

# A Comparison of Spatial Audio Loudspeaker Reproduction Techniques

Alexander Chiriboga  
Department of Music Research, Schulich School of Music, McGill University  
Montreal, Canada H3A 1E3  
December 5, 2023

## Abstract

Sound field synthesis (SFS) and transaural audio are spatial audio loudspeaker reproduction techniques that promise to overcome the limitations of stereophonic reproduction. While there exist many comparisons of SFS and other physically based reproduction methods, there exists no direct comparison of SFS and transaural audio. A brief history of spatial audio techniques and the theory of stereophony is presented. This paper constitutes a literature review of recent advancements in SFS and transaural and compares each approach in various situations.

## 1. Introduction

Humans have always been captivated by sound, whether as a means of communication or artistic expression. However, it was not until the invention of the phonograph that humans could recreate aural experiences. As technology has progressed, so has our ability to recreate more lifelike aural experiences. Perhaps the most notable innovation was stereophonic sound, our first attempt at creating spatial audio. Pioneered by Alan Blumley in 1931, Stereophony developed a new way to capture and reproduce three-dimensional sound using two speakers.<sup>1</sup> This breakthrough, however, could not recreate a sound field that enveloped the listener. To solve this problem, techniques such as the Acoustic Curtain by Steinberg and Snow in 1934,<sup>2</sup> a forerunner to today's wave field synthesis (WFS), and Gerzon's Ambisonics in the 1970s sought to recreate the sound field physically with more accuracy.<sup>3</sup> This was achieved by decomposing the sound field onto a new basis, such as planar wavefronts in the acoustic curtain/WFS case and spherical wavefronts in the case of Ambisonics.

Meanwhile, Dolby began developing surround sound by expanding the panning laws found in stereophony to include more speakers. Simultaneously, research into perceptual methods such as binaural recording and reproduction by companies like JVC, Sennheiser, and Sony in the 1970s introduced a new philosophy of spatial audio.<sup>4</sup> These advancements represented the three dominant approaches to recreating aural experiences: physically reproducing the sound field as accurately as possible, improving panning laws found in stereophony, or recreating the perception of sound as closely as possible. The 1980/1990s saw

the emergence of new techniques such as Wave-field Synthesis (WFS) by Berkhout in 1993<sup>5</sup> and transaural audio by Cooper and Bauck in 1989,<sup>6</sup> which constitute some of the best approaches of the physical and perceptual philosophies. As discussed above, WFS recreated the sound field as a series of plane waves, while transaural audio introduced a way to reproduce binaural audio over loudspeakers using a process known as crosstalk cancellation (CTC). In 1997, Pulkki introduced a new 3d panning method termed Vector-Based Amplitude Panning to allow stereophonic methods to reproduce a more 3d auditory image.<sup>7</sup> As new spatial audio technologies emerge, the ability to compare techniques has become more critical than ever.

Two such examples are Sound Field Synthesis (SFS) by Jens Arhens in 2012,<sup>8</sup> which proposed a new framework that directly compares both WFS and Ambisonics and Planewave-Based Angle Panning developed by Julius Smith in 2019,<sup>9</sup> which compares VBAP and WFS. Despite the growing body of literature comparing Sound Field Synthesis (SFS) with different spatial audio methods such as VBAP and Stereophony, there is no direct comparison of SFS with transaural audio. This paper seeks to bridge that gap by offering a literature review of both techniques and comparing them regarding ease of implementation and perceptual accuracy. The following sections will explore the shortcomings of stereophony, followed by a literature review of recent advancements in transaural audio and SFS. A comparison of the various transaural audio and SFS approaches will then be discussed to determine the best strategy for multiple situations.

## 2. Stereophonic Reproduction

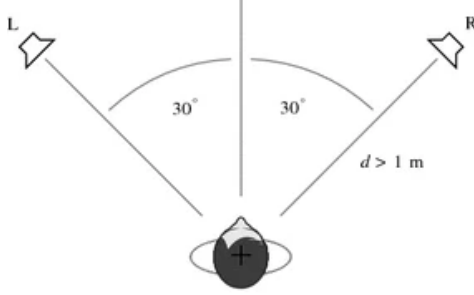


Figure 1: Standard stereo setup; the channels are termed left ('L') and right ('R'); the loudspeakers are at equal distance  $d$  from the listener.<sup>8</sup>

The most common method of spatial sound reproduction is stereophony. Stereophony can be defined as changing the amplitude and delay between the signals of two loudspeakers to create the perception of sound localization.<sup>10</sup> In stereophony, the loudspeakers are arranged in an equiangular triangle, with the listener positioned between the two loudspeakers in the space referred to as the "sweet spot." (see Figure 1) An auditory image can then be created through a psychoacoustic phenomenon known as summing localization,<sup>11</sup> which was later expanded into the association theory.<sup>12</sup> Summing localization and association theory describes how the auditory perception of multiple loudspeakers is translated into a single auditory event. In the figure above, when identical signals are sent to the two loudspeakers, a single sound source, known as the phantom source, is perceived to be located between the two loudspeakers. The relative position of this source can be changed using amplitude panning (summed localization) and delay panning (the precedence effect), where the precedence effect describes how two time-delayed signals can be perceived as a single auditory event.<sup>13</sup> By increasing the amplitude of the signal sent to a given loudspeaker, the perceived location of a phantom source can be shifted towards the respective loudspeaker (amplitude panning). Inversely, delaying a loudspeaker signal shifts the perceived location of the phantom source away from the respective loudspeaker (delay panning).

