Audio Engineering Society
Convention Paper 10493
Presented at the 150th Convention
2021 May 25–28, Online

A Steered-Beamforming Method for Low-Latency Direction-of-Arrival Estimation in Reverberant Environments Using Spherical Microphone Arrays

Jonathan Mathews[1] and Jonas Braasch[1]

[1]*Architectural Acoustics, Rensselaer Polytechnic Institute, Troy, New York 12180, United States of America*

Correspondence should be addressed to Jonathan Mathews (mathej4@rpi.edu)

ABSTRACT

This paper introduces a method to estimate the direction of arrival of an acoustic signal based on finding maximum power in iteratively reduced regions of a spherical surface. A plane wave decomposition beamformer is used to produce power estimates at sparsely distributed points on the sphere. Iterating beam orientation based on the orientation of maximum energy produces accurate localization results. The method is tested using varying reverberation times, source-receiver distances, and angular separation of multiple sources and compared against a pseudo-intensity vector estimator. Results demonstrate that this method is suitable for integration into real-time telematic frameworks, especially in reverberant conditions.

1 Introduction

Direction of Arrival (DOA) estimation is a vital element in myriad acoustic array processing applications, such as source localization, acoustic mapping, and speech enhancement. Spherical microphone arrays (SMAs) are excellent platforms for DOA estimation, since their compact footprint and directional independence confer a great deal of flexibility in application compared to linear or planar geometries [1, 2]. Several methods have been developed in the past few decades to localize sound sources using SMA data, but few have been tailored to address the constraints of telematic applications in reverberant spaces.

Telematic systems ideally produce low-latency streams of location data for real-time use. A localization method must be computationally efficient, since it may contribute to the overall complexity of a larger tracking and broadcast system. It must also be accurate despite reverberation or noise, since spaces like classrooms and multi-purpose spaces have large volumes, highly variable reverberation times [3, 4], and multiple, dynamic sources.

This paper introduces the sparse iterative search (SIS), which is a steered-beam power discrimination process, to address source localization for distant sources in reverberant spaces. Section 2 contains a brief background of SMA operation. Section 3 introduces the SIS. Finally, Section 4 provides an evaluation of the performance of SIS as compared to the pseudointensity vector (PIV) method [5] across a variety of tests. Although other methods for localization, including extensions to the PIV method itself, are more sophisticated and produce accurate results, the original PIV method was chosen in particular for its computational efficiency and simplicity of technique, which makes it

a suitable baseline for comparison

2 Technical Background

2.1 Spherical Harmonics

The localization methods discussed in this paper are dependent on efficient manipulation of the soundfield in spherical space. The following is a brief review of the theory of spherical harmonic decomposition. A thorough introduction to SMA theory and operation may be found in [6, 7].

A point in space is given in spherical coordinates (r,ϕ,θ), where r is the radius, ϕ is azimuth, and θ is elevation. A pressure field is measured at this point, $p(k,r,\theta,\phi)$, with wave number k. The spherical harmonic representation of the sound pressure at this point may be obtained by the spherical Fourier transform

$$p_{lm}(k,r) = \int_0^{2\pi}\int_0^{\pi} p(k,r,\theta,\phi)Y_l^m(\theta,\phi)^* \sin\theta d\theta d\phi, \tag{1}$$

where $(\cdot)^*$ is the complex conjugate. Y_l^m are the spherical harmonics of degree l and order m, given by:

$$Y_l^m = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_m^l(\cos(\theta))e^{im\phi}, \tag{2}$$

where P_m^l is the associated Legendre function, and $i=\sqrt{-1}$. A plane wave may be described incident upon the surface of a sphere with radius r_q using this spherical harmonic representation:

$$p_{lm}(k,r) = A(k)b_l(kr_q)Y_l^m(\theta,\phi), \tag{3}$$

where A is the amplitude and b_l is the modal gain, or mode strength. For a rigid spherical array, this modal gain term is described by

$$b_l(kr) = 4\pi i\big(j_l(kr) - \frac{j_l'(kr)}{h_l^{(2)}(kr)}h_l^{(2)}(kr)\big). \tag{4}$$

