# Dynamic Higher-Order Stereophony

Jacob Hollebon and Filippo Maria Fazi

***Abstract*—Higher-Order Stereophony is a new spatial audio approach which extends classic stereo to higher order soundfield reproduction and generalised loudspeaker arrays, in a similar manner to Higher-Order Ambisonics. Higher-Order Stereophony reproduces the soundfield accurately across a line only, which must be orientated to align with the listener's interaural axis. This work introduces Dynamic Higher-Order Stereophony, which expands the technique to include listener tracking using dynamic amplitude panning. This means the soundfield is reproduced across a dynamically moving line dependent on the listener's orientation, to ensure correct reproduction of a desired set of binaural signals. A number of classic stereo approaches are shown to be special cases of first order Stereophony, and decoders to reproduce Higher-Order Ambisonics content using the new technique presented. A listening test comparing Higher-Order Stereophony and Ambisonics reproduction, low-passed at 4 kHz to reduce spatial aliasing artefacts, shows that the former technique can perform equally well as Ambisonics to the same truncation order with respect to positional properties of a virtual sound source, while using a smaller number of loudspeakers. The approach can also produce rear virtual sources using only loudspeakers positioned in the front of the listener, however with the requirement of listener tracking.**

***Index Terms*—Spatial Audio, Panning, Stereophony, Higher-Order Stereophony (HOS), Higher-Order Ambisonics (HOA).**

## I. Introduction

SPATIAL audio reproduction systems aim to reproduce the acoustic illusion of virtual acoustic scenes to a listener. These systems may be classified into three categories [1]; panning laws (for example stereophony [2], [3], Vector Base Amplitude Panning [4]), soundfield reproduction (for example Higher-Order Ambisonics (HOA) mode matching [5]–[7]) and binaural reproduction (headphone-based [8] or Crosstalk Cancellation (CTC) over loudspeakers [9], [10]). Some techniques span multiple of these categories, such as HOA mode matching which is both soundfield reproduction and a panning approach [5]. It has also been demonstrated that at low frequencies a number of spatial audio techniques (including the stereo sine and tangent law, head-tracked stereo, crosstalk cancellation and Ambisonics) all perform first order soundfield reproduction and are subsets of First Order Ambisonics [11], [12]. Notably, the stereo approaches were shown to reproduce the soundfield to a first order approximation correctly across a line only, which is aligned with the listener's interaural axis. As First Order Ambisonics has a generalised extension to higher order reproduction, which extends the accuracy of the reproduced soundfield to a higher frequency or step size from the reproduction point, the research question was posed 'Does a generalised higher order extension to stereophony also exist?'

Such an approach, titled Higher-Order Stereophony (HOS), has been recently proposed [13]. HOS is a soundfield reproduction technique that reproduces the soundfield accurately across a line, unlike HOA which ensures accurate reproduction over a circle or sphere in 2D/3D respectively. This approach was derived using the Taylor expansion of a soundfield, where a loudspeaker array was used to reproduce the soundfield with respect to its derivatives about an expansion point. A key assumption is that the reproduction line coincides with the listener's interaural axis, and that this will lead to the correct binaural signals. The resulting loudspeaker gains are panning functions, and the approach generalises the classic stereo sine law (a first order HOS system) to higher order reproduction and any generalised loudspeaker array. HOS was also demonstrated to be intrinsically linked to HOA, extending the relationship previously formed beyond first order.

A limiting factor with HOS is that the listener's head orientation must be fixed. Rotations or translations by the listener will lead to the accurately reproduced line of soundfield not coinciding with the listener's interaural axis, which will introduce errors into the final binaural signals. It also means listener-induced dynamic localisation cues can not be recreated, which are important for resolving issues such as front-back confusions [14], [15]. Listener tracking has been applied to classic stereophony before, most notably through the head-tracked sine law (also known as Compensated Amplitude Panning) [16]–[18]. The head-tracked sine law dynamically adjusts the loudspeaker gains based on the listener orientation, ensuring first order soundfield reproduction across a line defined by the listener's ears. Loudspeaker array compensation assuming channel-based stereo content has also been proposed [19]. However, with no access to the mono audio objects the compensation was defined for a central virtual source only, and cannot be assumed to hold for other positions.

The main contribution of this paper is the extension of HOS to include listener tracking, titled Dynamic HOS. Through this extension, the line of accurately reproduced soundfield is dynamically adjusted to account for listener movements. Decoders that transition between HOA and HOS are also extended to account for the dynamic scenario. This extension reveals that a number of existing classic stereo techniques are first order Dynamic HOS systems, which further motivates the naming of the technique as Higher-Order Stereophony.

A secondary contribution is the comparison of HOS to HOA through a listening test. HOS requires less loudspeakers and simpler loudspeaker array geometries compared to HOA. Thus the test compares HOA and HOS to the same truncation order

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

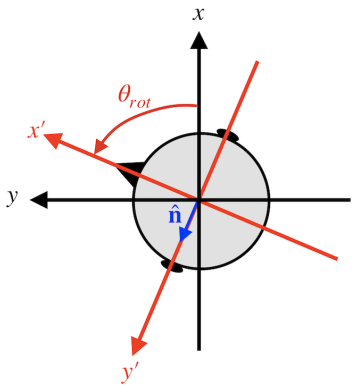



Fig. 1: Listener orientation with free rotation.

(but less loudspeakers with HOS) and the same number of loudspeakers (but a higher order of reproduction with HOS).

The article is arranged as follows. First the mathematical soundfield representation is presented and gain definitions utilising the HOS approach are extended to include listener compensation. The instability condition, an issue which arises due to certain loudspeaker array geometries and head orientations, is then discussed. A set of example HOS loudspeaker systems are considered and a number of classic stereo techniques are shown to be special cases of first order HOS systems. Decoders to transition between HOA and HOS are also extended to the dynamic case. Finally a listening test is presented to subjectively compare HOS of varying orders to HOA utilising loudspeaker-based reproduction.

## II. Dynamic Higher-Order Stereophony Order Matching

### A. Expansion Along A Generalised Axis

The HOS approach is to represent and reproduce a soundfield accurately along a single line only, in the expectation that if the line of reproduction aligns with the interaural axis this will lead to the correct reproduction of a desired set of binaural signals. Previously, the approach was derived through using the Taylor expansion of a plane wave soundfield to form a set of order-matching equations, leading to loudspeaker gain panning functions which reproduced the soundfield along this fixed line [13]. However, in this previous work the line was fixed to either the $x$ or $y$ axis of a static coordinate system, which means the listener must remain in a fixed position (with their ears aligned on said axis).

Here, the approach will now be generalised to allow the listener to move and rotate their head leading to a set of dynamic panning functions that dynamically change the direction of the line along which the soundfield is correctly reproduced, ensuring it always aligns with the listener's interaural axis. The Taylor expansion of a single plane-wave sound field is presented, expressions of the reproduced and target sound fields are derived by means of the Taylor expansion, and the order-matching approach is applied to derive the loudspeaker gains. This derivation considers the single target plane wave to be reproduced is a single sound object at a specific location in space, i.e. as a panning function to create a single phantom source. The method can be extended to the more general case when the sound field to be reproduced is more complex and is described by means of its Taylor series coefficients, or spherical harmonic coefficients, as will be made clear by the comparison to HOA in Section V).

Consider a 2D coordinate system defined by radial coordinate, $r$ and azimuthal angle, $\theta$. As in previous work defining the HOS approach, the derivation is simplest in 2D and may be later expanded to consider the 3D scenario. Let a listener be positioned with their head centred at the origin according to Fig. 1. The vector $\hat{\mathbf{n}}$ points from the head centre, $\mathbf{r}_c$ to the left ear, $\mathbf{r}_l$, defining the interaural axis, with the assumption that the listener's ears are diametrically opposed across the head. With $a$ the listener's head radius the ear positions are thus $\mathbf{r}_{l,r} = \pm a\hat{\mathbf{n}}$. The interaural axis, $\hat{\mathbf{n}}$, is defined by the head rotation angle $\theta_{\text{rot}}$, as the listener is allowed to rotate their head freely. A second frame of reference, $\mathbf{x}'$ and $\mathbf{y}'$ is defined as the listener's frame of reference such that $\hat{\mathbf{x}}'$ always points straight in front of the listener and $\hat{\mathbf{y}}' = \hat{\mathbf{n}}$, that is the interaural axis always coincides with the $\mathbf{y}'$ axis.

