# SOUND SOURCE LOCALISATION IN AMBISONIC AUDIO USING PEAK CLUSTERING

*Marc C. Green & Damian Murphy*

AudioLab, Department of Electronic Engineering, University of York

## ABSTRACT

Accurate sound source direction-of-arrival and trajectory estimation in 3D is a key component of acoustic scene analysis for many applications, including as part of polyphonic sound event detection systems. Recently, a number of systems have been proposed which perform this function with first-order Ambisonic audio and can work well, though typically performance drops when the polyphony is increased. This paper introduces a novel system for source localisation using spherical harmonic beamforming and unsupervised peak clustering. The performance of the system is investigated using synthetic scenes in first to fourth order Ambisonics and featuring up to three overlapping sounds. It is shown that use of second-order Ambisonics results in significantly increased performance relative to first-order. Using third and fourth-order Ambisonics also results in improvements, though these are not so pronounced.

***Index Terms***— sound source localisation, direction of arrival, spatial audio, beamforming, steered-response power, DBSCAN

## 1. INTRODUCTION

Sound Event Localisation and Detection (SELD) is the act of detecting and tracking individual sounds in an acoustic scene consisting of a mixture of sources, typically recorded or monitored using a microphone array. Such a system has applications including audio surveillance [1], vehicle tracking for the military [2], localisation of targets in robotics [3], and as a stage in source separation that has been proposed for use in evaluation of environmental soundscapes [4, 5]. Previous work involving SELD in Ambisonics is limited to FOA [6, 7]. These systems tend to work well when localising a single source, but performance drops when the complexity of the scene is increased by adding multiple sources overlapping in time. Higher-order Ambisonic (HOA) audio has much higher spatial resolution than FOA, and there are now several portable HOA microphones commercially available, including the second-order Core-Sound OctoMic [8] and fourth-order mh Acoustics Eigenmike [9], making it simple to gather high-order Ambisonic recordings.

The EigenScape database of acoustic scenes [10] was recorded in HOA using the Eigenmike. Analysis of this database has shown that spatial audio features can be useful in acoustic scene classification [10, 11], indicating that use of spatial audio could be a fruitful area of investigation in future work on soundscape analysis. Using spatial audio to consider individual sources will require a robust method for SELD.

In this paper we introduce a new method for estimation of onset/offset times and the DOA of active sound sources (covering the first two stages of SELD as defined in [6]) in Ambisonic recordings of acoustic scenes using spherical harmonic beamforming and unsupervised clustering of power peaks by the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [12]. An unsupervised approach such as that presented here could be especially useful when the amount of data available for training is small. The paper is organised as follows: Section 2 provides a detailed description of the different stages involved in the system, including beamforming, peak-finding, clustering and regression. Section 3 describes the evaluation procedure, including the data used to test the system and the metrics used for assessment and optimisation. Section 4 presents the results of the study, with Section 5 providing discussion. Section 6 concludes the paper.

## 2. METHOD

Our system uses four main steps that will be described in detail in this section. First, spherical harmonic beamforming is used to create steered-response power (SRP) maps. These maps are then analysed to extract a list of peak power positions. DBSCAN is used to create clusters from these peaks, which correspond to identified sound sources. Finally, regression models are fit to each cluster for smoothed trajectory estimates.

### 2.1. Steered-Response Power Map

The first stage of the system is the creation of a series of power maps describing how the sound power varies in the scene over time. Spherical harmonic beamforming is used to create an SRP map [13] for each frame of audio as follows:

$$Z(\theta,\phi) = \sum_{k}\sum_{n=0}^{N}\frac{1}{b_n(k)}\sum_{m=-n}^{n}W_{nm}(\theta,\phi)P_{nm}(k) \qquad (1)$$

where output power $Z$ for azimuth $\theta$ and elevation $\phi$ is calculated as the product of the spherical harmonic-wavenumber domain signals $P$ for spherical harmonic order $n$, degree $m$ and wavenumber $k$, and a weighting function $W$ that determines the look direction of the beam. This calculation is repeated and summed across all $n$ and $m$ [14], and across $k$ for a map describing power in all frequency bands. The $b_n(k)$ term is required to compensate for scattering of sound induced by the presence of the recording array [13, 14], here set to 1 as the system was tested using synthetic sound scenes. In this work, the simplest spherical harmonic beamformer was used, where the weights are simply substituted for the spherical harmonics for the given look direction, a process also known as plane-wave decomposition [15].