j_l is the spherical Bessel function, and $h_l^{(2)}$ is the spherical Hankel function of the second kind. Compensating for this term in the array output and incorporating a steering vector produces a plane-wave decomposition (PWD) beamformer, which spatially filters the array data to exclude soundfield information except in the direction of interest:

$$y(k) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{p_{lm}(k,r)}{b_l(kr)}Y_l^m(\theta_d,\phi_d), \tag{5}$$

where (θ_d,ϕ_d) is the orientation of interest. For a microphone array with Q discrete elements with locations given by (r_q,ϕ_q,θ_q), the spherical Fourier transform is an approximate sum over the pressure values from each element over the surface of the sphere:

$$p_{lm}(k,r_q) \approx \sum_{q=1}^{Q} p(k,r_q,\theta_q,\phi_q)Y_l^m(\theta_q,\phi_q)^*. \tag{6}$$

This approximation limits the order of harmonic decomposition, constraining the directivity pattern of the PWD beamformer:

$$y(k,\phi_q,\omega_q) = \sum_{l=0}^{L}\sum_{m=-l}^{l}\frac{p_{lm}(k,r)}{b_l(kr)}Y_l^m(\theta_d,\phi_d). \tag{7}$$

2.2 Steered Response Power Mapping

The steered response power (SRP) method for localization produces spatial maps of sound power by means of a grid search using a steered beamforming array. Although evaluation of the SRP method is not included, the methodology is given here for completeness. The map is generated with the output of the PWD beamformer and a set of all spherical grid points to be searched, Ω_S, in terms of azimuth and elevation:

$$\mathcal{M}(\Omega_S) = \sum_k |y(k,\Omega_S)|^2. \tag{8}$$

A single source may be localized by finding the maximum value in the map,

$$\Omega_{\max} = \arg\max_{\Omega_S}\mathcal{M}(\Omega_S), \tag{9}$$

whereas a peak detection algorithm may negotiate multiple-source discrimination for particular conditions.

The SRP method is accurate even under reverberant conditions [8], since its operation is dependent on a grid of steered beams. Because these beams are spatial filters, the response from room reflections and directional noise is greatly reduced except in the beams oriented along their directions of arrival. Ideally, an infinite-order beam would produce a response along only its oriented direction. However, for order-limited beamformers, the response from each beam corresponds to an angular arc surrounding the orientation of interest. This reduces the overall accuracy of SRP localization, but provides the basis of functionality for the SIS method.

3 Sparse Iterative Search

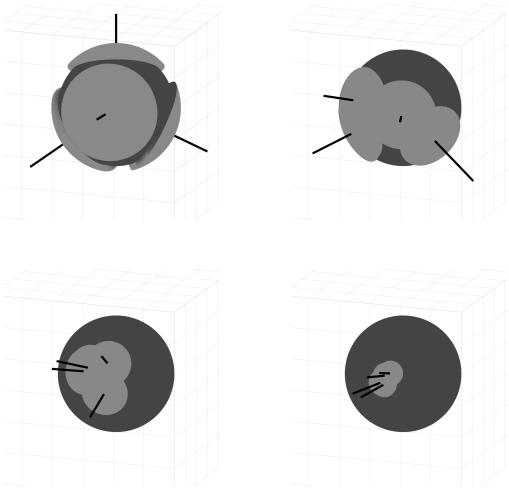

Fig. 1: A visual representation of the SIS method for 4 beam steering vectors and 4 iterations. Steering vector orientations are denoted by black lines, and search sectors are the light grey regions over the sphere in dark grey. Initially, beam orientations are evenly distributed around the sphere (a). On subsequent iterations (b–d) the beam orientation vectors are chosen via uniform sampling from the best of several search sectors corresponding to each beam.

SIS takes advantage of three major characteristics of speech energy distribution and spatial filtering to produce accurate DOA estimation. First, the energy distribution of a single frame of speech data is most likely to be convex upwards. Only a small percentage of frames of audio data, as would be generated for the Short-Time Fourier Transform, contain salient speech feature. This condition holds even for audio containing multiple simultaneous speakers, a principle known as Window-Disjoint Orthogonality [9].

