]Grant Agreement No.: 687645
Research and Innovation action
Call Topic: H2020 ICT-19-2015

Object-based broadcasting – for European leadership in next generation audio experiences

D3.5: Specification and implementation of reference audio processing for use in content creation and consumption based on novel broadcast quality standards

Version: 1.5

Deliverable type	DEM (Demonstrator, pilot, prototype, plan designs)
Dissemination level	PU (Public)
Due date	31/05/17
Submission date	09/06/17
Lead editor	Andrew Mason (BBC)
Authors	Andrew Mason (BBC), Tilman Herberger (Magix), Emanuël Habets (FHG), Matthias Geier (IRCAM), Anderas Silzle (FHG), Marius Volpel (Magix), Nicolas Epain (bcom), Niles Bogards (Ecandy), Michael Meier (IRT),
Reviewers	Werner Bleisteiner (BR), Uwe Herzog (EURES)
Work package, Task	WP 3, T3.1, T3.2, T3.3, T3.4
Keywords	

Abstract

There are several fundamental signal processing software blocks in the ORPHEUS object-based broadcasting (signal) chain. This document describes those that have been created already, the operating parameters where these have been established but the software not yet written. In addition, broad functional descriptions are given of some processes that might be desirable, but for which there are no definite plans within the ORPHEUS project. The most complex processes are microphone array signal processing, and the rendering of objects to loudspeaker signals.

Document revision history

Version	Date	Description of change	List of contributor(s)
V0.1	4/4/17	Creation of document outline	Andrew Mason
V0.3	11/4/17	Updated during webex; placeholders for allocated tasks	
V0.4	12/4/17	Additional text for sections 2, 4, 5 and 5	Andrew Mason
V0.5	13/4/17	Additional text on VST, section 3.1; corrections to cross-references	Tilman Herberger
V0.6	27/4/17	Additional text from Magix in §3, BBC in §3 and §4, FhG in §2, and IRCAM in §3	Andrew Mason, Emanuël Habets, Matthias Geier
V0.61	27/4/17	Additional text on personalisation, control room design and emission	Andreas Silzle
V0.7	28/4/17	Incorporation of references in V0.61	Andrew Mason
V0.8	28/4/17	Modifications immediately following the meeting on the 28/4; Magix in §3, BBC in §6	Marius Vopel, Andrew Mason
V0.9	2/5/2017	Addition of HOA mic. processing	Nicolas Epain
V1.0	2/5/2017	Executive summary added; tentative conclusions added; extra text for overview; binaural rendering in §6; corrections to §3.1 rendering	Andrew Mason, Marius Vopel
V1.1	22/5/2017	Removed reference to incomplete speculative text on personalisation based on MPEG-H profiles	Niels Bogaards
V1.2	23/5/2017	Results of object-based loudness measurement study; revision of control room design section; revision of personalisation in §6.3 "Consumption"	Michael Meier, Andrew Mason
V1.3	29/5/17	Deliverable review and final editing	Werner Bleisteiner, Uwe Herzog
V1.4	30/5/17	Changes and suggestions from reviewers incorporated	Andrew Mason
V1.5	01/6/17	Move most content from section 2.1 and 2.2 to D3.3	Andreas Silzle

Disclaimer

This report contains material which is the copyright of certain ORPHEUS Consortium Parties and may not be reproduced or copied without permission.

All ORPHEUS Consortium Parties have agreed to publication of this report, the content of which is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License[1].

Neither the ORPHEUS Consortium Parties nor the European Commission warrant that the information contained in the Deliverable is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using the information.

Copyright notice

[1] http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en_US

Executive Summary

This Deliverable specifies the reference audio processing for use in the ORPHEUS content creation and consumption chain, and documents the implementation of the respective tools required for that. ORPHEUS' object-based broadcasting signal chain is described as 5 links: capture, editing and mixing, distribution, provision, and reception.

In the *"capture"* part, there is considerable signal processing involved in converting microphone signals from compact microphone arrays into something suitable for handling by the production system. The use of RFID tags to improve the reliability of metadata and the speed and accuracy of configuring the processing is also described.

In the *"editing and mixing"* link, tools have been developed for recording, replaying, verifying and extracting ADM metadata. Rendering of object-based audio in the production workstation is expected to go through a small number of phases, depending on, for example, the availability of standardised renderers from the ITU-R.

Constraints on production techniques using personalisation are expected to draw on experience from previous work, and be evaluated during the pilots. Studies on intelligibility by organisations outside the project are being monitored such that the incorporation of measurement tools from them into the Orpheus chain might take place.

Control room design has to evolve, in particular to accommodate more loudspeakers, and more flexible monitoring.

In the *"distribution"* link, it is anticipated that there will be 'object funnelling', whereby the number of objects has to be reduced. The situation has not yet arisen where number of simultaneous objects in different positions exceeds the capacity available, but it is acknowledged that this must be handled, or prevented.

Legal compliance, and the requirements of archiving are relatively simply addressed in ORPHEUS.

In the *"provision"* link, most signal processing is defined by international standards, substantiated with formal test data. Work by ORPHEUS members continues in the ITU-R to improve the subjective test methods to be more suitable for assessing the performance of object-based spatial audio systems, and thereby provide additional guidance for setting up provision (coding) systems.

In *"reception"* the main flexibility is in personalisation. Work is still taking place to define the behaviour of devices in response to listener and broadcaster personalisation inputs (such as dynamic range control).

Table of Contents

Executive Summary 3

Table of Contents 4

List of Figures 6

List of Tables 7

Abbreviations 8

1 Overview of audio processing in content creation and consumption 9

2 Capture 10

2.1 3D Audio Capture with a horizontal circular array 10

2.2 Ambient sound capture using a spherical microphone array 10

2.3 Equalisation of HOA signals acquired with microphone arrays 11

2.4 Use of RFID tags to associate accurate metadata with objects 12

2.5 Use of RFID tags to configure default audio processing for a presenter 13

3 Editing and mixing 14

3.1 Rendering used in DAW 14

3.2 IRCAM ADM tools 15

3.2.1 ADM Player/Renderer 15

3.2.2 ADM Recorder + Monitoring 17

3.2.3 ADM ExtractXML 19

3.2.4 ToscA 19

3.3 Loudness control 19

3.3.1 Object-based loudness measurement 20

3.3.2 Higher-Order Ambisonics loudness measurement 24

3.4 Personalisation constraints 24

3.5 Intelligibility 25

3.6 Control room design 25

4 Distribution 27

4.1 Object funnelling 27

4.2 Recording of transmission for legal compliance 27

4.3 Archiving of finished programmes 28

5 Provision 29

6 Reception 30

6.1 Decoding 30

6.2 Rendering 30

6.3 Personalisation 30

7 Conclusions 32

References 33

List of Figures

Figure 1: Overview of ORPHEUS implementation 9

Figure 2: Example of a cylindrical microphone array 10

Figure 3: Block diagram of the 3D audio capturing algorithm 10

Figure 4: Block diagram of the ambient sound capturing algorithm 11

Figure 5: Average amplitude spectrum of the Ambisonic signals acquired with a Sennheiser AMBEO microphone, using filters based on calibration measurements and on a theoretical model 12

Figure 6: Main user interface of the ADM Player/Renderer 16

Figure 7: Scene display of the ADM Player/Renderer playing the file machine-aer_bwf.wav [5]. Left pane: view from above; Right pane: view from behind 17

Figure 8: Main window of the ADM Recorder 18

Figure 9: Metadata editor and routing matrix 19

Figure 10: Schematic view of generic object-based loudness measurement with position-dependent weighting 20

Figure 11: Schematic view of generic object-based loudness measurement without position-dependent weighting 21

Figure 12: Schematic view of an adaptive object-based loudness measurement with weighting based on a target setup-dependent loudness signature 21

Figure 13: Example of a loudness signature with weighting factors for each source position based on target loudspeaker setup and rendering method. Original high resolution version (left), and quadtree compressed example (right) 22

Figure 14: Boxplots of the deviations between the loudness variants tested and a measurement of the resulting loudspeaker signals in accordance to ITU-R BS. 1770 for completely random audio scenes with white noise objects. Boxes show median, quartiles, minimum, maximum, and arithmetic mean (cross). 23

Figure 15: Boxplots of the deviations between the tested loudness variants and a measurement on the resulting loudspeaker signals in accordance to ITU-R BS. 1770 for real audio scenes with natural audio signals. Boxes show median, quartiles, minimum, maximum, and arithmetic mean (cross). 23

Figure 16: Comparison of loudness estimations of 8 sound scenes encoded in the HOA format using an existing HOA loudness algorithm and the algorithm developed during the Orpheus project 24

List of Tables

Table 1: Levels for low complexity profile of MPEG-H Audio 29

Table 2: Recommended core bitrates for excellent audio quality for broadcast applications [13] 29

Abbreviations

AAC	Advanced Audio Coding
ADM	Audio Definition Model
ATSC	Advanced Television Systems Committee
BW64	Broadcast Wave 64 bit
DOA	Direction of Arrival
DRC	Dynamic range compression
DVB	Digital Video Broadcasting (an industry-led consortium of digital TV and technology companies)
EBU	European Broadcasting Union
EPG	Electronic Programme Guide
HOA	Higher Order Ambisonics
HRTF	Head Related Transfer Function
ITU	International Telecommunications Union
MATLAB	Matrix Laboratory – computing environment and programming language
MPEG	Moving Picture Expert Group
NGA	Next Generation Audio
OSC	Open Sound Control
RDF	Resource Description Framework
RFID	Radio-Frequency Identification
STFT	short-time Fourier transform
UHDTV	Ultra High Definition Television
VBAP	Vector base amplitude panning
VST	Virtual Studio Technology (audio software interface, created by Steinberg)

1 Overview of audio processing in content creation and consumption

This Deliverable specifies the reference audio processing for use in the ORPHEUS content creation and consumption chain, and documents the implementation of the respective tools required for that.

A large part of the signal processing required for the ORPHEUS' object-based broadcasting chain has now been defined, either by use of open international standards, or as a result of development by the partners. Tools have been created to record and replay ADM metadata, and it is possible to record, edit, and reproduce object-based audio in the studio to a high quality. These fit into the implementation, which is shown in Figure 1.

