# Decoding and Compression of Channel and Scene Objects for Spatial Audio

Dylan Menzies and Filippo Maria Fazi

***Abstract*—Sound fields can be encoded with a fixed number of signals, using microphones or panning functions. The sound field may later be reproduced approximately by decoding the signals to a loudspeaker array. The Stereo and Ambisonic systems provide examples. A framework is presented for addressing general questions about such encodings. The first problem considered is the conversion between encodings. The solution is applied to the decoding of scene encodings to a loudspeaker array. This is generalised to the decoding of *sub-scenes* where the resolution is focused in an angular window. Within an object based audio framework such sub-scenes are useful for representing complex objects without using all the channels required for a full scene. The second problem considered is the compression of a scene encoding to a smaller encoding, from which the original can be reconstructed. The spatial distribution of compression error can be controlled.**

***Index Terms*—IEEE, IEEEtran, journal, LaTeX, paper, template.**

## I. Introduction

A *channel-based* encoding consists of a fixed number of signals that can be used to drive a loudspeaker array in a specific configuration. No assumptions are made about how the signals are derived and what relationships they may have. A *scene-based* encoding consists of signals obtained from real or *virtual* microphones placed at the centre of a sound scene. This type of encoding may require decoding to produce loudspeaker feeds, using some form of gain or filter matrix. If this encoding is designed to be fed directly to an array, then it is also referred to as a channel encoding (A channel encoding doesn't have to be scene-based). Ambisonic encodings provide an example of scene encoding [1;2], based on spherical harmonic directivities. In this article general scene encodings are considered, for which there no assumptions are made about the microphones used.

One realisation of a virtual microphone is using *panning functions*: For plane wave sources a virtual microphone signal is the sum of the source signals weighted by panning functions, the input direction for each function set to the corresponding source. The *directivity* of the panning functions matches the directivity of the real microphone that produces the same output for the corresponding real scene, see Fig. 1.

For a given loudspeaker array, a set of microphones, decoding functions or panning functions are designed so that the original audio scene is reproduced faithfully, which is to say perceived images match the original recorded sources, or intended images. If a set of panning functions were replaced by microphones with the same directivities then the array would reproduce the sound field captured by the microphones. In this case the microphone decoding is trivial. Formal definitions for sound fields, microphone directivities and panning functions are given in Section II.

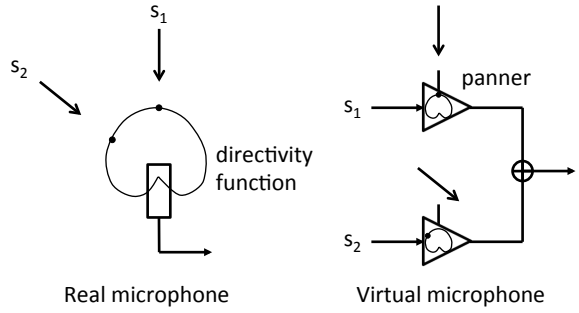



Fig. 1: Relationship between real and virtual microphones, and panning functions. $s_1$ and $s_2$ represent plane wave source signals. The microphone directivity function is shown as a polar plot.

Examples of channel-based systems include Stereo and Ambisonic systems. The Stereo system [3;4] provides the simplest and oldest example of channel reproduction. The listening arrangement assumes a $60^\circ$ loudspeaker separation from the listener, and the stereo channel signals feed directly to the loudpseakers. There are several variations of the panning functions. Tangent law panning is designed to produce the correct Interaural Time Difference (ITD) cue, 0s, when the listener faces an image. The stereo signals are derived using panning functions or from crossed-pair cardioid microphones that approximate their directivities.

In the Ambisonic system, the channels are generated either from microphone arrays or with multichannel panning functions, and the directivity functions are spherical harmonics, [1;2]. A variety of Ambisonic decoders have been designed according to different design criteria [5;6;7;8;9;10;11;12;13;14]. In the original mode matching approach, based on the Fourier-Bessel expansion, the decoding process attempts to reconstruct the sound field physically up to a given spherical harmonic order, 1st order for the original form of Ambisonics and higher for Higher Order Ambisonics (HOA). Higher order reproduction produces the sound field over a wider central region for a given frequency, and for a given order the region shrinks with increasing frequency. The reproduction can be made more accurate by modelling the loudspeaker source waves as point sources rather than plane waves. This is important if the reproduction extends over the whole interior region. The approach is known as Near-field compensated HOA (NFC-

D. Menzies and F.M. Fazi are with the University of Southampton.
Email: d.menzies@soton.ac.uk

HOA)[15]. The driving functions are complex filters, rather than the simple gains. NFC-HOA naturally accommodates near image sources, which also require complex filters under the assumptions for HOA.

Mode matching ensures the low frequency ITD cue is correctly reproduced for all listener head orientations, for a central listener. However mode matching, even with regularisation, may perform poorly for many sparse arrays at higher orders. Also the errors outside the central region restricts the audience size. Each of the decoding methods can be defined using equivalent loudspeaker panning functions: The panning gains for each plane wave are given by first encoding the plane wave to the Ambisonic channel encoding then decoding to the array. Channel-based encoding using Ambisonics enables a compact representation of the sound field, which may include reverberation and complex sources. However it is not possible to directly manipulate the component sounds in a channel-encoding independently from one another.

Vector base amplitude panning (VBAP)[16] is a widely used method for panning over a 3-dimensional loudspeaker array. VBAP is an extension of Stereo tangent law panning to 3D audio. For each target image VBAP provides non-zero panning gains for three near loudspeakers. Equivalently, for each loudspeaker, VBAP defines a continuous panning function. VBAP is usually presented in the context of *object-based audio*, in which each component sound is panned separately at the point of reproduction. However it can also be viewed in terms of channel reproduction using panning functions. From the respect of ITD, VBAP is inferior to mode matched Ambisonic decoding, since in the VBAP case ITD for each image is only accurate when the listener faces it. However for a non-central listener the ITD cues are disrupted in both cases, and the VBAP image is generally less distorted because the active loudspeakers are more localised.

An alternative to the mode matched HOA decoding approach is to prescribe target panning functions, without constraints, then calculate decoding functions which produce panning functions close to these[11]. This approach produces well behaved decoding functions given reasonable panning functions. Also, the perceptual performance of the overall reproduction system is more controllable since it is broadly separated into the perceptual performance of the panning functions and the quality of the encoding that supplies them. A natural choice is to use VBAP panning functions, since VBAP is robust, producing good images for a wide variety of arrays and listener positions, although the advantages of mode-matched decoding for a central listener are usually lost. In the limiting case of a dense symmetric array VBAP-Ambisonic decoding produces similar panning functions to the mode matched approach.

Wave Field Synthesis (WFS)[17], is a reproduction method based on the approximation of boundary source integrals. In its original form it can reproduce a sound field over an interior region up to the spatial aliasing frequency determined by the loudspeaker spacing. The total cost of driving functions for one individual point or plane wave source is low: a equalisation filter, a delay line, and one multiply for each loudspeaker.

If NFC-HOA and WFS are compared using equal numbers of loudspeakers, then up to the spatial aliasing frequency defined by loudspeaker spacing, NFC-HOA reproduction has slightly less error than WFS[18] over the interior. For frequencies above the aliasing frequency HOA error is low in a central region that shrinks with frequency, whereas for WFS the error is usually significant everywhere (Using focused sources it is possible, however, to achieve a low error with WFS in this case[20]). If the spacing is greater than head width, typically 0.17m, so that the alias frequency is less than the top of the ITD range, then the localisation performance of WFS is generally better than NFC-HOA outside the central region[19]. This is unsurprising considering that for each image source WFS driving energy is more localised on the array, and so produces less error outside the central region.

The plane wave driving functions define a set of panning functions. An Ambisonic encoding can be reproduced using WFS by decomposing it first into a plane wave set[21] . This can make sense if high order decoding is required over the full width of an array interior with less complexity than using NFC-HOA filters.

The reproduction methods each have advantages and disadvantages. Each can be expressed using panning functions that define how plane waves can be reproduced, and so the methods are all acceptable in the later examples where panning functions are used.

Unlike pure channel or scene encoding, an object-based audio encoding consists of a variable number of audio objects, each representing a single source of some kind. Each object includes signals and other metadata information, for example about source position and size. The audio objects are rendered to loudspeakers at the point of reproduction. In the context of spatial audio, object-based audio allows the reproduction to be modified and optimised according to the user's reproduction system and room. Object encoding has recently been developed in standards for cinema and interactive broadcasting[22]. As transmission bandwidth increases and reproduction hardware becomes more sophisticated, object-based audio has become more attractive.

Scene encodings can be embedded as objects within an object based encoding. This is useful because such a *scene object* can be used to efficiently capture an element of the sound field that is spatially complex, for which detailed independent control of the internal sound components is not required. An Ambisonic encoding is often used to provide a background scene, or *bed*, However it may not be possible or convenient to use such an encoding, either because of how the object was captured, or in order to minimise and manage the total channel count. For example it is often required to mix recorded Stereo to channel-based or object-based audio, and it is unclear how best to do this. Scene objects where the spatial information is focused in part of the overall scene will be referred to as *sub-scene* objects.

The contributions of this article are summarised in three stages as follows:

1. In Section II a framework is developed to formulate problems about sound fields, directivities, whether these refer to microphone directivities or panning functions, and the signals produced by applying directivities to sound fields. The

aim is to clarify existing discussions about these concepts, which is of general interest, and to prepare a language for formulating problems in the later parts.

2. Ambisonic decoding provides a way to decode an Ambisonic encoding to an array. A method is derived for converting between channel encodings, in Section III, and is used to give a decoding method for general scene objects. This is shown to be a generalisation of an Ambisonic decoding method based on prescribed panning functions[11], which is itself shown to be equivalent to the *AllRAD* decoding method[14]. Ambisonic encoding is shown to be useful in some cases as an intermediate encoding in the decoding process, by allowing more efficient decoding when the channel object is dynamically oriented. An example is given where a Stereo signal is decoded using three different methods. The results from a listening test support proposed approach.

3. Given a channel encoding for some set of directivities, it can be useful to re-encode using fewer channels, in order to reduce storage and transmission requirements, either for a channel encoding or a scene object. In Section II-C a method is given to calculate an optimal reduced encoding. The spatial distribution of error introduced by the re-encoding is controllable. Examples are provided of re-encoding a 7.0 type channel encoding to 5 channels.

