J. Mathews and J. Braasch, "Sparse Iterative Beamforming Using Spherical Microphone Arrays for Low-Latency Direction of Arrival Estimation in Reverberant Environments"
J. Audio Eng. Soc., vol. 69, no. 12, pp. 967–977, (2021 December).
DOI: https://doi.org/10.17743/jaes.2021.0057

Sparse Iterative Beamforming Using Spherical Microphone Arrays for Low-Latency Direction of Arrival Estimation in Reverberant Environments

JONATHAN MATHEWS, *AES Student Member,* **AND JONAS BRAASCH,** *AES Member*
(mathej4@rpi.edu) (braasj@rpi.edu)

Rensselaer Polytechnic Institute, Troy, NY

Acoustic direction of arrival estimation methods allows positional information about sound sources to be transmitted over a network using minimal bandwidth. For these purposes, methods that prioritize low computational overhead and consistent accuracy under non-ideal conditions are preferred. The estimation method introduced in this paper uses a set of steered beams to estimate directional energy at sparsely distributed orientations around a spherical microphone array. By iteratively adjusting beam orientations based on the orientation of maximum energy, an accurate orientation estimate of a sound source may be produced with minimal computational cost. Incorporating conditions based on temporal smoothing and diffuse energy estimation further refines this process. Testing under simulated conditions indicates favorable accuracy under reverberation and source discrimination when compared with several other contemporary localization methods. Outcomes include an average localization error of less than 10° under 2 s of reverberation time (T60) and the potential to separate up to four sound sources under the same conditions. Results from testing in a laboratory environment demonstrate potential for integration into real-time frameworks.

0 INTRODUCTION

Positional information about speech sources within a space is a vital property for internet-based audio applications. Such information may be used in object-based audio encoding schemes to efficiently reproduce the spatial qualities of a sound scene [1]. For scenarios such as teleconferencing, real-time positional information is required to accurately convey the scene as it evolves. Acoustic Direction of Arrival (DOA) estimators are ideal tools for generating this data, since they achieve high temporal resolution compared to their visual counterparts, and do not require computationally expensive models to infer acoustic activity [2]. Several microphone array geometries have been developed for the purpose of sound source localization, but spherical microphone arrays (SMAs) are particularly interesting since they are compact, are capable of generating high-resolution spatial data, and maintain their performance characteristics over the entire spherical surface [3, 4].

In the past couple decades, a variety of DOA estimation algorithms have been developed for applications like acoustic mapping, speech enhancement, or robot audition. The proliferation of smart devices and sophisticated telecommunications technology has broadened the scope of application for these algorithms beyond conference rooms into far more dynamic classrooms and multipurpose spaces, where reverberation and noise characteristics can vary wildly [5, 6]. Such contexts require an algorithm that is capable of producing high temporal accuracy and low error variance to capture the varying characteristics of speech activity despite non-ideal conditions.

Acoustic DOA estimation methods can be broadly sorted into grid search, region contraction, subspace analysis, and intensity vector. Grid search is performed by producing a uniform set of coordinates over the surface of interest and then measuring the power at each coordinate using a spatial filter. The Steered Response Power (SRP) method is the simplest grid-based technique. Although computationally expensive, it is robust to room reverberation [7].

Region-contraction techniques aim to lower the computational cost of grid search by either progressively reducing the search region or decomposing it into multiple smaller

regions. Coarse-to-Fine Region Contraction (CFRC) [8] performs estimation by first evaluating a coarse grid over the total search region for response power and then selecting a subset of grid coordinates with the largest response power values to define the boundaries of a reduced search region. Iterating these steps while increasing grid density as the search region is reduced produces accurate DOA estimates with fewer beams than a single search at a fixed resolution. Stochastic Region Contraction (SRC) [9] operates similarly, but beam steering vectors are randomly generated within the search region, rather than selected from a grid.

Subspace analysis methods like Multiple Signal Classification (MUSIC) [10] and Eigenbeam Estimation of Signal Parameters with Rotationally Invariant Techniques (EB-ESPRIT) [11] take advantage of the correlations between input vectors to extract signal parameters. Eigenbeam (EB)-MUSIC [12] and EB-ESPRIT [13] correlate subsets of the signals in the spherical harmonics domain to directly extract DOA information. Unlike EB-MUSIC, which requires a grid search, EB-ESPRIT directly obtains directional parameters from the array data, which affords significant computational efficiency. These methods tend to fail in the presence of room reflections because of rank-reduction in the spatial covariance matrix caused by coherent signals, though more recent formulations have attempted to address this [14–16].

Finally intensity vector methods utilize the pressure and particle velocity to determine the direction of energy flow [17]. Pseudo-intensity vector (PIV) methods approximate particle velocity with directional pressure measurements [18], negating the need for particle velocity sensors. The PIV method for spherical arrays [19] is extremely efficient in its use of eigenbeams for directional pressure measurement, and extensions further improve accuracy and efficiency [20, 21]. However, under multiple-source conditions, moderate reverberation, and large source-receiver distances, the performance of the PIV method degrades considerably [22].

This paper introduces the sparse iterative search (SIS) technique. This method aims to address the trade-off perceived in the above techniques, where algorithms may be efficient but fail in highly reverberant conditions or are acoustically robust but computationally expensive. Instead, this method utilizes modal beamforming to both produce accurate DOA estimates under high reverberation and incur minimal latency. After the technical background is addressed, the methodology will be discussed and compared to several other contemporary localization methods under a variety of conditions to assess performance.

1 TECHNICAL BACKGROUND

1.1 Spherical Harmonics

Efficient SMA operation is dependent on spherical harmonic decomposition. A brief review is given here, but a thorough introduction to SMA theory and operation may be found in [23] and [24]. A point in spherical coordinates is defined (r, ϕ, θ), where r is the radius, ϕ is azimuth, and θ is elevation. A pressure field at this point is described by $p(k, r, \theta, \phi)$, with wave number k. Using the Spherical Fourier Transform produces the spherical harmonic representation of the sound pressure at this point,

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

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

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

where P_m^l is the associated Legendre function, and $i = \sqrt{-1}$.

An incident plane wave upon a sphere with radius r_q can be described with this spherical harmonic representation,

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

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

$$b_l(kr) = 4\pi i \left(j_l(kr) - \frac{j_l\prime(kr)}{h_l^{(2)}\prime(kr)} h_l^{(2)}(kr) \right), \tag{4}$$

where j_l is the spherical Bessel function, $h_l^{(2)}$ is the spherical Hankel function of the second kind, and $(\cdot)'$ denotes the derivative.

