Subjective Evaluation of the Localization Performance of the Spherical Wavelet Format Compared to Ambisonics

1st Rubén Eguinoa
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
ruben.eguinoa01@estudiant.upf.edu

2nd Ricardo San Martin
Acoustics Group
Universidad Pública de Navarra
Pamplona, Spain
ricardo.sanmartin@unavarra.es

3rd Daniel Arteaga
ATG Soundtech
Dolby Iberia S.L.
Barcelona, Spain
daniel.arteaga@dolby.com

4th Davide Scaini
ATG Soundtech
Dolby Iberia S.L.
Barcelona, Spain
davide.scaini@dolby.com

***Abstract*—A common goal of most spatial audio techniques is to reproduce the precise location and size of sound sources. Ambisonics is a well-established spatial audio technique that renders sound sources with increasing accuracy as the Ambisonics' order increases. Recently, a novel spatial audio format that replaces spherical harmonics with a set of functions based on wavelets has been proposed. The Spherical Wavelet Format (SWF) aims to improve Ambisonics localization, especially at low orders. This study investigates the perceptual spatial properties of both techniques by means of a set of MUSHRA tests.**

***Index Terms*—Ambisonics, Spherical Wavelet Format, localization, listening tests, spatial audio**

I. Introduction

Ambisonics is a theory to record, manipulate and reproduce spatial audio, based on the spherical harmonics (SH) [1]. The Ambisonics channels can be understood as the coefficients of the series expansion of the distribution of sources in terms of SH: the higher the order of the expansion, the more spatial detail is obtained. The Spherical Wavelet Format (SWF) [2], [3] is a new spatial audio encoding which replaces the SH of Ambisonics with a more localized set of functions, the spherical wavelets. It aims to improve Ambisonics localization, especially at low orders. The spherical wavelets are wave-like oscillations on the sphere that can be associated to a certain angular direction, and they are zero or decay very fast outside the region of interest.

The goal of this paper is to assess and compare the spatial properties of SWF and Ambisonics in a full-sphere loudspeaker system, by means of a set of MUSHRA [4] listening tests. The tests compare first and third order of Ambisonics and the zeroth and first level of SWF. The experiments are carried out in a full-sphere layout consisting of 24 loudspeakers.

The paper is structured as follows: in Section II we present a brief introduction of Ambisonics and SWF. In Section III the experimental setup is described (hardware and software). In Section IV the listening tests performed are detailed. In Section V the experimental results are presented. Finally, conclusions and future work are drawn in Section VI.

II. Technologies

A. Ambisonics

Ambisonics is a theory for spatial audio recording and reproduction developed by Michael Gerzon during the 70s of the 20th century. From a theoretical point of view, Ambisonics is based on a perturbative series expansion of the sound field around the origin, in terms of SH. This way, each Ambisonics channel corresponds to a coefficient of the series expansion and, therefore, it is associated to a particular SH.

The number of spherical harmonics used in the encoding determines the Ambisonics order, and thus the number of channels (each l order has $2l + 1$ channels). Zeroth order Ambisonics consists of one channel, the W channel, the omnidirectional component of the field, which corresponds to the sound pressure. First order Ambisonics (FOA) adds the X, Y and Z channels, the directional components in three dimensions. *L*-th order Ambisonics adds other coefficients to the multipole expansion which amount to quantities proportional to derivatives (up to *L*-th order) of the pressure field. Higher order Ambisonics (HOA) is made of $K = (N+1)^2$ channels, where N is the Ambisonics order or spherical harmonic degree.

One of the challenges of Ambisonics is that the decoding to a loudspeaker layout is non-trivial, specially for non-regular setups. There are different approaches to the problem, such as decoding to an intermediate layout (AllRAD [5], [6]), mode matching [7], or finding the optimal decoding by solving a non-linear optimization problem [8]–[16]. For this work we adopt this latter approach by relying on the IDHOA decoder [15], [16].

B. Spherical Wavelet Format (SWF)

SWF is constructed in the framework of second generation wavelets [17], [18]. This implies that the wavelets used are defined on a recursive polygonal mesh that samples the sphere with increased precision as the SWF order increases. SWF itself is constructed over a discrete polygonal mesh; to go from the continous points on the sphere to the nodes on the mesh an interpolation method will be used.

This mesh is built recursively from a primitive polygonal mesh. Following the original SWF implementation, we will consider a octagonal polygonal mesh subdivided with the so-called Loop scheme [19]. In contrast to Ambisonics, where the encoding and decoding basis functions are the same (the SH), in SWF the encoding and decoding filters are different. Also in contrast to Ambisonics, in SWF the encoding and decoding filters are applied recursively.

1) Signal Discretization: The Spherical Wavelet Format decomposition starts with a continuous source distribution over the sphere. An example of such source distribution could be a delta function at the position of a given point source. First step is to sample the continuous source distribution over the sphere to the nodes on the mesh by using tri-linear interpolation, by means of which a given point source on the sphere will be encoded to a maximum of three nodes on the mesh. This will lead to a set of data defined over the finest level of a mesh, the sampled source distribution or *fine data* $\mathbf{f} = (f_1 \cdots f_N)^T$.

In our case, the fine data consists of 66 data elements, corresponding to the vertices of the mesh subdivision at level 2.

2) Signal Decomposition: The fine data enters a down-sampling process that decomposes it into two signals or sets of data: a *coarse* approximation $\mathbf{c}^{n-1}$ and the remainder, the *details* $\mathbf{d}^{n-1}$. The coarse data vector $\mathbf{c}^{n-1}$ represents a spatially low-passed and downsampled version of $\mathbf{f}$. This decomposition is carried out with the so called *decomposition*, *encoding* or *analysis filters* $\mathbf{A}^n$ and $\mathbf{B}^n$. The filters connect two levels: from the fine level n to a coarser level $n-1$. There are as many decomposition filters as mesh levels minus one.