Let the soundfield, $p(\mathbf{r})$, be an infinitely differentiable function at a point $\mathbf{r}_0$ where $\mathbf{r} = [x, y]^T$. The multi-variable Taylor expansion of this soundfield about a position $\mathbf{r}_0$ is [20]

$$p(\mathbf{r}) = \sum_{n=0}^{\infty} \frac{[a\hat{\mathbf{n}} \cdot \nabla]^n}{n!} p(\mathbf{r}_0) \tag{1}$$

where the step from the expansion point $\mathbf{r} - \mathbf{r}_0 = a\hat{\mathbf{n}}$. Previously, HOS was defined by evaluating this expression assuming the soundfield is an incident plane wave. Thus $p(\mathbf{r}) = e^{j\mathbf{k}_i \cdot \mathbf{r}}$ with $\mathbf{k}_i = k[\cos(\theta_i), \sin(\theta_i)]^T$, $\theta_i$ the incident angle of the plane wave, $k$ the wavenumber and $j$ the imaginary unit. For brevity the plane wave is normalised such that $p(\mathbf{r}_0) = 1$. Assuming that the listener's head aligns with the $y$ axis such that $\hat{\mathbf{n}} = \hat{\mathbf{y}}$ leads to the HOS expansion

$$p(ka) = \sum_{n=0}^{\infty} \frac{[jka\sin(\theta_i)]^n}{n!} \quad \text{if } \hat{\mathbf{n}} = \hat{\mathbf{y}}. \tag{2}$$

Truncation of this infinite summation is made to an order $N$, and the $n$-th order term is given by the $n$-th order derivative of the soundfield, which for a plane wave includes the sine of the incident angle to the power of $n$. For accurate reproduction to a higher $ka$ value (combining frequency and distance from the expansion point) then higher order terms are required as in HOA [7]. The mathematics holds for any generalised step $a$, however here it is conceptualised as the listener's head radius for the context of representing the soundfield at the listener's ears (that is, binaural signals). Thus the theory is presented here considering the point at which the listener's ears are situated however holds for any step size along the line. The simple representation above is only given when the expansion is made across the $y$ axis (requiring the listener to be aligned with $\hat{\mathbf{n}} = \hat{\mathbf{y}}$). Setting $\hat{\mathbf{n}} = \hat{\mathbf{x}}$ results in an identical representation except the sine terms become cosines, namely

$$p(ka) = \sum_{n=0}^{\infty} \frac{[jka\cos(\theta_i)]^n}{n!} \quad \text{if } \hat{\mathbf{n}} = \hat{\mathbf{x}}. \tag{3}$$

For any $\hat{\mathbf{n}}$ direction between these $x$ and $y$ axes, the evaluation is more complex due to the following product

$$[a\hat{\mathbf{n}} \cdot \nabla]^n = a^n \left[ n_x \frac{\partial}{\partial x} + n_y \frac{\partial}{\partial y} \right]^n \tag{4}$$

which considers the soundfield in any generalised direction $\hat{\mathbf{n}}$, but requires the evaluation of many cross-derivative products. To account for rotations of the listener's head, it might first be obvious to change the definition of $\hat{\mathbf{n}}$ such that the $n$-th order term for the multi-variable Taylor expansion will be dependant on $[\hat{\mathbf{n}} \cdot \nabla]^n$. However, as $\hat{\mathbf{n}}$ may not be aligned with a Cartesian axis, a simplified representation can not be achieved. When $\hat{\mathbf{n}}$ is aligned with one Cartesian axis then only the single variable Taylor expansion is required. That is if $\hat{\mathbf{n}} = [1, 0]^T$ or $[0, 1]^T$ the multi-variable Taylor expansion collapses to the single variable expansion, as only derivatives with respect to a single axis are required.

To create simple analytical expressions for the final panning gain functions, head rotations may be compensated for in a simple manner by considering the two separate frames of reference (the fixed frame of reference and the listener's frame of reference). This is demonstrated in Fig. 1. Measuring all necessary angles in the listener's frame of reference will therefore compensate for the head rotation, $\theta_{\text{rot}}$. In practice, this is easily implemented by converting between the two frames of reference for any given angle with $\theta' = \theta - \theta_{\text{rot}}$. In doing so, the Dynamic HOS representation is always such that $\hat{\mathbf{n}} = \hat{\mathbf{y}}'$ and the expansion

$$\begin{aligned} p(ka) &= \sum_{n=0}^{\infty} \frac{[jka \sin(\theta_i')]^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{[jka \sin(\theta_i - \theta_{\text{rot}})]^n}{n!} \end{aligned} \tag{5}$$

may always be used. This means HOS can be adapted for listener head rotations by using a head-tracker and compensating all angles in the fixed frame of reference by $\theta_{\text{rot}}$.

An interesting point arises when considering head rotation compensation. As noted previously, expansion along the $x$ axis results in cosine terms in the expansion instead of sine terms. Choice of the $x$ or $y$ axis to perform the expansion across (or indeed any arbitrary axis) is equally valid and corresponds to reproduction along two orthogonal axes in a given reference frame. However, these two representations using $\sin^n(x)$ or $\cos^n(x)$ may be seen to be equal by setting $\theta_{\text{rot}} = -90°$, such that $\hat{\mathbf{n}} = \hat{\mathbf{x}}$ and any given angle $\theta' = \theta + 90°$. Using the identity $\sin(x + 90°) = \cos(x)$ this shows the two representations may be equally used, as long as the angles are defined properly. Therefore, it is valid to use either sine or cosine terms when later defining the loudspeaker gains.

So far, only listener head rotations have been considered. To compensate for translations of the listener's head, a delay may be applied to each of the loudspeaker gains to keep them acoustically equidistant. In this sense, all loudspeakers in the array can be virtually kept at a constant radius away from the listener. The delays may be formulated and implemented independently to the HOS gain definitions to maintain the frequency-independence in the HOS gain calculations. Physically, a translation of the listener's head will move the point about which the expansion is performed, which is always kept as the listener's head centre. Practically, the listener should remain in the interior of the loudspeaker array.

Importantly, so far the mathematics represents the soundfield along a single line only. Whilst the context is that this line should align with the interaural axis of a listener, the effects of a complex Head-Related Transfer Function (HRTF) (e.g head shadowing, scattering of incident waves) is not included in the mathematics. In general, soundfield reproduction approaches such as HOS and HOA aim to reproduce the incident soundfield accurately within a region. This then ensures the acoustical effects due to the presence of the listener's head are as with the original target soundfield. However, as the accurate reproduction region with HOS is across a line only, analysis with complex HRTFs is left for future work.

With the Dynamic HOS soundfield representation defined to allow for listener movements, this representation will now be used as in previous work to define a set of loudspeaker panning functions. However, these will now be adaptive panning functions that depend on the listener head orientation, reproducing the soundfield accurately across the interaural axis regardless of the listener head movements.

### B. Target Soundfield

The target soundfield is that of the incident plane wave which the loudspeaker array is attempting to reproduce. The target soundfield, $p_T(ka)$, is given by the Dynamic HOS expansion from (5)

$$p_T(ka) = \sum_{n=0}^{\infty} \frac{[jka \sin(\theta_T')]^n}{n!} \tag{6}$$

where the listener rotates their head freely as defined by $\theta_{\text{rot}}$ and the incident target plane wave arrives at an angle $\theta_T$, therefore $\theta_T' = \theta_T - \theta_{\text{rot}}$.

### C. Reproduced Soundfield

The reproduced soundfield is given by the overall contributions of the loudspeaker array. Consider an array of $L$ radially-equidistant loudspeakers assumed to act as plane wave sources. The $\ell$-th loudspeaker is situated at $\theta_\ell$ and driven by a gain $g_\ell$. Compensating for listener head rotations then $\theta_\ell' = \theta_\ell - \theta_{\text{rot}}$. The reproduced soundfield is

$$p_R(ka) = \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty} \frac{[jka \sin(\theta_\ell')]^n}{n!}. \tag{7}$$

### D. Order Matching

For exact soundfield reproduction the following must be satisfied, $p_T(ka) = p_R(ka)$, which is found by defining the loudspeaker gains which minimise $||p_T(ka) - p_R(ka)||_2^2$. In practice this requires an infinite array of loudspeakers, however truncation to a finite order/finite loudspeaker array will be considered later. Therefore,

$$\sum_{n=0}^{\infty} \frac{[jka\sin(\theta'_T)]^n}{n!} = \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty} \frac{[jka\sin(\theta'_\ell)]^n}{n!}. \quad (8)$$

The order matching principle is now applied, where the $n$-th order terms from the expansion of the target and reproduced soundfields are equated. Traditionally the orthogonality property of the basis functions expressed over the domain of interest is applied to achieve this matching [7]. However, without invoking any orthogonality relation, if each $n$-th term of the sum in the right-hand side of (8) is equated to the corresponding $n$-th term on the left-hand side by choosing the appropriate loudspeaker gains $g_\ell$, then the sums on both sides of the equation will be equal. It is therefore required that

$$\frac{[jka\sin(\theta'_T)]^n}{n!} = \sum_{\ell=1}^{L} g_\ell \frac{[jka\sin(\theta'_\ell)]^n}{n!} \quad \forall\ n\ \in\ \mathbb{N}_0. \quad (9)$$

Finally removing all common terms leaves

$$\sin^n(\theta'_T) = \sum_{\ell=1}^{L} g_\ell \sin^n(\theta'_\ell) \quad \forall\ n\ \in\ \mathbb{N}_0 \quad (10)$$

which is the $n$-th Dynamic HOS order matching equation. Whilst due to the lack of an orthogonality property this is not the only possible solution to these equations, this specific solution leads to the elimination of the $(ka)^n$ dependency. This ensures the solution is valid for any frequency ($k$) or spatial coordinate ($a$) which will not likely be the case for other solutions (which may only be valid for a fixed $ka$).