A plane-wave decomposition beamformer is created by compensating for the modal gain and incorporating a steering vector composed of spherical harmonics. This spatially filters the signal from the array so that the array output is only acoustic information along the orientation 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 an SMA with Q discrete elements located at (r_q, ϕ_q, θ_q), the Spherical Fourier Transform is an approximate sum over 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. Although an infinite-order beamformer would produce a delta function along the orientation of interest, the order-limited case is described by

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

where L is the maximum harmonic order, and describes a main lobe with non-zero angular width.

An SRP map is generated by computing steered beams over a grid of angular coordinates over the surface of interest. The map may be described using

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

where $\mathscr{M}$ is the beamformer output over the set of orientations and Ω_S is a set of all grid coordinates to be searched in terms of azimuth and elevation. A single source is localized by finding the maximum value in the output map,

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

while multiple sources may be detected by using a peak-detection algorithm.

Since this technique is based on spatial filtering, array response at a grid point contains reduced noise from room reflections. An array capable of infinite-order harmonic decomposition could theoretically produce exact DOA coordinates but would require infinitely many spatial filters to do so. An order-limited array has reduced accuracy because of wider beampatterns, but fewer beams are required to fully map the surface. This provides part of the motivation for the SIS method.

2 SPARSE ITERATIVE SEARCH

SIS takes advantage of several basic properties to localize sources efficiently. First a frame of speech data, as taken by the Short-Time Fourier Transform, is most likely to have a spatial energy distribution that is convex upwards, even with multiple speakers present [25]. Second small changes in the steering vector produce changes in the energy recorded by a beam, with more drastic changes in higher-order beams. This allows steered-beam search techniques to improve in accuracy with increased grid density. Finally order-limited beams have a wider spatial pattern, which allows low-resolution sensitivity to the energy over the entire spherical surface with a small number of beams. The utilization of all these features is what differentiates SIS from other region contraction methods. Each beam defines its own search region, and each iteration selects the region with maximum energy for continued refinement.

For each frame of audio data, initial steering vectors are pre-selected to produce uniform coverage based on either equal-angular arc ($2\pi/N$, where N is the number of beams used), the Platonic solids, or nearly-uniform distributions [26]. The DOA of maximum energy is returned via Eq. (9). This is an extremely sparse form of the SRP method until this point. When the direction of maximum energy is returned, a new set of beams is generated using uniform random sampling within a spherical section centered on $\Omega_{\max}$ with conical angle $c = 2\pi/N$. Again the orientation of maximum energy is found, and the search region described by c is further reduced according to $c = 2\pi/IN$, where I is the iteration number. Fig. 1 demonstrates this process.

To further refine the accuracy of this process, two additional functions are used. First a primary beam is selected, in addition to the search beams, to be steered along the direction returned from the prior frame of audio data. By averaging the results of F previous frames and rejecting frames that produce large angular deviations D over short time intervals, a form of temporal smoothing that reduces the variance of the results may be achieved. Second, on the initial grid-based estimation step, a coarse estimate of diffuse energy over the sphere may be measured by taking the average energy recorded by all beams. Accuracy may then be further improved by rejecting frames that do not exceed a direct-to-diffuse ratio threshold R, as measured by comparing the maximum recorded energy to the average, since these frames are more likely to be dominated by room reflections.

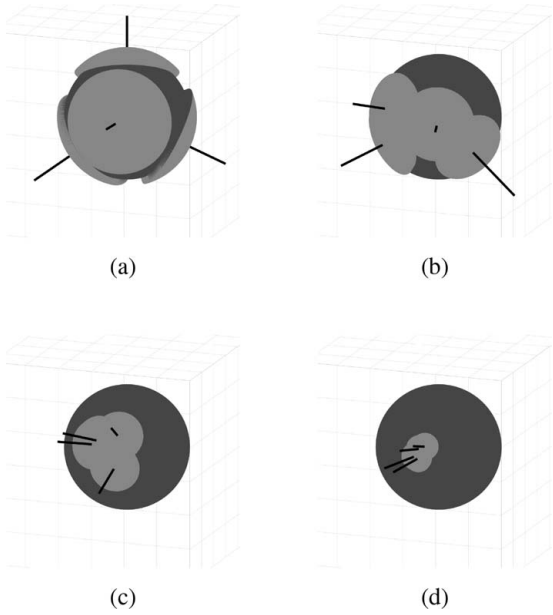

Fig. 1. A visual representation of the Sparse Iterative Search (SIS) method for four beam steering vectors and four iterations. Steering vector orientations are denoted by black lines, and search sectors are the light gray regions over the sphere in dark gray. Initially beam orientations are evenly distributed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

A step-by-step breakdown of the method is:

1. **Initialize:** Initialize N beams, I iterations, F frames in memory buffer, deviation threshold D, and energy threshold R.
2. **Generate:** Generate N uniformly spaced beams over sphere.
3. **Test:** If D or R criteria are not met, discard frame.
4. **Iterate:** Find $\Omega_{\max}$. In section $c = 2\pi/IN$ centered on $\Omega_{\max}$, generate new N beams. Repeat for I iterations.
5. **Return:** Save $\Omega_{\max}$ in buffer and keep result.

3 EVALUATION

The SIS algorithm was tested under simulated conditions to measure performance under varying reverberation time (RT), source-receiver distances, and angular separa-

Table 1. Direct-to-reverberant (DRR) ratios for the selected reverberation times (T60) used in this study.

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 2. Direct-to-reverberant ratios (DRR) for the selected source-receiver (S-R) distances used in this study for a T60 of 0.4 s.

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

tion with multiple sources. A real-time system was also developed to demonstrate practical operation.

3.1 Simulated Data

For all simulated trials, Spherical Microphone Array Impulse Response Generator [27] was used to produce room impulse responses (IRs). A 16 × 14 × 10–m room was generated, and a virtual 16-channel array with 2.5-cm radius was positioned at the center. The dimensions of the room and array geometry were chosen to achieve parity with the physical equipment and laboratory space used for real-world testing.