The decomposition may continue up to the coarsest level available (level 0). This process returns a set of $n-1$ detail signals or wavelet coefficients, $\mathbf{d}^0, \ldots, \mathbf{d}^{n-1}$, and one last coarse signal or scaling function coefficients $\mathbf{c}^0$. The representation $\{\mathbf{c}^0, \mathbf{d}^0, \ldots, \mathbf{d}^{n-1}\}$ constitutes the wavelet transform. The process is shown in Figure 1.

SWF is based on the wavelet transform but truncated to a certain level. The coarse data and details will constitute the channels of SWF. At level zero the coarse data $\mathbf{c}^0$ will constitute the SWF channels (one channel per each node on the base mesh); at level 1 there will be additional channels corresponding to the details $\mathbf{d}^0$.

3) Signal Reconstruction: If needed, the reconstruction of the signal can be done with an upsampling process that increases the spatial resolution of the coarse data $\mathbf{c}^0$ to the fine data $\mathbf{f}$. If the details $\mathbf{d}^0$ are added, the process will give back the original fine data. This reconstruction is done with the *reconstruction*, *decoding* or *synthesis filters* $\mathbf{P}^n$ and $\mathbf{Q}^n$ at level n.

Similarly to the decomposition filters, $\mathbf{P}^n$ and $\mathbf{Q}^n$ connect levels, but in the inverse path: from the coarser level $n-1$ to the finest level n. There are as many reconstruction filters as mesh levels minus one. The reconstruction of the original signal is a recursive procedure that starts from the coarse *level 0* and goes to the finest *level n*.

4) Spherical Wavelet Format: A SWF is defined to be each one of the spherical audio encodings determined by:

i) a recursive subdivision mesh over the sphere, ranging from the coarsest level 0 (the based mesh) to the finest level n;
ii) a set of bi-orthogonal filters $\{\mathbf{A}^j, \mathbf{B}^j, \mathbf{P}^j, \mathbf{Q}^j\}$, with $j = 1, \ldots, n$, defining a wavelet transform, and
iii) a truncation level $\ell \in [0, n]$, defining the level of the wavelet decomposition.

The SWF channels will be composed by the coarser data approximations, $\mathbf{c}^0$, in addition to details up to order $\ell - 1$ ($\mathbf{d}^0, \ldots, \mathbf{d}^{\ell-1}$); at level 0, only the coarser approximation $\mathbf{c}^0$ remains.

Therefore, there is not just one, but actually many possible SWF formats. In this paper we will rely on the original SWF proposal [3] based on i) a recursive subdivision mesh over the sphere based on the primitive octahedronal mesh; ii) the set of dual interpolating filters (the most important feature of them is that the decomposition filters interpolate on the first neighbours), and iii) truncation levels 0 and 1, having 6 and 18 channels, respectively.

5) Decoding to Loudspeakers: Similarly to Ambisonics, the decoding of SWF to non-regular layouts is non-trivial. We rely on the IDHOA decoder, which can generate both decodings for Ambisonics and SWF using a non-linear optimization approach. In SWF, IDHOA generates the loudspeaker feeds based on the SWF channels at level 0 ($\mathbf{c}^0$) or 1 ($\mathbf{c}^0, \mathbf{d}^0$) .

III. Experimental Setup

A. Hardware

The Acoustics Group of the Universidad Pública de Navarra developed a sound installation to work and experiment with three-dimensional audio.

The structure holding the speakers in place is a sphere, with a diameter of 2.9 meters, built with steel tubes. The five horizontal tubes and the twelve vertical tubes, each at 30°, give a lot of flexibility in the placement of the loudspeakers (Figure 3). The listening room hosting the structure meets the requirements for reverberation times set out in Recommendation ITU-R BS.1116-3 [20]. For the experiment, a 24

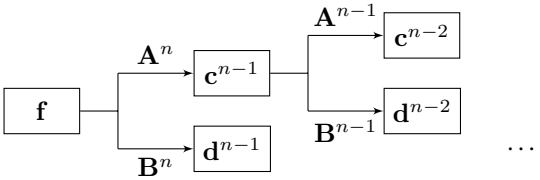


Fig. 1: Scheme of the signal decomposition, which illustrates the encoding process to wavelet space. The fine data $\mathbf{f}$ is decomposed by analysis filters $\mathbf{A}^n$ and $\mathbf{B}^n$ into the coarse $\mathbf{c}^{n-1}$ and details $\mathbf{d}^{n-1}$ signals respectively. The same operation of decomposition is repeated recursively on the coarse signal $\mathbf{c}^{n-1}$ up to the desired level. Reproduced from Ref. [3].

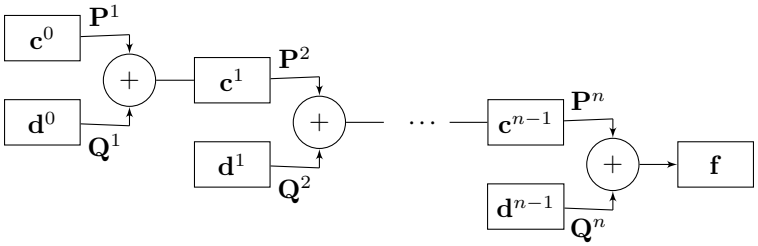


Fig. 2: Scheme of the signal reconstruction, which illustrates the decoding process from the wavelet space. The coarse and details signals at the lowest level, $\mathbf{c}^0$ and $\mathbf{d}^0$, are upsampled via the reconstruction filters $\mathbf{P}$ and $\mathbf{Q}$ respectively. The resulting signals are summed together to obtain the next level coarse signal $\mathbf{c}^1$. The process is repeated at each level with the contribution of the details $\mathbf{d}^\ell$, until the original fine data $\mathbf{f}$ is recovered.Reproduced from Ref. [3].

loudspeaker arrangement that corresponds to a spherical 7-design has been used (see positions in Table I).

