ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Supplement

ITU-T P Suppl. 29 (01/2023)

SERIES P: Telephone transmission quality, telephone installations, local line networks

ITU-T P.800 – Use cases

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	P.10–P.19
Voice terminal characteristics	P.30–P.39
Reference systems	P.40–P.49
Objective measuring apparatus	P.50–P.59
Objective electro-acoustical measurements	P.60–P.69
Measurements related to speech loudness	P.70–P.79
Methods for objective and subjective assessment of speech quality	P.80–P.89
Voice terminal characteristics	P.300–P.399
Objective measuring apparatus	P.500–P.599
Measurements related to speech loudness	P.700–P.709
Methods for objective and subjective assessment of speech and video quality	P.800–P.899
Audiovisual quality in multimedia services	P.900–P.999
Transmission performance and QoS aspects of IP end-points	P.1000–P.1099
Communications involving vehicles	P.1100–P.1199
Models and tools for quality assessment of streamed media	P.1200–P.1299
Telemeeting assessment	P.1300–P.1399
Statistical analysis, evaluation and reporting guidelines of quality measurements	P.1400–P.1499
Methods for objective and subjective assessment of quality of services other than speech and video	P.1500–P.1599

For further details, please refer to the list of ITU-T Recommendations.

Supplement 29 to ITU-T P-series Recommendations

ITU-T P.800 – Use cases

Summary

Supplement 29 to ITU-T P-series Recommendations describes ITU-T P.800 use case examples that include narrowband (NB), wideband (WB), super wideband (SWB) and fullband (FB) audio bandwidth, speech, music and mixed speech and music content, stereo and spatial quality evaluations.

This Supplement also provides guidance for using subjective listening methodology in Recommendation ITU-T P.800 for stereo and spatial speech and general audio. Anchor conditions, level normalization for stereo and multichannel signals, and listener screening methods are presented.

History

Edition	Recommendation	Approval	Study Group	Unique ID[*]
1.0	ITU-T P Suppl. 29	2023-01-26	12	11.1002/1000/15459

Keywords

Absolute category rating, degradation category rating, listening test, spatial localization, subjective evaluation, subjective testing.

[*] To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this publication may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the publication development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Page

1 Scope 1

2 References 1

3 Definitions 2

3.1 Terms defined elsewhere 2

3.2 Terms defined in this Supplement 2

4 Abbreviations and acronyms 2

5 Conventions 3

6 ITU-T P.800 use case examples 3

6.1 ITU-T P.800 use case examples from 3GPP EVS codec standardization 4

6.2 ITU-T P.800 DCR for stereo from ITU-T P.811 validation 7

6.3 Full-scale SWB experiments during ITU-T P.863 standardization 7

6.4 ITU-T P.800 DCR evaluation of parametric spatial speech 7

6.5 ITU-T P.800 evaluations of ambisonics spatial speech quality 7

7 ITU-T P.800 best practices especially related to stereo and spatial speech quality evaluation 8

7.1 General best practices 8

7.2 Best practices from specific experiments 13

8 ITU-T P.800 reference conditions and tools used in ITU-T P.800 use case examples 13

8.1 ITU-T P.50 fullband MNRU 14

8.2 SDRU 15

8.3 ESDRU 15

8.4 Multi-channel level normalization with BS1770DEMO 16

Appendix I – Collection of ITU-T P.800 test realizations 17

I.1 ITU-T P.800 evaluations of first order ambisonics (FOA) spatial speech quality with binaural headphone rendering 17

I.2 ITU-T P.800 DCR evaluation of parametric spatial speech 25

I.3 ITU-T P.800 DCR with higher-order scene-based audio 29

I.4 Comparison of DCR test experiments for FOA and HOA3 input in 7.0+4 and binaural listening set-ups 34

Supplement 29 to ITU-T P-series Recommendations

ITU-T P.800 – Use cases

1 Scope

This Supplement provides examples of experiments related to relevant standardization efforts in the ITU-T and 3rd Generation Partnership Project (3GPP) for which [ITU-T P.800] has been used, including quality assessments in the context of the standardization of enhanced voice services and immersive voice and audio services. The information provided by this Supplement on how these experiments were conducted is intended as guidance for ITU-T P.800 users when planning ITU-T P.800 quality assessments.

This Supplement also contains a collection of best practices for test design, preparation and execution.

In addition, specific reference conditions and tools used in the ITU-T P.800 use case examples are described in detail.

It should be noted that the use of [ITU-T P.800] is not limited to the examples, guidelines or reference conditions and tools described in this Supplement. Also note that there is no universal test method that can be applied under all circumstances. When designing an experiment, careful consideration is advised on the choice of a proper test method.

2 References

[ITU-T G.191] Recommendation ITU-T G.191 (2023), *Software tools for speech and audio coding standardization*, (see also https://github.com/openitu/STL).

[ITU-T P.78] Recommendation ITU-T P.78 (1996), *Subjective testing method for determination of loudness ratings in accordance with Recommendation P.76.*

[ITU-T P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality.*

[ITU-T P.810] Recommendation ITU-T P.810 (2023), *Modulated Noise Reference Unit (MNRU).*

[ITU-T P.811] Recommendation ITU-T P.811 (2019), *Subjective test methodology for evaluating Speech oriented stereo communication systems over headphones*.

[ITU-T P.863] Recommendation ITU-T P.863 (2018), *Perceptual objective listening quality prediction.*

[ITU-T P.913] Recommendation ITU-T P.913 (2021), *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment.*

[ITU-T P.917] Recommendation ITU-T P.917 (2019), *Subjective test methodology for assessing impact of initial loading delay on quality of experience*.

[ITU-T Handbook] ITU Publication (2011), *Practical procedures for subjective testing*.

[ITU-R BS.1770] Recommendation ITU-R BS.1770-4 (2015), *Algorithms to measure audio programme loudness and true-peak audio level*.

[3GPP EVS-7b] EVS Permanent Document EVS-7b, *Processing functions for selection phase*. Please refer to http://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/EVS_Permanent_Documents/

[3GPP EVS-7c] EVS Permanent Document EVS-7c, *Processing functions for characterization phase*. Please refer to http://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/EVS_Permanent_Documents/

[3GPP EVS-8b] EVS Permanent Document EVS-8b, *Test plans for selection phase including lab task specification*. Please refer to http://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/EVS_Permanent_Documents/

[3GPP EVS-8c] EVS Permanent Document EVS-8c, *Characterization Phase Test Plan including lab task specification*. Please refer to http://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/EVS_Permanent_Documents/

[3GPP S4-200158] 3GPP Tdoc S4-200158 (2020), *A Reference Audio Renderer for Qualification.* Please refer to http://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/TSGS4_107_Wroclaw/Docs

[3GPP S4-210848] 3GPP Tdoc S4-210848, *IVAS MASA C Reference Software package*. Please refer to https://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/TSGS4_114-e/Docs

[3GPP TR 26.918] 3GPP TR 26.918 (2022), *Virtual Reality (VR) media services over 3GPP.*

[3GPP TR 26.952] 3GPP TR 26.952 (2022), *Codec for Enhanced Voice Services (EVS); Performance characterization*.

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Supplement

This Supplement defines the following terms:

3.2.1 spatial audio (evaluation): (Quality assessment of) generic stereo/3D/immersive audio content offering a spatial sensation.

3.2.2 spatial speech (evaluation): (Quality assessment of) stereo/3D/immersive audio content offering a spatial sensation with predominant voice component(s).

4 Abbreviations and acronyms

This Supplement uses the following abbreviations and acronyms:

3GPP	3rd Generation Partnership Project
ACR	Absolute Category Rating
AllRAD	All-round Ambisonic Decoding
AMR	Adaptive Multi Rate (coding)
AMR-WB	Wide-Band Adaptive Multi Rate (coding)
DCR	Degradation Category Rating
ESDRU	Energy-based SDRU
EVS	Enhanced Voice Services
FB	Fullband
FER	Frame Erasure Rate
FOA	First Order Ambisonics

HOA	Higher Order Ambisonics
HOA3	3rd Order Ambisonics
IVAS	Immersive Voice and Audio Services
JBM	Jitter Buffer Management
LFE	Low Frequency Effects channel of multi-channel audio
LKFS	Loudness, K-weighted, relative to Full Scale
LPCC	Linear Pearson Correlation Coefficient
MASA	Metadata-assisted Spatial Audio
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
MUSHRA	Multi-Stimuli with Hidden Reference and Anchor points
NB	Narrowband
POLQA	Perceptual Objective Listening Quality Analysis
SDRU	Spatial Distortion Reference Unit
SWB	Super Wideband
VBR	Variable Bit Rate
VoIP	Voice over IP
VR	Virtual Reality
WB	Wideband

5 Conventions

The terms **reference conditions** and **anchor conditions** are used interchangeably in this Supplement.

6 ITU-T P.800 use case examples

This clause describes a collection of various [ITU-T P.800] use case examples that include narrowband (NB), wideband (WB), super wideband (SWB) and fullband (FB) audio bandwidth, speech, music and mixed speech and music content, stereo and spatial quality evaluations. It should be noted that the use of [ITU-T P.800] is not limited to the examples presented in this clause.

The use case examples include the following items:

- ITU-T P.800 use cases with SWB and FB audio bandwidths including mixed content quality evaluations during the standardization of the 3GPP enhanced voice services (EVS) codec, at all of qualification, selection, and characterization testing.
- ITU-T P.800 use cases with stereo speech quality evaluations as part of the ITU-T P.811 standardization, as published in Appendix II of [ITU-T P.811].
- ITU-T P.800 use cases of full-scale SWB experiments that included mixed-bandwidth samples and a set of anchor conditions for test-to-test consistency carried out during ITU-T P.863 standardization.
- ITU-T P.800 degradation category rating (DCR) experiment using EVS codec with parametric spatial speech.
- ITU-T P.800 quality assessments of ambisonics spatial speech:

- ITU-T P.800 quality assessments of first order ambisonics (FOA) spatial speech binaurally rendered over headphones.
- ITU-T P.800 DCR experiments using EVS codec to process 3rd order ambisonics (HOA3) and FOA channels of immersive speech conversation with music in background.
- ITU-T P.800 evaluation of ambisonics (spatial speech) quality with binaural headphone rendering.

6.1 ITU-T P.800 use case examples from 3GPP EVS codec standardization

3GPP SA4 carried out numerous ITU-T P.800 quality assessments during EVS codec qualification, selection and characterization testing. The test plans for selection and characterization [3GPP EVS-8b], [3GPP EVS-8c], the respective processing plans [3GPP EVS-7b], [3GPP EVS-7c] and the obtained experimental results [3GPP TR 26.952] are available through the corresponding references.

Below are important highlights of these experiments:

– Tests with SWB bandwidth included ITU-T P.50 modulated noise reference unit (MNRU) reference conditions.

– The processing for some tests included a network simulator to evaluate the tested codecs under voice over IP (VoIP) conditions.

– Tests involving mixed and music material rather than the conventional speech category used correspondingly adapted listener instructions.

An overview of these experiments is provided in the following tables. Table 1 presents a list of experiments in the enhanced voice services (EVS) codec selection test while Table 2 refers to the characterization tests.

