INVITED TUTORIAL

A tutorial on immersive three-dimensional sound technologies

Craig T. Jin*

*School of Electrical and Information Engineering, University of Sydney,
Building J03, Maze Crescent, Darlinghurst, Sydney, NSW, 2006 Australia*

Abstract: There is renewed interest in virtual auditory perception and spatial audio arising from a technological drive toward enhanced perception via mixed-reality systems. Because the various technologies for three-dimensional (3D) sound are so numerous, this tutorial focuses on underlying principles. We consider the rendering of virtual auditory space via both loudspeakers and binaural headphones. We also consider the recording of sound fields and the simulation of virtual auditory space. Special attention is given to areas with the potential for further research and development. We highlight some of the more recent technologies and provide references so that participants can explore issues in more detail.

Keywords: 3D sound technology, Virtual auditory space, Spatial hearing

PACS number: 43.38.Md, 43.38.Vk, 43.10.Ln, 43.10.Sv [doi:10.1250/ast.41.16]

1. INTRODUCTION

There is renewed interest in three-dimensional (3D) sound technologies driven in large part by the technological drive towards enhanced perception via mixed reality systems. A few commercial examples, as of 2018, are Microsoft 3D Soundscape and the HoloLens, Google's Resonance Audio, Sony Playstation headset with 3D sound, the Oculus headset and its support for 3D sound spatialization. There is also increasing support and awareness for 3D sound by broadcasting companies. A few broadcasting examples are the BiLi project in France, the Binaural Project by the BBC in the United Kingdom, the Orpheus European project, the NHK Super Hi-Vision theatre support for 3D sound with a 22.2 multichannel system. At the current time, the technologies for 3D sound are growing exponentially and so this tutorial will focus on fundamental principles. To begin, we should clarify our definition of 3D sound. By 3D sound, we refer to an immersive experience in which the listener has a clear and extended perception of a 3D sound space — that is sound objects and sound events clearly positioned relative to the listener and to each other in some ambient space that encompasses both the listener and the sound objects. It requires and involves something more than a transient illusion or perception of sound direction; it requires an extended and believable perceptual experience of auditory space that supports some version of reality.

This tutorial focuses on three primary technological areas: loudspeaker sound reproduction, binaural sound reproduction using earphones or headphones, and sound field recording. A related tutorial introduction is [1]. To a lesser extent we consider sound field simulation — the art of using engineering design to create a virtual sound environment. We will assume familiarity with the fundamental processes underlying human 3D sound perception. In particular, we assume familiarity with interaural time difference cues, interaural level difference cues, monaural spectral cues and the necessity for head-tracking when rendering 3D sound over headphones. We also assume familiarity with the following mathematical or signal processing terms: head-related impulse response filters, binaural room impulse responses and head-related transfer functions.

2. LOUDSPEAKER SOUND REPRODUCTION

Four loudspeaker reproduction methods shall be considered: vector-base amplitude and intensity panning, transaural cross-talk cancellation, ambisonics, and wave-field synthesis. Loudspeaker reproduction methods for 3D sound can be classified according to the listening conditions. Vector-base amplitude panning can provide a relatively stable acoustic image across a moderate-sized listening area with some compromises in spatial fidelity. Wave-field synthesis can provide higher spatial fidelity, but requires a substantially more dense loudspeaker array and is often limited in the frequency range for which high spatial fidelity can be achieved. The ambisonics method

*e-mail: craig.jin@sydney.edu.au

generally accommodates only a single listener and has a sweet-spot the size of a listener's head. An advantage of the ambisonic method is its rendering flexibility—it can accommodate wave-field approximation at lower frequencies as well as panning at higher frequencies; scene rotations are also easily handled. Transaural cross-talk cancellation systems attempt to directly control the signal at each ear and are generally targeted at a single or few listeners. These systems often come in the form of a single row of loudspeakers, often above or below a video screen. With the exception of transaural cross-talk cancellation, loudspeaker reproduction methods naturally or automatically encompass head movement. In other words, the loudspeaker array creates a sound field and when the listener moves his/her head the acoustic spatial impression changes accordingly. With transaural cross-talk cancellation, the loudspeaker array simulates a virtual headphone listening condition and head motion must be tracked using some form of head tracker.

2.1. Vector-Base Panning

We begin with vector-base amplitude panning [2] and vector-base intensity panning [3]. In three-dimensions, we pan the amplitude of an acoustic source across the three nearest loudspeakers as shown in Fig. 1. The three loudspeaker positions are associated with the vectors $\mathbf{x}_u$, $\mathbf{x}_v$, $\mathbf{x}_w$ and the virtual source position is indicated by the vector $\mathbf{x}_s$. The virtual source position, $\mathbf{x}_s$, can be expressed as a linear combination of the three loudspeaker positions using matrix algebra as follows:

$$\mathbf{x}_s = \begin{pmatrix} x_s \\ y_s \\ z_s \end{pmatrix} = \begin{pmatrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{pmatrix} \begin{pmatrix} \rho_u(\Omega_s) \\ \rho_v(\Omega_s) \\ \rho_w(\Omega_s) \end{pmatrix}, \quad (1)$$

Fig. 1 A virtual source is panned between three loudspeakers. We define vectors from a listening point to the three loudspeakers as $\mathbf{x}_u$, $\mathbf{x}_v$, $\mathbf{x}_w$ and a vector to the virtual source as $\mathbf{x}_s$.

where the vector $\boldsymbol{\rho}(\Omega_s)$ contains the component weights for the virtual source position vector expressed in terms of the three loudspeaker position vectors. Note that we label the plane-wave direction associated with the virtual source by Ω_s . It is straight-forward to compute the vector $\boldsymbol{\rho}(\Omega_s)$ given the virtual source position:

$$\boldsymbol{\rho}(\Omega_s) = \begin{pmatrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{pmatrix}^{-1} \mathbf{x}_s. \quad (2)$$

With vector base amplitude panning (VBAP), we select the three loudspeakers in the loudspeaker array that are closest to the intended virtual source position. We then set the three VBAP loudspeaker gains, g_u , g_v , and g_w according to $\boldsymbol{\rho}(\Omega_s)$. However, we normalize the loudspeaker gains so that the total energy is unity. In other words, we set the loudspeaker gains as follows:

$$\begin{aligned} g_u &= \frac{\rho_u(\Omega_s)}{\sqrt{\rho_u^2(\Omega_s) + \rho_v^2(\Omega_s) + \rho_w^2(\Omega_s)}}, \\ g_v &= \frac{\rho_v(\Omega_s)}{\sqrt{\rho_u^2(\Omega_s) + \rho_v^2(\Omega_s) + \rho_w^2(\Omega_s)}}, \\ g_w &= \frac{\rho_w(\Omega_s)}{\sqrt{\rho_u^2(\Omega_s) + \rho_v^2(\Omega_s) + \rho_w^2(\Omega_s)}}. \end{aligned} \quad (3)$$

The result of setting the loudspeaker gains according to this equation results in the acoustic *velocity vector* [4] pointing towards the direction of the virtual source. The physical significance of the acoustic velocity vector pointing towards the virtual source is that the interaural time difference cue for the listener will be appropriate at the low frequencies. We say that VBAP panning preserves the low-frequency interaural time difference cues for the virtual acoustic source. This leads naturally to the question of the interaural level difference cues—will they be veridical with VBAP panning? Unfortunately, the answer is not at the higher frequencies.

