MULTIPLE SOURCE LOCALISATION IN THE SPHERICAL HARMONIC DOMAIN

Christine Evers, Alastair H. Moore and Patrick A. Naylor

Dept. of Electrical & Electronic Engineering, Imperial College London, London, SW7 2AZ, United Kingdom

ABSTRACT

Spherical arrays facilitate processing and analysis of sound fields with the potential for high resolution in three dimensions in the spherical harmonic domain. Using the captured sound field, robust source localisation systems are required for speech acquisition, speaker tracking and environment mapping. Source localisation becomes a challenging problem in reverberant environments and under noisy conditions, leading to potentially poor performance in cocktail party scenarios. This paper evaluates the performance of a low-complexity localisation approach using spherical harmonics in reverberant environments for multiple speakers. Eigenbeams are used to estimate pseudo-intensity vectors pointing in the direction of sound intensity. This paper proposes a clustering approach in which the intensity vectors of active sound sources and strong reflections are extracted, yielding an estimate of the source direction in azimuth and inclination as an approach to source localisation.

***Index Terms*—** Spherical harmonics, multiple speaker localisation, reverberation, pseudo-intensity vectors, clustering.

1. INTRODUCTION

Source localisation is an important prerequisite for many acoustic systems, such as scene analysis, room geometry inference, blind source separation, and dereverberation [1]. Most sound localisation systems target the localisation of the sound source position by suppressing room reverberation and noise in order to extract an estimate of the Direction-of-Arrivals (DOAs) of only the direct path signals. Nonetheless, reflections off walls and obstacles create a unique impulse response of a room and can hence be exploited constructively to infer knowledge about the surrounding acoustic environment.

Localisation in conjunction with room inference becomes a particularly interesting challenge in robot audition. Sound sources must be localised, identified and tracked in real time, whilst building an acoustic map of the environment for situational awareness. To provide accurate sound source localisation and mapping in both the near and far field, accurate data in three dimensions are required. Spherical microphone arrays [2] are hence a natural choice for robot audition where the array can be mounted on a robot head to capture high resolution data in the Spherical Harmonic Domain (SHD).

This paper therefore proposes a source localisation approach using spherical harmonics estimating both the DOAs of active sound sources and reverberant reflections.

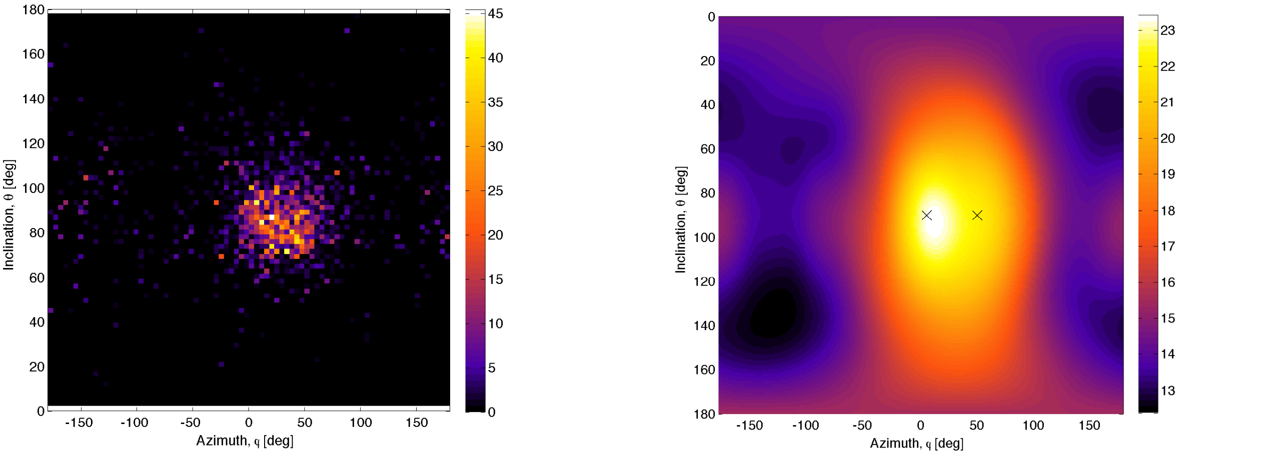
A number of source localisation approaches were proposed in the literature based on steered response power (see [3] for a review) in which a beamformer was steered in all possible directions before selecting those directions where the output fulfils the desired criteria [4]. With adequate resolution in the look directions and sufficiently narrow beam widths, the unique source DOAs can be identified. However, computational limitations in practice demand either a coarser resolution or adaptive processing to home in on dominant sources. Subspace domain methods were also formulated in the SHD [5, 6, 7] but remain computationally expensive.