### *E. Loudspeaker Gain Calculation*

To calculate the loudspeaker gains, formulate the Dynamic HOS order matching equations for all orders $n \in [0, N]$ as a set of linear equations. Truncation of the expansion to a finite order $N$ is now performed, with the assumption that $L \geq N+1$ to ensure an exact solution to the problem can be calculated. Let $\mathbf{p}_T$ be a length $(N+1)$ vector of target signals, $\Psi$ be an $(N+1) \times L$ plant matrix and $\mathbf{g}$ be a length $L$ vector of loudspeaker gains. The $n$-th entry of $\mathbf{p}_T$ is the $n$-th order term of the target signal, given by $\sin^n(\theta'_T)$. The $n$-th row and $\ell$-th column entry of $\Psi$ is the $n$-th order contribution due to the $\ell$-th loudspeaker, which is $\sin^n(\theta'_\ell)$. The $\ell$-th entry of $\mathbf{g}$ is the gain for the $\ell$-th loudspeaker, $g_\ell$.

$$\begin{aligned}
\mathbf{p}_T &= \begin{bmatrix} 1 & \sin(\theta'_T) & \sin^2(\theta'_T) & \dots & \sin^n(\theta'_T) \end{bmatrix}^T \\
\Psi &= \begin{bmatrix} 1 & 1 & \dots & 1 \\ \sin(\theta'_1) & \sin(\theta'_2) & \dots & \sin(\theta'_L) \\ \sin^2(\theta'_1) & \sin^2(\theta'_2) & \dots & \sin^2(\theta'_L) \\ \vdots & \vdots & \ddots & \vdots \\ \sin^N(\theta'_1) & \sin^N(\theta'_2) & \dots & \sin^N(\theta'_L) \end{bmatrix} \\
\mathbf{g} &= \begin{bmatrix} g_1 & g_2 & g_3 & \dots & g_L \end{bmatrix}^T.
\end{aligned} \quad (11)$$

The gains are found by solving an inverse problem:

$$\mathbf{p}_T = \Psi\mathbf{g} \implies \mathbf{g} = \Psi^\dagger \mathbf{p}_T \quad (12)$$

where the superscript $(\cdot)^\dagger$ indicates the Moore-Penrose pseudoinverse, which is commonly used to solve similar sets of linear equations in spatial audio reproduction [5], [7]. The problem is overdetermined when $(N+1) > L$ and as here an exact solution cannot be found, the pseudoinverse gives the least-squares solution that minimises the error between the target and reproduced soundfield. Alternatively when $(N+1) \leq L$ an infinite number of exact solutions exist, therefore the pseudoinverse chooses the minimum norm solution with respect to the $\mathrm{L}^2$ norm [21]. By performing this inversion, the loudspeaker panning functions may be calculated for any given head orientation and virtual source position. Furthermore, the loudspeaker gains have no frequency-dependence and therefore form dynamic panning functions.

## III. The Instability Condition

For every incident plane wave virtual source impinging from a given direction in 2D there exists a second virtual source position, reflected about the reproduction line, which creates an identical soundfield across said line. This is the 2D variant of the cone of confusion. In practice, this means the resulting HOS loudspeaker gains are equal for both positions. If the HOS loudspeaker array uses loudspeakers in front of the listener only, this may be considered as reproducing rear virtual sources through the equivalent position in front of the listener. Whilst this may lead to front-back ambiguities, it is expected that adding dynamic localisation cues through the Dynamic HOS approach will help resolve such issues, thus creating convincing virtual sources behind the listener using only loudspeakers in front of them.

Whilst HOS takes advantage of the cone of confusion for the virtual sources [13], it can lead to issues when defining the geometry of the loudspeaker array. As HOS only accounts for the contribution of the loudspeakers across the reproduction axis, $\sin(\theta - \theta_{\mathrm{rot}})$, for each loudspeaker position $\theta_i$ another position $\theta_j$ exists (in 2D) that results in the same soundfield along this axis. If expanded to 3D, for example loudspeakers with elevation, the locus of these positions becomes a ring (corresponding to the cone of confusion). Importantly, the soundfield from all of these positions is equal along the reproduction line only, not if evaluated elsewhere. When this occurs, the HOS system loses a degree of freedom as each of the two loudspeakers cannot contribute in a unique manner, or equivalently two columns of the plant matrix $\Psi$ are identical. Practically this means the number of loudspeakers available to the system is reduced by one, and if the number of uniquely contributing loudspeakers is less than $N+1$ then an exact solution can not be achieved. Here the matrix $\Psi$ becomes singular and the values of the loudspeaker gains computed with (12) diverges. The instability condition thus occurs when

$$\sin(\theta_i - \theta_{\mathrm{rot}}) = \sin(\theta_j - \theta_{\mathrm{rot}})\ \forall\ i, j\ \in [1, L]. \quad (13)$$

This issue is easily avoidable if the listener's head is fixed with respect to rotation as with the original HOS derivation [13], as the HOS loudspeaker system may be defined such that all loudspeakers contribute to the axis uniquely. However,

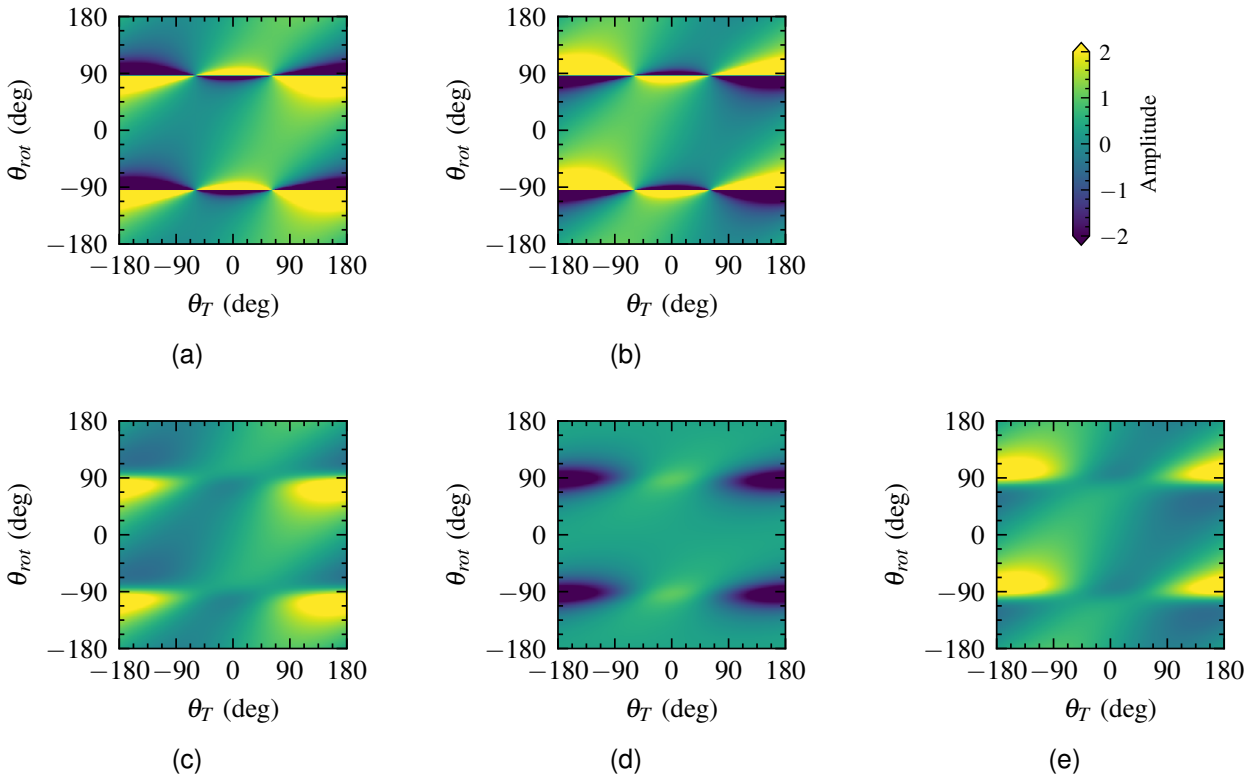



Fig. 2: Gains for two $N = 1$ loudspeaker arrays clipped to the range $\pm 2$ for illustrative purposes (a-b) Left right (LR) symmetric pair (c-e) Left centre right (LCR) trio.

issues arise when allowing the listener to rotate their head freely in Dynamic HOS, that is $\theta_{\text{rot}} \in [0°, 360°]$. For every pair of loudspeakers there exists a head position that gives rise to the instability, defined as

$$\theta_{\text{rot}}^{\text{instability}} = \arctan\left(\frac{\sin(\theta_j) - \sin(\theta_i)}{\cos(\theta_j) - \cos(\theta_i)}\right). \tag{14}$$

Note that this condition is valid for all permutations of every pair of loudspeakers used by the system, which for a large system can result in many angles where instabilities can occur.