In order to preserve the interaural level difference cues, we utilize vector base intensity panning. Vector base intensity panning (VBIP) is based on the acoustic energy of the loudspeakers, often referred to simply as *loudspeaker energy*. The loudspeaker energy for the loudspeaker at position, $\mathbf{x}_u$, is given by the square of the loudspeaker gain: g_u^2 . If we weight the three loudspeaker directions by the loudspeaker energies, rather than the loudspeaker gains, we obtain what is often referred to as the energy vector $\mathbf{r}_E$:

$$\mathbf{r}_E = \frac{g_u^2 \mathbf{x}_u + g_v^2 \mathbf{x}_v + g_w^2 \mathbf{x}_w}{g_u^2 + g_v^2 + g_w^2}. \quad (4)$$

Note that the energy vector is normalized by the total acoustic energy. In VBIP, the loudspeaker gains are calculated such that: 1) the vector $\mathbf{r}_E$ points toward the virtual source direction; 2) only the three loudspeakers

that are the most closely associated with the virtual source direction are employed; 3) the total acoustic energy is equal to one. In VBIP, we set the loudspeaker energies, e_u , e_v , and e_w , according to $\rho(\Omega_S)$ as follows:

$$[e_u(\Omega_S), e_v(\Omega_S), e_w(\Omega_S)] = \frac{[\rho_u, \rho_v, \rho_w]}{\rho_u + \rho_v + \rho_w}. \quad (5)$$

Correspondingly, we set the loudspeaker gains as:

$$\begin{aligned} g_u &= \sqrt{\frac{\rho_u(\Omega_S)}{\rho_u(\Omega_S) + \rho_v(\Omega_S) + \rho_w(\Omega_S)}}, \\ g_v &= \sqrt{\frac{\rho_v(\Omega_S)}{\rho_u(\Omega_S) + \rho_v(\Omega_S) + \rho_w(\Omega_S)}}, \\ g_w &= \sqrt{\frac{\rho_w(\Omega_S)}{\rho_u(\Omega_S) + \rho_v(\Omega_S) + \rho_w(\Omega_S)}}. \end{aligned} \quad (6)$$

One can easily verify that the total loudspeaker energy is unity and that:

$$\mathbf{r}_E = \frac{\mathbf{x}_S}{\rho_u + \rho_v + \rho_w}. \quad (7)$$

The difference between the acoustic velocity vector and the acoustic energy vector and the interaural time and level difference cues has been long understood. Nonetheless, when amplitude panning a virtual source between multiple loudspeakers one has a conflict between these two direction vectors — they are not the same, so what should one do? In this case, the duplex theory partially comes to the rescue. The duplex theory originates from Lord Rayleigh and his experiments with tuning forks [5] and indicates that interaural time difference (ITD) cue influences perception at lower frequencies and that the interaural level difference (ILD) cue influences perception at higher frequencies. However, there is no clear agreement on the boundary frequency and it depends on how one explores this boundary; for an entertaining and informative review, refer to Hartmann's recent paper [6]. Most practitioners of vector base panning split the usage of VBAP and VBIP gains at a boundary frequency between 500 Hz and 1,500 Hz. The boundary frequency, however, is not the only concern. There is also the perceived width of the source which is related to the norm of the energy vector, $\mathbf{r}_E$ [7]. In this regard, practitioners quantify the angular spread, $\sigma(\Omega_S)$, of the virtual source using [8]:

$$\sigma(\Omega_S) = 2 \cos^{-1}(2\|\mathbf{r}_E(\Omega_S)\| - 1). \quad (8)$$

With VBAP and VBIP, one does not have independent control of the angular spread of the virtual acoustic source. For example, with VBIP the loudspeaker gains are set by the acoustic energy vector and any variation in the individual loudspeaker energies will necessarily modify the perceived loudness of the virtual source. This implies

that as the source moves in space its perceived width will vary. Efforts to compensate for the variation in acoustic width are described here [8], but one can only really widen the source width by utilizing a greater number of loudspeakers.

2.2. Higher-Order Ambisonics

Higher-Order Ambisonics (HOA) provides a more sophisticated viewpoint for rendering a sound scene using loudspeakers. Nonetheless, at its essence, it still involves panning a virtual sound source across a number of loudspeakers. With HOA, a circular or spherical array of loudspeakers is generally used and one considers panning across a much larger number of loudspeakers than two or three loudspeakers. This approach is particularly effective for ambient sounds, but not so practical for direct sounds. Therefore, more recent approaches separate the direct and ambient components of the sound field [9–12]. A particular advantage of HOA rendering is its ability to work with sound field recordings and its ability to render ambient, reverberant environments. Before delving into the technical details we attempt to characterize the spirit and intuition behind HOA rendering. Consider a sound field together with a listening viewpoint. Without going into the details of sound field recordings, it stands to reason that if one records the sound field with a spherical microphone array at the listening viewpoint, then any loudspeaker reproduction of that sound field should ideally result in the same spherical microphone array recording when the spherical microphone array is positioned at the listening viewpoint of the loudspeaker array. Indeed, this provides one empirical method to validate loudspeaker reproduction techniques. The spherical microphone array enables a specific spherical harmonic decomposition of the sound field and sometimes this spherical harmonic decomposition is referred to as an ambisonic decomposition. Thus, the HOA rendering method attempts to identify the spherical harmonic decomposition of the virtual sound source or virtual sound scene and then reproduce this spherical harmonic decomposition using the loudspeaker array.