Anechoic speech recordings were taken from the Archimedes Project [28], of which 2-s segments were convolved with the IRs to simulate static speech sources within the room. White Gaussian noise was added to the simulated array data to lower the SNR to 25 dB to account for array noise. Since significant spatial aliasing [24] occurs at frequencies higher than 4 kHz, for this array design [29], simulations were processed at a sampling frequency of 8 kHz with a frame size of 32 samples and 50% overlap between frames. The Direct-to-Reverberant Ratio (DRR) for the varying test conditions are shown in Tables 1 and 2 for reverberation and source-receiver distances respectively. To produce error estimates, the true position of the source(s) relative to the virtual array, $\mathbf{u}$, was compared to the estimated position, $\hat{\mathbf{u}}$, using

$$\epsilon = \cos^{-1}(\mathbf{u}^T\hat{\mathbf{u}}). \tag{10}$$

To optimize SIS performance, trials were run to determine parameters that maximize accuracy while minimizing latency. Figs. 2 and 3 show the respective average angular error and processing time per frame of audio data as the number of iterations and generated beams are varied. The values were produced using 1,000 estimates generated from the simulated dataset under 0.4 s reverberation time (T60). Cases with only a single beam or iteration were omitted, since a single beam will not provide differential energy estimates, and SIS using a single iteration is simply a sparse SRP search.

These figures indicate that accuracy is dependent on both beams and iterations, and only a small number of beams and iterations are required to produce low estimation error. However the computational cost is dependent on the number of iterations performed. Optimal performance, with low latency and high accuracy, may then be achieved using a small number of iterations with a larger number of beams. Therefore two iterations ($I = 2$) and 12 first-order beams ($N = 12$) were used for all of the following experiments.

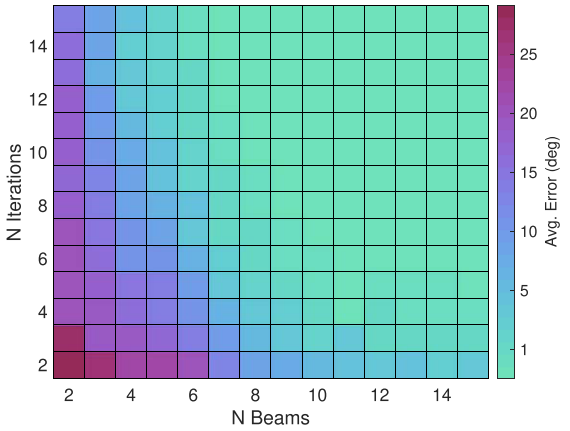

Fig. 2. Average error for varying number of steered first-order beams generated and number of iterations (color online). Lighter regions correspond to lower error, and darker regions indicate higher error.

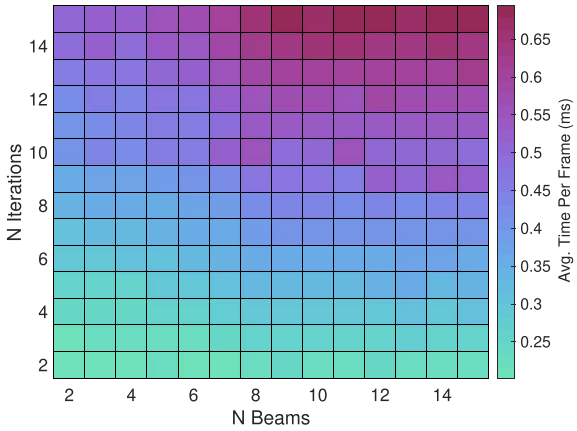

Fig. 3. Average frame processing time varying with number of steered first-order beams and iterations (color online). Lighter regions correspond to low latency, and dark regions indicate high latency. Values were generated in MATLAB and are expected to be much lower for real-time applications.

To evaluate the smoothing operation, 100 trials were run on the simulated dataset using 2.0-s RT (T60), with 1,000 frames per trial for each value of D and F. The average interquartile range of the estimation error and percentage of estimates returned relative to the total number of frames in each trial were calculated. Results are shown in Fig. 4.

Although the error range is small for small values of D, the percentage of estimates returned in each trial is very low, regardless of the number of frames stored in memory. Larger values of D ($>25^\circ$) are less effective at reducing error variance but return a higher percentage of frames, especially for smaller F values. This evaluation indicates that there is a trade-off in spatial versus temporal accuracy dependent on the value of D chosen, mitigated by select-

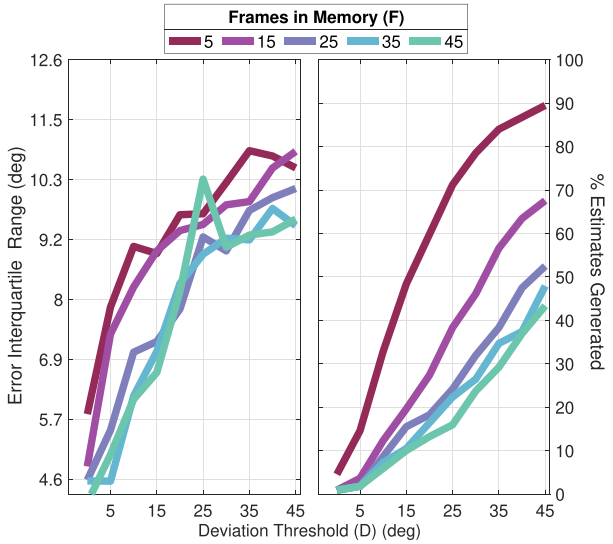

Fig. 4. Average interquartile error range and percentage of estimates generated relative to total frames of audio data for varying deviation threshold and frames used for smoothing (color online). Each line denotes a different number of frames in memory.

ing a small F. With this in mind, all subsequent experiments were conducted using a memory buffer of five frames ($F = 5$) to store and average prior coordinate data, corresponding to a time interval of 20 ms. The deviation threshold D was set to 17°.

SIS is evaluated against several other DOA estimation methods to benchmark its performance. SRP represents the basic grid-search technique, and CFRC and SRC are region-contraction algorithms. Each of these methods used second-order beampatterns. A 5° grid spacing (2,592 beams) was used for SRP since improvements in accuracy under the given test conditions are marginal and increased grid density results in excessive computational cost [30].

CFRC and SRC are region-contraction techniques that generate coarse maps of the spherical surface and iteratively reduce the search region by retaining points that return the largest power values and then using the points to describe a smaller region containing the source DOA. CFRC uses a coordinate grid to generate steering vectors, while SRC randomly generates steering vectors. CFRC had initial parameters of 162 grid points (based on a 20° grid spacing), and 50 retained points for region definition, with the number of retained points decreasing per iteration and grid points varying to a minimum value of 49 after five iterations based on grid density. Similarly, SRC was initialized with 100 randomly generated orientations and 50 retained points for region definition, varying to a minimum of 67 orientations after five iterations. These steered-beam methods used a threshold based on the range of beam power values recorded to minimize responses during reverberation.