### A. Notation

Signals and filters are represented in the frequency domain. For simplicity frequency variables are omitted, although signals, sound fields, and directivity functions have implicit frequency dependence. Vectors and matrices are in bold type. Either may also be represented in component form with normal type, for example the element of matrix $\boldsymbol{A}$ in the $i$-th row and $j$-th column is written $A_{ij}$. $j$ is also occasionally used for $\sqrt{-1}$ , but not simultaneously as an index, so its meaning is always clear. A hat is used to denote a spatial vector of unit length, for example $\hat{\boldsymbol{x}} = \boldsymbol{x}/|\boldsymbol{x}|$. Operators or matrices, vector spaces, and functions are all capitalised. The complex conjugate is represented with a bar, for example $\bar{p}$. The transpose is written $\boldsymbol{A}^T$ and the transpose conjugate is $\boldsymbol{A}^H$. A dual basis or space is represented with an asterix, like $S^*$.

## II. Representing Sound Fields And Microphones

### A. Sound Fields

There are several ways to represent a region of sound field in a 2D or 3D that is free of sources, and so which satisfies the homogeneous Helmholtz equation. The Herglotz expansion (HE), is built from a continuous set of plane wave basis functions. The pressure field as a function of position $\boldsymbol{x}$ and wave number $k$ is

$$p(\boldsymbol{x}, k) = \int_{\hat{\boldsymbol{k}} \in \Omega} e^{-j\boldsymbol{k}\cdot\boldsymbol{x}} s(\boldsymbol{k}) d\Omega \tag{1}$$

where the integration variable is $\hat{\boldsymbol{k}}$ ranging over a surface $\Omega$ of radius 1, a circle in 2 spatial dimensions and a sphere in 3 dimensions. The positive frequency convention, with time dependence $e^{j\omega t}$ is used here for wave representation. The wave vector of each plane wave component is $\boldsymbol{k} = k\hat{\boldsymbol{k}}$, where the direction of travel of the wave is $\hat{\boldsymbol{k}}$. The Herglotz density function $s(\boldsymbol{k})$ contains the information that uniquely represents and encodes the sound field, and can be thought of as a signal density function for the direction $\hat{\boldsymbol{k}}$ and wave number $k$.

For numerical calculation the HE integral cannot be used directly. It can be approximated by sampling over a set of uniformly distributed directions, $\{\hat{\boldsymbol{k}}_i\}$. The encoding is then represented by a function $s(\boldsymbol{k}_i) = s(\hat{\boldsymbol{k}}_i, k)$. For brevity we hide the frequency dependence from $k$ and write the encoding as a vector $\boldsymbol{s}$ with components $s_i = s(\hat{\boldsymbol{k}}_i) = s(\hat{\boldsymbol{k}}_i, k)$. For brevity $\boldsymbol{s}$ and $s(\boldsymbol{k})$ will be referred to *as the sound field*, since it contains the information content of the field, although the actual pressure field is given by (1). We also refer to the *space of sound fields* $S$, which is the vector space of all possible sound fields $\boldsymbol{s} \in S$. The continuous encoding $s(\boldsymbol{k})$ exists in an infinite dimensional space, however in this article we stick to the finite dimensional case and notation. The sound field pressure is now a sum

$$p(\boldsymbol{x}) = \sum_{i=1}^{L} e^{-j\boldsymbol{k}_i\cdot\boldsymbol{x}} s(\boldsymbol{k}_i) \Delta\Omega_i \quad . \tag{2}$$

$\Delta\Omega_i$ are weights that compensate for the arrangement of the plane waves. In the 2D case the directions $\{\hat{\boldsymbol{k}}_i\}$ can be spaced equally, with uniform $\Delta\Omega_j$. In 3D, the optimal choice of $\{\hat{\boldsymbol{k}}_i\}$ and $\{\Delta\Omega_j\}$ is not trivial in general. There exist direction sets for which uniform $\{\Delta\Omega_j\}$ is optimal, such as the *spherical designs* including *t-designs*[23;14]. In any case the reconstruction error can be made arbitrarily small for uniform $\{\Delta\Omega_j\}$ by choosing a uniform array with sufficient $L$. The suitable value for $L$ in the present context will be explained in Section II-B, once other other dependent factors have been explained.

The 2D or 3D sound field region, containing no sources, may also be expanded in terms of a countable set of localised regular harmonic basis functions $\{R_i\}$, or modes,

$$p(\boldsymbol{x}, k) = \sum_{i=1}^{\infty} y_i(k) R_i(\boldsymbol{x}, k) \quad , \tag{3}$$

where $y_i(k)$ are the coefficients or signals encoding the sound field. As with the HE encoding $\{s_i\}$, $\{y_i\}$ shall be referred to *as* the sound field without ambiguity.

The basis functions are othogonal by integration over space,

$$R_i \cdot R_j = \int_{\boldsymbol{x}\in V} R_i(\boldsymbol{x}) \bar{R}_j(\boldsymbol{x}) dV \quad = C\delta_{ij} \tag{4}$$

for a constant $C$. $V$ can be either the whole of 2D or 3D space.

The Fourier Bessel Expansion (FBE) provides a set of such basis functions that are spherically symmetric in 3D space. There are several variations including the *N3D* form[15], which has real valued functions. The basis functions are each the product of a spherical harmonic function of direction and spherical Bessel function of distance. Analogous functions exist in 2D based on sinusoidal functions of azimuth and Bessel functions, a common form is the *N2D* basis,[15]. The term FBE will be used to refer expansions of this type in either 3D or 2D.

An FBE can be approximated by truncation,

$$p(\boldsymbol{x},k) \approx \sum_{i=1}^{M} y_i(k) R_i(\boldsymbol{x},k) \quad , \tag{5}$$

For typical sound fields the approximation is very good within a radius $r$ depending on $M$ and $k$[24]. For 3D harmonics this relationship can be expressed as $N = kr$ where $N$ is the order of expansion and the number of harmonics is $M = (N+1)^2$. This property has been used in the Ambisonic reproduction method for sound field encoding. Variable truncation allows for variable resolution and compatibility of encodings with different resolution. The truncated encoding coefficients form a finite vector, $\boldsymbol{y} = \{y_i\}$, which is a representation of the sound field, like $\boldsymbol{s}$, but with a different basis.

Each plane wave used in an HE can be expanded in terms of an FBE. The expansion coefficients for a plane wave are equal to a set of spherical harmonic functions sampled in the direction of the plane wave[5]. So any sound field approximation $\boldsymbol{s}$ can be expressed with an FBE encoding $\boldsymbol{y}$, using a matrix $\boldsymbol{Y}$ composed from expansions of individual plane waves in $S$,

$$\boldsymbol{y} = \boldsymbol{Y}\boldsymbol{s} \tag{6}$$

(The plane wave weights $\Delta\Omega_i$ from (2) are still included in $\boldsymbol{s}$). As the number of plane waves in $S$ increases, each harmonic coefficient $y_i$ converges separately to the correct value for the actual sound field approximated by $\boldsymbol{s}$.

The spherical harmonics are orthogonal and can be normalised by scaling the FBE basis functions $R_i$, so that in the high order limit $\boldsymbol{Y}\boldsymbol{Y}^H \approx \boldsymbol{I}$. $\boldsymbol{Y}$ is approximately unitary if it is square. For $S$ based on a spherical design then for a maximum number of harmonics $M$ , less than $L$ and depending on the design, $\boldsymbol{Y}\boldsymbol{Y}^H = \boldsymbol{I}$ exactly.

What determines the physical resolution needed to encode a scene, either in terms of harmonics or plane waves? The sampling theorem would suggest that the angular resolution of the plane wave set should at least match the angular resolution of the scene. This is indeed true for pure physical reproduction in which the listener is not considered. Perceptually based panning methods however are able to exploit the human auditory system to achieve *image localisation* resolution that is much better than the encoding resolution. Even two channels, in the case of Stereo, are sufficient for good localisation resolution across a $60^\circ$ range. The main auditory mechanism providing this localisation is the ITD cue. With Ambisonics the picture is similar. The field can be physically reconstructed up to a chosen frequency limit so that it encloses the head and possibly other significant scattering body surfaces[25]. For 1st order Ambisonics ($N = 1$, $M = 4$) gives 500 Hz limit for an adult head, using $N = kr$, which ensures ITD cues are reproduced well. Clearly the information capacity of the encoding is very restricted, however the auditory system is focused on sparse signals, which can be encoded in this way, and is less sensitive to non-sparse information that cannot be. Increasing the encoding resolution allows auditory system to engage more cues, increasing localisation resolution and image quality.

The HE is a natural representation for the sound field in the sense that the signal content in each direction is represented directly. In the FBE the signal components are complex linear combinations of the plane wave signal components. However, as discussed later, there are fundamental and practical reasons why the FBE is sometimes preferable. In the next section the directivity of microphones is represented in a way similar to the HE. Sometimes it may be preferable to choose to represent microphones in a way that parallels the FBE.

### B. Directivity Functions

An ideal microphone is characterised by its directivity function. Specifically, the output $q$ of a perfectly linear microphone is the bilinear function of a complex-valued directivity function (DF) $Q(\hat{\boldsymbol{k}})$, and the sound field encoding $s(\hat{\boldsymbol{k}})$,

$$q = \int_{\hat{\boldsymbol{k}}\in\Omega} Q(\hat{\boldsymbol{k}}) s(\hat{\boldsymbol{k}}) d\Omega \tag{7}$$

The normalised wavevector $\hat{\boldsymbol{k}}$ provides the direction of travel of the wave. Note that microphone directivity is often stated as a function of the reverse direction $\hat{\boldsymbol{\theta}} = -\hat{\boldsymbol{k}}$, opposite the wave travel direction. In the following formula the wave directions are indexed, so we don't have to choose between variables, and no confusion arises.

A panning function provides a loudspeaker gain as a function of the desired image direction. A common goal is to find a set of panning functions for a loudspeaker array such that the output gains produce a perceived image close to the desired image, for a range of desired images. Microphones with the same directivities will produce the same loudspeaker feeds when exposed to the corresponding real scene, as shown before in Fig. 1. DF will be used to refer to both microphone directivities and panning functions.

Using a discretised sound field with components $s_j = s(\hat{\boldsymbol{k}}_j)$ and discretised DF with components $Q_j = Q(\hat{\boldsymbol{k}}_j)$ a microphone signal given by (7) can be approximated as

$$q \approx \sum_{j=1}^{L} Q_j s_j \Delta\Omega_j \tag{8}$$

The discretisation of the sound field and DFs is for computational purposes only, and should be high enough to represent the DFs well enough so that (8) is accurate. If the DFs are band limited as an FBE then there exists a sufficient spherical design that achieves this exactly. For t-designs $L \approx 4M$ is sufficient. Otherwise the error can be made as small as required by choosing $L$ high enough. For offline calculation this presents no practical issue. However if calculation is required across $S$ online then $L$ needs to be chosen considering the error / cost trade off.