2.2.1. Underlying theory

The HOA method of sound field reproduction is based on the idea that the sound field can be described as a sum of spherical harmonic components and can be considered as panning in the spherical harmonic domain. In the frequency domain, any sound field consisting of incoming sound waves can be expressed as a series of spherical harmonic functions [13]:

$$p(r, \theta, \varphi, k) = \sum_{l=0}^{\infty} \sum_{m=-l}^m i^l j_l(kr) Y_l^m(\theta, \varphi) b_{lm}(k), \quad (9)$$

where $p(r, \theta, \varphi, k)$ is the acoustic pressure corresponding to the wave number k and at the point with spherical

coordinates (r, θ, φ) , and where j_l is the spherical Bessel function of degree l and Y_l^m is the spherical harmonic function of order l and degree m . Equation (9), known as the spherical harmonic representation of the sound field, states that the sound field is fully described by the complex coefficients $b_{lm}(k)$. Describing the sound field at any point in space requires an infinite number of b_{lm} coefficients, which is impossible in practice. Nevertheless, a good approximation of the acoustic pressure in the vicinity of the origin can be obtained by truncating the series to an order L that depends upon the wave number, k , and radius, r , from the origin according to the formula [14]:

$$L \geq \frac{ekr - 1}{2}, \quad (10)$$

where e is the mathematical constant known as Euler's number. Hence, the higher the order, the larger the area where the sound field is accurately described, with regard to the wavelength.

2.2.2. Basic HOA decoding

Using HOA to record a sound field consists in obtaining the b_{lm} components up to order- L , which are often referred to as HOA signals. Conversely, using HOA to synthesise a sound field consists in controlling the loudspeakers so that they create a sound field having the desired order- L spherical harmonic representation. In the HOA method, it is usually assumed that the loudspeakers act on the sound field as would plane-wave sound sources. This is a reasonable approximation assuming the distance from the loudspeakers to the center of the listening area is large enough with regard to the order L . More precise methods that consider the distance of both the loudspeakers and the virtual sources more accurately are described in [15–17]. With these more accurate methods, one can even enable near-field sources [15]. Returning to the simplifying assumption that the loudspeakers may be modelled as plane-wave sources, the relation between the speaker signals and the spherical harmonic components of the sound field is a simple matrix-vector product:

$$\mathbf{b} = \mathbf{Y} \mathbf{g} \quad (11)$$

where $\mathbf{b}$ denotes the vector of the spherical harmonic components b_{lm} up to order L , $\mathbf{s}$ is the vector of the speaker complex gains, and $\mathbf{Y}$ is the matrix of the spherical harmonic function values at the loudspeaker angular positions:

$$\mathbf{Y} = \begin{bmatrix} Y_{00}(\theta_1, \varphi_1) & Y_{00}(\theta_2, \varphi_2) & \dots & Y_{00}(\theta_S, \varphi_S) \\ Y_{1-1}(\theta_1, \varphi_1) & Y_{1-1}(\theta_2, \varphi_2) & \dots & Y_{1-1}(\theta_S, \varphi_S) \\ \vdots & \vdots & \ddots & \vdots \\ Y_{LL}(\theta_1, \varphi_1) & Y_{LL}(\theta_2, \varphi_2) & \dots & Y_{LL}(\theta_S, \varphi_S) \end{bmatrix} \quad (12)$$

where S is the number of speakers. The speaker signals are calculated by *decoding* the desired (or recorded) spherical harmonic signals using a *decoding matrix* $\mathbf{D}$:

$$\mathbf{g} = \mathbf{D} \mathbf{b}^0 \quad (13)$$

where $\mathbf{b}^0$ denotes the vector of the desired spherical harmonic components. The sound field reconstruction error can be defined as the norm of the difference between the desired and reconstructed sound field:

$$E = \|\mathbf{b} - \mathbf{b}^0\| = \|(\mathbf{Y}\mathbf{D} - \mathbf{I})\mathbf{b}^0\| \quad (14)$$

where $\mathbf{I}$ denotes the identity matrix.

Clearly, the sound field will be perfectly reconstructed if the decoding matrix inverts perfectly the matrix $\mathbf{Y}$. This leads to the following results:

- A perfect sound field reconstruction is possible only if there at least as many loudspeaker as spherical harmonics. In 3D, this condition translates into the following relation:

$$S \geq (L + 1)^2 \quad (15)$$

- The rank of $\mathbf{Y}$ must be greater or equal to the number of spherical harmonics. This requires that the loudspeaker positions achieve a good angular sampling of the sphere, which means that the loudspeakers must be distributed as evenly as possible over the surface of the sphere.

An obvious solution to the HOA decoding problem is to use the Moore-Penrose pseudo inverse of matrix $\mathbf{Y}$:

$$\mathbf{D} = \text{pinv}(\mathbf{Y}) \quad (16)$$

This solution is known as the *basic* HOA decoding and is a direct analogue of vector-base amplitude panning in the spherical harmonic domain. It results in the acoustic velocity vector pointing toward the virtual source and correct low-frequency interaural time difference cues.

2.2.3. Sweet spot size and alternate decodings

The size of the area where the sound field is accurately described by the order- L spherical harmonic representation of the sound field decreases with frequency, as expressed in Eq. (10). A perfect sound field reconstruction requires the size of this area to be greater than or equal to the size of the listener's head. Assuming the listener's head radius is 10 cm, the maximum frequency value at which the sound field can be reconstructed perfectly for a single listener is given by:

$$f_{\max} = \frac{c}{2\pi} \frac{2L + 1}{0.1e} \quad (17)$$

The corresponding frequency values are presented in Table 1 for up to order 6. Clearly, the sound field will not be perfectly reconstructed for a single listener above a few kHz, unless a very large number of loudspeakers is

Table 1 Maximum frequency for perfect reconstruction as a function of the HOA order.

L	1	2	3	4	5	6
$f_{\max}$ (Hz)	602	1,004	1,406	1,807	2,209	2,611

used. Therefore, above the specified frequency values, it is preferable to use an alternative to the basic decoding procedure that improves the perceived quality of the sound field reconstruction. The most common alternate decoding method is referred to as the *maxrE decoding* and is an analogue of vector-basis intensity panning in the spherical harmonic domain. The maxrE decoding matrix is calculated by multiplying the basic decoding matrix with a weighting matrix, as follows:

$$\mathbf{D}_{\text{maxrE}} = \mathbf{D}\mathbf{W}_{\text{maxrE}} \quad (18)$$

where $\mathbf{W}_{\text{maxrE}}$ is a diagonal matrix whose diagonal components depend only on the order. These weights are calculated so that, when playing back a wave incoming from a particular direction, the speaker signal energy is concentrated in the corresponding direction. Mathematically, this maximizes the norm of the so-called *energy vector*, which can be expressed as:

$$\mathbf{r}_E = \frac{\sum_{i=1}^S \|g_i\|^2 \mathbf{x}_i}{\sum_{i=1}^S \|g_i\|^2} \quad (19)$$

where g_i is the gain of the i th loudspeaker and $\mathbf{x}_i$ is the unit vector pointing towards this loudspeaker from the listening point. For a given maximum order L , the maxrE weights are determined as follows. One first solves $P_{L+1}(x) = 0$, where $P_{L+1}(\cdot)$ is the Legendre polynomial of order $L + 1$. Let x^* be the largest value of x which solves this equation. Then the maxrE weights, w_l , for order $l \leq L$ are solved recursively as follows:

$$\begin{aligned} w_0 &= 1, \\ w_1 &= x^*, \\ w_{l+1} &= \frac{(2l+1)x^*w_l - lw_{l-1}}{l+1}. \end{aligned} \quad (20)$$

A detailed description of the benefits of maxrE decoding and the computation of the weighting matrix can be found in the PhD thesis by Jerome Daniel [4].

