PARAMETRIC ACOUSTIC CAMERA FOR REAL-TIME SOUND CAPTURE, ANALYSIS AND TRACKING

Leo McCormack, Symeon Delikaris-Manias and Ville Pulkki *

Acoustics Lab, Dept. of Signal Processing and Acoustics
School of Electrical Engineering
Aalto University, Espoo, FI-02150, Finland
leo.mccormack@aalto.fi

ABSTRACT

This paper details a software implementation of an acoustic camera, which utilises a spherical microphone array and a spherical camera. The software builds on the Cross Pattern Coherence (CroPaC) spatial filter, which has been shown to be effective in reverberant and noisy sound field conditions. It is based on determining the cross spectrum between two coincident beamformers. The technique is exploited in this work to capture and analyse sound scenes by estimating a probability-like parameter of sounds appearing at specific locations. Current techniques that utilise conventional beamformers perform poorly in reverberant and noisy conditions, due to the side-lobes of the beams used for the power-map. In this work we propose an additional algorithm to suppress side-lobes based on the product of multiple CroPaC beams. A Virtual Studio Technology (VST) plug-in has been developed for both the transformation of the time-domain microphone signals into the spherical harmonic domain and the main acoustic camera software; both of which can be downloaded from the companion web-page.

1. INTRODUCTION

Acoustic cameras are tools developed in the spatial audio community which are utilised for the capture and analysis of sound fields. In principle, an acoustic camera imitates the audiovisual aspect of how humans perceive a sound scene by visualising the sound field as a power-map. Incorporating this visual information within a power-map can significantly improve the understanding of the surrounding environment, such as the location and spatial spread of sound sources and potential reflections arriving from different surfaces. Essentially, any system that incorporates a video of the sound scene in addition to a power-map overlay can be labelled as an acoustic camera, of which several commercial systems are available today.

Capturing, analysing and tracking the position of sound sources is a useful technique with application in a variety of fields, which include reflection tracking in architectural acoustics [1, 2, 3], sonar navigation and object detection [4, 5], espionage and the military [6, 7]. An acoustic camera can also be used to identify sound insulation issues and faults in electrical and mechanical equipment. This is due to the fact that in many of these scenarios, the fault can be identified as an area that emits the most sound energy. Therefore, calculating the energy in multiple directions and subsequently generating a power-map that depicts their energies relative to each-other is an effective method of identifying the problem area. There have also been instances which incorporate a spherical video. One example is in [8], where a parabolic mirror with a single camera sensor is utilised and the image is then obtained after unwrapping. For a study using single and multiple cameras, the reader is directed to [9].

One approach to developing an acoustic camera is to utilise a rectangular or circular microphone array and then apply beamforming in several directions to generate the power-map. However, a more common approach is to use a spherical microphone array, as it conveniently allows for the decomposition of the sound scene into individual spatial components, referred to as spherical harmonics [10]. Using these spherical harmonics, it is possible to carry out beamforming with similar spatial resolution for all directions on the sphere. Therefore, these are preferred for use-cases in which a large field of view has to be analysed. The most common signal-independent beamformer in the spherical harmonic domain is based on the plane wave decomposition (PWD) algorithm, which (as the name would suggest) relies on the assumption that the sound sources are received as plane-waves, which makes it suited only for far-field sound sources [11]. These beam patterns can be further manipulated using in-phase [12], Dolph-Chebyshev [10] or maximum energy weightings [13].

Signal-dependent beamformers can also be utilised in acoustic cameras with the penalty of higher computational cost. A common solution is the minimum-variance distortion-less response (MVDR) algorithm [14]. This approach takes into account the inter-channel dependencies between the microphone array signals, in an attempt to enhance the beamformers performance by placing nulls to the interferers. However, the performance of such an algorithm is relatively sensitive in scenarios where high background noise and/or reverberation are present in the sound scene [15]. An alternative approach, proposed in [16], is to apply pre-processing in order to separate the direct components from the diffuse field. This subspace-based separation has been shown to dramatically improve the performance of existing super-resolution imaging algorithms [17]. Another popular subspace-based approach is the multiple signal classification (MUSIC) algorithm [18], which has been orientated as a multiple speaker localisation method in [19], by incorporating a direct-path dominance test.

A recent spatial filtering technique, which can potentially be applied to spherical microphone arrays, is the cross-pattern coherence (CroPaC) algorithm [20]. It is based on measuring the correlation between coincident beamformers and providing a post filter that will suppress noise, interferers and reverberation. The advantage of CroPaC, when compared to other spatial filtering techniques, is that it does not require the direct estimation of the microphone noise. The algorithm has recently been extended in the spherical harmonic domain for arbitrary combinations of beam-

* The research leading to these results has received funding from the Aalto ELEC school.

formers in [21].