The impact of this issue is that, to allow for full 360° listener head rotations, a minimum of $2N + 1$ loudspeakers are required, not $N + 1$. This is because, when considering symmetry about an axis, there are two given positions on the unit circle that give the same value when taking the sine (or equivalently the cosine) of those positions. Hence, if $N$-th order requires $N + 1$ loudspeakers, a maximum of $(N/2)$ of these loudspeakers can have symmetric counterparts at any one time. This means the largest possible reduction in the effective size of the loudspeaker array is reduction to size $(N/2) + 1$ loudspeakers. To ensure that there is always $N + 1$ loudspeakers available to the system (for an exact solution) there consequently must be $L = 2N + 1$ loudspeakers if full 360° head rotations are allowed.

This minimum number is the same as for an exact solution when employing 2D HOA. This is because, to allow for any head orientation, the system must be able to reproduce the sound field accurately over an area defined by all possible directions of the interaural axis (noting that HOS would only reproduce one of these lines correctly at any given time). The figure is the same as 2D HOA because this is what 2D HOA does - accurate reproduction across the region inside a circle (however all possible reproduction lines at the same time). An important difference between the two approaches is that 2D mode matching HOA ideally requires these $(2N + 1)$ loudspeakers distributed evenly across the circle [22]. However, this restriction does not exist for HOS which can handle more irregular loudspeaker distributions.

To overcome the instability condition careful design of the system and loudspeaker array is required. Tikhonov regularisation can be employed in the pseudoinverse to limit the effects of ill-conditioning that occurs when approaching an instability [21]. This allows for the minimal number of $L = N + 1$ loudspeakers to still be used, but introduces an error into the solution whilst at an instability position one of the order terms will also not be correctly reproduced. Alternatively, the number of loudspeakers used can be $L > N + 1$, therefore allowing for up to $L - N - 1$ loudspeakers to become unstable whilst ensuring an exact solution can still be achieved. Furthermore, if a smaller fixed range of head rotations is only required, then the loudspeaker layout can be optimised for that head rotation span to minimise the number of instabilities by using (14). This might occur naturally, for example, if the listener is watching content on a screen.

To demonstrate the issue of the instability condition, consider the simplest case of a first order system with two

loudspeakers situated at $\theta_1 = \theta, \theta_2 = -\theta$ (a symmetric left right (LR) pair). As per (14) there should be two instabilities which occur when $\theta_{\text{rot}} = \pm 90°$. The loudspeaker gains for this system across all virtual source incident angles and head rotations, with $\theta = 60°$, are shown in Fig. 2. At the instability when $\theta_{\text{rot}} = \pm 90°$ the loudspeaker gains tend towards infinity. This issue may be avoided by adding a third loudspeaker. Thus, consider a left centre right (LCR) system with a third loudspeaker at $\theta_3 = 0°$. Using this system but only to first order means the problem is underdetermined as $L > N + 1$, however avoids the instability as at $\theta_{\text{rot}} = \pm 90°$ there are still two uniquely contributing loudspeakers and an exact solution exists. Whilst the gains still increase around $\theta_{\text{rot}} = \pm 90$, they do not diverge to infinity. This also prevents the sudden $180°$ phase change that occurs in the 2-channel system when the listener's head passes by the instability positions.

## IV. Example Higher-Order Stereophony Systems

This section will introduce the gain definitions for a generalised first order Dynamic HOS system. Then, under certain conditions, a number of classic stereo approaches will be shown to be a subset of this generalised first order HOS approach. The minimum number of loudspeakers required is $L = N + 1 = 2$. Consider two loudspeakers positioned at arbitrary angles $\theta_1$ and $\theta_2$. In this case the target pressure vector, plant matrix and loudspeaker gains are

$$\begin{aligned} \mathbf{P}_T &= \begin{bmatrix} 1 \\ \sin(\theta'_T) \end{bmatrix}, \Psi = \begin{bmatrix} 1 & 1 \\ \sin(\theta'_1) & \sin(\theta'_2) \end{bmatrix} \\ \mathbf{g} &= \frac{1}{\sin(\theta'_2) - \sin(\theta'_1)} \begin{bmatrix} \sin(\theta'_2) - \sin(\theta'_T) \\ -\sin(\theta'_1) + \sin(\theta'_T) \end{bmatrix}. \end{aligned} \tag{15}$$

This is the head-tracked stereo sine law for a generalised loudspeaker geometry. The head-tracked stereo sine law has been previously derived through the work on Compensated Amplitude Panning (CAP) [16]–[18]. Furthermore, the low-frequency approximation of the CTC technique for two loudspeakers leads to the same gain definitions [11], [23].

From this system, classic stereo techniques can be derived under particular conditions. First, consider the scenario when the listener is assumed to face forward ($\theta_{\text{rot}} = 0$), this gives the stereo sine law for generalised loudspeaker geometries

$$\mathbf{g} = \frac{1}{\sin(\theta_2) - \sin(\theta_1)} \begin{bmatrix} \sin(\theta_2) - \sin(\theta_T) \\ -\sin(\theta_1) + \sin(\theta_T) \end{bmatrix}. \tag{16}$$

Next let the loudspeaker angles be symmetric, $\theta_1 = \theta = -\theta_2$, as with a traditional stereo loudspeaker arrangement. For arbitrary head rotations the loudspeaker gains are

$$\mathbf{g} = \frac{1}{2} \begin{bmatrix} 1 + \frac{\tan(\theta_{\text{rot}})}{\tan(\theta)} + \frac{\sin(\theta_T)}{\sin(\theta)} - \frac{\cos(\theta_T)\tan(\theta_{\text{rot}})}{\sin(\theta)} \\ 1 - \frac{\tan(\theta_{\text{rot}})}{\tan(\theta)} - \frac{\sin(\theta_T)}{\sin(\theta)} + \frac{\cos(\theta_T)\tan(\theta_{\text{rot}})}{\sin(\theta)} \end{bmatrix}. \tag{17}$$

This is the head-tracked stereo sine law [16]. Setting the listener to face forwards ($\theta_{\text{rot}} = 0$) gives the traditional stereo sine law as defined in [2], [24]

$$\mathbf{g} = \frac{1}{2} \begin{bmatrix} 1 + \frac{\sin(\theta_T)}{\sin(\theta)} \\ 1 - \frac{\sin(\theta_T)}{\sin(\theta)} \end{bmatrix}. \tag{18}$$

Finally, consider the case when the stereo symmetric loudspeaker angles are used, except now the listener always faces the virtual source such that $\theta_{\text{rot}} = \theta_T$. This gives the stereo tangent law as in [3], [25]

$$\mathbf{g} = \frac{1}{2} \begin{bmatrix} 1 + \frac{\tan(\theta_T)}{\tan(\theta)} \\ 1 - \frac{\tan(\theta_T)}{\tan(\theta)} \end{bmatrix}. \tag{19}$$

Hence, using the Taylor expansion of a plane wave soundfield, all of the most established stereo panning techniques have been derived. This shows that these key existing stereo methods can actually be interpreted as first order Taylor approximations of the reproduction of the target plane wave soundfield across a line, with assumptions about the listener's head orientation such that the reproduction line coincides with the interaural axis. Thus stereo is actually a soundfield reproduction technique to the first order and, as shown in this work, through the Taylor expansion may be expanded to higher orders for more accurate reproduction of the soundfield. As previously all common stereo techniques have been defined mathematically as low frequency methods [12], HOS thus generalises and expands the stereo theory to any order, for any loudspeaker array and reproduction across any frequency or spatial range (as per the truncation order).

## V. Relation To Higher-Order Ambisonics

Decoders to convert from both the 2D and 3D HOA soundfield representation (the soundfield coefficients of the basis expansions) to HOS have been defined in detail in previous work [13]. An example decoder for 3D HOA to Dynamic HOS is shown in Fig. 3. This figure shows an expanded signal chain to explain all the decoder steps, but the final most efficient implementation may combine all these steps into a single gain matrix operation where the matrix coefficients depend on the listener orientation.