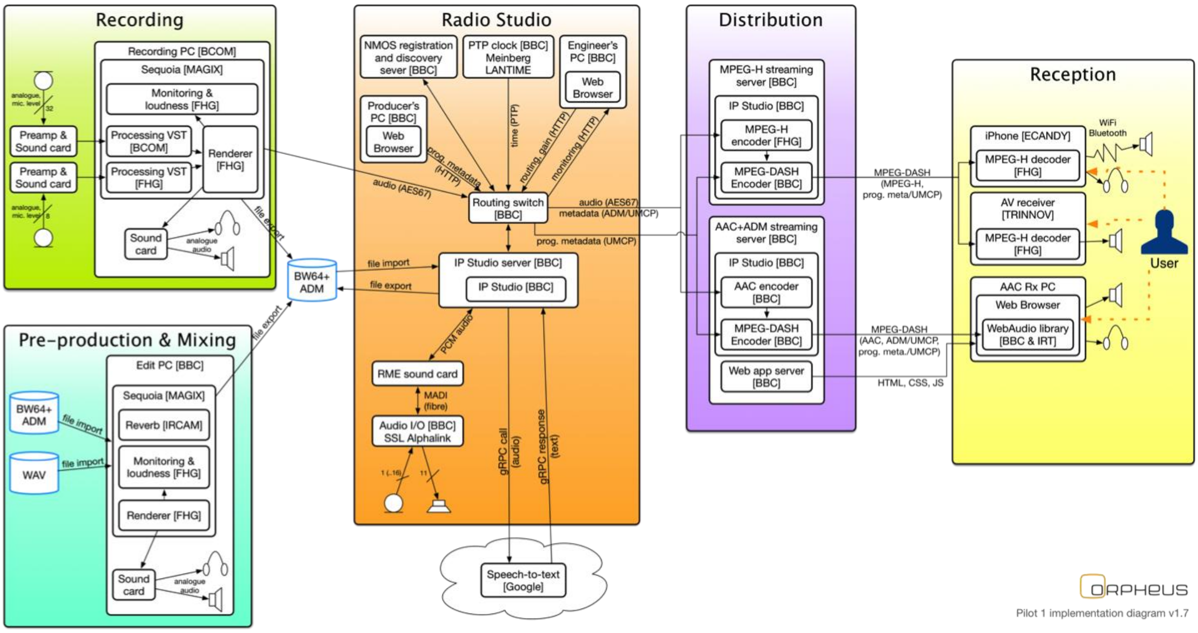


Figure 1: Overview of ORPHEUS implementation

There are five stages within the content creation and consumption process in which significant audio processing is expected. These are as follows:

Capture (signal acquisition)

Editing and mixing (producing the programme from the acquired signals)

Distribution (the programme path from studio to transmitter)

Provision (transmission or delivery to the audience of the programme)

Reception (listening by the audience)

The remainder of this document is structured along these five stages.

In capture, it is the conversion of microphone signal formats that is significant; in editing and mixing it is the creation and verification of metadata (including personalisation controls), and the fidelity of rendering that matters. Distribution and provision have requirements to maintain the technical quality that is necessary for the destinations of the signals (to the audience, and to the archive). Finally, on reception, although a lot of the signal processing is defined by existing new standards, there is work still to be done to explore how personalisation and interaction can, and should, be used to improve the listeners' quality of experience.

2 Capture

2.1 3D Audio Capture with a horizontal circular array

Explanations to the capturing with the circular microphone array in Figure 2 can be found in Deliverable 3.3.

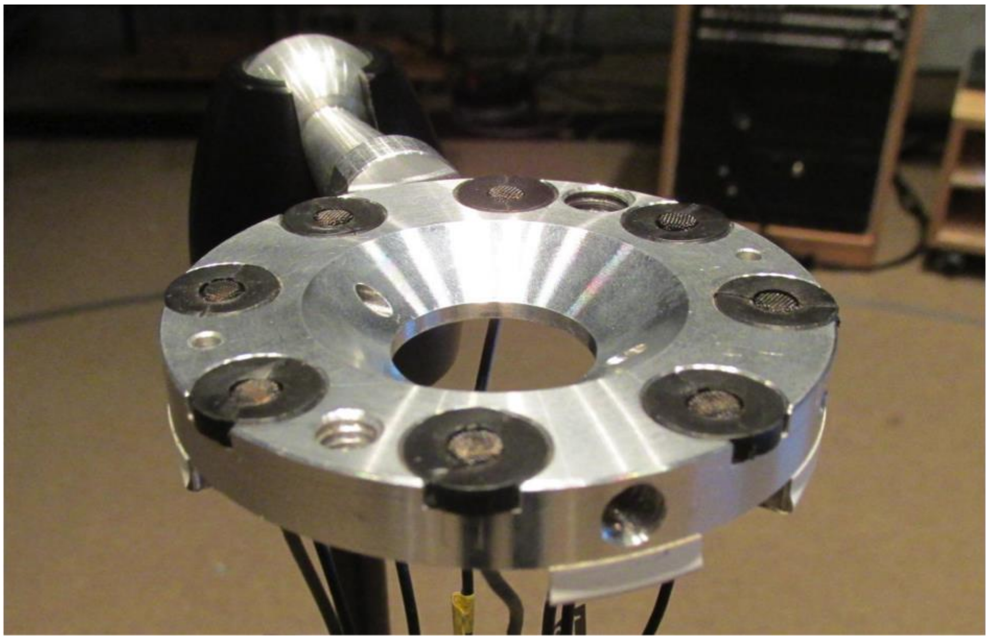


Figure 2: Example of a cylindrical microphone array

Also the parametric spatial sound processing depicted in Figure 3 is explained in Deliverable D3.3.

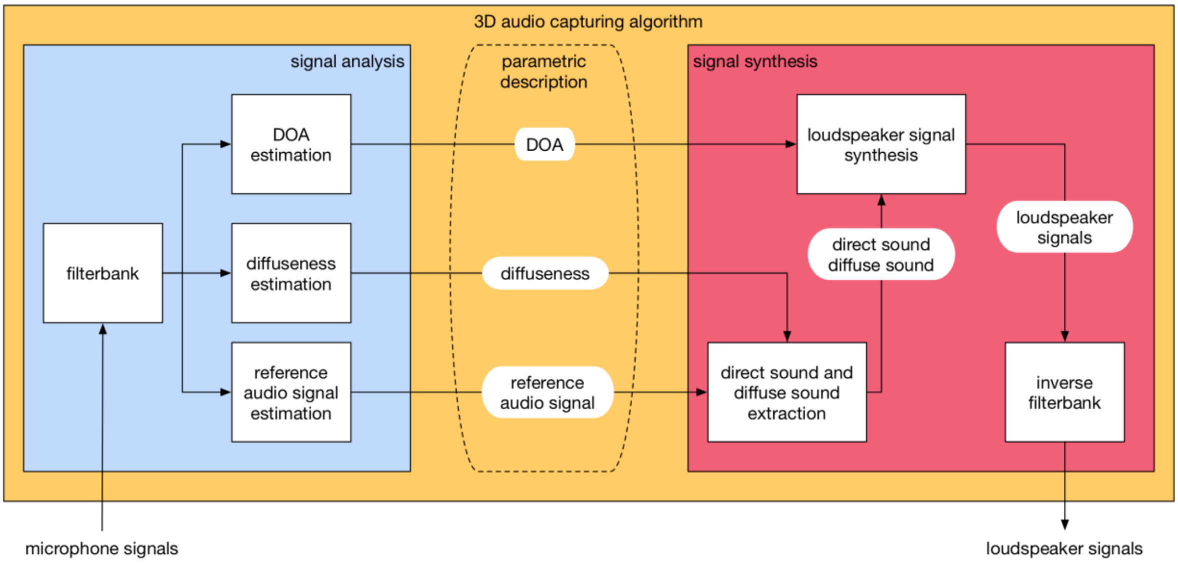


Figure 3: Block diagram of the 3D audio capturing algorithm

2.2 Ambient sound capture using a spherical microphone array

More details to the ambient sound capturing algorithm of Figure 4 can be found in Deliverable D3.3.

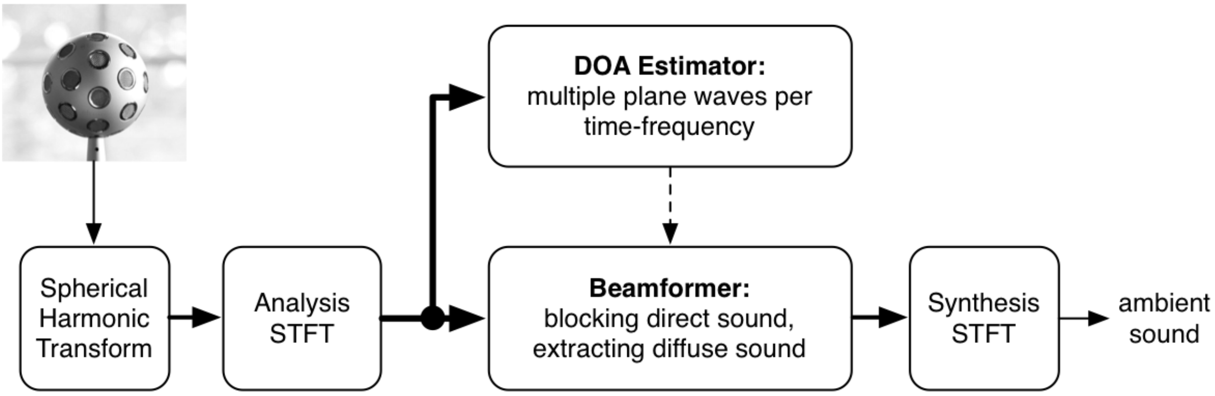

DOA Estimator:
multiple plane waves per time-frequency
Spherical Harmonic Transform
Analysis STFT
Beamformer:
blocking direct sound, extracting diffuse sound
Synthesis STFT
ambient sound

Figure 4: Block diagram of the ambient sound capturing algorithm

2.3 Equalisation of HOA signals acquired with microphone arrays

A frequent complaint with compact microphone arrays, and a major obstacle to their use in the production of audio content for broadcasting, is the quality or timbre of the recorded signals. This is primarily caused by the use of digital filters which convert the microphone signals to speaker or Higher-Order Ambisonics (HOA) signals. These filters are typically calculated with the hypothesis that the microphone array fits a simple mathematical model. For instance, in the case of rigid spherical microphone arrays (e.g. the Eigenmike®), the microphones are typically assumed to be dimensionless and baffled by a perfectly rigid sphere. However, calibration measurements show that the acoustic behaviour of such microphone arrays is significantly different from that predicted using the model. As the differences between the modelled and actual characteristics of the microphone array vary across frequencies, so does the overall amount of energy in the acquired signals. Equalisation problems can also result from a filter design that does not take into account the phenomenon of spatial aliasing properly.

In order to design HOA-encoding filters (filters converting microphone signals into HOA signals) that ensure an optimal equalisation, b<>com conducted calibration measurements for three commercially available microphone arrays: Sennheiser's AMBEO[2], Embrace Cinema's Brahma and MH Acoustics' Eigenmike. For each of these three microphone arrays, impulse responses were measured for hundreds of source directions in a facility designed for measuring Head Related Transfer Functions. The directional measurements were then used to design specially tailored HOA-encoding filters with the following process. First, for each frequency, a set of filters was calculated by solving the optimisation problem:

$$\text{minimise } |B - EG|_F \ ,$$

where B denotes the matrix of the expected HOA signals for every source direction, G denotes the matrix of the microphone frequency responses for the same directions, E denotes the matrix of the encoding filter frequency responses and $|.|_F$ denotes the Frobenius norm. Second, the amplitude of the filters is modified such that the overall energy of the sound scene represented by the HOA signals, averaged over every source direction, remains constant. This is equivalent to making sure that the energy of the recorded scene remains constant in the presence of a diffuse sound field.

