

Real-time Microphone Array Processing for Sound-field Analysis and Perceptually Motivated Reproduction

Leo Thomas McCormack

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of
Science in Technology.

Espoo 16.10.2017

Thesis supervisor:

Prof. Ville Pulkki


Thesis advisor:

MSc. Symeon Delikaris-Manias

Author: Leo Thomas McCormack		
Title: Real-time Microphone Array Processing for Sound-field Analysis and Perceptually Motivated Reproduction		
Date: 16.10.2017	Language: English	Number of pages: 7+54
Department of Signal Processing and Acoustics		
Professorship: Acoustics and Audio Signal Processing (S-89)		
Supervisor: Prof. Ville Pulkki		
Advisor: MSc. Symeon Delikaris-Manias		
<p>This thesis details real-time implementations of sound-field analysis and perceptually motivated reproduction methods for visualisation and auralisation purposes. For the former, various methods for visualising the relative distribution of sound energy from one point in space are investigated and contrasted; including a novel reformulation of the cross-pattern coherence (CroPaC) algorithm, which integrates a new side-lobe suppression technique. Whereas for auralisation applications, listening tests were conducted to compare ambisonics reproduction with a novel headphone formulation of the directional audio coding (DirAC) method. The results indicate that the side-lobe suppressed CroPaC method offers greater spatial selectivity in reverberant conditions compared with other popular approaches, and that the new DirAC formulation yields higher perceived spatial accuracy when compared to the ambisonics method.</p>		
Keywords: spatial audio, microphone arrays, sound-field analysis, sound-field reproduction, cross-pattern coherence, directional audio coding, time-frequency domain processing		

Preface

This work was carried out at the Department of Signal Processing and Acoustics at Aalto University in Finland.

Firstly, I would like to thank my supervisor Prof. Ville Pulkki  for his continuous support throughout this thesis work and for also providing me with an adjustable electronic desk to work on. I would also like to extend special thanks to Mr. Symeon Delikaris-Manias, for his limitless supply of knowledge and expertise regarding spherical harmonic domain processing in general; and also for his support during the reformulation of the original cross-pattern coherence algorithm detailed in this thesis. Special thanks is also extended to Dr. Archontis Politis, for helping me to understand the inner-intricacies of the latest directional audio coding reproduction methods; and also for his support during the formulation of the new headphone version detailed in this thesis.

I am also eternally grateful for the support and motivation I received from my wonderful girlfriend, Janani Fernandez, and also the Aalto Acoustics Lab researchers, my family, friends and the beverage known as beer; without which, this thesis would have been much more onerous.

Otaniemi, Espoo, Finland, 16.10.2017

Leo Thomas McCormack

Contents

Abstract	ii
Preface	iii
Contents	iv
Symbols and abbreviations	v
1 Introduction	1
1.1 Aims of the thesis	5
1.2 Organisation of the thesis	6
2 Microphone array processing in the time-frequency domain	7
2.1 Time-frequency transform	7
2.2 Spatial encoding	10
2.3 Static beamformers	11
2.4 Adaptive beamformers	13
3 Sound-field analysis and parameter estimation	14
3.1 Scanning beamformers	14
3.2 Cross-spectrum-based parameter estimation	16
3.2.1 Side-lobe suppression	18
3.2.2 Weighted-orthogonal beamforming-based formulation	20
3.3 Direction-of-arrival and diffuseness parameter estimation	22
4 Perceptually motivated sound-field reproduction	24
4.1 Spatial hearing	24
4.2 Ambisonics reproduction	27
4.3 Directional audio coding reproduction	28
4.3.1 Parametric analysis	29
4.3.2 Legacy first-order synthesis	29
4.3.3 Higher-order synthesis for headphones with optimal mixing	31
5 Real-time implementations	35
5.1 Mic2SH	37
5.2 AcCroPaC	38
5.2.1 Examples	40
5.3 OM-DirAC	43
5.3.1 Evaluation	43
6 Conclusion	47
References	48

Symbols and abbreviations

Symbols and operators

$\mathcal{A}[\cdot]$	parametric analysis operation on the enclosed signal
\mathbf{A}	optimal mixing matrix
b_n	modal coefficient of order n
\mathbf{B}	optimal mixing residual matrix
c	speed of sound
\mathbf{C}	covariance matrix
\mathbf{C}^w	covariance matrix with diagonal loading
$\mathcal{D}[\cdot]$	decorrelation operation on the enclosed signal
\mathbf{D}	ambisonic decoding matrix
e	energy density
$\mathbb{E}[\cdot]$	expectation operator
f	discrete frequency index
F_s	sampling frequency
\mathbf{G}	matrix of CroPaC pseudo-spectrum values / VBAP gains (contextual)
h_a^f	analysis mirror filter
h_s^f	synthesis mirror filter
$h_n^{(2)}$	spherical Hankel function of the second kind and order n
\mathbf{h}	vector of HRTF filters
\mathbf{H}	matrix of HRTF filters
\mathbf{i}_a	active intensity vector
\mathbf{I}	identify matrix
j	imaginary number
j_n	spherical Bessel function of the first kind and order n
k	wave number
K	number of frequency bands
L	number of loudspeakers
m	spherical harmonic degree
\mathbf{M}	mixing or rotation matrix
n	time-domain sample index / spherical harmonic order (contextual)
N	number of time-domain samples / spherical harmonic transform order (contextual)
p	sound pressure
\mathbf{p}	parameter vector
P_{nm}	orthonormal Legendre polynomials of order n and degree m
\mathbf{P}	matrix of power-map values
Q	number of microphones
S	number of sectors
r	radius
\mathbf{s}	vector of spherical harmonic signals
\mathbf{S}	matrix of pseudo-spectrum values
t	discrete time index
\mathbf{u}	particle velocity
\mathbf{U}	diffuse energy distribution matrix

w_a	analysis windowing function
w_s	synthesis windowing function
\mathbf{w}	vector of beamforming weights
\mathbf{W}	spatial encoding matrix
\mathbf{W}_n	equalisation matrix for signals of order n
x	time domain signal
\hat{x}	time-frequency domain signal
\mathbf{x}	vector of microphone signals
Y_{nm}	spherical harmonic of order n and degree m
\mathbf{y}	vector of output signals / spherical harmonic steering vector (contextual)
\mathbf{Y}	matrix of spherical harmonics
\mathbf{z}	DirAC audio stream
δ_{m0}	Kronecker delta
θ	elevation angle
ϕ	azimuthal angle
ψ	diffuseness
$\Phi_{\mathbf{SxSy}}$	cross-spectrum between signals \mathbf{x} and \mathbf{y}
λ	regularisation parameter
ρ_0	density of the medium
Σ	diagonal matrix of singular values

Abbreviations

AcCroPaC	acoustic camera software that utilises CroPaC
ALLRAD	all-round ambisonic decoding
ASW	apparent source width
BLAS	Basic Linear Algebra Library
CroPaC	cross-pattern coherence
DaS	delay-and-sum
DAW	digital audio workstation
DC	direct current
DFT	discrete Fourier transform
DirAC	directional audio coding
DoA	direction-of-arrival
FFT	fast Fourier transform
FIR	finite impulse response
FOV	field-of-view
HO-DirAC	higher-order directional audio coding
HRTF	head-related transfer function
IC	interaural coherence
ICC	inter-channel coherence
ICLD	inter-channel level difference
ICPD	inter-channel phase difference
ILD	inter-aural level difference
ITD	inter-aural time difference
IR	impulse response
LAPACK	Linear Algebra Package
LCMV	linearly-constrained minimum-variance
LEV	listener envelopment
Mic2SH	microphone signals to spherical harmonic signals conversion software
MSVC	Microsoft Visual Compiler
MKL	Mathematics Kernel Library
MUSIC	multiple signal classification
MVDR	minimum-variance distortion-less response
OM-DirAC	optimal-mixing directional audio coding
PQMF	pseudo-quadrature mirror filterbank
PWD	plane-wave decomposition
QMF	quadrature mirror filterbank
SHT	spherical harmonic transform
SMA	spherical microphone array
SNR	signal-to-noise ratio
STFT	short-time Fourier transform
SVD	singular-value decomposition
WOB-CroPaC	weighted-orthogonal beamforming-based CroPaC
VBAP	vector-base amplitude panning
VL-DirAC	virtual-loudspeaker directional audio coding
VST	Virtual Studio Technology

1 Introduction

The ability to capture and extract meaningful real-time information from a spatial sound scene, offers great potential for a variety of different applications; ranging from equipment diagnostics, construction and military uses, through to various entertainment platforms. In architectural acoustics, the ability to track the propagation of reflections or to visualise the distribution of sound energy in a room can be very helpful, for example, when optimising the placement of diffusers and sound insulating materials (O'Donovan et al., 2008; Farina et al., 2011). Similarly, faulty mechanical and electrical equipment components can often exhibit higher sound energy emissions; thus, determining the location of this higher energy can be helpful, especially in cases where the equipment is difficult to access. In the military sector, determining the direction of sound sources within a sound scene and presenting this information to the navigators (especially in low visibility conditions) can be of vital importance (Nielsen, 1991). On the other hand, the various entertainment industries go to great lengths to create immersive and authentic audio-visual content that can, for example, be experienced via the emerging virtual and augmented reality headsets. While these applications vary immensely in terms of their target audiences, all of the systems described above may be derived from a common foundation; which comprises of microphone array processing techniques, as this thesis will demonstrate.

A single omni-directional microphone placed within a sound scene is the starting point. In the presence of an ongoing sound source, the amplitude of the captured microphone signal will vary with time, and as such, the sound energy arriving at this point in space can be monitored. However, while this single microphone will aid the user in identifying the presence of a sound source, ascertaining certain spatial parameters from the perspective of the listening position, such as the direction of the sound source, is not possible without prior in-depth knowledge of the surrounding area. Introducing a second microphone, however, will allow a system to determine the direction of a sound source along a one-dimensional plane, by simply observing the phase differences between the two microphone signals; assuming that the distance between the sensors is known. This principle can be extended to two-dimensional (2D) or three-dimensional (3D) direction of arrival (DoA) estimation, by utilising three or four microphones, respectively; where the most typical orientation for the latter is in a tetrahedral fashion. The most popular four-capsule 3D sound-field microphone is the B-Format microphone (Gerzon, 1973).

While increasing the number of microphones beyond four will not allow a user to monitor sound sources in other dimensions, it will, however, permit increased spatial resolution when generating spatial filters; which are also referred to as *beamformers*. These spatially selective beamforming algorithms form a basis for many sound-field analysers and reproduction systems. One popular beamforming method is to simply delay-and-sum (DaS) the microphone signals in such a manner as to allow constructive interference to occur for a single target direction. Introducing more microphones into the array will further improve the spatial selectivity of the DaS beamformer, and also increase the signal-to-noise (SNR) ratio; which is especially important for submarine sonar systems, where sound sources may be located several kilometres away. However,

while the DaS beamformer will perform adequately for high-frequencies, low-frequency selectivity is generally poor when utilising compact microphone arrays, as the larger wavelengths will result in smaller differences in signal phase between the microphones.

Regarding the arrangement of these microphones, 2D arrays will generally orientate the microphone sensors uniformly (or nearly-uniformly) in a linear, circular or rectangular fashion, while 3D arrays will arrange the microphones around a cylinder or sphere. The advantage of a uniform sensor distribution is that beamformers will provide similar spatial resolution for all directions covered by the array. Whereas, the incentives for a cylindrical or spherical microphone distribution, is that they allow for both convenient and efficient methods of decomposing the sound-field at a listening point into *spherical harmonics* (Rafaely, 2015). The efficiency of this transform lies with the fewer microphone sensors that are required to obtain the spherical harmonic signals, when compared to arrays that are not arranged in a spherical or cylindrical manner; while the convenience aspect is due to the fact that this transform is well founded in theory and translates well in practice. The advantage of performing this additional spherical harmonic transform and generating beamformers using these spherical harmonic signals, rather than relying on the microphone signals directly, is that the microphone array specifications are not required to steer the beamformers to the target *look direction*. In other words, the beamforming algorithms do not require microphone array impulse responses (IR) for each possible steering direction.

It should be stressed at this point that once a system is capable of accessing these spherical harmonic signals, and can subsequently generate beamformers, the system can be orientated towards any of the aforementioned applications with reasonable results. For example, in applications which require a user to effectively *visualise* the relative sound energy distribution in a sound-field, multiple beamformers can be first generated and steered towards points on a pre-defined grid. The relative energy of these beamformed signals can then be displayed via a *power-map*; where, for example, the colour red may indicate high-energy regions, and the colour blue may indicate low-energy regions. If this power-map is subsequently overlayed onto a corresponding video of the same sound scene, then the system may be referred to as an *acoustic camera*. These systems may be used for source tracking in passive sonar systems, identifying problem areas in room acoustics, or used to locate faults in mechanical and electrical equipment which display higher sound energy. On the other hand, to accommodate applications which require a user to *auralise* a captured sound scene, beamformers may be generated for a pre-defined grid of loudspeaker positions and sent to their respective loudspeakers or, alternatively, convolved with their corresponding head-related transfer functions (HFRTs) and experienced over headphones (Wiggins et al., 2001; Noisternig et al., 2003; Davis et al., 2005; Melchior et al., 2009; Atkins, 2011; Shabtai and Rafaely, 2013; Bernschütz et al., 2014). Note that when operating in the spherical harmonic domain, this linear mapping of spherical harmonic signals to loudspeakers is more commonly referred to as ambisonic decoding (Gerzon, 1973), which is a popular non-parametric reproduction method discussed later in this thesis.

Regarding beamforming algorithms formulated in the spherical harmonic domain, perhaps the most common signal-independent approach, utilised for both the auralisa-

tion and visualisation of spatial sound-fields, is the plane wave decomposition (PWD) algorithm. As the name would suggest, this beamformer relies on the assumption that the waves emitted by sound sources present in the recorded sound scene, are received as plane-waves at the listening position; therefore, this algorithm is said to be suited only for capturing *far-field* sound sources (Erić, 2011). The beam patterns generated by the PWD algorithm can be further manipulated by utilising in-phase (Daniel, 2000), Dolph-Chebyshev (Rafaely, 2015) or maximum energy (Zotter et al., 2012) weightings, to suit better the specific application.

However, while the PWD beamformer can provide reasonable results for many applications, it often yields less than ideal results for many of the applications described above, mainly as a result of its unwanted side-lobes and large beam width (especially when utilising first-order or lower-order spherical harmonic signals). For example, in the case of sound-field analysing systems (such as acoustic cameras), this large beam-width can result in sound sources appearing spatially larger than they actually are, and the unwanted side-lobes may cause a user to erroneously identify non-existent sound sources/reflections in the power-maps. In the case of non-parametric sound-field reproduction methods (such as ambisonics), these problems may often result in directional blurring of point sources; localisation ambiguity; loss of externalisation; reduced sense of envelopment in reverberant conditions and strong colouration effects (Solvang, 2008; Santala et al., 2009; Braun and Frank, 2011; Kearney et al., 2012; Bertet et al., 2013; Bernschütz et al., 2014; Stitt et al., 2014; Yang and Bosun, 2014).

Sound-field analysis

In order to surpass the performance offered by the PWD algorithm for sound-field analysis and visualisation purposes, signal-dependent solutions can be employed; with the penalty of increased computational complexity. One common solution, is the minimum-variance distortion-less response (MVDR) algorithm (Capon, 1969), which operates in the time-frequency domain and takes into account the inter-channel dependencies between the microphone array signals per frequency band. It then utilises this information to enhance the performance of the beamformer, by adaptively placing nulls to the interferers. However, while this algorithm works well in free-field conditions, the algorithm is relatively sensitive in scenarios where high background noise and/or reverberation are present in the sound scene; resulting in sub-optimal results for many real-world situations (Zoltowski, 1988).

An alternative approach is to utilise subspace methods, such as the multiple signal classification (MUSIC) algorithm (Schmidt, 1986), which performs an eigenvalue decomposition on the covariance matrix or a singular value decomposition (SVD) operation on the data matrix of the microphone signals and subsequently extracts a *pseudo-spectrum* from the resulting signal and noise subspaces. The algorithm can be orientated as a multiple-speaker localisation method by incorporating a statistical direct-path dominance test, as described in (Nadiri and Rafaely, 2014). Another approach, proposed in (Epain and Jin, 2013), is to employ pre-processing, in order to separate the direct components from the diffuse field. This subspace-based

separation has been shown to improve the performance of existing super-resolution imaging algorithms in (Noohi et al., 2013). However, while these methods are generally more robust than their beamforming counterparts, they are still susceptible to reverberation and/or background noise to some extent. This is due to their reliance on the inter-channel dependencies of the signals and the difficulties that arise when attempting to determine the size of the sub-matrices which define the noise and signal subspaces. Therefore, a certain degree of bias may be expected when these algorithms are subjected to sound-fields that deviate from the initial assumptions. Furthermore, the increased implementation and computational complexity demanded by these algorithms, may render them unsuitable for some systems.

On the other hand, one approach that has been shown to be robust to interferers/reverberation and has reduced computational requirements, is the cross-pattern coherence (CroPaC) spatial filtering technique described in (Delikaris-Manias and Pulkki, 2013). It operates by measuring the correlation between coincident beamformers and subsequently deriving a post-filter that aims to suppress noise, interferers and reverberation. While originally intended for enhancing the spatial selectivity of other beamformers, it is possible to utilise multiple post-filters essentially as statistical likelihood parameters and to subsequently derive pseudo-spectrums. Additionally, the algorithm has recently been extended via a linearly-constrained minimum variance (LCMV) inspired solution, such that it can utilise arbitrary combinations of beamformers in (Delikaris-Manias, Pavlidi, Pulkki and Mouchtaris, 2016) which is defined here as *weighted-orthogonal beamforming-based* CroPaC (WOB-CroPaC).

However, the main limitation with the original CroPaC algorithm becomes apparent when utilising higher-order spherical harmonic signals, as unwanted side-lobes are generated by the cross-spectrum operation; producing similar problems to those attributed to the standard PWD beamformer. Therefore, a reformulation of the original CroPaC algorithm which applies the product of multiple rotated CroPaC beams to suppress the side-lobes, could be expected to yield improved performance when utilising these higher-order spherical harmonic signals. Furthermore, the new WOB-CroPaC algorithm presented in (Delikaris-Manias, Pavlidi, Pulkki and Mouchtaris, 2016), currently uses both the spherical harmonic signals and the microphone signals themselves; resulting in a large computational overhead, as significantly more signals are required to be transformed into the time-frequency domain. A reformulation of this algorithm, such that it operates solely on spherical harmonic signals, should be expected to yield lower computational requirements.

Perceptually motivated sound-field reproduction

To improve upon the poor spatial accuracy offered by the ambisonics method, parametrically enhanced sound-field reproduction techniques, such as directional audio coding (DirAC) (Pulkki, 2006, 2007), may be utilised. The DirAC approach operates on the basis of extracting a DoA parameter and a diffuseness parameter at each time-frequency index. These parameters can be derived from the energetic quantities of the active intensity vector and diffuseness estimate, which DirAC interprets as narrowband perceived DoA and interaural coherence (IC) cues. Another

parametric approach is the HARPEX method (Barrett and Berge, 2010), based on first order ambisonics and binaural rendering, with a dual plane-wave model and no diffuse component.