The purpose of this work is to detail a scalable acoustic camera system that utilises a spherical microphone array and a spherical video camera, which are placed in a near-coincident fashion. Several different static and adaptive beamforming techniques have been implemented within the system and care has been taken to ensure that the proposed system is accessible to a wide range of acoustic practitioners. Additionally, we investigate the use of a coherence-based parameter to generate the power-maps. The main contributions of this work can be summarised as:

- The capture and analysis of a sound scene using a microphone array and subsequently estimating a parameter to determine sound source activity in specific directions.
- The development of a real-time VST plug-in, for spatially encoding the microphone array signals into spherical harmonic signals.
- Devising a real-time acoustic camera, also implemented as a VST plug-in, by utilising a commercially available spherical microphone array and spherical camera.
- The use of vector-base amplitude panning (VBAP) [22], in order to interpolate the power-map grid.
- Optimal side-lobe suppression of the CroPaC spatial filters for analysis purposes.

This paper is organised as follows. In Section 2 we provide the necessary background on spherical microphone array processing, which includes the encoding of the microphone signals into spherical harmonic signals and common signal-dependent and signal-independent beamforming techniques. In Section 3 we elaborate the proposed theoretical background for generating the power-maps. In Section 4 the details of the hardware and software are shown in detail. Finally, in Section 5 we present our conclusions.

2. SPHERICAL MICROPHONE ARRAY PROCESSING

Spherical microphone arrays (SMA) are commonly utilised for sound field analysis for three-dimensional (3-D) spaces, as they provide a similar performance in all directions when sensors are placed uniformly or nearly-uniformly on the sphere. In this section we provide a brief overview of how to estimate the spherical harmonic signals from the microphone signals and how to perform adaptive and non-adaptive beamforming in the spherical harmonic domain. Only the details required for the current implementation are included here. For a detailed overview of these methods, the reader is referred to [12, 23, 10, 24, 25].

Note that matrices, $\mathbf{M}$, have been denoted using bold upper-case letters and vectors, $\mathbf{v}$, are denoted with bold lower-case letters.

2.1. Spatial encoding

The SMA may be denoted with Q sensors at $\Omega_q = (\theta, \phi, r)$ locations with $\theta \in [-\pi/2, \pi/2]$ denoting elevation angle, $\phi \in [-\pi, \pi]$ azimuthal angle and r the radius. A common approach is to decompose the microphone input signal, $\mathbf{x} \in \mathbb{C}^{Q\times 1}$, into a set of spherical harmonic signals for each frequency. The accuracy of this decomposition depends on the microphone distribution on the sphere, the type of the array and the radius [10]. The total number of microphones defines the highest order of spherical harmonic signals L that can be estimated. Please note that the frequency and time indexes are omitted for the brevity of notation.

The spherical harmonic signals can be estimated as

$$\mathbf{s} = \mathbf{W}\mathbf{x}, \tag{1}$$

where

$$\mathbf{s} = [s_{00}, s_{1-1}, s_{10}, \ldots, s_{LL-1}, s_{LL}]^T \in \mathbb{C}^{(L+1)^2\times 1}, \tag{2}$$

are the spherical harmonic signals and $\mathbf{W} \in \mathbb{C}^{(L+1)^2\times Q}$ is the frequency-dependent spatial encoding matrix. For uniform or nearly-uniform microphone arrangements, it can be calculated as

$$\mathbf{W} = \alpha_q \mathbf{W}_l \mathbf{Y}^\dagger, \tag{3}$$

where α_q are the sampling weights, which depend on the microphone distribution on the sphere [10]. The sampling weights can be calculated as $\alpha_q = \frac{4\pi}{Q}$. Furthermore, $\mathbf{W}_l \in \mathbb{C}^{(L+1)^2\times(L+1)^2}$ is an equalisation matrix that eliminates the effect of the sphere, defined as

$$\mathbf{W}_l = \begin{bmatrix} w_0 & & & & & \\ & w_1 & & & & \\ & & w_1 & & & \\ & & & w_1 & & \\ & & & & \ddots & \\ & & & & & w_L \end{bmatrix}, \tag{4}$$

where

$$w_l = \frac{1}{b_l}\frac{|b_l|^2}{|b_l|^2 + \lambda^2}, \tag{5}$$

where b_l are frequency and order-dependent modal coefficients, which contain the information of the type of the array, open or rigid, and the type of sensors, omnidirectional or directional. Lastly, λ is a regularisation parameter that influences the microphone noise amplification. For details of some alternative options for calculating the equalisation matrix $\mathbf{W}_l$, the reader is referred to [26, 27], or for a signal-dependent encoder [28]. $\mathbf{Y}(\Omega_q) \in \mathbb{R}^{Q\times(L+1)^2}$ is a matrix containing the spherical harmonics

$$\mathbf{Y}(\Omega_q) = \begin{bmatrix} Y_{00}(\Omega_1) & Y_{00}(\Omega_2) & \ldots & Y_{00}(\Omega_Q) \\ Y_{-11}(\Omega_1) & Y_{-11}(\Omega_2) & \ldots & Y_{-11}(\Omega_Q) \\ Y_{10}(\Omega_1) & Y_{10}(\Omega_2) & \ldots & Y_{10}(\Omega_Q) \\ Y_{11}(\Omega_1) & Y_{11}(\Omega_2) & \ldots & Y_{11}(\Omega_Q) \\ \vdots & \vdots & \vdots & \vdots \\ Y_{LL}(\Omega_1) & Y_{LL}(\Omega_2) & \ldots & Y_{LL}(\Omega_Q) \end{bmatrix}^T, \tag{6}$$

where Y_{lm} are the individual spherical harmonics of order $l \geq 0$ and degree $m \in [-l, l]$.