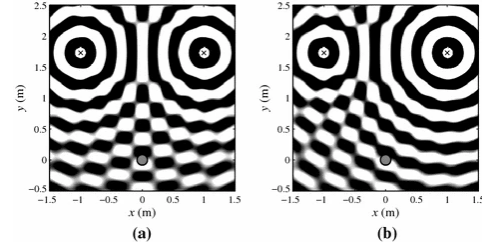


Figure 2: Sound field generated by two omnidirectional loudspeakers (represented by x) driven with a 1000hz signal. A grey disk represents the listener (located in the sweet spot). a) Both loudspeakers are driven with the same amplitudes. b) Right loudspeaker is driven with 6 dB higher amplitude<sup>8</sup>

While stereophony can recreate an accurate auditory experience, its generated sound fields are often inaccurate. Because stereophony relies on amplitude and delay panning to create an auditory event, the loudspeakers' placement heavily influences the perception of phantom sources. Because of this position-dependent panning, it's typically not possible to create the same perception in a location outside the sweet spot. For example, in the figure above, we see the generated sound field of a 1000hz signal using stereophony. If adequately synthesized, we should observe a 1000 hz plane wave propagating toward the listener. While the plane wave is correctly synthesized in the "sweet spot," listeners sufficiently far outside the sweet spot will perceive a positionally altered phantom source. Arhens found that "When the listening position is chosen such that the relative timing between the loudspeakers is altered by significantly more than 1 ms, then the precedence effect can take over and the spatial composition of the presented scene collapses completely into the closest loudspeaker."<sup>8</sup> This can be seen in the figure above; when sufficiently far outside the sweet spot, the wavefront can be observed as coming from the loudspeaker closest to the listener. As a result, the perception of the sound field is significantly altered if the listener were to rotate their head outside the sweet spot. Another consequence of the amplitude and delay panning is that phantom sources can only exist on a line between the two loudspeakers. While using advanced panning techniques such as VBAP, it is possible to constrain sources to a plane instead of a line.<sup>7</sup> It is impossible to create the perception of a source placed closer than the loudspeakers using panning techniques. For example, when trying to create the perception of raindrops falling around the listener's head, the listener would instead perceive the raindrops as falling near the loudspeakers.

ers. Due to the limitations of panning laws, expanding the listener's sweet spot is impossible as it relies heavily on the listener's position. Recent advancements in transaural audio, a perception-based reproduction technique, offer a solution to the sweet spot problem by creating multiple listening positions.

### 3. Transaural Audio

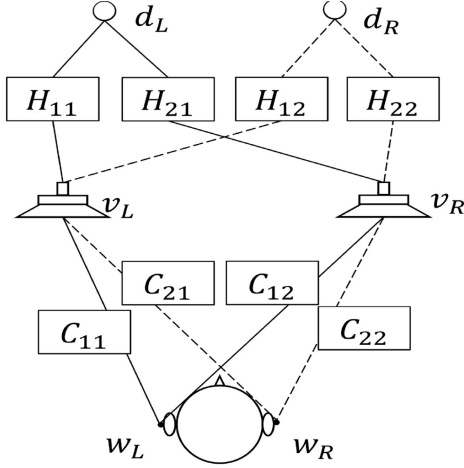


Figure 3: Block diagram for stereo crosstalk cancellation<sup>14</sup>

Schroeder and Atal proposed the first stereo crosstalk cancellation in 1963 as a perceptually motivated spatial audio method to deliver binaural audio at the listener's ear using speakers instead of headphones.<sup>15</sup> To realize this presentation method, the crosstalk, or the off-diagonal terms of the transmission path matrix,  $\mathbf{C}$  need to be canceled out. These can be seen in the Figure 3, where the crosstalk from the left speaker to the right ear is represented by  $(C_{21})$ , and the crosstalk from the right speaker to the left ear is represented by  $(C_{12})$ . By using an inverse filter matrix  $\mathbf{H}$  to cancel the speaker crosstalk, the intended audio  $\mathbf{d}$  could be replicated at the listener's ear  $\mathbf{w}$  using a loudspeaker driving function  $\mathbf{v}$  where

$$\mathbf{d} = \begin{bmatrix} d_L \\ d_R \end{bmatrix}, \mathbf{v} = \begin{bmatrix} v_L \\ v_R \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_L \\ w_R \end{bmatrix},$$

$$\mathbf{H} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}, \text{ and } \mathbf{C} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}.$$

Given that  $\mathbf{w} = \mathbf{C}\mathbf{v}$ , and  $\mathbf{v} = \mathbf{H}\mathbf{d}$ , the inverse filter matrix  $\mathbf{H}$  should be the inverse of transmission path the matrix  $\mathbf{A}$  to reproduce the intended audio signal at the ears. Cooper and Bauck later improved this concept in 1975 when they coined the term transaural audio to refer to the spatial

audio method of synthesizing the sound pressure at the ears of the listener to deliver binaural audio.<sup>6</sup>

### 3.1 Optimal Source Distribution

Unfortunately, the stereo crosstalk method proposed by Schroeder suffered from a variety of issues, such as loss of dynamic range, only being suitable in anechoic conditions, and the inability to account for individual differences in the head-related transfer functions. To solve these issues, Takeuchi and Nelson proposed an optimal source distribution (OSD) in 2002.<sup>16</sup>