While originally defined for first order spherical harmonic signals, DirAC has recently been extended to accommodate higher-order signals (HO-DirAC) (Pulkki et al., 2013; Politis, Vilkamo and Pulkki, 2015). The parametric model applied within DirAC has been shown to be effective in a number of listening tests (Vilkamo et al., 2009; Laitinen and Pulkki, 2009; Politis, Laitinen, Ahonen and Pulkki, 2015; Politis, Vilkamo and Pulkki, 2015), mitigating the perceptual deficiencies present in lower-order ambisonic decoded audio, with the compromise of increased computational cost and implementation complexity. DirAC has also been utilised for headphone reproduction in (Laitinen and Pulkki, 2009), by employing a virtual loudspeaker approach (VL-DirAC). However, while this method has been effectively demonstrated in real-time with head-tracking support, it is computationally demanding and prone to certain artefacts in scenarios which are challenging for the parametric analysis.

However, improvements in computational efficiency over the virtual loudspeaker approach, should be expected by utilising a binaural ambisonic decoder. Whereas, the artefacts could be mitigated by utilising the optimal adaptive mixing solution presented by (Vilkamo and Pulkki, 2013), in a similar manner to the loudspeaker implementation described in (Politis, Vilkamo and Pulkki, 2015) or (Vilkamo and Delikaris-Manias, 2015), which can synthesise the binaural cues directly from the spatial parameters.

1.1 Aims of the thesis

With a focus on real-time implementations, the main aim of this thesis was to first investigate methods for sound-field analysis and perceptually motivated reproduction. Secondly, to derive new methods that could potentially yield higher spatial accuracy in their respective target applications; while also remaining computationally feasible given realistic hardware constraints.

The main contributions of this thesis work with respect to sound-field analysis methods, may be summarised as:

- The development of a real-time acoustic camera framework, for which various different algorithms can be implemented and contrasted. The software utilises a commercially available spherical microphone array and a spherical camera, to produce the power-map overlay and video stream, respectively.
- Among the various power-map/pseudo-spectrum methods that were implemented into the real-time software are the: PWD, MVDR, MUSIC, and original CroPaC algorithms. In order to assess the relative performance of these algorithms for real-world situations, power-maps were generated using reverberant test sound-fields and compared.
- An investigation into the feasibility of utilising the proposed side-lobe cancellation technique with the original CroPaC algorithm (Delikaris-Manias and Pulkki, 2013).

- A minor reformulation of the WOB-CroPaC algorithm (Delikaris-Manias, Pavlidi, Pulkki and Mouchtaris, 2016) such that it may yield lower computational requirements, was also investigated.

The main contributions of this thesis work regarding perceptually motivated sound-field reproduction techniques, can be summarised as:

- The development of a novel real-time headphone DirAC implementation, utilising a new architecture that overcomes the drawbacks of the VL-DirAC approach. The proposed formulation incorporates support for higher order spherical harmonic signals and the optimal adaptive mixing solution of (Vilkamo and Pulkki, 2013), in order to synthesise the binaural cues directly from the spatial parameters.
- Listening test results comparing the new DirAC formulation with the ambisonics method are also presented and discussed.

1.2 Organisation of the thesis

This thesis is organised as follows: Section 2 provides details on how to transform microphone array signals into the time-frequency and spherical harmonic domains, and also how to generate both static and adaptive beamformers; Section 3 presents various means of extracting spatial parameters and generating power-maps/pseudo-spectrums of sound-fields, including a novel reformulation of the original CroPaC algorithm; Section 4 provides an overview of the principles of spatial hearing and also describes some perceptually-motivated means of reproducing sound-fields; including an efficient and robust reformulation of the DirAC algorithm for headphone playback; Section 5 provides details of the real-time implementations developed during the thesis for spatial encoding, sound-field imaging, and sound-field reproduction; and Section 6 concludes the thesis.

2 Microphone array processing in the time-frequency domain

This thesis is concerned with utilising microphone arrays to analyse a spatial sound-field and subsequently visualise or auralise it with high spatial accuracy. Therefore, this chapter provides the required background, regarding how spherical microphone array (SMA) signals may be transformed into the time-frequency and spherical-harmonic domains and how to generate spatially selective filters. This provides the foundation for the implementations which will be described later in the thesis.

Note that vectors, \mathbf{v} , have been given in bold font and lower-case letters; whereas, matrices, \mathbf{M} , are given in bold font and upper-case letters.

2.1 Time-frequency transform

A time-frequency transform is a means of dividing a time-domain signal into individual sub-bands via the discrete Fourier transform (DFT) or a bank of filters. Typically, these transforms are reversible and are categorised as either *perfect reconstruction* or *near-perfect reconstruction* transforms. The former indicates that there is no degradation in quality between the input signal and the output signal (provided that the signals were not altered whilst in the time-frequency domain). In the case of near-perfect reconstruction filterbanks, certain distortion artefacts are introduced during the transform; however, this loss in quality is generally perceptually negligible in practice.

Depending on the application, filterbanks may be *critically sampled* or *oversampled*. In the case of critically sampled filterbanks, input signals of sampling rate F_s are decomposed into K sub-bands of sampling rates F_s/K , which make these signals suitable for audio coding applications as redundancy is minimised. Oversampled filterbanks, on the other-hand, utilise overlapping windowing functions in order to make the transform more robust to signal manipulations.

Short-time Fourier transform (STFT)

The short-time Fourier transform (STFT) is perhaps the most popular means of transforming a time-domain sequence $x(n)$ of N samples, into its time-frequency domain counterpart $\hat{x}(t, f)$, where t refers to the down-sampled time index and f refers to the frequency index. The transform may be configured as either critically sampled or oversampled, depending on the chosen windowing function. The forward STFT transform is given as (Smith, 2011)

$$\hat{x}(t, f) = \sum_{i=0}^{N/K-1} \sum_{n=0}^{N-1} w_a(n) x(n + iK + tN) e^{-j2\pi fn/N}, \text{ for } t = 0, \dots, \infty \quad (1)$$

where $j^2 = -1$, K refers to the hop size in time-domain samples, and w_a is the analysis windowing function. The inverse STFT transform is expressed as (Smith,

2011)

$$y(n) = \sum_{i=0}^{N/K-1} \frac{1}{N} w_s(n) \sum_{f=0}^{N-1} \hat{x}(t + i + n[\frac{N}{K}], f) e^{j2\pi f n/N}, \text{ for } n = 0, \dots, \infty \quad (2)$$

where w_s is the synthesis windowing function, which suppresses aliasing artefacts which arise due to non-zero edges in the output frame that may have been introduced during manipulation of the STFT data. However, spectral processing algorithms should only manipulate the positive, DC and Nyquist frequency indices, as these are conjugate copied to the corresponding negative frequency indices before applying the inverse-DFT.

Note, however, that the fast Fourier transform (FFT) should be utilised for practical implementations, due to the vast performance gains incurred by the increase in computational efficiency. Furthermore, in order to attain the perfect reconstruction characteristic, the windowing functions w_a and w_s should conform to the condition laid down by Princen-Bradley, which ensures amplitude conservation. For example, when utilising a hop size $K = N/2$ (which would constitute an oversampling factor of 2) the windowing functions should satisfy

$$w_a(n)w_s(n) + w_a(n + N/2)w_s(n + N/2) = 1, \text{ for } n = 0, \dots, (N/2 - 1). \quad (3)$$

However, due to this windowing and/or zero padding prior to the FFT, the output is susceptible to both temporal and spectral aliasing if it is modified in the time-frequency domain. These aliasing components are typically suppressed via the application of a synthesis windowing function. However, one alternative is the alias-free STFT used in (Vilkamo and Pulkki, 2013), which has been designed to suppress these aliasing effects without the need for a synthesis windowing function and has been shown to be more perceptually robust than conventional STFT implementations.

This formulation avoids circular convolution effects (which lead to aliasing artefacts) by zero-padding the signal frame and pre-processing the complex magnitude and phase operators in (1) and (2), so that the non-zero section of the input signal frame is limited in length. This pre-processing consists of first applying an inverse-FFT, then windowing and zero padding, and then applying an FFT before applying the complex multipliers for the STFT transform. Providing that the combined length of the non-zero parts are the same length or shorter than the zero-padded sections, the circular convolution effects will be avoided; albeit, with the penalty of increased computational complexity.

Quadrature mirror filterbanks

For applications that do not require the manipulation of the intermediate data, the pseudo-quadrature mirror filterbank (PQMF) is a popular time-frequency transform, which has found its way into various audio codecs due to its critically sampled characteristic. It is implemented by first designing a low-pass prototype FIR filter, $h_{\text{lpf}}(n)$, of order N and then modulating it with cosine sequences to obtain analysis,

$h_a^f(n)$, and synthesis, $h_s^f(n)$, band-pass filters. The analysis filters for K sub-bands are defined as (Nguyen, 1994)

$$h_a^f(n) = h_{\text{lpf}}(n) \cos\left[\frac{\pi}{2K}(2f+1)\left(n - \frac{N}{2} - \frac{K}{2}\right)\right], \text{ for } f = 0, \dots, K-1, \quad (4)$$

which are then convolved with the time domain input frame to attain the sub-bands, which are subsequently down-sampled by a factor of K in order to reduce redundancy. After the time-frequency domain processing/analysis has been performed, the sub-bands are then up-sampled by a factor of K , before applying the synthesis filters, which are defined as (Nguyen, 1994)

$$h_s^f(n) = h_{\text{lpf}}(n) \cos\left[\frac{\pi}{2K}(2f+1)\left(n - \frac{N}{2} + \frac{K}{2}\right)\right], \text{ for } f = 0, \dots, K-1. \quad (5)$$

The prototype filter should be designed such that the energy is preserved in the inter-sections between adjacent sub-bands and that the non-adjacent sub-bands are sufficiently suppressed. Due to the down-sampling and up-sampling operations, spectral aliasing artefacts in adjacent bands are introduced in the analysis stage and then subsequently cancelled out completely in the synthesis stage; providing that no manipulations to the intermediate data are made. If manipulations are made, however, then the aliasing components will remain; although, typically they are sufficiently suppressed so that they are not perceivable by the listener, which is why the PQMF is referred to as a near-perfect reconstruction filterbank. However, this property of relying on adjacent sub-bands to cancel out aliasing components is not ideal. Therefore, to accommodate robust manipulations of the time-frequency domain signals, the complex valued quadrature mirror filterbank (QMF) is more preferable, as this reliance on adjacent bands is avoided.

The complex QMF allows each of the sub-bands to be altered independently from one another without the introduction of spectral aliasing by utilising complex modulators (Vaidyanathan and Hoang, 1988). The complex analysis filters for K sub-bands are defined as (Nguyen and Vaidyanathan, 1989)

$$h_a^f(n) = h_{\text{lpf}}(n) \exp\left[j\frac{\pi}{2K}(2f+1)\left(n - \frac{N}{2} - \frac{K}{2}\right)\right], \text{ for } f = 0, \dots, K-1, \quad (6)$$

which are applied to the time domain input frame and subsequently down-sampled by a factor of K in a similar manner to the PQMF; however, due to the complex modulation, this results in an oversampling by a factor of 2. The complex synthesis filters are defined as (Nguyen and Vaidyanathan, 1989)

$$h_s^f(n) = h_{\text{lpf}}(n) \exp\left[j\frac{\pi}{2K}(2f+1)\left(n - \frac{N}{2} + \frac{K}{2}\right)\right], \text{ for } f = 0, \dots, K-1. \quad (7)$$

These filters are favoured for perceptually-motivated spatial audio codecs, for example, which process the audio in the decoding stage to synthesise the original multi-channel audio from a down-mixed audio stream and the inter-channel level, phase, and coherence differences that were encoded into the bit-stream by the encoder.

2.2 Spatial encoding

Spatial encoding is the means by which microphone signals are decomposed into spatially selective signals, which sample the sound-field. In the case of spherical microphone arrays (SMA), these spatially selective components are typically spherical harmonic signals, which form an orthonormal basis over the unit sphere and are obtained via a spherical harmonic transform (SHT) or a spherical Fourier transform. The accuracy of this decomposition is dependent on: the arrangement of the sensors on the sphere, the radius of the sphere and the array type (Rafaely, 2015). The total number of sensors will dictate the highest order N of spherical harmonic signals that can be estimated.

In order to decompose the input signals \mathbf{x} from a SMA of Q sensors into an estimate of the spherical harmonic signals for each frequency, a frequency-dependent spatial encoding matrix $\mathbf{W} \in \mathbb{C}^{(N+1)^2 \times Q}$ can be applied

$$\mathbf{s}(t, f) = \mathbf{W}(f)\mathbf{x}(t, f), \quad (8)$$

where

$$\mathbf{x}(t, f) = [x_1(t, f), x_2(t, f), \dots, x_Q(t, f)]^T \in \mathbb{C}^{Q \times 1}, \quad (9)$$

are the microphone signals and

$$\mathbf{s}(t, f) = [s_{00}(t, f), s_{1-1}(t, f), s_{10}(t, f), \dots, s_{NN-1}(t, f), s_{NN}(t, f)]^T \in \mathbb{C}^{(N+1)^2 \times 1}, \quad (10)$$

are the spherical harmonic signals.

The frequency-dependent spatial encoding matrix is calculated as

$$\mathbf{W}(f) = \alpha_q \mathbf{W}_n(f) \mathbf{Y}^\dagger \quad (11)$$

where $\mathbf{W}_n \in \mathbb{C}^{(N+1)^2 \times (N+1)^2}$ is an equalisation matrix which attempts to mitigate the effects induced by the spherical sensor arrangement; $\mathbf{Y} \in \mathbb{R}^{Q \times (N+1)^2}$ is a matrix of spherical harmonics for the sensor positions; and $\alpha_q = 4\pi/Q$ is (in this case) the sampling weight for a uniform or nearly-uniform sensor distribution on the sphere (Rafaely, 2015). For arrangements that are not uniform, α_q should be replaced by a diagonal matrix and may be optionally incorporated into matrix \mathbf{Y} . The equalisation matrix is defined as

$$\mathbf{W}_n(f) = \begin{bmatrix} w_0(f) & & & & & \\ & w_1(f) & & & & \\ & & w_1(f) & & & \\ & & & w_1(f) & & \\ & & & & \ddots & \\ & & & & & w_N(f) \end{bmatrix}, \quad (12)$$

where

$$w_n(f) = \frac{1}{b_n(f)} \frac{|b_n(f)|^2}{|b_n(f)|^2 + \lambda^2}, \quad (13)$$

where $b_n(f)$ are frequency-dependent and order-dependent modal coefficients, which take into account certain characteristics of the SMA; including whether the construction is open or rigid and the directivity of the sensors (omnidirectional or directional); and λ is a regularisation parameter that influences the sensor noise amplification. For more details on calculating the equalisation matrix \mathbf{W}_n , the reader is referred to (Alon et al., 2015; Lösler and Zotter, 2015), or for a signal-dependent encoder (Schörkhuber et al., 2017). Additionally, for microphone arrays that incorporate sensors of differing directional characteristics, a more general formulation of (11) is detailed in (Rafaely, 2015).

The spherical harmonics \mathbf{Y} are given here in matrix form

$$\mathbf{Y} = \begin{bmatrix} Y_{00}(\Omega_1) & Y_{00}(\Omega_2) & \dots & Y_{00}(\Omega_Q) \\ Y_{-11}(\Omega_1) & Y_{-11}(\Omega_2) & \dots & Y_{-11}(\Omega_Q) \\ Y_{10}(\Omega_1) & Y_{10}(\Omega_2) & \dots & Y_{10}(\Omega_Q) \\ Y_{11}(\Omega_1) & Y_{11}(\Omega_2) & \dots & Y_{11}(\Omega_Q) \\ \vdots & \vdots & \vdots & \vdots \\ Y_{NN}(\Omega_1) & Y_{NN}(\Omega_2) & \dots & Y_{NN}(\Omega_Q) \end{bmatrix}^T, \quad (14)$$

where $\Omega_q = (\theta_q, \phi_q, r)$ are the locations of each of the sensors, where $\theta_q \in [-\pi/2, \pi/2]$ denotes the elevation angle, $\phi_q \in [-\pi, \pi]$ the azimuthal angle and r the radius of the SMA. Note, however, that due to the difficulties that arise when utilising complex-valued spherical harmonics for real-time applications (such as increased computational complexity and non-trivial sound field rotation); this thesis work has been developed using real-valued spherical harmonics of order $n \geq 0$ and degree $m \in [-n, n]$, which are calculated as

$$Y_{nm}(\theta, \phi) = \sqrt{(2 - \delta_{m0}) \frac{(2n+1)(n-|m|)!}{4\pi(n+|m|)!}} P_{n|m|}(\sin\theta) v_m(\phi), \quad (15)$$

with

$$v_m(\phi) = \begin{cases} \sin |m|\phi, & m < 0 \\ 1, & m = 0 \\ \cos m\phi, & m > 0, \end{cases} \quad (16)$$

where δ_{m0} is the Kronecker delta, $P_{n|m|}$ are the orthonormal Legendre polynomials and $!$ denotes the factorial operator. Real spherical harmonics up to fourth order are depicted in Fig. 1. Note that arbitrary rotations of real-valued spherical harmonics can be performed using a single matrix multiplication, detailed in (Blanco et al., 1997), or recursively, as described in (Ivanic and Ruedenberg, 1996, 1998). Whereas rotations of complex spherical harmonics may be performed using the Wigner-D weighting (Rafaely and Kleider, 2008) or by utilising projection methods (Atkins, 2011).