A set of DFs, representing a set of microphones or panning functions, can then be written as a matrix $\boldsymbol{Q}$ with components $Q_{ij}$ where $i$ indexes the DFs, and $j$ indexes the plane wave directions. $\boldsymbol{Q}_i$ will denote the $i$th DF as a vector. For the sake of convenience we redefine $s_j$ by absorbing the product with $\Delta\Omega_j$ into it. This keeps the signals and DFs, which are of most interest, constant whatever discretisation is used. Each DF $\boldsymbol{Q}_i$ produces a signal $q_i$ given by (8). The set of signals can be

written as a vector $\boldsymbol{q}$, and the corresponding set of equations is

$$\boldsymbol{q} = \boldsymbol{Q}\boldsymbol{s} \tag{9}$$

It might be considered more consistent to represent the mics as column vectors in $\boldsymbol{Q}$ so that $\boldsymbol{q} = \boldsymbol{Q}^T\boldsymbol{s}$ or $\boldsymbol{q} = \boldsymbol{Q}^H\boldsymbol{s}$, however to simplify notation we follow (9). $\boldsymbol{q}$ can be viewed as a *lossy encoding* of the sound field represented by $\boldsymbol{s}$. To represent the sound field accurately the number of elements in $\boldsymbol{s}$ is much greater than the number of elements in $\boldsymbol{q}$. $\boldsymbol{q}$ then contains less information than $\boldsymbol{s}$ and the original sound field, and $\boldsymbol{Q}$ has *full row rank*.

$\boldsymbol{Q}$ could be defined instead by substituting its transpose $\boldsymbol{Q}^T$ or transpose conjugate $\boldsymbol{Q}^H$. This would perhaps be a more symmetric presentation since DFs are then column vectors like $\boldsymbol{s}$. The choice does not affect the discussion here.

According to (7) DF can be viewed as a linear map from the space of sound fields $S$ to the complex numbers $\mathbb{C}$, so the space of all possible directivity functions is the *dual space* $S^*$, of $S$. $S$ and $S^*$ have the same internal structure but represent different types of object. The subspace of $S^*$ spanned by DFs in $\boldsymbol{Q}$ will be written

$$S^*_{\boldsymbol{Q}} \subseteq S^* \quad . \tag{10}$$

For any sensible choice of DFs they are linearly independent, and form a basis of a subspace in $S^*$. We assume this case unless stated otherwise. A linearly dependent set is a *frame*[26], and may arise for example when two different basis sets are joined together.

Equation (9) can also be viewed as an equation with an operator $\boldsymbol{Q}$ that acts on the sound field according to the original definition in (7). All subsequent expressions have analogs with this interpretation. We focus on the discrete case to give a presentation in terms of familiar matrix operations. For background on the linear algebra and matrix results employed here refer to[27].

### C. Compact Source Representation

Some technical difficulties arise when using plane wave expansions. $\boldsymbol{s}(\hat{\boldsymbol{k}})$ is not simply defined for any field where the exterior region contains a source at a finite distance. This can be seen for example by looking at the sequence of FBEs of a monopole field including orders up to $N$, for $N = 1, 2, 3...$ The FBE sequence converges at every point in the interior region. However if the sequence is then re-expanded as HEs we find the encoding functions $\boldsymbol{s}_N(\hat{\boldsymbol{k}})$ do not converge to a limit, in fact they are unbounded. This is explored in[28].

Even though $\boldsymbol{s}_N(\hat{\boldsymbol{k}})$ is non-convergent, if DFs in $\boldsymbol{Q}$ are spatially band-limited, meaning spherical harmonic coefficients of sufficient order are zero, then a physically valid signal vector $\boldsymbol{q} = \boldsymbol{Q}\boldsymbol{s}$ can still be found. This is because the corresponding sequence $\boldsymbol{q}_N$ then does converge, and therefore converges to the physically realised value. The difficulty arises because of the sound field representation rather than the microphone physics. This strict band-limit condition could be relaxed to allow for DFs with harmonic components that decay sufficiently fast, however there is no practical need to discuss this further here.

## III. Conversion Between Signal Sets

### A. Theoretical Background

A range of problems can be expressed as the task of converting one set of signals associated with a set of DFs to another set of signals associated with another set of DFs. The DFs are known, but the associated sound field is not. One specific example is finding signals for a standard multichannel microphone set given a non-standard set of microphone signals. Another example is finding loudspeaker feeds, for an array with defined panning functions, from signals of a possibly unrelated microphone set. This conversion is an example of decoding called since it is the process of reproducing the sound field from a set of encoding signals. This will be the focus of Sections III-B and III-D.

More precisely, given signals $\boldsymbol{q}$ and DFs $\boldsymbol{Q}$ such that $\boldsymbol{q} = \boldsymbol{Q}\boldsymbol{s}$ for an unknown sound field $\boldsymbol{s}$, what is the best estimate for signals $\boldsymbol{r}$ such that $\boldsymbol{r} = \boldsymbol{R}\boldsymbol{s}$ for known DFs $\boldsymbol{R}$? Generally there will be many sound fields satisfying $\boldsymbol{q} = \boldsymbol{Q}\boldsymbol{s}$, since $\boldsymbol{Q}$ has full row rank as noted previously. A natural estimate is the sound field $\tilde{\boldsymbol{s}}$ for which the $L^2$ norm $\|\boldsymbol{s}\|$ is a minimum, since this will have the least total energy. Nearly all possible sound fields $\boldsymbol{s}$ have excessively high energy, being distant from $\tilde{\boldsymbol{s}}$. Sparsity is another possible criteria useful for estimating $\boldsymbol{s}$, since natural sound fields are sometimes sparse. The $L^1$ norm is one way to select for sparsity, and can be used in combination with energy criteria, but this is not explored here. The kind of sound field that is usefully represented by channel encodings is dense and complex rather than sparse.

If $\boldsymbol{Q}$ has full row rank, as discussed in the previous section, then the least power estimate, written $\tilde{\boldsymbol{s}}_{\boldsymbol{q}}$ to indicate the dependence on known signals $\boldsymbol{q}$, can be calculated using the Moore-Penrose pseudo inverse $\boldsymbol{Q}^+$, which can be calculate in this case using $\boldsymbol{Q}^H(\boldsymbol{Q}\boldsymbol{Q}^H)^{-1}$

$$\tilde{\boldsymbol{s}}_{\boldsymbol{q}} = \boldsymbol{Q}^+\boldsymbol{q} \tag{11}$$

Usually the DFs $\boldsymbol{Q}$ are chosen to be linearly independent so that the pseudo inverse is well conditioned. This may not always be the case, for example microphones at low frequency become nearly linearly dependent. The pseudo inverse can be extended to include Tikhonov regularisation[29],

$$\boldsymbol{Q}^+ = \boldsymbol{Q}^H(\boldsymbol{Q}\boldsymbol{Q}^H + \beta\boldsymbol{I})^{-1} \tag{12}$$

$\beta$ controls the amount of regularisation, and can vary with frequency. This has the effect of reducing signal strength and leads to filters that are simpler and better behaved.

Equation (11) can also be viewed as an expansion of $\tilde{\boldsymbol{s}}_q$ with a set of *dual* vectors $\{\boldsymbol{Q}^*_j\}$ where $\boldsymbol{Q}^*_j = \{Q^+_{ij}\}$, the $j^{th}$ column vector of $\boldsymbol{Q}^+$,

$$\tilde{\boldsymbol{s}}_{\boldsymbol{q}} = \sum_j Q^+_{ij}q_j = \sum_i q_i\boldsymbol{Q}^*_i \tag{13}$$

The inner product of vectors $\boldsymbol{Q}_i$ with vectors $\boldsymbol{Q}^*_j$ defines a matrix

$$\boldsymbol{Q}_i \cdot \boldsymbol{Q}^*_j = \sum_k Q_{ik}Q^+_{kj} \tag{14}$$

$$= \boldsymbol{Q}\boldsymbol{Q}^+ = \boldsymbol{I} = \delta_{ij} \quad . \tag{15}$$

The equivalence to the identity follows because $\boldsymbol{Q}$ has full row rank. $\{\boldsymbol{Q}_i^*\}$ is therefore a *dual basis* for $\{\boldsymbol{Q}_i\}$. In the current context $\boldsymbol{Q}^+$ are referred to as the *dual sound fields* to the DFs $\boldsymbol{Q}$. The space spanned by the dual basis will be called

$$S_{\boldsymbol{Q}^*} \subseteq S \quad . \tag{16}$$

The subspaces $S_{\boldsymbol{Q}^*}$ and $S^*_{\boldsymbol{Q}}$ are dual to one another. The space of sound fields $S$ can be written as an orthogonal inner direct sum of $S_{\boldsymbol{Q}^*}$ and the null space $S_{\boldsymbol{Q}_0} = \{\boldsymbol{s} : \boldsymbol{Q}\boldsymbol{s} = 0\}$:

$$S = S_{\boldsymbol{Q}^*} \oplus S_{\boldsymbol{Q}_0} \quad , \tag{17}$$

since any sound field can be written as $\boldsymbol{s} = \tilde{\boldsymbol{s}_q} + (\boldsymbol{s} - \tilde{\boldsymbol{s}_q})$, where $\boldsymbol{q} = \boldsymbol{Q}\boldsymbol{s}$, $\tilde{\boldsymbol{s}_q} \in S_{\boldsymbol{Q}^*}$ and $(\boldsymbol{s} - \tilde{\boldsymbol{s}_q}) \in S_{\boldsymbol{Q}_0}$ ( since $\boldsymbol{Q}(\boldsymbol{s} - \tilde{\boldsymbol{s}_q}) = \boldsymbol{q} - \boldsymbol{q} = 0$ ).

Fig. 2 illustrates this relationship and other quantities discussed below. From the estimate $\tilde{\boldsymbol{s}_q}$ the corresponding estimate

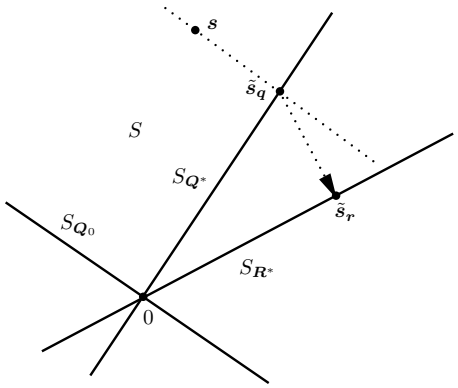



Fig. 2: Illustration of relationships between quantities in the sound field space $S$. The view is normal to the plane of $\tilde{\boldsymbol{s}_q}$ and $\tilde{\boldsymbol{s}_r}$. $\boldsymbol{s}$ may be out of the plane.

for $\boldsymbol{r}$ is

$$\tilde{\boldsymbol{r}} = \boldsymbol{R}\tilde{\boldsymbol{s}_q} = \boldsymbol{R}\boldsymbol{Q}^+\boldsymbol{q} \tag{18}$$

$\boldsymbol{R}\boldsymbol{Q}^+$ is then the transcoding / decoding matrix from signals $\boldsymbol{q}$ to $\tilde{\boldsymbol{r}}$. If there exists a DF in $\boldsymbol{R}$ that is proportional with a DF in $\boldsymbol{Q}$, $\boldsymbol{R}_k = \alpha\boldsymbol{Q}_i$, then signal $r_k$ can be recovered exactly, as we would hope: From (18) $\tilde{r}_k = \alpha\boldsymbol{Q}_i\boldsymbol{Q}^+\boldsymbol{q} = \alpha q_i$, since $\boldsymbol{Q}\boldsymbol{Q}^+ = \boldsymbol{I}$, so $\tilde{r}_k = r_k$.