2.2. Generating power-maps and pseudo-spectrums in the spherical harmonic domain

A power-map can be generated by steering beamformers in multiple directions, as dictated by some form of pre-defined grid. The energy of these beamformed signals can then be calculated and subsequently plotted with an appropriate colour gradient.

Static beamformers in the spherical harmonic domain can be generated using

$$y(\Omega_j) = \mathbf{w}_{\mathrm{PWD}}^H \mathbf{s}, \tag{7}$$

where y denotes the output signal for direction Ω_j and $\mathbf{w}_{\mathrm{PWD}} \in \mathbb{C}^{(L+1)^2 \times 1}$ is a vector containing the beamforming weights, calculated as

$$\mathbf{w}_{\mathrm{PWD}} = \mathbf{y}(\Omega_j) \odot \mathbf{d}, \tag{8}$$

where $\mathbf{y}(\Omega_j) \in \mathbb{C}^{1 \times (L+1)^2}$ are the spherical harmonics for direction Ω_j, $\odot$ denotes the Hadamard product and $\mathbf{d}$ is a vector of weights, defined as

$$\mathbf{d} = [d_0, d_1, d_1, d_1, \ldots, d_L]^T \in \mathbb{R}^{(L+1)^2 \times 1}. \tag{9}$$

The weights $\mathbf{d}$ can be adjusted to synthesise different types of axis symmetric beamformers: regular [10], in-phase [12], maximum energy [13, 23] and Dolph-Chebyshev [10]. A comparison of the performance of these beamformers as DOA estimators is given in [21]. The spherical harmonic signals or a spherical harmonic-domain beamformer can be steered to an arbitrary angle. Steering matrices for rotationally symmetric functions can be obtained using real multipliers [29]. Rotations for arbitrary angles can also be performed by utilising the Wigner-D weighting [30], or by utilising projection methods [31].

Adaptive beamformers may also be utilised for generating power-maps. Typically, these signal-dependent methods operate on the covariance matrix of the spherical harmonic signals $\mathbf{C}_{\mathrm{lm}} \in \mathbb{C}^{(L+1)^2 \times (L+1)^2}$, which can be estimated as

$$\mathbf{C}_{lm} = \mathbf{W}\mathbf{C}_{\mathrm{x}}\mathbf{W}^H, \tag{10}$$

where $\mathbf{C}_x = \mathrm{E}\left[\mathbf{x}\mathbf{x}^H\right] \in \mathbb{C}^{Q \times Q}$ is the covariance matrix of the microphone input signals and $\mathrm{E}[\cdot]$ represents a statistical expectation operator. The covariance matrix can be estimated using an average over finite time frames, typically in the range of tens of milliseconds, or by employing recursive schemes.

A popular signal-dependent beamforming approach is to solve the MVDR minimisation problem, which aims to synthesise a beam that adaptively changes according to the input signal. The response of this beamformer is constrained to have unity gain in the look direction, while the variance of the output is minimised [10]. This minimisation problem is defined as

$$\begin{aligned} &\text{minimise } \mathbf{w}\mathbf{C}_{\mathrm{lm}}\mathbf{w}^H \\ &\text{subject to } \mathbf{y}(\Omega_j)\mathbf{w}^H = 1, \end{aligned} \tag{11}$$

which can be solved to obtain the beamforming weights using

$$\mathbf{w} = \frac{\mathbf{y}(\Omega_j)\mathbf{C}_{\mathrm{lm}}^{-1}}{\mathbf{y}(\Omega_j)\mathbf{C}_{\mathrm{lm}}^{-1}\mathbf{y}^H(\Omega_j)}. \tag{12}$$

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.