In this paper we propose an approach to multiple source localisation based on feature estimation in the SHD. Previously, [8] presented a localisation algorithm for single sources by estimating the direction and magnitude of the acoustic intensity in each bin in the spherical time-frequency domain. The DOA is obtained by averaging all intensity vectors in order to estimate the source direction.

This contribution extends the approach in [8] to the estimation of multiple sound sources by clustering the intensity vectors. It is shown that multiple simultaneously active sound sources as well as strong reflections can be estimated from their intensities. A qualitative example is used to elaborate on the physical meaning of sound intensities in the context of source localisation. Intensity maps are compared to steered response power maps to highlight the suitability of the proposed approach for localisation. The quantitative evaluation of this paper demonstrates two-source DOA estimation to within 5 deg source separation under a range of simulated noise and reverberation conditions. Further results show that the algorithm is effective at localising at least five sources.

Sects. 2.1 and 2.2 provide an overview of spherical harmonics and pseudo-intensity vectors. A qualitative example is presented in sect. 2.3 to illustrate the processing of the pseudo-intensity vectors to estimate DOAs of the sources and reflections. Quantitative results are discussed in sect. 3 and conclusions drawn in sect. 4.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465



(a) Intensity map of pseudo-intensity vectors.

(b) Steered response power map, × denotes true source positions.

Fig. 1: Intensity map of pseudo-intensity vectors vs steered response power map

2. PROPOSED METHODOLOGY

2.1. Signal model

The eigenbeams, $P_{lm}(k)$, of order l and degree m, measured at a spherical microphone array can be modelled as:

$$P_{lm}(k) = \sum_{n=1}^{N_s+N_{\text{ref}}} X_{lm}(k, \mathbf{\Omega}_n)\, S_n(k) + V_{lm}(k) \qquad (1)$$

where $k = \omega/c$ is the wavenumber, N_s is the number of audio sources, N_{ref} is the number of reflections, $X_{lm}(k)$ is the sound intensity due to a unit amplitude plane wave arriving from angle $\{\mathbf{\Omega}_n\}_{n=1}^{N}$; $S_n(k)$ is the amplitude of the n^{th} plane wave; and $V_{lm}(k)$ models sensor noise and late reverberation.

The eigenbeams can also be expressed in terms of the microphone array elements at radius r_q and angle $\mathbf{\Omega}_q$ via [9]

$$P_{lm}(k) \approx \sum_{q=1}^{Q} Y_{lm}^{*}\left(\mathbf{\Omega}_q\right)\, p\left(k, r_q, \mathbf{\Omega}_q\right) \qquad (2)$$

where superscript $*$ denotes the complex conjugate, Q is the number of microphones in the spherical array, and $p(k, r_q, \mathbf{\Omega}_q)$ is the q^{th} microphone signal. $Y_{lm}(\mathbf{\Omega}_q)$ are the spherical harmonics for the microphone angles, $\{\mathbf{\Omega}_q\}_{q=1}^{Q}$:

$$Y_{lm}(\mathbf{\Omega}_q) = \sqrt{\frac{(2l+1)}{4\pi}\frac{(l-m)!}{(l+m)!}}\mathcal{L}_{lm}(\cos\theta_q)e^{\jmath m\phi_q} \qquad (3)$$

where $\mathbf{\Omega}_q = \begin{bmatrix}\phi_q & \theta_q\end{bmatrix}^T$; $\mathcal{L}_{lm}(\cos\theta)$ is the associated Legendre function; and the azimuth, ϕ_q, and inclination, θ_q, are evaluated from the Cartesian positions via $\phi_q = \arctan(y/x)$ and $\theta_q = \arccos(z/r)$ with range $r = \sqrt{x^2+y^2+z^2}$.

2.2. Pseudo-intensity vectors

The sound power per unit area is described by the acoustic intensity vector, $\mathcal{I}$, as a function of sound pressure and particle velocity. Sound intensity consists of an active and a reactive part relating to the stored parts and flow respectively of signal energy. In the far-field of an acoustic sensor, sound is described by the active sound intensity, expressed as [10]

$$\mathcal{I} = \frac{1}{2}\Re\left\{p(\mathbf{x},\omega)\,\mathbf{v}^{*}(\mathbf{x},\omega)\right\} \qquad (4)$$

where $p(\mathbf{x},\omega)$ is the sound pressure; $\mathbf{v}(\mathbf{x},\omega)$ is the particle velocity of the sound field at point, $\mathbf{x}$, and angular frequency, $\omega = 2\pi f$; and $\Re\{\cdot\}$ is the real part of the argument.