### B. Relation To Existing Decoding Methods

The sound fields can be approximated using the truncated FBE, instead of the sampled HE. The FBE encodes the sound field information with a set of coefficients or signals $\{y_i\}$, described in equation (3). An HE is converted to an FBE using (6),

$$\boldsymbol{y} = \boldsymbol{Y}\boldsymbol{s} \tag{19}$$

which can be interpreted as a set of microphones $\boldsymbol{Y}$ acting on $\boldsymbol{s}$. Substituting $\boldsymbol{Y}$ for $\boldsymbol{Q}$ in (18) gives

$$\tilde{\boldsymbol{r}} = \boldsymbol{R}\boldsymbol{Y}^+\boldsymbol{y} \tag{20}$$

This can be viewed as the microphone equation $\boldsymbol{r} = \boldsymbol{R}\boldsymbol{s}$ rewritten in terms the FBE, so $\tilde{\boldsymbol{r}} = \boldsymbol{R}_Y\boldsymbol{y}$, with microphones $\boldsymbol{R}_Y = \boldsymbol{R}\boldsymbol{Y}^+$

$\boldsymbol{R}\boldsymbol{Y}^+$ is equivalent to an Ambisonic decoding matrix based on defined panning functions $\boldsymbol{R}$ arrived at by previous authors[5;11]. Batke[11] chooses $\boldsymbol{R}$ to be the panning functions that arise in VBAP, providing a way to decode an Ambisonic encoding as if the component sounds were individually panned using VBAP. In the *AllRAD*[14] Ambisonic decoding method VBAP is also used. The method first decodes onto a regular virtual array. Each virtual feed is then panned onto the actual array using VBAP. This is equivalent to applying the output of the virtual array to virtual microphones defined by VBAP, and can be written using the same form as (20). If a spherical design array is used, as proposed for AllRAD, then $\boldsymbol{Y}\boldsymbol{Y}^H = \boldsymbol{I}$, so that $\boldsymbol{Y}^+ = \boldsymbol{Y}^H(\boldsymbol{Y}\boldsymbol{Y}^H)^{-1} = \boldsymbol{Y}^H$, and the decoding matrix in (20) can be simplified to $\boldsymbol{R}\boldsymbol{Y}^H$.

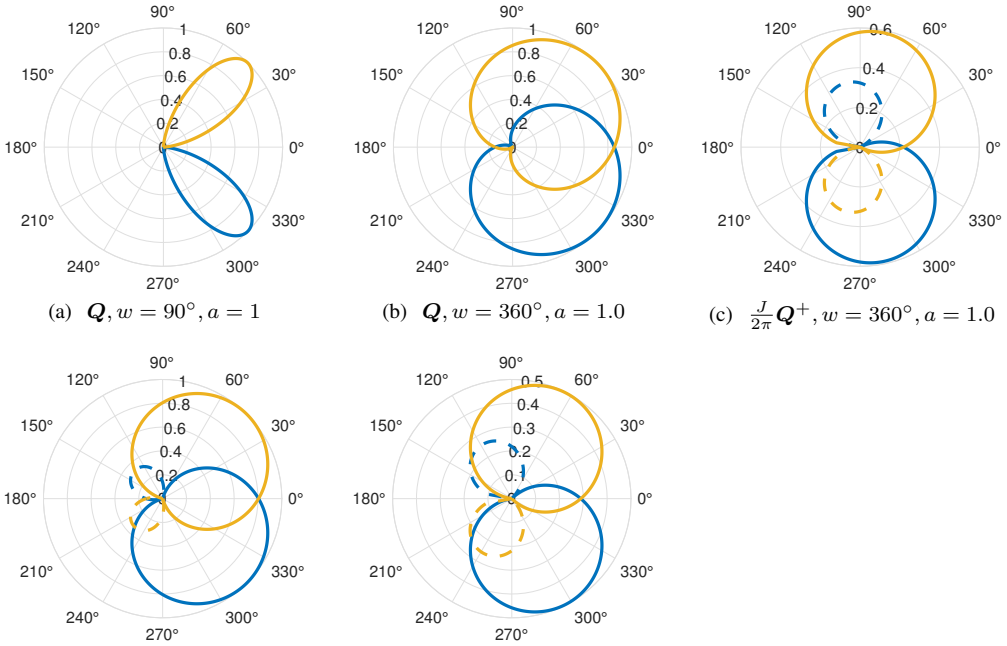
To reduce computation $\boldsymbol{R}\boldsymbol{Y}^+$ can be pre-calculated. In a listener position adaptive system[30], the VBAP panning matrix $\boldsymbol{R}$ is updated when the listener changes position relative to the loudspeaker array, and then $\boldsymbol{R}\boldsymbol{Y}^+$ should be recalculated.

### C. Transforming Encodings

Within an object-based context it may be useful to transform a scene object at the reproduction point, which should be done as efficiently as possible. Transformations might include moving the central direction of the object, rotating about this direction, and spreading the object. It may be convenient to include metadata with the object to set initial transformations. The central direction, and rotation about this direction, can be transformed by rotating the DFs $\boldsymbol{Q}$ used in (18). If $\boldsymbol{\Theta}$ is a rotation acting on a column vector microphone then the microphones $\boldsymbol{Q}$ rotated are $\boldsymbol{Q}\boldsymbol{\Theta}^T$, and $\boldsymbol{\Theta}^T = \boldsymbol{\Theta}^H = \boldsymbol{\Theta}^+ = \boldsymbol{\Theta}^{-1}$ (each microphone is a row vector in $\boldsymbol{Q}$). Consequently $\boldsymbol{Q}^+$ is $(\boldsymbol{Q}\boldsymbol{\Theta}^T)^+ = \boldsymbol{\Theta}\boldsymbol{Q}^+$, using the unitarity of $\boldsymbol{\Theta}$. The transcoding matrix is $\boldsymbol{R}\boldsymbol{\Theta}\boldsymbol{Q}^+$. The last form allows the pseudoinverse to be precalculated rather than calculated after manipulation at the reproduction point. Direct rotation in $S^*$ or $S$ is not very convenient. One approach is to oversample the DF then rotate using linear interpolation. If the DF is already oversampled this reduces computational cost, but requires more space. Another approach is to generate an intermediate FBE, with signals $\boldsymbol{y} = \boldsymbol{Y}\boldsymbol{Q}^+\boldsymbol{q}$, for which the transcoding matrix is $\boldsymbol{R}\boldsymbol{Y}^+\boldsymbol{\Theta}_{\boldsymbol{Y}}\boldsymbol{Y}\boldsymbol{Q}^+$, where $\boldsymbol{\Theta}_{\boldsymbol{Y}}$ is the rotation in a form that acts on the FBE signals $\boldsymbol{y}$. $\boldsymbol{Y}\boldsymbol{Q}^+$ can be pre-calculated for efficiency, and $\boldsymbol{\Theta}_{\boldsymbol{Y}}$ can be calculated efficiently using an iterative process. If the DFs are approximated well using less than the maximum order represented by $\boldsymbol{y}$ then the rotation can be speeded up by restricting to the this order. To further save computation it may be worthwhile to pre-encode the scene object using the significant signals in $\boldsymbol{y}$ rather than $\boldsymbol{q}$, provided the number of encoded signals is not significantly increased. More general spatial transformations can be incorporated in this framework[31].

### D. Sub-scene Object Encoding / Decoding with Examples

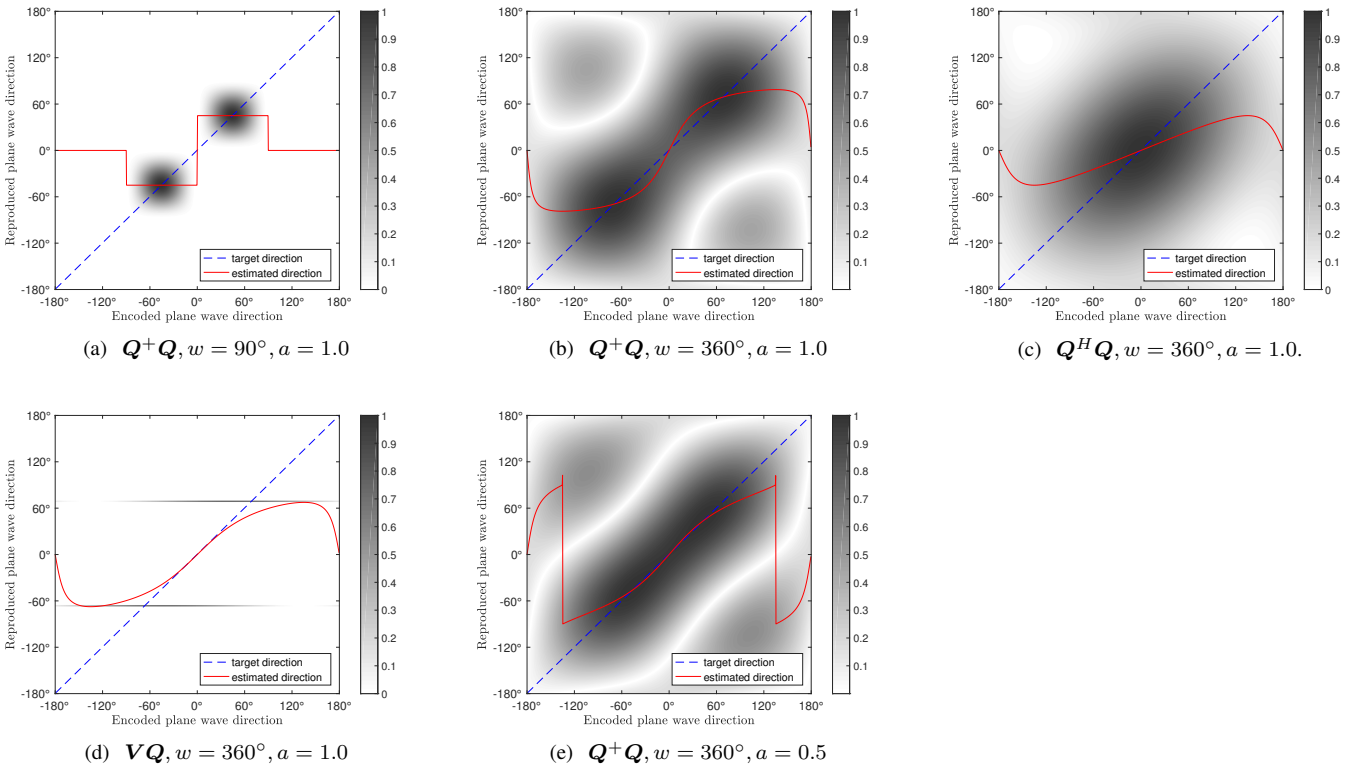
The existing decoding methods referred to in Section III-B act on encodings of an entire sound field, based either on the sampled HE or truncated FBE. The FBE encoding is often used to provide a background sound field, or bed, that spans



(a) $\boldsymbol{Q}, w = 90°, a = 1$ (b) $\boldsymbol{Q}, w = 360°, a = 1.0$ (c) $\frac{J}{2\pi}\boldsymbol{Q}^{+}, w = 360°, a = 1.0$

(d) $\boldsymbol{Q}, w = 360°, a = 0.5$ (e) $\frac{J}{2\pi}\boldsymbol{Q}^{+}, w = 360°, a = 0.5$

Fig. 3: Directivity patterns for various microphones, $\boldsymbol{Q}$, and dual sound fields $\boldsymbol{Q}^{+}$ .