To create the SRP map, the beam must be steered in multiple directions to sample the 3D space. We used the Fibonacci spiral [16] to distribute 600 points in a nearly-uniform spherical pattern.

Funding was provided by a UK Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Award, an Audio Engineering Society (AES) Educational Foundation grant, and the Department of Electronic Engineering at the University of York. Thanks to Sharath Adavanne and the team at TUNI Audio Research Group for assistance with the dataset.

https://doi.org/10.33682/05my-4306

This distribution was chosen as it can generate any number of points with the only major irregularities in spacing occurring near the poles [17], which are uncommon positions for sound sources. A map was generated for each frame of audio, describing how the sound power impacting on the measurement position changed over time. We split our audio into frames of 256 samples, using rectangular windows and no overlap.

### 2.2. Peak-finding

The series of SRP maps is then passed to a peak-finding algorithm. Peak-finding in a spherical function presents a challenge in that the wraparound of the sphere at the edges of the data i.e. $f(2\pi, \phi) = f(0, \phi)$ has to be taken into account, along with the fact that the sphere has not been sampled with a regular grid.

We used the peak-finding function from the *dipy* Python library [18]. Originally designed for analysis of MRI data, this peak-finder overcomes these issues by requiring the sampling directions as input as well as the power map. There are two parameters that govern the behaviour of the function. The first, *rel_pk*, is used to calculate a threshold below which to discard peaks by:

$$\text{threshold} = \wedge + \mathit{rel\_pk} \cdot R \tag{2}$$

where $R$ is the range of the data and $\wedge$ is either the minimum of the data or 0 if the minimum is negative. The second factor is *min_sep*, an angular distance that governs the minimum separation allowed between peaks. This helps avoid groups by discarding peaks that are found within this distance of each other - only the largest peak is retained. The algorithm returns an indeterminate number of peaks, so we allocated enough memory for a maximum of 20 per frame. This could easily be extended if the application demanded it, but in practise this limit was rarely approached. The output of this stage is a list of vectors containing the angle of the detected peaks along with the time in seconds.

### 2.3. Clustering

In order to estimate coherent sound sources, we used the DBSCAN algorithm [12] to intelligently cluster sets of peaks proximal in space and time. DBSCAN is an unsupervised algorithm that groups data into clusters based on their proximity, with points in low-density regions (having fewer neighbours) designated as outliers. This algorithm is very useful in terms of estimating which peaks belong to the same source sounds. Onset and offset times for each source can be predicted by considering the first and last-occurring peak points grouped into each cluster.

To once again avoid the problems mentioned in Section 2.2 involving spherical wraparound points, the spherical co-ordinate component of each peak vector is converted to Cartesian co-ordinates. Each peak is therefore mapped from 3D $(t, \theta, \phi)$ to 4D $(t, x, y, z)$, similar to the approach used in [6]. Without this process, there would be a disconnect in the clusters identified by DBSCAN as sources moved across or near to the spherical co-ordinate boundaries.

The spatial dimensions of the data were normalised, as is standard in machine learning, to zero mean and unit variance. The time dimension was not collapsed, as in testing this resulted in clusters being made of peaks occurring in similar spatial locations but separated by large amounts of time. There are two main input parameters for the DBSCAN algorithm:

- $\epsilon$ - The largest distance between two adjacent points before the algorithm considers assigning the points to different clusters.
- *MinPts* - The number of data points required within $\epsilon$ of a given point for that point to be considered a ‘core’ point. This affects how dense groups need to be in order to be clustered.