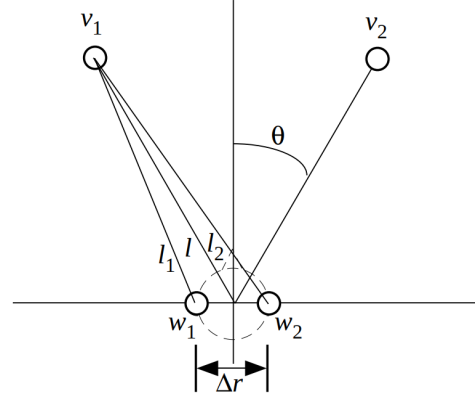


Figure 4: Crosstalk cancellation geometry<sup>16</sup>

Takeuchi and Nelson proposed to view the desired signals as delayed versions of the reproduced signals. Assuming that the two sources were free-field acoustic monopoles with volume velocities  $v_L$   $v_R$ , the transmission path matrix can be defined as

$$\mathbf{C} = \frac{\rho_0}{4\pi} \begin{bmatrix} \frac{e^{-jkl_1}}{l_1} & \frac{e^{-jkl_2}}{l_2} \\ \frac{e^{-jkl_2}}{l_2} & \frac{e^{-jkl_1}}{l_1} \end{bmatrix}$$

where  $k = \frac{\omega}{c_0}$  with  $c_0$  and  $p_0$  being the speed of sound and density of air, respectively. This matrix can then be rewritten as

$$\mathbf{C} = \frac{\rho_0 e^{-jkl_1}}{4\pi l_1} \begin{bmatrix} 1 & g e^{-jk\Delta l} \\ g e^{-jk\Delta l} & 1 \end{bmatrix} \text{ or } \mathbf{C} = \frac{\rho_0 e^{-jkl_1}}{4\pi l_1} \mathbf{C}_N$$

where  $\Delta l = l_2 - l_1$  and  $g = \frac{l_1}{l_2}$ .  $\mathbf{C}_N$  can then be further simplified using a far-field approximation  $\Delta l = \Delta r \sin \theta$ , where  $\Delta r$  is the width of the ears of the listener and

$$\mathbf{C}_N = \begin{bmatrix} 1 & g e^{-jk\Delta r \sin \theta} \\ g e^{-jk\Delta r \sin \theta} & 1 \end{bmatrix}.$$

The reproduced signal is therefore

$$\mathbf{w} = \frac{\rho_0 e^{-jk l_1}}{4\pi l_1} \mathbf{C}_N \mathbf{H} \mathbf{d}.$$

We can then define the reproduced target values as

$$\hat{\mathbf{w}} = \frac{\rho_0 e^{-jk l_1}}{4\pi l_1} \mathbf{d},$$

where  $\hat{\mathbf{w}}$  is the desired signal  $\mathbf{d}$  delayed by  $l_1/c_0$ . Similarly to Schroeder's formulation, we want  $\hat{\mathbf{w}} = \mathbf{w}$ , so the inverse filter matrix must be equal to the inverse of the transmission path matrix ( $\mathbf{H} = \mathbf{C}_N^{-1}$ ), where

$$\mathbf{H} = \frac{1}{1 - g^2 e^{-2jkr \sin \theta}} \begin{bmatrix} 1 & -ge^{-jk\Delta r \sin \theta} \\ -ge^{-jk\Delta r \sin \theta} & 1 \end{bmatrix}.$$

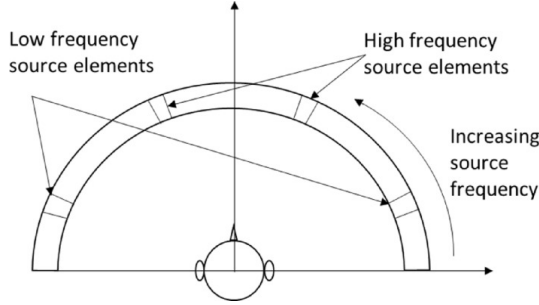


Figure 5: Continuous frequency dependent loudspeaker array<sup>16</sup>

This formulation can be proven to be well-conditioned, meaning that the crosstalk cancellation can be reproduced with minimal error.<sup>17</sup> Analyzing  $\mathbf{H}$ , it can be seen that the  $\|\mathbf{H}\|$  changes periodically and is well-conditioned when  $k\Delta r \sin \theta = n\pi/2$  where  $n$  is an odd integer. These positions represent where  $\|\mathbf{H}\|$  is minimized or when the out-of-phase and in-phase components can be replicated with the least amount of effort. We can maintain this well-conditioned case by setting the path lengths equal to one-quarter the wavelength. Thus, we get a continuous distribution of sources whose frequency varies with azimuth to preserve the path length condition. Under this formulation,  $\mathbf{H}$  can be defined similar to a rotation matrix with small amplitude changes where

$$\mathbf{H} = \frac{1}{1 + g^2} \begin{bmatrix} 1 & -jg \\ -jg & 1 \end{bmatrix}.$$

### 3.2 OSD Radiation

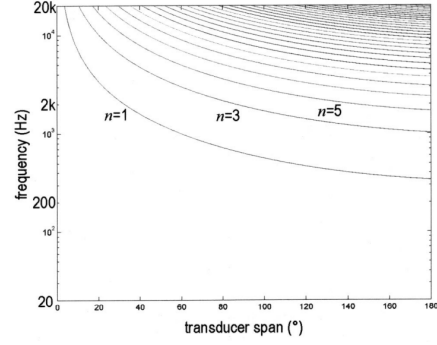


Figure 6: The source span for different frequencies and odd integer numbers  $n$ .<sup>16</sup>

In Figure 6, one can see the lowest frequency that can be correctly spatialized is a function of the range of the source span where the lowest value of  $n$  gives the lowest frequency limit. The lower limit for the frequency is given by

$$f = \frac{nc_0}{4\Delta r}$$

when the source span is  $180^\circ$ . For a hemispherical array (for  $180^\circ$  and  $n = 1$ ), the lowest frequency that can be spatialized is about 300-400hz. For frequencies below 300hz, Takeuchi and Nelson propose that a single subwoofer can reasonably well supplement the loss of dynamic range for out-of-phase components below 300hz.<sup>16</sup>