Table 1 – List of experiments in the EVS codec selection tests [3GPP EVS-8b]

Exp.	Content	Methodology	# of Exp.
N1	NB clean speech under clean channel condition including input level dependency	Absolute category rating (ACR)	1
N2	NB clean speech under impaired channel conditions including delay/jitter profiles	ACR	1
N3	NB noisy speech under clean channel condition and impaired channel conditions	Degradation category rating (DCR)	1
N4	NB mixed content and music under clean channel condition and impaired channel conditions including delay/jitter profiles	ACR	1
W1	WB clean speech under clean channel condition including input level dependency	ACR	1
W2	WB clean speech under impaired channel conditions including delay/jitter profiles	ACR	1
W3	WB noisy speech under clean channel condition	DCR	1
W4	WB noisy speech under impaired channel conditions including delay/jitter profiles	DCR	1
W5	WB mixed contents and music under clean channel condition	DCR	1

Table 1 – List of experiments in the EVS codec selection tests [3GPP EVS-8b]

Exp.	Content	Methodology	# of Exp.
W6	WB mixed contents and music under impaired channel conditions	DCR	1
W7	WB mixed contents and music under impaired channel conditions including delay/jitter profiles	DCR	1
I1	AMR-WB IO clean speech under clean channel condition including input level dependency	ACR	1
I2	AMR-WB IO clean speech under impaired channel conditions	ACR	1
I3	AMR-WB IO noisy speech under clean channel condition	DCR	1
I4	AMR-WB IO noisy speech under impaired channel conditions	DCR	1
I5	AMR-WB IO mixed contents and music under clean channel condition	DCR	1
I6	AMR-WB IO mixed contents and music under impaired channel conditions	DCR	1
S1	SWB clean speech under clean channel condition including input level dependency	DCR	1
S2	SWB clean speech under impaired channel conditions including delay/jitter profiles	DCR	1
S3	SWB noisy speech under clean channel condition	DCR	1
S4	SWB noisy speech under clean channel condition	DCR	1
S5	SWB noisy speech under impaired channel conditions	DCR	1
S6	SWB mixed contents and music under clean channel condition	DCR	1
S7	SWB mixed contents and music under impaired channel conditions including delay/jitter profiles	DCR	1
		Total	24

Table 2 – List of experiments in the EVS codec characterization tests [3GPP EVS-8c]

Exp.	Test #	Listening Lab	Language	Designator	Bandwidth	Number of tests	Content	Test type	Noise type (SNR)	Notes
N1	1	Delta	FIN	an1.1	NB	1	Clean speech	ACR		Rate switching, tandeming, jitter buffer management (JBM)
N2	1	Dynastat	NAE2	bn2.1	NB	1	Noisy speech	DCR	Street (20 dB)	Rate switching. Untested selection conditions. Tandeming.
N3	1	Mesaqin	FRN	cn3.1	NB	1	Noisy speech	DCR	Street (25 dB)	High frame erasure rate (FER). 16.4 and 24.4 modes.
N4	1	Delta	DANm	an4.1	NB	1	Music/mixed	ACR		Rate switching. Untested selection condition.
W1	1	Dynastat	NAE1	bw1.1	WB	2	Clean speech	ACR		Rate switching. Channel aware mode.
	2	Mesaqin	CHN	cw1.2	WB			ACR		
W2	1	Dynastat	SPN	bw2.1	WB	2	Noisy speech	DCR	Office (20 dB)	Rate switching.
	2	Mesaqin	SLV	cw2.2	WB			DCR	Office (20 dB)	
W3	1	Dynastat	NAE3	bw3.1	WB/IO	2	Clean speech	ACR		Rate switching. wide-band adaptive multi rate coding (AMR-WB) IO Case C.
	2	Mesaqin	SLV	cw3.2	WB/IO			ACR		
W4	1	Dynastat	NAEm	bw4.1	WB/IO	1	Music/mixed	DCR		Rate switching. Variable bit rate (VBR).
W5	1	Delta	DAN	aw5.1	WB	1	Clean speech	ACR		Tandem. High FER.
S1	1	Dynastat	NAE1	bs1.1	SWB	2	Clean speech	DCR		Rate switching. Channel aware mode. Tandeming.
	2	Delta	DAN	as1.2	SWB			DCR		
S2	1	Delta	FIN	as2.1	SWB	2	Noisy speech	DCR	Car (15 dB)	Rate switching. High FER.
	2	Mesaqin	FRN	cs2.2	SWB			DCR	Car (15 dB)	
S3	1	Dynastat	SPNm	bs3.1	SWB	1	Music/mixed	DCR		Rate switching. JBM. Untested selection phase conditions.
M1	1	Dynastat	NAE2	bm1.1	NB, WB, SWB	1	Clean speech	DCR		EVS (NB, WB, SWB), AMR, AMR-WB
M2	1	Delta	FIN	am2.1	NB, WB, SWB	1	Noisy speech	DCR	Car (20 dB)	EVS (NB, WB, SWB), AMR, AMR-WB
M3	1	Mesaqin	CHNm	am3.1	NB, WB, SWB	1	Music/mixed	DCR		EVS (NB, WB, SWB), AMR, AMR-WB
F1	1	Mesaqin	GER	cf1.1	SWB, FB	1	Clean speech	DCR		EVS (SWB, FB)
F2	1	Delta	DANm	cf2.1	SWB, FB	1	Music/mixed	DCR		EVS (SWB, FB)
Total						22				

6.2 ITU-T P.800 DCR for stereo from ITU-T P.811 validation

During validation of [ITU-T P.811], called P.SOSH at the time, two ITU-T P.800 listening tests were conducted in conjunction with the ITU-T P.811 validation tests. These ITU-T P.800 tests are published in Appendix II of [ITU-T P.811]. One test was run by a Chinese listening laboratory, and another test was run by a German listening laboratory. The tests comprised SWB stereo speech material in Mandarin Chinese and German respectively. ITU-T P.811 and ITU-T P.800 degradation category rating (DCR) tests were run on the same test material, including spatial distortion reference units and channel randomization synchronized MNRU that were developed for [ITU-T P.811]. In the test by the Chinese laboratory, the instructions were written to be close to the [ITU-T P.811] instructions, while the instructions for the test by the German laboratory were slightly shorter.

6.3 Full-scale SWB experiments during ITU-T P.863 standardization

Design of ITU-T P.800 SWB ACR experiments targeted validation of perceptual objective listening quality analysis (POLQA) objective model in ITU-T SG12 Q9/12. Proponents of SWB databases had to meet the specific design rules described in Appendix II of [ITU-T P.863]. An additional constraint was that each test covered the entire range of degradation dimensions.

6.4 ITU-T P.800 DCR evaluation of parametric spatial speech

An ITU-T P.800 DCR listening test of binaurally rendered parametric spatial speech was carried out. The test content is based on varied recordings in both real environments and controlled listening room environments representing spatial audio communications and user-generated content capture use cases. The detailed description of the experiment including the capture set-ups and the specific listener instructions is provided in Appendix I.2.

6.5 ITU-T P.800 evaluations of ambisonics spatial speech quality

6.5.1 ITU-T P.800 evaluations of first order ambisonics (FOA) spatial speech quality with binaural headphone rendering

Two ITU-T P.800 quality assessments were carried out of first order ambisonics (FOA) spatial speech binaurally rendered over headphones. The detailed description of these experiments is available in Appendix I.1.

Below are important highlights of these experiments:

– Tests with FOA spatial speech with SWB bandwidth, binaurally rendered over headphones.

– Tests with spatially adapted ITU-T P.50 MNRU and energy-based SDRU (ESDRU) reference conditions.

– Tests with talker interactions (overtalk) and with moderate to high spatial ambient noise levels.

– Tests in 2 languages, Polish and American English.

An overview of these experiments is provided in Table 3:

Table 3 – List of FOA spatial speech quality experiments

Exp.	Content	Methodology
Exp1	Use case 'immersive conferencing' with ambisonics (FOA) spatial speech, 6 content type categories	DCR
Exp2	Use case of immersive telephony while on the move (outside) with ambisonics (FOA) spatial speech, 6 content type categories	DCR

6.5.2 ITU-T P.800 evaluations of scene-based audio (ambisonics) spatial speech quality with binaural headphone rendering

Two listening tests of scene-based audio (up to 3rd order ambisonics) were carried out to obtain insights about the suitability of ITU-T P.800 DCR testing for clean and noisy speech in a binaurally rendered 3D audio scene. The two tests were conducted each with naïve subjects from the general population and also using listeners with previous experience in assessing spatial audio. The detailed description of these experiments is available in Appendix I.3.

The tests covered a wide range from EVS in mono to uncoded HOA3 (3rd order ambisonics) exposing multiple types of degradations.

6.5.3 Comparison of DCR test experiments for FOA and HOA3 input in 7.0+4 and binaural listening set-ups

Two DCR experiments were carried out using the EVS codec to process HOA3 and FOA channels of artificially created immersive speech conversation with music in background. The goal of this evaluation was to contribute to the collective experience with immersive listening testing using naïve listeners. The test was run twice to compare naïve listeners' perception using 7.0+4 loudspeaker listening set-up and a binaural listening set-up using headphones. The details of the tests, results and analysis can be found in Appendix I.4.

7 ITU-T P.800 best practices especially related to stereo and spatial speech quality evaluation

This clause provides guidance for future tests on the basis of past experience.

An important aspect in test experiment design is identified when multiple degradation types are mixed in the same experiment (e.g., signal quality and spatial degradations).

The next main aspect is related to the listeners. The influence of experience level of subjects when running complex (spatial audio) ITU-T P.800 tests is discussed, followed by considerations on instructions to listeners. Methods for listener pre-screening and more importantly post-screening are described in detail.

The following clauses describe the choice of anchor conditions, with focus on ITU-T P.50 MNRU and ESDRU, and level normalization for stereo and multichannel signals, as may be necessary, e.g., avoiding clipping and handling the fact that audio codecs may not be level neutral due to format conversions.

As far as application of ITU-T P.800 DCR for stereo speech is concerned, a best practice is addressed based on ITU-T P.811 validation.

7.1 General best practices

7.1.1 Experimental design

When mixing multiple degradation types in a single experiment, both the frequency and the intensity of a given degradation type may have a strong impact on the remaining conditions, rendering the comparison of different degradations more difficult. For example, the preference of noise over packet loss could differ from one experiment to another. This is an important aspect that should be considered when designing a subjective test. Depending on the purpose of the test, where possible, the experiment may want to reduce the number of degradation types in a single test, or implement tighter constraints on the design.

The numbers of conditions, content types (categories), samples per content type and listener panel and the construction of randomizations should be done following the "partially-balanced/randomized blocks" experimental design described in "Practical procedures for subjective testing", [ITU-T Handbook]. However, the Handbook does not take into account aspects of spatial audio testing.

7.1.2 Listener instructions for ITU-T P.800 DCR evaluation of spatial speech

In ITU-T P.800 DCR, a single vote is given following each pair (or repeated pair). For spatial audio, this single vote aims to cover all degradations (dimensions). In comparison, in [ITU-T P.811], three sub-trials for signal rating, spatial rating, and overall rating are conducted. It is important for the listeners to understand what they are listening to and what they are evaluating. This can be achieved by suitable listener instructions and familiarization or training for spatial audio content. For example, specific listener instructions for spatial audio evaluations may be used and training sessions can help listeners building their expectation and increase their awareness with spatial content and the types of distortions and spatial impairments that may appear.

Specific guidance on making the test subject aware of the location attribute of stereo speech is found in [ITU-T P.811] and should also be taken into account. Depending on the task, modified rating scales could be considered.

Generic listener's instructions

– It is explained to listeners that they should be using the full range of degradation scores. For example: "*The sound samples you will hear in the test are expected to span the complete range of degradation scores*".

Specific listener's instructions for spatial audio evaluation task

– It is explained to listeners that they are listening to spatial audio/binaural audio and what this means. For example: "*Binaural means that you can locate various sound sources around yourself while listening with headphones*".

– Examples are provided of how the listener can perceive spatial audio/binaural audio. For example: "*[A] first talker may appear to talk from the left-hand side and a second talker from the right-hand side*".

– It is explained to listeners how spatial audio/binaural audio differs from their possible previous experiences and expectations (e.g., in this case, from mono voice communications). For example, in the context of a listening test with headphones: "*The experiences in this test, with spatial audio, differ from traditional mono audio where you cannot hear the direction of talkers and where two talkers would appear to talk from the same position inside your head*".

– It is clarified what the listener is asked to consider when casting their vote. For example: "*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*".

– It is explained to listeners that the location of sound sources or of the speaker is an important spatial sound quality attribute. For example: "*You may hear that the location of a sound has moved compared to the reference, which would be a spatial impairment. There may also be other spatial impairments you that hear, e.g., a changed width of the sound scene*".

– It is explained to listeners that any heard difference should be rated as a degradation (where the experiment is designed such that the reference sample represents the optimum result). For example: "*In this test the first sample is the expected experience and therefore any difference of the second sample shall be regarded as a degradation even if it may sound better to you*".

Creation of awareness of spatial content and expectation building

- The type of content the listener will evaluate is described. For example: "*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*".
- Motivation is provided for the task with expectation on what this can relate to, which may for example be telling the listener that he or she may listen to systems that go beyond his/her current experience. For example: "*For the second sample there may have been some future mobile phone technology used*".

Familiarization with the content and the types of distortions and spatial impairments

- Before the listening test and the practice test, several introductory samples may be played back covering the full range of degradations appearing in the test. This familiarizes the listener with spatial audio and introduces the listener to the types of practical degradations that can appear in the evaluation. These can be difficult to explain to naïve listeners in text. Instead of playback of pure audio samples, an introduction video may serve the same purpose but may use the visual senses to explain certain degradations or novel perceptions.
- Before the listening test, a short practice test with samples covering the full range of degradations appearing in the test is recommended. The listener may learn here to use the full range of degradation scores. Also, some discussion on listeners' perception after the training session might help to ensure the listeners understand the task.