To adapt these decoders to be compatible with Dynamic HOS, head rotation compensation can be performed as with standard HOA by rotating the soundfield in the opposite direction to the listener rotation. The next steps for the decoders are then the same as presented previously. A second rotation is required which ensures a small subset of $(N+1)$ HOA basis functions can be used to represent the soundfield along a line only. For 2D and 3D respectively, the desired reproduction axis must be aligned with the $x$ or the $z$ axis which defines the basis functions. This results in a subset of basis functions, in 2D the set $\cos(n\theta)$ and in 3D the $m = 0$ spherical harmonics, which fully represent the soundfield along the $x$ or $z$ axis, respectively. The subsequent decimation step is the removal of all channels not corresponding to this subset of basis functions.

Next, a conversion is required to transform this subset of HOA coefficients into the equivalent HOS representation. Reproduction to the $N$-th order expressed using the HOS cosine representation requires terms of $\cos^n(\theta)$. To transform HOA to HOS the Chebyshev polynomials or the Legendre

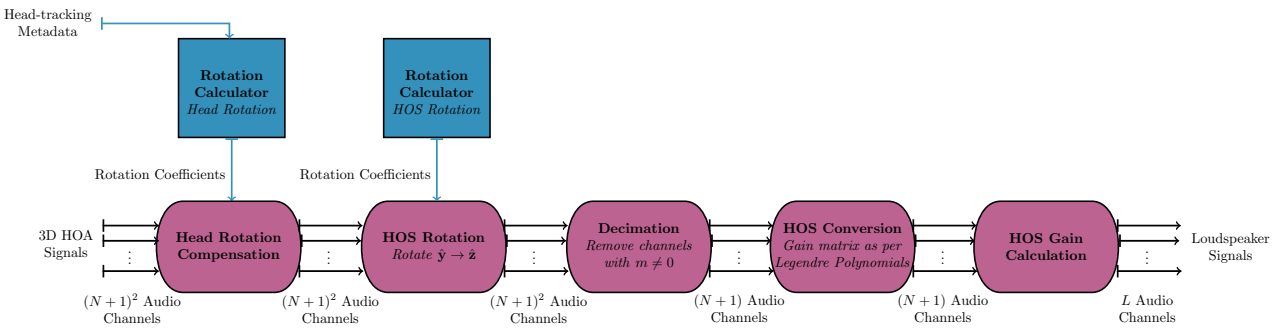
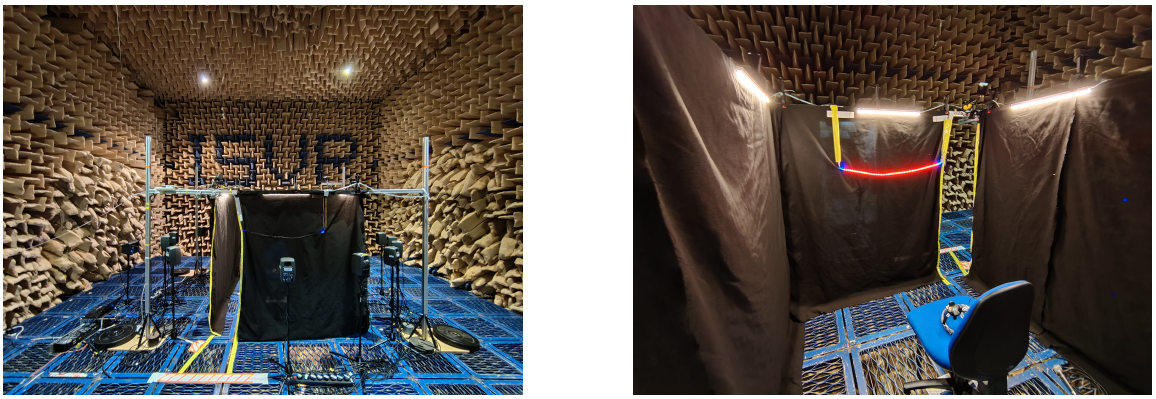



Fig. 3: Signal chain for reproducing 3D HOA signals using Dynamic HOS.



(a) (b)

Fig. 4: (a) Experimental setup for the listening test. (b) View of the subject's seat with the head rotation LED feedback active.

polynomials may be used, which each transform this specific set of $(N+1)$ HOA basis functions into a polynomial of cosine functions to the power of $n$ [13]. After this, the HOS approach may be used to create the resulting loudspeaker signals.

## VI. Comparative Listening Test

To investigate the viability of Dynamic HOS and to compare it to HOA, a modified MUltiple Stimulus with Hidden Reference and Anchor (MUSHRA) listening test was performed. The objective of the listening test was to compare Dynamic HOS to 2D HOA, to investigate perceived differences in the localisation of a virtual source between the techniques.

### A. Experimental Setup

The listening test was run in the large anechoic chamber at the Institute of Sound and Vibration Research (ISVR) to ensure free field conditions, which both HOS and HOA assume in their derivation. The experimental setup delivered real-time loudspeaker rendering of a single static virtual source about the listener in the horizontal plane. The listener was therefore surrounded by a number of loudspeakers, of which different subsets were activated for different renderers using either HOS or 2D HOA. The 2D HOA implementation used panning laws as defined by a mode matching decoder utilising circular harmonics [5]. Real loudspeakers were also placed at the chosen target virtual source positions so the listener could compare the virtual source to a real reference. All loudspeakers were hidden, as shown in Fig. 4.

The audio rendering was performed utilising the Versatile Interactive Scene Renderer (VISR) object-based audio production plugins for object and metadata control, with the VISR python environment used to render the resulting loudspeaker array signals [26]. Example real-time code for HOS is also made publicly available[1] The listener was able to seamlessly switch between each of the rendering approaches with a 5 ms fade out and 5 ms fade in of the audio using a cosine ramp [27]. The test was controlled by the subject using an iPad running TouchOSC. For the HOS approaches, listener tracking was performed using a head mounted HTC Vive tracker. The reproduction loudspeaker array consisted of Genelec 8020C loudspeakers. A normalisation process was run to ensure consistent sound levels for all real and virtual source positions and the different renderer types. Audio processing was performed at a sample rate of 48 kHz and a block size of 512 samples. The measured motion-to-sound latency was 33 ms [28].

The instability condition can lead to issues for HOS loudspeaker arrays at certain listener head orientations. Therefore, to avoid these issues the head rotation range of the listener was limited. A strip of LED pixels was used to give the listener feedback on their current head orientation as shown in Fig. 4. A block of blue pixels indicated the two limits the listener

[1] https://github.com/jacobhollebon/hos

| Renderer | Order | Num. Loudspeakers | Loudspeaker Positions (°) |
|---|---|---|---|
| HOS | 1 | 2 | $\pm 45$ |
| HOS | 1 | 2 | $\pm 90$ |
| HOS | 2 | 3 | $0, \pm 90$ |
| HOS | 3 | 4 | $\pm 45, \pm 90$ |
| HOS | 4 | 5 | $0, \pm 45 \pm 90$ |
| HOS | 5 | 6 | $\pm 45, \pm 76, \pm 90$ |
| HOA | 2 | 5 | $0, \pm 76, \pm 144$ |
| HOA | 4 | 9 | $0, \pm 45, \pm 76, \pm 120, \pm 160$ |

TABLE I: Spatial audio techniques and their configuration used in the listening test.

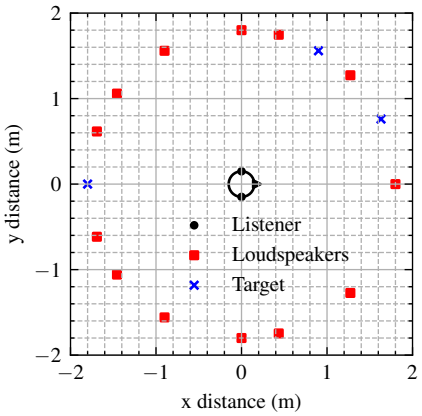



Fig. 5: Loudspeaker and target source positions.

could rotate their head within, whilst the rest of the strip flashed red when the listener exceeded these limits.

### B. Experimental Design

The renderers chosen for comparison are given in Table I. These approaches were chosen to investigate the following research questions:

1) Does increasing the truncation order for HOS increase the accuracy of the perceived virtual source position?
2) Can HOS reproduce virtual images out-of-span of the loudspeaker array or behind the listener, using loudspeakers in front only?
3) Is the accuracy of the perceived virtual source position equal for HOS and HOA renderers of the same order, but different number of loudspeakers?
4) Is the accuracy of the perceived virtual source position improved when the number of loudspeakers is fixed, leading to higher order reproduction using HOS compared to HOA?

To answer question 1, HOS systems of increasing order from 1 to 5 (HOS O1-5) were used. For question 3, two HOA systems of order 2 and 4 (HOA O2 and O4) were employed to compare to HOS at the same truncation order. To consider question 4, where a fixed number of loudspeakers is allowed, HOS O4 and HOA O2 both using the minimum requirement of 5 loudspeakers were included for comparison.