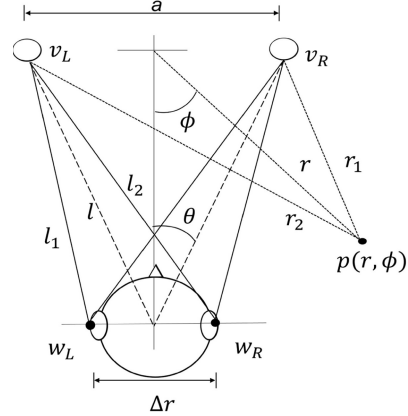


Figure 7: Continuous frequency dependent loudspeaker array<sup>14</sup>

One interesting property of the OSD is the ability to produce crosstalk cancellation at multiple locations. Following the derivation by Nelson and Takeuchi,<sup>14</sup> a desired signal of  $\mathbf{d} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

is used to view the far-field radiation properties of the OSD. The source signals  $\mathbf{v}$  can then be found to be

$$\mathbf{v} = \mathbf{H}\mathbf{d} = \frac{1}{1+g^2} \begin{bmatrix} 1 \\ -jg \end{bmatrix}.$$

The pressure field given by the two sources can be found to be

$$p(r, \phi) = \frac{\rho_0}{4\pi(1+g^2)} \left[ \frac{e^{-jkr_1}}{r_1} - jg \frac{e^{-jkr_2}}{r_2} \right].$$

Setting  $h = r_1/r_2$  we can further simplify this expression to

$$p(r, \phi) = \frac{\rho_0 e^{-jkr_1}}{4\pi r_1(1+g^2)} \begin{bmatrix} 1 & -jgh e^{-jk(r_2-r_1)} \end{bmatrix}.$$

Taking a far-field approximation, we can express  $r_1$  and  $r_2$  as

$$r_1 = r - \frac{a}{2} \sin \phi, \quad r_2 = r + \frac{a}{2} \sin \phi.$$

Taking the modulus squared of the pressure field, we can find the pressure field to be

$$|p(r, \phi)|^2 = \left( \frac{\rho_0}{4\pi r_1} \right)^2 \frac{1 + (gh)^2 - 2gh \sin(ka \sin \phi)}{1 + g^2},$$

and by assuming that  $r \approx r_1 \approx r_2$  we can further simplify the expression to get

$$|p(r, \phi)|^2 = \left( \frac{\rho_0}{4\pi r_1} \right)^2 (1 - \sin(ka \sin \phi)).$$

The pressure field produces a maxima and minima when  $ka \sin \phi = n\pi/2$  with minima at  $n_{min} = 4n + 1$  and maxima  $n_{max} = 4n + 3$  for ( $n \in \mathbb{Z}$ ). This can be seen in Figure 8, where the sound field intensity is plotted as a function of azimuthal angle. This condition applies for all frequencies as long as the quarter wavelength condition is upheld. Figure 8 shows that multiple crosstalk positions are an intrinsic property of the OSD.

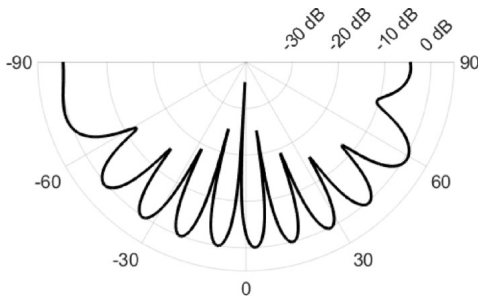


Figure 8: Far field sound pressure level as a function of the angle produced by a two-channel Optimal Source Distribution driven with a desired signal of  $\mathbf{d} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ .<sup>14</sup>

### 3.3 OSD Discretization

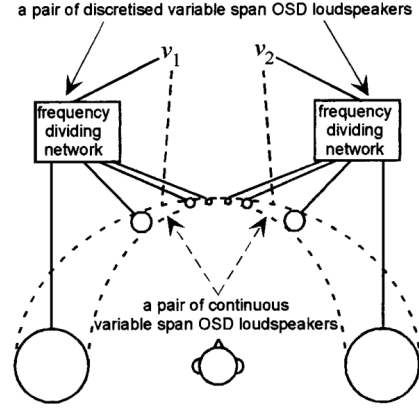


Figure 9: Discrete optimal source distribution implementation.<sup>16</sup>

Unfortunately, given today's technology, a continuous distribution is not possible; as such, the continuous source distribution must be approximated by a discrete array. Fortunately, a discrete array reasonably well approximates an OSD with minimal loss of dynamic range and slightly reduced robustness to changes in listening position. As seen in Figure 9, a discrete system can be realized by a frequency division matrix where the two desired signals are divided into multiple frequency bands and fed into the corresponding pairs of loudspeakers based on frequency. While a discretized array means that the lower frequency limit for the source span is discretized, for large source spans (such as  $180^\circ$ ), the lower limit is approximately the same. Takeuchi and Nelson have shown that a few pairs of transducers with differing transducer spans could cover the entire audible frequency range.<sup>14</sup>

Additionally, Takeuchi and Nelson found that there is a tradeoff between loss of dynamic range and crosstalk for each frequency range. For greater crosstalk performance, a more significant loss of dynamic range is required for the frequency range. This is problematic for the sub-bass/bass frequency range ( $< 300\text{Hz}$ ), where there is already a significant loss of dynamic range due to the lower frequency limit of the source span. The authors noted that even if reasonable crosstalk is unavailable for the lowest-frequency transducer pair, the dynamic range is reasonably well preserved when the two signals are in phase. This works well for transaural reproduction since lower frequency components are usually panned to sit in the center, meaning the out-of-phase components between the two channels are usually negligible for lower frequencies.