[2] AMBEO is a registered trademark of Sennheiser.

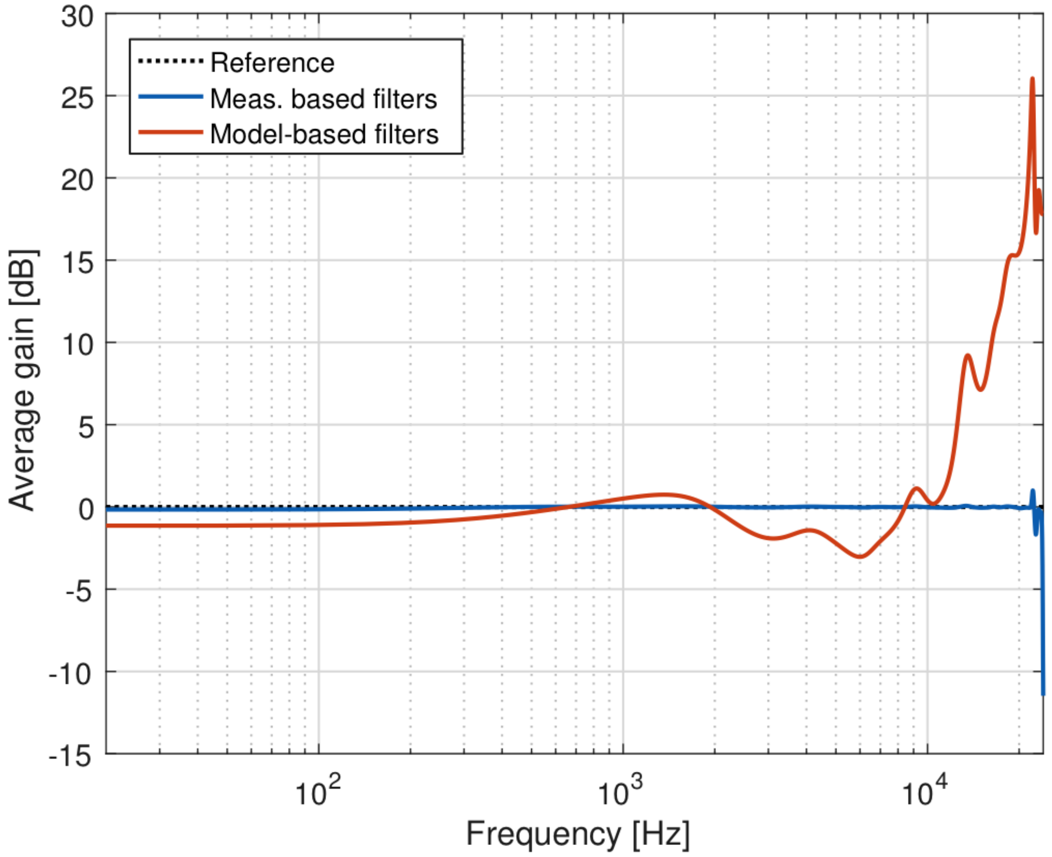


Figure 5: Average amplitude spectrum of the Ambisonic signals acquired with a Sennheiser AMBEO microphone, using filters based on calibration measurements and on a theoretical model

The result of this process for the case of the AMBEO microphone is illustrated in Figure 5. The orange curve represents the average sound field energy obtained when using the digital filters provided by the manufacturer in the VST plugin accompanying the microphone array. Major fluctuations can be observed: in particular, the energy is significantly higher than expected above 10 kHz. On the other hand, the sound energy obtained with the filters optimised using calibration measurements (blue curve) remains nearly constant over the entire frequency range, with the exception of minor fluctuations caused by the conversion of the filter frequency responses to the time domain.

2.4 Use of RFID tags to associate accurate metadata with objects

A system has been developed for use in the studio set up in BBC Broadcasting House whereby presenters can use their BBC ID card to indicate which microphone they are using.

A typical radio talk studio will have several microphones on a circular table, or on either side of a desk. Knowing which channel on the mixing console is associated with which microphone, and which microphone with which contributor, has always mattered, but object-based production brings new possibilities. Because the signals from the microphones are not necessarily mixed together, but rather are kept as separate objects, it is important to capture and preserve reliable identification information. This ensures that objects can be processed correctly throughout the production process, not only for rendering, but for linking to data in the EPG, and beyond into a “*resource description format*” (RDF).

BBC ID cards use a standard RFID protocol. Within the Orpheus project, an off-the-shelf card reader has been used to retrieve data from them, which has then been used to retrieve the corresponding identity information from the corporate database. This information can then be added to the metadata associated with the audio object. The advantage of this is that the ID card provides a unique identifier, avoiding confusion between contributors with the same or similar names. This increases the quality of the metadata.

2.5 Use of RFID tags to configure default audio processing for a presenter

It is commonplace for a presenter of a radio programme to have their own preferred settings for voice processing (equalisation and dynamics compression). The use of RFID described in Section 2.4 can also be used to configure default processing that is to be performed in the mixing console on the signal from their microphone (as opposed to being signalled by metadata).

Automatically configuring this processing reduces the possibility of errors that lead to lower audio quality.

The processing to that might be applied is beyond the scope of the ORPHEUS project, and this function has not yet been implemented.

3 Editing and mixing

3.1 Rendering used in DAW

The simplest form of renderer would be point source rendering using vector base amplitude panning (VBAP). The technique is described in Ville Pulkki's paper "Virtual Sound Source Positioning Using Vector Base Amplitude Panning" [1]. The BBC has implemented this technique in its own software library.

Further developments in rendering are anticipated. Studies in the ITU-R are expected to yield a specification of technical requirements for renderers to be used during programme production, and a set of standardised renderers that meet those requirements. A fundamental requirement will be that the renderers accept metadata input according to the ITU-R Audio Definition Model (ADM) [2]. Unfortunately, it is not possible to predict whether this study will be completed in time for its results to be incorporated into ORPHEUS.

It had been thought that a standardised renderer would also be used as reference in subjective evaluation of other renderers. The scope of the ITU-R study has had to adapt as a result of pressure from some participants, and this is no longer part of the scope of the current plan of work. However, work continues in ITU-R on the development of a subjective test method that does not require a reference. A working document now exists, enabling the method to be tried, but it is not clear whether or not the method will be available as a published recommendation before the end of the ORPHEUS project.

One anticipated outcome of the ITU-R renderer development study is a set of three algorithms that may be implemented, and an indicator to be conveyed in ADM metadata to show which algorithm has been used in the production process. A renderer from the set might be expected to indicate whether it is, or is not, the same renderer as is indicated by the metadata in the stream of data that it is rendering. It is not clear at this stage how interchangeable the different renderers in the set will be.

VBAP is not the only rendering technique considered suitable for use in ORPHEUS. An alternative exists in the MPEG-H 3D audio specification [3]. An implementation of the renderer in compliance with the standard has been provided by FHG and is being integrated into the Sequoia DAW by Magix as an internal VST plugin.

Metadata relating to the audio definition model published by ITU-R [2] such as gain, panning in 3 dimensions, language selection, and foreground/background balancing is prevented from being processed by the internal Sequoia audio engine. Instead of that, in each Sequoia track an instance of the renderer plugin collects the ADM metadata and sends it to the master instance, which is an insert in the surround master bus.

Here all metadata is processed and an audio signal is created that includes all the ADM effects, e.g. panning to various surround loudspeaker setups (switchable to 4+7+0[3], 4+5+0[3], 7.1, 5.1, or 2.0) and realizing the interactivity features.

[3] Notation from Recommendation ITU-R BS.2051[4] is used here. A ".1" subwoofer channel is also generated.

3.2 IRCAM ADM tools

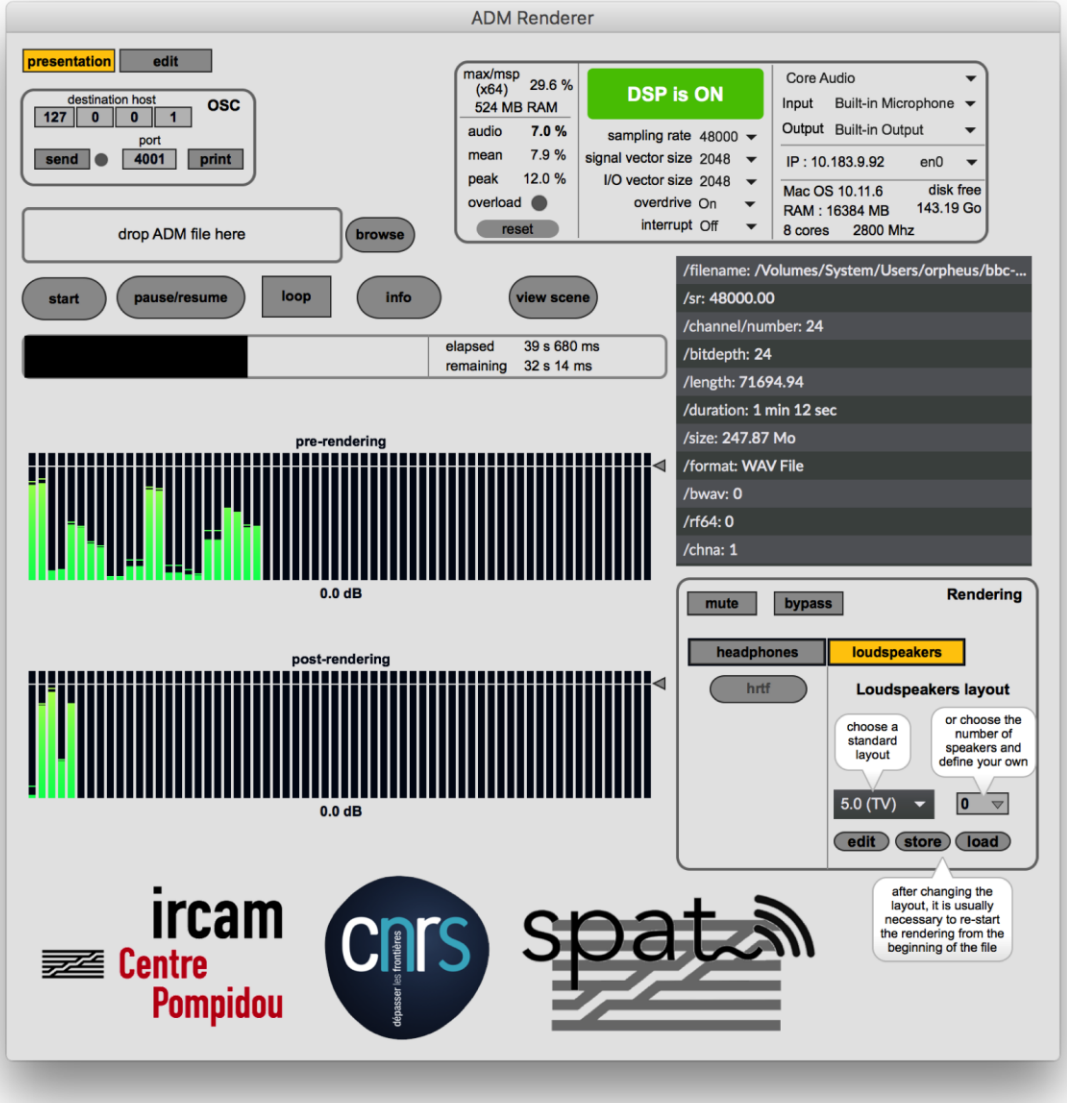
Several stand-alone tools were developed for experimenting with reverberation tracks in ADM files, but they can also be used for more general ADM recording, playback and rendering. These tools currently only support a subset of the ADM specifications, but the most common features are already available. All the tools mentioned below are available for macOS and MS Windows.