Example of instructions to listeners

Examples of instruction to listeners are found in the descriptions of experiments in Appendix I and in [ITU-T P.811], Appendix II.

7.1.3 Subject pool

ITU-T P.800 subjective listening tests are in general meant for assessment by the general population and therefore mainly subjects that can be considered as naïve subjects when it comes to the evaluation in an experiment should be recruited. While this requirement serves the purpose of evaluating technologies for the general public, such evaluations are only valid at the time the experiment was conducted and will not represent the experience in the future of such a service. This may penalize novel kinds of experiences that are commonly unknown and potentially irritating for naïve subjects for the moment but may be differently perceived in the future.

Therefore, experiments with conditions that are knowingly beyond the common experience for naïve subjects may use more experienced listeners that are familiar with such novel experiences, obtained, e.g., via extensive training sessions or by other relevant experiences.

7.1.4 Listener screening

7.1.4.1 Pre-screening

[ITU-T P.800] indicates that in some cases screening of subjects may be necessary and a method based on Annex B of [ITU-T P.78] may be applicable. The purpose of this pre-screening is to make sure that only subjects with suitable/normal hearing are admitted to the test.

7.1.4.2 Post-screening

Despite suitable/normal hearing it may turn out that the scores of certain listeners may be useless. There may be many reasons including for instance, lack of attention or lack of comprehension of the listener task. The effect on the opinion scores of such listeners is that they are statistically incompatible with the gross of other listener opinion scores such that including them in the calculation of mean opinion scores is not warranted. Procedures to analyse individual listener scores about their statistical compatibility with the gross listener scores with the possibility to reject them are called post-screening.

A general statistical post-screening method is described in Annex A of [ITU-T P.917]. It evaluates the (linear Pearson) correlation coefficient (LPCC) of the scores from one subject versus the scores from all subjects on certain sets of test items. A useful post-screening criterion in the context of this Recommendation, where it is a best practice to apply a balanced block design with multiple listener panels, is to evaluate the LPCCs for the full set of conditions. This means:

The linear Pearson correlation coefficient (LPCC) for one subject versus all subjects is calculated as:

$$\mathrm{LPCC}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i{}^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i{}^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}} \qquad (1)$$

where *x* and *y* are arrays of data and *n* is the number of data points.

To calculate LPCC on all conditions, compute:

$$r(x,y) = \mathrm{LPCC}(x,y) \qquad (2)$$

where in Equation (1):

x_i: condition MOS of all subjects (i.e., condition MOS is the average value across all listener scores from the same condition)

y_i: individual condition MOS of one subject for the corresponding condition

n: total number of conditions

i: condition number.

Screening analysis is performed using Equation (2). Subjects are rejected if *r* falls below a set threshold. Subjects should be discarded one at a time, beginning with the worst outlier (i.e., lowest *r*) and then recalculating *r* for each subject.

It is recommended to apply an appropriate discard threshold *t* on LPCC, such that a subject having an LPCC < *t* is discarded. A typical value is *t* = 0.75. However, different thresholds may be needed depending upon the specific test, where more demanding tests tend to require lowering the threshold to, e.g., *t* = 0.7, or tests with expert listeners may use a higher threshold, e.g., *t* = 0.8.

7.1.4.3 Alternative post-screening

The objectives of this alternative post screening method are:

– To eliminate listeners who are not able to perform the required task, i.e., to reliably detect degradations in the audio sample due to both coding artefacts and modifications in the spatial representation.

– Not to eliminate outlier listeners, i.e., listeners who do hear degradations, but their opinion is not aligned with average listeners. For example, a listener significantly more sensitive to a spatial impairment than the average listener should not be discarded. Similarly, a listener particularly sensitive to a given artefact should not be discarded just because average listeners do not perceive it as strongly, or at extremis do not hear it at all.

For this reason, the following screening method is based on correctly assessing the anchor conditions only. In order to be considered:

– A listener must use the whole voting scale during the session. In other words, the listener must have voted at least once "1" and at least once "5".

- A listener must vote, in average, the direct condition better than or equal to a high quality MNRU ITU-T P.811 condition.
- If spatial distortion reference units (SDRUs) of [ITU-T P.811] are used, a listener must vote, in average, the direct condition better than or equal to a high quality SDRU condition.
- A listener must vote for each MNRU, in average, a higher quality MNRU condition better than a lower quality MNRU condition.
- If SDRUs are used, a listener must vote for each SDRU, in average, a higher quality SDRU condition better than a lower quality SDRU condition.

The exact value of the MNRU and SDRU conditions used for elimination of listeners might depend on the test resolution, e.g., on the number and values of MNRUs and SDRUs used in the test. For this reason, the formulation above is kept generic, allowing an MNRU condition to be compared with next lower MNRU condition, or, e.g., with the second next lower MNRU condition, etc.

7.1.5 Anchor conditions

7.1.5.1 Background from anchor conditions in ITU-T P.800 ACR testing

A large number of ITU-T P.800 ACR SWB speech subjective assessments were designed during the development of [ITU-T P.863]. The aim for these experiments was to train and validate objective models on a wide variety of degradation types and intensity. In the effort to obtain consistent rating across subjective tests, the concept of "full-scale" experiment was introduced.

Based on ITU-T P.800 ACR methodology, these experiments presented additional design constraints to ensure that each experiment was balanced in many aspects, especially on the spread of distortions on each of four degradation types: background noise, filtering, temporal clipping and presentation level. A detailed description of the full-scale experiment design is provided in Appendix II of [ITU-T P.863].

This experiment design was largely successful with the introduction of anchor conditions for each degradation types, providing an overall stable score distribution across experiments. The subjective scores varied expectedly and consistently when degradation types were analysed separately. For example, an increase of packet loss or background noise would lower the subjective scores. This indicates that the ACR method is likely to provide repeatable assessments of super-wideband signals when the number of degradation types is limited.

7.1.5.2 Anchor conditions for ITU-T P.800 testing of spatial speech

Use of ITU-T P.50 MNRU and ESDRU for signal and spatial anchors seems suitable for spatial audio items with binaural headphone evaluation. The ITU-T P.50 MNRU anchors predominantly exhibit signal degradation. To this end, the ITU-T P.50 MNRU should be applied coherently to component signals of the spatial audio. Otherwise, there might be introduced decorrelation between component signals that may have a spatial widening effect. On the other hand, any practical spatial audio coding system can have spatial degradations. The ESDRU, affecting left and right stereo channels, is applied to binaural output and introduces specifically spatial impairments.

The calibration of the used ITU-T P.50 MNRU and ESDRU reference conditions should be done such that they span the complete range of impairments of the experimental main conditions. An experimental approach for calibration of the reference conditions is described in clause I.1.4.

7.1.6 Stereo and multichannel level normalization

Appropriate level normalization is essential for fair comparison of the conditions of a listening test. Most codecs and audio processing tools aim to preserve the input level, but in some cases the level of the final output signal may be affected. In such cases, the conditions presented to the listener with a higher level is typically favoured in the scoring. For mono signals the speech voltmeter algorithm implemented by the sv56demo tool in [ITU-T G.191] is commonly used. This is however not suitable for audio signals with more than one channel. In such cases, the multi-channel level measurement tool [ITU-R BS.1770] described in clause 8.4 is appropriate. There are several considerations regarding level normalization:

- When applying level normalization, care should be taken such the signal is not overloaded (clipped) in any part of the processing chain.
- Although audio codecs in general aim to preserve the signal level, some processing steps may alter the level. For example, format conversions such as ambisonics to multi-channel loudspeaker layouts are not generally reversible and may not be level neutral. Also, the spatial degradation anchors SDRU (clause 8.2) and ESDRU (clause 8.3) and to some extent ITU-T P.50 MNRU (clause 8.1) do not preserve the signal level and level normalization is recommended after applying these. The bs1770demo tool from [ITU-T G.191] reports the level in loudness, K-weighted, relative to full scale (LKFS) while the "Scaling factor" is in the linear domain, similar to the "Norm factor" of sv56demo.
- When the input signal is in a format different from the listening format, a level measurement can be done on the output format and a level normalization may be applied on the input signals. For example, a processing chain which inputs an ambisonics signal and outputs a loudspeaker or binaural signal may be measured on the output using bs1770demo. The obtained gain factor may be applied on the input ambisonics signal. However, it should be noted that [ITU-R BS.1770] includes a perceptual weighting filter and an adaptive gating function. For this reason, the obtained level may not exactly match the target level when applying a gain factor on the input signal.
- More generally, whenever the processing chain contains a renderer, which can be regarded as an audio format converter, special care should be taken to ensure that the output audio signal levels are correct and consistent across conditions. This may especially be of relevance if different renderers are used for different conditions. In such cases, it may be necessary to carry out post normalization of the audio levels on the output signals.
- Audio codecs are often designed to, under certain requirements, not change the output level outside a certain tolerance level. If, however, the processed conditions of the listening test do not preserve the level, a post-normalization of the levels before listening may be considered.

7.2 Best practices from specific experiments

7.2.1 ITU-T P.800 DCR for stereo from ITU-T P.811 validation

The tests showed a high correlation between the ITU-T P.800 DCR and the overall score of [ITU-T P.811]. Hence, the set-up outlined is a useful alternative when the overall score is of the main interest, requiring a shorter test time than for a corresponding ITU-T P.811 test.

Two spatial distortion reference units are presented in [ITU-T P.811], SDRU and ESDRU. The SDRU offers a somewhat stronger spatial distortion, which may be perceived a bit unnatural. The ESDRU aims to provide a more realistic distortion but may give a slightly weaker spatial distortion compared to the SDRU. It should be noted that the SDRU and ESDRU do not preserve the signal level, and it is recommended that the spatial distortion processing is followed by a level normalization.

8 ITU-T P.800 reference conditions and tools used in ITU-T P.800 use case examples

The reference or anchor conditions of a listening test provides controlled and known degradation of the audio samples presented to the listeners. These conditions give a reference point for making sensible comparisons of different listening tests, conducted in various listening labs or at a different time. It should however be noted that comparisons of listening tests should be made with caution and that the score of a reference condition will be affected by the accompanying listening test material. The reference conditions also serve as a guide for the listener subjects, making them aware of the dimensions of degradation to consider when casting their votes. As indicated in [ITU-T P.800], the reference conditions depend on what is being assessed. This section describes a set of reference conditions which were found useful and relevant in their respective listening tests.

8.1 ITU-T P.50 fullband MNRU

8.1.1 Description

ITU-T P.50 MNRU was introduced to accommodate the fact the energy at higher frequencies in super-wideband and fullband speech signals is much weaker that in the lower frequencies. As a results, the white noise traditionally used in MNRU creates an uneven SNR across the signal spectrum, generating unnatural and more noticeable distortions.

ITU-T P.50 fullband MNRU was introduced to address this issue by shaping the modulated noise to match the average spectrum of speech signals. The algorithm is defined in [ITU-T P.810] and an ANSI C implementation is available in the software tool library 2023 in [ITU-T G.191].

8.1.2 Applications

8.1.2.1 Super-wideband and fullband experiments

The ITU-T P.50 fullband MNRU was introduced in 2008 during the design of ITU-T P.800 ACR subjective tests for the validation of [ITU-T P.863], also known as POLQA. Since, it has been used as the MNRU reference conditions for most SWB and FB experiments, notably for the ITU-T P.800 tests designed for the validation and characterization of the 3GPP EVS codec.

8.1.2.2 ITU-T P.800 DCR for stereo from ITU-T P.811 validation

The MNRU adapted for SWB testing as described under ITU-T in clause 8.1.1 works well for mono testing but needs to be amended for stereo testing. Applying the MNRU on the channels separately may create a diffuse noise source which is spatially separated from the source material. By synchronizing the random number generator between the channels, the noise will appear to have the same location as the source material. This is implemented as a part of the SDRU executable provided as a software attachment to [ITU-T P.811]. It could also be created by applying the ITU-T P.50 MNRU on the left and right channels as in clause 8.1.2.3, provided that the random number generator is set to the same value for each channel. Note that the ITU-T P.50 MNRU includes a high-pass filter with a cut-off frequency at 115 Hz. In case of strong low-frequency content, this effect may be audible and a level re-normalization may be desirable after applying the ITU-T P.50 MNRU.