Second, despite a beampattern's sensitivity to a region of space rather than a single orientation, only slight changes in orientation are capable of affecting the recorded energy. This feature is what allows the SRP method to generate more accurate results with increased grid density.

The final characteristic is the PWD beamformer itself. Although an ideal beam is a delta function, finite-series spherical harmonics produce hypercardioid beampatterns. The output of the beamformer corresponds to a spatial area rather than a single direction. This allows sufficient sensitivity to energy distribution over the sphere with a small number of generated beampatterns.

This final feature is what distinguishes SIS from existing region-contraction methods [10, 11]. Rather than contract a search region around volumes containing large energy values, each beam is used to define its own search region, and subsequent iterations reject spatial regions while progressively converging on the orientation of maximum energy.

For a single time frame of speech data, the initial steering vectors for each beam are chosen such that the beams are evenly distributed over the sphere and maximally distant in orientation from each other. For a circular distribution around the equator, the angular arc between all beam steering vectors is $2\pi/N$, where N is the number of beams used per iteration. For spherical distributions, orientations derived from the Platonic solids or nearly-uniform spacing distributions [12] may be used.

The orientation that returns the maximum power from the PWD beamformer is obtained using

$$\Omega_{\max} = \arg\max_{\Omega_S} \mathcal{M}(\Omega_S), \tag{10}$$

where $\mathcal{M}$ is the beamformer output over the set of orientations. Conceptually, up to this point, the process is very similar to the SRP mapping method. However, the set of beam orientations are not a dense grid, but an extremely sparse set of points over the sphere. When the orientation of maximum power is found, a new set of steering vectors are selected via uniform random sampling from within the spherical section centered on Ω_s with conical angle $c = 2\pi/N$. Subsequent iterations repeat this random selection operation, generating a set

of orientations from successively smaller search sectors corresponding to the beam which produces maximum power according to $c = 2\pi/kN$, where k is the current iteration count. As the energy distribution is assumed convex, this approach converges to the global maximum of the energy distribution for a single time frame, which, ideally, corresponds to the source orientation relative to the array. A visualization of this method for four beams and four iterations is shown in Figure 1.

4 Evaluation

A variety of metrics were used to characterize the performance of this method. Experiments were performed on both simulated and recorded array data with varying reverberation time (RT), source-receiver distance, and angular separation for multiple sources.

4.1 Simulated Data

For each simulation, room impulse responses (IRs) were generated using SMIRGen [13], in a $16 \times 14 \times 10\,\mathrm{m}^3$ room with a virtual 16-channel array with 2.5 cm radius located at the center. Two-second segments of anechoic speech recordings from the Archimedes Project [14] were convolved with the generated IRs to produce virtual microphone audio data. Average Direct-to-Reverberant Ratio (DRR) values for the IRs with a source-receiver distance of 4 m are shown in Table 1. DRR values for varying source-receiver distance with an RT of 0.4 s are shown in Table 2

T60 (s)	0.4	0.6	0.8	1.0	1.5	2.0
DRR (dB)	11.72	6.69	4.27	2.91	1.02	0.02

Table 1: Direct-to-reverberant ratios for the selected reverberation times used in this study

S-R (m)	3	4	5	6	7
DRR (dB)	10.72	8.58	6.85	5.78	4.4

Table 2: Direct-to-reverberant ratios for the selected source-receiver distances used in this study for a T60 of 0.4 s

For all simulated trials, a sampling rate of 8 kHz, a frame length of 32 samples, and 50% overlap between frames were used. White Gaussian noise was added to the audio data to produce a signal-to-noise ratio of 25 dB. One hundred trials were generated for each test condition. To evaluate localization error, the SIS algorithm is performed with 3 iterations per frame of audio data, and 12 beam orientations generated per iteration. The PIV method uses only 0th- and 1st-order eigenbeams, while the SIS algorithm was further evaluated for 1st-, 2nd-, and 3rd-order beams.