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 [19]. First, the signal $\mathbf{U}_s \in \mathbb{C}^{1 \times 1}$ and noise $\mathbf{U}_n \in \mathbb{C}^{(L+1)^2 - 1 \times (L+1)^2 - 1}$ subspaces are obtained via a singular-value decomposition (SVD) of the spherical harmonic covariance matrix

$$\mathbf{C}_{lm} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H = [\mathbf{U}_s\mathbf{U}_n]\begin{bmatrix}\mathbf{\Sigma}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_n\end{bmatrix}\begin{bmatrix}\mathbf{U}_s \\ \mathbf{U}_n\end{bmatrix}, \tag{13}$$

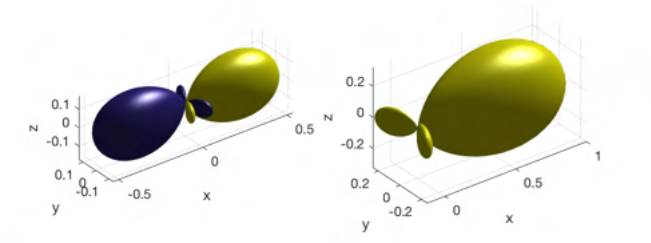


Figure 1: Visualisation of a CroPac beam for $L = 2$ before and after half-wave rectification, shown in the left and right plots, respectively.

where $\mathbf{\Sigma}$ denotes the singular values and $\mathbf{C}_{lm}$ is of unit effective rank.

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 value, σ_1 of matrix $\mathbf{\Sigma}$ is significantly larger than the second singular value, σ_2

$$\frac{\sigma_1}{\sigma_2} > \beta, \tag{14}$$

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 [19]. 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

$$S_{\mathrm{MAP}}(\Omega_j) = \frac{1}{\mathbf{y}(\Omega_j)\left(\mathbf{I} - \mathbf{U}_s\mathbf{U}_s{}^H\right)\mathbf{y}^H(\Omega_j)}, \tag{15}$$

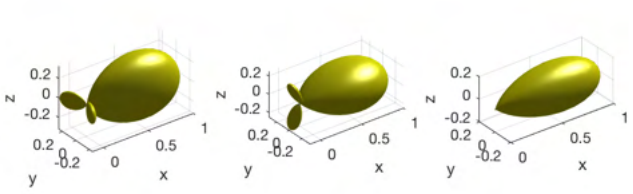
where S_{MAP} is the pseudo-spectrum value for direction Ω_j, and $\mathbf{I}$ is an identity matrix.

3. COHERENCE-BASED SOUND SOURCE TRACKING

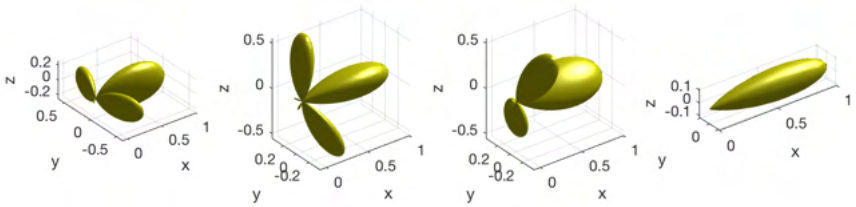
In this work, instead of utilising beamformers to generate an energy-based power-map or utilising subspace methods to generate a pseudo-spectrum, we estimate a parameter using the cross spectrum of different beamformers. This parameter, the cross pattern coherence (CroPaC), has been utilised for spatial filtering applications, where it has been shown to be effective in noisy and reverberant conditions [20, 32]. In this section we propose a generalisation of the algorithm presented in [20] for SMAs, using static beamformers and microphone arrays that define an arbitrary order L. A novel approach of suppressing the side-lobes of CroPaC beams is also explored.

3.1. Cross-spectrum-based parameter estimation

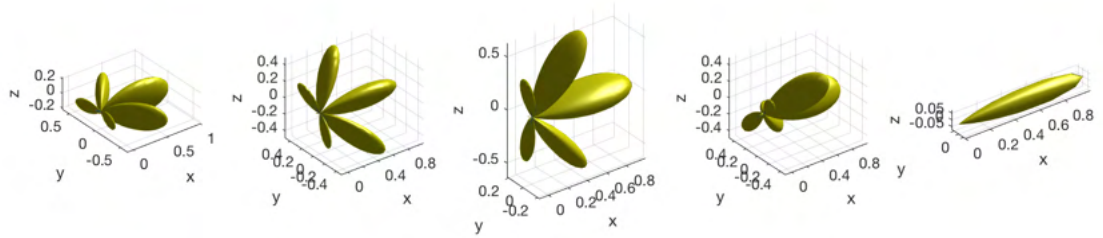
The CroPaC algorithm estimates the probability of a sound source emanating from a specific direction in a 3-D space. The time-domain microphone signals are initially transformed into the spherical harmonic domain according to the formulation shown in Section 2.1, up to order L. The spherical harmonic signals are



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



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



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

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

then transformed into the time-frequency domain and a parameter is estimated for each frequency, k, and time index, n. The cross spectrum is then calculated between two spherical harmonic signals of orders L and $L + 1$ and the same degree m

$$G(\Omega_j, k, n) = \lambda \frac{\Re[\mathbf{s}_L(\Omega_j, k, n)^* \mathbf{s}_{L-1}(\Omega_j, k, n)]}{\sum_L^{L+1} |\mathbf{s}_L(\Omega_j, k, n)|^2}, \qquad (16)$$

where $\Re$ denotes the real operator, $\mathbf{s}_L$ and $\mathbf{s}_{L-1}$ 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 $G_{\mathrm{MAP}} \in [0, 1]$. The normalisation factor can be calculated as

$$\lambda = \frac{(L+1)^2 - (L-1)^2 + 1}{2} = \frac{4L+1}{2}. \qquad (17)$$

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

$$G_{\mathrm{MAP}}(\Omega_j, n) = \max\left[0, \frac{1}{K} \sum_{k=1}^{K} G(\Omega_j, k, n)\right]. \qquad (18)$$

The half-wave rectification process ensures that only sounds arriving from the look direction are analysed. An illustration of the effect of the half-wave rectification process to the directional selectivity of the CroPaC beams is depicted in Fig. 1.