(a) $\boldsymbol{Q}^{+}\boldsymbol{Q}, w = 90°, a = 1.0$ (b) $\boldsymbol{Q}^{+}\boldsymbol{Q}, w = 360°, a = 1.0$ (c) $\boldsymbol{Q}^{H}\boldsymbol{Q}, w = 360°, a = 1.0.$

(d) $\boldsymbol{V}\boldsymbol{Q}, w = 360°, a = 1.0$ (e) $\boldsymbol{Q}^{+}\boldsymbol{Q}, w = 360°, a = 0.5$

Fig. 4: Sound field estimates from encoded plane waves. Also shown are the target direction and the estimated overall direction.

all directions. It may also be useful to use a sub-scene object that is focused in a particular range of directions. For example Stereo recordings obtained with crossed pair of microphones are very common. If we wish to transmit only part of a scene it could be costly to transmit the channels required for a full scene.

Sub-scene encodings can be produced either directly from microphones that are focused in one region, or by panning source signals. Sub-scene encodings can also be extracted from high order microphone signals by using (18) where in this case $\boldsymbol{R}$ are the sub-scene DFs, and $\boldsymbol{Q}$ are the high order microphone directivities.

The spatial information of a Stereo encoding is focused around the line between the microphone directions. To encode spatial information across a region, rather than a line, requires a minimum of 3 signals, using DFs that are directed to the corners of a triangle, rather than a line. In the following examples some general features of sub-scene decoding will be illustrated using Stereo decoded to a 2-dimensional loudspeaker array.

Equation (18) can be applied as a sub-scene decoder $\boldsymbol{R}\boldsymbol{Q}^+$ where $\boldsymbol{R}$ are the loudspeaker panning functions and $\boldsymbol{Q}$ are the sub-scene DFs. This will also be compared with two other naive decoding approaches. In the *virtual stereo loudspeaker* approach, the signals $\boldsymbol{q}$ are panned discretely using $\boldsymbol{R}$ in the directions given by stereo loudspeaker positions. The decoding function can be written as $\boldsymbol{R}\boldsymbol{V}$, where the columns of $V$ are delta vectors. It is not clear how optimal this reproduction is however, since it does not take into account the particular directivites $\boldsymbol{Q}$, nor does it make full use of the loudspeaker array to reproduce directions encoded by $\boldsymbol{Q}$. Another naive approach is to replace the dual sound fields $\boldsymbol{Q}^+$ in (18) with $\boldsymbol{Q}^H$. In other words the sound field driven by each encoded signal is equal to the DF for that signal. All three decoding methods use prescribed loudspeaker panning functions $\boldsymbol{R}$.

In the following examples several microphone directivities are used that have different degrees of overlap and shape. These are based on variations of the general cardioid, and have the form

$$Q(\theta) = \begin{cases} \frac{1}{(1+a)}(a + \cos((\theta - \theta_c)\frac{2\pi}{w}), & \text{if } |\theta - \theta_c| < w/2, \\ & 2\pi - |\theta - \theta_c| < w/2 \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

where $\theta_c$ is the angle of the beam centre, $w$ is the beam width angle, $a = 1$ for a cardioid type shape ($2\pi - |\theta - \theta_c| < w/2$ catches the case where the beam crosses $\theta = 0$). A pure cardioid response is given by $a = 1, w = 2\pi$, a hyper-cardioid by $0 < a < 1, w = 2\pi$, and figure of eight (velocity) response by $a = 0, w = 2\pi$. The angle variables in the formula use radians. They will be quoted in degrees when indicated by the degree symbol $^\circ$.

The encodings are decoded to a hexagonal 6 channel array using standard tangent law pairwise panning functions, which are equivalent to 2D Vector Base Panning (VBAP)[16]. This panning function is chosen because of it provides images that are well localised and robust to listener location. The tangent law, arranged for the ratio of a pair of loudspeaker gains is,

$$\frac{g_1}{g_2} = \frac{\tan\theta_0 + \tan\theta}{\tan\theta_0 - \tan\theta} \tag{22}$$

where $2\theta_0$ is the angle between the loudspeakers and $\theta$ is the angle of the desired image relative to the direction midway between the loudspeakers. The final gains are calculated by applying a normalisation $g_1^2 + g_2^2 = 1$.

As outlined in Section I, for a decoder that has prescribed loudspeaker panning functions, the overall performance can be separated into the performance of this panner $\boldsymbol{R}$, and the performance of the function driving the panner, either $\boldsymbol{Q}^+$, $\boldsymbol{Q}^H$ or $\boldsymbol{V}$ in this study. The focus in this section is on the performance of the function driving the panner. The performance of loudspeaker panners has been studied extensively.

The approximation introduced by the encoding process is measured by comparing the estimated sound field with the sound field that is encoded. For the proposed decoder $\boldsymbol{R}\boldsymbol{Q}^+$, the estimated sound field is $\boldsymbol{Q}^+\boldsymbol{q} = \boldsymbol{Q}^+\boldsymbol{Q}\boldsymbol{s}$. Each column of $\boldsymbol{Q}^+\boldsymbol{Q}$ is then the estimated sound field for a single plane wave. The closer $\boldsymbol{Q}^+\boldsymbol{Q}$ is to the identity $\boldsymbol{I}$, the identity, the more accurate is the estimated sound field.

To begin we check that the decoder $\boldsymbol{R}\boldsymbol{Q}^+$ is consistent with the trivial decoding for the ideal Stereo system. In this the DFs for the encoded stereo signals are the same as the loudspeaker panning functions, and the loudspeaker feeds are equal to the stereo signals. In this case $\boldsymbol{R} = \boldsymbol{Q}$, and the feeds are then $\tilde{\boldsymbol{r}} = \boldsymbol{Q}\boldsymbol{Q}^+\boldsymbol{q} = \boldsymbol{q}$, equal to the stereo signals as required. The same applies for any encoding - decoding system where $\boldsymbol{R} = \boldsymbol{Q}$.

If the non-zero regions of the encoding microphone directivity functions do not overlap then they are orthogonal, and the dual sound fields $\boldsymbol{Q}^+$ have the same shape as the encoding microphones. This is seen in the first example where two microphones are encoded with a $w = 90^\circ$ and separated by $90^\circ$, as shown in Fig. 3a.

Fig. 4a plots $\boldsymbol{Q}^+\boldsymbol{Q}$, the transfer matrix from the desired sound field to the estimated sound field via the microphone channels. Reading the plot vertically for one horizontal position in the plot gives the estimated sound field for a single encoded plane wave. If the transfer were ideal then the matrix would be the identity, as shown by the dashed line. In Fig. 4a there are two isolated diffuse regions. This does not take into account the psychoacoustics of the reproduction matrix $\boldsymbol{R}$ on the overall spatial effect, however it does provide a useful view of the quality of spatial information being fed into the reproduction stage.

For each encoded plane wave an overall estimated direction is calculated by finding the absolute maximum of each estimated sound field in $\boldsymbol{Q}^+\boldsymbol{Q}$, and comparing with the ideal case where the estimate direction is equal to the encoded wave direction. These are included in Fig. 4b. The maximum is representative because the sound field directivities are symmetrical about the maximum direction for these examples. The estimated direction is unreliable for plane wave directions where the estimated sound field is weak and diffuse.

The case $w = 360°$ corresponds to pure cardioid response. The cardioid configuration shown in Fig. 3b is common in Stereo microphones, with significant overlap of the DFs. The dual sound fields shown in Fig. 3c are differentiated from the microphone directivities, with significant negative lobes (The factor $\frac{J}{2\pi}$, where $J$ is the number of plane wave direction samples, makes the scale equal with the integral definition of sound field (1), prior to discretisation

The corresponding sound field estimate for encoded plane waves Fig. 4b shows agreement with the identity across a continuous range, with a varying degree of blur. The range extends outside the $(-45, +45)°$ range of the central encoding microphone directions. There is some relatively low level negative gain in this plot. If it were required as part of the overall strategy that there were no negative gains or reduced negative gains in $\boldsymbol{Q}^+\boldsymbol{Q}$ then this could possibly be built into the calculation of the pseudo inverse $\boldsymbol{Q}^+$.

Using the naive decoder $\boldsymbol{R}\boldsymbol{Q}^H$, the sound field estimates for plane waves are $\boldsymbol{Q}^H\boldsymbol{Q}$, and are less defined and accurate compared with those of $\boldsymbol{Q}^+\boldsymbol{Q}$, Fig. 4c.

For the virtual loudspeaker approach the decoder is $\boldsymbol{R}\boldsymbol{V}$. Then $\boldsymbol{V}\boldsymbol{Q}$ gives the mapping from measured plane wave to estimated sound field, Fig. 4d. This is non-zero across two horizontal lines, separated by a wide gap. Although the estimated direction is reasonable, the gap implies the image may be unstable or blurred. Stereo images quickly become more unstable as the loudspeaker separation increases beyond $60°$.

For a hypercardioid response with $a = 0.5$ (Fig. 3d), the spatial encoding measured by $\boldsymbol{Q}^+\boldsymbol{Q}$ (Fig. 4e) is improved when compared with the cardioid mics (Fig. 4b). .,. The decoding method is based on the assumption of equal weight for errors in all directions. However a sub-region encoding is only useful for a limited range of directions, and so we know a priori that sound field energy is not required for reproduction in directions away from this (ignoring complications of low order Ambisonics at low frequency). A simple way to address this is to apply zero gain in the reverse directions in (Fig. 3d), providing a small improvement in reproduction. Another possible approach would be to apply spatial regularisation in the sound field estimation process, adding greater weight to the cost of reverse reproduction.