A recent formulation of the EB-ESPRIT algorithm [16], which addresses error due to ambiguity and singularities at the array poles, was used to represent subspace localization methods. The PIV method was also utilized in its original form [19]. The Direct-Path Dominance (DPD) test [31] is a technique for selecting audio data that contains information from a single source by separating frames into frequency bins and averaging them over time. DPD evaluates the rank of the correlation matrix of the audio data in each time-frequency bin via Singular Value Decomposition, selects bins with reduced rank, and then performs DOA estimation in order to improve the accuracy of the estimates generated. To demonstrate the improvement this technique creates, PIV, EB-ESPRIT, and SIS were also wrapped into the DPD test. Smoothing was done over four successive time frames, with each frame divided into six frequency frames. Frames were selected for DOA estimation if the ratio between the first two singular values of the correlation matrix was greater than 6. These are referred to respectively as DPDPIV [32], DPDESPRIT, and DPDSIS.

Experiment 1 [Fig. 5(a)] tested angular error under varying RT, while Experiment 2 [Fig. 5(b)] tested angular error for varying source-receiver distances. Both experiments were performed using a single speech source. For each RT or source-receiver distance, 100 trials were run for each algorithm, with 1,000 frames of speech data per trial. A control test was generated using random data as a benchmark, though this was omitted from Experiment 2 for clarity. The average control error was approximately 90°, independent of RT or source-receiver distance. For both figures, the variation of estimation error is shown, where black dots denote the median, solid bars show the interquartile range, and dashed whiskers extend to the fifth and 95th percentile range. Based on these results, all methods produce favorable results at small distances or low RT, with EB-ESPRIT and PIV producing the lowest average error. As distance or RT increases, SIS continues to perform well, producing an interquartile error range of 9.4° under 2 s of reverberation.

Experiment 3 (Fig. 6) tested discrimination of multiple sources. Source-receiver distance was set to 4 m, RT values of 0.4 and 2 s were chosen, and angular separation of sources varied from 15°–180°. For each trial, a histogram was produced over 1,000 frames of speech data, and the number of peaks, corresponding to clusters of estimates with similar DOA values, was counted. One hundred trials were performed for each test case using the selected algorithms. The SRC and CFRC methods were omitted for clarity since their performance is generally comparable to SRP. For this scenario, all methods perform similarly under low reverberation. However, in the 2-s RT case, SIS consistently identifies distinct peaks, while other methods fail to do so for most cases.

To account for developments in incorporating frequency resolution into localization techniques, the DPD test was performed for the same reverberation dataset generated for Experiment 1. The DPDSIS, DPDPIV, and DPDESPRIT methods incorporate this time-frequency smoothing operation. Results are shown in Fig. 7. DPDPIV produces an increase in localization error variance at low RT but a significant reduction at higher RT, while DPDESPRIT maintains low-RT accuracy. DPDSIS shows consistent results over all RTs.

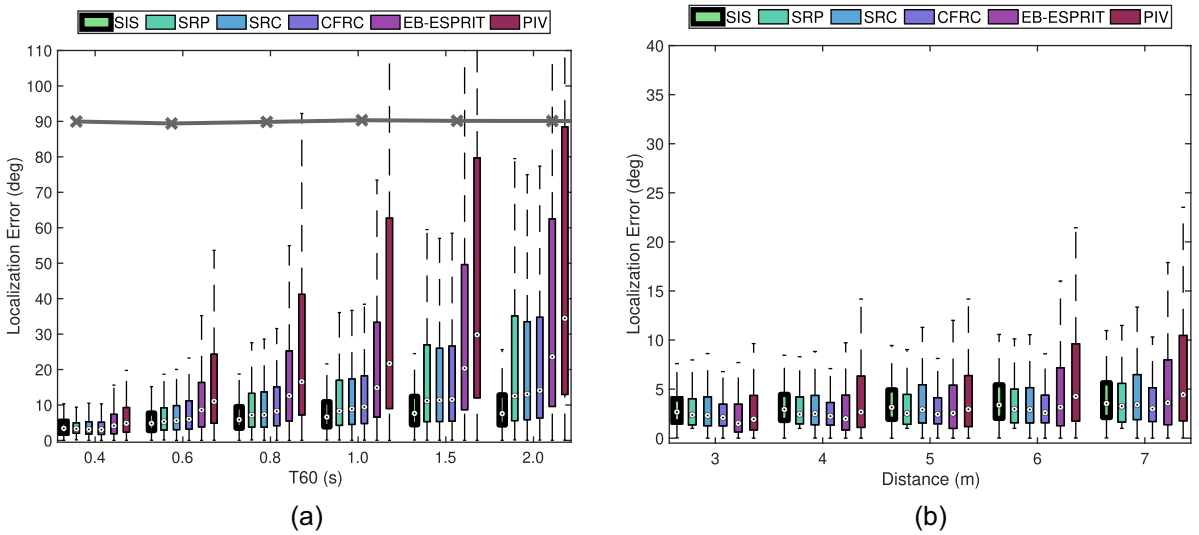

Fig. 5. Comparison of localization error between selected methods for (a) Reverberation Time (RT) (T60) varying between 0.4 and 2 s and (b) Source-Receiver Distance from 3–7 m (color online). Average error from randomly generated estimates is used as a control trial, and is indicated by the marked gray line. The line has similar values in both plots, though it is omitted from (b) due to y-axis scaling for clarity. CFRC, Coarse-to-Fine Region Contraction; EB-ESPRIT, Eigenbeam Estimation of Signal Parameters with Rotationally Invariant Techniques; PIV, Pseudo-intensity Vector; SIS, Sparse Iterative Search; SRC, Stochastic Region Contraction; SRP, Steered Response Power.

Fig. 8 uses the same datasets as the above trials but shows the percentage of total processed samples that produced an orientation estimate within 10° of the source position. This demonstrates the prevalence of both large angular deviations in the set of estimates produced and dropped frames due to smoothing techniques. Methods using the DPD test discard a vast majority of frames but return a consistent percentage of estimates regardless of RT. The other methods show a significant drop in percentage as reverberation increases, although SIS is the only method to exceed 50% at 2.0 s RT (T60), which indicates dense clusters of estimates over time.