2.3. Wave-Field Synthesis

With wave-field synthesis, we consider both a reproduction region of space, Λ , and a control region of space, V , as shown in Fig. 2. We also refer to the loudspeaker signals as driving point functions. The question is how to calculate the driving point functions given the desired sound field. Traditional descriptions of wave-field synthesis (WFS) develop the theory via the Rayleigh integrals and

Fig. 2 Two regions of space are identified: a reproduction region and a control region. Adapted from [23].

the Kirchhoff-Helmholtz integral [17–21] and initially considered planar and linear loudspeaker arrays.

The Rayleigh integral approach contains many approximations and will only be briefly touched upon here. Two readable accounts of WFS may be found in the short tutorial presented at the 2013 European Acoustics Association Winter School [18] and the book by Jens Ahrens [22]. The brief description given here follows the formulation provided in [22]. The Rayleigh I Integral describes the sound field, $S(\mathbf{z})$, in a source-free half-space defined by a boundary plane $\partial\Lambda$. It states that the sound field is uniquely defined by the gradient of the sound field on the planar boundary in the normal direction pointing into the half-space:

$$P(\mathbf{z}) = - \int_{\partial\Lambda} 2 \frac{\partial}{\partial \mathbf{n}} S(\mathbf{z}) \Big|_{\mathbf{z}=\mathbf{z}_0} G(\mathbf{z} - \mathbf{z}_0) dA(\mathbf{z}_0), \quad (21)$$

where $G(\mathbf{z} - \mathbf{z}_0)$ is the transfer function for an acoustic point source, detailed below. Thus, the driving point function, $D(\mathbf{z}_0)$, is given by:

$$D(\mathbf{z}_0) = -2 \frac{\partial}{\partial \mathbf{n}} S(\mathbf{z}) \Big|_{\mathbf{z}=\mathbf{z}_0}. \quad (22)$$

It is here that the complexities of WFS begin and we only provide a summary of these complexities (refer to [22] for details) in this tutorial introduction. What generally occurs next is that many approximations are made. The planar array of loudspeakers accommodating the driving point function degenerates to a linear array that accommodates listening only in the horizontal plane at a height of zero. Because the integral involves integration with a complex exponential, Laplace's method of stationary phase is used to make a high-frequency approximation that asserts that the only important complex exponential terms contributing to the integral occur where the phase of the exponential term is stationary. This approximation leads to a driving point function that depends on the listening position of the listener and so further approximations are made so that the driving point function is constant. In the end, the approximations lead to what is commonly referred to as a 2.5 dimensional solution in which the driving point function has the form:

$$D_{2.5D}(z) = \sqrt{\frac{2\pi d_{\text{ref}}}{ik}} D(z_0) \Big|_{\text{height}=0}, \quad (23)$$

where d_{ref} is a reference distance chosen to optimize the playback and k is the wavenumber, described in more detail below.

From a computational viewpoint, it is easier and more practical to follow a formulation of WFS that follows the simple source formulation as espoused by Fazi [23]. In the following mathematical considerations, we operate in the frequency domain and consider a single frequency only. For simplicity, we omit the frequency dependence of the variables. Using the simple source formulation, the pressure, $p(z)$, at a location, z , within the reproduction region is given by:

$$p(z) = \int_{\partial\Lambda} G(z, y) \mu(y) dA(y), \quad (24)$$

where $G(z, y) = \frac{\rho e^{ik|z-y|}}{4\pi|z-y|}$ is the transfer function for an acoustic point source and $k = \frac{2\pi f}{c}$ is the wavenumber, f is frequency, ρ is the air density, and c is the speed of sound. The transfer function for an acoustic point source is referred to as a Green's function in the literature. As well, the point source is referred to as a monopole loudspeaker in the wave-field synthesis literature. If one knows both the interior sound field, $p_-(z)$, and exterior sound field, $p_+(z)$, as shown in Fig. 3, then the driving point function, $\mu(y)$, is determined as:

$$\mu(y) = \nabla_n p_-(y) - \nabla_n p_+(y), \quad (25)$$

where $\nabla_n p(z) := \nabla p(z) \cdot \hat{n}(z)$ is the gradient of the pressure in the direction of the outward normal.

The complicated mathematical considerations above become significantly simpler when using a matrix formulation. This matrix formulation is similar to the Boundary Surface Control method developed by Ise [24]. Assume one has a set of loudspeaker locations indexed by i and a set of control points indexed by j . One then forms the transfer function matrix, S , where the element $S_{i,j} =$

Fig. 3 The interior sound field, $p_-(z)$ and exterior sound field, $p_+(z)$ are identified. Adapted from [23].

$G(z_i, y_j)$. Let the desired sound field at the control points be represented by the vector y and let the unknown loudspeaker driving point functions be represented by the vector μ . One then computes the loudspeaker driving point functions, μ , as a solution to the system of linear equations:

$$S\mu = y. \quad (26)$$

A common technique to obtain a least-norm solution is to use Tikhonov regularization:

$$\mu = [S^H S + \beta I]^{-1} S^H y. \quad (27)$$

A common value for the regularization parameter is $\beta = 0.03$; refer to [25] for a discussion of regularization. This is then repeated for each frequency. A complicated sound field can be simulated for a given configuration of loudspeakers and control zone using a sound field simulator such as MCRoomSim [26].

With WFS, there are a number of practical considerations. For example, when the number of loudspeakers is large, a sparse solution generally improves performance [27], i.e., solve:

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \{ \|\mu\|_1 : S\mu = y \}. \quad (28)$$

With a large array of loudspeakers one should also compensate for the directivity pattern of the loudspeakers. In addition, one should also scale the size of the control zone proportional to the frequency. One can use the rule-of-thumb that the approximate radius, R , of the control zone should scale as: $R \approx \frac{4\sqrt{N}}{ek}$, where N is the number of loudspeakers, k is the wavenumber, and e is the natural logarithm number. For robustness, one should include some control points inside the control zone. These “inside” control points are sometimes referred to as CHIEF-points in the boundary element method literature [28], where CHIEF refers to combined Helmholtz integral equation formulation. The CHIEF points resolve ambiguities in the sound field solution.