3.2. Side-lobe suppression

The calculation of the spectrum between different orders of beamformers results in the creation of unwanted side-lobes that exhibit different shapes depending on the order. A visual depiction of these aberrations, in Fig. 2, have been generated by multiplying the following spherical harmonics together: $Y_{LL} Y_{(L+1)(L+1)}$ for

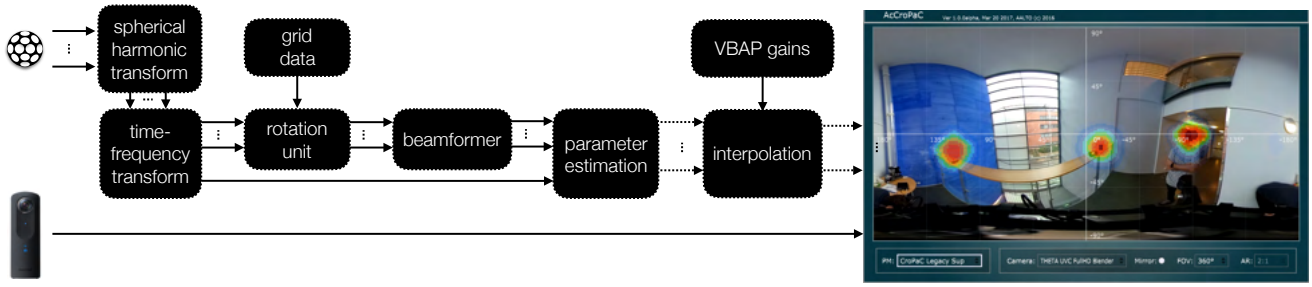


Figure 3: Block diagram of the proposed parametric acoustic camera system. The microphone array signals are processed in the time-frequency domain using the proposed parametric analysis. The output is then averaged across frequencies and the grid is interpolated with VBAP for visualisation. The resulting power-map is projected on top of the spherical video.

$L = 1$ (Fig. 2, (a), left), $L = 2$ (Fig. 2, (b), left) and $L = 3$ (Fig. 2, (c), left).

These side-lobes can potentially introduce biases in the power-map. Therefore, in this sub-section we propose a technique to suppress these side-lobes by multiplying rotated versions of the estimated beams. The number of estimated beams is determined by the order L. The side-lobe suppressing parameter G_{SUP} is estimated as

$$G_{\mathrm{SUP}}(\Omega_j, n) = \begin{cases} G_{\mathrm{MAP}}(\Omega_j, n) & \text{, if } L = 1 \\ \prod_{i=1}^{L} G_{\mathrm{MAP}}^{\rho_i}(\Omega_j, n) & \text{, if } L > 1, \end{cases} \tag{19}$$

where ρ_i is a parameter that defines an axis symmetric roll on the direction of the beam of $\frac{\pi}{L}$ radians. Such a roll can successfully suppress side-lobes that are generated by the multiplication of the different spherical harmonics. This concept is illustrated in Fig. 2. Each row illustrates the side-lobe suppression for different orders. In the top row $L = 1$, which results in a single roll of $\frac{\pi}{2}$ in (19). For $L = 2$ (middle row) and $L = 3$ (bottom row) three and four rolls of $\frac{\pi}{3}$ and $\frac{\pi}{4}$ are applied, respectively. The resulting enhanced beam patterns $G_{\mathrm{SUP}} \in [0, 1]$, which are derived from the product of multiple $G_{\mathrm{MAP}} \in [0, 1]$, are shown on the right-hand side of the figures.

4. IMPLEMENTATION DETAILS

In order to make the software easily accessible, the acoustic camera was implemented as a virtual studio technology (VST) audio plug-in[1] using the open-source JUCE framework. The motivation for selecting the VST architecture, is the wide range of digital audio workstations (DAWs) that support them. This enables an acoustic practitioner to select their DAW of choice, which will in turn act as the bridge between real-time microphone signal capture and the subsequent visualisation of the energy distribution in the sound scene. The rationale behind the selection of JUCE is the many useful classes it offers; most notably of which is camera support, which allows for real-time video to be placed behind the corresponding power-map. Additionally, the framework is developed with cross-platform support in mind and can also produce other audio plug-in formats, provided that the corresponding source development kits are linked to the project.