### 2.4. Regression

Each cluster is used to train a set of Support Vector Regressors (SVRs) [19], which create models of source trajectories. Since the clusters are labelled by the DBSCAN stage, this stage of learning is supervised. A separate regressor is trained for each spatial dimension, modelling $x$, $y$, and $z$ separately against $t$, and the outputs of these three models are combined for a final 4D trajectory. The regressors serve to smooth the raw data, which can exhibit a certain amount of ‘jitter’ as adjacent sample points are instantaneosly identified as peaks in a given frame. The model can also be used to fill in missing points in the cluster, as there might not necessarily be a peak identified in the cluster for every time step. In this way we provide some mitigation for interference.

The salient input parameter for the SVR algorithm in terms of this study is $C$, which is the cost associated with the distance of input data from the regression line. A higher value of $C$ causes overfitting as the cost associated with points not coinciding with the line is high. In this study we determined experimentally to use $1 \times 10^{-3}$ as the value for $C$, as this ensured smoother predicted trajectories with minimal jitter. The output of these regressors is calculated for every frame between the first and last points of each cluster. The predictions are then re-scaled back to the original spatial ranges and the Cartesian co-ordinates are converted back to spherical co-ordinates, giving the final output of the system.

## 3. EVALUATION

### 3.1. Dataset

The system was tested using an expanded version of the TUT Sound Events 2018 Ambisonic Anechoic Dataset [20]. This dataset features synthetic Ambisonic scenes with sounds at static locations in the full range of $\theta$ and at $\phi$ between $\pm 60^\circ$, with a resolution of $10^\circ$. Scenes are included with three levels of polyphony, up to a maximum of one, two or three simultaneous sounds active (denoted *OV1* to *OV3*). Using synthetic data, the level of polyphony, position, and movement of sounds is controllable and can therefore be precisely known. Real recordings would have to be labelled manually and this would be very labour-intensive. Indeed, it is not clear how one would go about labelling real recordings for DOA in a way that would be at all reliable.

Since the original dataset is only available in FOA, we re-synthesised it in fourth-order HOA using the original scene description files and source sounds from the DCASE 2016 task 2 dataset [21], which contains a variety of everyday sounds. See [7] for more detail on the method for synthesising the data. The dataset features 240 training examples and 60 testing examples for each *OV*. Our system has no need of training, so we used only the testing examples. The examples were all resampled to 16 kHz, as would be necessary with real-world Eigenmike recordings in order to avoid spatial aliasing artefacts due to the geometry of the array [22].

### 3.2. Metrics

To assess the system we employ the two frame-wise DOA metrics used in the DCASE 2019 Task 3 challenge [6, 7, 23]. The first is *DOA error*, defined as:

$$\text{DOA error} = \frac{1}{\sum_{t=1}^{T} D_E^t} \sum_{t=1}^{T} \mathcal{H}(\mathbf{DOA}_R^t, \mathbf{DOA}_E^t) \tag{3}$$

where $\mathbf{DOA}_R^t$ and $\mathbf{DOA}_E^t$ are lists of reference and estimated DOAs, respectively, in frame $t$. $D_E^t$ is the number of estimates in $\mathbf{DOA}_E^t$, and $\mathcal{H}$ is the Hungarian algorithm [24], used to assign predicted angles to reference angles based on optimising pair-wise costs using angular distances. This metric gives the average error between predicted and actual DOA angles.

The second metric is *frame recall* (FR), formally defined for DCASE 2019 as [23]:

$$\text{FR} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(D_R^t = D_E^t) \tag{4}$$

where $D_R^t$ is the ground-truth number of sources present in frame $t$, and $\mathbb{1}$ is an indicator function which outputs one where the bracketed condition is met, otherwise returning zero. FR indicates the proportion of frames where the estimated and reference number of active sounds are equal. A perfect system would have a FR of 1 and a DOA error of 0.