2.4. Transaural Cross-talk Cancellation

We now consider transaural cross-talk cancellation systems. These systems are commonly implemented in the form of a row of loudspeakers as shown in Fig. 4 and aim to control the signals at the two ears. A detailed description and further references are given in [29]. The mathematical description is similar to WFS except there are only two control points: the signals at the two ears. As previously, we operate in the frequency domain and for simplicity omit the frequency dependence of variables. Let the loudspeaker driving point functions be given by μ and the desired binaural pressure signals at the two ears be specified by $y = [y_L, y_R]$. We again solve:

$$S\mu = y, \quad (29)$$

Fig. 4 Each loudspeaker has a transfer function to the location of both the left and right ear. Adapted from [29].

where $\mathbf{S}$ is the transfer function matrix. A least-norm solution for the $M \times 2$ loudspeaker filter matrix, $\mathbf{H}$, is given by:

$$\mathbf{H} = [\mathbf{S}^H \mathbf{S} + \beta \mathbf{I}]^{-1} \mathbf{S}^H, \quad (30)$$

so that $\boldsymbol{\mu} = \mathbf{H}\mathbf{y}$. If one needs to compensate for slight movements of the listener, an adaptive solution for the loudspeaker filter matrix can be obtained from simple beamforming considerations [29]. Note that in all of the discussion above, we do not explicitly consider head-related impulse responses — we simply control the sound field at two points. One could, of course, amend the transfer function matrix to include a consideration of head-related impulse responses.

3. SOUND FIELD RECORDING

Spherical microphone arrays (SMAs) have been the focus of considerable recent research [30–39] and are especially useful for recording panoramic sound scenes. Two recent books [40,41] provide a fairly comprehensive overview. By virtue of their spherical symmetry, SMAs provide a natural framework for analyzing sound fields in the spherical harmonic domain, see Fig. 5.

We now derive the mathematical model used to describe the acoustic behavior of a rigid, baffled SMA. This description is taken from [42] and has been shortened in parts for this tutorial. We consider the case of an SMA consisting of N omnidirectional microphones located at various positions around a perfectly rigid sphere with radius R . As illustrated in Fig. 6, we define the position of the microphones by their spherical coordinates (r, θ, ϕ) . For simplicity, the mathematical expressions are derived in the frequency domain as a function of the dimensionless frequency kR , where k denotes the wavenumber, $k = 2\pi f/c$, f denotes the frequency and c denotes the speed of sound. As well, for a given radial distance, r , we introduce the dimensionless radius, ρ , defined by: $\rho = r/R$.

Consider the n -th microphone of the SMA, whose spherical coordinates are $(\rho_n R, \theta_n, \phi_n)$. In the case where

Fig. 5 The dual-radius SMA prototype described in [42].

Fig. 6 The notations used to describe the geometry of an SMA are illustrated. From [42].

the incident sound field consists of incoming waves, the acoustic pressure measured by this sensor is given by:

$$p_n = \sum_{l=0}^{\infty} \sum_{m=-l}^l w_l(kR, \rho_n) Y_l^m(\theta_n, \phi_n) b_{l,m}, \quad (31)$$

where

- b_{lm} is a complex coefficient depending only on the incident sound field, which we denote as the order- l and degree- m spherical harmonic component.
- Y_l^m denotes the order- l and degree- m real-valued spherical harmonic function:

$$Y_l^m(\theta, \phi) = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} P_l^{|m|}(\sin \theta) \dots$$

$$\times \begin{cases} \cos m\phi & \text{for } m \geq 0 \\ \sin |m|\phi & \text{for } m < 0 \end{cases}, \quad (32)$$

where P_l^m is the order- l , degree- m associated Legendre polynomial. Note that the $\sin \theta$ term arises from the spherical coordinate convention chosen in this paper (see Fig. 6).

- $w_l(kR, \rho_n)$ is the ‘modal strength’ associated with an ideal, rigid spherical baffle, of the order- l spherical harmonic modes at the microphone position and is given by:

$$w_l(kR, \rho_n) = i^l \left(j_l(\rho_n kR) - \frac{j'_l(kR)}{h_l^{(2)'}(kR)} h_l^{(2)}(\rho_n kR) \right), \quad (33)$$

where j_l and $h_l^{(2)}$ denote the order- l spherical Bessel function and spherical Hankel function of the second kind, respectively.

We refer to Eq. (31) as a Bessel-weighted spherical harmonic expansion of the acoustic pressure. In the audio engineering literature, this equation is sometimes referred to as a spherical Fourier transform or a Higher Order Ambisonic (HOA) transform and we use the term HOA synonymously with the expression ‘spherical harmonic.’

According to Eq. (31), the exact value of the pressure is determined by the summation of an infinite number of terms. This sum must be truncated for the pressure to be estimated numerically. In practice, we advise selecting the truncation order, Λ , as the minimum order such that the relative error in the pressure measured by the SMA’s farthest sensor is less than -80 dB (0.01%). Figure 7 shows the truncation order, Λ , as a function of kR for different values of ρ . Details for the calculations to determine the truncation order are provided in [42].

Assuming that an appropriate truncation order, Λ , has been determined, the value of the acoustic pressure measured by the n -th sensor can be approximated as:

Fig. 7 The truncation order Λ is shown as a function of kR for different values of ρ . From [42].

$$p_n \approx \sum_{l=0}^{\Lambda} \sum_{m=-l}^l w_l(kR, \rho_n) Y_l^m(\theta_n, \phi_n) b_{l,m}. \quad (34)$$

Equation (34) expresses the summation over a finite number of terms. It can therefore be rewritten as the following vector product:

$$p_n = \mathbf{t}_{\Lambda,n}^T \mathbf{b}_{\Lambda}, \quad (35)$$

where

$$\begin{aligned} \mathbf{t}_{\Lambda,n} &= [t_{0,0,n}, t_{1,-1,n}, t_{1,0,n}, \dots, t_{\Lambda,\Lambda,n}]^T, \\ t_{l,m,n} &= w_l(kR, \rho_n) Y_l^m(\theta_n, \phi_n), \\ \mathbf{b}_{\Lambda} &= [b_{0,0}, b_{1,-1}, b_{1,0}, \dots, b_{\Lambda,\Lambda}]^T. \end{aligned} \quad (36)$$