3.2.1 ADM Player/Renderer

The main user interface is shown in Figure 6. WAV/BWF/BW64/RF64 files with ADM metadata can be opened. Summary information about the content of the currently open file is displayed in the list on the right. Controls for playback are located on the left. Below them, there are level meters for each audio track in the file and below these there are level meters for the rendered output channels.

On the lower right, the rendering settings can be chosen. Both loudspeaker-based and headphone-based reproduction is possible. For rendering on loudspeakers, VBAP [1] is used. The loudspeaker setup can be chosen from a predefined set of common loudspeaker setups or by manually specifying custom loudspeaker positions. Up to 64 loudspeakers are supported. For binaural rendering, Head-Related Transfer Functions (HRTFs) can be selected from a large set of existing HRTF data.

In the top left corner of the user interface, Open Sound Control (OSC) messages can be enabled which can for example be used to control an external renderer (while switching the built-in renderer to “bypass” mode) or further process the metadata generated by the ADM Player. The OSC messages can either be sent to another application running on the same computer or to another computer on the network.



destination host 127 0 0 1
port 4001

max/msp (x64) 29.6 %
524 MB RAM
audio 7.0 %
mean 7.9 %
peak 12.0 %
overload ●

sampling rate 48000
signal vector size 2048
I/O vector size 2048
overdrive On
interrupt Off

Core Audio
Input Built-in Microphone
Output Built-in Output
IP : 10.183.9.92 en0
Mac OS 10.11.6
RAM : 16384 MB
8 cores 2800 Mhz
disk free 143.19 Go

elapsed 39 s 680 ms
remaining 32 s 14 ms

/filename: /Volumes/System/Users/orpheus/bbc-...
/sr: 48000.00
/channel/number: 24
/bitdepth: 24
/length: 71694.94
/duration: 1 min 12 sec
/size: 247.87 Mo
/format: WAV File
/bwav: 0
/rf64: 0
/chna: 1

choose a standard layout
5.0 (TV)

or choose the number of speakers and define your own
0

after changing the layout, it is usually necessary to re-start the rendering from the beginning of the file

Figure 6: Main user interface of the ADM Player/Renderer

The “view scene” button opens a window showing the current position of all audio “objects” in the ADM file, as shown in Figure 7. The scene window also allows interactively moving objects with the mouse, at least as long as there is no movement data recorded in the ADM file for the given object at the current time.

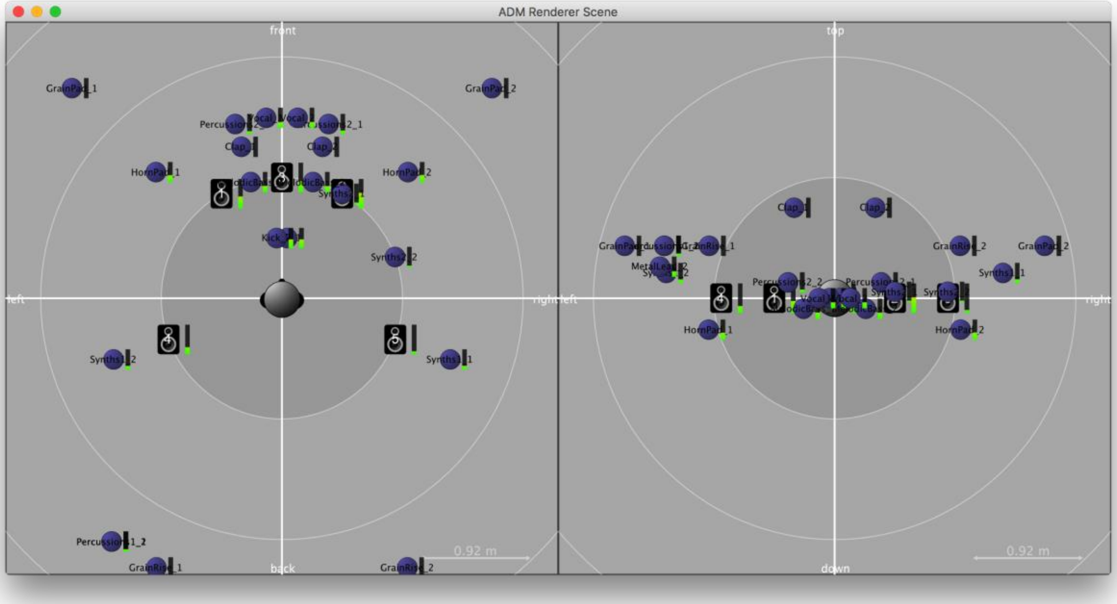


Figure 7: Scene display of the ADM Player/Renderer playing the file machine-aer_bwf.wav [5].
Left pane: view from above; Right pane: view from behind

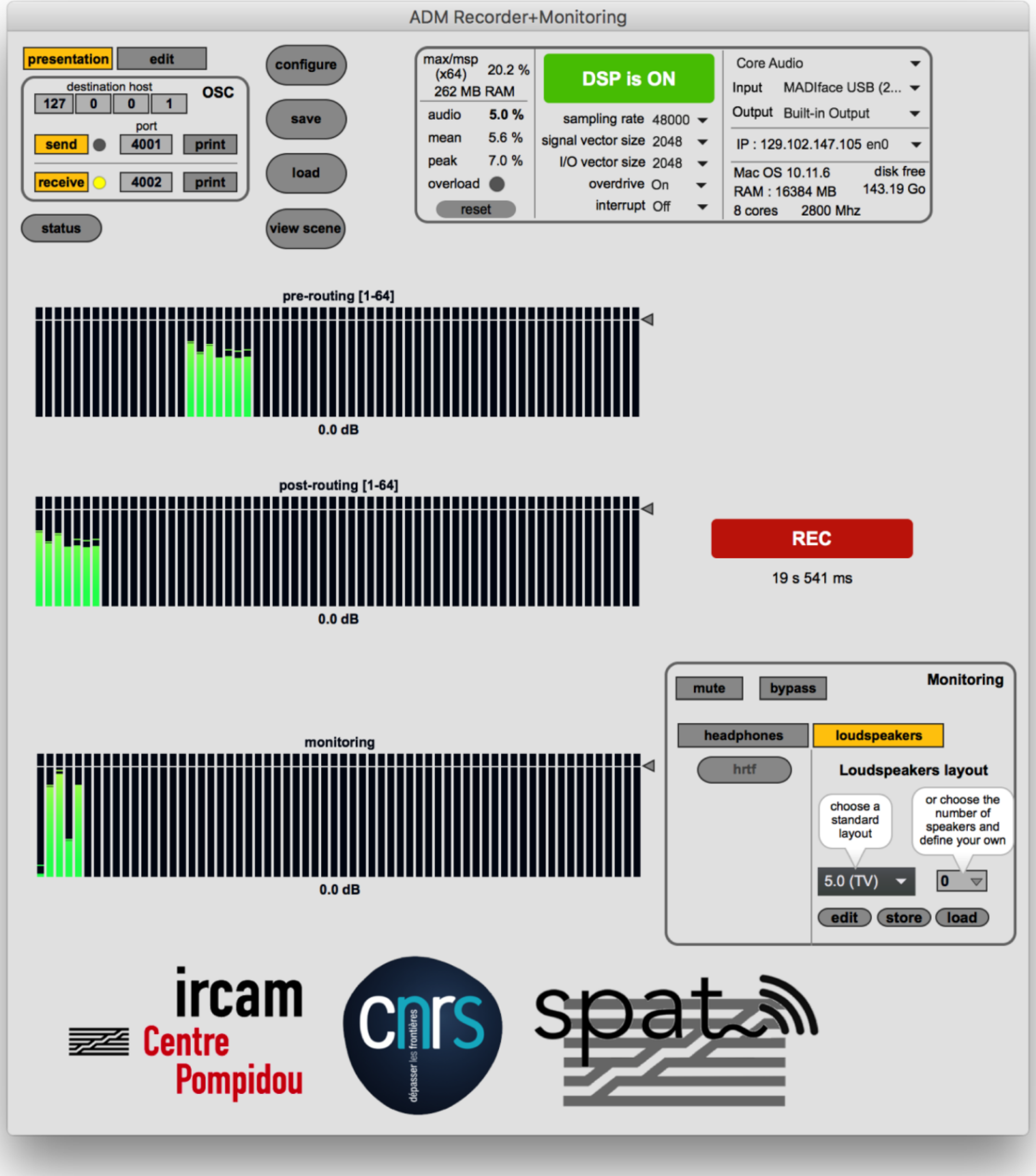
3.2.2 ADM Recorder + Monitoring

Figure 8 shows the main window of the ADM Recorder, which contains a built-in renderer for monitoring. The monitoring settings in the bottom right of the window are the same as for the ADM Renderer described above. The “configure” button can be used to open a window for editing ADM metadata as shown in Figure 9. On the top of this window, the programme and content name can be specified. Below that, a list of “packs” can be set up. Each pack has one of the ADM types “Objects”, “DirectSpeakers”, “Binaural”, “HOA” and “Matrix”. Note that the last two are currently not implemented. Each pack can contain one or more “channels”. For “DirectSpeakers”, one of a set of pre-defined channel-based setups can be selected which populates the channels appropriately. A “Binaural” pack automatically contains two channels, one for the left and one for the right ear.

Each channel has a unique channel number that can be used together with the routing matrix on the left side to assign input channels to the desired ADM tracks.

Once the packs and channels are set up as desired, the “arm” button can be used to enable recording. This will open a scene display window similar to Figure 7 which allows configuration of the initial positions of all audio objects.