2.3 Static beamformers

Once the SMA signals have been transformed into spherical harmonic signals, signal-independent spatial filters (or beamformers) can be generated without the need for

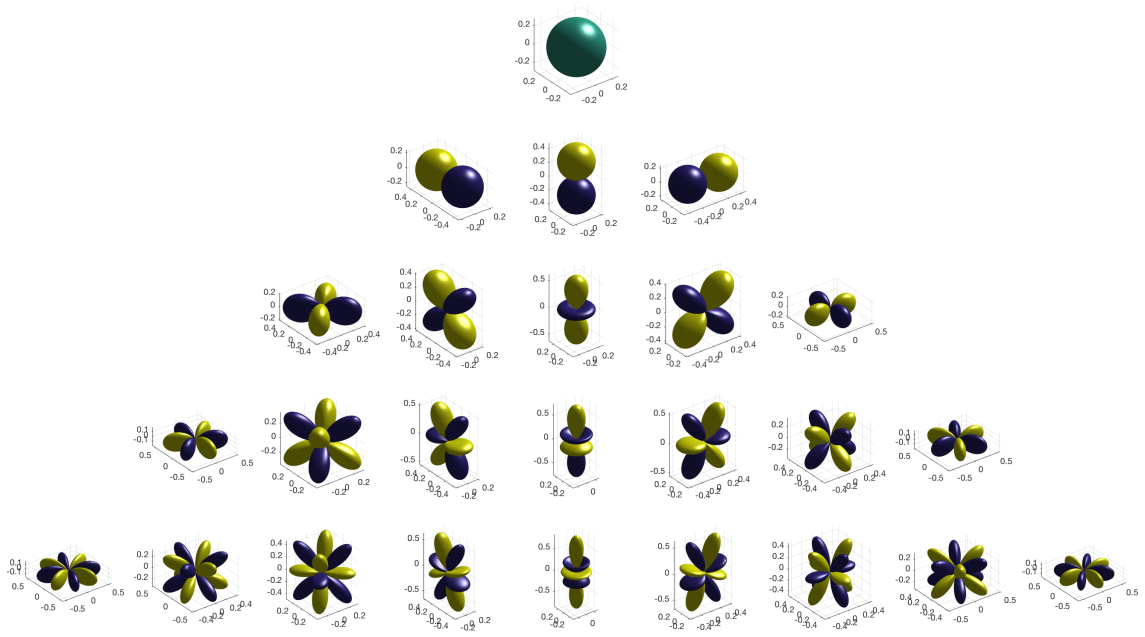


Figure 1: Real spherical harmonics from zeroth to fourth order (top to bottom). Positive and negative phase lobes are depicted with yellow/green and blue colours, respectively.

measuring the impulse response of the array. Beamforming in the spherical harmonic domain can be performed as

$$y(t, f) = \mathbf{w}^H \mathbf{s}(t, f), \quad (17)$$

where $\mathbf{w} \in \mathbb{C}^{(N+1)^2 \times 1}$ is (in this case) a frequency-independent steering vector, which is given as

$$\mathbf{w} = \mathbf{y}(\Omega) \odot \mathbf{d}, \quad (18)$$

where $\mathbf{y}(\Omega) \in \mathbb{C}^{1 \times (N+1)^2}$ are the spherical harmonic weights for any arbitrary direction $\Omega = (\theta, \phi)$, where $\theta \in [-\pi/2, \pi/2]$ denotes the elevation angle and $\phi \in [-\pi, \pi]$ the azimuthal angle; \odot denotes the Hadamard product; and \mathbf{d} is a vector of weights defined as

$$\mathbf{d} = [d_0, d_1, d_1, d_1, \dots, d_N]^T \in \mathbb{R}^{(N+1)^2 \times 1}. \quad (19)$$

The weights \mathbf{d} can be adjusted to generate a range of different types of axis-symmetric beamformers: regular (Rafaely, 2015), in-phase (Daniel, 2000), maximum energy (Zotter et al., 2012; Moreau et al., 2006) and Dolph-Chebyshev (Rafaely, 2015). A comparison of the performance of these beamformers for DOA estimation is given in (Delikaris-Manias, Pavlidi, Pulkki and Mouchtaris, 2016).

2.4 Adaptive beamformers

Signal-dependent beamformers will typically utilise the frequency-dependent covariance matrix of the input spherical harmonic signals, which can be estimated as

$$\mathbf{C}_{\text{SH}}(f) = \text{E} \left[\mathbf{s}(t, f) \mathbf{s}(t, f)^H \right] \in \mathbb{C}^{(N+1)^2 \times (N+1)^2} \quad (20)$$

where $\text{E}[\cdot]$ denotes a statistical expectation operator. The covariance matrix can be estimated by averaging this result over finite time frames, typically in the range of tens of milliseconds or by recursive schemes.

One popular adaptive beamformer is the MVDR algorithm, which is a special case of the linearly-constrained minimum-variance (LCMV) algorithm. The generated beam will adaptively change according to the input signals and its response is constrained to unity in the specified *look* direction, while minimising the variance of the output (Rafaely, 2015). The minimisation problem is defined as

$$\begin{aligned} & \text{minimise } \mathbf{w}^H \mathbf{C}_{\text{SH}}(f) \mathbf{w} \\ & \text{subject to } \mathbf{w}^H \mathbf{y}(\Omega) = 1. \end{aligned} \quad (21)$$

The resulting MVDR weights can then be derived as

$$\mathbf{w}(f) = \frac{\mathbf{y}(\Omega)^H \mathbf{C}_{\text{SH}}^{-1}(f)}{\mathbf{y}(\Omega)^H \mathbf{C}_{\text{SH}}^{-1}(f) \mathbf{y}(\Omega)}, \quad (22)$$

which should then be applied per frequency band.

Note that the main advantage of applying the MVDR algorithm in the spherical harmonic domain, instead of utilising the microphone signals directly, is that the steering vectors are simply the spherical harmonics for different angles $\mathbf{y}(\Omega)$ on the sphere.

3 Sound-field analysis and parameter estimation

This section is concerned with exploring various methods of extracting spatial information from a sound-field. The purpose of this may be to convert the information into a visual depiction of the surrounding sound energy, or to enhance the perceptually motivated sound-field reproduction methods described in Section 4. Therefore, various means of displaying the relative sound energy distribution in a sound-field, through the use of beamforming or via statistical means, are presented; as well as the extraction of spatial parameters, such as the DoA and diffuseness.

Additionally, this section details a novel approach which generates pseudo-spectrums by utilising the CroPaC algorithm, presented in (Delikaris-Manias and Pulkki, 2013), with an additional side-lobe suppression technique to improve spatial selectivity. This algorithm has also been generalised to operate on higher-order spherical harmonic signals. Furthermore, a minor reformulation of the WOB-CroPaC algorithm, detailed in (Delikaris-Manias, Pavlidi, Pulkki and Mouchtaris, 2016), is presented, which operates solely on the spherical harmonic signals; thus, reducing the computational complexity for real-time implementations.

3.1 Scanning beamformers

The term *scanning beamforming* is commonly used to refer to the act of steering beamformers in multiple directions, controlled by means of a predefined grid which samples an area of interest, and then calculating the sound energy or a statistical likelihood parameter for each of these directions. In the case of the former, the relative sound energies can be plotted as a power-map, where (for example) a red colour may indicate an area of high sound energy and a blue colour may indicate an area of low sound energy. Whereas the statistical likelihood parameter may be plotted as a pseudo-spectrum, which leads to similar depiction, as the likelihood of a sound arriving from each direction is plotted instead. A user may then observe these visual representations of the sound-field, in order to identify where sound sources and/or early reflections are located.

Generating power-maps

The simplest approach to obtain a power-map \mathbf{P} is to generate regular PWD beamformers and to calculate the energy of the resulting signals

$$\mathbf{P}_{\text{PWD}}(\Omega_j, t, f) = |\mathbf{y}(\Omega_j)^H \mathbf{s}(t, f)|^2, \quad (23)$$

where \mathbf{s} are the spherical harmonic signals and \mathbf{y} are the spherical harmonic weights for directions $\Omega_j = [\Omega_0, \Omega_1, \dots, \Omega_J]$, which are specified by a pre-determined grid that should (ideally) uniformly sample the area of interest. The result may then be averaged over the frequencies of interest and/or averaged over time to suit the application.

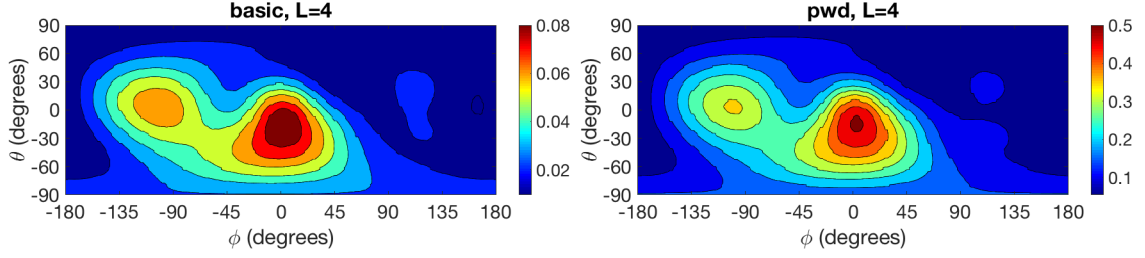


Figure 2: Power-maps generated off-line using fourth order spherical harmonic signals and the PWD algorithm, comparing (23) on the left with (24) on the right. A total of 1002 beamformers were steered uniformly around the unit sphere and then converted to a 2D representation via triangular interpolation. A pre-recorded sound-field was utilised, consisting of a female speaker located at $[90, 0]$ degrees and a male speaker located at $[0, 0]$ degrees, in a reverberant room. Results were averaged with equal weighting over frequency bands with centre frequencies between $[140, 8k]$ Hz.

Alternatively, the PWD power-map can be derived from the spherical harmonic covariance matrix \mathbf{C}_{SH}

$$\mathbf{P}_{\text{PWD}}(\Omega_j, t, f) = |\mathbf{y}(\Omega_j)^H \mathbf{C}_{\text{SH}}(f) \mathbf{y}(\Omega_j)|, \quad (24)$$

which will obtain near-identical results to (23), provided that the covariance matrix is estimated from the same signal frame and not averaged over time. A comparison of the two approaches is depicted in Fig. 2.

This same approach can also be extended to generate power-maps using adaptive beamformers, such as the MVDR algorithm (Rafaely, 2015)

$$\mathbf{P}_{\text{MVDR}}(\Omega_j, t, f) = |\mathbf{w}(\Omega_j)^H \mathbf{C}_{\text{SH}}(f) \mathbf{w}(\Omega_j)|, \quad (25)$$

where

$$\mathbf{w}(\Omega_j) = \frac{\mathbf{y}(\Omega_j)^H \mathbf{C}_{\text{SH}}^{-1}(f)}{\mathbf{y}(\Omega_j)^H \mathbf{C}_{\text{SH}}^{-1}(f) \mathbf{y}(\Omega_j)}. \quad (26)$$

However, since the covariance matrix is time-variant and the weights are required to be recalculated periodically, this method is more computationally demanding than the PWD algorithm; especially if you consider that (26) will require a matrix inversion, lower-upper decomposition or Gaussian elimination, to implement in practice. Furthermore, diagonal loading of the covariance matrix $\mathbf{C}_{\text{SH}}^{\text{w}}$ may be required, in order to mitigate the possibility of ill-conditioned estimates that subsequently lead to an unsuitable matrix inversion being performed

$$\mathbf{C}_{\text{SH}}^{\text{w}}(f) = \mathbf{C}_{\text{SH}}(f) + \left(\frac{\lambda_r}{(N+1)^2} \text{trace}[\mathbf{C}_{\text{SH}}(f)] \right) \mathbf{I}, \quad (27)$$

where λ_r is a regularisation parameter and \mathbf{I} is an identity matrix.

Generating pseudo-spectrums

Alternatively, instead of generating a traditional power-map using beamformers, a pseudo-spectrum may be obtained by utilising subspace methods, such as the MUSIC algorithm described in (Nadiri and Rafaely, 2014). First, the signal $\mathbf{U}_s \in \mathbb{C}^{1 \times 1}$ and noise $\mathbf{U}_n \in \mathbb{C}^{(N+1)^2-1 \times (N+1)^2-1}$ subspaces are obtained via a singular value decomposition (SVD) or eigenvalue decomposition of the spherical harmonic covariance matrix (please note that the frequency and time indexes are omitted for the brevity of notation)

$$\mathbf{C}_{SH} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H = [\mathbf{U}_s \mathbf{U}_n] \begin{bmatrix} \mathbf{\Sigma}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_n \end{bmatrix} \begin{bmatrix} \mathbf{U}_s \\ \mathbf{U}_n \end{bmatrix}, \quad (28)$$

where $\mathbf{\Sigma}$ denotes the singular/eigen values.

A direct-path dominance test is then performed, in order to ascertain which time-frequency bins provide a significant contribution to the direct path of a sound source. These time-frequency bins are selected by determining whether the first singular/eigen value, σ_1 of matrix $\mathbf{\Sigma}$ is significantly larger than the second singular/eigen value, σ_2 , when given in a descending order

$$\frac{\sigma_1}{\sigma_2} > \beta, \quad (29)$$

where $\beta \geq 1$ is a threshold parameter.

Essentially, this subspace method is based on the assumption that the direct path of a sound source will be characterised with higher energy than the reflecting path (Nadiri and Rafaely, 2014). However, unlike the PWD and MVDR approaches, where a power-map is generated by depicting the relative energy of beamformers, the MUSIC pseudo-spectrum is obtained as

$$\mathbf{S}_{MAP}(\Omega_j) = \frac{1}{\mathbf{y}^H(\Omega_j) \left(\mathbf{I} - \mathbf{U}_s \mathbf{U}_s^H \right) \mathbf{y}(\Omega_j)}, \quad (30)$$

where \mathbf{S}_{MAP} is the pseudo-spectrum value for direction Ω_j . For a more comprehensive comparison between various sound-field imaging approaches, the reader is directed to (Delikaris-Manias, Pavlidi, Pulkki and Mouchtaris, 2016).

3.2 Cross-spectrum-based parameter estimation

Instead of using beamformers to generate an energy-based power-map or employing subspace methods to generate a pseudo-spectrum, it is possible to derive statistical-likelihood parameters by utilising the cross-spectrum of different beamformers. This cross pattern coherence (CroPaC) parameter has been previously utilised for spatial filtering applications, where it has been shown to be effective in noisy and reverberant conditions (Delikaris-Manias and Pulkki, 2013; Delikaris-Manias, Vilkamo and Pulkki, 2016). In this subsection, the original CroPaC algorithm presented in (Delikaris-Manias and Pulkki, 2013), has been reformulated and generalised to use static beamformers and spherical harmonic signals of arbitrary order N .

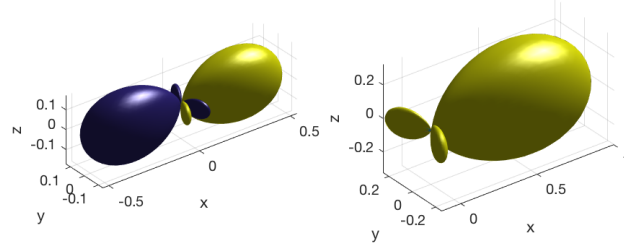


Figure 3: Visualisation of the CroPaC half-wave rectification and normalisation process, before (left) and after (right).

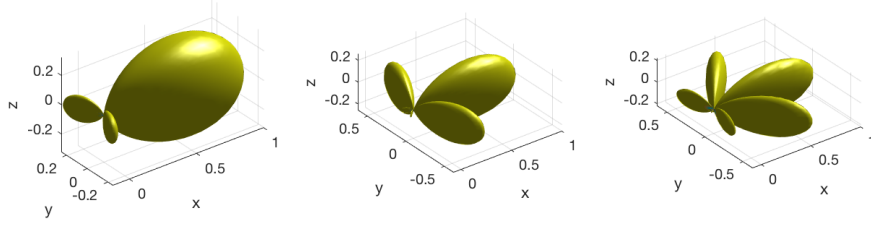


Figure 4: Visualisation of CroPaC beams for $N = 1$ (left), $N = 2$ (middle) and $N = 3$ (right).

The CroPaC algorithm estimates the probability of a sound source emanating from a specific direction in 3D space, by calculating the cross-spectrum between two spherical harmonic signals of orders N and $N + 1$

$$\mathbf{G}(\Omega_j, t, f) = \lambda \frac{\Re[\mathbf{s}_N(\Omega_j, t, f) * \mathbf{s}_{N+1}(\Omega_j, t, f)]}{\sum_N^{N+1} |\mathbf{s}_N(\Omega_j, t, f)|^2}, \quad (31)$$

where \Re denotes the real operator; \mathbf{s}_N and \mathbf{s}_{N+1} are the spherical harmonic signals for a look direction Ω_j and the same degree m , $*$ denotes the complex conjugate and λ is an order-dependent normalisation factor to ensure that $\mathbf{G}_{\text{MAP}} \in [0, 1]$. The normalisation factor can be calculated as

$$\lambda = \frac{(N + 1)^2 - (N - 1)^2 + 1}{2} = \frac{4N + 1}{2}. \quad (32)$$

The power-map is then estimated for a grid of look directions $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_J)$, averaged across frequencies and subjected to a half-wave rectifier. The resulting power-map is then given by

$$\mathbf{G}_{\text{MAP}}(\Omega_j, t) = \max \left[0, \frac{1}{K} \sum_{f=1}^K \mathbf{G}(\Omega_j, t, f) \right]. \quad (33)$$

The half-wave rectification process ensures that only sounds arriving from the look direction are analysed. An illustration of how this half-wave rectification process affects the directional selectivity of the CroPaC beams is depicted in Fig. 3.

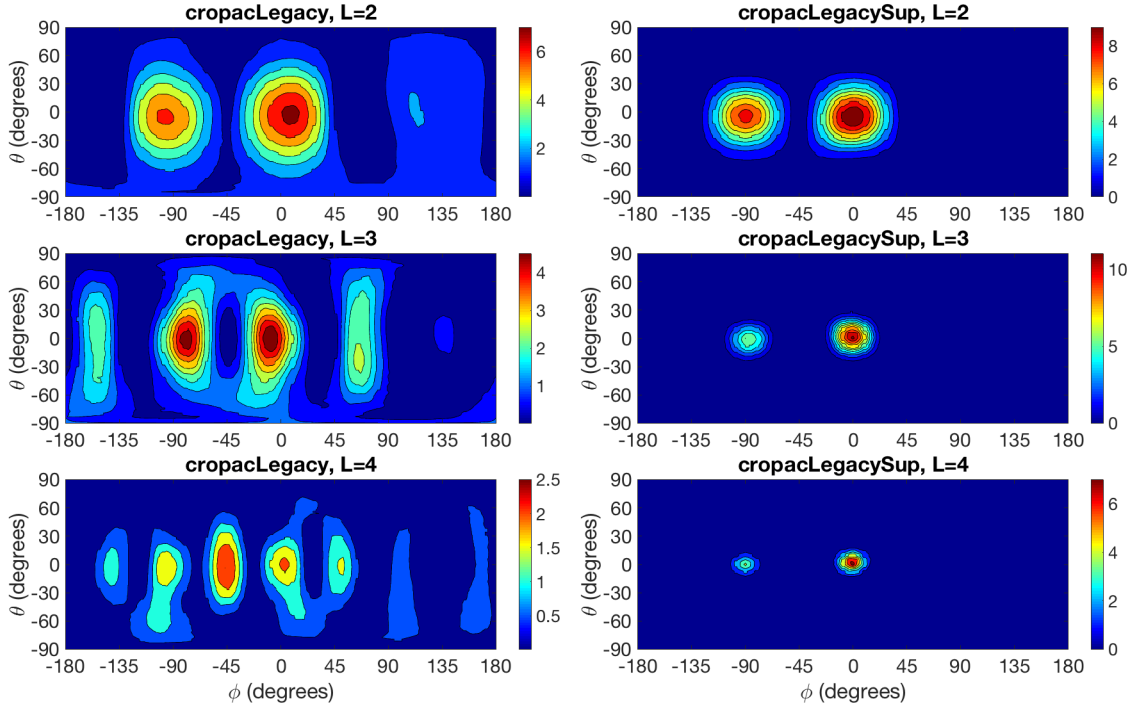
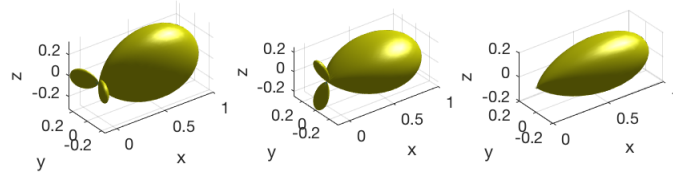


Figure 5: The original CroPaC algorithm (left) and the proposed side-lobe suppressed version (right), using second, third and fourth order spherical harmonic signals (top-bottom). A total of 1002 CroPaC beamformers were steered uniformly around the unit sphere and then converted to a 2D representation via triangular interpolation. A pre-recorded sound-field was utilised, consisting of a female speaker located at $[90, 0]$ degrees and a male speaker located at $[0, 0]$ degrees, in a reverberant room. Results were averaged with equal weighting over frequency bands with centre frequencies between $[140, 8k]$ Hz.