### E. Averaged Signal Estimate Error

The signal estimate error $\|\tilde{\boldsymbol{r}} - \boldsymbol{r}\|$, based on $(18)$, depends on the unknown sound field. However we can calculate an overall signal error estimate by averaging over a set of representative sound fields. The norm of the sound fields must be limited otherwise the error is unbounded. Here we consider the average over plane waves $\{\boldsymbol{s}_i\}$ of fixed amplitude in all directions, defined by $\boldsymbol{s}_i(\hat{\boldsymbol{k}}_j) = \delta_{ij}$. In terms of the sound field $\boldsymbol{s}$ the signal estimate is

$$\tilde{\boldsymbol{r}} = \boldsymbol{R}\boldsymbol{Q}^+\boldsymbol{q} = \boldsymbol{R}\boldsymbol{Q}^+\boldsymbol{Q}\boldsymbol{s} = \tilde{\boldsymbol{R}}\boldsymbol{s} \tag{23}$$

with the definition $\tilde{\boldsymbol{R}} = \boldsymbol{R}\boldsymbol{Q}^+\boldsymbol{Q}$.

Ignoring overall normalisation, the averaged signal error is

$$\sqrt{\sum_i \|\tilde{\boldsymbol{r}}(\boldsymbol{s}_i) - \boldsymbol{r}(\boldsymbol{s}_i)\|^2} \tag{24}$$

$$= \sqrt{\sum_{i,j,k} |(\tilde{R}_{kj} - R_{kj})\boldsymbol{s}_i(\hat{\boldsymbol{k}}_j)|^2} \tag{25}$$

$$= \sqrt{\sum_{k,i} |\tilde{R}_{ki} - R_{ki}|^2} = \|\tilde{\boldsymbol{R}} - \boldsymbol{R}\| \quad , \tag{26}$$

which is the Frobenius norm of the matrix $\tilde{\boldsymbol{R}} - \boldsymbol{R}$. Further more the estimator $\tilde{\boldsymbol{R}} = \boldsymbol{R}\boldsymbol{Q}^+\boldsymbol{Q}$ gives the lowest total error $\|\tilde{\boldsymbol{R}} - \boldsymbol{R}\|$ of all the possible estimators $\tilde{\boldsymbol{R}} \in S^*_{\boldsymbol{Q}}$. This is because $\boldsymbol{R}\boldsymbol{Q}^+\boldsymbol{Q}$ is the projection of the DFs $\boldsymbol{R}_i$ onto $S^*_{\boldsymbol{Q}}$, and so for each $i$ $\boldsymbol{R}_i\boldsymbol{Q}^+\boldsymbol{Q}$ is the closest DF in $S^*_{\boldsymbol{Q}}$ to $\boldsymbol{R}_i$. This is emphasised in the next section by writing the subscript $_{\boldsymbol{Q}}$ in $\tilde{\boldsymbol{R}}_{\boldsymbol{Q}} = \boldsymbol{R}\boldsymbol{Q}^+\boldsymbol{Q}$. In summary, the signal estimates $\tilde{\boldsymbol{r}} = \boldsymbol{R}\boldsymbol{Q}^+\boldsymbol{q}$ are optimal in this averaged sense.

### F. Listening test

A listening test was created to compare reproduction using the decoder based on (18), labeled RQ, with virtual stereo reproduction, labeled VL. For both conditions the stereo encoding used was produced using a simulated crossed pair of hypercardioid microphones, with $a = 0.5$ and separation $90°$. The position of the virtual stereo loudspeakers was at $\pm 30°$ relative to the listener. A third condition was included as a reference, consisting of reproduction by discrete panning of images (labeled VBAP).

A 7.0 horizontal array was used with loudspeaker positioned at $0°, \pm 30°, \pm 90°, \pm 135°$., in a room measuring 3 x 4 m, and meeting the acoustic conditions specified by ITU-R BS 1116-1[32]. For each condition 3 images were reproduced, a stream located at $50°$, a woman's voice at $0°$, and a violin at $-50°$.

Eight subjects participated in the experiment. All had some experience with audio engineering, and reported normal hearing. Each subject was provided with a keyboard, with which they can select between four short repeating sounds by pressing number keys, and toggle playback using the space key. The sounds could be listened to in any order as many times as required to reach a decision. The first sound was the reference, and the remaining three sounds were the two conditions and the reference, hidden, in random order. The subject was asked to rate each condition on a scale from 1 (worst) to 10 (best), for two criteria. First according to the overall image localisation, with the listener at the central listening position, and secondly according to stability of image localisation to movement of the listener about the central listening position. The reference is prescribed a score of 10 for both variables.

Two subjects scored the hidden reference lower than the reference for both variables, however all scores were retained. Boxplots from the results are shown in Fig. 5. At first glance these indicate a clear ranking between the conditions.

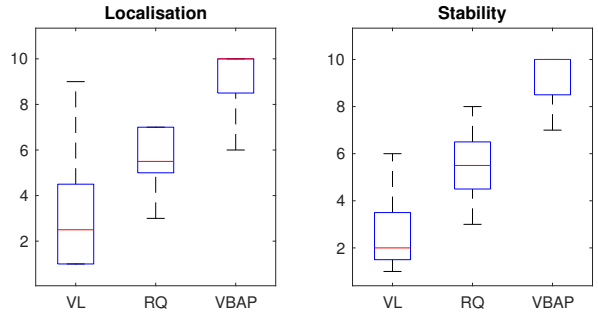



Fig. 5: Boxplots of scores for localisation and stability, for three conditions: virtual loudspeaker (VL), decoder (RQ), and discrete panning (VBAP).

The significance of the apparent rankings was tested statistically. Using a paired t-test[33], the p-values for comparison of localisation scores for pairs of conditions are:

| | |
|---|---|
| VL, RQ | $p = 0.0234$ |
| RQ, VBAP | $p = 0.0006$ |

The t-test p-values for comparison of stability scores for pairs of conditions are:

| | |
|---|---|
| VL, RQ | $p = 0.0052$ |
| RQ, VBAP | $p = 0.0022$ |

Since the sample is small and normal distribution is an assumption of the t-test, the scores were also tested using binomial test[33] on the rank of pairs of scores. The p-values for localisation are:

| | |
|---|---|
| VL, RQ | $p = 0.0352$ |
| RQ, VBAP | $p = 0.0352$ |

The binomial p-values for stability pair rank are:

| | |
|---|---|
| VL, RQ | $p = 0.0352$ |
| RQ, VBAP | $p = 0.0352$ |

Even though the sample is small, the p-values suggest the means of the underlying populations, for both variables, are ranked in the order VL $<$ RQ $<$ VBAP. Furthermore, the mean scores for RQ are positioned roughly midway between VL and VBAP. To this extent the original goal of improved sub-scene decoding has been met. Clearly many other scenarios can be investigated.

Comments by subjects indicate that RQ and VBAP reproduced the lateral images more accurately than VL, for which images were more restricted in range. Also the stability of the central image was less for VL: For VBAP this image is produced by the central loudspeaker, and for RQ the central loudspeaker contributes. However the lateral images produced by VL were stable since they were each focused on one loudspeaker. The images produced by RQ were all blurred to some extent because they all used two or more loudspeakers. In this respect the comparison with VL is less straightforward, and in retrospect could also be tested for.

## IV. Finding An Optimal Reduced Encoding

Given a set of DFs, it may be useful to find a smaller set of DFs so that the encoding can be reduced to a smaller encoding from which the original can be reconstructed. The reduction may be desirable because transmission is required over a lower bandwidth channel, or when several possibly overlapping sets are combined. The goal then is to find the reduced set for which the original signals can be reconstructed as well as possible.

### A. Theory

The reduced DF set will be chosen to minimise the difference between the signals from the original DFs and their estimates derived from the reduced DF signals, averaging over all plane waves. So, given N DFs $\boldsymbol{Q}$, how should M DFs $\boldsymbol{B}$ with $M < N$ be chosen to minimise the total estimated signal error, $\|\tilde{\boldsymbol{q}} - \boldsymbol{q}\|$, over all plane waves, where $\tilde{\boldsymbol{q}} = \boldsymbol{Q}\boldsymbol{B}^{+}\boldsymbol{b}$, $\boldsymbol{b} = \boldsymbol{B}\boldsymbol{s}$?

The discussion leading from (24) has shown that the best estimate for DFs $\boldsymbol{Q}$, constructed from DFs $\boldsymbol{B}$, is $\tilde{\boldsymbol{Q}}_{\boldsymbol{B}} = \boldsymbol{Q}\boldsymbol{B}^{+}\boldsymbol{B}$, with average error $\|\tilde{\boldsymbol{Q}}_{\boldsymbol{B}} - \boldsymbol{Q}\|$. Hence the problem is equivalent to finding the value $\boldsymbol{B}$ giving the least error for estimates $\tilde{\boldsymbol{Q}}_{\boldsymbol{B}}$,

$$\boldsymbol{B}_{\boldsymbol{Q}} = \arg\min_{\boldsymbol{B}}(\|\tilde{\boldsymbol{Q}}_{\boldsymbol{B}} - \boldsymbol{Q}\|) \tag{27}$$

which is the DF set, spanning $S^*_{\boldsymbol{B}} \subseteq S^*$, that minimises the overall distance from $S^*_{\boldsymbol{B}}$ to the vectors $\{\boldsymbol{Q}_{\boldsymbol{i}}\}$, illustrated in Fig. 6. The optimum set is not unique, since any basis for $S^*_{\boldsymbol{B}}$ provides an alternative set.

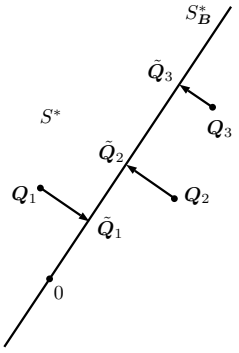



Fig. 6: Choice of $S^*_{\boldsymbol{B}}$ in $S^*$ minimising distance from mics $\boldsymbol{Q}$.