In the top left of the main window, host and port settings can be configured for receiving and sending OSC messages that can transport ADM metadata between applications and between computers. All messages that are received via this interface are stored as XML metadata in the recorded ADM file. This interface can also be used to establish a connection with the ToscA plugin, see below for more information.



ADM Recorder+Monitoring

OSC destination host 127 0 0 1
port 4001 (send)
port 4002 (receive)

max/msp (x64) 20.2 %
262 MB RAM
audio 5.0 %
mean 5.6 %
peak 7.0 %
overload

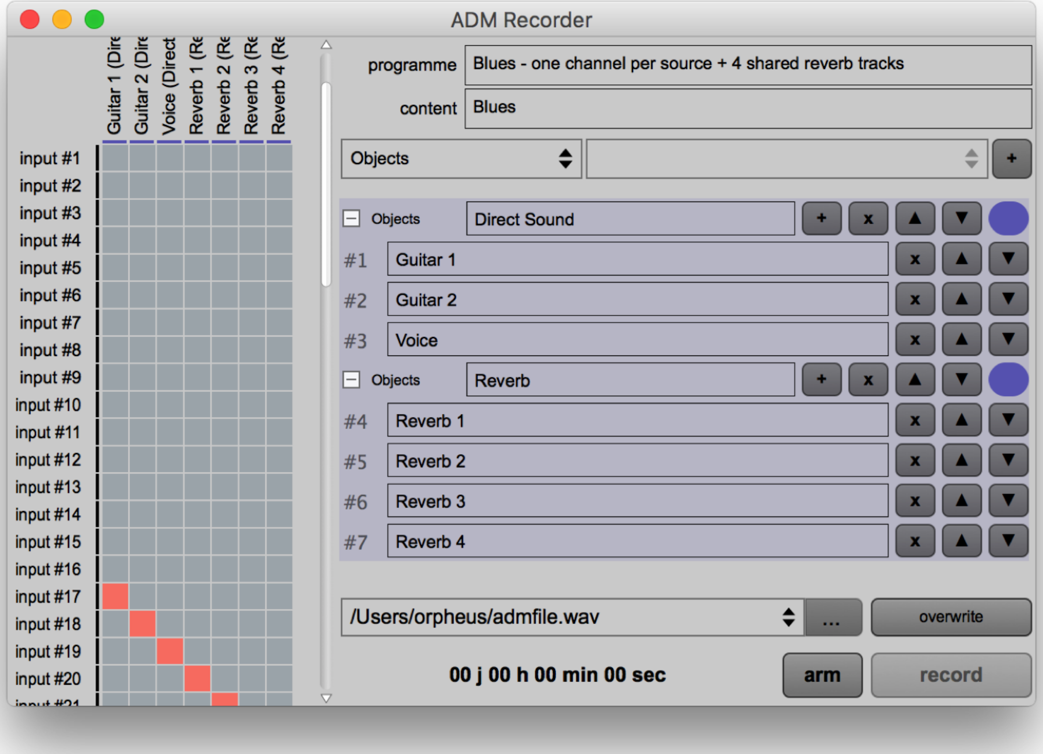
DSP is ON
sampling rate 48000
signal vector size 2048
I/O vector size 2048
overdrive On
interrupt Off

Core Audio
Input MADIface USB (2...
Output Built-in Output
IP : 129.102.147.105 en0
Mac OS 10.11.6
RAM : 16384 MB
8 cores 2800 Mhz
disk free 143.19 Go

REC
19 s 541 ms

Loudspeakers layout: 5.0 (TV), 0

Figure 8: Main window of the ADM Recorder



ADM Recorder

programme: Blues - one channel per source + 4 shared reverb tracks

content: Blues

Objects: Direct Sound
- #1 Guitar 1
- #2 Guitar 2
- #3 Voice

Objects: Reverb
- #4 Reverb 1
- #5 Reverb 2
- #6 Reverb 3
- #7 Reverb 4

/Users/orpheus/admfile.wav

00 j 00 h 00 min 00 sec

Figure 9: Metadata editor and routing matrix

3.2.3 ADM ExtractXML

Given a sound file with ADM metadata, this tool can be used to create a separate XML file containing only the ADM metadata.

3.2.4 ToscA

ToscA is a plugin for Digital Audio Workstations (DAWs) that allows the recording of automation tracks for arbitrary parameters that can be sent and received as OSC messages via a network interface. This tool is not specific to ADM, but it can be used together with the ADM Recorder to record movements and changing gain values as automation tracks in a DAW. The ADM Recorder can in turn record those data in the ADM file.

3.3 Loudness control

Loudness measurement in BBC'S ORPHEUS IP Studio is done according to Recommendation ITU-R BS.1770. Annex 3 to the 4th revision of this specification includes weighting values for all loudspeaker positions in Recommendation ITU-R BS.2051 [4], derived from a table of weightings for presentation from any azimuth and elevation angle.

Until the study in ITU-R to develop a truly object-based loudness meter is complete, loudness in the ORPHEUS project will be measured using a renderer, rendering the objects to channels for a standard loudspeaker constellation, and then using the channel-based loudness meter to measure the loudness of the channel-based audio.

Investigations have been done on the difference in loudness that there might be for typical programme material rendered to different loudspeaker constellations (two-channel stereo, 5-

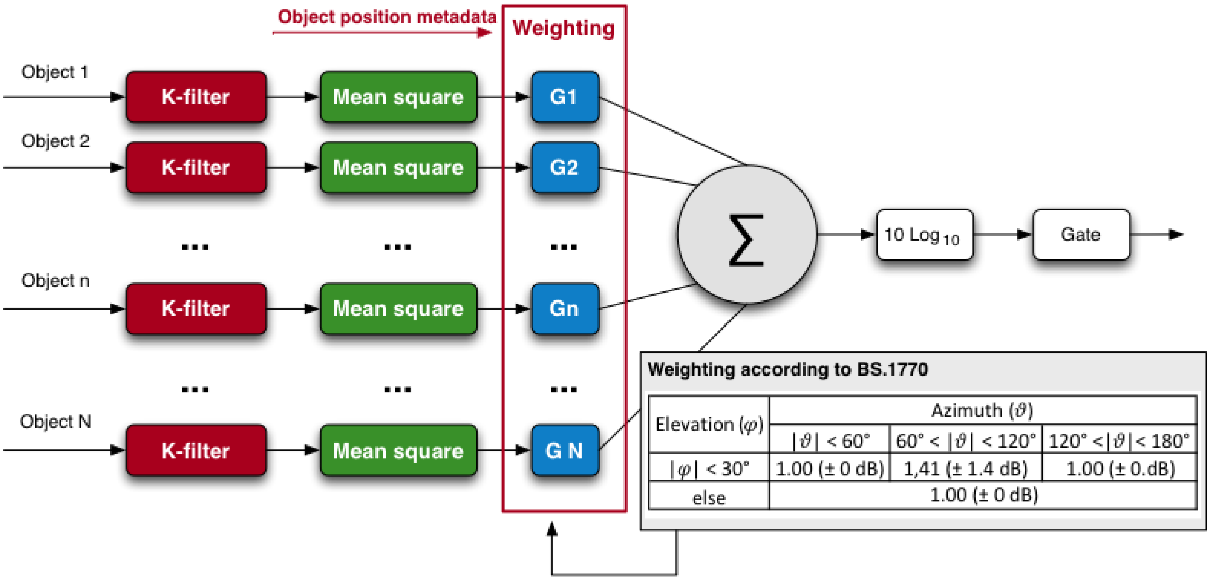
channel surround, 7+4 surround with height, and so on).

3.3.1 Object-based loudness measurement

One of the major challenges in object-based loudness measurement is that neither the target loudspeaker setup nor the renderer that will be used on the reproduction site might be known during production. Thus, a generic object-based loudness measurement algorithm dependent only on the audio scene itself would be preferred. If that's not possible or feasible for all potential configurations, a common measurement procedure for production including adaptation strategies for distribution and provision would be required. In all cases compatibility with the existing loudness measurement algorithm in accordance to Recommendation ITU-R BS.1770 is required, as currently the standard for current channel-based audio productions.

IRT conducted a study based on modifications of the ITU-R BS. 1770 as a starting point to work towards an object-based loudness measurement algorithm. Three variations of the channel-based BS.1770 were considered for this.

The first one is modelled very closely on the original loudness measurement algorithm for advanced sound systems as recommended by the ITU-R [6]. Its slightly simplified block diagram is depicted in Figure 10. The overall idea here is that the same basic structure as for channel-based loudness measurement will be used for object-based measurements while using the audio signals of the audio objects as input signals directly instead of the resulting channel signals of the rendering process. Each object's contribution to the loudness is weighted based on its position, with values given by the rules in BS.1770 for advanced sound systems. Thus, while slightly simplified, it could be argued that objects are measured like loudspeaker signals of an advanced sound system whose speaker positions can change over time.



Weighting according to BS.1770

Elevation (φ)	Azimuth (ϑ)		
	$\lvert\vartheta\rvert$ < 60°	60° < $\lvert\vartheta\rvert$ < 120°	120° < $\lvert\vartheta\rvert$ < 180°
$\lvert\varphi\rvert$ < 30°	1.00 (± 0 dB)	1,41 (± 1.4 dB)	1.00 (± 0.dB)
else	1.00 (± 0 dB)		

Figure 10: Schematic view of generic object-based loudness measurement with position-dependent weighting

The second and even simpler variant is depicted in Figure 11. Again, the object signals are directly used as input signals for the measurement algorithm, but this time no weighting based on the position is done. Thus, the resulting measured loudness depends solely on the audio signals of the individual objects.