The loudspeakers of the installation have been designed ad hoc for the structure. The loudspeaker boxes are made of PLA and created with the help of a 3D printer, having a spherical shape with a back horn. The transducers are the `Dayton Audio PS95-8` full-range woofer fed by two `Dayton Audio MA1240A` amplifiers and driven with the `MOTU 24AO` audio interface.

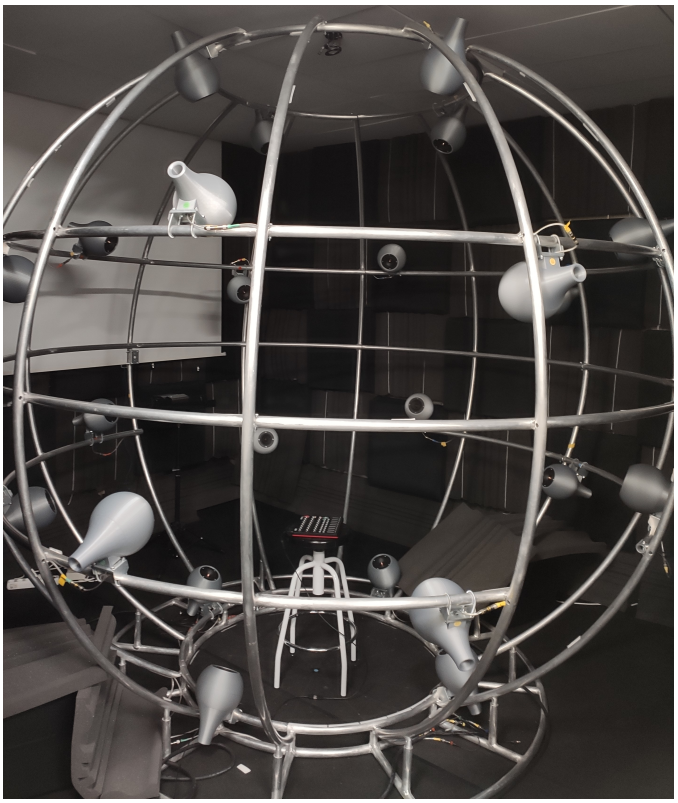


Fig. 3: A look of the sphere's structure. The 24 loudspeakers are mounted in a 7-design layout.

Figure 4 plots the frequency responses of the 24 loudspeakers at their respective positions on the sphere. They have been obtained with an omnidirectional microphone located in the center of the sphere, using a logarithmic sweep as the excitation signal. The frequency response of the loudspeakers is consistent across most of the spectrum.

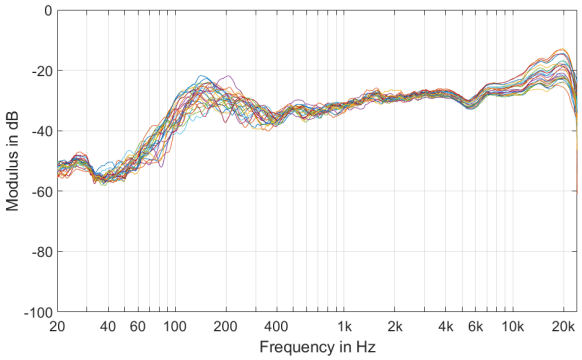


Fig. 4: Frequency response for the 24 loudspeakers in the layout. The loudspeakers' frequency response is very consistent in the range between 400 Hz and 7 kHz. At low frequencies the differences in frequency response between different loudspeakers are more evident due to room modes. At frequencies higher than 10 kHz we might be observing the shadowing effect of the measurement microphone itself.

B. Software

To carry out the listening tests a customized desktop application has been created. The application has been developed with the open-source JUCE framework [21]. JUCE provides a wide variety of libraries and it is specifically optimized for developing audio applications and plug-ins. The application is able to perform the two types of tests described in Section IV: a MUSHRA test to evaluate the localization, and a modified MUSHRA test to assess the apparent width of an audio source. The code is available as a GitHub repository [22].

The evaluations are presented in groups. For each group up to six multi-channel audio stimuli plus the reference and the anchor can be evaluated. The number of stimuli has to be the same in all groups. The playback is done in a synchronous way so, when the stimulus is changed, the new sound will continue to play from the same point in time. The stimuli are presented randomly each time a new group is introduced.

TABLE I: Positions of the loudspeakers in the spherical structure, arranged as a 7-design.

Loudspeaker	Azimuth (°)	Elevation (°)
1	−34.4	−25.0
2	−25.5	15.5
3	−19.3	60.0
4	6.2	−60.0
5	12.5	−15.5
6	21.3	25.0
7	55.6	−25.0
8	64.5	15.5
9	70.7	60.0
10	96.2	−60.0
11	102.5	−15.5
12	111.3	25.0
13	145.6	−25.0
14	154.5	15.5
15	160.7	60.0
16	−173.8	−60.0
17	−167.5	−15.5
18	−158.7	25.0
19	−124.4	−25.0
20	−115.5	15.5
21	−109.3	60.0
22	−83.8	−60.0
23	−77.5	−15.5
24	−68.7	25.0

It is also possible to customize the test via a JSON file providing some basic information, e.g.: the number of audio channels, number of stimuli, name of the audio files.