### 3.3. Optimisation

To assess the system, we ran it on all 60 test files for each order of Ambisonics from first to fourth-order (denoted *N1* to *N4*) and each level of polyphony available in the dataset. Metrics were calculated for each file, with their means calculated to characterise the system's performance across the whole dataset.

To find the best possible performance for each *N* and *OV*, we used the *hyperopt* library [25] to run 1000 iterations using various combinations of hyperparameters, optimising for FR. Following preliminary tests to find appropriate ranges, we set the search space as follows:

- $\{\epsilon \in \mathbb{R} \mid 0.1 \leq \epsilon \leq 1.25\}$
- $\{\mathit{MinPts} \in \mathbb{Z} \mid 3 \leq \mathit{MinPts} \leq 10\}$
- $\{\mathit{rel\_pk} \in \mathbb{R} \mid 0.0 \leq \mathit{rel\_pk} \leq 1.0\}$
- $\{\mathit{min\_sep} \in \mathbb{Z} \mid 0 < \mathit{min\_sep} < 90\}$

*Hyperopt* uses the Tree Parzen Estimator (TPE) algorithm [26] to focus on optimal values over time. This enables more fine-tuning of the system's performance compared to the same number of iterations in a random search.

## 4. RESULTS

### 4.1. Optimised Systems

Figure 1 shows the performance metrics recorded for each level of overlap and Ambisonic order from systems optimised for maximum FR, along with results from SELDnet reported in [6], as a comparison. It can be seen that performance on *OV1* audio is very good regardless of *N*, with almost perfect FR and low DOA error of around 3°. For *OV2* there is a clear pattern of improvement in performance with increase of *N*. *OV2/N4* yields FR of 0.85 with DOA error of 2.9°, with *OV2/N2* yielding FR of 0.81 with a DOA error of 4.2°, remarkably close to the *OV2/N4* performance. There is a larger gap between *OV2/N1* and *OV2/N2* results than between the other orders. This pattern of performance difference across *N* appears to be more pronounced as *OV* increases. For *OV3/N1* the DOA error is a relatively poor 19.7°, with a FR of 0.55. *OV3/N2* reduces the DOA error to 5.7°, an improvement of 14°, whilst the gap between *OV3/N2* and *OV3/N4* is just 2.5°. FR also increases with *N*, though the difference is not so marked as that of DOA error.

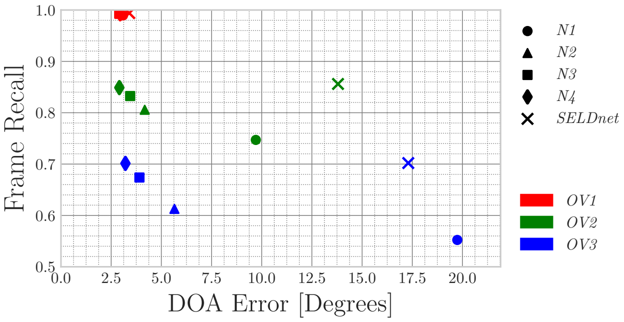


Figure 1: Plot of DOA Error against FR for systems using parameters maximising FR, as well as reported SELDnet results [6].

