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

Leo McCormack, Symeon Delikaris-Manias and Ville Pulkki*

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 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 [10, 11], sonar navigation and object detection [4, 5], espionage [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 [6], 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 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 and subsequently estimating a parameter to determine source sound activity in specific directions.
- The development of a real-time VST plug-in, for spatially encoding the microphone input signal, \( x \), into a set of spherical harmonic signals from the microphone signals and how to perform adaptive and non-adaptive beamforming in the spherical harmonic domain.
- 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, \( M \), have been denoted using bold uppercase letters and vectors, \( v \), are denoted with bold lowercase 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, \( x \in \mathbb{C}^{Q \times 1} \), 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

\[
    s = W x, \tag{1}
\]

where

\[
    s = [s_0, s_1, s_2, \ldots, s_{LL}, s_{LL-1}]^T \in \mathbb{C}^{(L+1)^2 \times 1}, \tag{2}
\]

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

\[
    W = \alpha Q W_1 Y^T, \tag{3}
\]

where \( \alpha_q \) 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, \( W_1 \in \mathbb{C}^{(L+1)^2 \times (L+1)^2} \) is an equalisation matrix that eliminates the effect of the sphere, defined as

\[
    W_1 = \begin{bmatrix}
    w_0 & w_1 & \cdots & w_L \\
    w_1 & w_1 & \cdots & w_L \\
    \vdots & \vdots & \ddots & \vdots \\
    w_L & w_L & \cdots & w_L \\
    \end{bmatrix}, \tag{4}
\]

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, \( \lambda \) is a regularisation parameter that influences the microphone noise amplification. For details of some alternative options for calculating the equalisation matrix \( W_1 \), the reader is referred to \([26, 27]\), or for a signal-dependent encoder \([28]\). \( Y(\Omega_q) \in \mathbb{R}^{Q \times (L+1)^2} \) is a matrix containing the spherical harmonics

\[
    Y(\Omega_q) = \begin{bmatrix}
    Y_{00}(\Omega_1) & Y_{00}(\Omega_2) & \cdots & Y_{00}(\Omega_Q) \\
    Y_{-11}(\Omega_1) & Y_{-11}(\Omega_2) & \cdots & Y_{-11}(\Omega_Q) \\
    Y_{10}(\Omega_1) & Y_{10}(\Omega_2) & \cdots & Y_{10}(\Omega_Q) \\
    Y_{11}(\Omega_1) & Y_{11}(\Omega_2) & \cdots & Y_{11}(\Omega_Q) \\
    \vdots & \vdots & \ddots & \vdots \\
    Y_{LL}(\Omega_1) & Y_{LL}(\Omega_2) & \cdots & Y_{LL}(\Omega_Q) \\
    \end{bmatrix}^T, \tag{6}
\]

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

\[
    y(\Omega_j) = \mathbf{w}_p^H \mathbf{w}_p^T y, \tag{7}
\]
where $y$ denotes the output signal for direction $\Omega_j$ and $w_{\text{PWD}} \in \mathbb{C}^{(L+1)^2 \times 1}$ is a vector containing the beamforming weights, calculated as

$$w_{\text{PWD}} = y(\Omega_j) \odot d,$$

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

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

The weights $d$ 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 microphone signals are initially transformed into the spherical harmonic domain according to the formulation shown in [12]. The spherical harmonic signals are defined as

$$\mathbf{c}_{\text{SH}} = \mathbf{W} \mathbf{c},$$

where $\mathbf{c}_{\text{SH}}$ is the spherical harmonic domain microphone signals and $\mathbf{c}$ is the microphone input signals and $\mathbf{W}$ 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

$$\min w \mathbf{C}_{\text{tm}} w^H$$

subject to $y(\Omega_j) w^H = 1,$

which can be solved to obtain the beamforming weights using

$$w = \frac{y(\Omega_j) \mathbf{C}_{\text{tm}}^{-1}}{y(\Omega_j) \mathbf{C}_{\text{tm}}^{-1} y^H(\Omega_j)}.$$ (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 $\mathbf{u}_n \in \mathbb{C}^{(L+1)^2 \times 1}$ subspaces are obtained via a singular-value decomposition (SVD) of the spherical harmonic covariance matrix

$$\mathbf{C}_{\text{tm}} = \mathbf{U} \mathbf{\Sigma} \mathbf{U}^H = \begin{bmatrix} \mathbf{u}_s & \mathbf{u}_n \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_s & 0 \\ 0 & \mathbf{\Sigma}_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_s \\ \mathbf{u}_n \end{bmatrix},$$

where $\mathbf{\Sigma}$ denotes the singular values and $\mathbf{C}_{\text{tm}}$ 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, $\sigma_1$ of matrix $\mathbf{\Sigma}$ is significantly larger than the second singular value, $\sigma_2$.

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

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

$$S_{\text{MAP}}(\Omega_j) = \frac{1}{y(\Omega_j) \mathbf{I} - \mathbf{u}_s \mathbf{u}_s^H} y^H(\Omega_j),$$

where $S_{\text{MAP}}$ is the pseudo-spectrum value for direction $\Omega_j$, 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

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

$$G(\Omega_j, k, n) = \frac{\Re[s_L(\Omega_j, k, n) \ast s_{L-1}(\Omega_j, k, n)]}{\sum_{L=0}^{L+1} |s_L(\Omega_j, k, n)|^2},$$

(16)

where $\Re$ denotes the real operator, $s_L$ and $s_{L-1}$ are the spherical harmonic signals for a look direction $\Omega_j$ and the same degree $m$, $\ast$ denotes the complex conjugate and $\lambda$ is an order-dependent normalisation factor to ensure that $G_{\text{MAP}} \in [0, 1]$. The normalisation factor can be calculated as

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

(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

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

(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: $Y_{L,L} Y_{(L+1),(L+1)}$ for...
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.

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

\[ \hat{G}_{SUP}(\Omega_j, n) = \alpha \hat{G}_{SUP}(\Omega_j, n) + (1 - \alpha)G_{SUP}(\Omega_j, n - 1), \]  

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 \( G_{SUP} \in [0, 1] \). 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 Eigen-mike 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.
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


DAFX-7
Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17), Edinburgh, UK, September 5–9, 2017