The application is designed so it can be controlled via the AKAI MIDIMIX MIDI controller. This avoids having a computer, a monitor and an input interface inside the structure.

The graphical interface can be divided in four zones (see Figure 5):

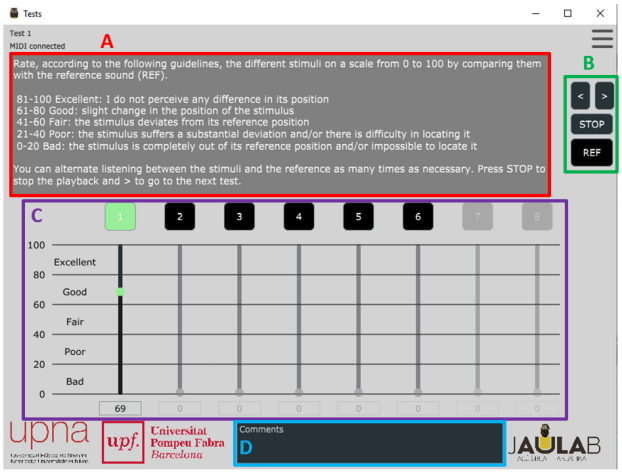
- A Help box: guidelines to assist in the grading of the stimulus.
- B Buttons: $<$ and $>$ to move across groups, *STOP* button to stop the playback and the *REF* button to reproduce the reference sound.
- C Sliders: eight sliders with their respective play button on the top. Out of the eight sliders, only the ones corresponding to an existing stimulus are active, and their value can be modified only if the stimulus is playing.
- D Observations box: the subjects can leave comments to motivate their evaluation.

If the current test group is a source-width test, the reference button is divided in two reference buttons for the narrowest and widest sounds (B) and two circles with different size (E) are drawn next to the help box (A).

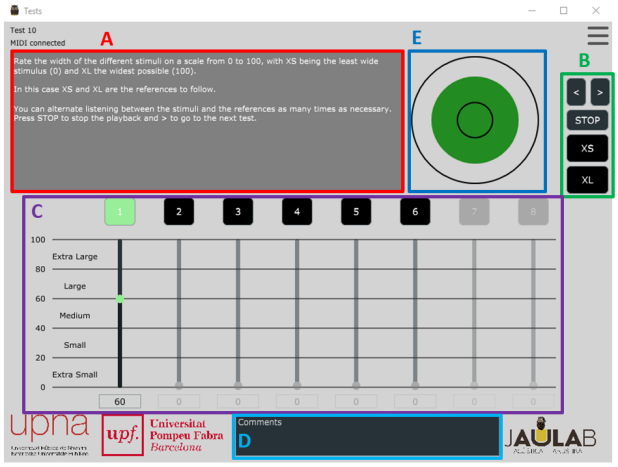
All the results are saved in an xml file, where, for each group, the technologies and their marks are listed.

IV. Listening Tests

The listening tests target two different characteristics of the sound generated by Ambisonics and SWF: the positioning and the width of the sound sources. All the experiments use pink noise at -40 LUFS, 48 kHz, as stimulus. The resulting sound



(a) Localization tests interface.



(b) Source width tests interface. This interface has a dedicated section, E, to help visualizing the source width with respect to the two references.

Fig. 5: User interfaces for the tests. The A block hosts the help box and shows the guidelines for grading the stimulus. The B block hosts the control buttons and the *REF* button to reproduce the reference sound. In the C block it is possible to see eight sliders with their respective play button on the top. The box D is dedicated to user comments.

level at the center of the sphere for all the stimuli is 65 dBA. The stimulus is encoded and decoded in Ambisonics of first and third order and SWF of zeroth and first level.

The encoding for both techniques has been done using MATLAB libraries [23]: in the case of Ambisonics the `Higher Order Ambisonics (HOA)` library [24] has been used; and for the initial interpolation to the finest mesh of 66 nodes in SWF, the `Vector Base Amplitude Panning` library [25]. The downsampling of SWF from the finest mesh at level 2, to level 1 and 0 has been done using the interpolating matrices published here [26].

The matrices for the decoding to loudspeakers have been generated with IDHOA [15] for both Ambisonics and SWF. In Figures 6 and 7 we report the values for the reconstructed

energy and radial intensity over the whole sphere, both for Ambisonics (order 1 and 3) and SWF (levels 0 and 1). We can look at the values of reconstructed energy and intensity over the 24 loudspeakers of the setup, and calculate their mean, maximum and minim values, see Table II. Energy is reconstructed with great consistency across the whole sphere by Ambisonics, with a $\Delta E \approx 0.2$ dB. SWF is less uniform in the energy reconstruction but it is still acceptable, with a $\Delta E \approx 2.6$ dB for SWF at level 0 and $\Delta E \approx 1.4$ dB at level 1. If we focus on the values of the reconstructed radial intensity, that correlates well with the apparent source width, we see that SWF at level 1 and Ambisonics at order 3 should perform similarly, and that SWF at level 0 performs better than Ambisonics at order 1.

The listening panel (LP) consists of 18 people (4 female, 14 male) with different musical training. The age of the subjects ranges between 21 and 43. No remarkable hearing problems have been reported.

In addition to the instructions given in the help box of the interfaces (Figure 5), oral directives on how to carry on the tests were given to the subjects.