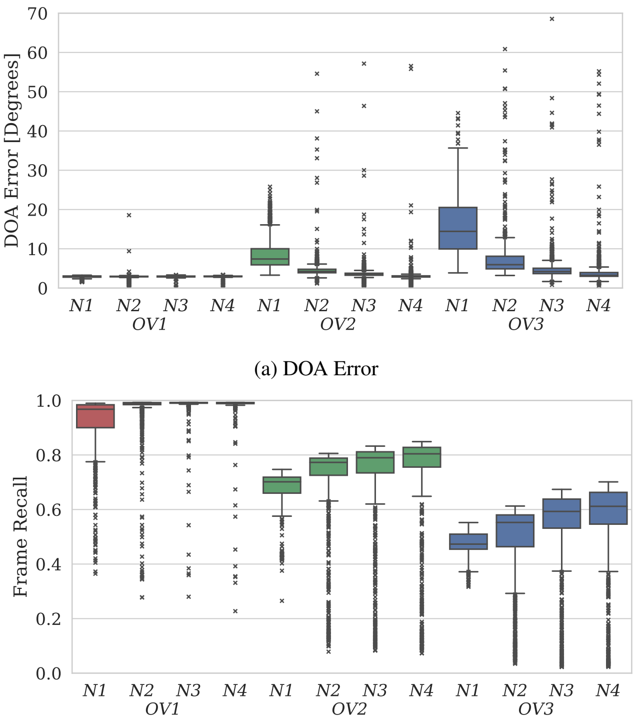
The results achieved for *OV1* are very closely aligned with those achieved by SELDnet. DOA error for *OV2* is smaller than the SELDnet result in all orders, but FR does not begin to approach the SELDnet result until higher orders are used. SELDnet outperforms this system for *OV3/N1*, but is outperformed in terms of DOA error for *OV3* using all higher orders. Again, the FR achieved by SELDnet is only approached using higher orders.

### 4.2. Performance Variance

Figure 2 shows the distribution of results returned by all 1000 iterations of the system using various hyperparameters. It should be noted that due to the use of the TPE algorithm these results will be skewed, with more data on performance with hyperparameters set close to optimal values. This accounts for the large number of visible outliers, although they represent only a small proportion of the 1000 iterations.

It can clearly be seen in Figure 2(a) that increasing the order of the beamformer decreases the median and variance of DOA Error for *OV2*, and to an ever greater degree for *OV3*. Similarly to the pattern of results in Figure 1, the largest reduction in both is between *N1* and *N2*, with higher orders yielding diminishing returns in this regard. The comparatively low variance in systems using *N2* or above indicates a degree of robustness to varying hyperparameters, at least within a certain range. This could be a benefit when using the system on real-world audio in which the precise number of sources would usually be unknown.

Returning attention to the outliers, it can be seen that there are iterations of the system that achieved very low DOA error values. These are not, however, the results shown in Figure 1, as achieving this low DOA error incurs a trade-off whereby FR becomes poor. Further investigation indicated that these metrics were recorded on iterations where both peak-finding parameters discussed in Section 2.2 were set very low. This results in clusters of peaks being identi-



(a) DOA Error

(b) Frame Recall

Figure 2: Distributions of DOA Error and Frame Recall system performance metrics across all 1000 *hyperopt* iterations.

fied in the directions of the sound sources, as opposed to the single peaks enabled when the parameters are set more appropriately. The presence of these clusters may enable the regression stage to interpolate and find a truer DOA for the source lying in the centre of these peaks. Unfortunately, lowering these hyperparameters also results in an increase of spurious peaks, leading to greatly reduced FR. Despite these considerations, in a context where precise DOA measurement was of greater importance than an exact number of active sources, such optimisation for DOA error might be desirable.

The distributions of FR values are shown in Figure 2(b). Once again, higher Ambisonic orders tend to have better performance both in median and maximum values. There is once again a larger increase in performance between *N1* and *N2* than for other orders, though this is not as pronounced as with DOA error values. Unlike DOA error, FR declines consistently given increasing levels of sound overlap. The variance in FR for *OV1* decreases between *N1* and *N2*, but this pattern is not repeated for *OV2*, where variance remains consistent regardless of *N*, or *OV3*, where variance actually increases between *N1* and *N2*.