Error values were generated by comparing the true source orientation in Cartesian coordinates relative to the array, $\mathbf{u}$ to the estimated orientation generated by each method $\hat{\mathbf{u}}$ using

$$\varepsilon = \cos^{-1}(\mathbf{u}^T \hat{\mathbf{u}}). \tag{11}$$

For scenarios with multiple sources present, the relative error between each source orientation and the estimate was computed, and the minimum value was chosen. This method of evaluating error was chosen to achieve parity with other error estimation techniques seen in the literature [5, 15].

Experiment 1 evaluates SIS against the PIV method for RTs varying from 0.4 s to 2 s. The source-receiver distance is fixed at 4 m, and source separation for multiple sources is 45°. For each test condition, a control trial was produced using randomly generated orientation data. Figure 2 shows the error distribution, where the black dot denotes median, the boxes show upper and lower quartiles, and the whiskers extend to the 5th and 95th percentile range. The localization error range for the SIS algorithm is less than half the error obtained by the PIV method for T60 values over 0.8 s for a single source condition. The control trial error is approximately 90° for a single source, 70° for two sources, and 53° for three sources.

Experiment 2 fixes the RT to 0.4 s and varies the source-receiver distance from 3 m to 7 m, while maintaining angular separation between sources at 45°. Results are shown in Figure 3. A control trial of random data was generated for this experiment as well, but is not displayed in the plot for clarity. Error values for the control trial are comparable to those in Experiment 1. For distances greater than 5 m, SIS produces a reduction in error of 15° relative to PIV.

In experiment 3, RT values of 0.4 s and 2 s were used, and the source-receiver distance is 4 m while the angular separation varies from 15° to 180°. Figure 4 shows the number of sources estimated by counting the number of peaks in the spatial spectrum of DOA estimates.

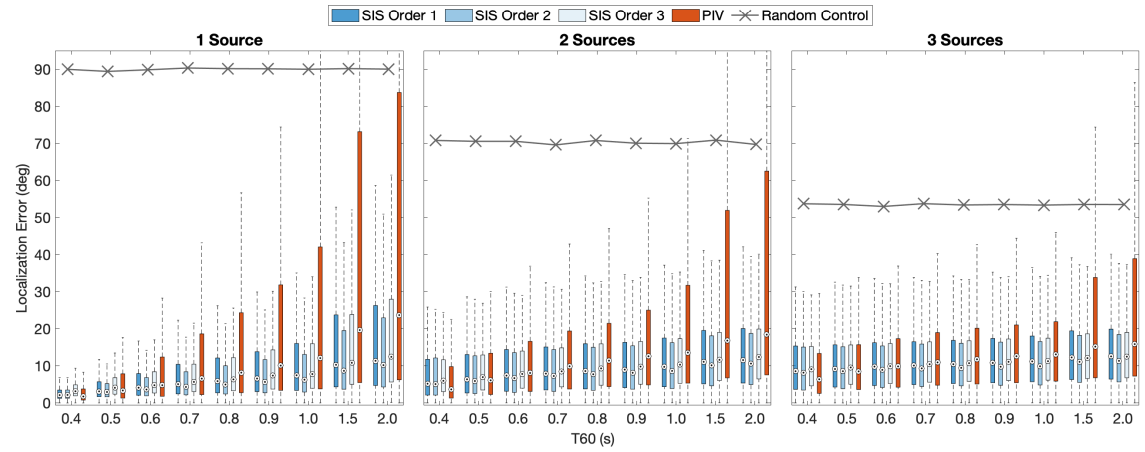

Fig. 2: Comparison of localization error between SIS and the PIV method for T60 varying between 0.4 s and 2 s (color online). Simulated trials were carried out for 1-3 simultaneously active sources. Source-receiver distance was 4 m. The marked grey line denotes average error from randomly generated data to demonstrate stochastic influence on the representation of accuracy due to an increase in ground truth data.