In practice, it is difficult to measure the particle velocity, $\mathbf{v}$, without specialised transducers. Pseudo-intensity vectors [8] are conceptually similar to the active sound intensity but are calculated from the zero- and first-order eigenbeams, $P_{lm}(k)$. The pseudo-intensity vector, $\mathbf{I}(k)$, is thus defined as

$$\mathbf{I}(k) = \frac{1}{2}\Re\left\{\tilde{P}_{00}(k)^{*}\mathbf{d}(k)\right\}. \qquad (5)$$

where the omnidirectional pressure at the centre of the array, $\tilde{P}_{00}$, is approximated as

$$\tilde{P}_{00}(k) = \left(\frac{P_{00}(k)}{b_0(k)}\right) \qquad (6)$$

where $b_0^{-1}(k)$ compensates for the 0-order mode strength, which is a function of the array geometry [11]. The particle velocity in each of the axial directions is approximated by a vector of dipoles steered in the negative x, y and z directions,

$$\mathbf{d}(k) = \begin{bmatrix}D_x(k) & D_y(k) & D_z(k)\end{bmatrix}^T \qquad (7)$$

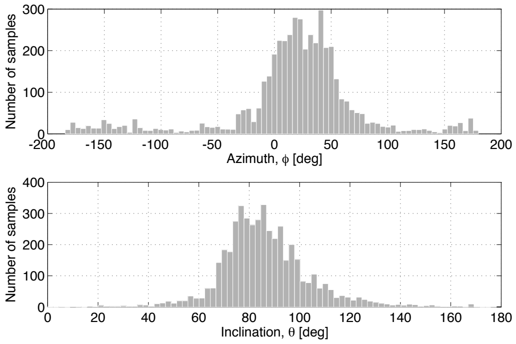


Fig. 2: 2D histogram of intensity vectors in azimuth and inclination.

where superscript T denotes the transpose. The dipoles are

$$D_a(k) = \sum_{m=-l}^{l} \frac{Y_{lm}\left(\mathbf{\Omega}_a\right)}{b_l(k)} P_{lm}(k),\ a \in x, y, z \tag{8}$$

where the steering angles $\mathbf{\Omega}_a$ are given by $\mathbf{\Omega}_x = (\pi/2, \pi)$ and $\mathbf{\Omega}_y = (\pi/2, -\pi/2)$ and $\mathbf{\Omega}_z = (\pi, 0)$, where $1/b_l(k)$ compensates for the l-order mode strength.

2.3. Processing of intensity vectors

For a qualitative example, the pseudo-intensity vectors are calculated for a simulated room with sources located at $\mathbf{\Omega}_1 = (5, 90)$ deg and $\mathbf{\Omega}_2 = (50, 90)$ deg relative to a 32-element spherical microphone array as discussed in Sect. 2.2. The weighted histogram of the resulting pseudo-intensity vectors is evaluated with a resolution of 4.5 deg in azimuth and 2.25 deg in inclination, where the vector lengths are used as histogram weights. The results are shown in the intensity map in Fig. 1a. The figure shows that a strong cluster of vectors is formed around the DOAs of the sound sources.

To elaborate on the information reflected in the intensity map, the steered response power of the scene is plotted in Fig. 1b for a resolution of 1 deg in azimuth and inclination. This is equivalent to a raster scan of the entire environment using a beamformer. The result shows a strong power field around the source location. Comparing Figures 1a and 1b, the concept of pseudo-intensity vectors hence bears a clear relation to the information reflected in steered response power maps. However, the pseudo-intensity vector requires significantly less computational effort. The computational complexity of approaches based on steered response power is generally governed by the required resolution. For instance, assuming a range in azimuth and elevation of $[-180, 180]$ deg and $[0, 180]$ deg respectively, 64800 beams are formed to achieve a 1 deg resolution. In contrast, the approach proposed in this paper is equivalent to the formation of a single beam and can hence be implemented in real-time.

Furthermore, for this example of sources separated by 45 deg, the power map in Fig. 1b only indicates one peak, rendering resolution between two sources difficult. However, the 2D plot of the intensity vector histograms in Fig. 2 shows that two peaks can be identified around the source locations.