The loudspeaker positions for each system are given in Table I and the final loudspeaker array used by all systems is shown in Fig. 5. All loudspeakers were positioned at head height and radially equidistant, 1.8 m away from the listener. All renderers used a small amount of Tikhonov regularisation ($\beta = 0.01$) in the inversion to define the loudspeaker signals, to help stabilise the solutions in the presence of experimental errors (for example small misalignments of the loudspeakers).

For HOS, at each order $(N + 1)$ loudspeakers were required and chosen by sampling a semicircle in front of the listener (maximum position $\pm 90°$). To consider whether the loudspeaker span affected the performance, two HOS O1 systems with varying spans of $\pm 45°$ and $\pm 90°$ (HOS O1-45 and HOS O1-90 respectively) were included. Thus for the two matching order systems, a virtual source position larger than $45°$ meant an out-of-span source could be compared to an in-span source. To investigate question 2, whether increasing order does increase virtual source position accuracy, an increasing number of loudspeakers is required for each higher order system. Therefore, to isolate whether a difference in performance was just due to using more loudspeakers or to increasing the order of the reproduction, HOS O1-45, O3 and O5 were designed to add loudspeakers outside of $\pm 45°$ only as order increased. This meant a virtual source positioned within $\pm 45°$ could assess this hypothesis without being biased by a new loudspeaker being added closer to the virtual source.

The loudspeaker positions for the HOA systems were chosen by equally sampling a circle around the listener. The minimum number of loudspeakers, $(2N+1)$, was used. Due to practicalities in building the loudspeaker array, some positions for the HOA approaches deviated by a maximum of $5°$ from exact circular sampling. These small shifts in position should have minimal impact on performance, as the loudspeakers remained very close to the optimal sampling positions.

The highest order system used was order 5. As per the $N = kr$ rule this corresponds to a frequency limit of approximately 3500 Hz for a head radius of 8 cm above which aliasing will occur for the highest order system [7]. Furthermore, at high frequencies an intensity panning approach is often favoured over mode matching techniques [29]. Therefore, a low-pass filter with a cut-off at 4000 Hz was applied to all techniques.

### C. Stimuli

Three different source stimuli were tested at three different positions. Therefore, there were 9 pages in the test. The order of these pages was randomised for all subjects and on average the test took one hour including a break in the middle. The stimuli were looped anechoic recordings of a male speech sample, a short rock drum beat, and a flamenco acoustic guitar recording from [30]. Drums have been shown to be a particularly appropriate critical signal for listening tests, combining both broadband content and sharp transients [31]. The source positions used may be viewed in comparison to the reproduction loudspeaker positions in Fig. 5 and were

$$\theta_1 = 25°, \qquad \theta_2 = 60°, \qquad \theta_3 = 180°. \tag{20}$$

$\theta_1$ meant the effect of increasing the HOS order without adding loudspeakers close to the virtual source could be assessed for HOS O1-45, O3 and O5 (question 2). The source at $\theta_2$ meant in-span and out-of-span sources could be compared for HOS O1-45 and O1-90, while the rear position of $\theta_3$ was

chosen to see whether HOS could create the illusion of a rear virtual source using frontal loudspeakers only (question 2).

The reference sources were chosen as real loudspeakers at the same positions as the virtual sound sources, thus the participants could always compare a real source to a virtual one. The anchor was designed to be both spatially and tonally impeded. Therefore, the anchor used an equal sampling of loudspeakers positioned around the listener, with all loudspeakers active with equal gain to create a large ambiguity in the position of the sound source. The anchor was also low-passed [27], however the cut-off frequency was set to 1000 Hz. This was lower than the standards require as all audio had already been low-passed, therefore the anchor was redefined to ensure it was sufficiently impaired.

### D. Procedure

A modified MUSHRA approach focusing only on positional audio attributes was chosen for the listening test as it allows for simultaneous comparison of a large number of stimuli with respect to a reference sound source. Furthermore, inclusion of a hidden reference and anchor allows rating of the reliability and quality of each subject's results.

The participants were asked to rate the similarity of the positional properties of the virtual source created by each of the renderers, in comparison to the real reference sound source. All participants underwent a training session, during which they were encouraged to consider at least the following positional properties of the virtual source compared to the reference: absolute position, apparent source width and stability with head rotations. It was nevertheless explained to the subjects that they should provide a holistic rating. The decision to not compare tonal impairments of the audio was motivated by preliminary tests, which showed all renderers sounded significantly coloured with respect to the real reference, whilst focusing on positional properties helps answer the research questions posed above. Furthermore, the presence of the global 4 kHz low-pass filter required to minimise aliasing effects would mean any conclusions drawn from comparing colouration would be limited. The rating was on a scale from 0 to 100 with labels *'huge change'*, *'large change'*, *'moderate change'*, *'small change'* or *'no change'*. The training phase included a representative set of the renderers (HOS O1-45, HOS O5, HOA O2, HOA O4, reference, anchor) and was three pages, using the three source stimuli at the three positions.

Head rotations are key to ensuring Dynamic HOS performs properly, therefore the participants were strongly encouraged to rotate their head through-out the test and whenever they selected a new renderer. To minimise the effect of the instability condition, the participant's head rotation was limited to a $\pm 25^\circ$ span with the LEDs used to give feedback if they exceeded this value, which was active for all renderers equally.

### E. Results

24 subjects took part in the listening test, 21 of which were between the ages of 22-50 and 3 were over 50. 18 of the subjects were male and 6 female. All subjects had self-reported normal hearing. A single participant's results were discarded as they incorrectly rated the Hidden Reference with a score of less than 90 for more than 15% of the test pages.

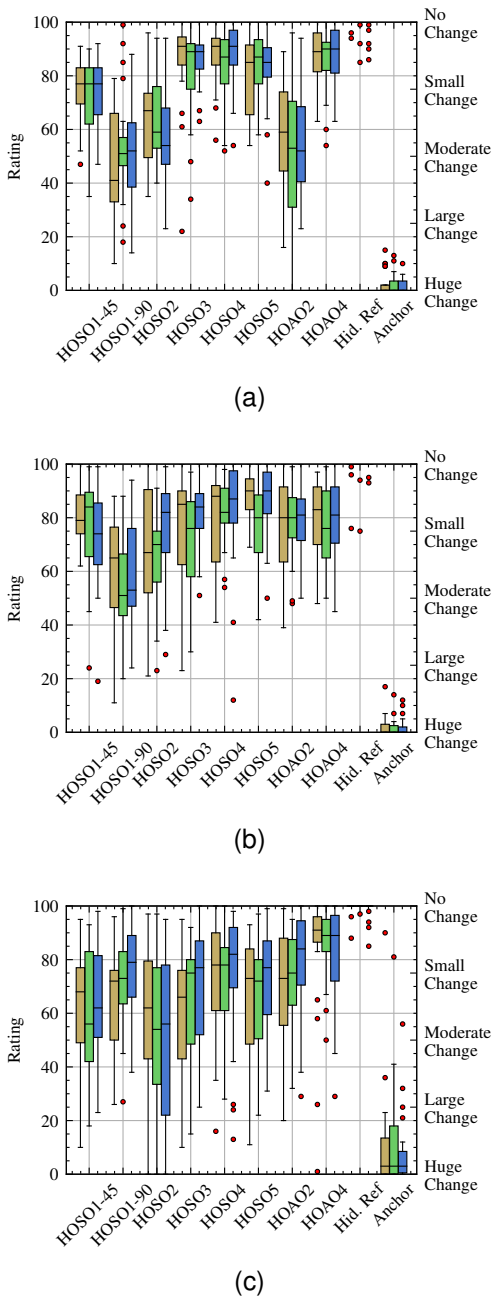



Fig. 6: Raw results with the brown, green and blue boxes representing the speech, drums and guitar signal types respectively. (a) $\theta_1 = 25^\circ$. (b) $\theta_2 = 60^\circ$. (c) $\theta_3 = 180^\circ$.