Singular Value Decomposition (SVD) solves this problem directly[27]: Given mics $\boldsymbol{Q}$, SVD provides unitary $\boldsymbol{U}$,$\boldsymbol{V}$ and diagonal $\boldsymbol{\Sigma}$ such that

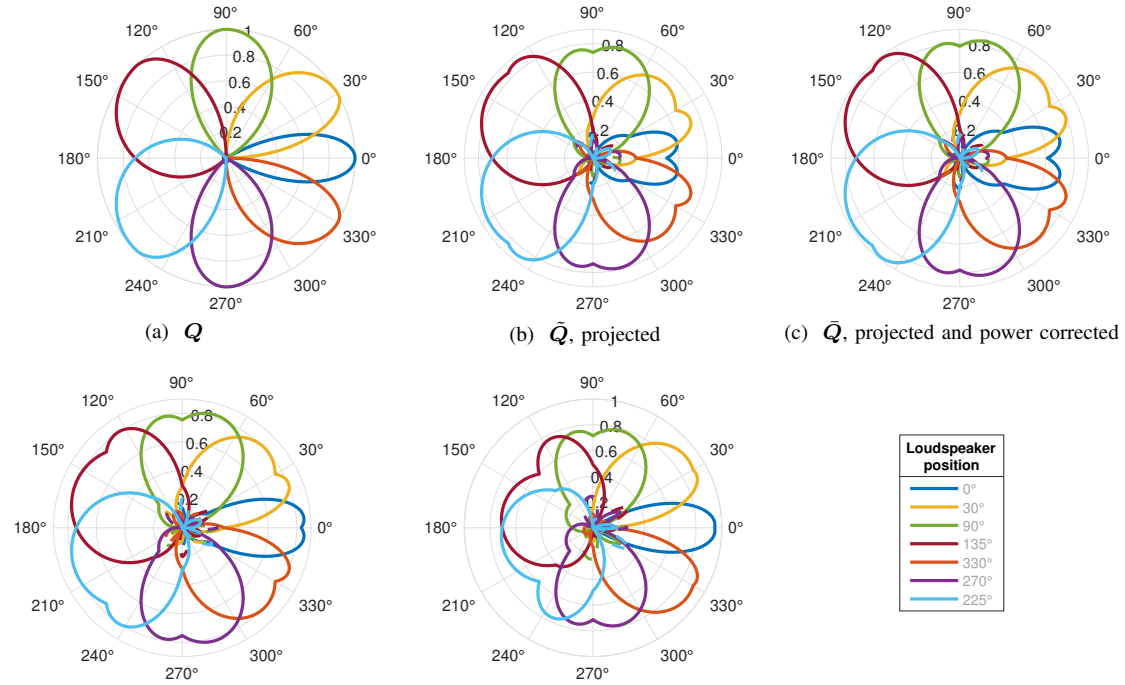
$$\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{H} \tag{28}$$

and the diagonal entries of $\boldsymbol{\Sigma}$, the singular values, are real valued, non-negative, and ordered by decreasing size, $\boldsymbol{V}^{H}$ the complex conjugate of $\boldsymbol{V}$. Then a solution to (27) is given by

$$\boldsymbol{B}_{\boldsymbol{Q}} = \left[\boldsymbol{V}^{H}\right]_{M} \tag{29}$$

denoting the restriction of $\boldsymbol{V}^{H}$ to the first M rows, which form an orthonormal basis, with the most significant vector first. The projection of $\boldsymbol{Q}$ into the space spanned by basis DFs $\boldsymbol{B}_{\boldsymbol{Q}}$ is

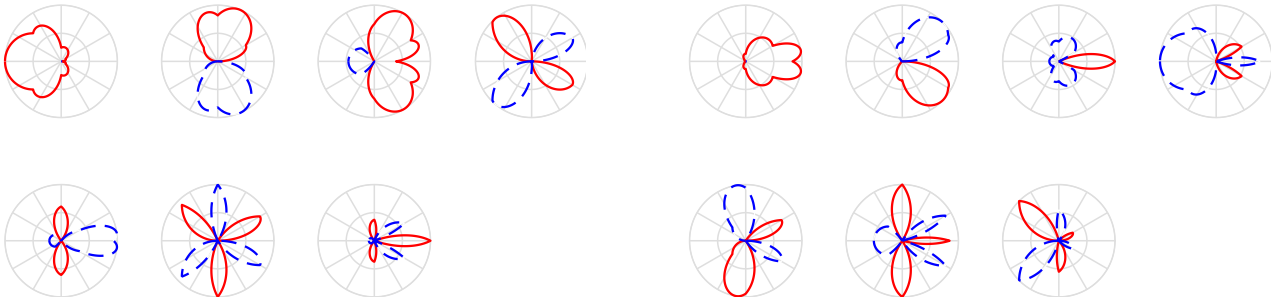
$$\tilde{\boldsymbol{Q}} = \boldsymbol{Q}\boldsymbol{B}_{\boldsymbol{Q}}^{H}\boldsymbol{B}_{\boldsymbol{Q}} \tag{30}$$




(a) $\boldsymbol{Q}$ (b) $\tilde{\boldsymbol{Q}}$, projected (c) $\bar{\boldsymbol{Q}}$, projected and power corrected

(d) $\bar{\boldsymbol{Q}}$, pre-weighted, projected and power corrected (e) $\bar{\boldsymbol{Q}}$, pre-weighted with front emphasis, projected and power corrected

Fig. 7: Polar plots of panning functions for a 2D 7-loudspeaker array, targets $\boldsymbol{Q}$, and reconstructed $\tilde{\boldsymbol{Q}}, \bar{\boldsymbol{Q}}$.



(a) $\boldsymbol{B_Q}$ for target microphones $\boldsymbol{Q}$, Fig. 7a (b) $\boldsymbol{B_Q}$ for front weighted target microphones $\boldsymbol{Q}$, Fig. 7e

Fig. 8: Polar plots of basis panning functions $\boldsymbol{B_Q}$, ordered left to right, from the top, most significant first. Dashed lines indicate negative values.

where $\boldsymbol{B}_{\boldsymbol{Q}}^H \boldsymbol{B_Q}$ is a projection operator, since $\boldsymbol{B_Q} \boldsymbol{B}_{\boldsymbol{Q}}^H = \boldsymbol{I}$ by orthonormality of $\boldsymbol{B_Q}$. Furthermore $\boldsymbol{B}_{\boldsymbol{Q}}^H = \boldsymbol{B}_{\boldsymbol{Q}}^+$ as $\boldsymbol{B_Q}$ has full row rank. The projection can also be evaluated as

$$\tilde{\boldsymbol{Q}} = [\boldsymbol{U\Sigma}]^M \boldsymbol{B_Q} \tag{31}$$

where $[\boldsymbol{U\Sigma}]^M$ is the restriction of $\boldsymbol{U\Sigma}$ to the first $M$ columns.

Finally, the reduced signals are given by

$$\boldsymbol{b} = \boldsymbol{B_Q} \boldsymbol{Q}^+ \boldsymbol{q} \quad , \tag{32}$$

and to reproduce estimates of the original signals from the encoded signals,

$$\tilde{\boldsymbol{q}} = \boldsymbol{Q} \boldsymbol{B}_{\boldsymbol{Q}}^H \boldsymbol{b} \quad . \tag{33}$$

### *B. An Example*

Fig. 7a shows panning functions $\boldsymbol{Q}$ for a 2D 7-loudspeaker array, of the type used in 7.0 reproduction systems. The panning functions are derived from the tangent law (22), used previously. In this case the panning function for each loudspeaker is asymmetric because each is comprised of two neighbouring pair-wise panning segments of differing angular extent. In practice a set of 7.0 channel signals may not be obtainable strictly using these panning functions or similar, but it is likely to be close.

The method of the previous section is applied to find an encoding with less than 7 signals. The resulting basis

panning functions $\boldsymbol{B_Q}$ are shown in Fig. 8a. These resemble distorted 2D harmonics. Interestingly, if the number of panning functions $\boldsymbol{Q}$ are increased and arranged uniformly, then $\boldsymbol{B_Q}$ converge on 2D circular harmonic functions.

Using just the first five of the basis panning functions, the projected panning functions $\tilde{\boldsymbol{Q}}$ resemble the original functions $\boldsymbol{Q}$, with some expected distortions (Fig. 7b).

### C. Power Correction

In Fig. 7b the panning functions, especially the narrower ones, are broadened in comparison with the originals. Even taking this into account there is a non-uniform change of relative power $\|\tilde{\boldsymbol{Q}}_i/\boldsymbol{Q}_i\|^2$ The subspace $S_B^*$ is drawn closer to the more powerful $\boldsymbol{Q}_i$ that have greater $\|\boldsymbol{Q}_i\|$. The power balance can be restored by applying gain to the projected DFs to form new encoding DFs $\bar{\boldsymbol{Q}}_i$ that all have the same power as the originals,

$$\bar{\boldsymbol{Q}}_i = \tilde{\boldsymbol{Q}}_i \frac{\|\boldsymbol{Q}_i\|}{\|\tilde{\boldsymbol{Q}}_i\|} \tag{34}$$

The panning functions $\bar{\boldsymbol{Q}}_i$ are shown in Fig. 7c.

This process is not ideal. $\bar{\boldsymbol{Q}}$ will not in general have the least error among constant power DFs. For this the SVD process would need to be replaced by an optimisation that has the equal power constraint built in, and this is not so easy to obtain in a simple form.

### D. Detailed Error

The remaining error between $\bar{\boldsymbol{Q}}_i$ and $\boldsymbol{Q}_i$ is due to difference of shape. This can be measured in detail by subtracting them first. The detailed relative power error is defined as

$$\epsilon_i = \|\bar{\boldsymbol{Q}}_i - \boldsymbol{Q}_i\|/\|\boldsymbol{Q}_i\| \tag{35}$$

The values of $\epsilon_i$ for each loudspeaker angular position are:

| | |
|---|---|
| $0°$ | $0.32$ |
| $30°, 330°$ | $0.16$ |
| $90°, 270°$ | $0.11$ |
| $135°, 225°$ | $0.06$ |

Detailed error for the centre loudspeaker panning function is particularly significant, as this is the most important channel in 7.0 reproduction.

### E. Weighting Between DFs

The less powerful DFs suffer greater detailed error $\epsilon_i$. This is because the error function in (27) is a sum across all the DF error, so less power DFs carry relatively less weight. Pre-emphasising the less powerful DFs should give more uniform detailed error. This can be achieved by pre-weighting the DFs before calculating the basis using SVD,

$$\boldsymbol{U\Sigma V}^* = \boldsymbol{\Delta Q} \tag{36}$$

where $\boldsymbol{\Delta} = \mathrm{diag}(\delta_i)$. Increasing the weight $\delta_i$ for each DF will reduce the error $\|\tilde{\boldsymbol{Q}}_i - \boldsymbol{Q}_i\|$ for each DF $\tilde{\boldsymbol{Q}}_i$ approximated using the optimised DF set $\boldsymbol{B_Q}$. Appropriate weights for balancing the strengths of the DFs are given by the inverse of the DF lengths,

$$\delta_i = \frac{1}{\|\boldsymbol{Q}_i\|} \tag{37}$$

The resulting projected DFs, carry the emphasis forward. This is rebalanced by power correction, as before. The results are shown in Fig. 7d

The corresponding detailed errors $\epsilon_i$ for each loudspeaker angular position are:

| | |
|---|---|
| $0°$ | $0.10$ |
| $30°, 330°$ | $0.14$ |
| $90°, 270°$ | $0.13$ |
| $135°, 225°$ | $0.12$ |

which shows are much more even spread of error, compared with the previous errors without pre-weighting.