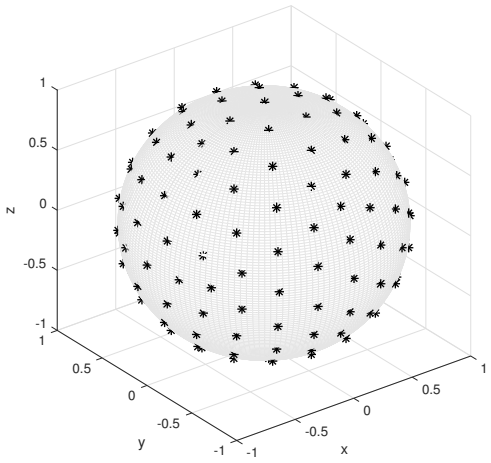
[1]The VST plug-ins are available for download on the companion webpage: http://research.spa.aalto.fi/publications/papers/acousticCamera/

The algorithms within the acoustic camera have been generalised to support spherical harmonic signals up to the 7th order. These signals can be optionally generated by using the accompanying Mic2SH VST, which accepts input signals from spherical microphone arrays such as A-format microphones (1st order) or the Eigenmike (up to 4th order). In the case of the Eigenmike, Mic2SH will also perform the necessary frequency-dependent equalisation, described in Section 2.1, in order to mitigate the radial dependency incurred when estimating the pressure on a rigid sphere. Different equalisation strategies have been implemented that are common in the literature, such as the Tikhonov-based regularised inversion [23] and soft limiting [33].

In order to optimise the linear algebra operations, the code within the audio plug-in has been written to conform to the basic linear algebra library (BLAS) standard. Other operations such as the lower-upper (LU) factorisation and SVD are addressed by the linear algebra package (LAPACK) standard; for which Apple's accelerate framework and Intel's MKL are supported for the Mac OSX and Windows versions, respectively.

The overall block diagram of the proposed system is shown in Fig. 3. The time-domain microphone array signals are initially transformed into spherical harmonic signals using the Mic2SH audio plug-in, which are then transformed into the time-frequency domain by the acoustic camera. For computational efficiency reasons, the spherical harmonic signals are rotated after the time-frequency transform towards the points defined by the pre-computed spherical grid. These signals are then fed into a beamformer unit, which forms the two beams that are required to compute the cross-spectrum based parameter for each grid point. Note that when the side-lobe suppression mode is enabled, one parameter is estimated per roll and the resulting parameters are multiplied, as defined in (19). For visualisation, the parameter value at each of the grid points is interpolated using VBAP and projected on top of the spherical video.

The user-interface for the acoustic camera consists of a view window and a parameter editor (see Fig. 3). 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 user definable in the parameter editor, which allows the VST to accommodate a wide range of different web-cam devices. Additionally, the image frames from the camera can be optionally mirrored using an appropriate affine transformation (left-right, or up-down); in order to accommodate for a variety of different camera orientations.



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

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

Figure 4: Spherical and interpolated grids.

4.1. Time-frequency transform

The time-frequency transform utilised in this work is filter-bank-based and was implemented originally for the study in [21] [2]. The filter-bank was configured to use a hop size of 128 samples and an FFT size of 1024. Additionally, the optional hybrid filtering mode offered by the filter-bank was enabled, which allows for more resolution in the low frequency region by dividing the lowest four bands into eight; thus, attaining 133 frequency bands in total. A sampling rate of 48 kHz was chosen, and the power-map analysis utilises frequency bands with centre frequencies between [140, 8000] Hz. The upper limit of 8000 Hz was selected due to the spatial aliasing of the microphone array used [34].

4.2. Power-map modes and sampling grids

The power-map is generated by sampling the sphere with a spherical grid. A precomputed almost-uniform spherical grid was chosen that provides 252 nearly-uniformly distributed data points on the sphere. The grid is based on the 3LD library [35], 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 of the spherical grid. Two different power-map modes and two pseudo-spectrum methods were implemented in the spherical harmonic domain: conventional signal-independent beamformers (PWD, minimum side-lobe, maximum energy and Dolph-Chebyshev); MVDR beamformers; multiple signal classification (MUSIC); and the proposed cross-spectrum-based with the additional side-lobe suppression. The power-map/pseudo-spectrum values are then summed over the analysis frequency bands and averaged over time slots using a one-pole filter

$$\hat{G}_{\mathrm{SUP}}(\Omega_j, n) = \alpha \hat{G}_{\mathrm{SUP}}(\Omega_j, n) + (1-\alpha) G_{\mathrm{SUP}}(\Omega_j, n-1), \tag{20}$$