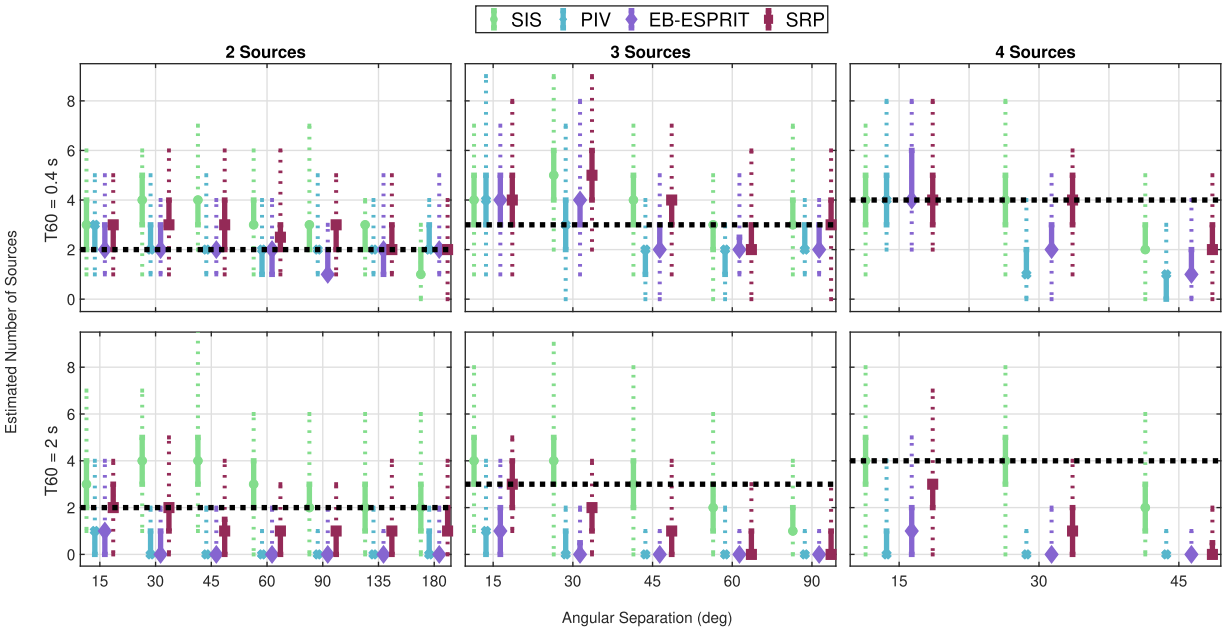

Fig. 6. Comparison between selected methods for varying angular separation between multiple simultaneous sources (color online). The markers denote the average number of sources detected over 100 trials for each test condition. Solid bars indicate the interquartile range, and dashed lines denote the range. The horizontal dashed line is the correct number of sources for each condition. EB-ESPRIT, Eigenbeam Estimation of Signal Parameters with Rotationally Invariant Techniques; PIV, Pseudo-intensity Vector; SIS, Sparse Iterative Search, SRP, Steered Response Power.

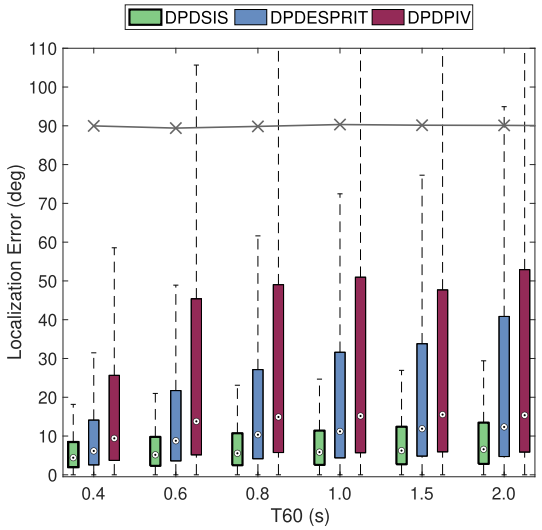

Fig. 7. Comparison between SIS, PIV, and EB-ESPRIT using the Direct-Path Dominance (DPD) test for T60 varying between 0.4 and 2 s (color online). The markers denote average number of sources detected over 100 trials for each test condition. Average error from randomly generated estimates is used as a control trial and indicated by the marked gray line. EB-ESPRIT, Eigenbeam Estimation of Signal Parameters with Rotationally Invariant Techniques; PIV, Pseudo-intensity Vector; SIS, Sparse Iterative Search.

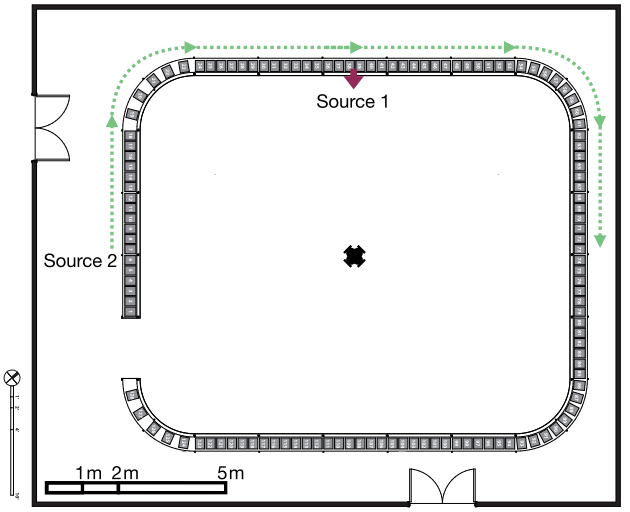

Fig. 9. Representation of the test environment (color online). Experimentation was performed at the Collaborative Research-Augmented Immersive Virtual Environment (CRAIVE) Lab at Rensselaer Polytechnic Institute in Troy, New York. The arrow labeled "Source 1" marks the position of the stationary sound source, while the dashed line labeled "Source 2" shows the direction of motion of the moving sound source around the speaker array.