Pre-weights can be increased further to selectively boost the accuracy of some reconstructed DFs at the expense of others. For example the front stage is usually more important for the listener, so the accuracy should be increased at the front. Combining several DF sets that are each normally used separately, it may be useful to assign a weight to each set reflecting its relative importance. The following additional pre-weight factors were applied to each $\boldsymbol{Q}_i$ identified by loudspeaker angular position:

| | |
|---|---|
| $0°$ | $1.6$ |
| $30°, 330°$ | $1.6$ |
| $90°, 270°$ | $1.2$ |
| $135°, 225°$ | $1.0$ |

The basis functions for this case, shown in Fig. 8b, reflect the increased significance of the front panning functions. Fig. 7e shows that the resulting panning functions are much closer to the target functions, Fig. 7a, where they have been pre-emphasized.

The detailed errors $\epsilon_i$ for each loudspeaker angular position are :

| | |
|---|---|
| $0°$ | $0.02$ |
| $30°, 330°$ | $0.04$ |
| $90°, 270°$ | $0.20$ |
| $135°, 225°$ | $0.27$ |

The error for the front panning functions at $0°, 30°, 330°$ is low enough that the perceptual difference from the target would be very small. For the other panning functions, at the sides and rear, the errors are small enough that the perceptual difference is expected to be small, especially for diffuse sound sources that are typically applied in those regions.

An advantage of this method over using an Ambisonic encoding with harmonic panning functions is that some of the encoding functions can be made as close as desired to the target functions, using only a few channels. This is because the basis set, $\boldsymbol{B_Q}$, transforms according to the pre-emphasis applied, and is not fixed. This is particularly useful where the target panning functions are of varying widths or importance.

### F. Direction Weighting

An alternative approach is to weight by direction within DFs, rather than weight between DFs. This is possible by transforming to a weighted sound field description before calculating the basis, then unweighting the sound field after calculating the projected DFs. Using SVD we find $\boldsymbol{U}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{V}^H$ such that

$$\boldsymbol{Q}\boldsymbol{\Lambda} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H \tag{38}$$

with weights $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_i)$, with larger entries emphasising the importance of the corresponding direction. This has the equivalent effect to applying a weight to the sum over direction in (24). The DF estimates are then formed by unweighting the projections,

$$\tilde{\boldsymbol{Q}} = \boldsymbol{Q}\boldsymbol{B}_{\boldsymbol{Q}}^{*}\boldsymbol{B}_{\boldsymbol{Q}}\boldsymbol{\Lambda}^{-1} \tag{39}$$

DF and direction weighting can be combined in the SVD decomposition,

$$\boldsymbol{\Delta}\boldsymbol{Q}\boldsymbol{\Lambda} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{*} \tag{40}$$

Yet another emphasis method is to sample the directivity functions $Q_i(\tilde{\boldsymbol{k}})$ with a non-uniform set of directions $\{\boldsymbol{k_i}\}$, which can be considered as a form of pre-warping. The error $\|\tilde{\boldsymbol{Q}}_i - \boldsymbol{Q}_i\|$ is then reduced for regions that are relatively more densely sampled. Further details are not considered here.

## V. Conclusions

A framework was presented for representing sound fields, microphone directivity functions and panning functions, and the resulting signals. A method was found for converting signals from one directivity set to another, based on intermediate estimation of the sound field. This is compatible with conventional decoding methods including Stereo and Ambisonics, and allows the decoding of general scene encodings, including sub-scenes, to arbitrary loudspeaker arrays in a rational manner. It was shown how some existing Ambisonic decoding methods are included as particular cases. An important general feature is that the psychoacoustical content of the loudspeaker panning functions is separated from the process of mapping the encoding functions on to the panning functions.

The overlap between encoding directivity functions allows channel signals to be compressed into fewer channels and restored approximately. While this causes some loss of spatial accuracy, it is a linear process, so the temporal fine structure of the signals is preserved. Weighting can be used to distribute spatial encoding accuracy non-uniformly in the compressed signals.

The work was motivated initially by the need to develop better representations and reproduction methods for object based audio. Sub-scenes are a way to encode complex sources or create scene building blocks that minimize channel count to practical levels. These ideas, and others, will be incorporated in production and reproduction tools that are currently in development. It will then be possible to further evaluate and develop the methods within a working end to end environment. All data supporting this study are openly available via https://doi.org/10.5258/SOTON/D0200.

## Acknowledgment

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

## References

[1] P. G. Craven and M. A. Gerzon, "Coincident microphone simulation covering three dimensional space and yielding various directional outputs," Aug. 16 1977, uS Patent 4,042,779.

[2] M. A. Gerzon, "General metatheory of auditory localisation," in *92nd Audio Engineering Society Convention, Vienna*, no. 3306, 1992.

[3] B. Bernfeld, "Attempts for better understanding of the directional stereophonic listening mechanism," in *Audio Engineering Society Convention 44*, no. C-4, March 1973. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=1743

[4] A. D. Blumlein, "British patent specification 394,325 (improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems)," *Journal of the Audio Engineering Society*, vol. 6, no. 2, pp. 91–130, 1958.

[5] M. Poletti, "Robust two-dimensional surround sound reproduction for nonuniform loudspeaker layouts," *Journal of the Audio Engineering Society*, vol. 55, no. 7/8, pp. 598–610, 2007.

[6] D. Arteaga, "An ambisonics decoder for irregular 3-d loudspeaker arrays," in *Audio Engineering Society Convention 134*, May 2013. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16818

[7] E. Benjamin, A. Heller, and R. Lee, "Design of ambisonic decoders for irregular arrays of loudspeakers by non-linear optimization," in *Audio Engineering Society Convention 129*, Nov 2010. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=15665

[8] M. A. Gerzon, "Design of ambisonic decoders for multispeaker surround sound," NRDC, Tech. Rep., 1977.

[9] D. Moore and J. Wakefield, "The design and analysis of first order ambisonic decoders for the itu layout," in *Audio Engineering Society Convention 122*, May 2007. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14038

[10] D. Scaini and D. Arteaga, "Decoding of higher order ambisonics to irregular periphonic loudspeaker arrays," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*. Audio Engineering Society, 2014.

[11] J.-M. Batke and F. Keiler, "Using vbap-derived panning functions for 3d ambisonics decoding," in *2nd Ambisonics Symposium, Paris*, 2010.

[12] N. Epain, C. Jin, and F. Zotter, "Ambisonic decoding with constant angular spread," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 928–936, 2014.

[13] B. Wiggins, "The generation of panning laws for irregular speaker arrays using heuristic methods," in *Audio Engineering Society Conference: 31st International Conference: New Directions in High Resolution Audio*. Audio Engineering Society, 2007.

[14] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *J. Audio Eng. Soc*, vol. 60, no. 10, pp. 807–820, 2012. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16554

[15] J. Daniel, "Spatial sound encoding including near field effect," in *Proc. AES 23nd International Conference, Helsinger, Denmark*, 2003.

[16] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=7853

[17] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *Preprint 7358, AES 124th Convention, Amsterdam*, May 2008.

[18] S. Spors and J. Ahrens, "A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling," in *In Proc. AES 125th Convention*, October 2008.
[19] H. Wierstorf, A. Raake, and S. Spors, "Assessing localization accuracy in sound field synthesis a," *The Journal of the Acoustical Society of America*, vol. 141, no. 2, pp. 1111–1119, 2017.
[20] S. Spors and J. Ahrens, "Local sound field synthesis by virtual secondary sources," in *In Proc. AES 40th International Conference, Tokyo*, 2010.
[21] J. Ahrens and S. Spors, "Wave field synthesis of a sound field described by spherical harmonics expansion coefficients," *The Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2190–2199, 2012.
[22] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers *et al.*, "Spatial audio object coding (saoc)-the upcoming mpeg standard on parametric object based audio coding," in *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.
[23] R. H. Hardin and N. J. Sloane, "Mclaren's improved snub cube and other new spherical designs in three dimensions," *Discrete & Computational Geometry*, vol. 15, no. 4, pp. 429–441, 1996.
[24] R. A. Kennedy, P. Sadeghi, T. D. Abhayapala, and H. M. Jones, "Intrinsic limits of dimensionality and richness in random multipath fields," *IEEE Transactions on Signal processing*, vol. 55, no. 6, pp. 2542–2556, 2007.
[25] D. Menzies, "Nearfield binaural synthesis report," in *Proc. Acoustics08, Paris*, 2008.
[26] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.
[27] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.
[28] F. M. Fazi, M. Noisternig, and O. Warusfel, "Representation of sound fields for audio recording and reproduction," *Acoustics 2012 Nantes*, 2012.
[29] O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," *J. Audio Eng. Soc*, vol. 47, no. 7/8, pp. 583–595, 1999. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=12098
[30] M. Simon, D. Menzies, F. M. Fazi, T. de Campos, and A. Hilton, "A listener position adaptive stereo system for object based reproduction," in *Proc. AES 138th Convention, Warsaw*, May 2015.
[31] M. Kronlachner and F. Zotter, "Spatial transformations for the enhancement of ambisonic recordings," in *2nd International Conference on Spatial Audio, Erlangen*, 2014.
[32] R.-B. ITU-R, "1116-1, methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. 1997, intern," *Telecom Union: Geneva, Switzerland*, p. 26.
[33] S. Bech and N. Zacharov, "Perceptual audio evaluation-theory, method and application," 2002.

**Dylan Menzies** Dylan Menzies is a Senior Research Fellow in the Institute of Sound and Vibration, at the University of Southampton. Areas of interest include spatial audio synthesis and reproduction, sound synthesis for virtual environments, and musical synthesis and interfaces. He holds a PhD in Electronics from the University of York, an MA in Mathematics from Cambridge University, and has worked as a research engineer for several companies including Sony Professional Audio.

**Filippo Maria Fazi** Filippo Maria Fazi graduated in Mechanical Engineering from the University of Brescia (Italy) in 2005. He obtained his PhD in acoustics from the Institute of Sound and Vibration Research (ISVR) of the University of Southampton, UK, in 2010, with a thesis on sound field reproduction. In the same year, he was awarded a research fellowship by the Royal Academy of Engineering and by the Engineering and Physical Sciences Research Council. He is currently an Associate Professor at the University of Southampton. Dr Fazi s research interests include Audio technologies, Electroacoustics and Digital Signal Processing, with special focus on acoustical inverse problems, multi-channel systems, virtual acoustics, microphone and loudspeaker arrays. He is a member of the Audio Engineering Society and of the Institute of Acoustics.