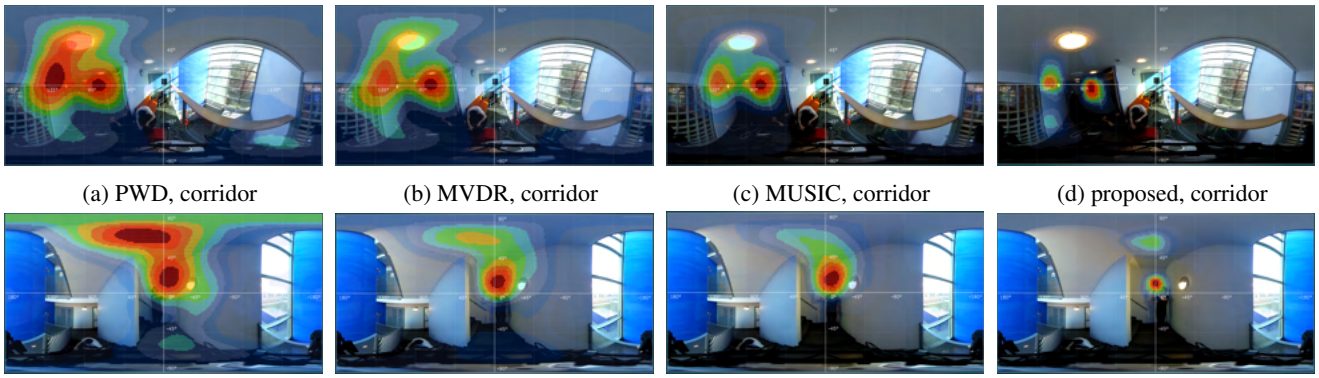
where $\alpha \in [0, 1]$ is the smoothing parameter. The spherical power-map values are then interpolated to attain a two-dimensional (2-D) power-map, using pre-computed VBAP gains. The spherical and interpolated grids are shown in Fig. 4. These 2-D power-maps are then further interpolated using bi-cubic interpolation depending on the display settings and are normalised such that $\hat{G}_{\mathrm{SUP}} \in [0, 1]$. The pixels that correspond to the 2-D 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 energy valued beams to ensure that they do not unnecessarily detract from the video stream.

4.3. Example power-maps

Power-maps examples are shown in Fig. 5 for four different modes: the basic PWD beamformer, the adaptive MVDR beamformer, the subspace MUSIC approach, and the proposed technique. The recordings were performed by utilising the Eigenmike microphone array and a RICOH Theta S spherical camera. Fourth order spherical harmonic signals were generated using the accompanying Mic2SH VST plugin, which were then used by all four power-map modes. The video was unwrapped using the software provided by RICOH and then combined with the calculated power-map to complete the acoustic camera system. However, since the camera may not be facing the same look direction as the microphone array, a calibration process is required in order to align the power-map with the video stream. However, it should be noted that since the two devices do not share a common origin, sources that are very close to the array may not be correctly aligned. The resulting power-maps are shown for two different recording scenarios: a staircase of high reverberation time of approximately 2 seconds (Fig. 5, bottom) and a corridor of approximately 1.5 seconds (Fig. 5, top).

It can be seen from Fig. 5(top) 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

[2] `1https://github.com/jvilkamo/afSTFT`



(a) PWD, corridor (b) MVDR, corridor (c) MUSIC, corridor (d) proposed, corridor

(e) PWD, stairs (f) MVDR, stairs (g) MUSIC, stairs (h) proposed, stairs

Figure 5: Images of the acoustic camera VST display, while using fourth order spherical harmonic signals and the four processing modes in reverberant environments.

than they actually are. The distinction between the two paths is improved slightly when using MVDR beamformers, which is improved further when utilising the MUSIC algorithm. However, in the case of the proposed technique, the two paths are now completely isolated and a second early reflection with lower energy is now visible; which is not as evident in the other three methods. PWD also indicates a sound source that is likely the result of the side-lobes pointing towards the real sound source; thus, highlighting the importance of side-lobe suppression for acoustic camera applications. Fig. 5(bottom) indicates similar performance; however, in the case of MUSIC, the ceiling reflection is more difficult to distinguish as a separate entity.

5. CONCLUSIONS

This paper has presented an acoustic camera that is easily accessible as a VST plug-in. Among the possible power-map modes available, is the proposed coherence-based parameter, which can be tuned to this particular use case via additional suppression of the side-lobes. This method presents an intuitive approach to attaining a power-map, and is potentially easier and computationally cheaper to implement than MVDR or MUSIC, as it does not rely on lower-upper decompositions, Guassian Elimination, or singular value decompositions. It is also demonstrated that in the simple recording scenarios, the proposed method can be inherently tolerant to reverberation.

6. REFERENCES

[1] Adam O'Donovan, Ramani Duraiswami, and Dmitry Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 5284–5287.

[2] Angelo Farina, Alberto Amendola, Andrea Capra, and Christian Varani, "Spatial analysis of room impulse responses captured with a 32-capsule microphone array," in *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

[3] Lucio Bianchi, Marco Verdi, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "High resolution imaging of acoustic reflections with spherical microphone arrays," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*. IEEE, 2015, pp. 1–5.