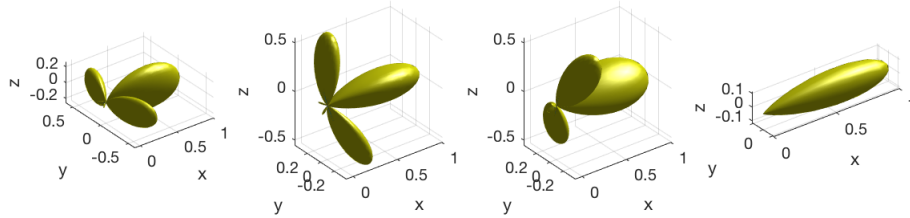
3.2.1 Side-lobe suppression

The fundamental problem with the original CroPaC algorithm (Delikaris-Manias and Pulkki, 2013), and the generalised formulation (31), is that the calculation of the cross-spectrum between different orders of spherical harmonics results in the creation of unwanted side-lobes, which vary depending on the order. A visual depiction of these aberrations, shown in Fig. 4, may be generated by multiplying the following spherical harmonics together: $Y_{NN}Y_{(N+1)(N+1)}$ for $N = 1$ (left), $N = 2$ (middle) and $N = 3$ (right).

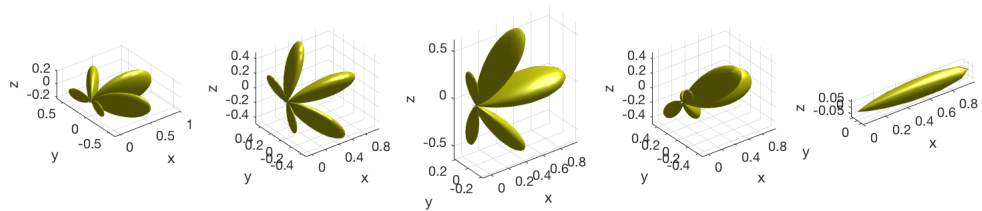
These side-lobes can introduce biases in the power-map, as is depicted in Fig. 5. This thesis therefore, proposes a technique that suppresses these side-lobes by multiplying rotated versions of the estimated CroPaC beams. The number of rotations/estimated beams is determined by the order N . This novel side-lobe suppressing



(a) Side-lobe suppression for order $N = 1$. The two beams on the left show the rotating beam patterns and the beam on the right is the resulting beam pattern.



(b) Side-lobe suppression for order $N = 2$. The three beams on the left show the rotating beam patterns and the beam on the right is the resulting beam pattern.



(c) Side-lobe suppression for order $N = 3$. The four beams on the left show the rotating beam patterns and the beam on the right is the resulting beam pattern.

Figure 6: Visualisation of side-lobe cancellation for $N = 1, 2, 3$.

parameter \mathbf{G}_{SUP} can be estimated as

$$\mathbf{G}_{\text{SUP}}(\Omega_j, t) = \begin{cases} \mathbf{G}_{\text{MAP}}(\Omega_j, t) & , \text{ if } N = 1 \\ \prod_{i=1}^N \mathbf{G}_{\text{MAP}}^{\rho_i}(\Omega_j, t) & , \text{ if } N > 1, \end{cases} \quad (34)$$

where ρ_i is a parameter that defines an axis-symmetric roll of $\frac{\pi}{N}$ radians of the spherical harmonic signals prior to the calculation of the CroPaC parameter. This rotation can successfully suppress side-lobes that are generated when using (31). This concept is illustrated in Fig. 6; where each row illustrates the side-lobe suppression for different orders. In the top row $N = 1$, which results in a single roll of $\frac{\pi}{2}$ in (34). For $N = 2$ (middle row) and $N = 3$ (bottom row) three and four rolls of $\frac{\pi}{3}$ and $\frac{\pi}{4}$ are applied, respectively. The resulting enhanced beam patterns $\mathbf{G}_{\text{SUP}} \in [0, 1]$, which are derived from the product of multiple $\mathbf{G}_{\text{MAP}} \in [0, 1]$, are shown on the right-hand side of the figures.

3.2.2 Weighted-orthogonal beamforming-based formulation

Much in the same vein as the original CroPaC algorithm, the more recent LCMV inspired reformulation also relies on determining the cross-spectrum between two coincident beamformers (Delikaris-Manias, Vilkamo and Pulkki, 2016), in order to formulate a post-filter and subsequently enhance an existing robust beamformer. The cross-spectrum operation is given as (please note that the time and frequency indices (t, f) have been omitted for the brevity of notation)

$$\Phi_{\text{SaSo}}(\Omega_j) = \mathbf{w}_a^H(\Omega_j) \text{diag}(\mathbf{C}_{\text{SH}}) \mathbf{w}_o(\Omega_j), \quad (35)$$

where $\mathbf{w}_a(\Omega_j) \in \mathbb{C}^{(N+1)^2 \times 1}$ is the steering vector of a signal-independent beamformer specified by the user; and $\mathbf{w}_o(\Omega_j) \in \mathbb{C}^{(N+1)^2 \times 1}$ is a steering vector of an adaptive beamformer derived from an LCMV minimisation problem

$$\begin{aligned} \hat{\mathbf{w}}_o(\Omega_j) &= \arg \min_{\mathbf{w}_o(\Omega_j)} \mathbf{w}_o^H(\Omega_j) \mathbf{C}_{\text{SH}} \mathbf{w}_o(\Omega_j) \\ &\text{subject to } \mathbf{A}(\Omega_j) \mathbf{w}_o(\Omega_j) = \mathbf{b}, \end{aligned} \quad (36)$$

using

$$\mathbf{A}(\Omega_j) = [\mathbf{y}(\Omega_j) \text{diag}[\mathbf{C}_{\text{SH}}] \mathbf{w}_a(\Omega_j)]^H, \quad (37)$$

and

$$\mathbf{b} = [1 \ 0]^T, \quad (38)$$

where $\mathbf{y} \in \mathbb{C}^{(N+1)^2 \times 1}$ is a spherical harmonic steering vector for direction Ω_j , which should be the same look-direction as the robust beamformer that is to be enhanced.

Applying the Lagrange multipliers method (Balanis and Ioannides, 2007), the solution to (36) is given as

$$\mathbf{w}_o(\Omega_j) = \frac{\mathbf{C}_{\text{SH}}^{-1} \mathbf{A}^H(\Omega_j)}{\mathbf{A}(\Omega_j) \mathbf{C}_{\text{SH}}^{-1} \mathbf{A}^H(\Omega_j)} \mathbf{b}. \quad (39)$$

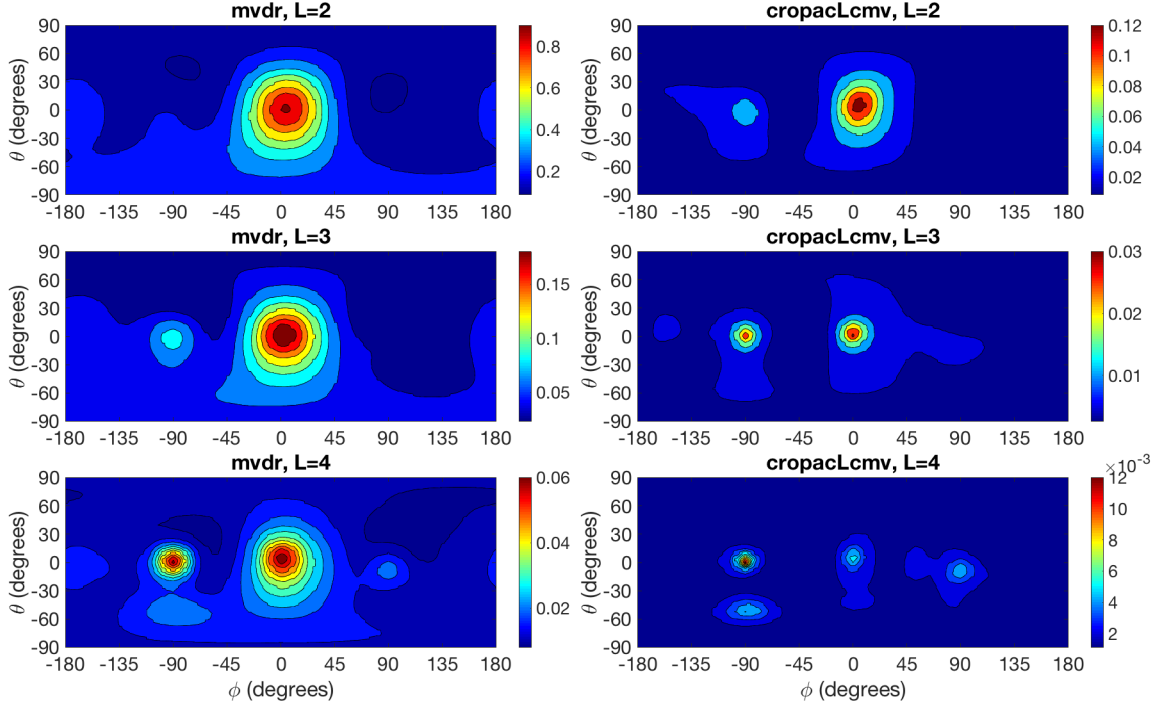


Figure 7: Power-map utilising MVDR beamformers (left) and the revised WOB-CroPaC formulation (right); using second, third and fourth order spherical harmonic signals (top-bottom). A total of 1002 beamformers were steered uniformly around the unit sphere and then converted to a 2D representation via triangular interpolation. A pre-recorded sound-field was utilised, consisting of a female speaker located at $[90, 0]$ degrees and a male speaker located at $[0, 0]$ degrees, in a reverberant room. Results were averaged with equal weighting over frequency bands with centre frequencies between $[140, 8k]$ Hz.

In (Delikaris-Manias, Vilkamo and Pulkki, 2016) the post-filter was applied to a robust beamformer derived in the space domain; whereas for this thesis work, it was modified to operate solely in the spherical harmonic domain. This reduces the computational complexity, as the time-frequency transform needs only be applied to the spherical harmonic signals and not also to the microphone signals. Therefore, the final gain factor is given as

$$\mathbf{G}_{\text{WOB}}(\Omega_j) = \sqrt{\frac{\min[E_b(\Omega_j), |\Phi_{\text{Sas}}(\Omega_j)|]}{E_b(\Omega_j)}}, \quad (40)$$

where $E_b(\Omega_j)$ is the energy of the robust beamformer that is to be adaptively attenuated to improve its performance, calculated as

$$E_b(\Omega_j) = |\mathbf{w}_b^H(\Omega_j) \mathbf{C}_{\text{SH}} \mathbf{w}_b(\Omega_j)|, \quad (41)$$

where $\mathbf{w}_b(\Omega_j)$ are beamforming weights of the robust beamformer formulated in the spherical harmonic domain. Note that (40) can also be constrained by a spectral

floor parameter, such as in (33), to allow the user the control over how harshly the post-filter is applied to the robust beamformer.

A comparison between (40) and MVDR beamformers, is depicted in Fig. 7; where $\mathbf{w}_b(\Omega_j)$ are MVDR beamforming weights calculated using (26) and $\mathbf{w}_a(\Omega_j)$ are PWD beamforming weights with Dolph-Chebyshev weighting (Rafaely, 2015). It is important to note that since the diffuse noise of the spherical harmonic signals is not known, the assumption made in (35), is that the diagonal elements of \mathbf{C}_{SH} can be utilised to provide an estimation of this noise. Therefore, the post-filter will act as a form of spatial-gate during periods when the signal extracted from the specified look direction is on-going. This is also the reason why the post-filter should theoretically work better in high-noise and reverberant environments, as the estimate effectively becomes more accurate as this noise level increases. The algorithm also has the added benefit of deriving the post-filter from one set of beamforming weights, as the $\mathbf{w}_o(\Omega_j)$ weights are derived from the user specified $\mathbf{w}_a(\Omega_j)$ weights; whereas the original CroPaC algorithm requires the user to specify both beam patterns.

3.3 Direction-of-arrival and diffuseness parameter estimation

Extracting estimates of the DoA of sound sources and the diffuseness within a sound-field is useful for many sound-field visualisers and parametric reproduction methods that are available today.

DoA estimation via peak finding

To determine the DoA of the most prominent sound source in a sound scene, it is possible to use an existing power-map/pseudo-spectrum and subsequently determine the highest energy point and note down this position. Such an approach is commonly referred to as *peak finding* and can be performed simply as

$$\text{DoA}(\theta_s, \phi_s) = \Omega \in \max(\mathbf{P}_{\text{MAP}}). \quad (42)$$

where $\theta_s \in [-\pi/2, \pi/2]$ denotes the elevation angle and $\phi_s \in [-\pi, \pi]$ the azimuthal angle of the sound source; and \mathbf{P}_{MAP} is an arbitrary power-map or pseudo-spectrum.

However, extracting multiple DoA estimates from a single power-map is not a trivial matter as this will involve removing the area of the power-map containing the first DoA estimate, before determining the next highest peak. Therefore, ascertaining the optimal amount of area to remove from the power-map, such that the same sound source is not identified twice or a second sound source is unintentionally removed, can be problematical; especially, if the spatial width of the sound source is not known. For an in-depth description of a power-map based multiple sound source localisation approach, the reader is directed to (Pavlidis et al., 2015).

The act of deriving DoA estimates from power-maps, however, is inherently computationally demanding, as numerous beamformers are required to sufficiently sample the area of interest; in order to subsequently obtain these estimates with adequate precision. Therefore, while these approaches are well suited for scenarios

which do not require real-time analysis, less computationally demanding methods are preferred for certain applications. Furthermore, in the case of parametric reproduction methods (such as the original DirAC formulation) only a single DoA estimate is required.

DoA and diffuseness estimation using first-order spherical harmonic signals

Due to the large computational complexity associated with power-map-based DoA estimation, cheaper and more flexible alternatives are often employed for real-time applications. Perhaps the most popular approach, is to extract the DoA and diffuseness, ψ , estimates from the active intensity vector \mathbf{i}_a , which can be estimated per frequency using the zeroth and first order spherical harmonic signals, \mathbf{s} , as (Pavlidis et al., 2015)

$$\mathbf{i}_a(t, f) = \frac{1}{2} \Re[p(t, f) \mathbf{u}^*(t, f)], \quad (43)$$

where, p , is the sound pressure, estimated as

$$p(t, f) \simeq \mathbf{s}_{00}(t, f), \quad (44)$$

and, \mathbf{u} , is the particle velocity. Assuming that the sound sources are received as plane-waves, the particle velocity can be estimated as

$$\mathbf{u}(t, f) \simeq -\frac{1}{\rho_0 c \sqrt{2}} \begin{bmatrix} \mathbf{s}_{1-1}(t, f) \\ \mathbf{s}_{10}(t, f) \\ \mathbf{s}_{11}(t, f) \end{bmatrix}, \quad (45)$$

where ρ_0 is the density of the medium in question and c is the speed of sound.

The DoA estimate can then be obtained simply as

$$\text{DoA}(\theta_s, \phi_s, t, f) = \angle(\mathbf{E}[-\mathbf{i}_a(t, f)]), \quad (46)$$

where, \angle , represents the angle operator. Whereas, the diffuseness can be obtained as

$$\psi(t, f) = 1 - \frac{|\mathbf{E}[\mathbf{i}_a(t, f)]|}{c \mathbf{E}[e(t, f)]}, \quad (47)$$

where e is the energy density, calculated as

$$e(t, f) = \frac{\rho_0}{2} |\mathbf{u}(t, f)|^2 + \frac{|p(t, f)|^2}{2\rho_0 c^2}. \quad (48)$$

Both the DoA and ψ estimates can then be averaged over frequencies or treated individually, depending on the application. For a more detailed and up-to-date description of how to utilise the active intensity vector for DoA estimation, the reader is directed to (Pavlidis et al., 2016) and (Delikaris-Manias et al., 2017).

4 Perceptually motivated sound-field reproduction

Before delving into perceptually motivated spatial sound-field reproduction methods, it is first helpful to have an understanding of how the human auditory system perceives a spatial sound-field to begin with. Therefore, this section begins with a brief introduction into spatial hearing. This is followed by formulations for both non-parametric and perceptually motivated parametric means of reproducing sound-fields, in the spherical harmonic domain. For the former, the popular ambisonics method is detailed, whereby spherical harmonic signals are linearly mapped to loudspeaker positions via a decoding matrix for loudspeaker or headphone playback. Whereas the parametric reproduction method which this thesis work focused upon, is the DirAC approach; which operates by extracting perceptually meaningful parameters from the sound-field, in order to enhance the spatial accuracy of the decoded audio which is delivered to the listening position.

A new DirAC formulation for headphones is then presented, which aims to reduce the computational complexity of DirAC implementations that preceded it; while also reducing artefacts by utilising the optimal adaptive mixing solution presented in (Vilkamo and Pulkki, 2013). This approach is similar to the loudspeaker implementation described in (Politis, Vilkamo and Pulkki, 2015), whereby binaural cues are synthesised directly from the analysed spatial parameters.

4.1 Spatial hearing

During human evolution, the sensory ability for locating the direction of sound sources was essential for survival in the pre-modern world, which is largely the reason why humans possess relatively advanced auditory systems. Today, spatial hearing remains useful for orientation in one's environment and for localising sound sources that may or may not be visible (Pulkki and Karjalainen, 2015). The term *localisation* describes the ability to make judgements regarding the direction of a sound source and also its distance relative to the listener. According to Moore (Moore, 2012), the most reliable cues utilised in the localisation of sounds sources are derived from the comparison between the signals arriving at the two ear canals; however, localisation can still be partly influenced by the signal at a single ear canal. It is at this point that a distinction is made between *binaural* (both ears) and *monaural* (one ear) directional cues.

