical procedures for subjective testing”, [13].
* Every condition has 30 different samples passed through it (6 categories x 5 panels). Each of these are voted on by the 6 subjects in the panel, giving: (30 samples x 6 subjects/panel) = 180 (150) votes per condition.
* 30 test conditions x 6 categories = 180 DCR trials.
* Average trial duration: 16 s (6.5 s reference sample +0.5 s silence + 6.5 s test sample + 2.5 s voting period).
* Test duration: ~1.6 h per listening panel. Test duration comprises 50% of actual listening/voting time (48 min) and 50% test overhead including orientation, instructions, preliminaries, and rest breaks
* The listening sessions were split into a number of sub-sessions with breaks in between to allow for the subject to relax. This was to avoid listener fatigue.
* Test platform: Dolby-internal

Degradation references (anchors)

According to ITU-T Rec. P.811 Appendix II, P.811 [9] overall quality scores strongly correlate with P.800 DCR scores if the latter is run with modified instructions and degradation references that span both signal and spatial quality dimensions. P.811 suggests using P.50 MNRU for signal degradation anchors and SDRU/ESDRU for spatial degradation anchors. P.50 MNRU is a modulated noise reference unit with P.50-artificial voice weighting. SDRU/ESDRU are spatial degradation reference units defined for stereo signals that gradually, depending on a degradation parameter α, impair the stereo image without substantially causing signal distortions. A random process additionally introduces temporal fluctuations ranging from the original to the maximally degraded stereo image. The ESDRU applies a more sophisticated random process.

We followed this recommendation and adapted the P.50 MNRU and the ESDRU to derive degradation anchors for our P.800 experiments with binauralized FOA content.

For the P.50 MNRU the adaptation is that it is coherently applied (same seed) to all 4 FOA signals. This has the perceptual effect that the spatial direction of the introduced signal distortion coincides with the spatial signal direction. Thus, the introduced signal distortion does not significantly affect the spatial image.

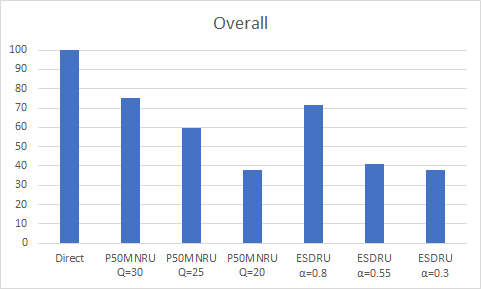
The ESDRU on the other hand is directly applied to the two binaural channels after binaural rendering of the FOA signal.

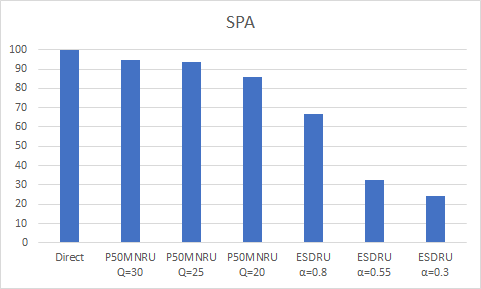
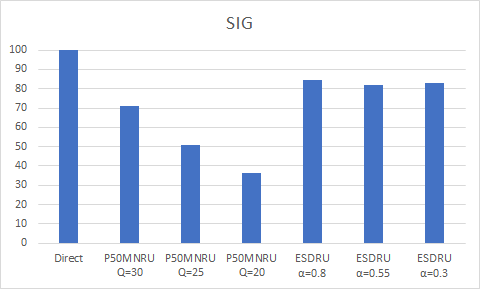
A limited subjective experiment was carried out to

* verify the suitability of these degradation anchors,
* to verify the basic assumption that the P.50 MNRU has little impact on spatial distortion and vice-versa that the ESDRU has little impact on perceived signal distortion, and
* to find suitable P.50 MNRU and ESDRU degradation parameters Q and, respectively, α.

In the experiment 6 FOA voice vectors were degraded either with P.50 MNRU values of Q=30, 25, and 20 dB or with ESDRU parameter values of α = 0.8, 0.55, and 0.3. These vectors were evaluated in a Mushra test (with 3 expert listeners) with the three quality attributes overall quality (Overall), signal quality (SIG), and spatial quality (SPA).

The results are displayed in the following plots:





From the plots, the following observations can be made:

* The P.50 MNRU degradation affects mainly signal (SIG) and Overall quality while spatial quality (SPA) is less impacted.
* The ESDRU degradation affects mainly spatial (SPA) and Overall quality while signal quality (SIG) is less impacted.
* The P.50 MNRU induced signal degradation appears a bit too strong and should be softened for the P.800 tests.
* The ESDRU induced degradation is too strong, which results in that spatial and overall quality start to saturate at the lower end. Consequently, for the P.800 tests, it was decided to increase the α parameters.