### 3.4 MIMO-CTC

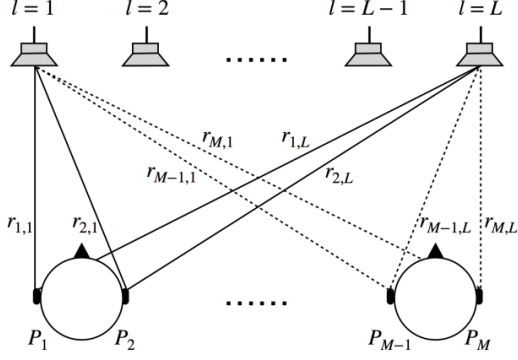


Figure 10: MIMO crosstalk cancellation geometry<sup>18</sup>

While OSD offers an attractive solution to the multi-listener problem, the lack of choice over the listener position is problematic for implementations that require arbitrary listening positions. This can be solved through the use of uniform line arrays, where we preclude the use of a well-conditioned inverse filter matrix for more control over the listener position. In multiple-input multiple-output (MIMO) crosstalk cancellation  $L$  loudspeakers are used to control the pressure at  $M$  control points (listener's ears). The notation of MIMO is slightly different than OSD where the reproduced pressure at the listeners ear is  $\mathbf{p}$ , the desired signals are  $\mathbf{b}$ , and the loudspeaker signals are  $\mathbf{v}$ . These vectors are of size  $M$  where

$$\begin{aligned}\mathbf{p} &= [p_1(\omega), p_2(\omega), \dots, p_M(\omega)]^T, \\ \mathbf{b} &= [b_1(\omega), b_2(\omega), \dots, b_M(\omega)]^T, \text{ and} \\ \mathbf{v} &= [v_1(\omega), v_2(\omega), \dots, v_L(\omega)]^T.\end{aligned}$$

where  $(\omega = 2\pi f)$  is the angular frequency. As seen before the reproduced signals  $\mathbf{p}$  are equal to the product of the transmission path matrix and the loudspeaker driving signals ( $\mathbf{p} = \mathbf{C}\mathbf{v}$ ) where  $\mathbf{C}$  is the transmission matrix of  $M \times L$  acoustic transfer functions. Assuming the loudspeakers radiate as free-field acoustic monopoles, then  $\mathbf{C}$  is given by:

$$\mathbf{C} = \frac{1}{4\pi} \begin{bmatrix} e^{jk r_{1,1}} & \dots & e^{jk r_{1,L}} \\ \vdots & \ddots & \vdots \\ e^{jk r_{M,1}} & \dots & e^{jk r_{M,L}} \end{bmatrix}$$

where  $k = \frac{\omega}{c_0}$  is the wavenumber and  $r_{m,l}$  is the distance between loudspeaker  $L$  and listener's ear  $M$ . Similarly, the loudspeaker signals are defined as  $\mathbf{v} = \mathbf{H}\mathbf{b}$  where  $\mathbf{H}$  is the  $L \times M$  matrix of CTC

filters. As this filter is not well-conditioned, a cost function is used where

$$J = \|\mathbf{p} - \mathbf{b}\|_2^2 = \|\mathbf{C}\mathbf{v} - \mathbf{b}\|_2^2.$$

This cost function minimizes the error between the reproduced pressure at the listener's ears and the desired signals. In the case where  $L > M$ , the solution to this minimization is given by a Moore-Penrose (pseudo) inverse denoted by a  $+$  where

$$\mathbf{H} = \mathbf{C}^+.$$

The inverse filter matrix is ill-conditioned for specific frequencies, making it sensitive to small errors and giving solutions with large loudspeaker gains. Tikhonov regularization is used at these frequencies where the regularization inherently introduces errors to the solution, giving a modified cost function of

$$J = \|\mathbf{C}\mathbf{v} - \mathbf{b}\|_2^2 + \beta \|\mathbf{v}\|_2^2.$$

and an inverse filter matrix

$$\mathbf{H} = \mathbf{C}^H [\mathbf{C}\mathbf{C}^H + \beta \mathbf{I}_M]^{-1}$$

where  $\beta$  is the regularization parameter, and  $\mathbf{I}_M$  is the  $M \times M$  identity matrix.

To selectively control the contribution of each loudspeaker to the pressure at the control points, a diagonal weighting matrix  $\mathbf{\Gamma}$  is introduced, leading to a modified cost function of

$$J = \|\mathbf{C}\mathbf{v} - \mathbf{b}\|_2^2 + \beta \|\mathbf{\Gamma}\mathbf{v}\|_2^2.$$

with a solution of

$$\mathbf{H} = \mathbf{C}^H [\mathbf{C}\mathbf{C}^H + \mathbf{\Gamma}]^{-1}$$

This formulation provides a flexible way to apply regularization on a per loudspeaker and frequency basis. This allows for different types of loudspeaker distributions. In a study about loudspeaker distributions for MIMO CTC, Hollenbon found that more sparsely distributed loudspeakers had increased low-frequency crosstalk cancellation performance. In contrast, the more dense distribution of loudspeakers in front of each listener had better the mid-frequency and high-frequency crosstalk cancellation performance.<sup>18</sup> It was ultimately found that uniform linear loudspeaker arrays (ULA) were a good compromise between crosstalk cancellation performance and practical implementation.

## 4. Sound Field Synthesis

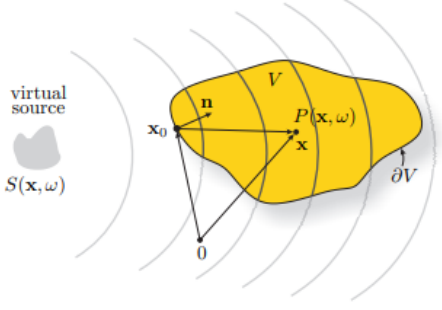


Figure 11: Sound Field Synthesis geometry<sup>8</sup>

In 2012, following the work of his doctoral thesis, Jens Arhens proposed a unified framework to compare the two most popular methods of loudspeaker sound field control: near-field compensated higher-order ambisonics (NFC-HOA) and wavefield synthesis (WFS).<sup>8</sup> NFC-HOA, proposed by Daniel in 2003,<sup>19</sup> built upon Gazon's Ambisonic formulation and solved the issue of infinite bass response for sources in the near field. Arhens proposed a sound field synthesis equation