The raw data from the subject's responses along with the statistical analysis is made publicly available[2]. Boxplots in Fig. 6 show the subject's ratings of the different rendering approaches, for the three different source signals and three virtual source positions. It is clear that regardless of the source position or source stimuli, the hidden reference and anchor were both consistently identified and scored appropriately. One interesting deviation is for the anchor at the rear source

[2]https://doi.org/10.5258/SOTON/D2535

| Effect | $df$ | $F$ | $\epsilon_{HF}$ | $\eta_p^2$ | $p_{HF}$ |
|---|---|---|---|---|---|
| T | $2, 154$ | 35.449 | .660 | .617 | $< .001$* |
| S | $2, 44$ | 4.158 | 1.000 | .159 | .022* |
| P | $2, 44$ | 12.364 | .880 | .360 | $< .001$* |
| T $\times$ S | $14, 308$ | .985 | .892 | .043 | .465 |
| T $\times$ P | $14, 308$ | 16.125 | .602 | .423 | $< .001$* |
| S $\times$ P | $4, 88$ | 1.55 | .792 | .066 | .207 |
| T $\times$ S $\times$ P | $28, 616$ | 1.644 | .07 | .717 | .04* |

$\epsilon_{HF}$: Huynh-Feldt correction
$\eta_p^2$: Partial eta squared
$p_{HF}$: Huynh-Feldt corrected $p$ values
Statistical significance at 5% indicated by an asterisk

TABLE II: Within-subject effects from the three-way repeated measures ANOVA for technique, stimuli and position.

position of $180°$, which was scored higher than expected by some subjects. This is likely due to the confusing nature of the soundfield created by the anchor approach, with some participants reporting a larger concentration of energy arriving from behind them. Despite this anomaly, the impact on the overall experimental design is minimal as the overall anchor rating remained very low.

There appears to be a trend that increasing the truncation order of the HOS system leads to a higher rating. The HOS O1-45 system however does not follow this pattern, and scores consistently higher than might otherwise be expected. This is observable most clearly for $\theta_1 = 25°$ and $\theta_2 = 60°$. For the rear source at $\theta_3 = 180°$ the ratings across all participants are more variable, signified by the larger boxplots indicating a large spread in the ratings for many of the techniques. In general, there are few observable trends when considering the results for different types of source stimuli.

Repeated measures ANalysis Of VAriance (ANOVA) was utilised for statistical analysis of the results. A Kolmogorov-Smirnov normality test was applied to the residuals of the dataset. The test rejected the null hypothesis for 22 out of the 72 tests at a significance level of $p = 0.05$. Whilst the majority of conditions passed the normality test, 22 did not which indicates some degree of non-normality. However, approaches such as ANOVA are fairly insensitive to such violations [27].

Within-subject effects were tested for every combination of the three independent variables, technique (T), stimuli (S) and source position (P). All but two interactions (stimuli, position) violated sphericity therefore for all combinations the Huynh-Feldt corrected values are reported. The results of the three-way repeated measures ANOVA to a $5\%$ significance level are detailed in Table II. The strong main effect of technique, combined with the significant main interactions for position and technique $\times$ position support the trend that there were significant differences between the renderers and that these vary with source position. Whilst the main effect of stimuli and the interaction technique $\times$ stimuli $\times$ position were significant, both first order interactions technique $\times$ stimuli and stimuli $\times$ position were not. This suggests some stimuli were more critically revealing than others, however the trends between different techniques and positions did not vary across stimuli.

Mean ratings of the scores for each technique are given in Fig. 7. The overall means demonstrate that increasing the order of the HOS approach leads to a higher overall rating corroborating with the boxplots and ANOVA result for the technique interaction. Considering the means when varying the stimuli, it is apparent the scores for each technique are similar across all types of source stimuli, agreeing with the insignificant interaction technique $\times$ stimuli in the ANOVA. This is not the case when considering the means across technique and position, which combined with the ANOVA finding said interaction significant suggests that the different renders performed variably depending on the virtual source position. Notably, comparing the $\theta_1 = 25°$ position for HOS O1-45 and O3 where loudspeakers were added outside of the $\pm 45°$ span to increase the order, an increase in the score is observed showing that increasing the order does increase the accuracy of the source localisation. This is also clearly observable in Fig. 6. Furthermore, HOS O1-45 scored significantly higher than HOS O1-90 except for the rear position. Overall, the HOS renderers perform poorest for the rear position $\theta_3 = 180°$, corroborating with the boxplots which additionally suggest a larger variability in the ratings at this source position. Regardless, both increasing the order aids the performance and a significant rating is still observed, showing that HOS can reproduce a rear virtual source with just frontal loudspeakers.

Finally, post-hoc pairwise comparisons using a Bonferroni test were also performed considering each independent variable in turn pooled over all over variables. Considering the technique variable the majority of the tests showed a significant difference for each given combination of the renderers. Although notably comparing each of HOS O3 ($74.889 \pm 2.468, p = .742$) and HOS O4 ($80.019 \pm 2.185, p = 1.0$) to HOS O5 ($77.527 \pm 2.145$) resulted in no significant difference. Comparing HOA to HOS with a matched order showed no significant difference for both order 2 ($69.029 \pm 2.278$ and $62.198 \pm 3.065$, $p = .854$) and order 4 ($83.473 \pm 1.921$ and $80.019 \pm 2.185$, $p = .842$). Likewise with a comparison of HOS O1-45 to HOA O2 ($71.072 \pm 2.490$ and $69.029 \pm 2.278$, $p = 1.0$). The post-hoc tests for source stimuli were all insignificant except for the drums and guitar comparison ($70.717 \pm 2.081$ and $73.656 \pm 2.090, p = .041$), whilst for the source position all comparisons were significant verifying the trends seen in Fig. 6 and 7 that different positions lead to significantly different ratings from the participants.

### *F. Discussion*

The listening test was designed to compare different versions of HOS as well as a direct comparison to HOA. These covered whether increasing the truncation order of HOS increased the accuracy of the virtual source position, including in-span versus out-of-span and rear positions, as well as comparing HOS to HOA when the truncation order or the number of loudspeakers is fixed. The test focused on rating positional attributes of the virtual sound source excluding colouration, with the limitation of a 4 kHz low-pass filter to reduce artefacts from spatial aliasing.

A clear trend is observed between increasing the order of the truncation and the perceived localisation of the virtual source. This verifies that HOS is indeed the higher order extension of classical stereo, in a similar manner as HOA is

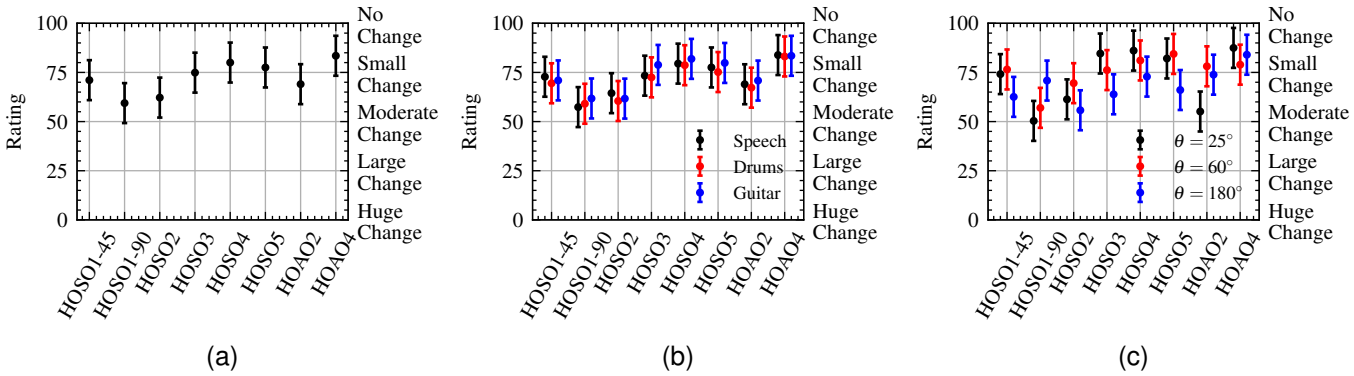



Fig. 7: Mean ratings with 95% within-subject confidence intervals marked [32] for the main effect of technique. (a) Overall means. (b) Means for each source stimuli. (c) Means for each source position.

the extension of First Order Ambisonics. This trend remains across different source material and source positions, and is particularly proven by considering the $\theta_1 = 25°$ position for HOS O1-45 and O3. Here, when increasing the order between these two systems, it was ensured that no new loudspeakers were added closer to the virtual source position, with the higher order system still scoring higher. A drop in performance is observed when moving from O4 to O5 which suggests a perceptual limit beyond which, for this scenario, a higher order is not advantageous. Furthermore, a first order system with a small span, HOS O1-45, performed considerably better than expected, in some cases outperforming higher order systems. This suggests that using HOS over readily available classic stereo arrangements could be advantageous, noting this setup was not statistically different in performance to HOA O2.

Moreover the data demonstrates that when using the dynamic version of HOS with loudspeakers situated only in front of the listener, convincing virtual sources can also be placed in the rear, albeit with some reduction in accuracy compared to frontal sources. This is a significant result as loudspeaker arrays that fully enclose the listener as required for optimal performance of HOA can be impractical. Nevertheless HOS comes at the cost of requiring dynamic head-tracked rendering.

Comparing HOS and HOA implementations at matching orders (O2 and O4 for both techniques) showed that there was no statistical difference between the two rendering approaches. Notably, this means Dynamic HOS presents an alternative to HOA requiring a significantly smaller number of loudspeakers arranged only in front of the listener. In the scenario where a subject may have a fixed number of loudspeakers available, for example only 5 loudspeakers, a HOS O4 implementation would perform better than the HOA O2 alternative, as well as requiring frontal loudspeakers only (but also head-tracking).