Binaural and monaural cues

Binaural directional cues arise when the auditory system decodes the differences in sound between the two ear canals, in order to help ascertain the azimuth and elevation of sound sources. Spectral differences are referred to as frequency-dependent *interaural level differences* (ILDs) and the temporal differences are referred to as *interaural time differences* (ITDs) (Pulkki et al., 2011). However, it is important to note that due to the physical properties of acoustical wavefronts, these ILDs and ITDs are not equally effective for localisation at all frequencies (Moore, 2012). This

is best explained by describing how these interaural differences occur. Firstly, ILDs arise due as a result of the shadowing effect of the head, where the contra-lateral ear signal is attenuated. This is because high frequencies, with wavelengths shorter than the dimensions of the head, do not diffract around the head or pass through it easily (Moore, 2012). ITDs on the other hand, simply occur due to the fact that the ears are located on different sides of the skull; thus, arrival times of wavefronts are dependent on the direction of the sound source relative to the position of the head (Pulkki and Karjalainen, 2015). Additionally, ITDs will also result in an interaural phase difference (IPD) when the auditory system is presented with on-going or sinusoidal sound sources (Moore, 2012). Psychoacoustic tests (Blauert, 1997) have shown that the auditory system is sensitive to ILD at all frequencies. However, it is generally acknowledged that ILDs that are sufficiently large to provide useful localisation cues, will occur for frequencies above 2 kHz (Moore, 2012). On the other hand, the auditory system is sensitive to ITDs mainly at frequencies below 750 Hz; although, humans may also be sensitive to ITDs between the signal envelopes at higher frequencies (Pulkki et al., 2011). The *duplex theory* introduced by Rayleigh (Rayleigh, 1907), describes the extent of the relevance of these ITD and ILD cues for specific frequency ranges. However, while this theory may be sufficiently accurate when applied to pure tones, complex sounds will deviate from these specified boundaries in practice (Moore, 2012).

Monaural spectral cues are developed within the pinna of the listener and the torso itself may also filter the incoming sound depending on the DoA and introduce spectral changes from 1-2 kHz upwards (Pulkki et al., 2011). However, frequencies above 6 kHz are the most affected, as it is at these frequencies where the wavelengths are sufficiently short for the sound waves to diffract and reflect strongly within the cavities of the pinna (Moore, 2012). This direction dependent filtering is especially important for the localisation of elevated sound sources (Blauert, 1997). Additionally, this filtering provides information about the source direction within the *cone of confusion* (Pulkki and Karjalainen, 2015); which is a perceptually ambiguous region, where the difference in distance from both ears to any point on the surface of an imaginary cone is constant. Although, these ambiguities related to the cone of confusion may also be resolved by head movements (Moore, 2012).

Monaural and binaural cues for a specific direction can be collectively described by a head-related transfer function (HRTF), where the ILD corresponds to the difference in frequency-dependent magnitudes between the left and right spectra (Xie, 2013). The ITD are encoded into the phase characteristics of the HRTFs (Xie, 2013); and the monaural cues are derived from the ratio between the spectral content of the sound source and the spectral content of the left and right channels (Vorländer, 2007).

Localisation in enclosed spaces

When subjected to a sound source in a normal listening situation, the sound will reach the two ears from a number of different directions, due to reflections from the surfaces in the surroundings. These reflections will provide the listener with

some impression of the listening space and a sense of the objects and architecture around them (Valimaki et al., 2012). This interaction between the sound source and the room is referred to as reverberation and is comprised of both early and late reflections.

Despite the fact that many reflected sound waves arrive at the listening position from multiple directions, the auditory system is still capable of localising sound sources based on the directional cues from the direct sound with little additional difficulty (Moore, 2012). This is due to assisting mechanisms in the auditory system that suppress early reflections, so that localisation is largely determined by the direct sound. After psychoacoustic experiments, such as in (Litovsky et al., 1999), it has been established that this phenomenon is due to an *echo suppression* and a subsequent *precedence effect*; historically, referred to as the *Haas effect* (Haas, 1951) or the *law of the first wavefront* (Blauert, 1997). Essentially, if two correlated sounds reach the ear in close succession, then an echo suppression occurs to ensure that the listener perceives them as a single fused sound; provided that the interval between them is below the *echo threshold* (which is roughly between 30-40 ms (Pulkki and Karjalainen, 2015)). This single fused sound will then be localised in the direction of the earliest sound, due to this precedence effect phenomenon; provided that the sound is of a transient nature (Moore, 2012). In situations where the arrival time between the two correlated sounds is less than 1 ms, this precedence effect no longer operates; instead, the two sounds are localised in a direction that is somewhere in-between the two sounds, which is due to a phenomenon referred to as *summing localisation* (Moore, 2012); exploited prolifically in sound reproduction systems.

The reverberation that follows a direct sound is comprised of both early reflections and spatially diffuse late reflections. The early reflections are less likely to be masked than the later reflections; therefore, psychoacoustic evaluation is often biased to the analysis of the first 80 ms (Begault and Trejo, 2000). While it has been acknowledged that the direct path is largely responsible for localisation of a sound source, as a result of the assisting mechanisms described above, the early reflections are more responsible for conveying a sensation of both the geometry and the materials of the surrounding space (Valimaki et al., 2012). A typical room has several physical features, which have an influence on reverberation and consequently the perception of the environmental context. These include: the volume of the room, which is evident in the length and magnitude of the reverberation; the absorptivity of the room surfaces, which are frequency-dependent; and the shape of the enclosure, including the orientation of the objects within it (Begault and Trejo, 2000). Furthermore, certain physical parameters of a room are often calculated, as they can be roughly translated into perceptual characteristics; these include the reverberation time (T60), the reflected-to-direct ratio (R/D) and the spatial distribution of the early reflections (Begault and Trejo, 2000).

The T60 is defined as the duration that is required for the reflected sound energy to fall 60 dB below that of the direct sound energy. This T60, to some extent, is proportional to the listener's perception of the size of the surrounding environment or enclosure (Begault and Trejo, 2000). Whereas, the R/D ratio is simply the ratio of the reverberant sound energy to direct sound energy and can be attributed to the

extent of the perceived distance between the sound source and the receiver (Politis, Delikaris-Manias and Pulkki, 2015). Interestingly, it has been found that even a single reflection from a wall can provide sufficiently reliable distance cues (Sheeline, 1983; Von Békésy and Wever, 1960). Increasing the reverberant sound energy that arrives at the receiver position, will also have an influence on the perceived *apparent source width* (ASW) (Von Békésy and Wever, 1960); which equates to the listener’s perception of the physical extent of a sound source. Additionally, the early lateral reflections can also have an influence on the perceived ASW, while potentially adding some uncertainty to sound source localisation (Vorländer, 2007). Whereas, late lateral reflections are considered to be responsible for a dimension of spatial impression known as *spaciousness* or *listener envelopment* (LEV) (Toole, 2008); which may be described as quite literally how enveloped in sound the listener perceives to be. While research regarding the cues that contribute to this perceived ASW and LEV is still ongoing, these sensory concepts are commonly described collectively by the frequency-dependent normalised cross-correlation between the two ear canal signals; referred to as the *interaural coherence* (IC).

4.2 Ambisonics reproduction

Ambisonics reproduction is a method which aims to recreate a captured sound-field, by utilising spherical harmonic signals as the input and a decoder that takes into account the target playback system. This decoder is essentially a matrix of beamforming coefficients, with one steering vector per loudspeaker direction or filter per headphone transducer.

Ambisonic decoding to loudspeakers can be performed via a single matrix multiplication with $\mathbf{D}_{\text{ls}} \in \mathbb{R}^{L \times (N+1)^2}$, where L refers to the number of loudspeakers

$$\mathbf{y}_{\text{ls}}(t, f) = \mathbf{D}_{\text{ls}} \mathbf{s}_N(t, f). \quad (49)$$

The ambisonic decoding matrix depends on the available order N ; where the simplest decoder is formulated as

$$\mathbf{D}_{\text{ls}} = \frac{1}{L} \mathbf{Y}_{\text{ls}}^T, \quad (50)$$

where $\mathbf{Y}_{\text{ls}} \in \mathbb{R}^{(N+1)^2 \times L}$ is a matrix of spherical harmonics for each loudspeaker position.

However, a more perceptually motivated decoder is the all-round ambisonic decoder (ALLRAD) detailed in (Zotter and Frank, 2012), which maps signals to a uniformly-distributed loudspeaker layout with energy preserving properties in a linear manner, before routing them to an arbitrary loudspeaker set-up via vector-base amplitude panning (VBAP) (Pulkki, 1997)

$$\mathbf{D}_{\text{ls}} = \frac{1}{L_{\text{td}}} \mathbf{G}_{\text{td}} \mathbf{Y}_{\text{td}}^T, \quad (51)$$

where L_{td} is the number of directions of a uniform spherical t-design grid (Hardin and Sloane, 1996); $\mathbf{Y}_{\text{td}} \in \mathbb{R}^{(N+1)^2 \times L_{\text{td}}}$ is a matrix of spherical harmonics for the

t-design grid positions; and $\mathbf{G}_{\text{td}} \in \mathbb{R}^{L \times L_{\text{td}}}$ is a matrix of VBAP loudspeaker gains which assign the t-design signals to virtual source positions. For a more expanded description of ambisonic decoding in general, the reader is referred to (Zotter and Frank, 2012).

Binaural decoding

Ambisonic decoding to headphones, on the other hand, can be performed with a matrix of filters $\mathbf{D}_{\text{bin}} \in \mathbb{C}^{2 \times (N+1)^2}$ based on a set of individualised or generic HRTFs

$$\mathbf{y}_{\text{bin}}(t, f) = \mathbf{D}_{\text{bin}}(f) \mathbf{s}_N(t, f). \quad (52)$$

This binaural ambisonic decoding matrix also depends on the available order N and can be designed based on a virtual loudspeaker approach (Wiggins et al., 2001; Melchior et al., 2009), or by directly expressing the HRTFs in the spherical harmonic domain (Shabtai and Rafaely, 2013); for a description of the two approaches see e.g. (Politis and Poirier-Quinot, 2016).

Additionally, rotation of the sound-field, necessary for head-tracking, can be performed through an appropriate spherical harmonic rotation matrix $\mathbf{M}_{\text{rot}} \in \mathbb{R}^{(N+1)^2 \times (N+1)^2}$

$$\mathbf{y}_{\text{bin}}^{(\text{rot})}(t, f) = \mathbf{D}_{\text{bin}}(f) \mathbf{M}_{\text{rot}}(\alpha, \beta, \gamma) \mathbf{s}_N(t, f), \quad (53)$$

where α, β, γ are rotation angles sent by the head-tracker. For implementation details on the rotation matrices, the reader is again referred to (Politis and Poirier-Quinot, 2016).

4.3 Directional audio coding reproduction

While the ambisonics method (much like the signal-independent PWD beamformers in Section 3) does not introduce any distortion to the resulting signals, its performance is completely determined by the spatial resolution of the input format. In the case of first order and lower-order spherical harmonic signals, this limitation can affect the perceived spatial quality of the output; resulting in directional blurring of point sources, localisation ambiguity, and strong colouration effects (Solvang, 2008; Santala et al., 2009; Braun and Frank, 2011; Kearney et al., 2012; Bertet et al., 2013; Bernschütz et al., 2014; Stitt et al., 2014; Yang and Bosun, 2014).

On the other hand, perceptually motivated approaches, such as DirAC, are designed to improve upon the quality offered by ambisonic decoding, while still utilising the same input spherical harmonic signals. The DirAC method can be divided into two main processes, the *analysis* of spatial parameters and the *synthesis* of the signals which accurately exhibit these analysed traits via parametric enhancement. The end goal, is to reproduce a sound-field with a higher degree of perceived spatial accuracy compared to ambisonics and other non-parametric reproduction methods; while minimising any detrimental artefacts that may be introduced in the process.

4.3.1 Parametric analysis

The spatial parameter vector which is utilised by DirAC comprises of a DoA estimate, the total energy and the diffuseness. For first order spherical harmonic signals \mathbf{s}_1 this process can be denoted as

$$\mathbf{p}_1(t, f) = \mathcal{A}[\mathbf{s}_1(t, f)] = [\theta, \phi, e, \psi], \quad (54)$$

where $\mathcal{A}[\cdot]$ refers to the extraction of spatial parameters from the signal in question (refer to Section 3.3 for more details).

Note that in the case of first order spherical harmonic signals, only one estimate of the parameter vector is possible per time and frequency index. However, if higher-order spherical harmonic signals are available, then spatially selective patterns may be imposed on the zeroth and first order spherical harmonic signals, in order to extract multiple parameter vectors per time and frequency index. This results in improved spatial accuracy, in cases where multiple sound sources are located in their own individual sectors.

The extraction of a parameter vector for S sectors can be written as

$$\mathbf{p}_N(t, f) = \mathcal{A}[\mathbf{W}_N \mathbf{s}_N(t, f)] = [e_1, \psi_1, \theta_1, \phi_1, \dots, e_S, \psi_S, \theta_S, \phi_S], \quad (55)$$

where $\mathbf{W}_N \in \mathbb{C}^{4S \times (N+1)^2}$ is a beamforming matrix that generates the appropriate sector analysis signals, and S is the order-dependent number of sectors; for more details on the structure of \mathbf{W}_N , the reader is referred to (Politis, Vilkamo and Pulkki, 2015). A depiction of these spatially selective sectors utilising second order spherical harmonic signals is provided in Fig. 8. Furthermore, it is important to note that if only first order spherical harmonic signals are available, the analysis reduces to the basic parameter vector analysis in (54).

4.3.2 Legacy first-order synthesis

Once the parameter vectors have been extracted from the spherical harmonic signals, preliminary first-order ambisonic decoding for loudspeakers is performed with matrix $\mathbf{D}_{ls} \in \mathbb{R}^{L \times (N+1)^2}$ as in (49).

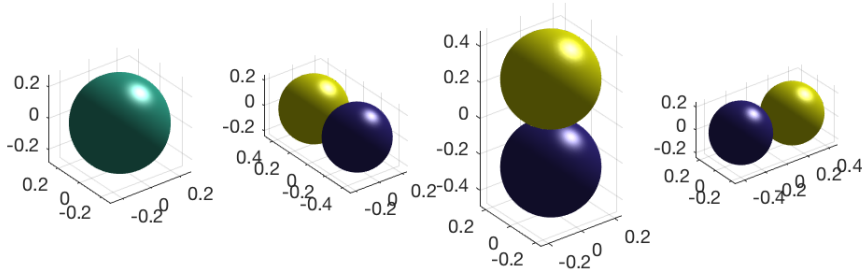
These loudspeaker signals are then parametrically enhanced, where the analysed parameters are utilised to separate the signals into directional and diffuse streams, which are then re-distributed appropriately between the loudspeakers. The directional stream \mathbf{z}_{dir} can be obtained by employing vector-base amplitude panning (VBAP) (Pulkki, 1997)

$$\mathbf{z}_{\text{dir}}(t, f) = \frac{\sqrt{1 - \psi}}{L} \mathbf{G}(\theta, \phi) \mathbf{y}_{ls}(t, f), \quad (56)$$

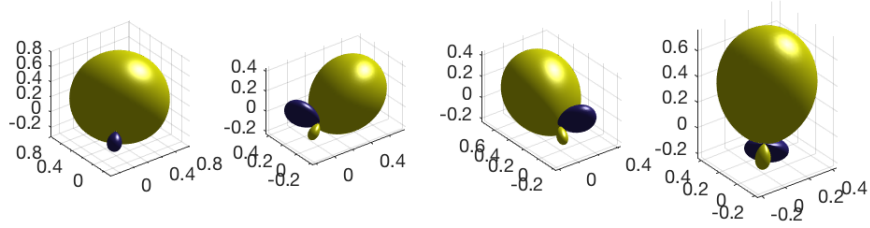
where \mathbf{G} is a diagonal matrix containing the VBAP gains for the estimated DoA, and ψ is the diffuseness estimate.

The diffuse stream can then be obtained via suitable decorrelation of the ambisonic decoded signals

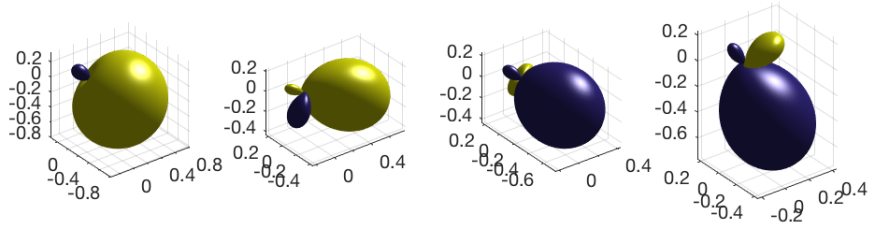
$$\mathbf{z}_{\text{diff}}(t, f) = \mathcal{D} \left[\sqrt{\psi} \mathbf{y}_{ls}(t, f) \right], \quad (57)$$



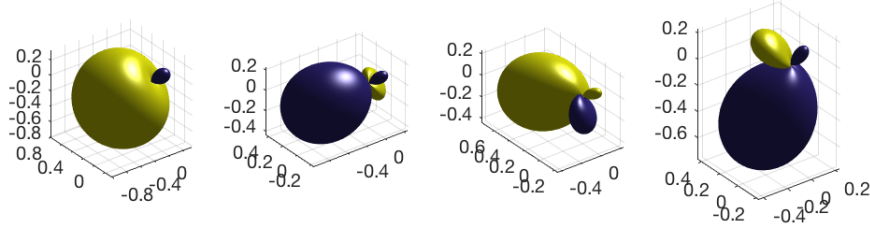
(a) Spherical harmonics up to the first-order.



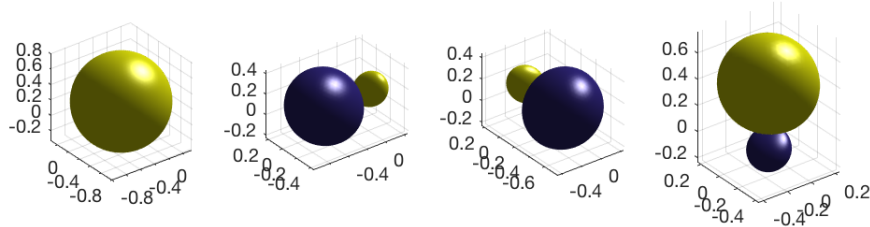
(b) First sector patterns.



(c) Second sector patterns.



(d) Third sector patterns.



(e) Fourth sector patterns.

Figure 8: Visualisation of the spherical harmonics up to the first-order (a), and after the second-order spherical harmonics have been spatially filtered to obtain the sectors used by the higher-order DirAC analysis (b)-(e).

where $\mathcal{D}[\cdot]$ denotes a decorrelation operation of the enclosed signals.

The resulting direct and diffuse loudspeaker streams may then be summed together and reproduced over loudspeakers, or alternatively, they can be convolved with their respective HRTFs in order for them to be experienced via headphones

$$\mathbf{y}_{\text{bin}}(t, f) = \mathbf{H}_{\text{ls}}[\mathbf{z}_{\text{dir}}(t, f) + \mathbf{z}_{\text{diff}}(t, f)], \quad (58)$$

where the binaural rendering of the virtual loudspeaker signals is performed with a set of HRTFs contained within a matrix of filters $\mathbf{H}_{\text{ls}} \in \mathbb{C}^{2 \times L}$. Head-tracking may then be integrated by rotating the analysed DoAs and the spherical harmonic signals prior to the ambisonic decoding (Laitinen and Pulkki, 2009).