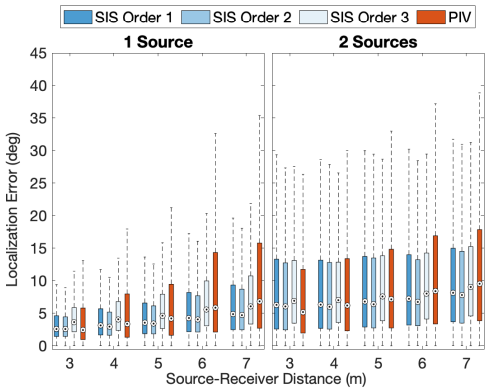

Fig. 3: Comparison of localization error for source-receiver distances varying from 3 m to 7 m under a T60 of 0.4 s (color online). To preserve axis limits for clarity, the line denoting angular error generated by random data is omitted. For a single source, the random trials averaged 90° error for a single source, and 70° error for two sources.

For this scenario, SIS shows no significant improvement in performance over PIV for low RT. However, in the 2 s RT case, PIV fails to produce distinct peaks, leading to zero sources counted, whereas SIS accuracy remains similar to the 0.4 s case.

4.2 Real-World Data

Evaluation in real-world conditions took place in a large multi-purpose room with dimensions $16.1 \times 13.7 \times 5.6\,\mathrm{m}^3$ and broadband RT of 0.89 s. Two anechoic speech recordings were used from the Archimedes project. Both audio files source were broadcast simultaneously from a semi-rectangular speaker array with dimensions $12 \times 10\,\mathrm{m}^2$ positioned at the height of 1.7 m. A 16-channel SMA was located in the center of the room, positioned at the same height as the speaker array. Source 1 maintained a stationary position at the 0° azimuth point relative to the microphone at a distance of 5 m. Source 2 traveled in a 180° arc relative to the microphone at a velocity of approximately $0.9\,\mathrm{m{\cdot}s}^{-1}$, varying radial distance from 5 m at the closest point to the array to 7.1 m at its most distant. A diagram of the experimental setup is shown in Figure 5. Playback of the audio files over the speaker array and recording of microphone array data was performed using standard studio-grade audio hardware operating at a sampling frequency of 48 kHz, and frame size of 128 samples, or 2.6 ms.

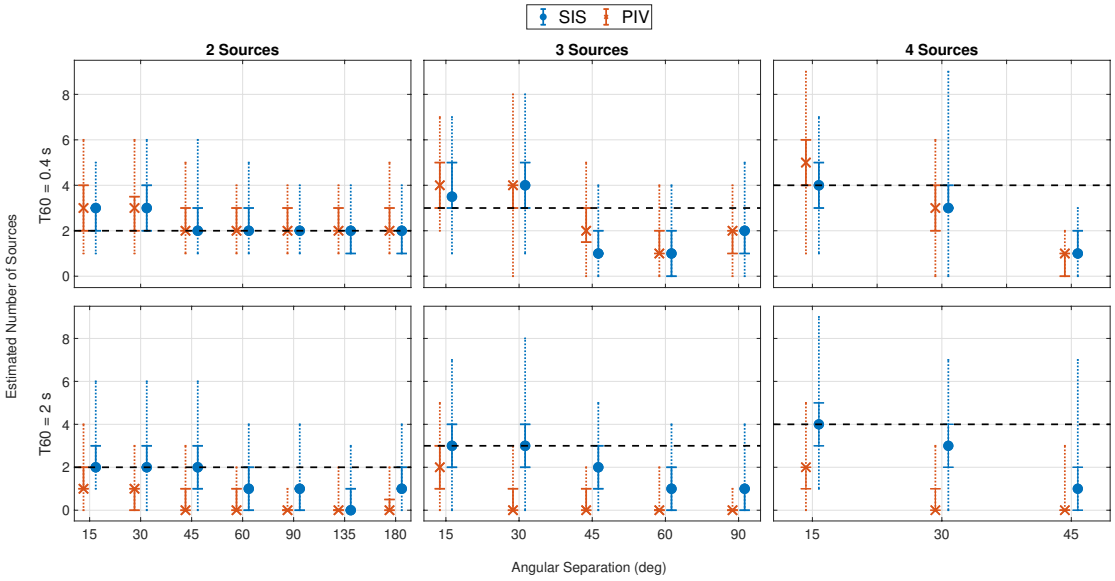

Fig. 4: Comparison of SIS against PIV for varying angular separation between multiple simultaneous sources (color online). The markers denote average number of sources detected over 100 trials for each test condition. The solid error bars show the interquartile range, while the dashed bars show the range of sources detected. The correct number of sources is shown by the horizontal dashed line. T60 is 0.4 s for the top row of figures, and 2.0 s for the bottom row. Source-receiver distance is 4 m.