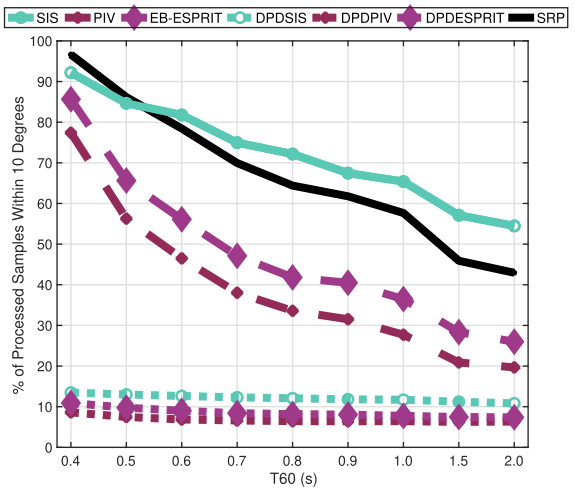

Fig. 8. Percentage of total frames that returned Direction of Arrival (DOA) estimates within 10° of source position. One hundred trials were run for each Reverberation Time (RT) (T60) value. DPD, Direct-Path Dominance; EB-ESPRIT, Eigenbeam Estimation of Signal Parameters with Rotationally Invariant Techniques; PIV, Pseudo-intensity Vector; SIS, Sparse Iterative Search; SRP, Steered Response Power.

3.2 Real-World Data

Real-world testing took place at the CRAIVE Lab at Rensselaer Polytechnic Institute, within a multi-purpose space measuring 16.1 × 13.7 × 5.6 m and broadband RT of 0.89 s. Two speech audio files, taken from the Archimedes project, were broadcast simultaneously from a semi-rectangular speaker array measuring 12 × 10 m at a height of 1.7 m. Source 1 maintained a stationary position at the 0° azimuth point relative to the microphone at a distance of 5 m, while Source 2 traveled in a 180° arc relative to the microphone at a velocity of approximately 0.9 m · s^{-1}, varying radial distance from 5 m at the closest point to the array to 7.1 m at its most distant point. The moving source was panned across the array using Vector-Base Amplitude Panning [33], with the virtual source positioned relative to the array such that no more than two speakers were simultaneously active.

A 16-channel SMA was positioned at the center of the room at the same height as the speaker array and was used to generate real-time DOA estimates using the SIS algorithm. The algorithm was implemented in C++, using identical parameters as in the MATLAB simulations, and incorporated into a patch developed in Cycling74's Max 8. A diagram of the experimental setup is shown in Fig. 9. Six thousand four hundred frames of audio data were recorded with standard audio capture hardware at a sampling rate of 48 kHz and frame size of 128 samples.

Results are shown in Fig. 10. The overall average error was 8.4°, with an interquartile range of 9.1° and 75th percentile of 11.4°. The algorithm returned a DOA estimate for 64% of all frames processed. Two simultaneously active speech sources with varying angular separation generate large angular deviations in the set of estimates returned, which are then filtered by the smoothing operation. Speech segments cover relatively long periods of time compared to the audio sample rate, which allows a sufficient number of estimates to be generated to capture salient speech information, despite a large number of dropped frames.

Table 3. Relative computational cost of DOA estimation methods.

Method	SRP	PIV	SIS	EB-ESPRIT	DPDSIS	DPDESPRIT	DPDPIV	SRC	CFRC
Cost (% SRP)	100	4.1	25.7	17.6	66.2	51.4	40.4	57.6	62.5
Cost (× PIV)	24.3	1	6.3	4.3	16.1	12.5	9.8	14.0	15.2

EB-ESPRIT, Eigenbeam Estimation of Signal Parameters with Rotationally Invariant Techniques; DPD, Direct-Path Dominance; PIV, Pseudo-intensity Vector; SIS, Sparse Iterative Search; SRP, Steered Response Power; SRC, Stochastic Region Contraction; CFRC, Coarse-to-Fine Region Contraction.

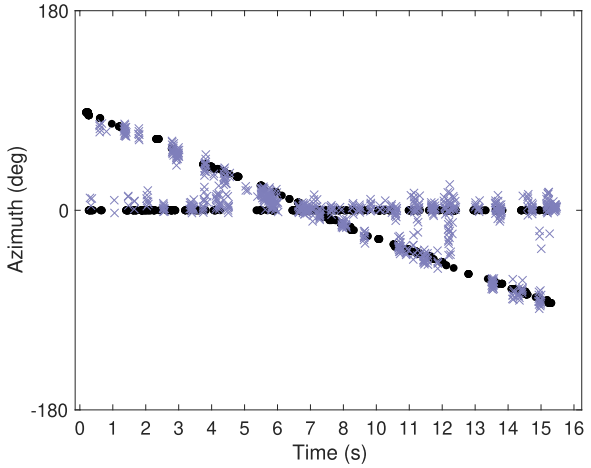

Fig. 10. Sparse Iterative Search (SIS) localization of speech in a real environment (color online). Results are displayed here as a map of azimuth over time. The black lines denote ground truth trajectory of each source. Gaps in the lines denote periods of voice inactivity. The X markers are the localization estimates.

3.3 Computational Complexity and Efficacy

Of all the algorithms used in this study, the PIV method requires the fewest operations to compute—only one zeroth-order and three first-order eigenbeams. SRP is the slowest, since a beam must be generated for each orientation in the search grid. A 5° search grid requires 2,592 beams per frame of audio data, while a 1° grid would require 64,800 beams. Region-contraction algorithms are variable in their complexity but still require hundreds of beams to describe and assess the search volume. The SIS algorithm is variable in terms of complexity—the N beams and I iterations performed per time frame dictate the number of operations performed. EB-ESPRIT does not require beam generation, but evaluation of the spatial correlation matrix and inverse matrix calculation add to the computational cost.

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 compare the cost of these methods. Table 3 shows the average processing time for all the methods utilized in this study. For clarity, results are shown both as a percentage of processing time relative to SRP and as a factor relative to PIV. The values presented were produced by evaluating 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 SRP, CFRC, and SRC methods were processed using a pre-compiled library of spherical harmonic coefficients for beam generation to maximize speed for comparison.

SRP and PIV are the slowest and fastest methods, respectively. The region-contraction methods are significantly faster than SRP but are still an order of magnitude slower than PIV. SIS and EB-ESPRIT are comparable in their performance. The performance advantage of PIV and EB-ESPRIT disappears when incorporated with the DPD test, since SIS is less costly than DPDESPRIT and DPDPIV. DPDSIS is about as fast as the region-contraction methods, although the improvement in the accuracy of estimates is minimal.