4.3.3 Higher-order synthesis for headphones with optimal mixing

The primary drawback with the legacy first-order DirAC synthesis approach, is that it does not take into account what has already been achieved by the ambisonic decoding, in terms of directional spatialisation and diffuse reproduction. Rather, time-variant panning and diffuse gains are applied to the linearly-decoded signals $\mathbf{D}_{\text{ls}}\mathbf{s}_1$ without jointly considering the combined linear and parametric rendering; with the exception of a mean correction for the purpose of energy preservation (Vilkamo et al., 2009). Therefore, an improvement in spatial accuracy and signal fidelity should be attained, if the panning gains and decorrelation are applied only to the extent that is necessary following the linear ambisonic decoding. Such a reformulation of the DirAC synthesis was presented by Vilkamo et al. in (Vilkamo and Pulkki, 2013), referred to here as *optimal mixing*.

Another limitation of the virtual loudspeaker approach detailed in (Laitinen and Pulkki, 2009), is its inability to make use of higher-order spherical harmonic signals at both the analysis and synthesis stages; which would be expected to obtain a more accurate model of the sound-field and subsequently improve its performance in scenarios that are challenging for the basic intensity-diffuseness model. The HO-DirAC formulation detailed in (Politis, Vilkamo and Pulkki, 2015) takes advantage of the higher spatial resolution of higher-order spherical harmonic signals in order to extract multiple intensity-diffuseness estimates from spatially separated sectors (see Section 4.3.1). The formulation also fully integrates the optimal mixing approach in (Vilkamo and Pulkki, 2013), which aims to preserve the single-channel quality of the linear ambisonic decoding, while also enhancing the spatialisation cues that ambisonics may fail to deliver; such as sharp point source sounds and spatially incoherent diffuse sounds.

In (Politis, Vilkamo and Pulkki, 2015), HO-DirAC was formulated for arbitrary loudspeaker set-ups, while in this section it has been reformulated for efficient headphone reproduction.

Optimal mixing solution

The optimal mixing formulation, presented in (Vilkamo and Pulkki, 2013), essentially solves the problem of finding a mixing matrix which when applied to the ambisonic

decoded signals, will yield outputs that have the desired binaural cues, while also preserving the high single-channel quality of the decoded signals as much as possible.

Since the ILD, ITD and IC cues relate directly to inter-channel level differences (ICLD), phase differences (ICPD) and coherence (ICC) during headphone listening, they are all encoded into the covariance matrix of the binaural signals. Therefore, the problem can be set as

$$\mathbf{y}_{\text{bin}} = \mathbf{A}\mathbf{y}_{\text{lin}} + \mathbf{B}\mathcal{D}[\mathbf{y}_{\text{lin}}], \quad (59)$$

where

$$\mathbf{y}_{\text{lin}}(t, f) = \mathbf{D}_{\text{bin}}(f)\mathbf{s}_N(t, f). \quad (60)$$

The mixing matrices \mathbf{A} and \mathbf{B} are then the solution to

$$\mathbf{A}\mathbf{C}_{\text{SH}}\mathbf{A}^H + \mathbf{B}\tilde{\mathbf{C}}_{\text{SH}}\mathbf{B}^H = \mathbf{C}_{\text{target}}, \quad (61)$$

where \mathbf{C}_{SH} is the covariance matrix of the spherical harmonic signals; $\tilde{\mathbf{C}}_{\text{SH}}$ is a diagonal matrix composed of the diagonal elements of the covariance matrix for the decorrelated decoded signals, calculated as

$$\tilde{\mathbf{C}}_{\text{SH}} = \text{diag}\{\mathbb{E}[\mathcal{D}[\mathbf{y}_{\text{lin}}]\mathcal{D}[\mathbf{y}_{\text{lin}}]^H]\}; \quad (62)$$

and $\mathbf{C}_{\text{target}}$ is the target covariance matrix, derived from the analysed parameters, and contains information related to the ICLDs, ICPDs and ICCs which the linearly-decoded ambisonic signals should exhibit.

Firstly, the optimal mixing solution attempts to meet the energies and coherence targets imposed by $\mathbf{C}_{\text{target}}$, via linear mixing of the spherical harmonic signals through matrix \mathbf{A} ; therefore, avoiding decorrelation and the subsequent potential decorrelation artefacts, such as the temporal smearing of transients. However, if the targets cannot be satisfied in this first step, the spherical harmonic signals are decorrelated and then further mixed to the outputs through matrix \mathbf{B} . For example, such a case could arise whereby the sound-field is analysed as being completely diffuse and \mathbf{y}_{lin} fails to meet the appropriate binaural correlations, due to high inter-channel coherence of the ambisonic decoding.

Apart from satisfying the constraint in (61), the solution also aims to minimise phase differences between the enhanced \mathbf{y}_{bin} of (59) and \mathbf{y}_{lin} as much as possible; hence preserving the high single-channel quality of the linear decoding and increasing robustness. The overall structure of the proposed method is presented in Fig. 9. The full solution to (61) can be found in (Vilkamo et al., 2013).

Target covariance matrix definition

The final step to complete the optimal mixing (OM-DirAC) method is to define the target covariance matrix $\mathbf{C}_{\text{target}}$. Assuming S sectors are used in the analysis, S sets of parameters $\mathbf{p}_s = [e_s, \psi_s, \theta_s, \phi_s]$ are extracted. The following assumptions are made:

- a) the energy of the diffuse part for the s th sector is $\psi_s e_s$,

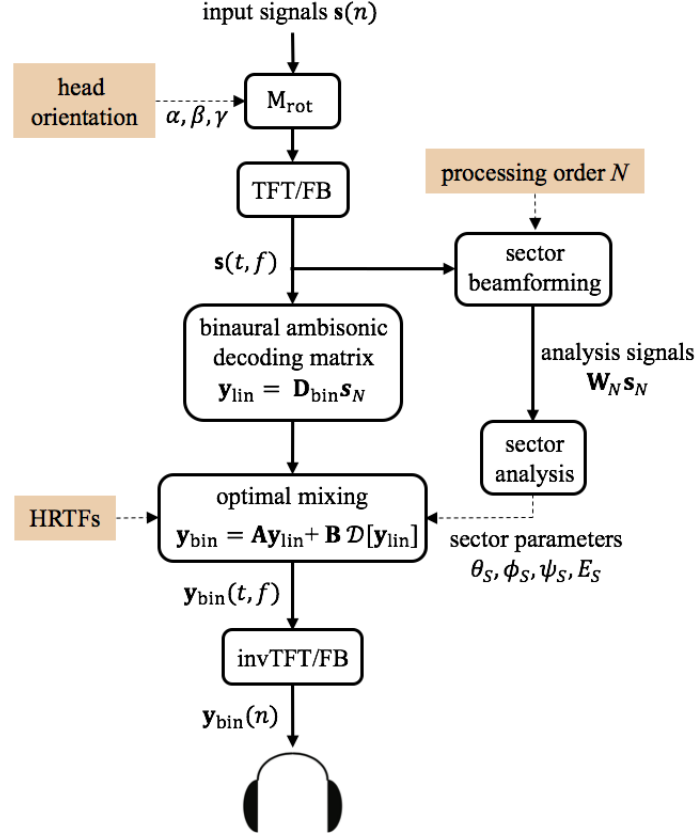


Figure 9: Diagram of the proposed headphone DirAC formulation. Where TFT/FB refers to a time-frequency transform or perfect-reconstruction filterbank.

- b) the energy of the directional part for the s th sector is $(1 - \psi_s)e_s$,
- c) the diffuse components are uncorrelated between sectors, and
- d) the directional components are uncorrelated with the diffuse components.

Based on these assumptions, the covariance matrix of the directional and diffuse components of a single sector can be calculated as

$$\mathbf{C}_{\text{dir}}^{(s)} = (1 - \psi_s)e_s \mathbf{h}(\theta_s, \phi_s) \mathbf{h}^H(\theta_s, \phi_s), \quad (63)$$

$$\mathbf{C}_{\text{diff}}^{(s)} = \psi_s e_s \mathbf{U}, \quad (64)$$

where $\mathbf{h} = [h_L, h_R]^T$ is a vector of HRTFs for the analysed direction θ, ϕ ; matrix \mathbf{U} is a diffuse energy distribution matrix dependent on the reproduction system.

In the case of binaural signals, \mathbf{U} should conform to the binaural coherence curve $c_{\text{bin}}(f)$ under diffuse-field excitation. This coherence curve can be computed from the set of HRTFs, e.g. as proposed in (Politis, 2016), or it can be approximated via a parametric model such as in (Borß and Martin, 2009). A suitable distribution matrix is given here

$$\mathbf{U} = \begin{bmatrix} \alpha & c_{\text{bin}} \\ c_{\text{bin}} & \beta \end{bmatrix}, \quad (65)$$

where α, β are factors that distribute the diffuse energy between the left and right ears where $\alpha + \beta = 1$, and determined by

$$[\alpha, \beta] = \frac{\text{diag} [\tilde{\mathbf{C}}_{\text{SH}}]}{\text{trace} [\tilde{\mathbf{C}}_{\text{SH}}]}. \quad (66)$$

The total target covariance matrix $\mathbf{C}_{\text{target}}$ combines all of the individual sector contributions

$$\mathbf{C}_{\text{target}} = \sum_{s=1}^S \mathbf{C}_{\text{dir}}^{(s)} + \mathbf{C}_{\text{diff}}^{(s)} \quad (67)$$

which essentially contains the binaural ILDs, ITDs and ICs determined by the directional analysis and its definition, along with (61), concludes the proposed solution.

5 Real-time implementations

For the bulk of the thesis project work, real-time implementations of various sound-field analysis and perceptually-motivated reproduction methods were developed and subsequently compared to the proposed formulations in Section 3.2.1 and 4.3.3, for their respective use cases. Therefore, to effectively demonstrate the various sound-field analysis algorithms (including the approach proposed in this thesis), an acoustic camera framework was developed; which utilises a commercially available spherical microphone and spherical camera to produce the necessary power-map overlays and video streams, respectively. Whereas the proposed OM-DirAC formulation was implemented within a binaural ambisonic decoder with head-tracking support.

This section first provides details of the design considerations and the external libraries which were employed in the real-time software implementations. It then describes the real-time implementations in detail; providing examples or listening test results as validation of the implemented algorithms and their relative performance.

Software design considerations

The chosen platform for the real-time implementations was the Virtual Studio Technology (VST) audio plug-in format, which is a standardised application extension that can be utilised by a variety of different digital audio workstations (DAWs); thus, ensuring support for a wide range of hardware configurations and operating systems. Both Windows and MacOSX versions of the software implementations were developed, utilising Visual Studio 2015 and Xcode projects, respectively. Furthermore, the JUCE framework was employed for the graphical user interface (GUI) designs, and written in the C++ programming language. The rationale for selecting this additional framework also included its video camera and Open Sound Control (OSC) message support, which were utilised by the acoustic camera implementation and the ambisonic decoder for head-tracking purposes, respectively.

Since the algorithms presented in this thesis all operate in the spherical harmonic domain, an additional VST audio plug-in was created for the real-time spatial encoding of microphone signals into spherical harmonic signals (see Section 2). Therefore, the three real-time VST plug-ins developed for this thesis work are as follows:

- *Mic2SH*, which takes spherical microphone array signals (namely the commercially available Eigenmike32) and encodes them into spherical harmonic signals.
- *AcCroPaC*, which is an acoustic camera that features the power-map/pseudo-spectrum generating algorithms described in Section 3, including the novel CroPaC reformulation in Section 3.2.1. However, it does not include the algorithm described in Section 3.2.2, as the computational complexity proved too high to operate in real-time.
- *OM-DirAC*, which is a binaural ambisonic decoder that is parametrically enhanced via the novel DirAC formulation that is detailed in Section 4.3.3.

Due to the overlap in implementation requirements and in the interest of efficiency, various standalone software libraries were developed and utilised by the three VST plug-ins; these are as follows:

- *utilib*, which is a separate utility library for various helpful functions, such as memory allocation for multi-dimensional arrays. Additionally, due to the primitive C language support offered by Microsoft’s Visual Compiler (MSVC), a wrapper for complex numbers was required for the same code to compile correctly on both operating systems.
- *shlib*, which is a library containing functions for generating real-valued spherical harmonics and rotation matrices. Additionally, various types of axis-symmetric beamforming weightings were implemented, such as in-phase (Daniel, 2000), maximum energy (Zotter et al., 2012; Moreau et al., 2006) and Dolph-Chebyshev (Rafaely, 2015).
- *cdf4saplib*, which is a C implementation of the covariance domain framework for spatial audio processing, originally presented as a MatLab function in (Vilkamo et al., 2013), utilised for the optimal-mixing solution in the OM-DirAC implementation.

Optimisations

The internal algorithms for the software implementations were all written in the C programming language and underwent rigorous optimisations in order to operate in real-time. For example, the linear algebra operations that are used prolifically throughout the thesis work, were all written to conform to the basic linear algebra subroutines (BLAS) standard. Whereas, operations such as the lower-upper (LU) factorisation and singular value decomposition (SVD) were addressed by the linear algebra package (LAPACK) standard. The optimised variants of these aforementioned libraries which were utilised in this thesis work, were Apple’s Accelerate framework and Intel’s Mathematics Kernal Library (MKL), for the Mac OSX and Windows versions, respectively.

Time-frequency transform

The time-frequency transform selected for the real-time implementations is based on the STFT (see Section 2), and utilises analysis and synthesis windows which are optimised to suppress temporal aliasing; note that the source code of the implementation can be found in (Vilkamo, 2015). The temporal resolution of the transform was determined by a hop size of 2.7 msec (128 samples at 48kHz sample rate).

The chosen time-frequency transform initially provides a uniform resolution of 128 bands, before employing additional filters which increase the low-frequency resolution, resulting in a total of 133 bands with centre frequencies which are roughly equivalent to the Bark scale at the lower frequencies (Pulkki and Karjalainen, 2015). Note that this practice is similar to the hybrid-QMF filterbanks utilised by many spatial

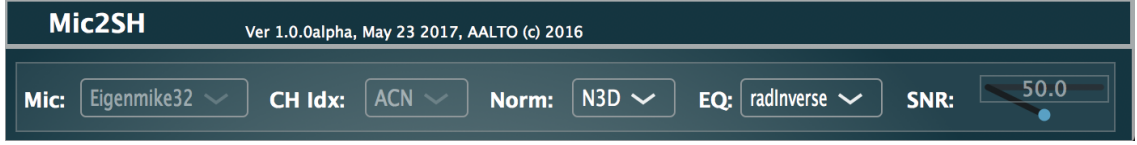


Figure 10: The Mic2SH VST user interface.

audio codecs (Schuijers et al., 2004) and is popular with implementations of other perceptually-motivated time-frequency domain algorithms.

5.1 Mic2SH

The first VST implemented was the microphone to spherical harmonic signals (Mic2SH) converter, which spatially encodes the microphone signals from the Eigenmike32 by utilising the transformation matrices described in Section 2; the GUI for the plug-in is shown in Fig. 10.

The plug-in also performs the necessary equalisation required to mitigate the effect that a rigid sphere has on the initial estimate of the pressure on the sphere, which can be first determined by calculating the theoretical modal coefficients as (Rafaely, 2015)

$$b_n(kr) = 4\pi j^n [j_n(kr) - \frac{j'_n(kr_a)}{h_n^{(2)'}(kr_a)} h_n^{(2)}(kr)], \quad (68)$$

where j_n is a spherical Bessel function of the first kind and order n ; $h_n^{(2)}$ is a spherical Hankel function of the second kind and order n ; r is the radius of the sphere; and $k = 2\pi f/c$ is the wave number.

This influence of the rigid sphere on the spherical harmonic signals estimate can be removed by inverting these b_n coefficients and applying them to the encoded signals. However, it is important to note that this approach would result in a large amplification in microphone noise, especially at the lower frequencies. Therefore, a regularised inversion is recommended [such as (13)], where a regularisation parameter λ can be applied in order to make a compromise between noise amplification and the accuracy of the transform. These regularised inverted modal coefficients can then be utilised in (12), to complete the spatial encoding. Note that different regularisation strategies common in the literature have been implemented in Mic2SH, such as the Tikhonov-based regularised inversion (Moreau et al., 2006) and soft limiting (Bernschütz et al., 2011). The plug-in can also accommodate the various spherical harmonic format conventions which are popular today, such as the SID/ACN channel orderings and the N3D/SN3D normalisation schemes.



Figure 11: The spherical microphone array (Eigenmike32) and spherical camera (RICOH Theta S) orientation.

5.2 AcCroPaC

Various scanning beamformers described in Section 3.1 and the proposed side-lobe suppressed CroPaC formulation detailed in Section 3.2.1, were introduced into the second VST plug-in, AcCroPaC¹. This plug-in relies on spherical harmonic signals and spherical video stream as input, for which the commercially available Eigenmike32 and RICOH Theta S were utilised, respectively. The set-up is depicted in Fig. 11.

The algorithms within the acoustic camera have been generalised to support spherical harmonic signals up to 7th order and can be optionally generated using the Mic2SH VST described above. The overall block diagram of the system is shown in Fig. 12. The time-domain microphone spherical harmonic signals are first transformed into the time-frequency domain. For computational efficiency the spherical harmonic signals are rotated after the time-frequency transform towards all the points defined by the grid, in the case of the proposed CroPaC algorithm. These signals are then fed into either beamforming units or pseudo-spectrum generators. For the proposed algorithm, the cross-spectrum based analysis parameter is estimated for each grid point by using the spherical harmonics of the maximum available order, and the order before it. Please note that when the side-lobe suppression mode is enabled, one parameter is estimated per roll and the resulting parameters are multiplied.

The user-interface consists of a view window, and a parameter editor (see Fig. 13). The view window displays the camera feed and overlays the user selected power-map in real-time. The field-of view (FOV) and the aspect ratio are also user definable, which accommodates a wide array of different web-cam devices. Additionally, the image frames from the camera can be optionally mirrored using an appropriate affine transformation (left-right, up-down etc.); in order to accommodate a variety

¹Special thanks is extended to Symeon Delikaris-Manias for his help during the design and development of the AcCroPaC VST plug-in.

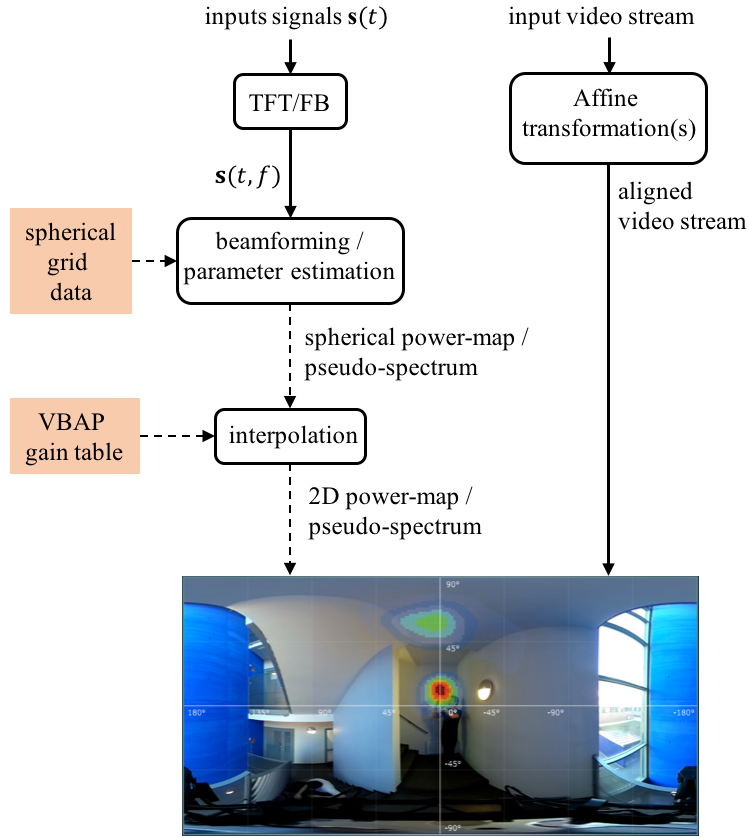


Figure 12: Block diagram for the AcCroPaC VST internal processing; where TFT/FB refers to an STFT or perfect reconstruction filter-bank.