Factors and conditions

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 11 | Dolby-internal FOA coding system |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 12 | Multi-mono 4xEVS operated at  4\*8, 4\*9.6, 4\*13.2, 4\*16.4, 4\*24.4, 4\*32, 4\*48, 4\*64, 4\*96 kbps with DTX off and  4\*13.2, 4\*16.4, 4\*24.4 kbps with DTX on |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | Nominal input level |
| P.50 MNRU (applied to all FOA components) | 3 | Q=22, 27, 32 dB (all: nominal level) |
| ESDRU [9] | 3 | α = 0.55, 0.7, 0.85 (output loudness forced to nominal level) |
|  |  |  |
| **Common Conditions** |  |  |
| Test item generation: pre-processing incl. spatialization | 1 | Model-based relying on convolution of raw mono clean speech sentences convolved with (FOA) Spatial Room Impulse Responses respective various talker positions relative to a capture point and spatial (FOA) ambient noise mixing |
| Binaural renderer | 1 | FOA to binaural rendering according to [14] |
| Audio sampling frequency/bandwidth | 2 | 48 kHz/SWB except for 4xEVS@4\*8kbps which is 48 kHz/WB |
| Content types (categories) | 6 | Exp1: 6 Different conference rooms and talker interactions  Exp2: 6 Different background noise types and levels |
| Kind of samples | 1 | Sentence pair uttered by different talkers and genders (3 male and 3 female) |
| Number of samples | 5 | per content type |
| Input frequency mask | 1 | Flat |
| Nominal output loudness | 1 | -26 LKFS (ITU-R BS.1770-4 [12]) |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve Listeners |
| Randomizations | 5 | 5 panels of 6 listeners |
| Rating Scale | 1 | DCR with modified instructions |
|  |  |  |
| Languages | 1 | Exp1: Polish, Exp2: American English |
| Listening System | 1 | High-quality headphone for diotic presentation |
| Listening Environment | 1 | No room noise |

Preliminaries (familiarization of listeners)

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 0 |  |
| Codec references | 5 | Multi-mono 4xEVS operated at  4\*8, 4\*13.2, 4\*24.4, 4\*48, 4\*64, with DTX off |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | Nominal input level |
| P.50 MNRU (applied to all FOA components) | 3 | Q=22, 27, 32 dB (all: nominal level) |
| ESDRU  [9] | 3 | α = 0.55, 0.7, 0.85 (output loudness forced to nominal level) |
|  |  |  |
| **Common Conditions** |  |  |
| Test item generation: pre-processing incl. spatialization | 1 | Model-based relying on convolution of raw mono clean speech sentences convolved with (FOA) Spatial Room Impulse Responses respective various talker positions relative to a capture point and spatial (FOA) ambient noise mixing |
| Audio sampling frequency/bandwidth | 1 | 48 kHz/SWB except for 4xEVS@4\*8kbps which is 48 kHz/WB |
| Content types (categories) | 6 | Exp1: 6 Different conference rooms and talker interactions  Exp2: 6 Different background noise types and levels |
| Number of samples | 1 | per content type |
| Input frequency mask | 1 | Flat |
| Nominal output loudness | 1 | -26 LKFS (ITU-R BS.1770-4 [12]) |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve Listeners |
| Randomizations | 1 | Same randomization for the 5 panels of 6 listeners |
| Rating Scale | 1 | DCR with modified instructions |
|  |  |  |
| Languages | 1 | Exp1: Polish, Exp2: American English |
| Listening System | 1 | High-quality headphone for diotic presentation |
| Listening Environment | 1 | No room noise |

Instructions to listeners and Degradation Scale

The following presents the modified DCR test instructions given to the subjects and the five-point degradation category scale used in the test:

**"Evaluation of the quality of future 3D audio telephony and conferencing systems"**

In this experiment you will hear pairs of speech samples that have been recorded through various experimental 3D audio telephone and conferencing equipment. You will listen to these samples through a set of stereo headphones.

What you will hear is a first sample containing one pair of sentences from two talkers, a short period of silence, and a second sample. You will evaluate the OVERALL quality of the second sample compared to the quality of the first sample.

You should listen carefully to each pair of samples. As soon as a sample pair has been completely played back, you should register your opinion on ANY kind of degradation of the second sample compared to the first sample. Please consider in your vote, besides, e.g., the quality of the speech or other sounds, also any change in the perceived location of voices or sounds or changes in spatial width.

