**Source: Editor**[[1]](#footnote-2)\*

**Title: IVAS Permanent Document IVAS-8a: Test Plan for Selection Phase**

**Version: v.0.4.0**

**Agenda Item: 7.5**

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 |

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 | EVS 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.2051-1 (06/2017): 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

* 1. Key Acronyms

BIT Beijing Institute of Technology

CL Cross-check Laboratory

CuT Codec under Test

DCR Degradation Category Rating

ESDRU E Spatial Distortion Reference Unit

EVS Enhanced Voice Services

FB Full Band

FoA First-order Ambisonics

GAL Global Analysis Laboratory

HL Host Laboratory

IVAS Immersive Voice and Audio Services

LKFS Loudness, K-weighted, relative to Full Scale

LL Listening Laboratory

MASA Metadata-Assisted Spatial Audio

MNRU Modulated Noise Reference Unit

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 and ETSI will contract the HL, the CL, and the GAL to perform respective tasks defined in this document.

[The LLs shall provide unprocessed 48 kHz sampled raw speech, mixed content and music samples to the HL.]

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, FORCE Technology [1], HEAD acoustics GmbH, IKS [2]]

HL: [HEAD acoustics GmbH, IKS]

CL:

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).

Editor’s note: Possibly integrating Annexes with laboratory tasks here if there is no particular reason for keeping it in annexes

* + 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
* Interaction with the HL to cross-check the HL’s implementation of its CuT executable
	+ 1. Listening Laboratories
		2. Host Laboratory
		3. Cross-check Laboratory
		4. Global Analysis Laboratory
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
	2. Opinion Scales
	3. 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 interleaved [headerless] format, 16-bit PCM in little endian files following the naming convention provided in the IVAS Processing Plan (IVAS-7a).

[The following three categories of audio content will be used in IVAS Selection Test:

* Clean speech: each sample contains two (or more) different talkers in conversation scenario. The talkers transition from one to another as in natural conversation, i.e. without a pause, possibly with partial overlap.
* Speech with background: the background comprises e.g. background music, noise, interfering talkers.
* Music and Mixed content]

Editor’s note: What each category comprises is for further discussion

* + 1. Speech Material
		2. Noise Material
		3. Captured Music and Mixed Content Material
	1. Listening Systems and Listening Environments

The IVAS Selection Test will use the following listening systems:

* Stereo headphones, both for static binaural listening and binaural listening with simulated head-tracking (scene rotation is predefined)
* Loudspeaker listening system - 7.1+4 loudspeaker setup [1].
	1. Experimental Procedure
	2. 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

[Table 1: High-level overview of Selection experiments

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Exp | Input Format | Source Material | Output Format | Bitrates | FER/Jitter | Headtracking | Methodology | Nb of References | Nb of test Conditions |
| 1 | Stereo | Clean speech | Stereo |  | ≤ x%  |  |  |  |  |
| 2 | Stereo | Speech+Background | Stereo |  | ≤ x% |  |  |  |  |
| 3 | Stereo | Mixed & Music | Stereo |  | ≤ x% |  |  |  |  |
| 4 | 5.1 | Clean speech |  |  | ≤ x% |  |  |  |  |
| 5 | 5.1 | Speech+Background |  |  | ≤ x% |  |  |  |  |
| 6 | 5.1 | Mixed & Music |  |  | ≤ x% |  |  |  |  |
| 7 | 7.1.4 | Clean speech |  |  | ≤ x% |  |  |  |  |
| 8 | 7.1.4 | Speech+Background |  |  | ≤ x% |  |  |  |  |
| 9 | 7.1.4 | Mixed & Music |  |  | ≤ x% |  |  |  |  |
| 10 | FOA | Clean speech |  |  | ≤ x% |  |  |  |  |
| 11 | FOA | Speech+Background |  |  | ≤ x% |  |  |  |  |
| 12 | FOA | Mixed & Music |  |  | ≤ x% |  |  |  |  |
| 13 | HOA3 | Clean speech |  |  | ≤ x% |  |  |  |  |
| 14 | HOA3 | Speech+Background |  |  | ≤ x% |  |  |  |  |
| 15 | HOA3 | Mixed & Music |  |  | ≤ x% |  |  |  |  |
| 16 | Objects | Clean speech |  |  | ≤ x% |  |  |  |  |
| 17 | Objects | Speech+Background |  |  | ≤ x% |  |  |  |  |
| 18 | Objects | Mixed & Music |  |  | ≤ x% |  |  |  |  |
| 19 | MASA | Clean speech |  |  | ≤ x% |  |  |  |  |
| 20 | MASA | Speech+Background |  |  | ≤ x% |  |  |  |  |
| 21 | MASA | Mixed & Music |  |  | ≤ x% |  |  |  |  |

Notes:

* Currently considered methologies are P.800 DCR and MUSHRA
* Bitrates will be specified later, depending on the methodologies etc.
* Stereo may include binauralized samples (without head tracking)
* Stereo and Binaural Output is listened via headphones
* For inputs 7.1+4, FOA, HOA3, Objects & MASA vertical dimension is assumed in the samples
* If listening is done with headphones, headtracking might be used, and is assumed simulated
* Maximum Frame Error Rate (FER) *x*% depends on whether channel error conditions are mixed with clean channel conditions in the same experiment (as assumed in the above table), or whether separate experiments are designed specifically for testing channel errors. In the former case, *x* should not be too high to prevent compressing results for clean channel conditions, e.g. *x*=3.
* DTX on/off is assumed within the same experiment, where DTX on is used for relevant conditions.
* ESDRUs are assumed only for listening via headphones as currently not defined for loudspeaker listening
* FOA can possibly be merged with HOA, more information needed whether this is appropriate
* Tandem is not tested in Selection
* Rate switching is not tested in Selection
* Operating points not tested in Selection will be addressed in the Characterization test]

Editor’s note: The table is just a template, basis for further discussion.

1. Sample Instructions to Subjects and Data Collection
2. Presentation Orders
3. Data to be Provided by LL
4. Obligations and Task for the Listening Laboratories
5. Host Laboratory Tasks
	1. Included tasks
	2. Excluded tasks
6. Cross check Laboratory Tasks
	1. Included tasks
	2. Excluded tasks
7. GAL Tasks
	1. Tasks
	2. Statistical analysis of results
8. Selection Testing Timeline

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 inaudible4 Degradation is barely audible3 Degradation is audible but not annoying2 Degradation is slightly annoying1 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 boothsDo 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 for Experiment 1 in the original document [11].

Editor’s note: A Tdoc reference for the results of Experiment 2 will be provided in future when the experiment is completed and the results are available.

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

1. verify the suitability of these degradation anchors,
2. 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
3. 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 and4\*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 [4]  | 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 interactionsExp2: 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 |
| Replications | 1 |  |
| 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 at4\*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 [4]  | 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 interactionsExp2: 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 |
| Replications | 1 |  |
| 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:

1. a down-mix (collapse) of the stereo image for $α≈0.5$ and a full reversal of the channels for $α≈1$.
2. 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 1 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 1: 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 $\frac{L+R}{2}$, 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 [17] are shown in Table 2. 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 2: 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.

1. \* Milan Jelinek, VoiceAge Corporation; Milan.Jelinek@USherbrooke.ca [↑](#footnote-ref-2)