## VII. Conclusions

This article has introduced the dynamic head-tracked extension of HOS, allowing a listener to rotate their head freely resulting in adaptive loudspeaker panning. This maintains consistent virtual source imaging with a minimum of only $(N+1)$ loudspeakers for $N$-th order reproduction. The loudspeaker gains remain as panning functions (i.e., frequency independent) as with the original HOS implementation. HOS was demonstrated to be a higher order extension of classic stereophonic techniques, with the stereo sine, tangent and generalised head-tracked sine law all first order HOS systems.

An issue called the instability condition arises from dynamic HOS, which occurs when two loudspeakers fall on the same cone of confusion. Approaches to overcome the instability condition have been introduced, including limiting the listener's head rotation range, adding more loudspeakers and applying Tikhonov regularisation. HOS has also been demonstrated to be intrinsically linked to both 2D and 3D HOA through presenting decoders between the techniques, which were expanded to accommodate dynamic head rotations.

Finally, a listening test was performed to compare a number of HOS and 2D HOA systems with respect to localisation of a virtual source in comparison to a real sound source. The results demonstrate that increasing the order of HOS generally results in higher accuracy of the perceived virtual source location, albeit with some exceptions. HOS and HOA systems of matching order were shown to perform very similarly, which is advantageous as HOS requires both less loudspeakers and positioning of the loudspeakers in front of the listener only, although requires the use of listener tracking. Whilst HOS could reproduce rear virtual source positions with just loudspeakers in the front, generally HOA did perform better in these regions. Furthermore, simple first order HOS using a classic stereo loudspeaker pair performed beyond expectations suggesting that Dynamic HOS could bring significant advantages using this readily available standard setup.

For future work, a new approach for binaural rendering using HOS will be presented, derived from considering the inclusion of complex HRTFs. The ability of HOS to reproduce elevated sources using horizontal-only loudspeakers will also be considered, building on previous work where gain definitions for this scenario were presented [13]. A key limitation in our study was that the effect of colouration was not investigated. Furthermore, all renderers were subject to a 4 kHz low-pass filter to mitigate spatial aliasing. Future work could consider the effect of colouration in a more broadband study, as well as comparing HOS to other relevant alternatives such as VBAP or more state-of-the-art improvements to HOA.

## ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) through the University of Southampton's Doctoral Training Partnership Grant 2106106. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

## REFERENCES

[1] J.-M. Jot, V. Larcher, and J.-M. Pernaux, "A comparative study of 3-d audio encoding and rendering techniques," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, 03 1999.

[2] B. B. Bauer, "Phasor analysis of some stereophonic phenomena," *The Journal of the Acoustical Society of America*, vol. 33, no. 11, pp. 1536–1539, 1961.

[3] D. M. Leakey, "Some measurements on the effects of interchannel intensity and time differences in two channel sound systems," *The Journal of the Acoustical Society of America*, vol. 31, no. 7, pp. 977–986, 1959.

[4] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[5] F. Zotter and M. Frank, *A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer International Publishing, 2019, vol. 19.

[6] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc*, vol. 53, no. 11, pp. 1004–1025, 2005.

[7] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 697–707, 2001.

[8] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3, pp. 171–218, 1992.

[9] P. Damaske, "Head-related two-channel stereophony with loudspeaker reproduction," *The Journal of the Acoustical Society of America*, vol. 50, no. 4B, pp. 1109–1115, 1971.

[10] M. F. Simón Gálvez, D. Menzies, and F. M. Fazi, "Dynamic audio reproduction with linear loudspeaker arrays," *J. Audio Eng. Soc*, vol. 67, no. 4, pp. 190–200, 2019.

[11] J. Hollebon and F. M. Fazi, "Generalised low frequency 3d audio reproduction over loudspeakers," in *AES 148th Convention*, 2020.

[12] ——, "Experimental study of various methods for low frequency spatial audio reproduction over loudspeakers," in *I3DA: International Conference on Immersive and 3D Audio*, 2021.

[13] ——, "Higher-order stereophony," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2872–2885, 2023.

[14] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization." *Journal of Experimental Psychology*, vol. 27, no. 4, p. 339, 1940.

[15] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2841–2853, 1999.

[16] D. Menzies, M. F. S. Gálvez, and F. M. Fazi, "A low-frequency panning method with compensation for head rotation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 304–317, 2018.

[17] D. Menzies and F. M. Fazi, "A complex panning method for near-field imaging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1539–1548, 2018.

[18] ——, "Surround sound without rear loudspeakers: Multichannel compensated amplitude panning and ambisonics," in *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx-18), Portugal*, 2018.

[19] S. Merchel and S. Groth, "Adaptively adjusting the stereophonic sweet spot to the listener's position," *J. Audio Eng. Soc*, 2010.

[20] G. B. Arfken and H. J. Weber, *Mathetical Methods For Physicists*, sixth edition ed. Boston: Academic Press, 2005.

[21] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 189–194, 03 1998.

[22] M. Poletti, "Robust two-dimensional surround sound reproduction for nonuniform loudspeaker layouts," *J. Audio Eng. Soc*, vol. 55, no. 7/8, pp. 598–610, 2007.

[23] E. C. Hamdan and F. Maria Fazi, "Low frequency crosstalk cancellation and its relationship to amplitude panning," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 566–570.

[24] H. A. M. Clark, G. F. Dutton, and P. B. Vanderlyn, "The 'stereosonic' recording and reproducing system. a two-channel system for domestic tape records," *Proceedings of the IEE - Part B: Radio and Electronic Engineering*, vol. 104, no. 17, pp. 417–432, 09 1957.

[25] B. Bernfeld, "Attempts for better understanding of the directional stereophonic listening mechanism," in *Audio Engineering Society Convention 44*, 03 1973.

[26] A. Franck, G. Costantini, C. Pike, and F. M. Fazi, "An open realtime binaural synthesis toolkit for audio research," in *Audio Engineering Society Convention 144*, 2018.

[27] I.-R. BS.1534-2, "Method for the subjective assessment of intermediate quality level of audio systems," ITU-R, Standard, 2014.

[28] N. Meyer-Kahlen, M. Kastemaa, S. J. Schlecht, and T. Lokki, "Measuring motion-to-sound latency in virtual acoustic rendering systems," *J. Audio Eng. Soc*, vol. 71, no. 6, pp. 390–398, 2023.

[29] M. A. Gerzon, "General metatheory of auditory localisation," in *Audio Engineering Society Convention 92*, 1992.

[30] M. Woirgardt, P. Stade, J. Amankwor, B. Bernschütz, and J. Arend, "Cologne university of applied sciences," Anechoic Recordings, Tech. Rep., 2012.

[31] J. Ahrens and C. Andersson, "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2783–2794, 2019.

[32] G. Loftus and E. Masson, "Using confidence intervals in within-subject designs," *Psychonomic Bulletin And Review*, vol. 1, no. 4, pp. 476–490, 1994.

**Jacob Hollebon** Dr Jacob Hollebon is a Postdoctoral Researcher at the Institute of Sound and Vibration Research at the University of Southampton. Jacob graduated from the University of Warwick in 2017 with a BSc in Physics. He then completed the MSc in Acoustical Engineering at the University of Southampton in 2018, with a thesis on spatial audio reproduction for multiple listeners. In 2018, he joined the Virtual Acoustics and Audio Engineering team to begin a PhD in 3D audio reproduction over loudspeaker arrays, working from existing methods such as Soundfield Reproduction, Ambisonics and Crosstalk Cancellation to develop a new spatial audio technique, Higher-Order Stereophony. Jacob is the holder of the 2018 ISVR Elsevier prize and supported by two grants from the AES Educational Foundation, where he is also the 2018 Emil Torick Scholar.

**Filippo Mari Fazi** Dr Filippo Maria Fazi is Professor of Acoustics and Signal Processing at the Institute of Sound and Vibration Research of the University of Southampton, where he also serves as head of the Acoustics Group and leads the Virtual Acoustics and Audio Engineering Team. His research interests include acoustics, audio technologies, electroacoustics and digital signal processing, with special focus on acoustical inverse problems, multi-channel systems (including Ambisonics and Wave Field Synthesis), virtual acoustics, and microphone arrays. He is the author of more than 150 scientific publications and several patents. Dr Fazi was awarded a research fellowship by the Royal Academy of Engineering (2010) and the Tyndall Medal by the Institute of Acoustics (2018). He is a fellow of the Audio Engineering Society, a member of the Institute of Acoustics and is co-founder and chief scientist at Audioscenic, a start-up company that develops and commercialises 3D audio and loudspeaker array technologies.