Then, when the system requests your vote, please record your opinion on the OVERALL quality using the following scale:

The OVERALL quality DEGRADATION of the Second Compared to the First is:

5: Inaudible

4: Audible but not annoying

3: Slightly annoying

2: Annoying

1: Very annoying

You will have five seconds to record your answer by pushing the button corresponding to your choice. There will be a short pause before the presentation of next pair of sentences.

We will begin with a short practice session to familiarize you with the test procedure. The actual tests will take place during multiple sessions with short breaks in between.

**Degradation Scale**

The OVERALL quality DEGRADATION of the Second Compared to the First is:

5: Inaudible

4: Audible but not annoying

3: Slightly annoying

2: Annoying

1: Very annoying

Example 3: Experience of P.800 for stereo testing [15]

Test description

As a part of Ericsson’s involvement in the development of P.811 [9] standard, a listening test according to the draft P.811 specification was done in a collaboration between Ericsson and Beijing Institute of Technology (BIT). The test was conducted in October 2018 and was done in conjunction with a P.800 [8] DCR test on the same test material. The purpose was to evaluate the proposed P.811 standard (called P.SOSH at the time) and to compare the overall score of the P.811 test with a P.800 DCR test which requires shorter test time. These tests were performed:

* Experiment 1: P.800 Degradation category rating (DCR) with spatial distortion reference units and listener instructions similar to the P.811 instructions, see appendix A
* Experiment 2: Subjective test methodology for evaluating speech oriented stereo communication systems over headphones (P.811)

The test design and processing were carried out by Ericsson, while BIT handled recording of the test material and execution of the test itself.

**Test material**

The test was conducted using stereo speech samples in Mandarin Chinese recorded at BIT. The talkers were 4 female and 4 male talkers recruited from the BIT students. The talkers were all native Mandarin Chinese speakers and were selected to have a rather neutral dialect. The stereo capture was done using a Sabinetek® SMIC Panoramic Microphone and the recordings were made using 48 kHz sampling rate.

Out of the 20 test items in total, 10 items contained one talker with a split of 5 female and 5 male talkers. The remaining 10 items contained two concatenated talkers at different positions, where each item contained one male and one female talker. The concatenation of the talkers was done with a short pause between each talker, i.e. no overlapping talk. The talkers were positioned at the angles of -90, -45, 0, 45 and 90 degrees relative to the front pickup of the microphone.

**Listener subjects**

Each of the experiments was performed with 32 naïve listeners (balanced between male and female). All of them were BIT adult students between 20-24 years old. In total, 64 different native listeners of Chinese were selected as test subjects.

The listeners were selected randomly from native Chinese persons in the BIT campus. After the pre-tests, the staff checked the subjects' scores to make sure they understood the rating criterion. If the listener gave inconsistent or confusing votes, they were asked to do the pre-test session again. If the inconsistencies were not resolved in the second pre-test session, the listener was excluded from the main test session.

**Experiments Procedure**

For both the P.800 and P.811 tests, the subjects were divided into 4 listening panels of 8 persons each. Each panel used its own randomization sequence files.

Preliminary tests (pre-tests) were held before the main tests. In the pre-test, 4 trials were run to make the subjects familiar with test methodology. The main test was divided into 4 sessions of 20 trials each. A break was inserted between each test session, of 5, 10 and 5 minutes respectively.

The processed speech material was presented to groups of listeners, who were seated in separate listening stations in an acoustically conditioned sound room meeting the requirements recommended in ITU-T P.800. A photo of the test room is shown in Figure 1.



Figure 1: Listening laboratory

All test stimuli were presented to the subjects using Sennheiser® HD 280 Pro headphones. Tablets were used to collect votes during the two experiments. The voting table interfaces are shown in Figure 2 and Figure 3.