To demonstrate the potential utilization of clustering algorithms with the SIS method, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was used, with a neighborhood radius of 5° and minimum point density of 30. Figure 6 shows the azimuthal localization performed by the SIS algorithm with source identification via DBSCAN over time.

4.3 Computational Complexity

The PIV localization method requires the fewest operations to localize a sound source, since only one zeroth-order and three first-order eigenbeams are required to generate the vector. The SIS algorithm is variable in terms of complexity – the number of beams chosen and the number of iterations performed per time frame dictate the number of operations performed. Modern computing hardware and practices allow for a large variation in efficiency in executing operations; therefore a relative evaluation of processing time per frame of data is used to evaluate the cost of these methods. The results in Figure 7 show the ratio of processing time of SIS to PIV for varying iterations with 12 beams of selected harmonic orders processed per iteration. The average of 100 trials for each SIS test condition was compared against a 100-trial average of PIV. Processing time for the SIS method is within a single order of magnitude of PIV performance for most cases. The minimum value obtained is a ratio of 7.3 for the 1st-order case and 2 iterations, and a ratio of 130.9 for the 3rd-order case and 15 iterations. Figures 8 and 9 show heat maps of the average angular error and relative latency, respectively, given the number of beams and iterations used, with an RT of 0.4 s and source-receiver distance of 4 m. Relatively few beams and iterations are required for localization to converge with the minimum angular error as dictated by array geometry, which allows us to select parameters that preserve accuracy and minimize latency. The values presented in these figures were produced by evaluating the computation time using MATLAB R2020b on a 2013 Apple MacBook Pro with an Intel Core i7 processor with a clock speed of 2.3 GHz. The algorithm will have significantly lower latency values when operating within a proper real-time processing framework.

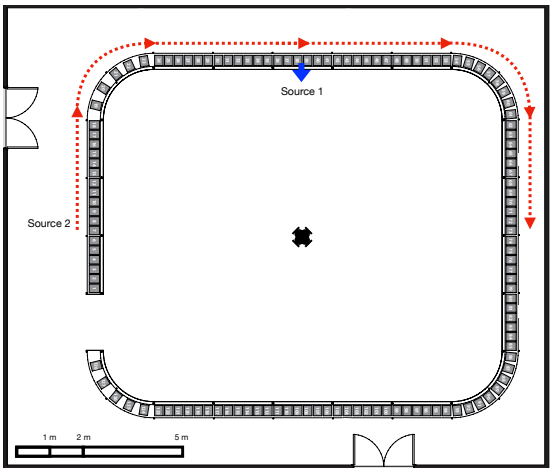

Fig. 5: Visual representation of the test environment (color online). Experimentation was performed at the CRAIVE-Lab at Rensselaer Polytechnic Institute in Troy, NY. The arrow labeled "Source 1" marks the position of the stationary sound source, while the dashed line labeled "Source 2" shows the direction of motion of the moving sound source around the speaker. array.

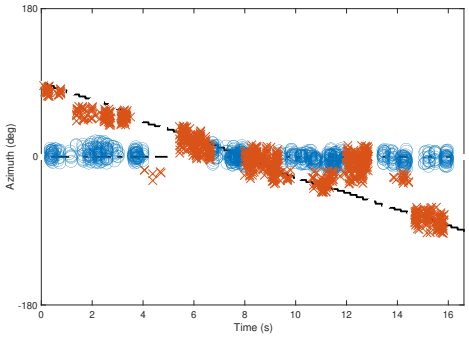

Fig. 6: SIS localization of speech in a real environment (color online). Results are displayed here as a map of azimuth over time. The raw localization data was processed using the DBSCAN algorithm to produce rudimentary point clustering and source counting. The black lines denote ground truth trajectory of each source. Gaps in the lines denote periods of voice inactivity. The 'x' and 'o' markers are the clustered localization estimates. These two markers are used to distinguish where multiple clusters are identified concurrently.