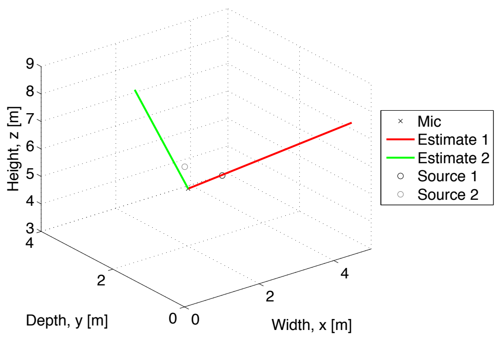


Fig. 3: Source positions vs DOA estimates.

To identify the dominant directions in the set of pseudo-intensity vectors, K-means clustering [12] is employed using randomly generated initial cluster centers (centroids). In order to capture intensity vectors due to both direct path wavefronts and reflections, the number of clusters is set to $C = 10$. The distance between each source position, $\{\mathbf{s}_n\}_{n=1}^{N_s}$, and each cluster centroid $\{\mathbf{c}_\ell\}_{\ell=1}^{C}$ is calculated as the angular distance:

$$d\left(\mathbf{s}_n, \mathbf{c}_\ell\right) = \frac{1}{\pi} \cos^{-1} \frac{\mathbf{s}_n \cdot \mathbf{c}_\ell}{\|\mathbf{s}_n\| \|\mathbf{c}_\ell\|}. \tag{9}$$

The Hungarian assignment algorithm [13] is used to extract two cluster centers that are closest to the true target positions, yielding an estimate of the source DOA. Fig. 3 shows the plot of the true target positions and their DOA estimates. Note that, in practice, the source localisation can be performed independent of any knowledge of the true target position by evaluating the model selection scheme over the number of cluster centers, e.g., by minimising the Akaike Information Criterion (AIC) over different choices of C [14].

3. RESULTS

An evaluation of the proposed algorithm is conducted using simulated data with multiple talkers active at all times and a realistic level of sensor noise. The Acoustic Impulse Responses (AIRs) of a 32-element rigid spherical microphone array to sources in 5x4x6 m shoebox room were simulated using a modification of the image-source method [15, 16, 17]. The spherical microphone array with radius 4.2 cm is placed arbitrarily at (2.54, 2.55, 4.48 m) and N_s sources are distributed on a circle of radius 1 m at the same height as the microphone. In each trial the azimuth of the first source are chosen randomly and the subsequent sources are placed at regularly spaced intervals, $\Delta\phi_s$ (45 deg in experiment 1, $5 - 180$ deg in experiments 2 and 3). Anechoic speech samples are drawn randomly for each trial from the APLAWD database [18] and convolved with the simulated AIRs. The

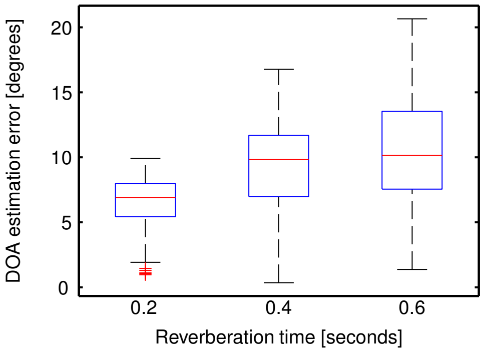


Fig. 4: Distribution of absolute angular error as a function of T_{60} for two sources 45° apart on the horizontal plane of the microphone array.

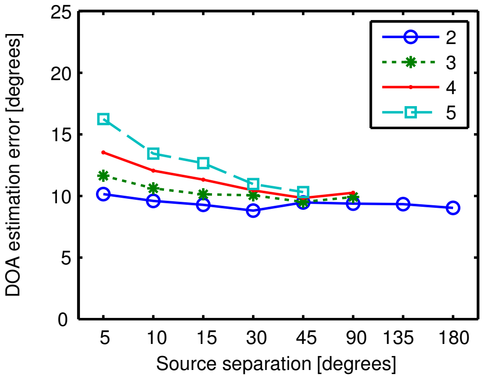


Fig. 6: Mean absolute angular error as a function of number of simultaneous sources on the horizontal plane of the microphone array and source spacing (T_{60}=0.4 s).

active level of each speech source according to ITU-T P.56 [19], as measured at $\tilde{P}_{00}$, is set to be equal. Spatio-temporally white Gaussian noise is added to the microphone signals to produce a signal to incoherent noise ratio (iSINR) at $\tilde{P}_{00}$ for each speaker of 25 dB. It should be noted that the noise at $\tilde{P}_{00}$ is approximately 10 dB lower than at the individual sensors due to the individual signals being summed.

Pseudo-intensity vectors are calculated for each time-frequency bin as described in sect. 2.2, using a sampling rate of 8 kHz, frame length of 8 ms and 50% overlap. To ensure an excess of available clusters for the maximum number of sources, the $C = 10$ centroids are estimated as described in sect. 2.3. The angular error between each centroid and source position is calculated according to eqn. (9).