Similarly, the vector of the acoustic pressures received by the N microphones of the SMA can be expressed as:

$$\mathbf{p} = \mathbf{T}_{\Lambda} \mathbf{b}_{\Lambda}, \quad (37)$$

where $\mathbf{T}_{\Lambda}$ is the transfer matrix between the HOA components up to order- Λ and the pressure received by the N microphones, given by:

$$\mathbf{T}_{\Lambda} = [\mathbf{t}_{\Lambda,1}, \mathbf{t}_{\Lambda,2}, \dots, \mathbf{t}_{\Lambda,N}]^T. \quad (38)$$

3.1. Higher-Order Ambisonic Encoding

We refer to the process of retrieving the up-to-order- L HOA components from the microphone signals as order- L HOA encoding. In practice, this operation consists in convolving the microphone signals with a matrix of Finite Impulse Response (FIR) encoding filters. In this section, we derive the mathematical expression for calculating the encoding filter frequency responses.

The problem of finding the values of the HOA components from the microphone signals essentially consists in solving Eq. (37) for $\mathbf{b}_{\Lambda}$. In other words, the matrix of the order- L encoding filter frequency responses, $\mathbf{E}_L$, must invert the matrix $\mathbf{T}_{\Lambda}$ as accurately as possible. Mathematically, this means that $\mathbf{E}_L$ must minimize the quantity $\|\mathbf{E}_L \mathbf{T}_{\Lambda} - \mathbf{C}\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm and $\mathbf{C}$ is given by:

$$\mathbf{C} = [\mathbf{I}_{(L+1)^2} \mathbf{0}_{(L+1)^2 \times ((L+1)^2 - (L+1)^2)}], \quad (39)$$

where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix and $\mathbf{0}_{m \times n}$ denotes the $m \times n$ zero matrix.

In practice we solve for $\mathbf{E}_L$ using Tikhonov regularization. In other words, we define $\mathbf{E}_L$ as:

$$\mathbf{E}_L = \arg \min_{\mathbf{A}} (\|\mathbf{A} \mathbf{T}_{\Lambda} - \mathbf{C}\|_F^2 + \beta^2 \|\mathbf{A}\|_F^2), \quad (40)$$

where β is the regularization parameter setting the relative importance given to minimizing the norm of the matrix with respect to minimizing the error. Matrix $\mathbf{E}_L$ is unique and given by:

$$\mathbf{E}_L = \mathbf{C} \mathbf{T}_{\Lambda}^H (\mathbf{T}_{\Lambda} \mathbf{T}_{\Lambda}^H + \beta^2 \mathbf{I}_N)^{-1}$$

Fig. 8 The amplitude of the modal weights, $w_l(kR, \rho)$ are shown as a function of kR and for orders 0 to 4 when $\rho = 1$ (top) and $\rho = 2$ (bottom). From [42].

$$= \mathbf{T}_L^H (\mathbf{T}_A \mathbf{T}_A^H + \beta^2 \mathbf{I}_N)^{-1}. \quad (41)$$

This solution for $\mathbf{E}_L$ incorporates two critical aspects. First, it applies regularization to prevent measurement noise from being amplified unreasonably at low kR values. Second, it optimally weights the microphone signals to minimize spatial aliasing by employing a matrix $\mathbf{T}_A$ of sufficiently high order.

With regard to the amplification of measurement noise, Fig. 8 shows the amplitude of the modal weights w_l as a function of kR and for orders 0 to 4. At low kR values, the weights corresponding to HOA signals of order greater than one have very small amplitudes. As well, in the case where the microphones are not located on the rigid spherical surface ($\rho > 1$), the weights are equal to zero at certain kR values. Without regularization, the encoding filters would amplify the microphone signals greatly to acquire the HOA signals at these frequencies and would thus be very non-robust to measurement noise. The amount of measurement noise in the encoded signals is bounded by the regularization parameter β as follows:

$$\alpha \leq \left(\frac{1}{2\beta} \right)^2, \quad (42)$$

where α is the ratio of the energy of the measurement noise in the HOA signals to the energy of the measurement noise in the microphone signals. The proof of this result is given in [42].

Consider now the issue of spatial aliasing. At high frequencies, microphones can receive energy from undesired, greater-than-order- L , HOA components of the incoming sound. This results in spatial aliasing where the encoded HOA signals are polluted by these undesired HOA components. Figure 7 shows that the microphones with larger ρ receive more energy from the undesired, greater-than-order- L , HOA components than those with smaller ρ . Therefore, at high frequencies it is desirable to apply greater weight to the signals from microphones located closer to the rigid spherical surface. In summary, the high-order information contained in $\mathbf{T}_A$ is essential for optimal calculation of the encoding filters.

4. BINAURAL VIRTUAL SPATIAL AUDIO

4.1. Head-tracking and Individualization

Binaural spatial audio is based on head-related impulse responses and/or binaural room impulse responses. The first and most important point to make is that immersive binaural virtual spatial audio requires head-tracking. Without head-tracking the virtual acoustic sources move with the head and destroy the perceptual illusion. It is the advent of mixed-reality devices with head-tracking that provide a significant impetus for immersive spatial audio in the consumer industry. This then raises the issue of just how important is individualized head-related impulse responses.

In order to illustrate these issues we consider some recent data [43] for an experiment contrasting individualized HRIRs and generic HRIRs. This description is taken from [43]. There were four listening conditions: (1) binaural rendering with individualized HRIR filters and head-tracking; (2) binaural rendering with generic HRIR filters and head-tracking; (3) binaural rendering with individualized HRIR filters and no head-tracking; and (4) normal headphone listening without binaural spatial rendering. There were twenty-three self-reporting normally-hearing listeners participate in the listening test. Listeners were asked to listen to six sound excerpts:

- Mono: drums, Radiohead — Weird Fishes/Arpeggi
- Mono: guitar, Tarrega — Capriccio Arabe
- Stereo: Pop, Radiohead — Jigsaw Falling Into Place
- Stereo: Bossa-Nova, Stan Getz, João Gilberto — Vivo Sonhando
- 5.1 Surround: Rock, Pink Floyd — Money
- 5.1 Surround: Pop Jazz, Norah Jones — Come Away With Me

Sounds were played to the listener using the AKG 1000 open headphones and also a loudspeaker array consisting

Fig. 9 Results from the binaural listening test are shown for the six sound stimuli. The average population scores for the four listening conditions are shown using a bar plot. The legend labels are as follows: Indiv. + HT – Individualized HRIRs with head-tracking; Generic + HT – generic HRIRs with head-tracking; Indiv. no HT – Individualized HRIRs with no head-tracking; No 3D – no binaural spatial rendering. Figures taken from [43].