8.1.2.3 ITU-T P.800 evaluations of First order ambisonics (FOA) spatial speech quality with binaural headphone rendering

ITU-T P.50 MNRU reference conditions were used with the following adaptation to FOA audio. The monophonic ITU-T P.50 MNRU (as defined above) was coherently applied to all 4 FOA component signals. To achieve coherent application, the ITU-T P.50 MNRU was operated with the same random generator seed for all component signals. In practice, this is achieved by starting each instance of the ITU-T P.50 MNRU executable on the same frame of the component signals. The perceptual effect of the synchronization is 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. It is to be noted that it was also considered to apply the ITU-T P.50 MNRU after binauralization as it is done in clause 8.1.2.2. However, this option was discarded as it did not equally well retain the spatial image as when applying the ITU-T P.50 MNRU to the ambisonics component signals.

8.2 SDRU

8.2.1 Definition

SDRU is a spatial distortion reference unit defined in Figure A.2 of [ITU-T P.811], and in the software attachment. It modulates the position of the source using a triangular wave with a period of 1 second. The strength of the spatial distortion is controlled through the amplitude of the modulation. The triangular wave offers a smooth and somewhat stronger spatial distortion compared to the ESDRU. On the other hand, this modulation is perhaps less realistic and the sensation of rotation may induce dizziness among the listener subjects.

8.2.2 Applications

8.2.2.1 ITU-T P.800 DCR for stereo from ITU-T P.811 validation

The SDRU was applied to the pre-processed stereo input signal, and the level normalization was reapplied afterwards.

8.3 ESDRU

8.3.1 Definition

The ESDRU is an alternative spatial distortion reference unit defined in [ITU-T G.191]. It is similar to the SDRU, but the triangular wave modulation function is replaced with a stepwise random function. This pattern intends to simulate errors in a parametric representation of a stereo signal. To avoid discontinuities in the pattern, leading to distortions in the signal domain, the transitions between each step have been smoothened. Further, an energy analysis of the signal is done to avoid changes in the modulation function during high energy segments. While the ESDRU aims to reproduce a more realistic distortion, it may give a slightly weaker spatial distortion than the SDRU.

The ESDRU tool is controlled by two parameters. The first parameter is the overall modification strength $\alpha \in [0,1]$, where $\alpha = 1$ gives no modulation and $\alpha = 0$ gives maximum modulation. The second parameter $e_{step} \in [0,1]$ is the allowed modulation step during high energy segments. It was observed that a reduced modulation step during high energy segments is desirable to minimize the signal distortion. The default value is $e_{step} = 0.5$.

Example command lines for ESDRU C-code implementation in [ITU-T G.191]:

- To run ESDRU with alpha=0.5:
```
esdru 0.5 input.pcm output.pcm
```
- To run ESDRU with alpha=0.0, 32 000 Hz sampling rate, modulation step during high energy = 1.0, seed 10:
```
esdru -sf 32000 -e_step 1.0 -seed 10 0.0 input.pcm output.pcm
```

8.3.2 Applications

8.3.2.1 ITU-T P.800 DCR for stereo from ITU-T P.811 validation

The ESDRU was applied to the pre-processed stereo input signal, and the level normalization was reapplied afterwards.

8.3.2.2 ITU-T P.800 evaluations of first order ambisonics (FOA) spatial speech quality with binaural headphone rendering

The ESDRU was directly applied to the two binaural channels after binaural rendering of the FOA signal, followed by level re-normalization.

8.4 Multi-channel level normalization with BS1770DEMO

BS1770DEMO is an implementation of the method to measure multi-channel audio signals described in [ITU-R BS.1770]. It has been included in [ITU-T G.191]. If the tool is run only with an input file, it runs in measurement mode and outputs the estimated level in LKFS. If an output file is specified, the level of the output file is set to the target level. The algorithm is based on level normalization for a multi-channel set-up, where each audio channel has a position described in terms of azimuth θ and elevation ϕ. In case the loudspeaker position falls in the side region defined as:

$$\begin{cases} |\phi| < 30^\circ \\ 120^\circ \leq |\theta| \leq 60^\circ \end{cases}$$

where the forward direction is defined as $\theta = 0$, the level is scaled with a factor of 1.41. For channels corresponding to loudspeaker positions outside of the side region, the factor is 1.0. Finally, LFE channels are omitted by setting the factor to 0. The channel configuration may be given as a string argument to the tool using the flag "-conf STRING" where STRING is a sequence of characters according to:

- '0' – Not in the side region.
- '1' – Within the side region.
- 'L' – LFE channel.

Example command line for clause 5.1 audio input file with channel order [FL FR C LFE BL BR]:

```
bs1770demo -nchan 6 -conf 000L11 -lev -26 input.pcm output.pcm
```

with:

- FL = Front Left
- FR = Front Right
- C = Centre
- LFE = Low Frequency Effects
- BL = Back Left
- BR = Back Right

Appendix I

Collection of ITU-T P.800 test realizations

This appendix is a collection of reports of ITU-T P.800 test realizations that are not referenced to other standards documents.

I.1 ITU-T P.800 evaluations of first order ambisonics (FOA) spatial speech quality with binaural headphone rendering

Two ITU-T P.800 quality assessments of first order ambisonics (FOA) spatial speech binaurally rendered over headphones were carried out. The following contains the detailed description of these experiments.

Important highlights of these experiments are:

– Tests with spatial FOA speech with SWB bandwidth.

– Tests with spatially adapted ITU-T P.50 MNRU and ESDRU reference conditions.

– Tests in 2 languages, Polish and American English.

– Tests with talker interactions (overtalk) and with moderate to high spatial ambient noise levels.

An overview of these two experiments is provided in Table I.1:

Table I.1 – List of FOA spatial speech quality experiments

Exp.	Content	Methodology
Exp1	Use case 'immersive conferencing' with ambisonics (FOA) spatial speech, 6 content type categories	DCR
Exp2	Use case of immersive telephony while on the move (outside) with ambisonics (FOA) spatial speech, 6 content type categories	DCR

I.1.1 Test purpose

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

I.1.2 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 after binaural rendering according to [ITU-R BS.1770]). The scaling factors were derived by applying the binaural rendering to the speech signals and background noises separately and then normalizing them according to [ITU-R BS.1770]. The signals where then scaled accordingly before the mixing.
  - 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, 25 dB SNR, based on level normalization after binaural rendering according to [ITU-R BS.1770], see Exp1)
  - 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)

I.1.3 General consideration of the 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", [ITU-T Handbook].
- Every condition has 30 different samples passed through it (6 categories × 5 panels). Each of these are voted on by the 6 subjects in the panel, giving: (30 samples × 6 subjects/panel) = 180 votes per condition.
- 30 test conditions × 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

I.1.4 Degradation references (anchors)

According to Appendix II of [ITU-T P.811], ITU-T P.811 overall quality scores strongly correlate with ITU-T P.800 DCR scores if the latter is run with modified instructions and degradation references that span both signal and spatial quality dimensions. [ITU-T P.811] suggests using ITU-T P.50 MNRU for signal degradation anchors and SDRU/ESDRU for spatial degradation anchors. ITU-T P.50 MNRU is a modulated noise reference unit with ITU-T 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 ITU-T P.50 MNRU and the ESDRU to derive degradation anchors for our ITU-T P.800 experiments with binauralized FOA content.

For the ITU-T 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 ITU-T 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 ITU-T P.50 MNRU and ESDRU degradation parameters Q and, respectively, α.

In the experiment 6 FOA voice vectors were degraded either with ITU-T 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 multi-stimuli with hidden reference and anchor points (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 shown in Figures I.1a, I.1b and I.1c:

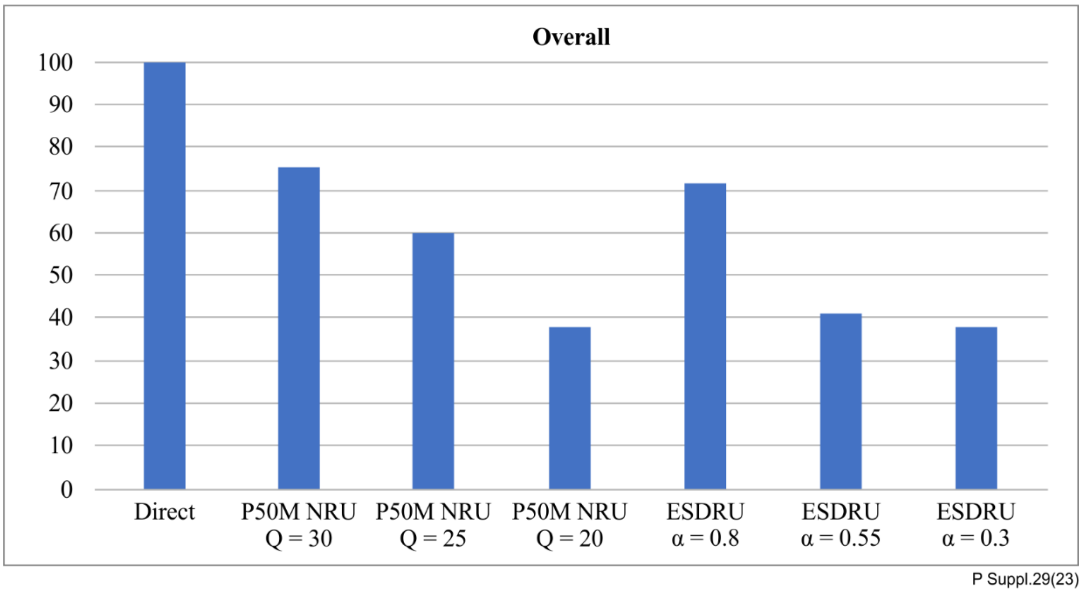


Figure I.1a – Overall quality

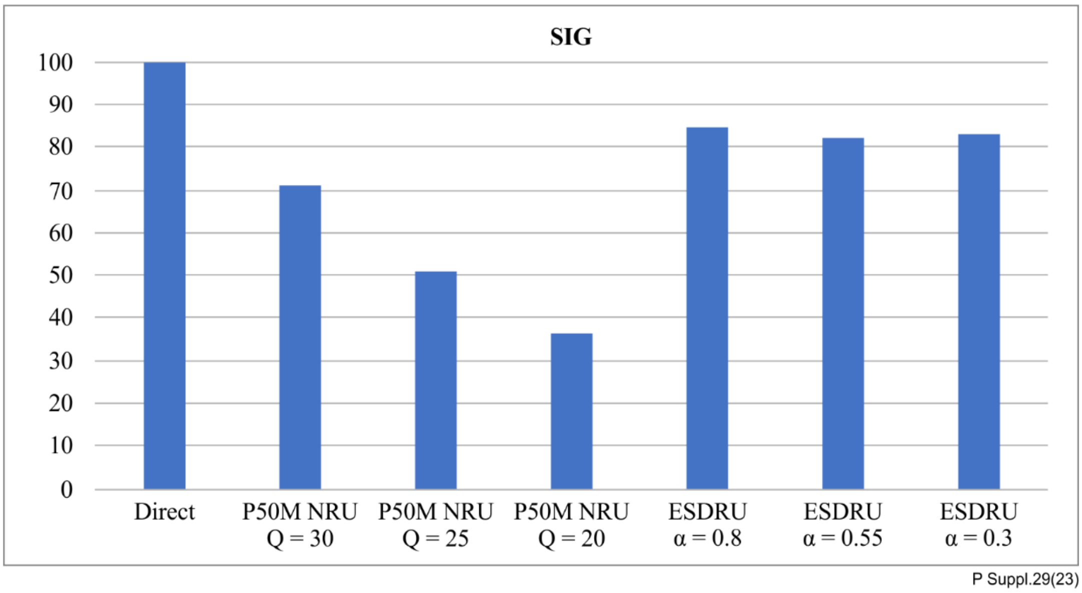


Figure I.1b – Signal quality (SIG)

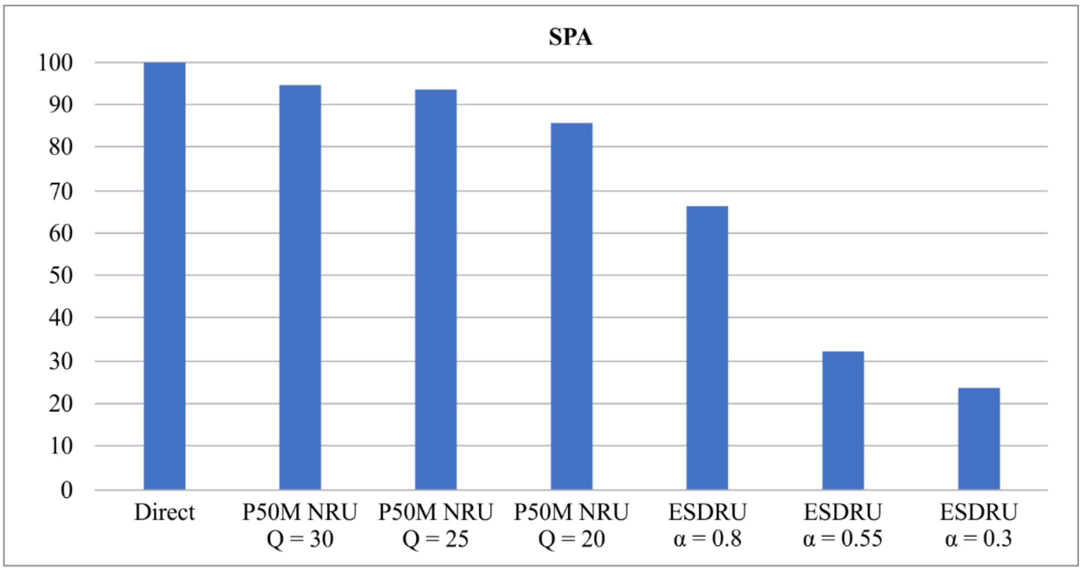


Figure I.1c – Spatial quality (SPA)

From the plots, the following observations can be made:

- The ITU-T 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 ITU-T P.50 MNRU induced signal degradation appears a bit too strong and should be softened for the ITU-T 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 ITU-T P.800 tests, it was decided to increase the α parameters.