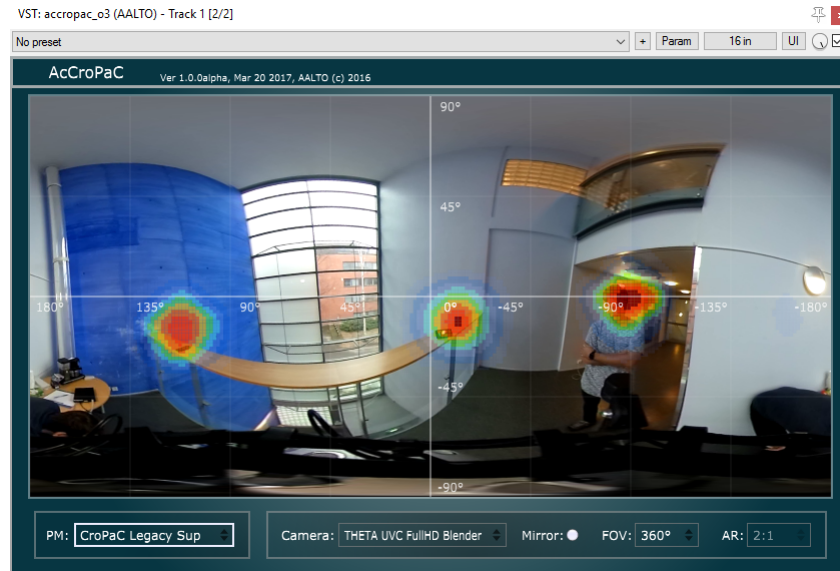


Figure 13: The AcCroPaC VST user interface. Third-order spherical harmonic signals and the proposed CroPaC formulation were utilised in a reverberant environment.

of different camera orientations.

Processing modes and sampling grids

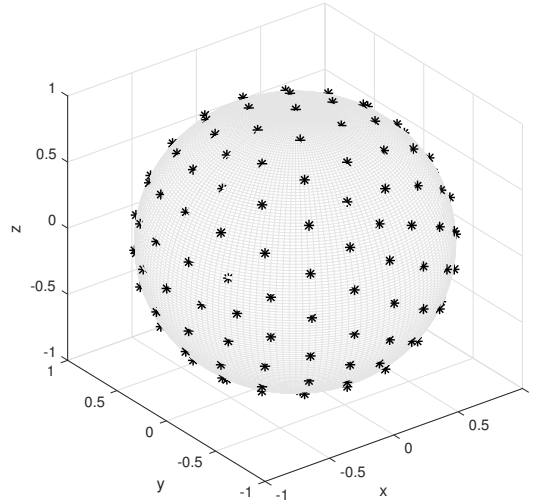
The AcCroPaC VST generates the power-map/pseudo-spectrums by sampling the sphere via a spherical grid. A pre-computed almost-uniform spherical grid was chosen for the implementation, which provides 252 nearly-uniformly distributed data points on the sphere. The grid is based on the 3LD library (Hollerweger, 2006), where the points are generated by utilising geodesic spheres. This is performed by tessellating the facets of a polyhedron and extending them to the radius of the original polyhedron. The intersection points between them are the points on the spherical grid. The two power-map approaches and two pseudo-spectrum methods described in Section 3 were implemented within the plug-in: conventional signal-independent beamformers (PWD, minimum side-lobe, maximum energy and Dolph-Chebyshev) and signal-dependent MVDR beamformers; and the MUSIC and proposed cross-spectrum-based methods. The power-map/pseudo-spectrum values are then summed over the analysis frequency bands and averaged over time slots using a one-pole filter

$$\hat{\mathbf{P}}_{\text{MAP}}(\Omega_j, t - 1) = \alpha \hat{\mathbf{P}}_{\text{MAP}}(\Omega_j, t) + (1 - \alpha) \mathbf{P}_{\text{MAP}}(\Omega_j, t), \quad (69)$$

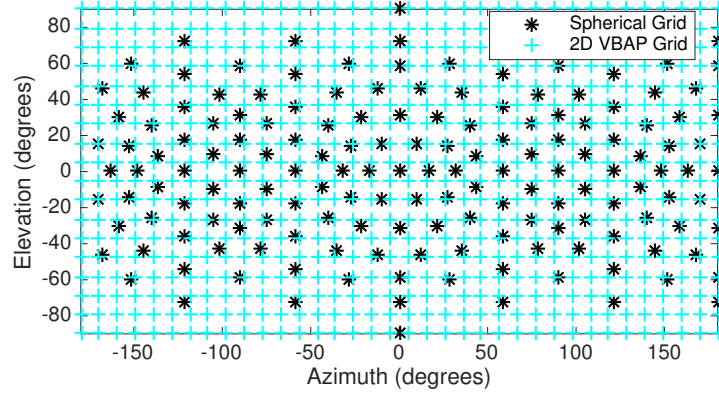
where $\alpha \in [0, 1]$ is the smoothing parameter. The spherical power-map values are then interpolated to attain a 2D power-map, using pre-computed VBAP gains. The spherical and interpolated grids are shown in Fig. 14. These 2D power-maps are then further interpolated using bi-cubic interpolation depending on the display settings and are normalised such that $\hat{\mathbf{P}}_{\text{MAP}} \in [0, 1]$. The pixels that correspond to the 2D interpolated results are then coloured appropriately, such that red indicates high energy and blue indicates low energy. Additionally, the transparency factor is gradually increased for the lower values to ensure that they do not unnecessarily detract from the video stream.

5.2.1 Examples

Examples of power-maps/pseudo-spectrums are shown in Fig. 15 for four different modes: PWD beamformer, MVDR, MUSIC and the technique proposed in this thesis. Fourth order spherical harmonic signals were generated using the Mic2SH VST plug-in, and these were used for all four methods. The video was unwrapped using the software provided by RICOH and then combined with the calculated power-map/pseudo-spectrum. Since the camera is not at the same position as the microphone array, a calibration process is performed to align the power map with the image. Note, however, this will affect sources that are very close to the array. The resulting acoustic camera outputs are shown for two different recording scenarios: a staircase of high reverberation time of approximately 2 seconds (Fig. 15, e-h) and a corridor with slightly shorter reverberation time (Fig. 15, a-d), approximately 1.5 seconds.



(a) Nearly-uniform spherical grid on a unit sphere.



(b) Nearly-uniform spherical grid plotted on a 2D equirectangular plane (black colour) and a 2D VBAP interpolated grid overlay (cyan colour).

Figure 14: Spherical and interpolated grids.

It can be seen from Fig. 15(a-d) that there is one direct source and at least one prominent early reflection. However, in the case of PWD, the distinction between the two paths is the least clear, and also erroneously indicates that the sources are spatially larger than they actually are. The distinction between the two paths is improved slightly when using MVDR beamformers, which is improved further when using MUSIC. In the case of the proposed technique, the two paths shown in the other three power-map modes are now isolated completely, and a second early reflection with lower energy is now visible; which is not as evident in the other three methods. The PWD algorithm also shows evidence of an erroneous sound source, which is probably the result of the side-lobes pointing towards the real sound source; thus, highlighting the importance of side-lobe suppression for acoustic camera

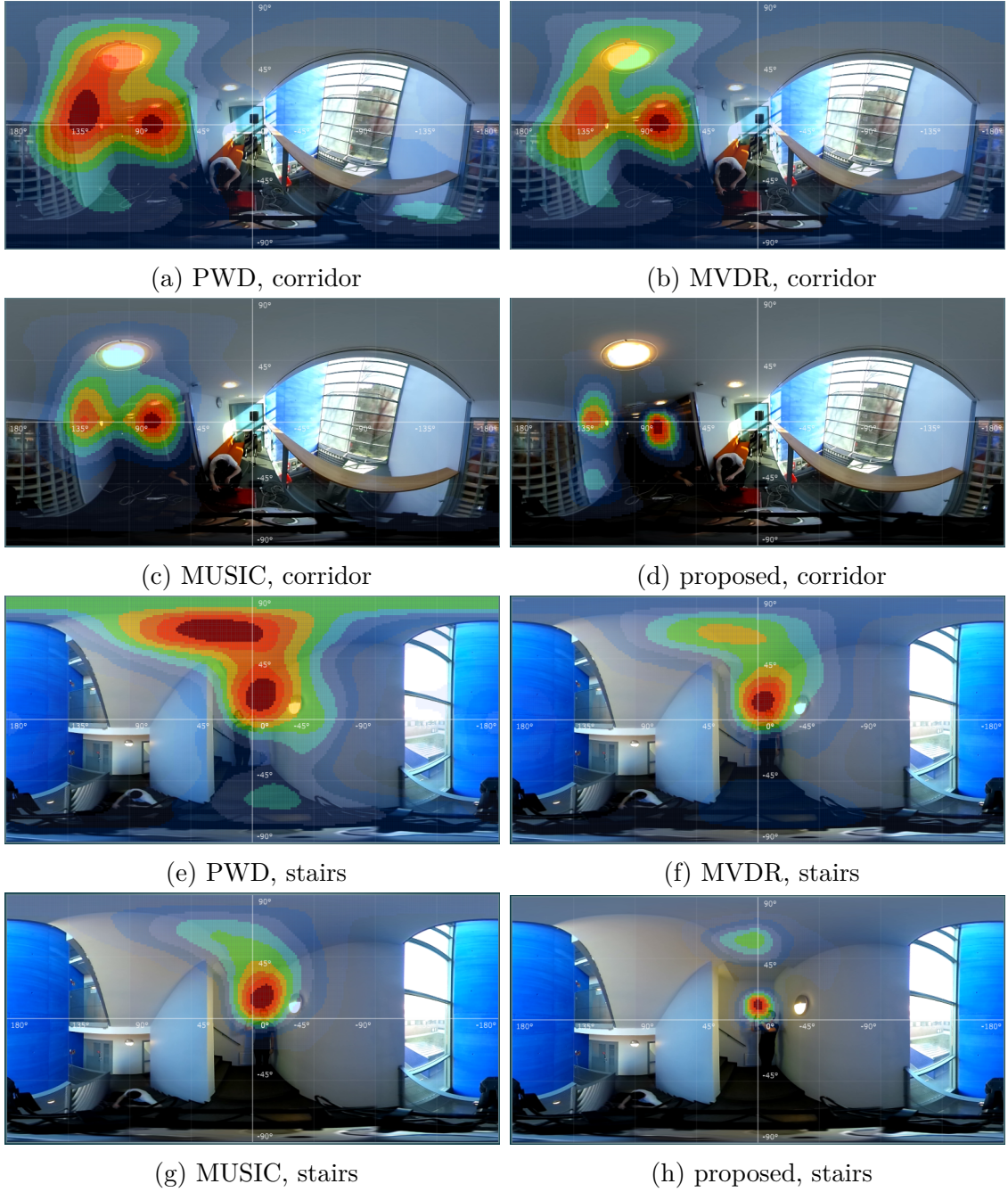


Figure 15: Images of the acoustic camera VST, while using fourth order spherical harmonic signals and the four processing modes in reverberant environments.

applications. The images in Fig. 15(e-h) indicate a similar performance; however, in the case of MUSIC, the ceiling reflection is more difficult to distinguish as a separate entity.

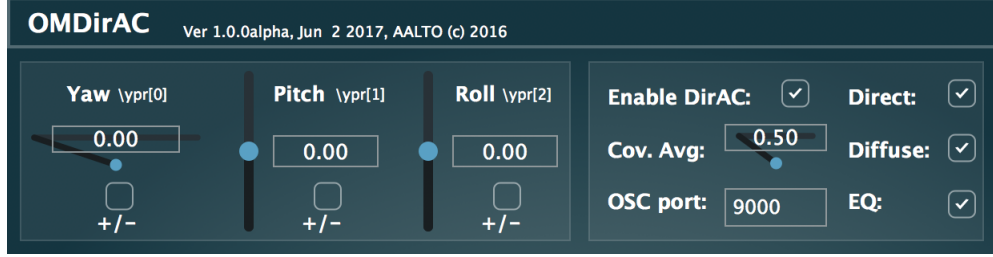


Figure 16: The OM-DirAC decoder user interface.

5.3 OM-DirAC

A real-time implementation of the binaural DirAC formulation proposed in Section 4.3.3 was developed as a VST audio plug-in, OM-DirAC². The implementation supports spherical harmonic signals up to 7th-order and head-tracking. Contrary to the rotation of the analysed DoAs, as was previously carried out in the VL-DirAC implementation [detailed in (Laitinen and Pulkki, 2009)] it was found to be more efficient to rotate the spherical harmonic signals directly, due to the faster update rates. The rotation angles are updated at every analysis frame.

The spatial parameters \mathbf{p}_N for (55) are obtained instantaneously to capture rapid variations of the sound scene, while the definition of input and target covariance matrices \mathbf{C}_{lin} , $\mathbf{C}_{\text{model}}$ are computed across multiple windows to capture and provide meaningful signal statistics for a robust synthesis. The beamforming coefficients \mathbf{W}_N in (55), the ambisonic decoding matrix \mathbf{D}_{bin} and the binaural coherence matrix \mathbf{U} for (65) are precomputed and stored within the plug-in. The decoding matrix \mathbf{D}_{bin} is computed using a densely measured set of HRTFs measured at Aalto University. The decoding filters and HRTFs are pre-processed with the same time-frequency transform as the one applied during runtime and converted to spectral coefficients. During the construction of the target covariance matrix, the HRTF vector \mathbf{h} in (63) is interpolated from the measured set to the respective analysed direction (θ_s, ϕ_s) using triangular interpolation on the measurement grid (Gamper, 2013).

The GUI is depicted in Fig. 16, which allows the user to manipulate the rotation angles via sliders or by utilising an external head-tracker and OSC messages. It also offers control over the degree of averaging performed in the calculation of the input spherical harmonic signals covariance matrix (via a one-pole filter) and allows the user to disable the direct and/or diffuse component contributions to the target covariance matrix. The OM-DirAC enhancement can also be disabled entirely, which means that the processing reverts back to a binaural ALLRAD ambisonic decoder.

5.3.1 Evaluation

A multiple-stimulus listening test was conducted in order to assess the performance of the OM-DirAC implementation.

²Special thanks is extended to Archontis Politis for his help during the design and development of the OM-DirAC VST plug-in.

Identifier	Description
groove_dry	Four instruments in a free-field
groove_small	Four instruments in a small room
mix_dry	Female speech, fountain, piano, claps/ free-field
mix_small	Female speech, fountain, piano, claps/ small room
speech_large	Female speech in front in a large hall

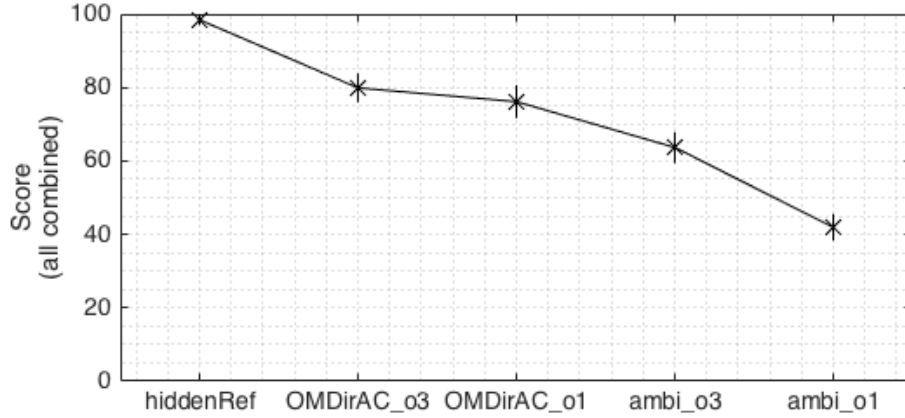
Table 1: Sound scenes identifiers.

Identifier	Description
hiddenRef	Reference 28-channel loudspeaker signals convolved with HRTFs
OMDirAC_o3	Third-order OM-DirAC
OMDirAC_o1	First-order OM-DirAC
ambi_o3	Third-order ambisonic decoding \mathbf{D}_{bin}
ambi_o1	First-order ambisonic decoding \mathbf{D}_{bin}

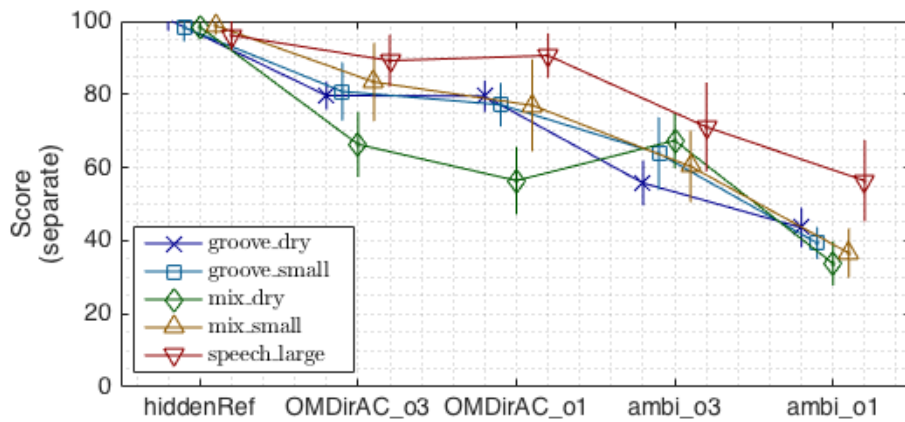
Table 2: Test cases identifiers.

Five synthetic sound scenes were simulated with a varying number of sound sources in anechoic and reverberant environments (see Table. 1). Room reverberation was simulated with the image source method. All direct paths and image sources were quantised directly to 28 plane wave signals covering the sphere, without employing panning. These 28-channel signals were played back in an anechoic chamber through real loudspeakers from their corresponding directions, to assess the naturalness of the synthetic scenes. The 28-channel scenes served as a reference to assess the different methods and were specifically designed to be critical of basic parametric analysis. There are two free-field sound scenes, labelled here as *groove_dry* and *mix_dry*. The former consists of individual dry recordings of a band distributed on the front hemisphere horizontally, while the latter incorporates clapping, a fountain, piano and female speech, with three of the sound sources placed horizontally and one above the listener. *groove_small* and *mix_small*, are the reverberant versions of their free-field counterparts. The final sample *speech_large*, comprises of female speech in front of the listener simulated in a large hall.