of 12 loudspeakers: 5 Tannoy System 15 loudspeakers forming a 5.1 arrangement and 7 additional Tannoy V6 loudspeakers forming a circular array spaced every 45 degrees. The loudspeaker playback provided a reference for the headphone listening. Because the headphones are open, the loudspeakers could be heard without distortion. Every listener had HRIRs recorded using a blocked-ear recording method [44] in an anechoic chamber using a semi-circular robotic arm (methods were similar to those presented here [45]). A MUSHRA-like [46] test paradigm was used in which there was no hidden reference, but an anchor was included. The explicit reference was loudspeaker playback and the anchor was headphone presentation with no spatial audio rendering. Listeners participated in two different trials. In one trial listeners were asked to rate overall preference and in another trial listeners were asked to rate the clarity of the frontal image. Head-tracking was implemented using a Polhemus G4 head-tracking device mounted on the headphones.

Results of the listening test are shown in Fig. 9. As expected, head-tracking contributed significantly to the listeners' scores because it provides a consistent listening environment in which sound sources are robustly and consistently localized when the head moves. Interestingly, listeners also showed a small, but consistent bias for individualized binaural rendering over generic binaural rendering. The added benefit of individualized binaural rendering is small compared to the benefit of head-tracking. Nevertheless, in listening conditions without a visual reference, there does seem to be a small benefit for individualization in binaural rendering. This would suggest that individualized binaural rendering will play some

role when visual stimuli are absent—for example, in augmented spatial hearing conditions using hearables.

4.2. Binaural Reverberation and Ambiance

4.2.1. Binaural rendering of HOA signals

Head-related impulse responses are generally recorded under anechoic, free-field conditions. This means that one must consider the issue of how reverberation and ambiance will be handled during binaural rendering. A perceptually reasonable (the author is unaware of precise psychophysical validation) encoding of reverberant and ambient sound fields is provided by order-3 HOA signals. Thus, we consider binaural rendering of HOA signals. Let the pressure, p , at the ears in the presence of N plane waves be given by: $p = Hs$, where H contains the head-related transfer function information and s specifies the plane-wave sources. If the plane-wave sources are encoded as HOA signals, Y , then we require a decoding matrix, F , so that: $\hat{p} = FY$, where ideally, $\hat{p} = p$. A possible mathematical solution can be found using the pseudo-inverse: $F = H \text{pinv}(Y)$. Nevertheless, this does not work well for many reasons: the pseudo-inverse will smear the source signals across directions, as well the phase component of the matrix H is complicated and this solution will result in strong spectral coloration. Therefore, it is recommended that one smoothly fix the phase of the head-related transfer functions above a certain frequency to a constant to obtain H_{mod} . Using $F = H_{\text{mod}} \text{pinv}(Y)$ provides a much improved binaural rendering for order-3 HOA signals provided one has a dense recording of head-related transfer functions.

4.2.2. Manipulation of the covariance matrix

There are many methods to simulate room impulse

responses and also many methods to decorrelate audio signals and it is not uncommon to apply these for binaural rendering of spatial audio. The crux of the issue is that the covariance of the signals at the ears determines many binaural perceptual properties and so it is best to preserve or control the covariance structure of the binaural signals. These matters are well described in [47]. A brief description is provided here. Assume there is a decoding method to obtain binaural signals, p_Q , from a set of input signals: $p_Q = Qx$, where Q is some decoding matrix and x is a vector of input signals and that somehow one knows the desired covariance matrix $C_p = E[pp^H]$ for some true or desired p . In general, $C_{p_Q} \neq C_p$, so what can one do? It turns out there is a method to obtain new signals $\hat{p} = Mx + M_D D[Qx]$, where $D[\cdot]$ is a decorrelation operator and the matrices M and M_D are to be determined, such that $C_{\hat{p}} = C_p$. Unfortunately, these computations are sufficiently complicated that one is referred to [47] for the details. The important point is that there are methods to control the covariance of the signals at the two ears.

4.3. Improving Static Binaural Spatial Audio

One of the more interesting research directions for binaural spatial audio is to enable static binaural recordings to be rendered with head-tracking in an attempt to create an immersive experience. This would be relevant to the broadcasting industry which likely has many historical, static binaural recordings. As well, this would be interesting for personal binaural recordings where listeners record a sound scene using earphones capable of making recordings, such as the Roland CS-10EM (available in 2018). This technique requires the extraction of acoustic sources and source directions from the static binaural recording and the creation of a new binaural rendering with head-tracking. This could likely be done using personal mobile devices. An interesting reference to work in this area is [48]. As well, refer to [49,50] for methods to decompose audio into direct and ambient sound as well as determine the direction of acoustic sources.