I.1.5 Factors and conditions

Main codec conditions		
Codec under test (CuT)	11	Dolby-internal FOA coding system
Codec references		
Codec references	12	Multi-mono 4×EVS 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
ITU-T P.50 MNRU (applied to all FOA components)	3	Q = 22, 27, 32 dB (all: nominal level)
ITU-T P.811 ESDRU	3	α = 0.55, 0.7, 0.85 (output loudness forced to nominal level after application of ESDRU)
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 [3GPP Tdoc S4-200158]
Audio sampling frequency/bandwidth	2	48 kHz/SWB except for 4×EVS@4×8 kbit/s 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]
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

I.1.6 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
ITU-T P.50 MNRU (applied to all FOA components)	3	Q = 22, 27, 32 dB (all: nominal level)
ITU-T P.811 ESDRU	3	α = 0.55, 0.7, 0.85 (output loudness forced to nominal level after application of ESDRU)
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 4×EVS@4×8 kbit/s 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])
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

I.1.7 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

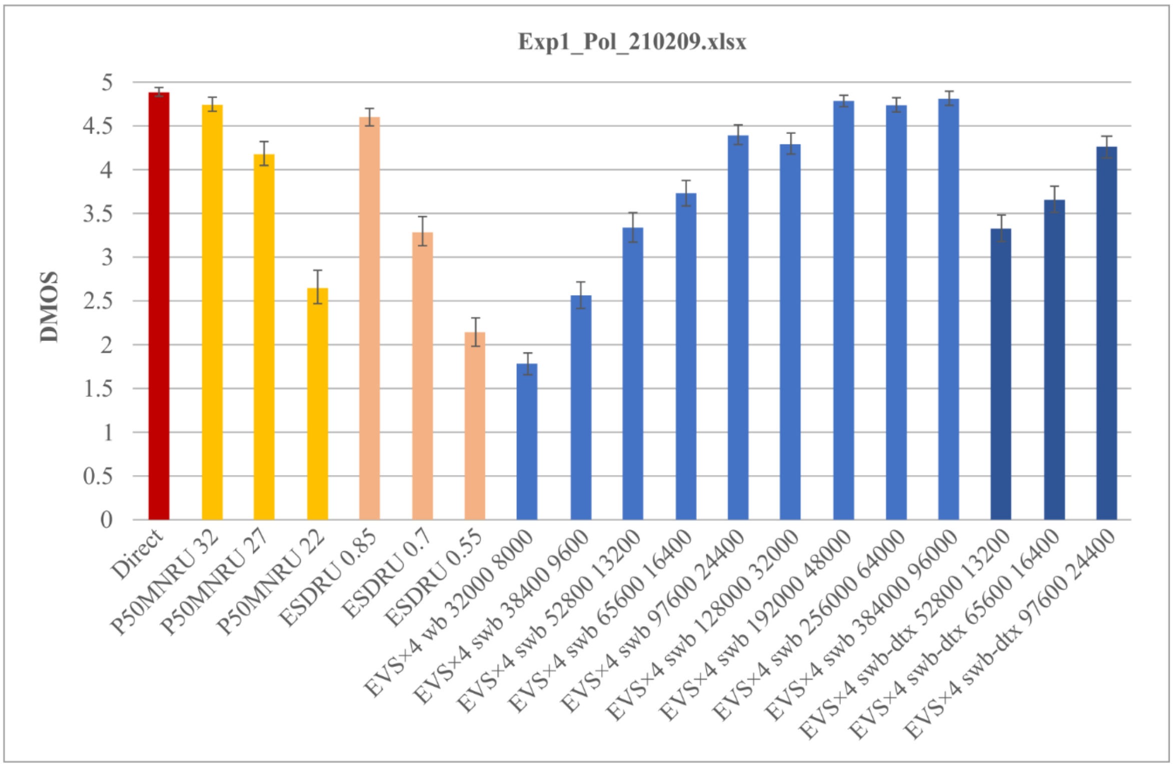
1: Very annoying

I.1.8 Results

I.1.8.1 Exp1: Use case 'immersive conferencing' with FOA speech (Polish)

Figure I.2 is a graph showing the mean opinion scores (MOS) with 95%CI of the direct, the degradation reference and the EVS multi-mono conditions. The results of the codec under test (CuT) are not presented as they are not relevant for the purpose of this contribution.

I.1.8.1.1 Degradation mean opinion scores



P Suppl.29(23)

Figure I.2 – Experiment 1: 'Immersive conferencing' with FOA speech (Polish)

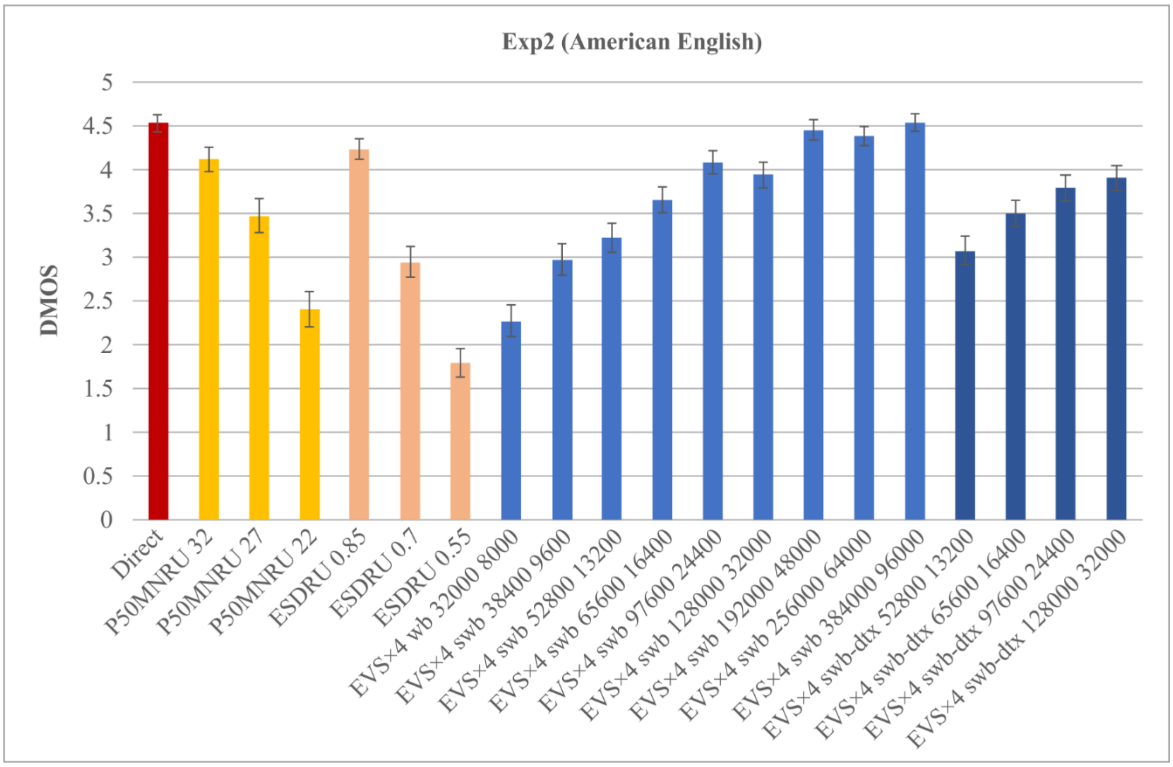
I.1.8.1.2 Observations

The mean opinions scores observed for the reference systems and especially the EVS reference system are very consistent with the expectations that build on the performance characterization of that codec [3GPP TR 26.952]. This is an indicator of the good resolution of the test. At high bit rates there is some non-surprising saturation effect.

I.1.8.2 Exp2: Immersive telephony while on the move (outside) with FOA audio (American English)

Figure I.3 is a graph showing the MOS scores with 95%CI of the direct, the degradation reference and the EVS multi-mono conditions. The results of the CuT are not presented as they are not relevant for the purpose of this contribution.

I.1.8.2.1 Degradation mean opinion scores



P Suppl.29(23)

Figure I.3 – Experiment 2: Immersive telephony while on the move (outside) with FOA audio (American English)

I.1.8.2.2 Observations

As in Exp1, the mean opinions scores observed for the reference systems and especially the EVS reference system are very consistent with the expectations that build on the performance characterization of that codec [3GPP TR 26.952]. This is an indicator of the good resolution of the test. At high bit rates the test resolution may run into saturation, however, the results still indicate a performance dip for the condition based on EVS operated at 64 kbit/s (EVS×4 swb 256000 64000), which coincides with a behaviour also seen in other EVS codec evaluations.

I.2 ITU-T P.800 DCR evaluation of parametric spatial speech

I.2.1 Test purpose

This test was originally designed for evaluation of potential reference conditions for the parametric metadata-assisted spatial audio (MASA) format in the context of 3GPP immersive voice and audio services (IVAS) codec standardization. The test further builds understanding about suitability of ITU-T P.800 DCR test methodology with ITU-T P.50 MNRU signal quality and ITU-T P.811 ESDRU spatial quality anchors for quality assessments of parametric spatial speech.

I.2.2 Test outline

This is an ITU-T P.800 DCR test based on binaural headphone listening of spatial speech items.

Generation of all the original items, as well as the binaural rendering for listening, is based on the IVAS MASA C reference software package [3GPP S4-210848], unless otherwise indicated in the following.

Content types and material generation:

- Realistic spatial speech items in real environments and controlled listening room 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".
- Original audio was recorded in various indoor and outdoor environments using Eigenmike, Eigenmike + an external microphone, ambisonic microphone + an external microphone, or a multi-microphone smartphone mockup device.
- Majority of the captured signals were analysed with the IVAS MASA C reference software with the sole exception of the smartphone mockup samples that were analysed using an in-house parametric analysis method.
- Binaural rendering was performed using the IVAS MASA C reference software package for all conditions.

Evaluation and listening system/environment:

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

I.2.3 Detailed test description

The following provides detailed description of the test:

- DCR test methodology;
- 16 test subjects;
- Six sample categories;
- Four randomizations for each 4-listener panel;
- Four samples per category (one for each listening panel);
- 96 votes casted for each condition;
- Total of 24 conditions:
  - 7 reference conditions, direct reference had no quantization and unquantized (UQ) spatial metadata;
  - 8 coded references using 3GPP EVS with UQ spatial metadata;
  - 9 CuTs (omitted from results);
- Degradation references: ITU-T P.50 MNRU and ESDRU;
- ITU-T P.50 MNRU Q values of 30, 24, and 18 dB;
- ESDRU values of 0.85, 0.70, and 0.55;
- Average trial duration: 20 s:
  - 8 s reference sample + 0.5 s silence + 8 s test sample + 3.5 s voting period;
- Test duration: ~1.4 h per listening panel including instructions, preliminaries, and breaks;
- Lab: Nokia Technologies (Tampere, Finland).