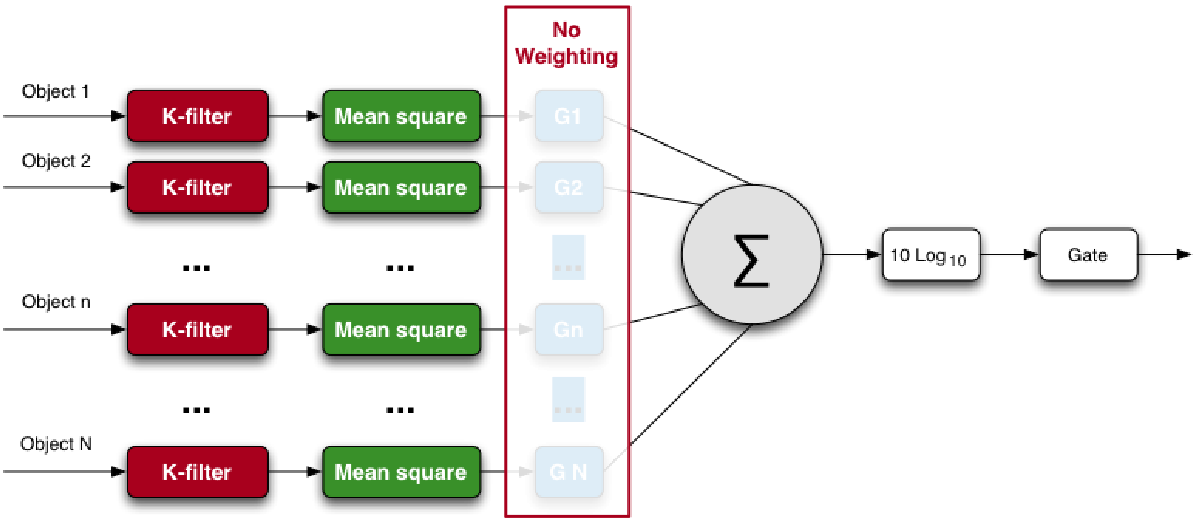
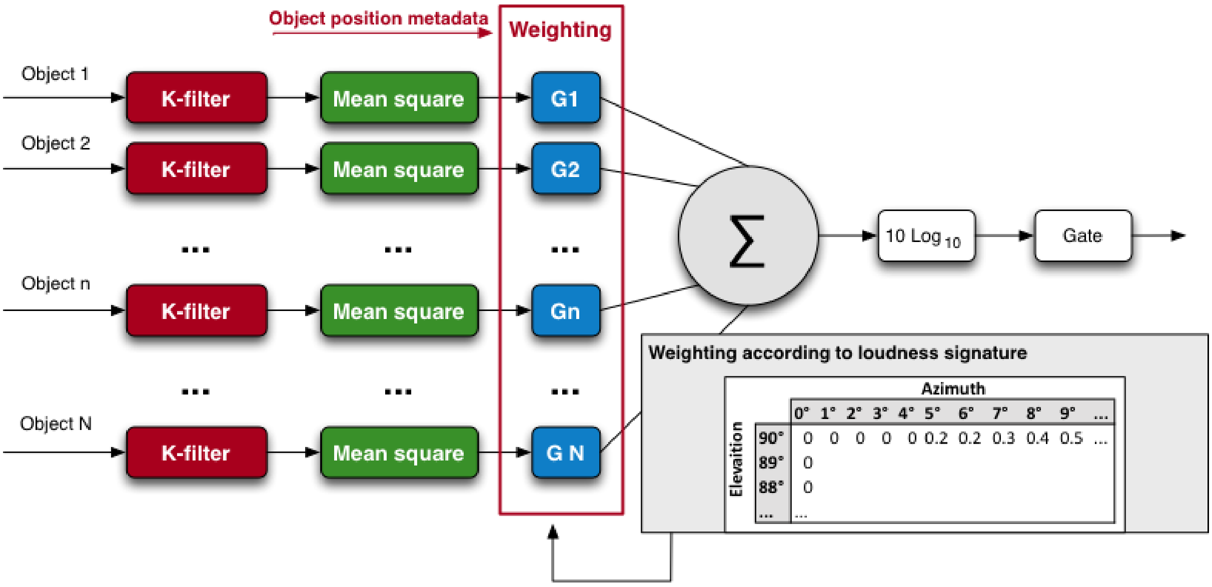


Figure 11: Schematic view of generic object-based loudness measurement without position-dependent weighting

Finally, to investigate how close one could estimate the resulting loudness of the rendered audio signal based on the signals of the audio objects, another variant based on a so-called loudness signature has been introduced, as shown in Figure 12.

A loudness signature in this context is essentially a fine-grained map of weighting values that has been especially tailored for use with a specific target speaker setup and object-panning algorithm. Because of this awareness of the target rendering parameters, this approach was termed '*adaptive loudness measurement variant*'. The basic idea behind using a loudness signature is that, when proven to be useful, it could be used to identify similarities between different configurations and use the obtained data to simplify the signatures as far as possible. One example of such a potential simplification that has been part of the study is based on *quadtree image compression*, as can be seen in Figure 13, which simplifies the weighting values by thresholding the weighting value in spherical coordinate space.



Elevaition	Azimuth 0°	1°	2°	3°	4°	5°	6°	7°	8°	9°	...
90°	0	0	0	0	0	0.2	0.2	0.3	0.4	0.5	...
89°	0										
88°	0										
...	...										

Figure 12: Schematic view of an adaptive object-based loudness measurement with weighting based on a target setup-dependent loudness signature

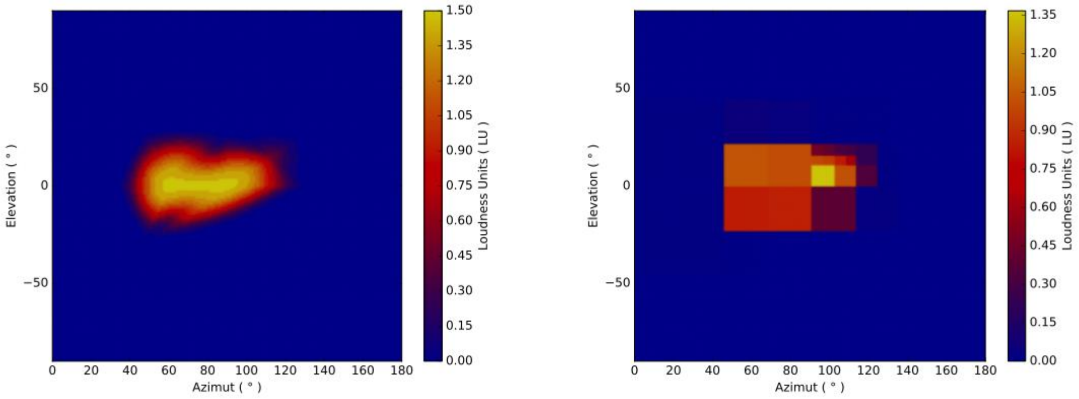


Figure 13: Example of a loudness signature with weighting factors for each source position based on target loudspeaker setup and rendering method.
Original high resolution version (left), and quadtree compressed example (right)

The evaluation was carried out for all three variants with real object-based audio scenes with natural audio signals and randomized audio scenes with audio objects carrying white noise. Multiple point source panning variants in combination with all speaker setups defined in ITU-R BS. 2051 have been evaluated. The performance has been measured by calculating the deviation between the estimate of the object-based loudness algorithm variant and the resulting program loudness of rendered loudspeaker signals as currently defined by ITU-R BS.1770.

The results can be seen in Figure 14 for completely randomized audio scenes and Figure 15 for real audio scenes. It can be seen that – independent of the audio signals or measurement variant – the error is very small in general. This means that, while differences can be found for specific configurations, even the simplest generic variant without weighting seems to provide reasonably accurate loudness estimates for real world applications. This is especially true when the required algorithmic complexity of the evaluated variants is considered.

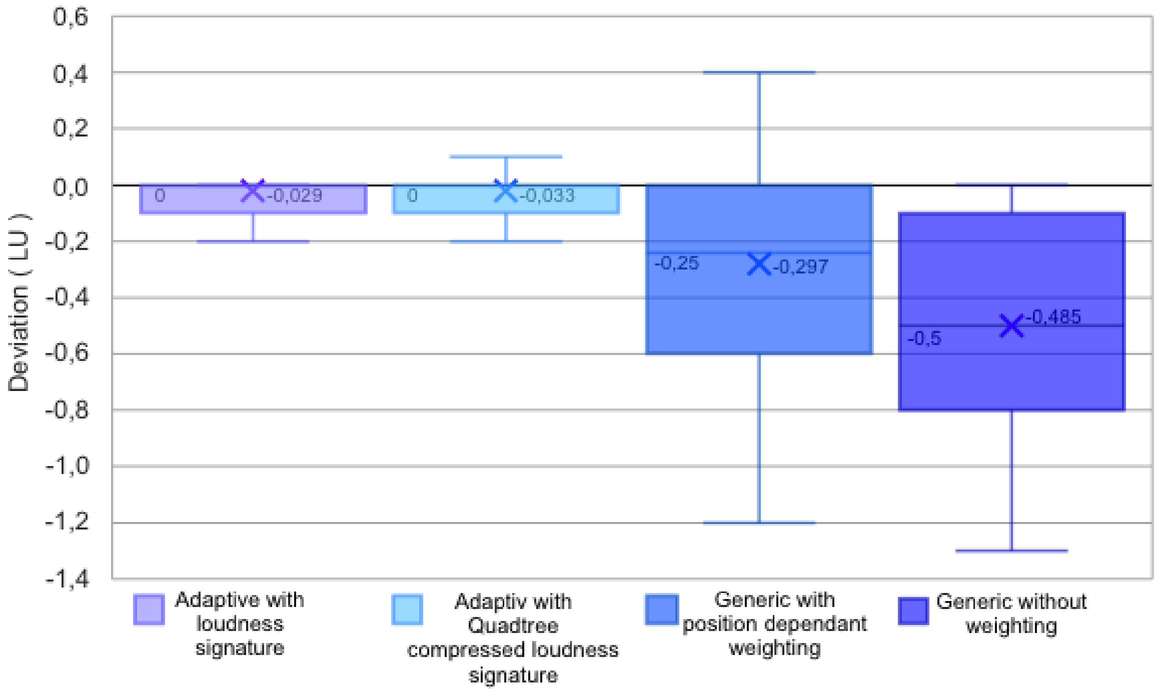


Figure 14: Boxplots of the deviations between the loudness variants tested and a measurement of the resulting loudspeaker signals in accordance to ITU-R BS. 1770 for completely random audio scenes with white noise objects. Boxes show median, quartiles, minimum, maximum, and arithmetic mean (cross).

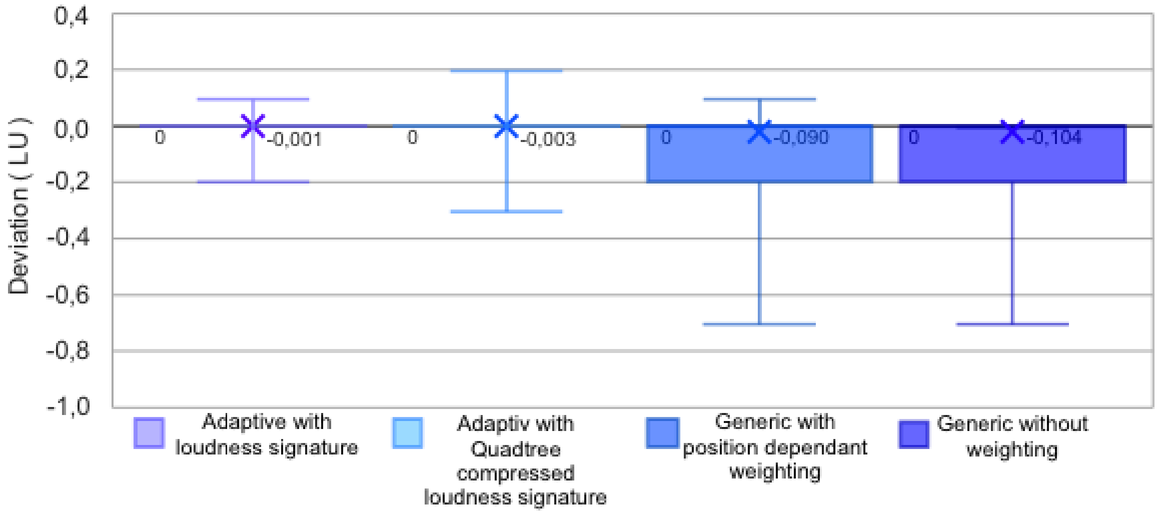


Figure 15: Boxplots of the deviations between the tested loudness variants and a measurement on the resulting loudspeaker signals in accordance to ITU-R BS. 1770 for real audio scenes with natural audio signals. Boxes show median, quartiles, minimum, maximum, and arithmetic mean (cross).