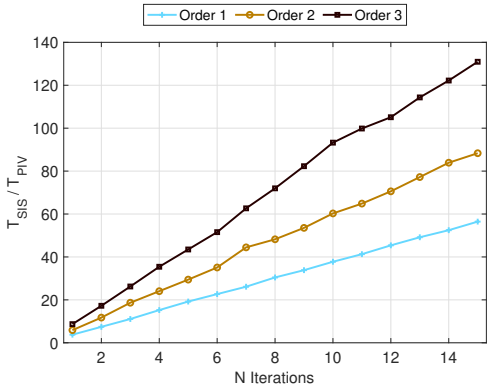

Fig. 7: Ratio of SIS computing time to PIV for varying number of iterations in a single time frame. To compute SIS performance, N iterations were used, and 12 beam patterns of 1st-, 2nd-, or 3rd-order were evaluated during each iteration. Each data point indicated is an average of 100 simulations.

5 Discussion

In the context of the presented research, it is important to understand how the different methods, SIS, PIV, and SRP, compare in terms of accuracy, robustness and computational efficiency. The results show that the SIS method trades computational speed for more accurate performance under high reverberation. In this sense, it may be viewed as a balance between SRP- and PIV-based methods. Comparison in other literature [5, 15, 16] of SRP mapping against other DOA estimation techniques has revealed its favorable performance under reverberant conditions, but also its dependence on grid density for accuracy and considerable computational load. Although it is more complex than the PIV method in terms of operations required, the computational latency incurred by SIS iteration is no greater than an order of magnitude of PIV computational speed for most cases, as seen in Figure 7.

Despite the efficiency of the PIV method, it is inherently susceptible to room reflections, significantly decreasing accuracy under even moderate reverberation, as seen in Figure 2. Room reflections produce stochastic variation in the direction of flow. The influence of this stochastic behavior is apparent in the localization results under varying reverberation when comparing the PIV method with the random control trial. Although derivatives of this method address this shortcoming

[17], the additional processing significantly increases the computational load and reduces the quantity of estimates generated over time.

The latency produced by SIS in MATLAB analysis is below 1 millisecond per frame for certain conditions, enabling continuous real-time operation. The low-latency localization data stream produced is suitable for further processing, such as incorporation within clustering algorithms, such as K-Means or DBSCAN, as seen in Figure 6.

Time-Frequency analysis, such as the coherence test [18] or the Direct Path Dominance test [19], to further improve multiple-source discrimination is also easily applicable, as it is for the PIV method.

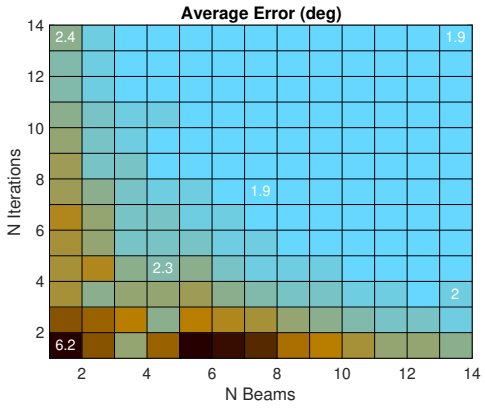

Fig. 8: Average error for varying number of steered beams generated and number of iterations (color online). Lighter regions denote lower error values, while darker regions correspond with high error. Selected trials have numerical values represented in degrees.

6 Conclusions

In this study, it was shown that the SIS algorithm is less sensitive to reverberation when compared to the PIV method. Accurate estimation may be achieved using a small number of beam orientations and iterations. The increased computational load for such cases relative to PIV is significant, but marginal given the additional processing required to improve PIV performance. SIS also maintains source identification performance in reverberant conditions. Incorporation of more sophisticated clustering and source counting methods may further improve identification accuracy.