Main codec conditions		
Codec under test (CuT)	9	Nokia-internal IVAS MASA coding system
Codec references		

Codec references	8	EVS with unquantized MASA metadata operated at 8(WB), 9.6, 13.2, 16.4, 24.4, 32, 48, 64 kbit/s. Rendering with the IVAS MASA C reference binaural renderer.
Other references		
Direct	1	Analysed with the IVAS MASA C reference software. No transport stream nor MASA spatial metadata compression. Rendering with the IVAS MASA C reference binaural renderer.
ITU-T P.50 MNRU (applied to MASA transport stream)	3	Q = 18, 24, 30 dB (output loudness set to nominal level)
ESDRU [4] (applied to binaural rendering)	3	α = 0.55, 0.7, 0.85 (output loudness set to nominal level)
Common conditions		
Test item generation	4	Recordings analysed using the IVAS MASA C reference software in various configurations or using an in-house system.
Binaural rendering	1	Rendering with the IVAS MASA C reference binaural renderer.
Audio sampling frequency / bandwidth	2	48 kHz/SWB except for reference condition EVS@8 kbit/s which used 48 kHz/WB
Rating scale	1	DCR with instructions for binaural/spatial telephony
Languages	1	Finnish
Listening system	1	Sennheiser HD650 headphones for binaural presentation
Listening environment	1	No room noise

I.2.4 Instructions to listeners

The following set of instructions were given to all listeners as printouts. The original instructions are in Finnish, and they are here translated into English.

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 Audible but not annoying

3. Degradation is Slightly annoying

2. Degradation is Annoying

1. Degradation is Very 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.

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. In this way, the user is made aware of the nature of the samples and is instructed to evaluate the different dimensions at the same time. 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 test.

I.2.5 Results and observations

The graph shown in Figure I.4 presents the DCR MOS scores of the listening test with 95% confidence intervals.

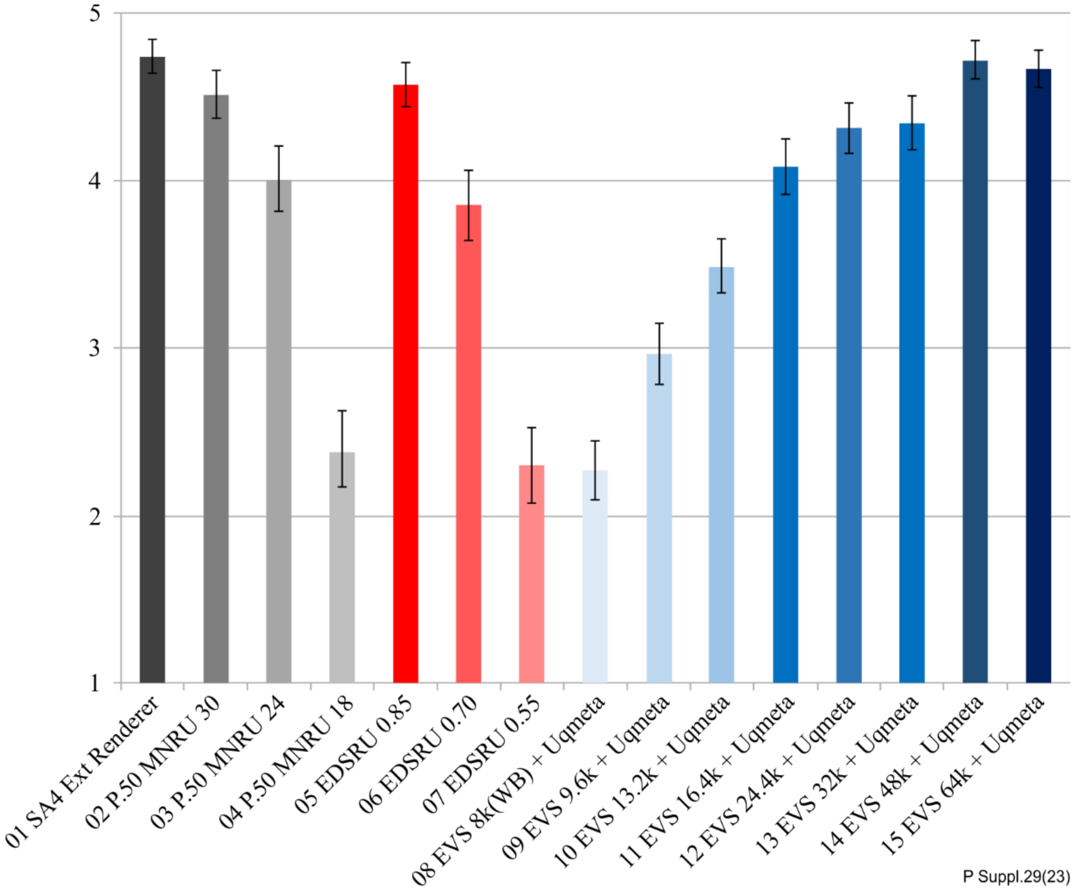


Figure I.4 – DCR MOS scores of the listening test with 95% confidence intervals

In this test, both the signal quality degradations and any spatial degradations of the reference conditions are caused by the encoding of the audio channels (using 3GPP EVS), whereas the spatial metadata itself is unquantized and therefore does not cause additional spatial degradation. Some saturation effect can be observed, including some indications relating to known characteristics of the EVS codec.

The CuT results are omitted from the presentation. It is however noted that the CuT conditions follow a similar curve based on the overall bit rate. For the CuT conditions, spatial degradations relate to both encoding of the audio and the lossy compression of the spatial metadata.

I.3 ITU-T P.800 DCR with higher-order scene-based audio

I.3.1 Test purpose

Listening testing of scene-based audio (up to 3rd order ambisonics) was performed to obtain insights about the suitability of DCR testing for higher-order scene-based audio.

I.3.2 Test outline

In the Figures I.5a, I.5b and I.5c, results from three ITU-T P.800 experiments conducted are shown. There is one clean speech experiment with 17 naïve subjects covering everything from EVS in mono to uncoded HOA3 (3rd order ambisonics). In addition, two noisy speech experiments are shown, one conducted with 14 naïve subjects, the other one with 10 experienced listeners.

The items are sentence pairs (non-overlapping) or triplets (overlapping) from different talkers for each sentence, where each talker is at a random spatial position in the audio scene. The item duration is ~6-7 seconds and the language of all sentence pairs is German. Each listener was exposed to each condition six times, thus the number of votes per condition is the number of subjects times six. Subjects were allowed to repeat each trial once.

The conditions can be described as follows:

- The DIRECT HOA3 signal (the reference) is the binauralised version of the uncoded HOA3 (3rd order ambisonics) signal.
- FOA (1st order ambisonics) and HOA0 (Mono) are generated by truncation of the higher order coefficients of the HOA3 signal.
- Anchors are either band-limited (7 kHz, as typically used in MUSHRA tests) or the previously discussed ESDRU, applied on the already binauralised output.
- EVS is used in mono or multi-mono configuration, using as input the HOA0 (Mono), FOA, or DIRECT HOA3 signal.

Instructions are the general ITU-T P.800 instructions as used during EVS testing, with a minor modification to hint to the listeners that the items are "stereo". Unfortunately not the full text as in clause II.2.3 of [ITU-T P.811], was used (which goes beyond mentioning stereo to also some further explanation: "Bitte beachten Sie dabei neben z.B. der Qualität der Sprache auch die Veränderung des Stereoklangs, z.B. die Breite des Stereoklangs oder wie gut die Position der Sprecher wiedergegeben wird.", i.e., example for stereo impairments such as width of the stereo image and reproduction of talker localization).

The noisy speech experiments used the same speech material as the clean speech test, mixed with HOA3 recordings of car, street, and office ambient noise at an SNR of 5, 10, and 10 dB, respectively. Those SNR numbers were picked so that speech is clearly understandable but also the noise can be well heard. All three noise types are part of the test, each noise is heard twice per condition.

The experiments with naïve subjects were run at the Fraunhofer IIS facilities in Erlangen, Germany, following the general ITU-T P.800 principles such as calibrated listening level in a controlled environment.

Experienced subjects were staff members of Fraunhofer IIS, which did mostly not have a background in codec development but at least a general understanding of multimedia and the concept of spatial and binaural audio. The listeners used their work equipment for listening, i.e., their generic headphones used for daily work connected to their PC without any explicit level calibration. Most experienced subjects did the test from home in an uncontrolled environment because of the pandemic.

It should be noted that there were more conditions in the experiment, they are however not relevant for this contribution and are omitted. The results are also not necessarily "final" because the tests did not reached the desired number of subjects (at least 16 were planned) because of the pandemic (at the time the tests were run) the tests could not get populated with the normal number of subjects within the available test slots.