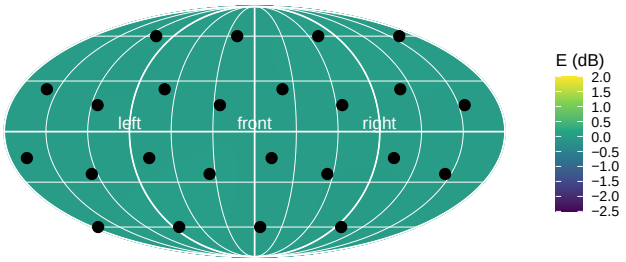
A. *Localization test*

The first test is a localization test. The subjects had to evaluate the accuracy of the source location with respect to a reference and an anchor. The subjects had to evaluate nine different source locations, which are listed in Table III. The stimulus sound was chosen to be a 500 milliseconds pink noise burst windowed by a raised cosine Hanning window, to create a 50 milliseconds fade-in and fade-out. The burst is followed by 250 milliseconds of silence and it is looped. The subjects could listen to the stimuli, reference and anchor as many times as needed.

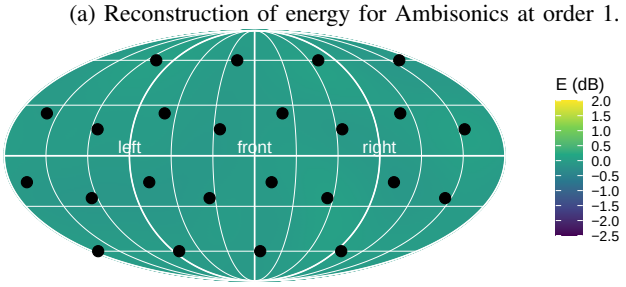
The reference is one physical loudspeaker placed in the position of the source, playing the stimulus in isolation. It is considered a point source encoding without near-field compensation. The anchor is built by decorrelating the mono stimulus signal and reproducing it from every loudspeaker, creating 24 decorrelated sources emitting from the loudspeakers [27] and giving the impression of a completely delocalized source.

TABLE II: Summary of energy and radial intensity reconstruction for Ambisonics and SWF. Mean, maximum and minimum values are calculated over the positions of the loudspeakers.

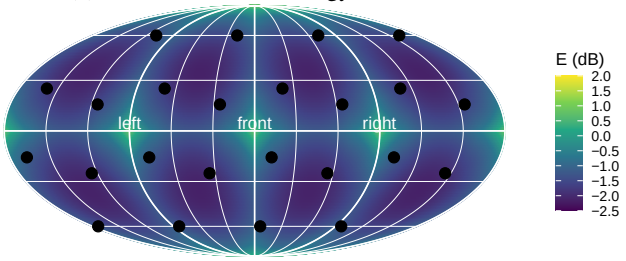
Technique	$\text{Mean}^{\text{Max}}_{\text{Min}}$ Energy (dB)	$\text{Mean}^{\text{Max}}_{\text{Min}}$ Radial Intensity
AMB 1	$-0.02^{-0.01}_{-0.03}$	$0.57^{0.58}_{0.56}$
AMB 3	$-0.04^{+0.05}_{-0.12}$	$0.82^{0.85}_{0.78}$
SWF 0	$-0.84^{+0.61}_{-1.97}$	$0.67^{0.77}_{0.59}$
SWF 1	$-0.82^{-0.31}_{-1.07}$	$0.80^{0.86}_{0.72}$



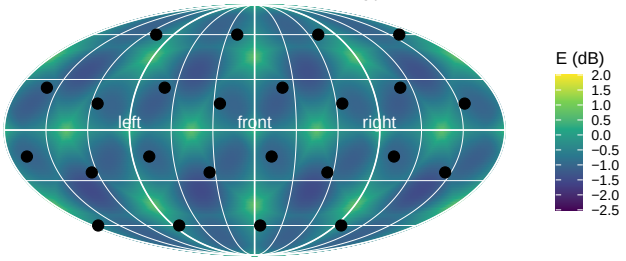
(a) Reconstruction of energy for Ambisonics at order 1.



(b) Reconstruction of energy for Ambisonics at order 3.



(c) Reconstruction of energy for SWF at level 0.

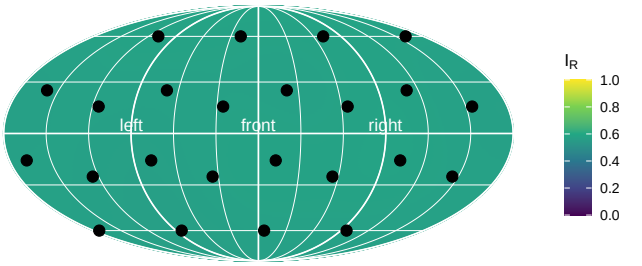


(d) Reconstruction of energy for SWF at level 1.

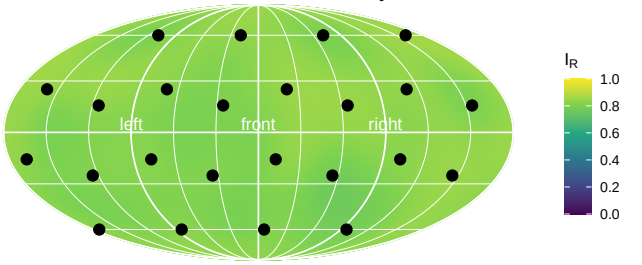
Fig. 6: Comparison of energy reconstruction performances over the sphere for Ambisonics and SWF decoders generated with IDHOA. The black dots represent the loudspeakers' location. Energy is reconstructed with great consistency across the whole sphere by Ambisonics, with a $\Delta E \approx 0.2$ dB. SWF is less uniform in the energy reconstruction but it is still acceptable, with a $\Delta E \approx 2.6$ dB for SWF at level 0 and $\Delta E \approx 1.4$ dB at level 1.