$$S(\mathbf{x}, \omega) = \oint_{\partial\Omega} D(\mathbf{x}_0, \omega) G(\mathbf{x}, \mathbf{x}_0, \omega) dA(\mathbf{x}_0)$$

where  $D(x_0, \omega)$  represents the driving signal of the secondary source (loudspeaker) located at a point on the boundary enclosing a volume  $x_0 \in \partial\Omega$  and  $G(x - x_0, \omega)$  represents the transfer function (Greens function) between a point in space  $\mathbf{x}$  and the secondary source. This synthesis equation can be thought of similarly to the Kirchhoff-Helmholtz integral, which states that if you know the pressure on a continuous, simply connected boundary of a source-free volume the you know the sound field within the volume. In other words, if you control the sound field on the boundary of a source-free volume, you can dictate the sound field inside. This can be seen in the synthesis equation where all the contributions from the sound field on the boundary are summed together, and Green's function is used to determine how the pressure created by each secondary source propagates to a point in space  $\mathbf{x}$ . This synthesis equation can be solved in two ways: explicitly and implicitly. The explicit formulation for spherical secondary source distributions can be shown to be equivalent to NFC-HOA. Conversely, the implicit solution is equivalent to the WFS formulation.

### 4.1 Wave Field Synthesis

The implicit solution can be found by analyzing the SFS problem from a physical point of view using the relationship between the sound field on the boundary and the sound field inside. While there are many ways to derive the implicit solution, the most straightforward solution can be found using the Rayleigh I Integral. The Rayleigh I Integral describes the sound field  $P(\mathbf{x}, \omega)$  in a target half-space  $\Omega$  that is bounded by a infinite simply connected planar surface  $\partial\Omega$  and is given by

$$P(\mathbf{x}, \omega) = \oint_{\partial\Omega} -2 \frac{\partial}{\partial \mathbf{n}} S(\mathbf{x}, \omega) \Big|_{\mathbf{x}=\mathbf{x}_0} G(\mathbf{x}, \mathbf{x}_0, \omega) dA(\mathbf{x}_0).$$

If we set  $P(\mathbf{x}, \omega) = S(\mathbf{x}, \omega)$  (ie the target half-space is source-free), we get a formulation that is nearly identical to the sound field synthesis equation proposed by Arhens where the driving function is

$$D(\mathbf{x}_0, \omega) = -2 \frac{\partial}{\partial \mathbf{n}} S(\mathbf{x}, \omega) \Big|_{\mathbf{x}=\mathbf{x}_0}.$$

Unfortunately, this formulation requires an infinite continuous planar array. Ideally, we would want a finite array to immerse the listener(s).

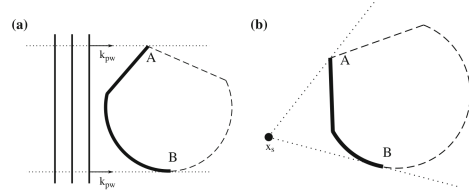


Figure 12: Secondary Source Illumination a) for a plane wave b) for a monopole source<sup>8</sup>

If we assume a far-field/high-frequency solution, we can apply an approximation from optics(Kirchhoff approximation).<sup>20</sup> With this approximation, a curved surface may be considered locally planar for sufficiently short wavelengths (high frequencies). We can then locally apply the Rayleigh-based solution where only the secondary sources that are virtually illuminated by the desired sound field are driven. (See Figure 15) The driving function is then

$$D(\mathbf{x}_0, \omega) = -2a(\mathbf{x}_0) \frac{\partial}{\partial \mathbf{n}} S(\mathbf{x}, \omega) \Big|_{\mathbf{x}=\mathbf{x}_0}$$

where  $a(\mathbf{x}_0)$  is the secondary source selection.

### 4.2 2.5D WFS

While the local planar approximation is quite good, implementing a planar surface of secondary

sources is difficult and computationally expensive. To solve this issue, a 2.5D solution is proposed. The reasoning is as follows: instead of using a surface to synthesize a 3D volume, use a line to synthesize a 2D listening plane that intersects the listener's ears. This solution is termed 2.5D because while it is only focused on synthesizing in the 2D plane, the formulation's and reproduction's physics are inherently based on 3D space, meaning the solution is incomplete.

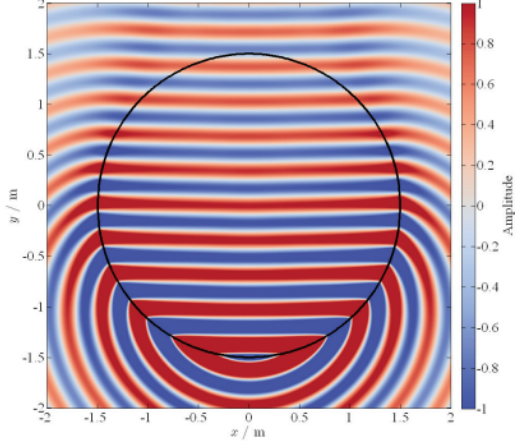


Figure 13: 2.5D synthesis of a 1000hz plane wave by a continuous distribution of sources<sup>21</sup>

A consequence of this incompleteness is that the synthesized sound field amplitude decays faster than desired, which is especially problematic for large systems. This can be seen in Figure 13, where the wavefronts decay as a function of distance from the secondary source distribution. Fortunately, the system still preserves the curvature of the wave fronts in the horizontal plane.