I.3.3 Results

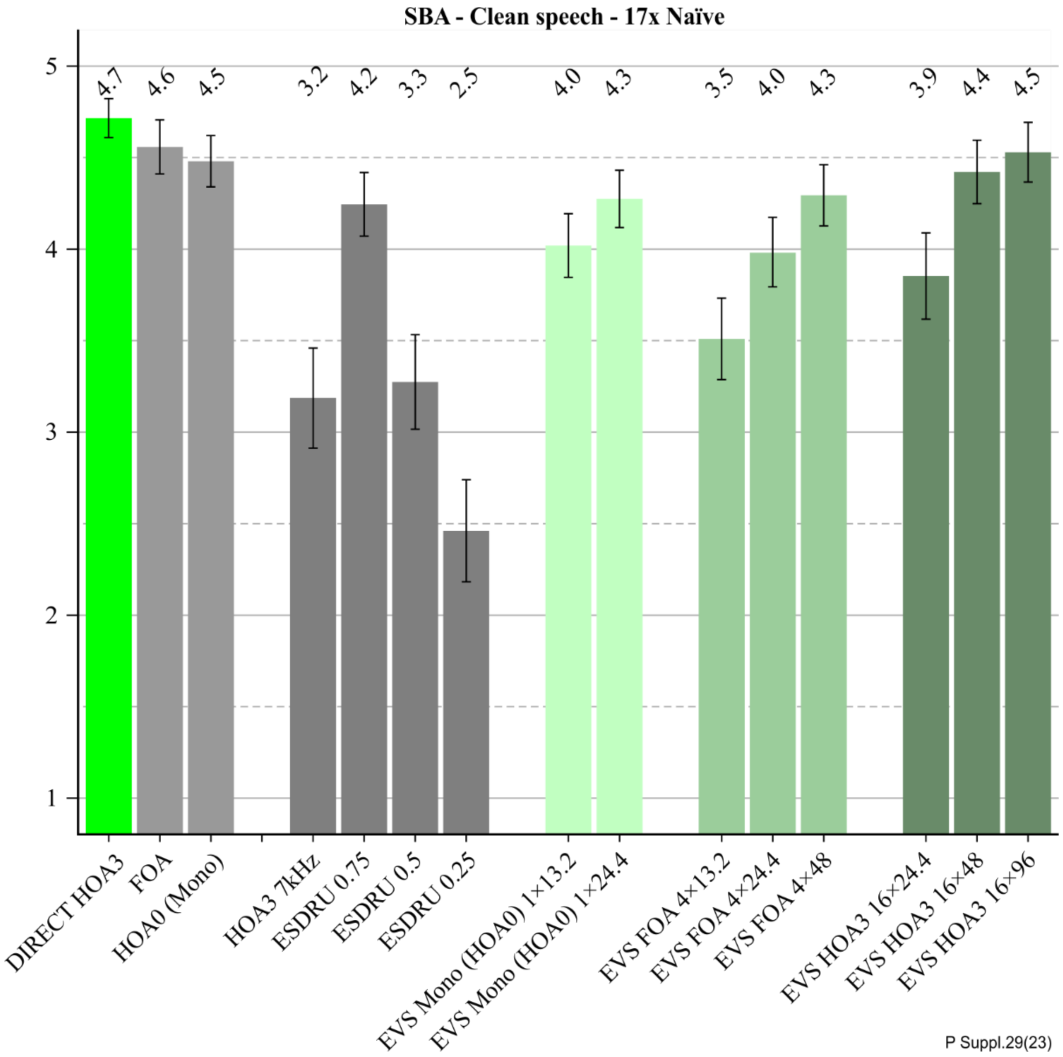


Figure I.5a – SBA – Clean speech – 17x Naïve

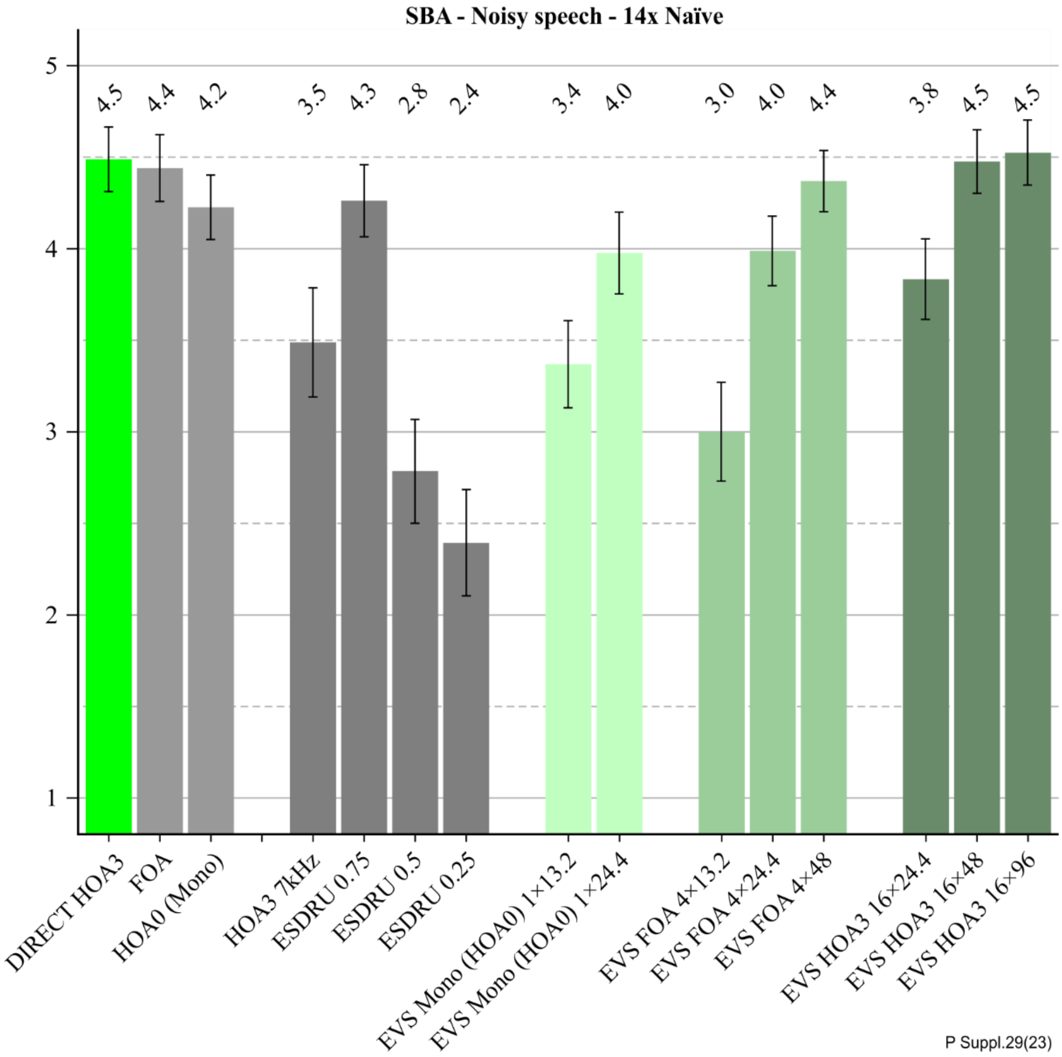


Figure I.5b – SBA – Noisy speech – 14x Naïve

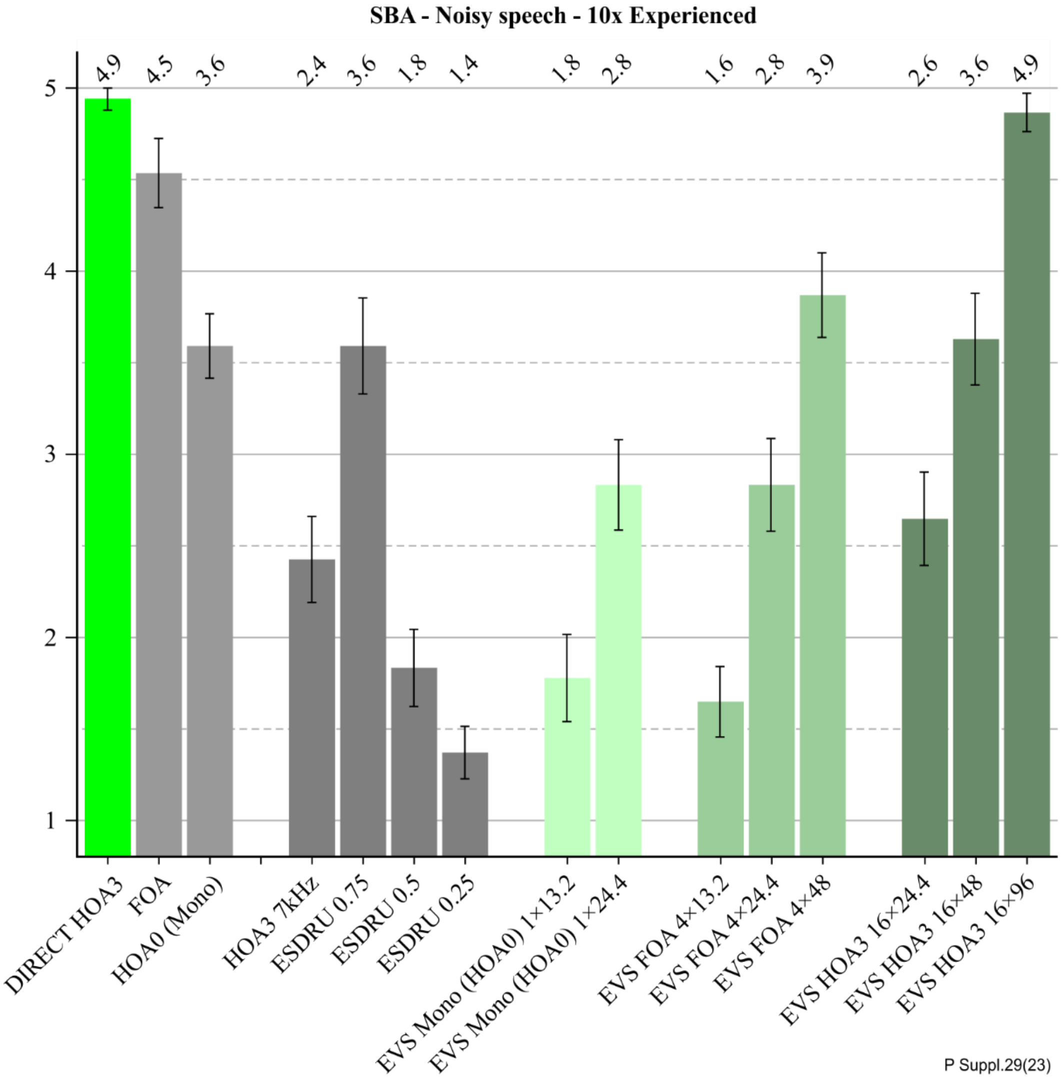


Figure I.5c – SBA – Noisy speech – 10x Experienced

I.3.4 Observations

While the results are somewhat expected in the sense that the conditions with less signal distortion get better scores, there was one unexpected observation that can be considered concerning: Naïve subjects can hardly (or are unable to) differentiate HOA3 from FOA and even HOA0, which is a non-spatial mono signal. However the three uncoded reference conditions were very well differentiated by experienced listeners.

Other observations from the test can be summarized as follows:

- EVS mono as a reference codec in multi-mono configuration in general works and scales with bit rate
- EVS mono at lower rates in higher complexity multi-mono configuration performs worse than the same EVS rate use for a lower complexity multi-mono configuration or mono
- EVS mono at higher rates scales towards the uncoded reference and also exposes the differences in terms of spatial resolution
- The ESDRU currently proposed for ITU-T STL can be used for binauralised content
- Experienced listeners tend to use the full voting scale, while naïve subjects rarely consider items to be "very annoying".

I.4 Comparison of DCR test experiments for FOA and HOA3 input in 7.0+4 and binaural listening set-ups

I.4.1 Test purpose

The goal of this evaluation is to contribute to the collective experience with immersive listening testing using naïve listeners. A listening test using an artificially created database was run twice, using a 7.0+4 loudspeaker listening set-up and a binaural listening set-up. A third test was run using a live-recorded database and 7.0+4 loudspeaker listening set-up. The tests were run in the immersive listening laboratory of the university of Sherbrooke, in North-American French.

I.4.2 Test outline

I.4.2.1 Artificially created audio database

- Artificially created spatial samples from mono recordings adjusted to −26 dBOvl.
- Phonetically balanced single sentences concatenated into pairs with similar meaning.
- 4 male and 4 female talkers, always a male and a female talker in a stereo 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 placed at the nominal height at different configurations using regular pattern:
  - 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
  - Instrumental music at 15 dB SNR.
  - Different music sample and position used for each speech sentence pair.
  - Elevation: 20°, 40°, 60° 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]

I.4.2.2 Live-recorded audio database

- Real-life recorded samples using Zylia ambisonic microphone (HOA3). The recording was done in a large and rather reverberant room.
- The audio scenes were more complex than in case of the artificially created database as some talkers were moving.
- On the other hand, the sound distribution was more limited spatially than in the case of the artificially created database, covering only a part of the whole range of azimuth and elevation.

− Phonetically balanced single sentences concatenated into pairs with related meaning.

− 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. The targeted overlap was about 30% of a sentence.

− Length of the samples – 6 or 7 s.

− 48 kHz sampling rate.

− HOA3 and FOA input codec format.

− Talkers scenarios

  1. 2 talkers sitting at a table, at 90° separation with respect to the microphone. The utterances were non-overlapping.
  2. 2 talkers sitting at a table, face-to-face, with the microphone in between. The utterances were overlapping.
  3. 1 talker sitting at a table, second talker walking around the table. The utterances were non-overlapping.
  4. 2 talkers sitting at a table at 90° separation, one talker stands up and sits down while talking. The utterances were overlapping.
  5. 2 talkers walking side-by-side around the table. The utterances were non-overlapping.
  6. 2 talkers walking around the table in opposite directions. The utterances were overlapping.

– Background

– Reproduced stereo music with speakers in a usual stereo set-up. The level of the background music was set up experimentally to be perceptually at similar level as in the case of the artificially created database.

– Different music was used for each speech sentence pair.

I.4.2.3 Test set-up

– ITU-T P.800 DCR test, instructions mentioning spatial aspect.

– 4 categories corresponding to the different talker pairs.

– 6 panels, each using different audio samples.

– Naïve listeners.

– 4 listeners per panel (one listener at a time in the loudspeaker set-up).

– 24 listeners in total.

– Each condition was evaluated 24 × 4 = 96 times.

– Anchors – ITU-T P.50 MNRUs (modulated noise reference units) [ITU-T P.811] – applied coherently (using the same seed) to all ambisonic channels

– 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).
  • Rendering was done on concatenated files.

– Level adjustment

- The level was adjusted to −26 LKFS.
- The direct signal level was first measured on the 7.0+4 signal using [ITU-R BS.1770] 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.

– Loudspeaker listening set-up – 7.0+4 Genelec SAM 3031 speaker set-up 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

I.4.2.4 Instructions to listeners

The following instructions to listeners were designed to emphasize the immersive aspect of the test. In particular, the instructions use the word *Alteration* instead of the word *Degradation* to avoid confusion of some listeners as to whether reducing spatial image should be considered as a degradation.

INSTRUCTIONS TO NAÏVE LISTENERS FOR IMMERSIVE 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 was possibly altered after it has passed through a telecommunication system.