It must be noted, though, that these results can only serve as a first indication, giving hints towards a potential future truly object-based loudness measurement variant. Most importantly, it should be highlighted that the results are only valid if the audio signals are incoherent with respect to the full audio program runtime. Whilst this constraint was met by all the object-based audio mixes that were available for the study, future productions might yield different results due to the continuing evolution object-based recording and production techniques.

Finally, it can be concluded from the results at hand that the programmatic approach to loudness monitoring taken by the ORPHEUS project is a very reasonable approach to working on live object-based audio productions

3.3.2 Higher-Order Ambisonics loudness measurement

In future object-based audio productions, it is also envisioned that compact microphone arrays be used to capture sound scenes. Although the signals recorded by such microphone arrays can be converted and transmitted as speaker signals (channel-based format), it may be advantageous to use the Higher-Order Ambisonics (HOA) format. One major obstacle to the use of the HOA format in production is the lack of a standardised, specifically designed loudness estimation algorithm. In order to fill this lack, b<>com conducted a study on HOA loudness estimation.

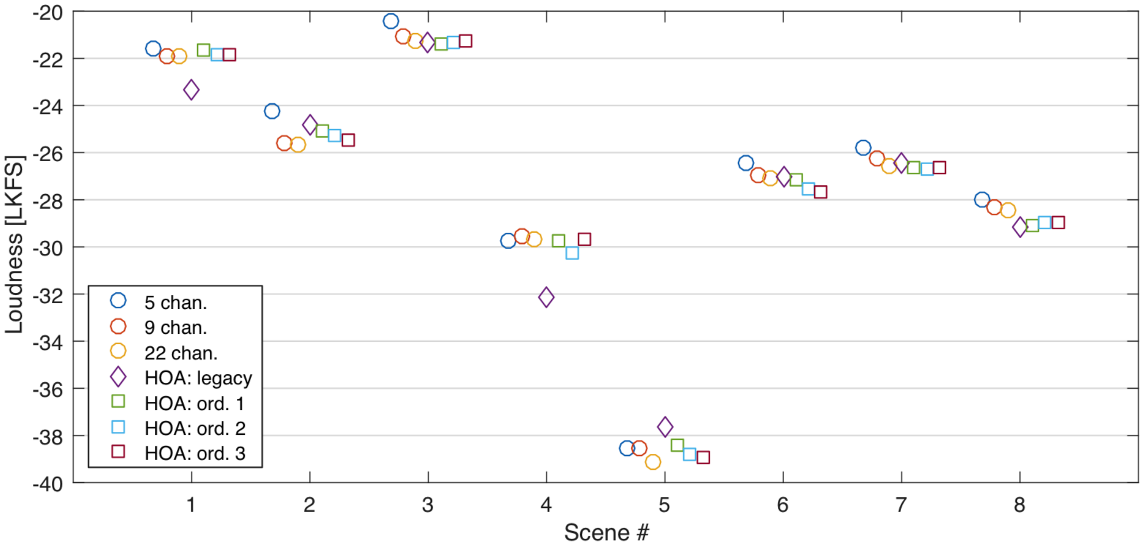


Figure 16: Comparison of loudness estimations of 8 sound scenes encoded in the HOA format using an existing HOA loudness algorithm and the algorithm developed during the Orpheus project

Figure 16 presents some results of the investigation. In this figure, we compare loudness values calculated for eight object-based sound scenes. On the one hand, the sound scenes were rendered to 5-channel, 9-channel and 22-channel speaker setups using amplitude panning. The method described in Recommendation ITU-R BS.1770 was then applied to the speaker signals. The corresponding loudness values are indicated as circles. On the other hand, the sound scenes were encoded to HOA signals of order 1, 2 and 3 (4, 9 and 16 signals, respectively). Two different loudness estimation algorithms were applied to the obtained HOA signals: an existing algorithm, which consists in calculating the loudness of the "omni" signal (values indicated as diamonds), and an algorithm designed by b<>com during the course of this study (values indicated as squares). In almost every case, the newly designed algorithm yields loudness values that are closer to those obtained with the speaker signals. As well, at every HOA order, the loudness calculated using the novel algorithm is within 1 LU of the value calculated for the 22-channel version of the scene.

3.4 Personalisation constraints

Dialogue intelligibility of audio content is an issue especially for people with hearing impairments. In [7] a Dialogue Enhancement technology as an advanced Clean Audio solution is presented to address this problem. This technology enables the audience to individually adjust the volume of dialogue, music, or sound effects within a single broadcast audio stream for improved speech intelligibility or customized listening control. It reports results of a BBC and Fraunhofer IIS experiment using the content from a Wimbledon lawn tennis tournament to test the integration of the technology into the production workflow to find out about user reactions and to verify whether it is as helpful as intended for a hearing-impaired audience. A user survey was linked to the playback client for collection feedback. The user reaction to the new technology was very positive, with users reporting the new feature as a useful extension for TV and radio broadcasting, especially for sports content, but also for drama.

Two uses of a single personalisation control are envisaged in the first implementations of the ORPHEUS chain. This control can be used to alter the balance between two elements of the audio presentation. In the first application, the balance is between speech and background music, principally to reduce the level of background in order to increase intelligibility. In the second, it is between commentary and crowd effects in sports events to suit two different audiences – those who want commentary, and those who do not. This second case is of limited relevance to the ORPHEUS project because sports events on the radio that make sense without commentary are few.

A constraint is desirable in the first use case because the complete absence of background music will lead to a loss of dramatic content, and might even cause an editorial problem if reference is made in the speech to sounds in the background.

In the second use case, it has previously been felt to be desirable to have a constraint in order to prevent complete suppression of the commentary. Accidental complete suppression of commentary (whether by misuse of a user interface control, or by persistence of a setting from one programme to another) could lead to audience dissatisfaction if the listener wants commentary, and it is not apparent that commentary is available. For this reason, overall audience satisfaction could be higher if the commentary is always audible. However, this might be offset by the irritation caused to the listener by commentary that is audible but unintelligible.

To avoid these problems, objects that have personalisation enabled in the DAW will, by default, have a constraint generated as well.

For objects tagged as “background” the limit will be [0 .. -20dB]

For objects tagged as “commentary” the limit will be [0 .. -20dB]

The experience from the “5 live Football Experiment” [8] showed three clear preferences for the football mix: (1) a mix of commentary and crowd with the fader in the middle of the range, (2) as little commentator as permitted, (3) as little crowd noise as possible.

3.5 Intelligibility

Personalisation can be a useful feature to help listeners adjust programme balance to their needs. Setting the default balance, to suit the largest proportion of the audience possible, could be aided by the use of an objective measurement of audibility. Two projects (outside ORPHEUS) are being monitored in the hope that they might provide tools to measure intelligibility.

The first project involves a partnership between the manufacturer RTW, well known for audio metering products, and a division of FhG in Oldenburg (rather than at Erlangen). This is supported by the German Ministry for Economy and Energy (BMWI), and goes by the name of “Speech Intelligibility for Broadcast” (SI4B) [9].

The second project is at the University of Salford, as part of the S3A “Future Spatial Audio in the Home” partnership [10]. S3A is funded by the Engineering and Physical Sciences Research Council (EPSRC), and involves the Universities of Surrey, Salford and Southampton, and the BBC Research & Development.

Should the opportunity arise to integrate either or both of these tools into the ORPHEUS programme chain to assess their performance, it is intended to do so.

3.6 Control room design

A control room should provide optimal conditions to produce and evaluate the audio content of a radio program. Therefore, it has to be isolated from outside noise, and any unwanted noise sources inside, e.g. fans, must be below a defined noise floor. The reverberation characteristic should be frequency neutral and the reverberation time should be low, to hear the details in the direct sound.

Often-cited room acoustic requirements are those in ITU-R BS.1116 for use in formal subjective evaluation tests [11]. This specification also recommends the technical parameters of the loudspeakers to be used. However, it is unusual for radio studios to meet these, as they are very stringent.

Even so, studio control rooms are equipped with high quality loudspeakers in optimal positions. This is necessary to ensure that problems with the audio signals will be audible during production, such that they can be corrected. In general there will be more, and better, loudspeakers for the sound engineer in the control room than an audience member will have.

There is a need therefore to extend a practice that has applied to stereo and 5.1 surround sound for many years: the provision of sub-optimal monitoring options. It is common for mixing consoles to provide outputs to drive multiple sets of loudspeakers. These might be identified as “large” and “small”, for example. The purpose of the “small” output is to drive loudspeakers that are more representative of low-cost domestic installations, whilst the “large” output drives the high quality studio monitor loudspeakers. Most of the work would be done using the monitoring loudspeakers, and a check for compatibility with the small loudspeakers would be done from time to time, as the programme content dictates. In addition to alternative loudspeakers, a mix to mono may also be auditioned.

Spatial audio production in the experimental ORPHEUS studio uses 11 loudspeakers, in a “4+7+0” layout (4 above head height, 7 at head height, none below head height).

It is anticipated that compatibility with stereo rendering, and intermediate formats needs to be checked, and so alternative renderings will be available from the IP Studio system, selectable easily from the mixer interface. Initially, it will provide “2.0” (conventional stereo) and “5.1” (5-channel, planar surround) as the “sub-optimal” renderings. The 2.0 rendering can be presented on a pair of smaller loudspeakers. It is not possible to equip the studio with a second set of 5.1 surround loudspeakers.

It is felt currently that this provides enough options to check on quality, without presenting more than it will be reasonable for someone to have time to check.

The choice of when to monitor the different renderings is left to the studio engineer.

4 Distribution

In this context, “Distribution” is used to encompass the process of transferring the programme from the studio to the provision encoder, but also to include the associated processes that take place at this stage where the programme is finished.

The reduction in the number of simultaneous objects may be needed in this process is considered, then the requirements for legal compliance checks, and those of archiving.

4.1 Object funnelling

During production, there need be no practical limits on the number of audio objects within a programme – a DAW could manage thousands. Some objects – an acoustic ambience for example - could last for a long time, whilst others – a sound effect of a gun shot – would only be short. During live distribution of the object-based audio, using one audio stream for each object would result in many streams unnecessarily using bandwidth during the time before and after the object exists.