While this solution is only an approximation and suffers from amplitude decay, it has a handful of properties that make it attractive for real-time applications. As the solution is solved implicitly, it is a more efficient way to synthesize a sound field. This is especially true for signals such as music or speech, which can be implemented very similarly to a phased line array. For these applications, it only requires a single filter that determines the weight and delay of the signal, which can be used to determine the driving functions.<sup>22</sup> Lastly, it only requires secondary source distributions that bound the synthesized region. We can use a half-bound distribution, such as a hemi-circular array if we only want our sources to propagate in one direction.

### 4.3 Higher Order Ambisonics

Alternatively, the driving functions can be found explicitly by solving for them directly and evaluating the integral. Fredholm's Theorem tells us that an exact solution exists for an arbitrary sound field when a boundary encloses the volume. The general solution for the explicit formula can be thought of as a convolution onto the basis of the boundary where

$$D(\mathbf{x}, \omega) = \sum_{n=0}^{\infty} \check{D}_n(\omega) \psi_n(\mathbf{x})$$

and

$$\check{D}_n(\omega) = \frac{\check{S}_n(\omega)}{a_n \check{G}_n(\omega)}.$$

$\psi_n(\mathbf{x})$  is the orthogonal basis being expanded on,  $\check{D}_n(\omega)$  being formulated as a convolution relationship and  $a_n$  being a constant related to the orthogonal basis. While we can solve the integral for any arbitrary boundary, simple geometries are the easiest to solve and implement. In the case of a spherical geometry for NFC-HOA, the explicit solution is the only exact solution. The solution for a sphere is an expansion on the spherical harmonics, which can also be interpreted as a convolution on the surface of a sphere where

$$D(\alpha, \beta, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \underbrace{\frac{1}{2\pi R^2} \sqrt{\frac{2n+1}{4\pi}} \frac{\check{S}_n^m(\omega)}{\check{G}_n^m(\omega)}}_{=\check{D}_n^m(\omega)} Y_n^m(\beta, \alpha).$$

In this case  $Y_n^m(\beta, \alpha)$  is an orthogonal basis known as the spherical harmonics with  $\beta$  and  $\alpha$  being parameters that specify the basis.

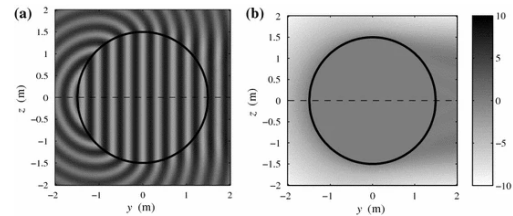


Figure 14: NFC-HOA synthesis of a 1000hz plane wave in the z-y plane by a continuous spherical distribution of sources<sup>21</sup>

As seen in Figure 14, NFC-HOA Does not suffer from the amplitude decay seen in WFS. This is because NFC-HOA forms an exact solution, while WFS is only an approximate solution. While 2.5D solutions are possible with the explicit formulation, they are more computationally expensive to implement and suffer from amplitude decay problems similar to those of the implicit case.



#### 4.4 SFS Discretization

So far, we have assumed that our secondary source distributions are continuous. However, as seen with OSD, the technology for continuous loudspeaker distributions does not exist, meaning practical implementations will employ a finite number of discrete loudspeakers. When using a finite number of evenly spaced loudspeakers, the synthesized sound field is well approximated to a particular frequency termed the spatial aliasing frequency. For loudspeaker spacings of 9 to 15cm, the spatial aliasing frequency is 2000hz to 1500hz, respectively. Blauert found this spatial aliasing frequency was a good compromise between accuracy and practicability.<sup>10</sup>

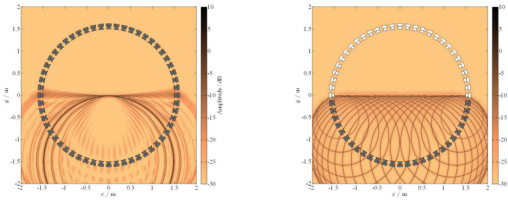


Figure 15: Synthesis of a 100hz plane wave using a distribution of 56 discrete loudspeakers a) NFC-HOA b) WFS<sup>21</sup>

The consequences of a discrete distribution are different for WFS and NFC-HOA. In the case of NFC-HOA, the synthesis of the sound field is only accurate in a centralized location, similar to the sweet spot seen in stereophony, where outside this location, additional artifacts, such as a curvature of the wavefront, are noticed. In the case of WFS, we notice additional wavefronts arising due to discretization. These artifacts are known as spatial aliasing and hurt the perceptual localization of sources.

#### 5. Comparison of SFS and Transaural Audio

The previous sections presented an introductory glance at the recent advancements in SFS and transaural audio. This section will present a summary of each method's strengths and weaknesses and the ideal situation in which each method should be employed.

Wavefield synthesis is perhaps the most researched spatial audio presentation method. It provides somewhat simple real-time implementation and offers a good auditory perception across an extended listening area without needing a loudspeaker distribution to completely surround the listening area. Unfortunately, it is sensitive to

room acoustics, experiences amplitude decay for linear distributions, requires a large number of speakers, and is only accurate below a spatial aliasing frequency range of 1500-2000 Hz. Despite its sensitivity to room acoustics, this is the best solution for presenting audio to a large distributed audience. The sphere in Las Vegas has implemented this method for spatial audio presentation with a 150,000-planar speaker array.