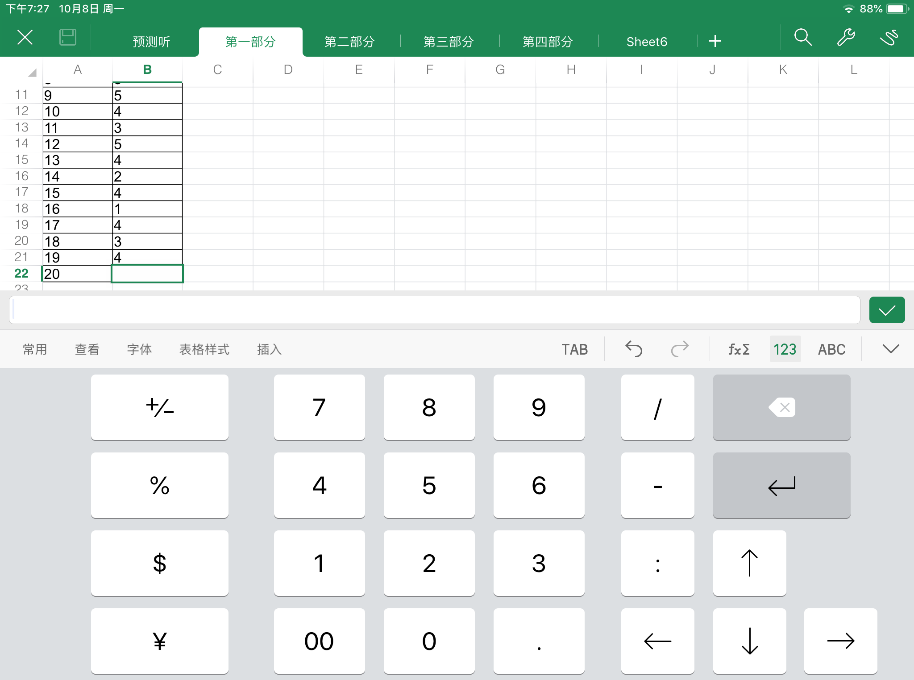
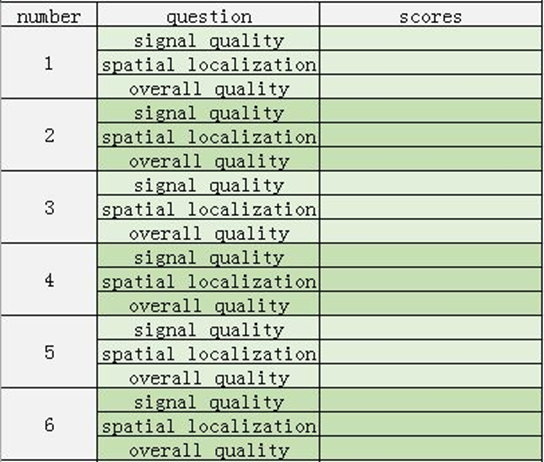


Figure 2: Voting interface on a tablet with spreadsheet for collecting votes in the P.800 test.



**Figure 3**: Spreadsheet for collecting votes in the P.811 test. The three rows for votes of a specific test file were marked with the same color to minimize the risk of confusion.

The voting time was 5 seconds after the completed presentation of each new stimulus. All seated listeners were required to vote prior to the subsequent presentation of a new stimulus. Comments, experiences and suggestions from listeners were collected at the end of each experiment.

The average test time per session was 18 minutes for the P.811 test and 6 minutes for the P.800 test.

**Scoring**

Both experiments used the Degradation Category Rating (DCR) method where the reference is played first followed by a test sample to be judged in comparison to the reference.

In the P.800 DCR test, listeners gave their opinion on any degradation in Overall Quality they could perceive on the second sample compared to the first one (the reference). The instructions for the P.800 test with P.811 inspired instructions can be found in appendix A.

In the P.811 test, listeners gave their opinion of any signal degradation, difference in spatial localization and overall quality degradation they could perceive on the second sample compared to the reference, according to the instruction below:

**Signal (SIG) degradation**

Attending ONLY to the SIGNAL (SPEECH and BACKGROUND NOISE or MUSIC), select the category that best describes the DEGRADATION in the second sample compared to the first sample.

Signal degradation in this sample was,

5 INAUDIBLE

4 AUDIBLE BUT NOT ANNOYING

3 SLIGHTLY ANNOYING

2 ANNOYING

1 VERY ANNOYING

**Spatial localization (SPA)**

Attending ONLY to the TALKER/SOURCE LOCATIONS, select the category that best describes the DIFFERENCE in the second sample compared to the first sample.

There was

5 NO DIFFERENCE

4 SMALL DIFFERENCE

3 MODERATE DIFFERENCE

2 LARGE DIFFERENCE

1 VERY LARGE DIFFERENCE

**Overall (OVRL) quality degradation**

Attending to the OVERALL impression, including but not limited to signal quality and spatial localization, select the category that best describes the OVERALL Quality degradation of the sample compared to the reference.

Overall quality degradation was,

5 INAUDIBLE

4 AUDIBLE BUT NOT ANNOYING

3 SLIGHTLY ANNOYING

2 ANNOYING

1 VERY ANNOYING

**Anchors used in the test**

To span the signal degradation dimension, MNRU anchors at Q-levels 16, 23 and 30 were used. The Direct signal and Direct-Downmix to mono were also used in the test. In addition, there were two versions of spatial anchors, SDRU and ESDRU.

**SDRU and ESDRU**

The effect of the SDRU can be summarized as:

* a down-mix (collapse) of the stereo image for and a full reversal of the channels for .
* an amplitude modulation (panning) of the signal with a triangle wave with a period of 1 second.