4 DISCUSSION

At a distance of 4 m, the arc length described by an angle of 10° is approximately 0.7 m. This angular threshold is sufficient for differentiating the DOA of one individual from another in all except the most intimate of circumstances. Figs. 5(a) and 5(b) indicate that all estimation algorithms are capable of performing this differentiation for low RT, with EB-ESPRIT and PIV generating the most accurate results. However, under high reverberation, these methods fail to produce precise estimates, since EB-ESPRIT is dependent on the relative incoherence of incident waves and PIV inherently produces a vector sum of all incident waves.

The large variance in estimates produced by these methods under high RT impedes generation of distinct clusters of estimates over time, leading to failure in differentiating multiple sources, as seen in Fig. 6. Although the DPD test greatly reduces estimation error for these methods under high RT (Fig. 7), both the time-frequency differentiation and Singular Value Decomposition operations for each bin incur significant computational load, which negates the performance advantage these methods have over beamforming-based techniques, as seen in Table 3.

SRP, CFRC, and SRC produce far more accurate estimates compared to EB-ESPRIT and PIV in high-RT scenarios because of their reliance on beamforming to find power from a set of given steering vectors. Despite this, precision is reduced significantly if no additional processing is done to remove errors on reverberation-dominant frames. For SRP, the evaluation of the complete search region results in a high computational load, and additional processing would only increase it. Although CFRC and SRC have a reduced load relative to SRP, the number of beams required to iteratively define the search region still incurs significant cost.

The SIS algorithm benefits from the steered-beamformer approach and further improves precision and accuracy of DOA estimation by incorporating a smoothing filter to reduce the effect of reverberation on performance. Only a small number of frames are required to effectively reduce the variance in estimation error (Fig. 4). The number of beams and iterations required is relatively small (Figs. 2 and 3) compared to the region-contraction techniques since steering vectors are not used for volume definition. This produces computational performance closer to that of EB-ESPRIT and PIV.

Furthermore, accuracy is only minimally improved by application of the DPD test. Without it, a far greater number of precise estimates may be generated over time, as seen in Fig. 8. This affords temporally and spatially dense clusters of estimates on salient speech features, enabling differentiation between multiple simultaneously active speech sources, as shown by both Figs. 6 and 10. However, because SIS is not a mapping method and provides no adaptive mechanism for search region definition, it is ill-suited for scenarios with multiple continuous sound sources, such as noise or musical instruments. In addition, because it is a beamforming-based method, sources in extremely close proximity may not be differentiated, unlike EB-MUSIC or EB-ESPRIT, which are capable of generating high-resolution maps. This is demonstrated in Fig. 10, where the paths of the two simultaneous sources intersect. Despite these caveats, the computational performance and resilience to non-ideal conditions indicate utility in a wide variety of spaces. This method is ideally suited for real-time applications involving speech data in reverberant environments, such as classrooms or multipurpose spaces.

5 CONCLUSION

Sparse Iterative Search is a conceptually simple localization algorithm that utilizes iterated steered-beam power estimates to identify Direction of Arrival of a sound source. It produces computational performance on par with more sophisticated algorithms like EB-ESPRIT while retaining accurate performance under high reverberation like other steered-beamforming methods. These factors, combined with simplicity of implementation within audio frameworks like Max, are ideal for integration and usage in large-volume immersive spaces that may benefit from real-time acoustic localization systems.

6 ACKNOWLEDGMENT

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

7 REFERENCES

[1] J. Herre, H. Purnhagen, J. Koppens, et al., "MPEG Spatial Audio Object Coding—The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes," *J. Audio Eng. Soc.*, vol. 60, no. 9, pp. 655–673 (2012 Oct.).

[2] J. Herre, C. Falch, D. Mahane, et al., "Interactive Teleconferencing Combining Spatial Audio Object Coding and DirAC Technology," *J. Audio Eng. Soc.*, vol. 59, no. 12, pp. 924–935 (2012 Feb.).

[3] J. Meyer and G. W. Elko, "A Spherical Microphone Array for Spatial Sound Recording," *J. Acoust. Soc. Am.*, vol. 111, no. 5, p. 2346 (2002 May).

[4] T. D. Abhayapala and D. B. Ward, "Theory and Design of High Order Sound Field Microphones Using Spherical Microphone Array," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1949–1952 (Orlando, FL) (2002 May). https://doi.org/10.1109/ICASSP.2002.5745011.

[5] J. S. Bradley, "Speech Intelligibility Studies in Classrooms," *J. Acoust. Soc. Am.*, vol. 80, no. 3, pp. 846–854 (1986 Sep.). https://doi.org/10.1121/1.393908.

[6] H. A. Knecht, P. B. Nelson, G. M. Whitelaw, and L. L. Feth, "Background Noise Levels and Reverberation Times in Unoccupied Classrooms," *Am. J. Audiol.*, vol. 11, no. 2, pp. 65–71 (2002 Dec.). https://doi.org/10.1044/1059-0889(2002/009).

[7] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust Localization in Reverberant Rooms," in M. Brandstein and D. Ward (Eds.), *Microphone Arrays*, Digital Signal Processing, pp. 157–180 (Springer, Berlin, Germany, 2001). https://doi.org/10.1007/978-3-662-04619-7_8.

[8] H. Do and H. F. Silverman, "A Fast Microphone Array SRP-PHAT Source Location Implementation Using Coarse-to-Fine Region Contraction (CFRC)," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 295–298 (New Paltz, NY) (2007 Oct.). https://doi.org/10.1109/ASPAA.2007.4392976.

[9] H. Do, H. F. Silverman, and Y. Yu, "A Real-Time SRP-PHAT Source Location Implementation Using Stochastic Region Contraction (SRC) on a Large-Aperture Microphone Array," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 121–124 (Honolulu, HI) (2007 Apr.). https://doi.org/10.1109/ICASSP.2007.366631.

[10] R. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280 (1986 Mar.). https://doi.org/10.1109/TAP.1986.1143830.

[11] R. Roy and T. Kailath, "ESPRIT-Estimation of Signal Parameters via Rotational Invariance Techniques," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, no. 7, pp. 984–995 (1989 Jul.). https://doi.org/10.1109/29.32276.

[12] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical Microphone Array Beamforming," in I. Cohen, J. Benesty, and S. Gannot (Eds.), *Speech Processing in Modern Communication: Challenges*

and Perspectives, Springer Topics in Signal Processing, vol. 3, pp. 281–305 (Springer, Berlin, Germany, 2010). https://doi.org/10.1007/978-3-642-11130-3_11.