B. *Source width test*

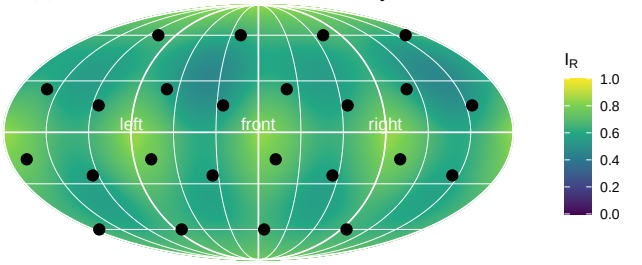
The second test aims at evaluating the apparent source width, with respect to a reference and an anchor. To measure how the width of the source is perceived, subjects had to evaluate the stimuli in comparison to a completely decorrelated reference (XL) and a single-loudspeaker reference (XS), for the six positions listed in Table III. In this case the stimuli are continuous pink noise. The subjects could listen to the stimuli



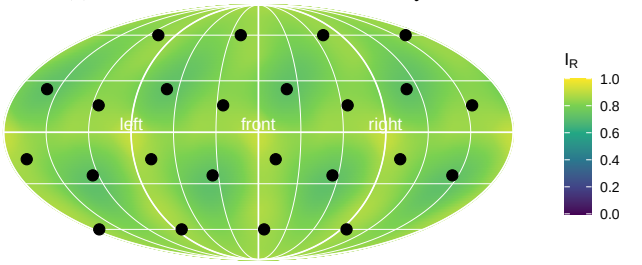
(a) Reconstruction of radial intensity for Ambisonics at order 1.



(b) Reconstruction of radial intensity for Ambisonics at order 3.



(c) Reconstruction of radial intensity for SWF at level 0.



(d) Reconstruction of radial intensity for SWF at level 1.

Fig. 7: Comparison of radial intensity reconstruction performances over the sphere for Ambisonics and SWF decoders generated with IDHOA. The black dots represent the loudspeakers' location. The reconstructed radial intensity correlates well with the apparent source width: SWF at level 0 performs better on average than Ambisonics at order 1, and SWF at level 1 and Ambisonics at order 3 should perform similarly. Also, Ambisonics performance is more uniform across the sphere than SWF is.

and references as many times as needed.

To help visually with the evaluation, the width of the stimuli is represented by circles of different diameter (Figure 5b). The narrowest reference sound (XS) is compared to the smallest circle, and the widest reference sound (XL) to the biggest circle.

The labeling approach is different from the standard MUSHRA tests. To resemble the width size, labels have been chosen to be: *Extra small*, *Small*, *Medium*, *Large* and *Extra large* (from the narrowest to the largest size possible).

V. Results

With a panel of 18 listeners, we obtained 162 evaluations for the localization test and 108 for the source width test (18 for each source position). For the localization test, the raw data has been filtered to keep only those evaluations where the reference is evaluated above 95, and the anchor below 40 MUSHRA points. For the source width test, the evaluation of the XS reference had to be below 5 MUSHRA points and the XL above 95. After filtering the raw data, 133 evaluations for the localization and 90 for the source width are left. This process of filtering is known as post-screening. The median and interquartile range (IQR) for each test and stimuli are shown in Table IV.

For the localization test (Figure 8a), third-order Ambisonics reaches *fair* with a negative difference of 12 MUSHRA points from first-level SWF. In its zeroth level, SWF is rated similarly to third-order Ambisonics, but with a bigger spread of the data. First-order Ambisonics is the worst rated technology, getting an evaluation of *poor* with a median of 36 MUSHRA points.

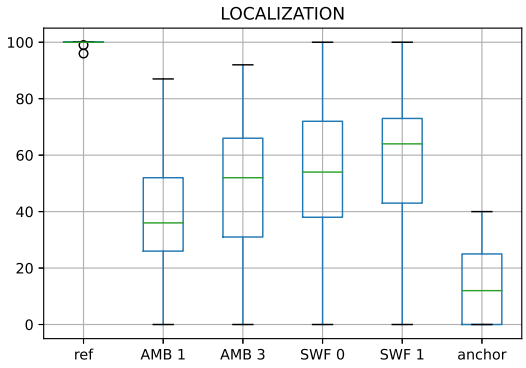
In the source width test (Figure 8b) both third-order Ambisonics and first-level SWF are rated as *good*, with a minimal difference of 3.5 points. Both zeroth-level SWF and first-order Ambisonics get a score of *fair* with a median of 55 and 46.5 MUSHRA points respectively.

Subjects have rated first-level SWF as the best technology both for localization and source width reproduction, getting in both cases a score of *good*, with a median value of 64 in the localization and of 68.5 for the source width.

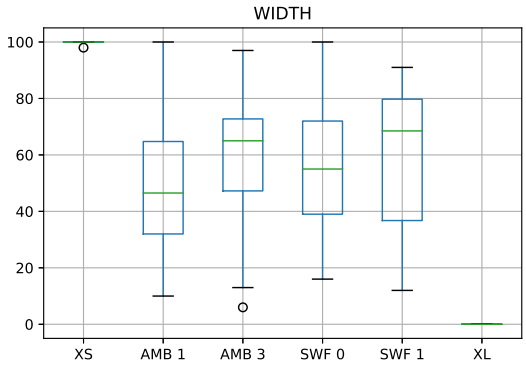
Figure 9 shows the results as a difference MUSHRA using first order Ambisonics as the reference method. This type of plot clearly highlight the relative differences between the 4 renderings of the stimuli. For the localization test, a gap arises between SWF and third order Ambisonics, with a significant difference of 10 points with level zero and even bigger for first level. The difference between SWF and third order Ambisonics for the source width test is not meaningful enough.