Nearfield compensated higher-order ambisonics (NFC-HOA) is the most accurate solution for spatial sound reproduction. However, it is challenging to implement in real-time, highly sensitive to room acoustics, requires a large number of loudspeakers (more than WFS) for good sound field presentation,<sup>23</sup> and requires a 3D spherical loudspeaker distribution. Like WFS, NFC-HOA is only accurate up to a spatial aliasing frequency of 1500-2000 Hz. As seen by the lack of commercial NFC-HOA systems, NFC-HOA only succeeds in situations where a large number of loudspeakers are available in an anechoic environment. It is for this reason NFC-HOA is mainly used by research institutions.<sup>8</sup> This, however, promises to be a good solution for testing devices in adverse acoustic environments.<sup>21</sup>

Optimal source distribution (OSD) crosstalk cancellation requires fewer speakers than Wavefield Synthesis and does not require a generalized Head-Related Transfer Function (HRTF). It is well-conditioned, which minimizes the computation of the inverse filter matrix, and is accurate for most of the auditory range (above 300 Hz). However, it is somewhat sensitive to room acoustics and cannot provide arbitrary listening positions. This solution is ideal for applications where the listening position coincides with the OSD radiation pattern. Some examples of this application are theme park rides or car audio, where the listening positions can be set to coincide with the radiation pattern.

Multiple-input multiple-output (MIMO) CTC allows for arbitrary listener positions and is mostly immune to reflections caused by room acoustics. The sparsity/density of the array can be tuned for different performance characteristics, either for improved low-frequency CTC or better mid/high-frequency CTC. However, it requires approximately five to seven speakers per listener and should be improved when head tracking is used.<sup>18</sup> This method succeeds in situations with a low number of listeners. For the case of a singular listener, this method seems the most promising, especially in mobile devices and computers where cameras or infrared sensors could be used for head tracking. Even without head tracking, this seems

a good option for home cinema where soundbars could present spatial audio, especially given its robustness to room acoustics.

## 6. Conclusion

A comparison of recent advancements in sound field synthesis (SFS) and transaural audio for spatial audio loudspeaker reproduction was discussed. A brief history of spatial audio techniques and the theory of stereophony was presented. The theory of wavefield synthesis (WFS), near-field compensated higher-order ambisonics (NFC-HOA), optimal source distribution crosstalk cancellation (OSD), and multiple-input multiple-output crosstalk cancellation (MIMO), was introduced and the ideal situations for each implementation were proposed. Further research should include a perceptual study of localization accuracy comparing SFS and transaural audio in both anechoic and non-anechoic situations. The use of head tracking for adaptive crosstalk cancellation in MIMO should be explored, and all four approaches need more work on minimizing the sensitivity to reflections caused by room acoustics. A further understanding of how spatial aliasing artifacts in SFS and dynamic range loss effect in OSD CTC, and frequency dependent CTC performance in MIMO should be explored to further research the larger question of how spatial audio loudspeaker systems influence the sound quality experienced by the listener.

## References

- [1] F. Rumsey. *Spatial Audio*. Focal Press, 2001.
- [2] J. C. Steinberg and W. B. Snow. Auditory perspective — physical factors. *Electrical Engineering*, 53(1):12–17, 1934.
- [3] Michael A. Gerzon. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10, 1973.
- [4] Stephan Paul. Binaural recording technology: A historical review and possible future developments. *Acta Acustica united with Acustica*, 95:767–788, 2009.
- [5] A. Berkhout, Diemer Vries, and P. Vogel. Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America*, 93:2764–2778, 1993.
- [6] Duane H. Cooper and Jerald L. Bauck. Prospects for transaural recording. *Journal of The Audio Engineering Society*, 37:3–19, 1989.
- [7] Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of The Audio Engineering Society*, 45:456–466, 1997.
- [8] Jens Ahrens. *Analytic Methods of Sound Field Synthesis*. Springer, 2012.
- [9] Julius Orion Smith. A spatial sampling approach to wave field synthesis: Pbap and huygens arrays. *arXiv*, abs/1911.07575, 2019. n. pag.
- [10] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, revised edition, 1997.
- [11] H. Warncke. Die grundlagen der raumbezüglichen stereophonischen Übertragung im tonfilm (the fundamentals of room-related stereophonic reproduction in sound films). *Akustische Zeitschrift*, 6:174–188, 1941.
- [12] G. Theile. *On the localisation in the superimposed soundfield*. PhD thesis, Technische Universität Berlin, 1980.
- [13] H. Haas. Über den einfluss eines einfachechos auf die horsamkeit von sprache (on the influence of a simple echo on the comprehensibility of speech). *Acustica*, 1:49–58, 1951.
- [14] P. A. Nelson, T. Takeuchi, P. Couturier, and X. Zhou. Sound field control for multiple listener virtual imaging. *Journal of Sound and Vibration*, 539:117259, 2022.
- [15] M. R. Schroeder and Bishnu Atal. Computer simulation of sound transmission in rooms. *Proceedings of the IEEE*, 2:536–537, 1963.
- [16] T. Takeuchi and P. A. Nelson. Optimal source distribution for binaural synthesis over loudspeakers. *Journal of the Acoustical Society of America*, 112(6):2786–2797, 2002.
- [17] M. Yairi, T. Takeuchi, K. R. Holland, D. G. Morgan, and L. Haines. Binaural reproduction capability for multiple off-axis listeners based on the 3-channel optimal source distribution principle. In *Proceedings of the 23rd International Congress on Acoustics*, Aachen, Germany, 2019.
- [18] Jacob Hollebon, Filippo Maria Fazi, and Marcos F. Simón Gálvez. A multiple listener crosstalk cancellation system using loudspeaker-dependent regularization. *Journal of the Audio Engineering Society*, 2021.

- [19] Jerome Daniel. *Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format*. PhD thesis, 2003.
- [20] D. L. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*. Springer, 1992.
- [21] Jens Ahrens, Rudolf Rabenstein, and Sascha Spors. Sound field synthesis for audio presentation. *Acoustics Today*, 10(2), 2014.
- [22] E. N. G. Verheijen. *Sound reproduction by wave field synthesis*. PhD thesis, Delft University of Technology, 1997.
- [23] Hagen Wierstorf, Alexander Raake, and Sascha Spors. Assessing localization accuracy in sound field synthesis. *The Journal of the Acoustical Society of America*, 141(2):1111–1119, 02 2017.