In experiment 1 the effect of T_{60} is considered for two sources with a fixed angular separation of $45\,\text{deg}$. Fig. 4 shows boxplots of the angular error obtained for 2 sources with $45\,\text{deg}$ spacing for $T_{60} = \{0.2, 0.4, 0.6\}$ s. The median errors are $\{6.9, 9.8, 10.2\}\,\text{deg}$ respectively. It can be seen that the median error and the interquartile range increase with T_{60}. However, the amount of increase is relatively modest.

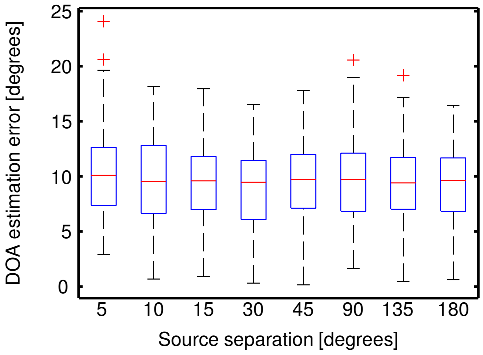


Fig. 5: Distribution of absolute angular error as a function of source spacing for two sources on the horizontal plane of the microphone array (T_{60}=0.4 s).

In experiment 2 the effect of the angular separation of two sources is investigated with $T_{60} = 0.4$ s. Fig. 5 shows boxplots of the angular error obtained for angular spacings of $\Delta\phi_s = \{5, 10, 15, 30, 45, 90, 135, 180\}\,\text{deg}$ for 2 sources with $T_{60} = 0.4$ s. The median error is approximately $10\,\text{deg}$ in all cases suggesting that with 2 sources the clustering approach is reasonably independent of $\Delta\phi_s$.

In experiment 3 the results of experiment 2 are extended to include up to 5 simultaneous sound sources. Fig. 6 shows the mean angular error obtained for $N_s = \{2, 3, 4, 5\}$ with the same values of $\Delta\phi_s$ as in experiment 2. It can be seen that, for relatively narrow spacings, increasing the number of sources tends to increase angular error. As the spacing widens to about $45\,\text{deg}$ the error tends to that obtained with 2 sources (i.e. $10\,\text{deg}$). The overall median error is between $10-13\,\text{deg}$ in all cases suggesting that the clustering approach is reasonably independent of $\Delta\phi_s$ for 2 sources. Note that the lines are truncated as the maximum spacing decreases with increasing number of sources.

4. CONCLUSIONS

This paper proposed an approach for the localisation of multiple sound sources using spherical microphone arrays in reverberant environments. Pseudo-intensity vectors were estimated and clustered to extract the DOAs of the direct path soundwaves and strong reflections. A practical example demonstrated the conceptual similarity between pseudo-intensity vectors and steered response power maps [3] and highlighted the advantage of pseudo-intensity vectors in terms of computational requirements. Multiple experiments showed robust source localisation with T_{60} up to 0.6 s, and for up to 5 sources.

5. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[2] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[3] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *J. Acoust. Soc. Am.*, vol. 131, pp. 2828–2840, 2012.

[4] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[5] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2005, vol. 3, pp. iii/89–iii/92.

[6] H. Teutsch and W. Kellermann, "Detection and localization of multiple wideband acoustic sources based on wavefield decomposition using spherical apertures," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2008, pp. 5276–5279.

[7] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds., chapter 11. Springer, Jan. 2010.

[8] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Simulating room impulse responses for spherical microphone arrays," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 129–132.

[9] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.

[10] M. J. Crocker, Ed., *Handbook of Acoustics*, Wiley-Interscience, 1998.

[11] J. Meyer and G. W. Elko, "Position independent close-talking microphone," *Signal Processing*, vol. 86, no. 6, pp. 1254–1259, June 2006.

[12] J. Wu, *Advances in K-means Clustering*, Springer Theses. Springer, Berlin Heidelberg, 2012.

[13] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1995.

[14] S. T. Buckland, K. P. Burnham, and N. H. Augustin, "Model selection: An integral part of inference," *Biometrics*, vol. 53, no. 2, pp. 603–618, June 1997.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[16] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sept. 2012.

[17] D. P. Jarrett, "Spherical Microphone array Impulse Response (SMIR) generator," http://www.ee.ic.ac.uk/sap/smirgen/.

[18] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Technical report, University College London, June 1987.

[19] ITU-T, "Objective measurement of active speech level," Dec. 2011.