## 5. DISCUSSION

Results indicate, in terms of best performance as well as average performance and variance, that there is a larger gap between *N1* and *N2* than between *N2* and *N3* or *N4*. Increasing *N* increases the computational complexity of the beamforming stage, as well as the amount of storage space required for the recorded data and the number of microphone capsules required to capture real-world audio. Since this increases cost at all stages, we have an incentive to keep *N* low. The jump in performance between *N1* and *N2* indicates that *N2* may be worth the increased cost, yet limited performance gains at higher orders suggests that second-order Ambisonics might mark a good compromise point for this application. On the other hand, the results do show that the improvements in performance with increased *N* become more pronounced as *OV* increases. Since real-world acoustic scenes are far more complex than the synthesised scenes tested here, use of higher orders may still be useful dependent on context. It is interesting to note that the lowest DOA error achieved at each *OV* in the optimised systems shown in Figure 1 are very similar, at around $3^\circ$. This corresponds closely to the average angular distance between pairs of adjacent points in the 600-point Fibonacci spiral, which is $2.72^\circ$, indicating that the system could achieve even lower DOA values if a finer grid pattern or some method of interpolation were employed (though this would complicate the peak-finding stage).

The fact that FR appears to decrease linearly with increasing *OV* is interesting, especially given that the synthetic scenes used here are anechoic. The diminishing returns in terms of improvements with increasing Ambisonic order indicate a trend towards a maximum performance level which is clearly less than perfect. The best results achieved here are very closely aligned with the results from SELDnet, which provides some evidence there may be a ceiling inherent in either the dataset used or more fundamentally with this type of approach. Increasing *N* will likely result in smaller and smaller improvements whilst at the same time increasing computational complexity exponentially. This indicates that improvements to FR will probably require an improved or alternative method for producing the power map than the plane-wave decomposition SRP method used here or the neural network-generated spatial pseudospectrum used in [6, 7]. It is also possible that given dynamic scenes with multiple moving sources that when the trajectories of two or more sources intersect, the DBSCAN algorithm used here would link them together, thus causing the regression stage to produce wildly erroneous DOA estimates. One potential solution to this could be utilising the different frequency bands present in the power map calculation (e.g. not summing over $k$ in Equation 1) to add another dimension that would make it less likely for overlapping sounds to be clustered provided they remained in different frequency ranges ($\omega$-disjoint orthogonality [4]).

## 6. CONCLUSION

In this paper, we have specified and tested a system using spherical harmonic beamforming and unsupervised peak clustering for conducting sound event localisation in Ambisonic recordings of acoustic scenes. The system has been tested on synthetic scenes it has been shown that performance given a single active source is consistently very good across all Ambisonic orders, with reductions in performance occurring as the number of concurrent sources is increased. Increasing Ambisonic order improves performance, especially between first and second-order.

Future work on this system could seek to improve frame recall by using an alternative method for calculating the power map. DOA error performance could be improved by introducing interpolation to the peak-finding stage. Apart from these improvements, the obvious next step would be to add a labelling stage, making this a fully-fledged SELD system.

## 7. REFERENCES

[1] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio Surveillance: A Systematic Review," *ACM Computing Surveys*, vol. 48, no. 4, Feb 2016. [Online]. Available: https://arxiv.org/abs/1409.7787

[2] M. R. Azimi-Sadjadi and N. R. A. Pezeshki, "Wideband DOA Estimation Algorithms for Multiple Moving Sources using Unattended Acoustic Sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, no. 4, pp. 1585–1598, October 2008.

[3] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor Auditory Scene Analysis Using a Moving Microphone Array Embedded in a Quadrocopter," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[4] O. Bunting and D. Chesmore, "Time frequency source separation and direction of arrival estimation in a 3D soundscape environment," *Applied Acoustics*, vol. 74, no. 2, pp. 264–268, Feb 2013. [Online]. Available: http://dx.doi.org/10.1016/j.apacoust.2011.05.018

[5] O. Bunting, J. Stammers, D. Chesmore, O. Bouzid, G. Y. Tian, C. Karatsovis, and S. Dyne, "Instrument for soundscape recognition, identification and evaluation (ISRIE): technology and practical uses," in *Euronoise 2009*, October 2009.