TABLE III: Positions (azimuth, elevation) for each group and test type, in degrees.

	Localization test		**Source width test**	
Group	**Azimuth (º)**	**Elevation (º)**	**Azimuth (º)**	**Elevation (º)**
g1	−25.5	15.5	−25.5	15.5
g2	−158.7	25.0	−158.7	25.0
g3	6.2	−60.0	6.2	−60.0
g4	−34.4	−25.0	−77.5	−15.5
g5	−77.5	−15.5	12.5	−15.5
g6	12.5	−15.5	−19.3	60.0
g7	−124.4	−25.0	—	
g8	21.3	25.0	—	
g9	−19.3	60.0	—	

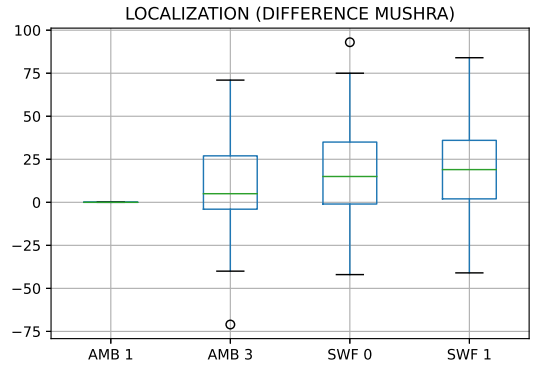


(a) Results for the localization MUSHRA tests.

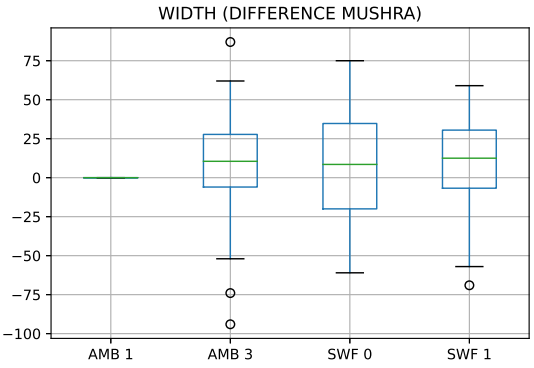


(b) Results for the source width tests.

Fig. 8: Results for the MUSHRA tests. The results are shown with the boxplot graphical method, a standardized method of graphically representing a series of numerical data through their quartiles. The blue box represents the range between the 25th percentile (Q_1) and the 75th percentile (Q_3). The median of the data (Q_2) is marked with the green line. The lines extending from the box are called whiskers, and they mark the maximum or minimum value of the series as long as they do not exceed 1.5 times the interquartile range (IQR). If so, $1.5 \cdot IQR$ is added or subtracted to the box limits, respectively, to the 25th percentile and to the 75th percentile. Any data not included between the whiskers is plotted as an outlier with a small circle.



(a) Results for localization with difference MUSHRA.



(b) Results for the source width with difference MUSHRA.

Fig. 9: Results for the tests displayed as difference MUSHRA, with the first order Ambisonics as reference.

TABLE IV: Summary of the obtained results, detailing the number of evaluations after post-scanning (N), median and interquartile ranges (IQR). Median and IQR expressed in MUSHRA points (0-100).

	Localization tests			**Source width tests**		
	N	**Median**	**IQR**	***N***	**Median**	**IQR**
Ref/XS	133	100.0	0.0	90	100.0	0.0
Anchor/XL	133	12.0	25.0	90	0.0	0.0
AMB 1	133	36.0	26.0	90	46.5	32.8
AMB 3	133	52.0	35.0	90	65.0	25.5
SWF 0	133	54.0	34.0	90	55.0	33.0
SWF 1	133	64.0	30.0	90	68.5	43.0

Since the distribution of the data cannot be assumed to be normal, a two-sided Wilcoxon test [28] has been performed with the `Stats` module from the `Scipy` library [29] and the `Pingouin` statistical package [30], to assess if there is significant difference in the evaluation of the different technologies. The Wilcoxon test can be understood as the non-parametric equivalent to the paired t-test.

The comparisons tested with this method are: level 0 SWF vs first order Ambisonics, level 0 SWF vs third order Ambisonics, level 1 SWF vs third order Ambisonics. To handle the multiple comparisons problem, the Holm–Bonferroni method [31] has been applied to the results of the statistical test. Table V shows p-values and the difference in MUSHRA points (diff) for the compared technologies. Setting a significance threshold for the p-value at 0.05, the Wilcoxon test shows that the difference in favour of SWF for the three comparisons in the localization tests are statistically significant, whereas the difference in the source width test are not, due to the wider spread of the results in the latter case.

TABLE V: p-values obtained with the Wilcoxon test and median difference (diff) on MUSHRA points (SWF - AMB). An asterisk marks those differences which are statistically significant once the correction for multiple comparisions has been applied.

	Localization test		**Source width test**	
	p**-value**	**diff**	p**-value**	**diff**
SWF 0 vs. AMB 1	9.1×10^{-10}*	18.0	0.05	8.5
SWF 0 vs. AMB 3	2.2×10^{-2} *	2.0	0.38	−10.0
SWF 1 vs. AMB 3	3.5×10^{-4} *	12.0	0.99	3.5

VI. Conclusions and discussion

A set of MUSHRA listening tests have been designed to compare and assess two spatial audio techniques, Ambisonics and SWF, in a common full-sphere setup. To our knowledge this is the first study addressing the auditory properties of the novel wavelet-based SWF format.