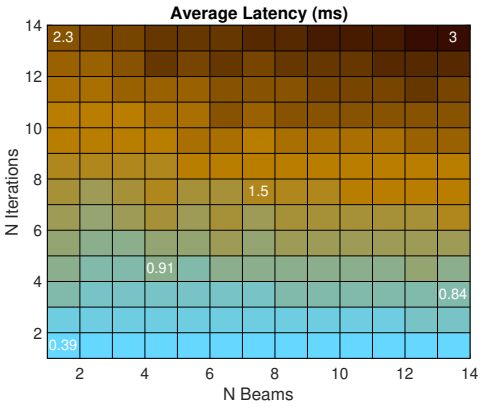

Fig. 9: Average latency for varying number of steered beams generated and number of iterations (color online). Lighter regions denote lower latency values, while darker regions correspond with high latency. Selected trials have numerical values represented in milliseconds. Numerical values displayed were generated in MATLAB and are expected to be significantly lower when using the algorithm within a real-time framework.

7 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants #1631674 & #1909229, the RPI Cognitive and Immersive Systems Laboratory, and the RPI Humanities, Arts, and Social Sciences Fellowship.

References

[1] Meyer, J. and Elko, G. W., "A spherical microphone array for spatial sound recording," *The Journal of the Acoustical Society of America*, 111(5), pp. 2346–2346, 2002.

[2] Abhayapala, T. D. and Ward, D. B., "Theory and design of high order sound field microphones using spherical microphone array," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. II–1949, IEEE, 2002.

[3] Bradley, J. S., "Speech intelligibility studies in classrooms," *The Journal of the Acoustical Society of America*, 80(3), pp. 846–854, 1986.

[4] Knecht, H. A., Nelson, P. B., Whitelaw, G. M., and Feth, L. L., "Background noise levels and reverberation times in unoccupied classrooms," *American Journal of Audiology*, 2002.

[5] Jarrett, D. P., Habets, E. A., and Naylor, P. A., "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *2010 18th European Signal Processing Conference*, pp. 442–446, IEEE, 2010.

[6] Williams, E. G., *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Elsevier, Amsterdam, 1999.

[7] Rafaely, B., *Fundamentals of Spherical Array Processing*, volume 8, Springer, Berlin, 2015.

[8] DiBiase, J. H., Silverman, H. F., and Brandstein, M. S., "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180, Springer, 2001.

[9] Rickard, S. and Yilmaz, O., "On the approximate W-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. I–529, IEEE, 2002.

[10] Do, H., Silverman, H. F., and Yu, Y., "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 1, pp. I–121, IEEE, 2007.

[11] Do, H. and Silverman, H. F., "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 295–298, IEEE, 2007.

[12] Fliege, J. and Maier, U., "The distribution of points on the sphere and corresponding cubature formulae," *IMA Journal of Numerical Analysis*, 19(2), pp. 317–334, 1999.

[13] Jarrett, D., Habets, E., Thomas, M., and Naylor, P., "Rigid sphere room impulse response simulation: algorithm and applications," *The Journal of the Acoustical Society of America*, 132(3), pp. 1462–1472, 2012.

[14] Hansen, V. and Munch, G., "Making recordings for simulation tests in the Archimedes project," *Journal of the Audio Engineering Society*, 39(10), pp. 768–774, 1991.

[15] Hafezi, S., Moore, A. H., and Naylor, P. A., "Augmented intensity vectors for direction of arrival estimation in the spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10), pp. 1956–1968, 2017.

[16] Çöteli, M. B., Olgun, O., and Hacıhabiboglu, H., "Multiple sound source localisation with steered response power density and hierarchical grid refinement," *in IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2018.

[17] Moore, A. H., Evers, C., and Naylor, P. A., "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1), pp. 178–192, 2016.

[18] Mohan, S., Lockwood, M. E., Kramer, M. L., and Jones, D. L., "Localization of multiple acoustic sources with small arrays using a coherence test," *The Journal of the Acoustical Society of America*, 123(4), pp. 2136–2147, 2008.

[19] Nadiri, O. and Rafaely, B., "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*, 22(10), pp. 1494–1505, 2014.