[13] H. Teutsch and W. Kellermann, "Detection and Localization of Multiple Wideband Acoustic Sources Based on Wavefield Decomposition Using Spherical Apertures," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 5276–5279 (Las Vegas, NV) (2008 Apr.). https://doi.org/10.1109/ICASSP.2008.4518850.

[14] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust Localization of Multiple Sources in Reverberant Environments Using EB-ESPRIT With Spherical Microphone Arrays," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 117–120 (Prague, Czech Republic) (2011 May). https://doi.org/10.1109/ICASSP.2011.5946342.

[15] A. Herzog and E. A. P. Habets, "On the Relation Between DOA-Vector Eigenbeam ESPRIT and Subspace Pseudointensity-Vector," in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5 (A Coruña, Spain) (2019 Sep.). https://doi.org/10.23919/EUSIPCO.2019.8902715.

[16] B. Jo and J.-W. Choi, "Parametric Direction-of-Arrival Estimation With Three Recurrence Relations of Spherical Harmonics," *J. Acoust. Soc. Am.*, vol. 145, no. 1, pp. 480–488 (2019 Jan.). https://doi.org/10.1121/1.5087698.

[17] F. J. Fahy, *Sound Intensity* (Spon Press, London, UK, 1995), 2nd ed.

[18] Y. Yamasaki and T. Itow, "Measurement of Spatial Information in Sound Fields by a Closely Located Four-Point Microphone Method," *J. Acoust. Soc. Jpn. (E)*, vol. 10, no. 2, pp. 101–110 (1989). https://doi.org/10.1250/ast.10.101.

[19] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D Source Localization in the Spherical Harmonic Domain Using a Pseudointensity Vector," in *Proceedings of the 18th European Signal Processing Conference (EUSIPCO)*, pp. 442–446 (Aalborg, Denmark) (2010 Aug.).

[20] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of Arrival Estimation in the Spherical Harmonic Domain Using Subspace Pseudointensity Vectors," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 1, pp. 178–192 (2017 Jan.). https://doi.org/10.1109/TASLP.2016.2613280.

[21] L. McCormack, S. Delikaris-Manias, A. Politis, et al., "Applications of Spatially Localized Active-Intensity Vectors for Sound-Field Visualization," *J. Audio Eng. Soc.*, vol. 67, no. 11, pp. 840–854 (2019 Nov.). https://doi.org/10.17743/jaes.2019.0041.

[22] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple Source Localisation in the Spherical Harmonic Domain," in *Proceedings of the 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 258–262 (Antibes, France) (2014 Sep.). https://doi.org/10.1109/IWAENC.2014.6954298.

[23] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Academic Press, London, UK, 1999). https://doi.org/10.1016/B978-0-12-753960-7.X5000-1.

[24] B. Rafaely, *Fundamentals of Spherical Array Processing*, Springer Topics in Signal Processing, vol. 8 (Springer, Berlin, Germany, 2015). https://doi.org/10.1007/978-3-662-45664-4.

[25] S. Rickard and O. Yilmaz, "On the Approximate W-Disjoint Orthogonality of Speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 529–532 (Orlando, FL) (2002 May). https://doi.org/10.1109/ICASSP.2002.5743771.

[26] J. Fliege and U. Maier, "The Distribution of Points on the Sphere and Corresponding Cubature Formulae," *IMA J. Numer. Anal.*, vol. 19, no. 2, pp. 317–334 (1999 Apr.). https://doi.org/10.1093/imanum/19.2.317.

[27] 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 (2012 Sep.). https://doi.org/10.1121/1.4740497.

[28] V. Hansen and G. Munch, "Making Recordings for Simulation Tests in the Archimedes Project," *J. Audio Eng. Soc.*, vol. 39, no. 10, pp. 768–774 (1991 Oct.).

[29] S. W. Clapp, *Design and Evaluation of a Higher-Order Spherical Microphone/Ambisonic Sound Reproduction System for the Acoustical Assessment of Concert Halls*, Ph.D. thesis, Rensselaer Polytech. Inst., Troy, NY (2014 Aug.).

[30] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, Springer Topics in Signal Processing, vol. 9 (Springer, Cham, Switzerland, 2017). https://doi.org/10.1007/978-3-319-42211-4.

[31] O. Nadiri and B. Rafaely, "Localization of Multiple Speakers Under High Reverberation Using a Spherical Microphone Array and the Direct-Path Dominance Test," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 10, pp. 1494–1505 (2014 Oct.). https://doi.org/10.1109/TASLP.2014.2337846.

[32] A. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of Arrival Estimation Using Pseudo-Intensity Vectors With Direct-Path Dominance Test," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, pp. 2296–2300 (Nice, France) (2015 Sep.). https://doi.org/10.1109/EUSIPCO.2015.7362794.

[33] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466 (1997 Jun.).

THE AUTHORS

Jonathan Mathews

Jonas Braasch

Jonathan Mathews is a Ph.D. Candidate in the Architectural Acoustics Program at Rensselaer Polytechnic Institute. He obtained both his B.S. in Electrical Engineering (2013) and Master's degree in Architectural Sciences (2016) from Rensselaer Polytechnic Institute (RPI). He has worked as an Electrical Design Engineer for Boeing, modeling radar system processes and designing object classification methods. His contributions at RPI include research in acoustic localization and sensor platforms for collaborative systems at the Cognitive and Immersive Systems Laboratory and Lighting Enabled Systems and Applications Center. His research interests include multi-user virtual reality systems and applications, spatial audio paradigms, audio digital signal processing, aural architecture, and music history. Jonathan is a student member of the Audio Engineering Society and Acoustical Society of America.

•

Jonas Braasch is a Professor at the School of Architecture and Director of Operations of the Cognitive and Immersive Systems Laboratory at Rensselaer Polytechnic Institute. He teaches in the Graduate Program in Architectural Acoustics. His research interests span collaborative virtual reality systems, binaural hearing, auditory modeling, multimodal integration, sensory substitution devices, aural architecture, and creative processes in music improvisation. For his work, he has received funding from the National Science Foundation, Natural Sciences and Engineering Research Council of Canada, DFG (German Science Foundation), European Research Council, New York State Council on the Arts, Christopher and Dana Reeve Foundation, and Craig H. Neilsen Foundation. He obtained a Master's degree from Dortmund University (Germany, 1998) in Physics and two Ph.D. degrees from Ruhr-University Bochum, Germany (2001, 2004) in Electrical Engineering/Information Science and Musicology. He is a fellow of the Acoustical Society of America and Associate Editor for its journal.