The second dimension of this distortion reference unit creates a “ping-pong” effect between the channels which was regarded a bit unnatural in relation to the typical distortions introduced by stereo codecs. In addition, some listeners reported the effect induced dizziness. While dizziness may be an unavoidable side-effect of spatial distortion, it was found relevant to try a different variant of the modulation function. The formulation of the ESDRU, an alternative spatial distortion reference unit, is the same as the SDRU apart from the definition of the modulation function. Instead of a periodic triangle wave, a random stepwise pattern was introduced. The idea behind this was that the random deviation would be more similar to a stereo codec which may introduce quantization errors on a parametric description of the stereo image. It would also avoid the periodic panning which may give the illusion that the listener’s head is spinning.

**Test conditions**

The input speech items were processed for the 20 conditions listed in Table 7 below. The same test material was used in both the P.800 DCR test and the P.811 test. The processing bandwidth in the test was Super Wideband (SWB) sampled at 32 kHz. The SDRU in conditions c06 - c08 operate on 48 kHz, which means a sampling rate change was necessary. All sampling rate changes were implemented using the ITU-T STL filter tool with SHQ2 and SHQ3 resampling filters and delay compensation as described in Table 6 of [16].

Table 8: Processed conditions

|  |  |
| --- | --- |
| **Label** | **Condition** |
| c01 | DIRECT |
| c02 | DIRECT downmix (L+R)/2 |
| c03 | MNRU Q=16 |
| c04 | MNRU Q=23 |
| c05 | MNRU Q=30 |
| c06 | SDRU 0.0 |
| c07 | SDRU 0.3 |
| c08 | SDRU 0.6 |
| c09 | ESDRU 0.0 |
| c10 | ESDRU 0.3 |
| c11 | ESDRU 0.6 |
| c12-c20 | Stereo codec conditions |

**Preprocessing**

The stereo signals were split using

* stereoop.exe -split <input> <outputL> <outputR>

Each channel was then high-pass filtered using filter, followed by a delay compensation of 839 samples

* filter.exe HP50\_48KHZ <input> <output> 960

The sampling rate was then changed from 48 kHz to 32 kHz and the level was normalized to   
-26 dBov using the following procedure:

* stereoop -maxenval <input> maxenval32
* sv56demo -log log.txt -lev -26 -sf 32000 maxenval32 dummy 640
* scale=`cat log.txt | grep "Norm factor" | awk '{print $6}'`
* scaldemo -gain $scale <input> <output>

**DIRECT**

Preprocessed input signal without further modification.

**DIRECT downmix (L+R)/2**The passive downmix realized as , using the tool CopyAudio [17]:

* CopyAudio.exe --chanA="0.5\*A+0.5\*B" -P integer16,,32000,,2 -F noheader <stereo> <output>

**MNRU**The MNRU conditions were generated using the SDRU tool [9], where the modulated noise generators are synchronized between left and right channels:

* BG\_MNR07.exe <input> <output> <Q-value> H 1

**SDRU**The SDRU conditions were generated using SDRU tool [9]:

* BG\_MNR07.exe <input> <output> 100 H <alpha-value>

**ESDRU**ESDRU conditions generated using the ESDRU tool [9]. The random seed may be set to get deterministic results for each processing run:

* matlab /minimize /nosplash /nodesktop /r "esdru('<input>', '<output>', 32000, <alpha-value>, 0.5, <random seed>);exit"

**Post-processing level normalization**

While the stereo coding normally preserves the level of the signal, the signal levels of SDRU, ESDRU and the DIRECT downmix often deviates from the input level. For this reason, the level was normalized for the SDRU and ESDRU conditions following the same normalization procedure as in the preprocessing:

* stereoop -maxenval <input> maxenval32
* sv56demo -log log.txt -lev -26 -sf 32000 maxenval32 dummy 640
* scale=`cat log.txt | grep "Norm factor" | awk '{print $6}'`
* scaldemo -gain $scale <input> <output>

The DIRECT downmix condition results in a dual mono representation, which tends to get a too high level with the described procedure. For this condition a separate normalization procedure was used. The procedure matches the energy of the down-mix signal with half the energy of left and right channels combined.

* sv56demo -rms -sf 32000 -blk 1280 -log tmp.log <stereo input> dummy.raw
* A=`cat tmp.log | grep "Norm factor" | gawk '{print $6}'`
* sv56demo -rms -sf 32000 -blk 640 -log tmp.log <downmix input> dummy.raw
* B=`cat tmp.log | grep "Norm factor" | gawk '{print $6}'`
* fac=`echo "$B/$A" | bc -l`
* scaldemo -gain $fac <downmix input> <downmix output>

Test results

The results of the listening tests are illustrated in Figure 4 and Figure 5 below. As seen in Figure 4, the signal distortion induced by the MNRU has the main impact on the SIG dimension (a) while keeping a fairly constant rating in the SPA dimension (b). Conversely, the spatial distortion of the SDRU and ESDRU has a strong effect on the SPA dimension (b) while it the showing less impact on the SIG dimension (b).



(b)

(a)

Figure 4: The scores of the signal degradation (a) and spatial localization (b) of the P.811 test.



(b)

(a)

Figure 5: The scores of the overall dimension (a) of the P.811 test and the P.800 DCR scores (b).

Turning to Figure 5, the overall scores of the P.811 test in the OVRL dimension (a) show a high degree of similarity with the P.800 DCR scores (b). The correlation coefficient between these scores is 0.966. As a comparison, the correlation between the scores of the two listening labs for each experiment in the EVS selection SWB conditions tests [23] are shown in Table 9. Here the two labs used the same test configuration and processing scripts but carried out their tests in different labs and in different languages.

Table 10: Correlation between scores from lab (a) and lab (b) in SWB experiments of the EVS selection tests.

|  |  |
| --- | --- |
| **Experiment** | **Corrcoef** |
| **s1** | 0.985 |
| **s2** | 0.972 |
| **s3** | 0.956 |
| **s4** | 0.960 |
| **s5** | 0.959 |
| **s6** | 0.888 |
| **s7** | 0.977 |

The relations between the scores may also be illustrated in the form of scatter plots. The relation between the SIG and SPA dimensions is shown in Figure 6. The scores of the MNRUs remains fairly stable for varying SIG scores, while the SDRU and ESDRU show a robustness in the SPA dimension.



Figure 6: The scores of the SIG dimension on the x-axis versus the scores of the SPA dimension on the y-axis.

The relation between the OVRL dimension and the SIG and SPA dimension is illustrated in Figure 7 (a) and (b) respectively.



(a)

(b)

Figure 7: Scores in the OVRL dimension (y-axis) compared to the scores of the SIG dimension (a) and the SPA dimension (b).

The relation between the P.811 overall score and the P.800 DCR scores is illustrated in Figure 8, indicating that the scores are highly correlated.



Figure 8: Scores of the P.811 OVRL dimension (x-axis) versus the scores of the P.800 DCR test (y-axis).

Focusing on the scores of the anchor conditions in Figure 9, one can see that the MNRU remains stable in the SPA dimension while declining in the SIG dimension (a). Conversely, the SDRU and ESDRU are stably in the SIG dimension while declining in the SPA dimension for increasing levels of distortion.



Figure 9: P.811 results for the MNRU (a), SDRU (b) and ESDRU (c).

**Additional small test about dizziness**

The test participants were encouraged to write comments after the tests about how they perceived the test methodology and the test material. These comments revealed that 8 out of the 64 test participants in the P.800 DCR and P.811 tests felt somewhat dizzy or uncomfortable during the test when the voices changed position between left and right channel. This behavior can be found for the spatial anchors. To examine if the SDRU and ESDRU anchors were perceived differently the test participant that had commented that they felt dizzy were invited to an extra test with only the SDRU and ESDRU conditions.

The 8 students were divided into two groups, A and B with 4 persons in each group. All 20 speech files used in the P.800 and P.811 tests were also used in this test. Group A listened to sentence pairs 1-10 and Group B listened to speech examples 11-20. Each group listened to 5 samples with one speaker and 5 samples with two speakers.

The processed samples were presented after the reference samples as in the main tests, but in this test the test subjects should quantify how dizzy they felt while listening to the test samples according to this scale:

5 Not dizzy.

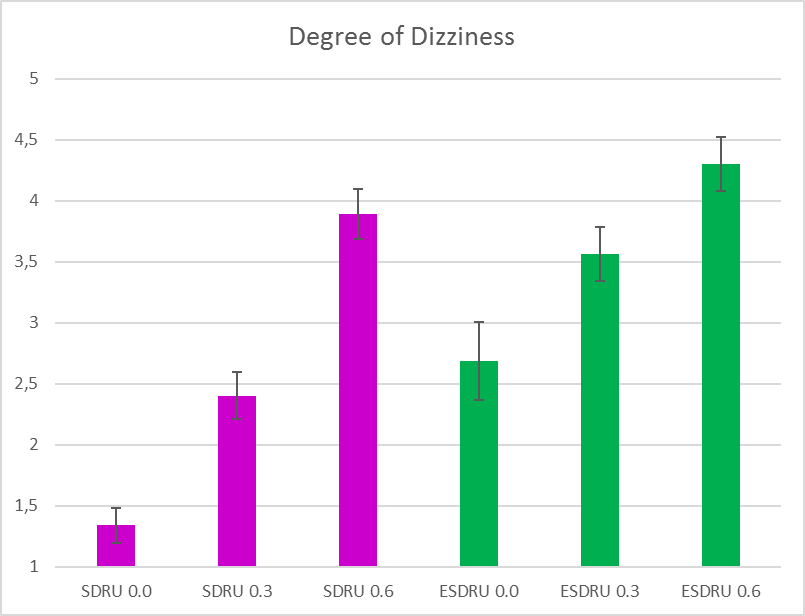
4 The degree of dizziness is very small

3 The degree of dizziness is moderate

2 The degree of dizziness is large

1 The degree of dizziness is very large

The results in Figure 10 reveal that the ESDRUs made the test subjects less dizzy than the SDRUs. This test was done only with persons that had reported that they got dizzy during the main tests. Most persons will however probably not get dizzy during a test as only 8 persons out of the 64 test participants commented that they became dizzy during the main tests.



4.5

2.5

1.5

3.5

Figure 10: Mean scores for the degree of dizziness in the test with only the spatial anchor conditions. In this additional test, higher scores indicated lower degree of dizziness. The confidence intervals (95%) are indicated using black lines.

**Comments and suggestions**

The main comment from test participants regarding the P.811 test methodology was that it was boring with so many repetitions. They suggested that the test material should be more enriched and varying. Some commented that the test was too long and monotonous and that they felt tired and thought it was hard to focus at the end of the test.

As this can be a problem there should be a careful selection of conditions to keep the test as short as possible.

A suggestion from two listeners was that it would be enough to listen to each speech sample two times instead of three times. After hearing the speech sample, the first time they could judge the signal degradation and after hearing the sample the second time they could vote for both the spatial and the overall quality. This suggestion would of course shorten the total test time but might lead to less focus on each of the spatial and overall dimensions and possibly less accurate results.

Another suggestion was that it would be better to use a more automatic collection of the votes as that would certify that the vote is connected to the correct speech example. Then it would also be possible to hide the previous votes, so that the judgments are not so easily influenced by previous votes.

Conclusions and proposal

The listening experiments shows that the P.811 method does give a relevant rating in the different dimensions specified by the test, but the prolonged test time from asking three questions may result in listener fatigue and puts limitations on the test size (e.g., number of conditions). Further, the results for the overall quality in the P.811 and P.800 DCR tests were highly correlated which indicates that P.800 DCR with adapted instructions and spatial anchors is an attractive alternative to P.811 for tests where the main interest is the overall score.

The spatial anchor ESDRU received similar quality ratings as the SDRU while inducing less dizziness. Hence, the ESDRU is considered a good alternative to the SDRU.

Example 4: DCR test experiments for FOA and HOA3 input in 7.0+4 and binaural listening setup [19]

Test Purpose

The purpose of the experiments was to evaluate suitability of P.800 DCR test [8] for immersive listening using naïve listeners, and to compare test results of 7.0+4 loudspeaker-rendered listening with test results of binaurally rendered listening via headphones.

Audio database

* Artificially created spatial samples from phonetically balanced mono recordings adjusted to -26 dBOvl.
* Language: North American French
* Two mono recordings with similar meaning were combined in HOA3 domain to create spatially separated sentence pairs.
* 4 male and 4 female talkers, always a male and a female talker in a sentence pair.
* Sentence pairs simulating a conversation with natural transition from one talker to another. Half of the samples partially overlapped.
* Length of the samples - 6 s.
* 48 kHz sampling rate.
* HOA3 and FOA input format.
* All talkers were placed at the nominal height at different configurations using regular pattern using:
  + 3 different speaker separations: 60, 90, 135
  + 24 different combinations:

|  |  |
| --- | --- |
| Separation [°] | 1st talker position [°] |
| 60 | -15 : 45: 300 |
| -90 | 30 : 45 : 345 |
| 135 | -15 : 45 : 300 |

* Background
  + Mono recordings of instrumental music at 15 dB SNR.
  + Different music sample and position used for each speech sentence pair.
  + The mono samples were encoded into HOA3 domain using elevation of 20°, 40°, and 60°. Exact azimuths and elevations were distributed as follows:

Azimuth = [15, 60, 115, 155, -155, -115, -60, -15, 15, 60, 115, 155, -155, -115, -60, -15, 15, 60, 115, 155, -155, -115, -60, -15]

Elevation = [20, 20, 20, 20, 20, 20, 20, 20, 40, 40, 40, 40, 40, 40, 40, 40, 60, 60, 60, 60, 60, 60, 60, 60]

Test Setup

* P.800 DCR test, instructions mentioning spatial aspect.
* 4 categories, each category corresponding to the different talker pair.
* 6 panels, each using different audio samples and randomizations.
* Naïve listeners.
* 4 listeners per panel
  + One listener at a time in the loudspeaker setup.
  + Four listeners at a time in the binaural setup.
* 24 listeners in total.
* 29 conditions. The randomization was done using a flexible algorithm, selecting conditions randomly with some constraints to balance distribution of conditions within each panel.
* Each condition was evaluated 24 x 4 = 96 times.
* Anchors - P.50 MNRUs (Modulated Noise Reference Units) [9] – applied coherently (using the same seed) to all ambisonic channels. 4 MNRU levels were used – 34, 29, 24, and 19 dB.
* No SDRUs (Spatial Distortion Reference Unit) or ESDRUs spatial anchors were used as they are not defined for loudspeaker listening.
* CuT – multi-mono EVS applied on FOA and HOA3 channels.
* Rendering
  + All conditions rendered to 7.0+4 loudspeaker system or to binaural representation using *All-Round Ambisonic Decoding* (AllRAD) [20].
  + Rendering was done on concatenated files.
* Level adjustment
  + The level was adjusted to -26 LKFS.
  + The direct signal level was first measured on the signal rendered to 7.0+4 loudspeaker system using B.1770 [12] and level difference was computed with -26 LKFS (Loudness, K-weighted, relative to Full Scale). The corresponding gain was then applied to the original HOA3 input channels. No level readjustment was done on the coded signals.
* Listening laboratory - Immersive listening laboratory at the University of Sherbrooke.
* Loudspeaker listening setup - 7.0+4 Genelec SAM 3031 speaker setup in the following configuration:

|  |  |  |
| --- | --- | --- |
| Speakers | Azimuth | Elevation |
| Left front | 30 | 0 |
| Right front | -30 | 0 |
| Centre front | 0 | 0 |
| LFE | - | - |
| Left rear surround | 135 | 0 |
| Right rear surround | -135 | 0 |
| Left side surround | 90 | 0 |
| Right side surround | -90 | 0 |
| Left front height | 30 | 35 |
| Right front height | -30 | 35 |
| Left rear surround height | 135 | 35 |
| Right rear surround height | -135 | 35 |

* Binaural listening setup used Beyer Dynamic DT 770 Pro headphones.

Screening of listeners

Listeners were post-screened as follows. In order to be considered, a listener had:

* To use the whole voting scale during the session. In other words, he must have voted at least once “1” and at least once “5”.
* To vote, in average, the direct condition better than or equal to the MNRU 29 dB condition. To reflect the fact that the perceptual quality of MNRU 29 dB is close to Direct, the listener was still kept if the median of his votes for all anchor conditions was below 4.
* To vote, in average, the MNRU 29 dB condition better than the MNRU 24 dB condition.
* To vote, in average, the MNRU 24 dB condition better than the MNRU 19 dB condition.
* To vote, in average, the MNRU 19 dB condition better than the MNRU 14 dB condition.

Comments

* The tests took about 2 weeks.
* Overall, naïve listeners could reliably detect coding deficiencies.
* When coding ambisonic channels with EVS at low bitrates (below 24.4 kb/s), more ambisonic channels seem to degrade the perceptual experience rather than improve it.
* Naïve listeners do not seem to be too sensitive to the spatial aspect, e.g., differentiating between FOA and HOA3. Nevertheless, they were still able to discriminate the direct HOA3 from FOA with statistical significance in both tests.
* Despite clear and explicit instructions, and standard DCR voting labels used in the listening software interface, some listeners still did not understand the task.

Comparisons of results between the loudspeaker rendering and binaural rendering

* Good correlation between the binaural listening test results and the loudspeaker listening test results.
* In binaural listening, the listeners were able to distinguish HOA3 and FOA direct conditions better than in the loudspeaker listening.
* Larger dynamics of results are observed for binaural listening than for loudspeaker listening.
* Overall, the multi-mono EVS processing conditions were voted noticeably lower in binaural listening than in the loudspeaker listening.
* For multi-mono EVS processing, at 24.4 kbps/channel, an advantage for FOA over HOA3 input is observed in binaural listening, but the opposite tendency is observed for loudspeaker listening

Conclusions

* With some adjustments, the DCR test with naïve listeners seems to be a good trade-off between accuracy and efficiency.
* EVS multi-mono seems to be a good reference, able to cover practically the whole range of perceptual quality.
* More explicit initiation of naïve listeners to spatial aspects would be beneficial, e.g., an extended training session at the very least. Also, some discussion on listeners’ perception after the training session might help.
* Agreed methodology for systematic post-screening of listeners would be useful.