The reference test cases *hiddenRef*, were obtained by convolving the 28-channel signals with their respective HRTFs and summing the resulting binaural signals. Ambisonic encoders were applied to each of the 28-channel sound scenes, in order to obtain both first-order and third-order spherical harmonic signals. The sum of the omnidirectional signals served as a low-quality anchor. The test cases for both first and third-order ambisonics and OM-DirAC *ambi_o1*, *ambi_o3*, *OMDirAC_o1*, *OMDirAC_o3*, were obtained by passing the corresponding spherical harmonic signals through their respective off-line decoders (see Table. 2). Note that the standalone Ambisonic decoders are identical to the one used within the OM-DirAC implementation.



(a) results averaged across all sound scenes.



(b) results for each sound scene.

Figure 17: The means and 95 % confidence intervals of the listening test results.

The test subjects were instructed to rate the test case they perceived to be closest to the reference, in terms of *overall quality* and *spatial accuracy*, as 100; to rate the test case furthest from the reference as 0; and to rate the remaining four test cases relative to each-other, the reference and anchor. Since previous studies (Vilkamo et al., 2009; Laitinen and Pulkki, 2009; Politis, Laitinen, Ahonen and Pulkki, 2015; Politis, Vilkamo and Pulkki, 2015), have found that lower-order ambisonics may colour the output spectrum compared to the reference, all of the test cases were equalised to spectrally match their reference; thus, reducing the likelihood of large variances in the results due to the easily remedied spectral differences between methods.

There were 13 test participants in total. It can be seen that for the majority of sound scenes (Fig. 17b) the first and third-order variants of OM-DirAC are perceived as being closer to the reference (in terms of overall quality and spatial accuracy) when compared to their respective first and third-order variants of ambisonics. However, it is evident in the *mix_dry* sound scene that spatial artefacts induced by the lack of individual sectors in *OMDirAC_o1* have negatively impacted scores; although, the

scores are still significantly higher than that of the *ambi_o1* test case, while using the same first-order spherical harmonic signals. It can also be seen that the increased number of sectors in the third-order case *OMDirAC_o3*, has reduced spatial artefacts to a certain extent; however, the performance is not significantly different to the *ambi_o3* test case for the *mix_dry* sound scene. Regardless, it must be stressed that this particular sound scene does not represent a likely recording scenario and most recordings will contain some degree of reverberation, which can be seen to mask these spatial artefacts to some degree in the reverberant counterpart *mix_small*.

6 Conclusion

This thesis has provided details regarding real-time software implementations for spatially encoding microphone signals, and for visualising and auralising spatial sound-fields at the listening position. These systems were realised as an acoustic camera framework and a DirAC enhanced ambisonic decoder, respectively, for which novel reformulations of the CroPaC and DirAC algorithms have been integrated.

Regarding the acoustic camera, the novel coherence-based parameter with additional suppression of the side-lobes, represents an intuitive approach for visualising sound-fields. It also yields a reduction in implementation and computational requirements when compared to MVDR or MUSIC, as it does not rely on lower-upper decompositions, Gaussian Elimination, or singular value decompositions. As is demonstrated in the simple test recording scenarios, this proposed method can be inherently tolerant to reverberation, providing greater spatial selectivity than the other methods explored in this thesis.

Regarding the novel DirAC formulation for head-tracked headphone playback, this proposed algorithm represents a clear improvement over existing implementations by reducing the computational requirements and artefacts arising from model mismatch; while also improving the robustness and overall perceived spatial accuracy. According to the listening test results, based on comparisons with a binaural reference, the method described outperforms first-order ambisonic decoding for all tested sound scenes and third-order ambisonic decoding in a number of cases, while using only first-order spherical harmonic signals. When using third-order spherical harmonic signals, the method is improved further and performs better than ambisonic decoding for all cases bar one; attaining scores which more closely match the reference.

The End

References

- Alon, D. L., Sheaffer, J. and Rafaely, B. (2015), ‘Robust plane-wave decomposition of spherical microphone array recordings for binaural sound reproduction’, *The Journal of the Acoustical Society of America* **138**(3), 1925–1926.
- Atkins, J. (2011), Robust beamforming and steering of arbitrary beam patterns using spherical arrays, *in* ‘Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on’, IEEE, pp. 237–240.
- Balanis, C. A. and Ioannides, P. I. (2007), ‘Introduction to smart antennas’, *Synthesis Lectures on Antennas* **2**(1), 1–175.
- Barrett, N. and Berge, S. (2010), A new method for b-format to binaural transcoding, *in* ‘Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space’, Audio Engineering Society.
- Begault, D. R. and Trejo, L. J. (2000), ‘3-d sound for virtual reality and multimedia’.
- Bernschütz, B., Giner, A. V., Pörschmann, C. and Arend, J. (2014), ‘Binaural reproduction of plane waves with reduced modal order’, *Acta Acustica united with Acustica* **100**(5), 972–983.
- Bernschütz, B., Pörschmann, C., Spors, S., Weinzierl, S. and der Verstärkung, B. (2011), ‘Soft-limiting der modalen amplitudenverstärkung bei sphärischen mikrofonarrays im plane wave decomposition verfahren’, *Proceedings of the 37. Deutsche Jahrestagung für Akustik (DAGA 2011)* pp. 661–662.
- Bertet, S., Daniel, J., Parizet, E. and Warusfel, O. (2013), ‘Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources’, *Acta Acustica united with Acustica* **99**(4), 642–657.
- Blanco, M. A., Flórez, M. and Bermejo, M. (1997), ‘Evaluation of the rotation matrices in the basis of real spherical harmonics’, *Journal of Molecular Structure: THEOCHEM* **419**(1), 19–27.
- Blauert, J. (1997), *Spatial hearing: the psychophysics of human sound localization*, MIT press.
- Borß, C. and Martin, R. (2009), An improved parametric model for perception-based design of virtual acoustics, *in* ‘Audio Engineering Society Conference: 35th International Conference: Audio for Games’, Audio Engineering Society.
- Braun, S. and Frank, M. (2011), Localization of 3d ambisonic recordings and ambisonic virtual sources, *in* ‘1st International Conference on Spatial Audio,(Detmold)’.
- Capon, J. (1969), ‘High-resolution frequency-wavenumber spectrum analysis’, *Proceedings of the IEEE* **57**(8), 1408–1418.

- Daniel, J. (2000), Représentation de champs acoustiques, application à la reproduction et à la transmission de scènes sonores complexes dans un contexte multimédia, PhD thesis, Ph. D. thesis, University of Paris 6, Paris, France.
- Davis, L. S., Duraiswami, R., Grassi, E., Gumerov, N. A., Li, Z. and Zotkin, D. N. (2005), High order spatial audio capture and its binaural head-tracked playback over headphones with hrtf cues, *in* ‘Audio Engineering Society Convention 119’, Audio Engineering Society.
- Delikaris-Manias, S., Pavlidi, D., Mouchtaris, A. and Pulkki, V. (2017), Doa estimation with histogram analysis of spatially constrained active intensity vectors, *in* ‘Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on’, IEEE, pp. 526–530.
- Delikaris-Manias, S., Pavlidi, D., Pulkki, V. and Mouchtaris, A. (2016), 3d localization of multiple audio sources utilizing 2d doa histograms, *in* ‘Signal Processing Conference (EUSIPCO), 2016 24th European’, IEEE, pp. 1473–1477.
- Delikaris-Manias, S. and Pulkki, V. (2013), ‘Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays’, *IEEE Transactions on Audio, Speech, and Language Processing* **21**(11), 2356–2367.
- Delikaris-Manias, S., Vilkamo, J. and Pulkki, V. (2016), ‘Signal-dependent spatial filtering based on weighted-orthogonal beamformers in the spherical harmonic domain’, *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* **24**(9), 1507–1519.
- Epain, N. and Jin, C. T. (2013), Super-resolution sound field imaging with sub-space pre-processing, *in* ‘Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on’, IEEE, pp. 350–354.
- Erić, M. M. (2011), Some research challenges of acoustic camera, *in* ‘Telecommunications Forum (TELFOR), 2011 19th’, IEEE, pp. 1036–1039.
- Farina, A., Amendola, A., Capra, A. and Varani, C. (2011), Spatial analysis of room impulse responses captured with a 32-capsule microphone array, *in* ‘Audio Engineering Society Convention 130’, Audio Engineering Society.
- Gamper, H. (2013), ‘Head-related transfer function interpolation in azimuth, elevation, and distance’, *The Journal of the Acoustical Society of America* **134**(6), EL547–EL553.
- Gerzon, M. A. (1973), ‘Periphony: With-height sound reproduction’, *Journal of the Audio Engineering Society* **21**(1), 2–10.
- Haas, H. (1951), ‘Über den einfluß eines einfachechos auf die hörsamkeit von sprache’, *Acta Acustica united with Acustica* **1**(2), 49–58.

- Hardin, R. H. and Sloane, N. J. (1996), ‘McLaren’s improved snub cube and other new spherical designs in three dimensions’, *Discrete & Computational Geometry* **15**(4), 429–441.
- Hollerweger, F. (2006), *Periphonic sound spatialization in multi-user virtual environments*, Citeseer.
- Ivanic, J. and Ruedenberg, K. (1996), ‘Rotation matrices for real spherical harmonics. direct determination by recursion’, *The Journal of Physical Chemistry* **100**(15), 6342–6347.
- Ivanic, J. and Ruedenberg, K. (1998), ‘Rotation matrices for real spherical harmonics. direct determination by recursion’, *The Journal of Physical Chemistry A* **102**(45), 9099–9100.
- Kearney, G., Gorzel, M., Rice, H. and Boland, F. (2012), ‘Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields’, *Acta Acustica united with Acustica* **98**(1), 61–71.
- Laitinen, M.-V. and Pulkki, V. (2009), Binaural reproduction for directional audio coding, in ‘Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA’09. IEEE Workshop on’, IEEE, pp. 337–340.
- Litovsky, R. Y., Colburn, H. S., Yost, W. A. and Guzman, S. J. (1999), ‘The precedence effect’, *The Journal of the Acoustical Society of America* **106**(4), 1633–1654.
- Lösler, S. and Zotter, F. (2015), ‘Comprehensive radial filter design for practical higher-order ambisonic recording’, *Fortschritte der Akustik, DAGA* pp. 452–455.
- Melchior, F., Thiergart, O., Del Galdo, G., de Vries, D. and Brix, S. (2009), Dual radius spherical cardioid microphone arrays for binaural auralization, in ‘Audio Engineering Society Convention 127’, Audio Engineering Society.
- Moore, B. C. (2012), *An introduction to the psychology of hearing*, Brill.
- Moreau, S., Daniel, J. and Bertet, S. (2006), 3d sound field recording with higher order ambisonics—objective measurements and validation of a 4th order spherical microphone, in ‘120th Convention of the AES’, pp. 20–23.
- Nadiri, O. and Rafaely, B. (2014), ‘Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test’, *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* **22**(10), 1494–1505.
- Nguyen, T. Q. (1994), ‘Near-perfect-reconstruction pseudo-qmf banks’, *IEEE Transactions on signal processing* **42**(1), 65–76.

- Nguyen, T. Q. and Vaidyanathan, P. (1989), ‘Two-channel perfect-reconstruction fir qmf structures which yield linear-phase analysis and synthesis filters’, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **37**(5), 676–690.
- Nielsen, R. O. (1991), *Sonar signal processing*, Artech House, Inc.
- Noisternig, M., Musil, T., Sontacchi, A. and Holdrich, R. (2003), 3d binaural sound reproduction using a virtual ambisonic approach, *in* ‘Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2003. VECIMS’03. 2003 IEEE International Symposium on’, IEEE, pp. 174–178.
- Noohi, T., Epain, N. and Jin, C. T. (2013), Direction of arrival estimation for spherical microphone arrays by combination of independent component analysis and sparse recovery, *in* ‘Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on’, IEEE, pp. 346–349.
- O’Donovan, A., Duraiswami, R. and Zotkin, D. (2008), Imaging concert hall acoustics using visual and audio cameras, *in* ‘Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on’, IEEE, pp. 5284–5287.
- Pavlidis, D., Delikaris-Manias, S., Pulkki, V. and Mouchtaris, A. (2015), 3d localization of multiple sound sources with intensity vector estimates in single source zones, *in* ‘Signal Processing Conference (EUSIPCO), 2015 23rd European’, IEEE, pp. 1556–1560.
- Pavlidis, D., Delikaris-Manias, S., Pulkki, V. and Mouchtaris, A. (2016), 3d doa estimation of multiple sound sources based on spatially constrained beamforming driven by intensity vectors, *in* ‘Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on’, IEEE, pp. 96–100.
- Politis, A. (2016), ‘Diffuse-field coherence of sensors with arbitrary directional responses’, *arXiv preprint arXiv:1608.07713*.
- Politis, A., Delikaris-Manias, S. and Pulkki, V. (2015), Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays, *in* ‘Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on’, IEEE, pp. 6–10.
- Politis, A., Laitinen, M.-V., Ahonen, J. and Pulkki, V. (2015), ‘Parametric spatial audio processing of spaced microphone array recordings for multichannel reproduction’, *Journal of the Audio Engineering Society* **63**(4), 216–227.
- Politis, A. and Poirier-Quinot, D. (2016), ‘Jsambisonics: A web audio library for interactive spatial sound processing on the web’.
- Politis, A., Vilkkamo, J. and Pulkki, V. (2015), ‘Sector-based parametric sound field reproduction in the spherical harmonic domain’, *IEEE Journal of Selected Topics in Signal Processing* **9**(5), 852–866.

- Pulkki, V. (1997), ‘Virtual sound source positioning using vector base amplitude panning’, *Journal of the Audio Engineering Society* **45**(6), 456–466.
- Pulkki, V. (2006), Directional audio coding in spatial sound reproduction and stereo upmixing, *in* ‘Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond’, Audio Engineering Society.
- Pulkki, V. (2007), ‘Spatial sound reproduction with directional audio coding’, *Journal of the Audio Engineering Society* **55**(6), 503–516.
- Pulkki, V. and Karjalainen, M. (2015), *Communication acoustics: an introduction to speech, audio and psychoacoustics*, John Wiley & Sons.
- Pulkki, V., Lokki, T. and Rocchesso, D. (2011), ‘Spatial effects’, *DAFX: Digital Audio Effects, Second Edition* pp. 139–183.
- Pulkki, V., Politis, A., Del Galdo, G. and Kuntz, A. (2013), Parametric spatial audio reproduction with higher-order b-format microphone input, *in* ‘Audio Engineering Society Convention 134’, Audio Engineering Society.
- Rafaely, B. (2015), *Fundamentals of spherical array processing*, Vol. 8, Springer.
- Rafaely, B. and Kleider, M. (2008), ‘Spherical microphone array beam steering using wigner-d weighting’, *IEEE Signal Processing Letters* **15**, 417–420.
- Rayleigh, L. (1907), ‘On the dynamical theory of gratings’, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **79**(532), 399–416.
- Santala, O., Vertanen, H., Pekonen, J., Oksanen, J. and Pulkki, V. (2009), Effect of listening room on audio quality in ambisonics reproduction, *in* ‘Audio Engineering Society Convention 126’, Audio Engineering Society.
- Schmidt, R. (1986), ‘Multiple emitter location and signal parameter estimation’, *IEEE transactions on antennas and propagation* **34**(3), 276–280.
- Schörkhuber, C., Zotter, F. and Höldrich, R. (2017), ‘Signal-dependent encoding for first-order ambisonic microphones’, *Fortschritte der Akustik, DAGA, Kiel*.
- Schuijers, E., Breebaart, J., Purnhagen, H. and Engdegard, J. (2004), Low complexity parametric stereo coding, *in* ‘Audio Engineering Society Convention 116’, Audio Engineering Society.
- Shabtai, N. R. and Rafaely, B. (2013), Binaural sound reproduction beamforming using spherical microphone arrays, *in* ‘Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on’, IEEE, pp. 101–105.
- Sheeline, C. W. (1983), An investigation of the effects of direct and reverberant signal interaction on auditory distance perception, PhD thesis, Stanford University.

- Smith, J. O. (2011), *Spectral audio signal processing*, W3K publishing.
- Solvang, A. (2008), ‘Spectral impairment of two-dimensional higher order ambisonics’, *Journal of the Audio engineering Society* **56**(4), 267–279.
- Stitt, P., Bertet, S. and van Walstijn, M. (2014), ‘Off-centre localisation performance of ambisonics and hoa for large and small loudspeaker array radii’, *Acta Acustica united with Acustica* **100**(5), 937–944.
- Toole, F. E. (2008), *Sound reproduction: Loudspeakers and rooms*, Taylor & Francis.
- Vaidyanathan, P. P. and Hoang, P.-Q. (1988), ‘Lattice structures for optimal design and robust implementation of two-channel perfect-reconstruction qmf banks’, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **36**(1), 81–94.
- Valimaki, V., Parker, J. D., Savioja, L., Smith, J. O. and Abel, J. S. (2012), ‘Fifty years of artificial reverberation’, *IEEE Transactions on Audio, Speech, and Language Processing* **20**(5), 1421–1448.
- Vilkamo, J. (2015), ‘Alias-free short-time Fourier Transform’, <https://github.com/jvilkamo/afSTFT>.
- Vilkamo, J., Bäckström, T. and Kuntz, A. (2013), ‘Optimized covariance domain framework for time–frequency processing of spatial audio’, *Journal of the Audio Engineering Society* **61**(6), 403–411.
- Vilkamo, J. and Delikaris-Manias, S. (2015), ‘Perceptual reproduction of spatial sound using loudspeaker-signal-domain parametrization’, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **23**(10), 1660–1669.
- Vilkamo, J., Lokki, T. and Pulkki, V. (2009), ‘Directional audio coding: Virtual microphone-based synthesis and subjective evaluation’, *Journal of the Audio Engineering Society* **57**(9), 709–724.
- Vilkamo, J. and Pulkki, V. (2013), ‘Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering’, *Journal of the Audio Engineering Society* **61**(9), 637–646.
- Von Békésy, G. and Wever, E. G. (1960), *Experiments in hearing*, Vol. 8, McGraw-Hill New York.
- Vorländer, M. (2007), *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*, Springer Science & Business Media.
- Wiggins, B., Paterson-Stephens, I. and Schillebeeckx, P. (2001), The analysis of multi-channel sound reproduction algorithms using hrtf data., in ‘Audio Engineering Society Conference: 19th International Conference: Surround Sound-Techniques, Technology, and Perception’, Audio Engineering Society.

- Xie, B. (2013), *Head-related transfer function and virtual auditory display*, J. Ross Publishing.
- Yang, L. and Bosun, X. (2014), Subjective evaluation on the timbre of horizontal ambisonics reproduction, *in* ‘Audio, Language and Image Processing (ICALIP), 2014 International Conference on’, IEEE, pp. 11–15.
- Zoltowski, M. D. (1988), ‘On the performance analysis of the mvdr beamformer in the presence of correlated interference’, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **36**(6), 945–947.
- Zotter, F. and Frank, M. (2012), ‘All-round ambisonic panning and decoding’, *Journal of the audio engineering society* **60**(10), 807–820.
- Zotter, F., Pomberger, H. and Noisternig, M. (2012), ‘Energy-preserving ambisonic decoding’, *Acta Acustica united with Acustica* **98**(1), 37–47.