Your task is to evaluate the overall alteration of the second sample compared to the first sample, comprising both alteration of the sound quality (e.g., due to additional noise, roughness, clicks or other distortions), and alteration of 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, please record your **overall** opinion about the amount of any alteration you can perceive in the second sample relative to the first sample using the following rating scale:

5 – Alteration is not audible

4 – Slight alteration is sometimes audible

3 – Alteration is audible

2 – Alteration is clearly audible

1 – Strong alteration is clearly audible

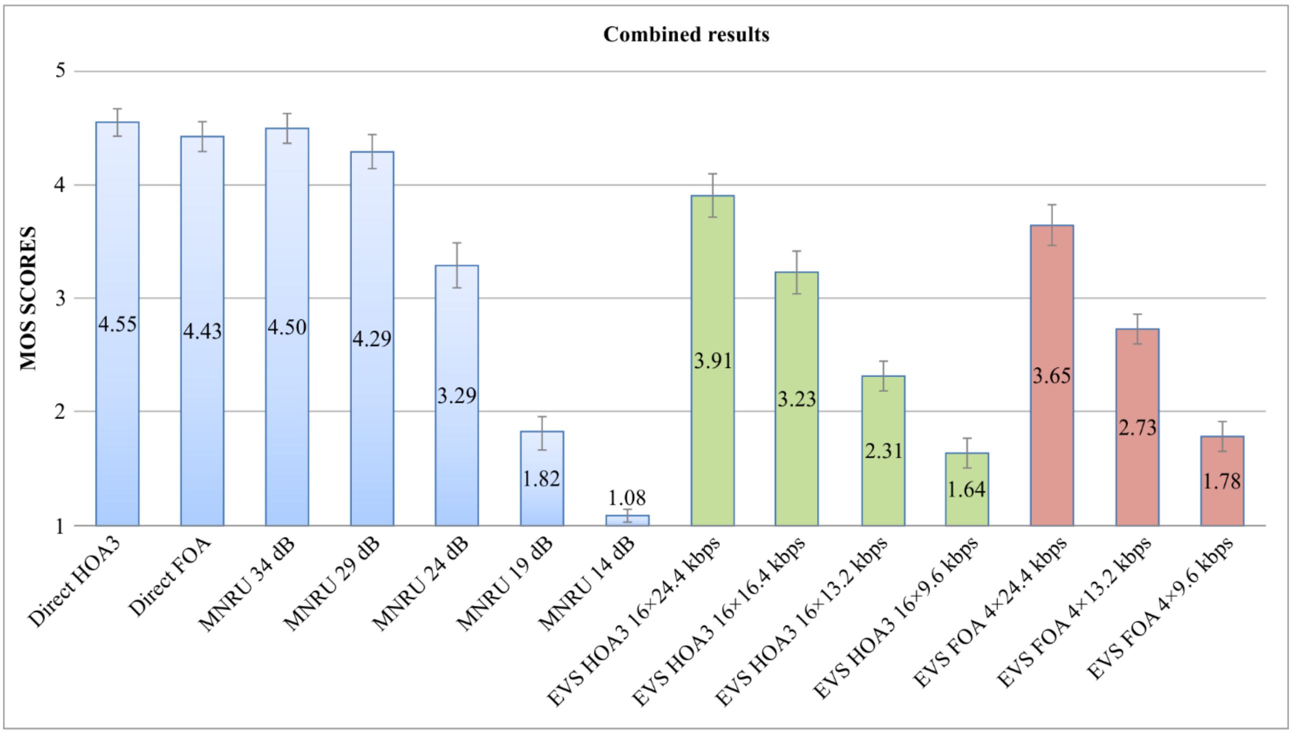
Note that the level of alteration present in different *test* samples is expected to span the complete range of the rating scale during the experiment.

Please do not discuss your opinions with other listeners participating in the experiment. If you have any questions, please ask the test administrator.

I.4.3 Test results

I.4.3.1 Loudspeaker listening test results for artificially created database

Figure I.6 shows loudspeaker listening test results for an artificially created database.

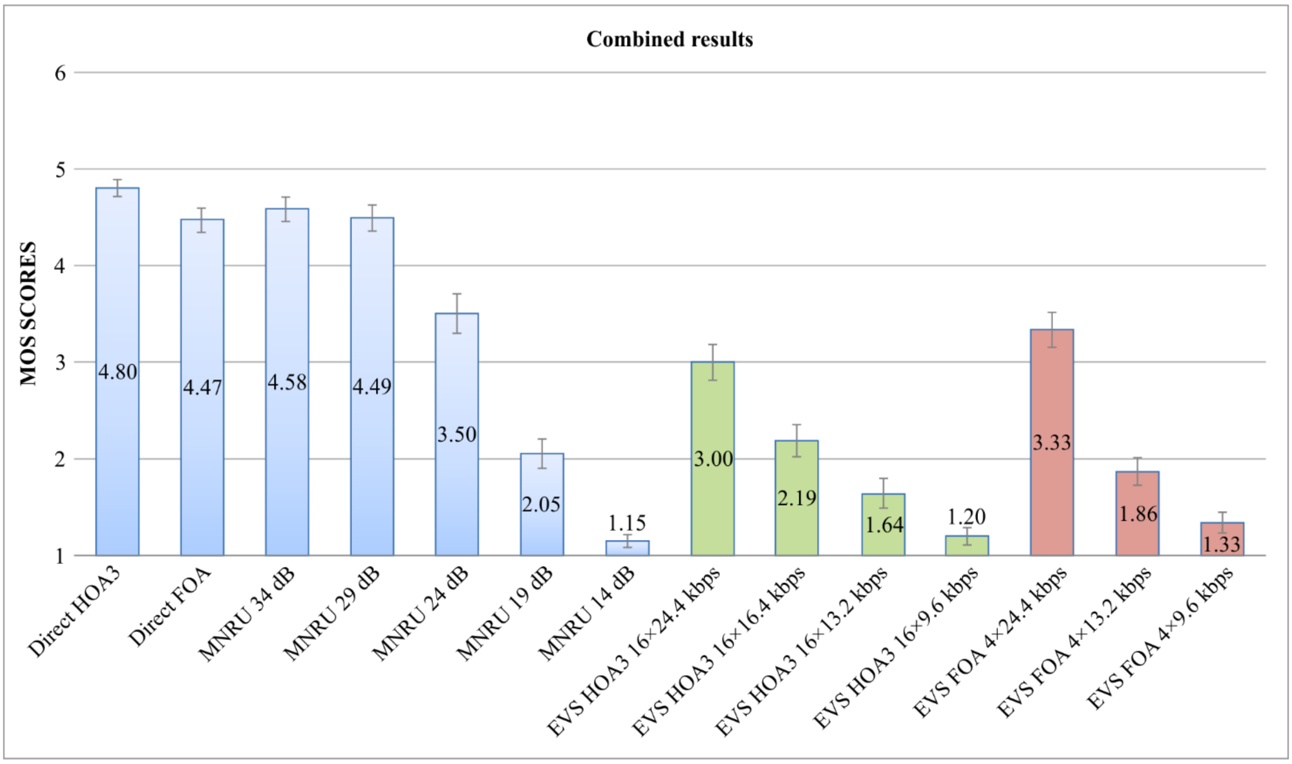


P Suppl.29(23)

Figure I.6 – Loudspeaker listening test results for artificially created database

I.4.3.2 Binaural listening test results for artificially created database

Figure I.7 shows binaural listening test results for artificially created database.

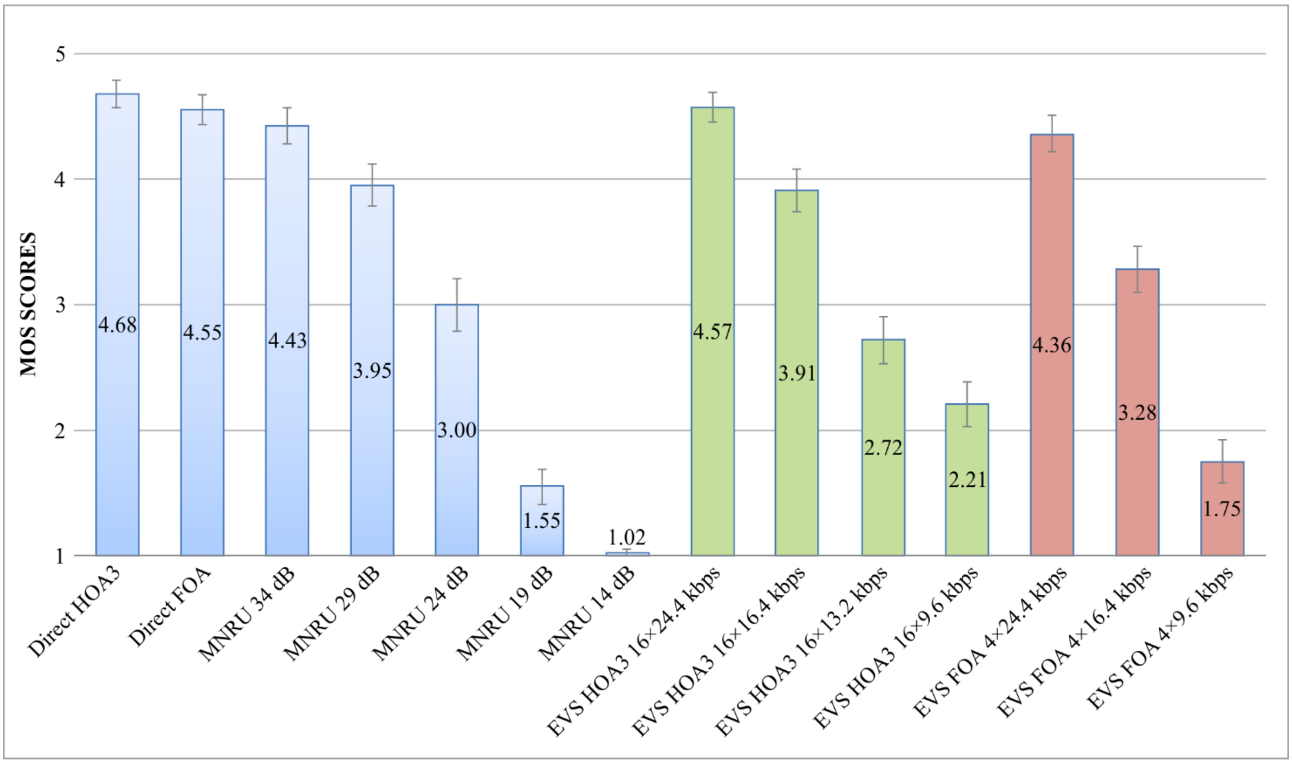


P Suppl.29(23)

Figure I.7 – Binaural listening test results for artificially created database

I.4.3.3 Loudspeaker listening test results for live-recorded database

Figure I.8 shows loudspeaker listening test results for live-recorded database.



P Suppl.29(23)

Figure I.8 – Loudspeaker listening test results for live-recorded database

I.4.3.4 Screening of listeners

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".
- vote, in average, the direct condition better than or equal to the MNRU 29 dB condition. 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 was below 4.
- vote, in average, the MNRU 29 dB condition better than the MNRU 24 dB condition.
- vote, in average, the MNRU 24 dB condition better than the MNRU 19 dB condition.
- vote, in average, the MNRU 19 dB condition better than the MNRU 14 dB condition.

I.4.4 Observations

- One loudspeaker plus one binaural test took about 2 weeks.
- Overall, naïve listeners could reliably detect coding deficiencies.
- EVS multi-mono seems to be a good reference, able to cover practically the whole range of perceptual quality.
- At low bitrates, more ambisonic channels seem to degrade the perceptual experience rather than improve it when the artificially created database was used.
- 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.
- Despite clear and explicit instructions, and standard DCR voting labels used in the listening software interface, some listeners still did not understand the task. We collected for instance, the following comments:
  - "I did not vote 1 or 2 as I was always able to understand the meaning".

  - "As I did not vote 1 during the training session, I decided not to vote 1 during the test session either".

I.4.4.1 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 HO3 and FOA direct conditions better than in the loudspeaker listening.
- For multi-mono EVS processing, at 24.4 kbit/s/channel, an advantage for FOA input is observed in binaural listening, but the opposite tendency is observed for 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.

I.4.4.2 Comparisons of results for the live-recorded database and the artificially created database

- Addition of ambisonic channels (i.e., HOA3 vs FOA) consistently increased the perceptual experience for all bitrates.

I.4.4.3 Issues, potential areas of improvement

- Leakage of intermittent noise into the listening room.
- For the first two tests (with an artificially created database), randomization did not take into account the regular pattern of spatial distribution of the samples. This issue was considered in the design of the 3rd test (live-recorded database).
- For the first two tests (with an artificially created database), instructions were presented only verbally; no instruction sheets were provided to the listeners. For the 3rd test using a live-recorded database, a sheet with instructions was given to all listeners.
- Recalibration of the loudspeaker listening room using the Genelec GLM system could be done before each test.
- Pivoting chair in the immersive loudspeaker room – some listeners were turning on the chair.

I.4.5 Conclusions

- With some adjustments, the DCR test with naïve listeners seems to be a good trade-off between accuracy and efficiency.
- More explicit initiation of naïve listeners to spatial aspects seems 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.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Published in Switzerland
Geneva, 2023