[4] GC Carter, "Time delay estimation for passive sonar signal processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 463–470, 1981.

[5] H Song, WA Kuperman, WS Hodgkiss, Peter Gerstoft, and Jea Soo Kim, "Null broadening with snapshot-deficient covariance matrices in passive sonar," *IEEE journal of Oceanic Engineering*, vol. 28, no. 2, pp. 250–261, 2003.

[6] RK Hansen and PA Andersen, "A 3d underwater acoustic camera: properties and applications," in *Acoustical Imaging*, pp. 607–611. Springer, 1996.

[7] Russell A Moursund, Thomas J Carlson, and Rock D Peters, "A fisheries application of a dual-frequency identification sonar acoustic camera," *ICES Journal of Marine Science: Journal du Conseil*, vol. 60, no. 3, pp. 678–683, 2003.

[8] Leonardo Scopece, Angelo Farina, and Andrea Capra, "360 degrees video and audio recording and broadcasting employing a parabolic mirror camera and a spherical 32-capsules microphone array," *IBC 2011*, pp. 8–11, 2011.

[9] Adam O'Donovan, Ramani Duraiswami, and Jan Neumann, "Microphone arrays as generalized cameras for integrated audio visual processing," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[10] Boaz Rafaely, *Fundamentals of spherical array processing*, vol. 8, Springer, 2015.

[11] Miljko M Erić, "Some research challenges of acoustic camera," in *Telecommunications Forum (TELFOR), 2011 19th*. IEEE, 2011, pp. 1036–1039.

[12] J Daniel, *Représentation de champs acoustiques, applicationa la reproduction eta la transmission de scenes sonores complexes dans un contexte multimédia*, Ph.D. thesis, Ph. D. thesis, University of Paris 6, Paris, France, 2000.

[13] Franz Zotter, Hannes Pomberger, and Markus Noisternig, "Energy-preserving ambisonic decoding," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 37–47, 2012.

[14] Michael Brandstein and Darren Ward, *Microphone arrays: signal processing techniques and applications*, Springer Science & Business Media, 2013.

[15] Michael D Zoltowski, "On the performance analysis of the mvdr beamformer in the presence of correlated interference," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 6, pp. 945–947, 1988.

[16] Nicolas Epain and Craig T Jin, "Super-resolution sound field imaging with sub-space pre-processing," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 350–354.

[17] Tahereh Noohi, Nicolas Epain, and Craig T Jin, "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, 2013, pp. 346–349.

[18] Ralph Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[19] O Nadiri and B Rafaely, "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)*, vol. 22, no. 10, pp. 1494–1505, 2014.

[20] Symeon Delikaris-Manias and Ville Pulkki, "Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2356–2367, 2013.

[21] Symeon Delikaris-Manias, Despoina Pavlidi, Ville Pulkki, and Athanasios Mouchtaris, "3d localization of multiple audio sources utilizing 2d doa histograms," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1473–1477.

[22] Ville Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[23] Sébastien Moreau, Jérôme Daniel, and Stéphanie Bertet, "3d sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone," in *120th Convention of the AES*, 2006, pp. 20–23.

[24] Thushara D Abhayapala, "Generalized framework for spherical microphone arrays: Spatial and frequency decomposition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 5268–5271.

[25] Heinz Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, vol. 348, Springer, 2007.

[26] David L Alon, Jonathan Sheaffer, and Boaz Rafaely, "Robust plane-wave decomposition of spherical microphone array recordings for binaural sound reproduction," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1925–1926, 2015.

[27] Stefan Lösler and Franz Zotter, "Comprehensive radial filter design for practical higher-order ambisonic recording," *Fortschritte der Akustik, DAGA*, pp. 452–455, 2015.

[28] Franz Zotter Christian Schörkhuber and Robert Höldrich, "Signal-dependent encoding for first-order ambisonic microphones," *Fortschritte der Akustik, DAGA*, 2017.

[29] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2002, vol. 2, pp. II–1781–II–1784.

[30] Boaz Rafaely and Maor Kleider, "Spherical microphone array beam steering using wigner-d weighting," *IEEE Signal Processing Letters*, vol. 15, pp. 417–420, 2008.

[31] J. Atkins, "Robust beamforming and steering of arbitrary beam patterns using spherical arrays," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2011, pp. 237–240.

[32] Symeon Delikaris-Manias, Juha Vilkamo, and Ville Pulkki, "Signal-dependent spatial filtering based on weighted-orthogonal beamformers in the spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 9, pp. 1507–1519, 2016.

[33] Benjamin Bernschütz, Christoph Pörschmann, Sascha Spors, Stefan Weinzierl, and Begrenzung der Verstärkung, "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, 2011.

[34] Boaz Rafaely, Barak Weiss, and Eitan Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1003–1010, 2007.

[35] Florian Hollerweger, *Periphonic sound spatialization in multi-user virtual environments*, Citeseer, 2006.