DVB standards for “next-generation audio” have been built in commercial requirements that stipulate that systems need support no more than 16 simultaneously decoded elements.

These two factors lead to a requirement for a process of “funnelling” to take place between programme production and programme provision: several objects that (in the simpler case) do not exist simultaneously are combined into a single channel.

The initial assumption in the ORPHEUS project is that this funnelling process can take place easily and losslessly because experimental programmes will have relatively few simultaneous objects. As complexity increases, rules may be built into production tools (the DAW, for example) such that a limit is enforced. The limit may be derived from knowledge of the transmission system that is in use, such as the value 16, known from DVB.

The more complex case, where the number of simultaneous objects in the programme is greater than the number supported by the distribution (or provision) system, is yet to be explored fully. It can be anticipated that there will be circumstances where several simultaneous objects can be combined into a single object, with some loss, perhaps of spatial resolution, or of the ability to apply separate processing to the objects.

A need is foreseen in the ORPHEUS project for guidance on how best to perform funnelling when loss of information is unavoidable. This might take the form of advice on the signal processing to apply to audio waveforms when combining objects, or on the manipulation of the metadata.

4.2 Recording of transmission for legal compliance

The BBC and BR are required to keep records of their broadcasts for a defined period of time for legal purposes (the investigation of complaints of misleading information, for example, or failure to meet required technical standards). These recordings need to fit for the purpose.

The recording is made of the signal that is sent *to* the transmitters (off-air monitoring of transmitter operation is a separate matter).

The process has involved tape in the past, but is now computerised. In the BBC, automated recording takes the output of a radio network’s “presentation” mixing console, after it has been subject to dynamics processing for broadcast. The files are kept uncompressed for 1 week, and then are compressed to save storage space. The BBC has different regulations for TV and radio, so the retention periods are different (90 and 42 days, respectively).

For object-based audio, it will be necessary to record the audio objects and their metadata, because the different renderings and interactivity could profoundly affect the audible result. Recording only a stereo rendering, for example, might conceal problems with a rendering for a 3-dimensional constellation of 11 loudspeakers.

Calculations have not yet been done of the amount of storage that the recording of transmission will require. Because the legal obligation is of a finite duration, it is not expected that the cost of storage will be a significant problem.

4.3 Archiving of finished programmes

Distribution is a suitable point in the programme chain at which to take finished programmes to be stored in the archive. The Orpheus project advocates the use of open standards throughout, and BW64 with ADM is the assumption for archiving. This is to ensure that the files remain easy to interpret and decode in future, as codecs (which might be proprietary) come and go.

The use of compressed formats (such as MPEG-H) is not recommended for archiving. Even though high bit rates offer excellent audio quality, the possible loss of quality that concatenated coding would incur means that these are not suitable.

As stated in another Deliverable, sampling at 48kHz, with 24 bits, is necessary and sufficient for programme production.

5 Provision

MPEG-H 3D Audio, specified as ISO/IEC 23008-3 (MPEG-H Part 3) [3], is an audio coding standard developed by the ISO/IEC Moving Picture Experts Group (MPEG) which was designed to meet the requirements of so called Next Generation Audio (NGA). It is adopted in broadcast application standards such as ATSC 3.0 and DVB and has been selected for terrestrial UHDTV services in Korea for which test streams are on air since April 2017.

The rendering of audio objects on arbitrary trajectories and the format conversion of loudspeaker configurations is an integral part of MPEG-H 3D Audio standard, which provides end-to-end control of the resulting audio quality.

Profile Level	1	2	3	4	5
Maximum sampling rate [kHz]	48	48	48	48	96
Maximum core codec channels in bit stream	10	18	32	56	56
Maximum simultaneously decoded core codec channels	5	9	16	28	28
Maximum loudspeaker outputs	2	8	12	24	24
Example loudspeaker configuration	2.0	7.1	7.1+4H	22.2	22.2
Maximum decoded objects	5	9	16	28	28

Table 1: Levels for low complexity profile of MPEG-H Audio

The MPEG-H 3D Audio standard defines subsets functionality and features for specific applications and complexity constraints. ORPHEUS follows the *Profile and Level* which has been defined in ATSC 3.0 and DVB, namely the Low Complexity Profile at Level 3 [12], see Table 1 (taken from [13]). Level 3 limits the codec to 32 core codec channels from which only 16 can be decoded simultaneously. For example, this allows decoding and rendering of 16 simultaneous audio objects or 3D speaker layouts such as 7.1+4H with 3 additional objects. Recommended core bitrates for the different channel configuration are listed in Table 2 (taken from [13]). These bitrate-channel configuration combinations are the results of the ISO/IEC MPEG-H verification test [14].

Channel Configuration	Bitrate [kbs]
2.0 Stereo	96
5.1 Multi-channel surround	192
7.1+4H Immersive Audio with 4 height speakers	384
22.2 Immersive audio	768

Table 2: Recommended core bitrates for excellent audio quality for broadcast applications [13]

6 Reception

The signal processing required for reception and consumption is

- Decoding of bit-rate reduced audio object
- Rendering of objects to loudspeakers or headphones
- Application of personalisation features that have been transmitted (e.g. foreground-background balance)
- Application of personalisation features that are entirely under listener control (e.g. environment-aware loudness control)

6.1 Decoding

Decoding of bit streams will be done using one or other standardised decoders from the MPEG family. MPEG 4 AAC and MPEG-H are the two main ones that will be used for object-based audio. Legacy, channel-based, systems (such DAB, DVB) will use MPEG 1 Layer II or AAC. The behaviour of these decoders is well specified in the respective standards.

6.2 Rendering

The case of rendering for the audience is similar to that in Section 3 above, in that several renderers might be used, however, the scope of application of the ITU-R recommended does not extend beyond programme production. The renderer in MPEG-H (from FhG) is used in the mobile app developed by Elephant Candy. As noted in Section **Error! Reference source not found.**, we will be using the low-complexity level 3 mode of MPEG-H. Receivers based on web browsers will use the VBAP renderer from BBC R&D.

A problem still to be solved for the ORPHEUS project (and more generally) is that of the possibility of using a binaural renderer during programme production different from the one used in the listener’s device. During production, as mentioned in Section 3.6, checks are made for compatibility with different listening environments. Loudspeakers would normally be used for that, but the expected prevalence of binaural listening on portable devices suggests that binaural rendering should also be monitored. Currently, there are several possibilities at these two ends of the Orpheus chain. It is not yet clear which will finally be used, and tests have not yet been done to establish how great the differences in sound might be. Given that quality of experience of binaural sound can vary quite widely between individuals, this is an area that will need careful consideration.

6.3 Personalisation

Constraints created and signalled by the broadcaster as described in Section 3.4 above, should normally be applied: failure to do so could result in unsatisfactory experiences by the audience. However, it is possible that manufacturers would benefit from the freedom to differentiate their products from others by having different personalisation processing. One example foreseen is for hard-of-hearing listeners, where the foreground/background balance might be changed beyond what is signalled as permissible to create something more distinctive.

Dynamic range compression (DRC) has two distinct approaches: one lends itself to personalisation much less than the other. The first requires the broadcaster to signal how much compression to apply, either by indicating an input-output transfer characteristic - a "profile", or by transmitting a gain coefficient for each coded frame of audio. The second passes responsibility and control entirely to the replay device and the listener.

Dynamic range compression controlled by the replay device allows much more individual personalisation, and can be done in a way that is aware of the environment. BBC R&D, in collaboration with Queen Mary University of London, has developed a technique that uses the microphone in the replay device, such as a mobile phone, to measure environment noise and adapt dynamic range control accordingly [15].

Initial assessments of the technique show a high degree of satisfaction. It is planned to incorporate the algorithm into the mobile app, such that its impact on quality of experience can be assessed further, in more realistic scenarios.

7 Conclusions

A large part of the signal processing required for ORPHEUS' object-based broadcasting chain has now been defined, either by use of open international standards, or as a result of development by the partners. Tools have been created to record and replay ADM metadata, and it is possible to record, edit, and reproduce object-based audio in the studio to a high quality. The distribution and provision links are largely defined by standards.

Work remains to be done in some areas, such as

Standardisation of object-based audio renderers;

Optimal processing for object funnelling in difficult cases;

Implementation and verification of dynamic range control in receiving devices;

Development of subjective test methods for cases where there is no defined reference.

References

[1] Pulkki, V., “Virtual Sound Source Positioning Using Vector Base Amplitude Panning”, J. Audio Eng. Soc., Vol.45, No. 6, June 1997.

[2] ITU-R, “Recommendation ITU-R - BS.2076 Audio Definition Model”, ITU-R, 2015.

[3] MPEG-H ISO/IEC 23008-3 Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D audio. 2016, ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Audio.

[4] ITU-R, "Recommendation ITU-R BS.2051 Advanced sound system for programme production", ITU-R, 2014

[5] D. Marston, C. Pike, F. Melchior, "SAQAS - Spatial Audio Quality Assessment Scenes", http://www.bbc.co.uk/rd/publications/saqas

[6] ITU-R, “Recommendation ITU-R BS.1770-4 - Algorithms to measure audio programme loudness and true-peak audio level,” ITU-R, 2015

[7] Fuchs, H. and D. Oetting, Advanced Clean Audio Solution: Dialogue Enhancement. SMPTE Motion Imaging Journal, 2014 (July/August).

[8] Mann, Churnside, Bonney, Melchior, “Object-Based Audio Applied to Football Broadcasts – The 5 live Football Experiment” BBC R&D White Paper WHP 272, November 2013

[9] SI4B, https://www.idmt.fraunhofer.de/en/Press_and_Media/press_releases/2016/si4b.html

[10] S3A, http://www.s3a-spatialaudio.org/

[11] ITU-R Recommendation BS.1116-3, Methods for the Subjective Assessment of Small Impairments in Audio Systems. 2014, Intern. Telecom Union, Geneva, Switzerland. p. 33.

[12] ATSC, A/342 Part 3:2017, MPEG-H System. 2017. http://atsc.org/wp-content/uploads/2017/03/A342-3-2017-MPEG-H-System-1.pdf.

[13] Bleidt, R.L., et al., Development of the MPEG-H TV Audio System for ATSC 3.0. IEEE Transactions on Broadcasting, 2017. 63(1), DOI: https://doi.org/10.1109/TBC.2017.2661258.

[14] ISO/IEC N16584, MPEG-H 3D Audio Verification Test Report, JTC1/SC29/WG11, Editor. 2017. http://mpeg.chiariglione.org/sites/default/files/files/standards/parts/docs/w16584_%283D_Audio_Verification_Test_Report%29.docx.

[15] Mason, A.J., Jillings, N., Ma, Z., et al., "Adaptive audio reproduction using personalised compression", BBC R&D White Paper 310, October 2015

[end of document]