Subjective listening tests have shown that zeroth and first level of SWF performs better than third order Ambisonics in terms of the accuracy in source localization, marginally in the case of zeroth level SWF and moderately in the case of first level SWF, but in both cases within the domain of statistical significance. In terms of source width, there was a wider variability of the results and the results are statistically inconclusive.

It is remarkable that SWF at level 0, with only 6 channels, was rated better in our localization tests than 3rd order Ambisonics, with 16 channels. However, the points in which we evaluated the two spatial audio techniques coincided with the position of loudspeakers in the setup. Whereas the performance of Ambisonics is largely independent of the presence of a loudspeaker in the region evaluated, SWF may have a small performance boost next to a loudspeaker, so SWF may have had a slight advantage in this comparison.

Conversely, tests were performed in a full-sphere layout designed specifically for Ambisonics, so in this regard Ambisonics may have had an advantage in the test. Most real-world layouts are hemispherical and irregular from the Ambisonics point of view. It remains to be seen if the advantage of SWF with respect to Ambisonics is increased in real-world layouts.

Future work could consist in repeating a similar experiment, but considering rendering to points far from loudspeakers intervening in the rendering, evaluating other layouts representative of real-world setups, and also evaluating the performance of SWF with moving sources, and, more generally, with realistic spatial audio mixes.

References

[1] Franz Zotter and Matthias Frank. *Ambisonics*. Springer International Publishing, 2019.

[2] Davide Scaini. Wavelet-based spatial audio framework : from ambisonics to wavelets: a novel approach to spatial audio. *TDX (Tesis Doctorals en Xarxa)*, dec 2019.

[3] Davide Scaini and Daniel Arteaga. Wavelet-Based Spatial Audio Format. *J. Audio Eng. Soc*, 68(9):613–627, 2020.

[4] ITU-R BS.1534-3. Method for the subjective assessment of intermediate quality level of audio systems (mushra). *International Telecommunication Union*, 3:34, 2015.

[5] Franz Zotter and Matthias Frank. All-round ambisonic panning and decoding. *J. Audio Eng. Soc*, 60(10):807–820, 2012.

[6] Franz Zotter and Matthias Frank. Ambisonic decoding with panning-invariant loudness on small layouts (allrad2). In *Audio Engineering Society Convention 144*, May 2018.

[7] Franz Zotter, Hannes Pomberger, and Markus Noisternig. Ambisonic Decoding with and without Mode-Matching: A Case Study Using the Hemisphere. In *2nd Int. Symposium on Ambisonics and Spherical Acoustics*, pages –, Paris, France, May 2010. cote interne IRCAM: Zotter10a.

[8] Bruce Wiggins, Iain Paterson-Stephens, Val Lowndes, and S Berry. The design and optimisation of surround sound decoders using heuristic methods. 04 2003.

[9] Bruce Wiggins. *An Investigation into the Real-Time Manipulation and Control of Three-Dimensional Sound Fields*. PhD thesis, 01 2004.

[10] David Moore and Jonathan Wakefield. The design and detailed analysis of first order ambisonic decoders for the itu layout. 05 2007.

[11] David Moore and Jonathan Wakefield. Designing ambisonic decoders for improved surround sound playback in constrained listening spaces. In *Audio Engineering Society Convention 130*, May 2011.

[12] K.W.K. Cheung and P.W.M. Tsang. Development of a re-configurable ambisonic decoder for irregular loudspeaker configuration. *IET Circuits, Devices & Systems*, 3(4):197–203, aug 2009.

[13] Aaron Heller, Eric Benjamin, and Richard Lee. Design of ambisonic decoders for irregular arrays of loudspeakers by non-linear optimization. 2, 01 2010.

[14] Aaron Heller, Eric Benjamin, and Richard Lee. A toolkit for the design of ambisonic decoders. *Linux Audio Conference*, 2012.

[15] Davide Scaini and Daniel Arteaga. Decoding of Higher Order Ambisonics to Irregular Periphonic Loudspeaker Arrays. In *Proceedings of the AES International Conference*, volume 2014, 2014.

[16] Davide Scaini and Daniel Arteaga. An evaluation of the IDHOA Ambisonics decoder in irregular planar layouts. In *AES convention 138*, May 2015.

[17] Maarten Jansen and Patrick Ooninc. *Second Generation Wavelets and Applications*. Springer, London, 2005.

[18] Wim Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM Journal on Mathematical Analysis*, 29(2):511–546, Mar. 1998.

[19] Charles Loop. *Smooth Subdivision Surfaces Based on Triangles*. PhD thesis, January 1987.

[20] ITU-R BS.1116-3. Methods for the subjective assessment of small impairments in audio systems BS Series Broadcasting service (sound). *International Telecommunication Union*, 3, 2015.

[21] ROLI. Juce. https://juce.com/, 2004. Accessed: 2021-05-15.

[22] Rubén Eguinoa. mushratests. https://github.com/iRubec/mushraTests, May 2021.

[23] Archontis Politis. *Microphone array processing for parametric spatial audio techniques*. Doctoral thesis, School of Electrical Engineering, 2016.

[24] Archontis Politis. Higher order ambisonics (hoa) library. https://github.com/polarch/Higher-Order-Ambisonics, 2015.

[25] Archontis Politis. Vector base amplitude panning library. https://github.com/polarch/Vector-Base-Amplitude-Panning, 2015.

[26] Davide Scaini. Swf data. https://github.com/davrandom/swf/, April 2020.

[27] Gary Kendall. The decorrelation of audio signals and its impact on spatial imagery. *Computer Music Journal*, 19, 12 1996.

[28] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[29] Pauli Virtanen et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[30] Raphael Vallat. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026, 2018.

[31] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.