REFERENCES

- [1] W. Zhang, P. N. Samarasinghe, H. Chen and T. D. Abhayapala, "Surround by sound: A review of spatial audio recording and reproduction," *Appl. Sci.*, **7**, 1–19 (2017).
- [2] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, **45**, 456–466 (1997).
- [3] J.-M. Pernaux, P. Boussard and J.-M. Jot, "Virtual sound source positioning and mixing in 5.1 implementation on the real-time system genesis," *Proc. Conf. Digital Audio Effects (DAFx-98)*, Barcelona, November, pp. 76–80 (1998).
- [4] J. Daniel, *Représentation de Champs Acoustiques, Application à la Transmission et à la Reproduction de Scènes Sonores Complexes dans un Contexte Multimédia*, Ph.D. thesis, Université Paris 6 (2000).
- [5] J. W. Strutt, "On our perception of sound direction," *Philos. Mag.*, **13**, 214–232 (1907).
- [6] W. M. Hartmann, B. Rakerd and Z. D. Crawford, "Transaural experiments and a revised duplex theory for the localization of low-frequency tones," *J. Acoust. Soc. Am.*, **139**, 968–985 (2016).
- [7] M. Frank, *Phantom Sources using Multiple Loudspeakers in the Horizontal Plane*, Ph.D. thesis, University of Music and Performing Arts, Graz (2013).
- [8] N. Epain, C. T. Jin and F. Zotter, "Ambisonic decoding with constant angular spread," *Acta Acust. united Ac.*, **100**, 928–936 (2014).
- [9] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, **55**, 503–516 (2007).
- [10] C. T. Jin, N. Epain and T. Noohi, "Sound field analysis using sparse recovery," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias and A. Politis, Eds. (John Wiley & Sons, Hoboken, New York, 2018), Chap. 3.
- [11] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo and J. Ahonen, "First-order directional audio coding (dirac)," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias and A. Politis, Eds. (John Wiley & Sons, Hoboken, New York, 2018), Chap. 5.
- [12] P. Archontis and V. Pulkki, "Higher-order directional audio coding," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias and A. Politis, Eds. (John Wiley & Sons, Hoboken, New York, 2018), Chap. 6.
- [13] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography* (Academic Press, London, 1999).
- [14] N. A. Gumerov and R. Duraiswami, *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions* (Elsevier, Amsterdam, 2005).
- [15] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," *Proc. 23rd Int. Conf. Audio Eng. Soc.: Signal Processing in Audio Recording and Reproduction*, May (2003).
- [16] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, **53**, 1004–1025 (2005).
- [17] J. Ahrens and S. Spors, "An analytical approach to sound field reproduction using circular and spherical loudspeaker distributions," *Acta Acust. united Ac.*, **94**, 988–999 (2008).
- [18] S. Spors and F. Zotter, "Spatial sound synthesis with loudspeakers," in *Cutting Edge in Spatial Audio*, F. Zotter, Ed. (EAA Documenta Acustica, 2013).
- [19] A. J. Berkhout, "A holographic approach to acoustic control," *J. Audio Eng. Soc.*, **36**, 977–995 (1988).
- [20] A. J. Berkhout, D. de Vries and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, **93**, 2764–2778 (1993).
- [21] S. Spors, H. Teutsch and R. Rabenstein, "High-quality acoustic rendering with wave field synthesis," *Proc. Vision, Modeling, and Visualization Conf. 2002*, November (2002).
- [22] J. Ahrens, *Analytic Methods of Sound Field Synthesis* (Springer-Verlag, Berlin, 2012).
- [23] F. M. Fazi, *Sound Field Reproduction*, Ph.D. thesis, University of Southampton (2010).
- [24] S. Enomoto, Y. Ikeda, S. Ise and S. Nakamura, "Three-dimensional sound field reproduction and recording systems based on boundary surface control principle," *Proc. 14th Int. Conf. Auditory Display*, Paris, June, pp. 1–8 (2008).
- [25] S. J. Elliot, J. Cheer, J.-W. Choi and Y. Kim, "Robustness and regularization of personal audio systems," *IEEE Trans. Audio Speech Lang. Process.*, **20**, 2123–2133 (2012).

- [26] A. Wabnitz, N. Epain, C. T. Jin and A. van Schaik, "Room acoustics simulation for multichannel microphone arrays," *Proc. AES 127th Conv.*, Sydney, August, pp. 1–6 (2010).
- [27] N. Epain, C. T. Jin and A. van Schaik, "The application of compressive sampling to the analysis and synthesis of spatial sound fields," *Proc. AES 127th Conv.*, New York, October, pp. 1–8 (2008).
- [28] P. M. Juhl, *The Boundary Element Method for Sound Field Calculations*, Ph.D. thesis, Technical University of Denmark (1993).
- [29] S. Gálvez, F. Marcos, T. Takeuchi and F. M. Fazi, "Low-complexity, listener's position-adaptive binaural reproduction over a loudspeaker array," *Acta Acust. united Ac.*, **103**, 847–857 (2017).
- [30] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the sound-field," *Proc. ICASSP 2002*, Vol. 2, pp. 1781–1784 (2002).
- [31] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," *Proc. ICASSP 2002*, Vol. 2, pp. 1949–1952 (2002).
- [32] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, **13**, 135–143 (2005).
- [33] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Trans. Speech Audio Process.*, **9**, 697–707 (2001).
- [34] A. Wabnitz, N. Epain and C. T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," *Proc. ICASSP 2012*, Kyoto, Japan, March (2012).
- [35] S. Bertet, J. Daniel, E. Parizet and O. Warusfel, "Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources," *Acta Acust. united Ac.*, **99**, 642–657 (2013).
- [36] H. H. Chen and S. C. Chan, "Adaptive beamforming and doa estimation using uniform concentric spherical arrays with frequency invariant characteristics," *J. VLSI Signal Process.*, No. 46, pp. 15–34 (2007).
- [37] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty and S. Gannot, Eds. (Springer, Berlin, Heidelberg, 2010).
- [38] H. Sun, H. Teutsch, E. Mabande and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," *Proc. ICASSP 2011*, Prague, Czech Republic, May (2011).
- [39] N. Epain and C. T. Jin, "Independent component analysis using spherical microphone arrays," *Acta Acust. united Ac.*, **98**, 91–102 (2012).
- [40] B. Rafaely, *Fundamentals of Spherical Array Processing* (Springer-Verlag, Berlin, 2015).
- [41] D. P. Jarrett, E. A. P. Habets and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing* (Springer International Publishing, Cham, 2017).
- [42] C. T. Jin, N. Epain and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **22**, 193–204 (2014).
- [43] C. T. Jin, R. Zolfaghari, X. Long, A. Sebastian, S. Hossain, Glaunés, A. Tew, M. Shahnawaz and A. Sarti, "Considerations regarding individualization of head-related transfer functions," *Proc. ICASSP 2018*, Calgary, Canada, April, pp. 6787–6791 (2018).
- [44] H. Moller, "Fundamentals of binaural technology," *Appl. Acoust.*, **36**, 171–218 (1992).
- [45] C. T. Jin, A. Corderoy, S. Carlile and A. van Schaik, "Contrasting monaural and interaural spectral cues for human sound localization," *J. Acoust. Soc. Am.*, **115**, 3124–3141 (2004).
- [46] ITU-R BS.1534-1:2003, Method for the subjective assessment of intermediate quality level of coding systems, ITU-R (2003).
- [47] J. Vilkamo and T. Bäckström, "Time-frequency processing: Methods and tools," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias and A. Politis, Eds. (John Wiley & Sons, Hoboken, New York, 2018).
- [48] S. Nagel and P. Jax, "Dynamic binaural cue adaptation," *2018 Int. Workshop on Acoustic Signals Enhancement*, Tokyo, Japan, September (2018).
- [49] C. Faller, "Multi-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, **54**, 1051–1064 (2006).
- [50] C. T. Jin and N. Epain, "Super-resolution sound field analysis," in *Cutting Edge in Spatial Audio*, F. Zotter, Ed. (EAA Documenta Acustica, 2013).

Craig Jin received a B.S. in Physics from Stanford in 1990, an M.S. in Applied Physics from Caltech, and a Ph.D. from the University of Sydney in 2001. He is currently an Associate Professor within the School of Electrical and Information Engineering at the University of Sydney and is the director of the Computing and Audio Research Laboratory.