[6] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources in three dimensions using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, 2018. [Online]. Available: https://arxiv.org/abs/1807.00129

[7] S. Adavanne, A. Politis, and T. Virtanen, "Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network," in *EUSIPCO 2018*, 2018. [Online]. Available: https://arxiv.org/abs/1710.10059

[8] "Core sound octomic," http://www.core-sound.com/OctoMic/1.php, accessed: 2019-06-20.

[9] mh Acoustics, *em32 Eigenmike® microphone array release notes*, mh acoustics, 25 Summit Ave, Summit, NJ 07901, April 2013. [Online]. Available: https://mhacoustics.com/sites/default/files/EigenmikeReleaseNotesV15.pdf

[10] M. C. Green and D. Murphy, "Eigenscape: A database of spatial acoustic scene recordings," *Applied Sciences*, vol. 7, no. 11, p. 1204, Nov 2017. [Online]. Available: http://dx.doi.org/10.3390/app7111204

[11] ——, "Acoustic scene classification using spatial features," in *DCASE 2017*, 2017. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2017/documents/workshop_papers/DCASE2017Workshop_Green_126.pdf

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[13] B. Rafaely, *Fundamentals of spherical array processing*. Heidelberg Germany: Springer, 2015.

[14] D. P. Jarrett, "Spherical Microphone Array Processing for Acoustic Parameter Estimation and Signal Enhancement," Ph.D. dissertation, Department of Electrical & Electronic Engineering, Imperial College London, 2013.

[15] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2149–2157, Oct 2004. [Online]. Available: http://dx.doi.org/10.1121/1.1792643

[16] E. B. Saff and A. B. J. Kuijlaars, "Distributing many points on a sphere," *The Mathematical Intelligencer*, vol. 19, no. 1, pp. 5 – 11, 1997.

[17] T. McKenzie, D. Murphy, and G. Kearney, "Diffuse-field equalisation of binaural ambisonic rendering," *Applied Sciences*, vol. 8, no. 10, p. 1956, Oct 2018. [Online]. Available: http://dx.doi.org/10.3390/app8101956

[18] E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. van der Walt, M. Descoteaux, and I. Nimmo-Smith, "Dipy, a library for the analysis of diffusion mri data," *Frontiers in Neuroinformatics*, vol. 8, Feb 2014. [Online]. Available: http://dx.doi.org/10.3389/fninf.2014.00008

[19] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004. [Online]. Available: https://alex.smola.org/papers/2004/SmoSch04.pdf

[20] S. Adavanne, A. Politis, and T. Virtanen, "TUT Sound Events 2018 - Ambisonic, Anechoic and Synthetic Impulse Response Dataset," Apr. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1237703

[21] "IEEE DCASE 2016 task 2 dataset," https://archive.org/details/dcase2016_task2_train_dev, accessed: 2019-06-20.

[22] mh Acoustics, *Eigenbeam Data Specification for Eigenbeams Eigenbeam Data Specification for Eigenbeams Eigenbeam Data Specification for Eigenbeams Eigenbeam Data: Specification for Eigenbeams*, 2016. [Online]. Available: https://mhacoustics.com/sites/default/files/Eigenbeam%20Datasheet_R01A.pdf

[23] "IEEE DCASE 2019 task 3," http://dcase.community/challenge2019/task-sound-event-localization-and-detection, accessed: 2019-06-20.

[24] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, 1955.

[25] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, Atlanta, Georgia, 2013. [Online]. Available: http://proceedings.mlr.press/v28/bergstra13.pdf

[26] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kegl, "Algorithms for hyper-parameter optimization," in *Neural Information Processing Systems*, Granada, Spain, Dec 2011. [Online]. Available: https://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf