A REAL-TIME SRP-PHAT SOURCE LOCATION IMPLEMENTATION USING STOCHASTIC REGION CONTRACTION(SRC) ON A LARGE-APERTURE MICROPHONE ARRAY

Hoang Do, Harvey F. Silverman, Ying Yu.

LEMS
Division of Engineering
Box D, Brown University, Providence, RI 02912

ABSTRACT

In most microphone array applications, it is essential to localize sources in a noisy, reverberant environment. It has been shown that computing the steered response power(SRP) is more robust than faster, two-stage, direct time-difference of arrival methods. The problem with computing SRP is that the SRP space has many local maxima and thus computationally-intensive grid-search methods are used to find a global maximum. Grid search is too expensive for a real-time system. Several papers have addressed this issue. In this paper we propose using stochastic region contraction(SRC) to make computing the SRP practical. We discuss one important SRP method, computing it from the phase transform (SRP-PHAT), review SRC, and show the computational saving. Using real data from human talkers, we show that SRC saves computation by more than two orders of magnitude with almost no loss in accuracy.

Index Terms – Optimization methods, microphones, arrays, acoustic position measurement

1. INTRODUCTION

An open problem for large-aperture microphone arrays is to obtain the location of sound sources in the focal area using the data acquired from selected microphones. In this paper, we are concerned with algorithms in which all sources are assumed to be effectively modeled as point sources and that nothing is known *a priori* from previous location estimations; full search is required, rather than updating a tracker.

The pros and cons of many proposed location algorithms have been discussed in earlier work [1, 2, 3, 4, 5]. For real-time systems in realistic environments, only a few locators, all of them depending on time-difference-of-arrival(TDOA) estimation, have stood the test of practical implementation and satisfactory performance. Currently, a two-stage (microphone-signal-to-TDOA-vector, spatial search of the TDOA vector) generalized cross correlation (GCC)-based technique called ”LEMSalg” is being used in our real-time system [6]. This and similar two-stage algorithms are fast but rely on making first-stage TDOA decisions, which are often quite poor estimates when SNRs decrease, causing the second-stage search to fail. Our research [6, 2] and that of others [7] has shown that a one-stage method whose functional is the steered-response power(SRP) is more robust than a two-stage algorithm. However, computational cost is a real issue because the SRP space to be searched has many local maxima. Thus, a simplex or gradient technique, such as used in the spatial search with TDOA's, is not feasible.

In this paper, we present a method that makes the computation of a typical and proven-robust one-stage algorithm, SRP-PHAT, [2], practical. A global-maximum-finding algorithm, called stochastic region contraction(SRC), is applied and tested experimentally using our real system. We have seen no loss in accuracy, yet the computation is usually reduced by 2-3 orders of magnitude.

2. STEERED RESPONSE POWER(SRP) USING THE PHASE TRANSFORM (SRP-PHAT)

For time frame n, The SRP, $P_n(\vec{x})$, is the real-valued functional for the 3-D spatial vector $\vec{x}$ obtained by *steering* a delay-and-sum beamformer. The hypothesis is that high maxima in $P_n(\vec{x})$ will occur at the actual a set of k point sources at locations $x_s^{(n)}(k)$ even under very noisy and highly reverberant conditions [6]. The high maxima form the set $\hat{x}_s^{(n)}(k)$. For example for a single source, the location estimate, $\hat{x}_s^n(1)$, is

$$\hat{x}_s^n(1) = \underset{\vec{x}}{\operatorname{argmax}}\, P_n(\vec{x}). \qquad (1)$$

Given, $m_i(t)$ is the signal from microphone i in an M microphone system, then the SRP for some finite-length frame of length T is defined as

$$P_n(\vec{x}) \equiv \int_{nT}^{(n+1)T} |\sum_{i=1}^{M} w_i m_i(t - \tau(\vec{x}, i))|^2 \, dt \qquad (2)$$

where w_i is a weight and $\tau(\vec{x}, i)$ is the direct time of travel from location $\vec{x}$ to microphone i. It has been shown[8] that an SRP may be exactly computed by summing the generalized cross-correlations for all possible pairs of the set of microphones. Expanding Equation 2, going to the frequency domain using more general, frequency-dependent weights $W_l^*(\omega)$ and using Parseval's theorem we obtain,

$$\begin{aligned} P_n(\vec{x}) \quad = \quad & \sum_{k=1}^{M}\sum_{l=1}^{M}\int_{-\infty}^{\infty} \qquad (3) \\ & W_k(\omega)W_l^*(\omega)M_k(\omega)M_l^*(\omega)e^{j\omega(\tau(\vec{x},l)-\tau(\vec{x},k))}d\omega. \end{aligned}$$

A combined weighting function is defined,

$$\Psi_{kl}(\omega) \equiv W_k(\omega)W_l^*(\omega). \qquad (4)$$

The integral is seen to be the cross power spectrum for microphones k and l with the direct waves in alignment. Noting the elements summing to $P_n(\vec{x})$ form a symmetric matrix with fixed energy terms on the diagonal, the part of $P_n(\vec{x})$ that changes with $\vec{x}$ is defined as $P_n'(\vec{x})$, i.e.,

$$\begin{aligned} P_n'(\vec{x}) \quad \equiv \quad & \sum_{k=1}^{M}\sum_{l=k+1}^{M}\int_{-\infty}^{\infty} \qquad (5) \\ & \Psi_{kl}(\omega)M_k(\omega)M_l^*(\omega)e^{j\omega(\tau(\vec{x},l)-\tau(\vec{x},k))}d\omega. \end{aligned}$$

The phase transform (PHAT) [1] is an especially effective weighting of a GCC [9] for finding a TDOA from speech signals in highly-reverberant environment. Weights are the inverse of the magnitudes of the frequency components, i.e.,

$$\Psi_{kl}(\omega) \equiv \frac{1}{|M_k(\omega)M_l^*(\omega)|}. \qquad (6)$$

The process is thus to explore $P'(\vec{x})$ over the whole focal volume and ultimately find the set of one or more distinct maxima $\hat{x}_s^n(k)$. The calculation of any *particular* point of $P'(\vec{x})$ will be called a functional evaluation(fe). For the SRP-PHAT functional, we want to determine a point-source location in the room that gives the maximum value of $P_n'(\vec{x})$. Instead of a grid-search, which requires fe's on a fine grid throughout the room, we advocate using stochastic region contraction(SRC) to find the global maximum.

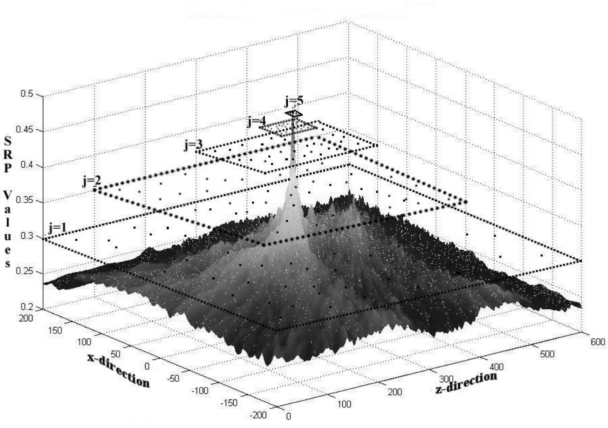

Fig. 1. 2D example of SRC: The surface is $P'(\vec{x})$. **j** is the iteration index. The rectangular regions show the contracting search regions

3. STOCHASTIC REGION CONTRACTION (SRC)

First presented in [10], the basic idea of the SRC algorithm is, given an initial rectangular search volume containing the desired global optimum and perhaps many local maxima or minima, gradually, in an iterative process, contract the original volume until a sufficiently small subvolume is reached in which the global optimum is trapped (the uncertainty voxel (volume V_u). The contraction operation on iteration i is based on a stochastic exploration of the $P_n'(\vec{x})$ functional in the current subvolume. While this can also be done using a refining grid-search, which ideally can have a computational advantage of up to 4, SRC features 1) a simple way to program and parameterize the optimization procedure, 2) a more robust procedure against an early wrong decision, and 3) an allowance of the optimum being on the continuum. The first step is to determine the number of random points, J_0, that need to be evaluated to ensure that one or more is likely to be in the volume, V_{peak}, of higher values (than the rest of the focal volume) surrounding the global maximum of $P_n'(\vec{x})$. see, e.g. Figure 1. Unfortunately, V_{peak} is not easy to determine and in our data changes substantially as the source is farther from the microphones. However, if V_{room} is the original search volume, we can estimate the number of fe's needed to ensure that the probability of missing V_{peak} altogether is less than a given percent (Table 1).

$\frac{\mathbf{V_{peak}}}{\mathbf{V_{room}}}$ **P(miss V_{peak})**	0.1	0.01	0.001	0.0001
1%	44	459	4603	46050
0.1%	66	688	6905	69,075
0.01%	88	917	9206	92,099

Table 1. Number of fe's required for three probabilities of missing V_{peak} and four values of the ratio $\frac{V_{peak}}{V_{room}}$.

Defining J_i as the number of random points evaluated for iteration i, N_i, the number of points used to define the new source volume, V_{i+1}, having a rectangular boundary vector $\vec{B}_{i+1} \equiv [x_{max}(i+1)\ x_{min}(i+1)\ y_{max}(i+1)\ y_{min}(i+1) z_{max}(i+1) z_{min}(i+1)]$, and I the number of iterations, and FE_i the total number of fe's evaluated as of iteration i,with Φ the maximum number of fe's allowed to be computed, the SRC algorithm for finding the global maximum is,

1. Initialize iteration: $i = 0$
2. Set initial parameters: J_0, N_0 and $V_0 = V_{room}$.
3. Calculate $P_n'(\vec{x})$ for J_i points.
4. Sort out the best $N_i \ll J_i$ points.
5. Contract the search region to the smaller region $V_{i+1}, \vec{B}_{i+1}$ that contains these N_i points.
6. IF: $V_{i+1} < V_u$, or $FE_i > \Phi$ and $V_{i+1} < T_1 V_u$, where T_1 is a parameter (about 10); determine $\hat{x}_s^n(i^*)$, $I = i$, STOP, KEEP RESULT.

7. ELSE IF $FE_i > \Phi$, STOP, DISCARD RESULT.
8. ELSE: Among the N_i points, keep a subset G_i points that have values greater than the *mean,* μ_i of the N_i points.
9. Evaluate J_{i+1} new random points in V_{i+1}.
10. Form the set of the N_{i+1} as the union of G_i and the best $N_{i+1} - G_i$ points from the J_{i+1} just evaluated. This gives N_{i+1} high points for iteration $i + 1$.
11. $i = i + 1$. GO TO STEP 5.

There are several variants of SRC for selecting the parameters J_i and N_i. We have found it very effective for our problem to set a fixed value for N_i based on experimentation as shown for our problem in Figure 2. Here we see that a value of $N = 100$ gives the best results with the lowest cost for all cases. That is, let $N_i \equiv N$, we define SRC-I, II and III.

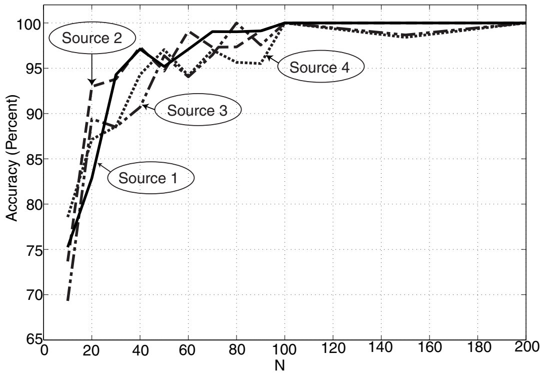

Fig. 2. Performance of SRC-I as a function of parameter N for four different source locations

- **SRC-I.** Let J_i be that number of random fe's needed find G_i points greater than μ_i. Guarantees monotone increasing μ_i. Use finite value of Φ.
- **SRC-II.** Let J_i be that number of random fe's needed to find G_i points higher than the minimum of the full set N_i. μ_i increases for **almost** all iterations. Use finite value of Φ.
- **SRC-III.** Fix $Ji = J$. Select the highest G_i points for each iteration. Does not guarantee monotone increasing μ_i. Set $\Phi \to \infty$.

4. COMPUTATIONAL COST

SRP-PHAT requires frequency-domain processing to do the phase transforms. For M microphones used in a locator, the computation of the $P'_n(\vec{x})$ requires $Q = M(M-1)/2$ phase transforms. For a DFT size of L, counting additions and multiplications as separate arithmetic operations,

1. **DFT:** A real FFT per microphone: $M \times \frac{5}{2} L \log_2 L$.
2. **Spectral Processing:** Phase transform, cross-power spectrum: $\approx 10QL$.
3. **IDFT:** Q real IFFT's or $\frac{5}{2} QL \log_2 L$.

In the current HMA system we use $M = 24$ microphones for a locator, implying $Q = 276$. A reasonable compromise among sufficient data, worst-case TDOA, and potential movement of the source is $L = 2048$. This totals $22.6x10^6$ ops/frame or 22.6mo/f (million operations per frame).

The focal area for the HMA is 4 m x 1 m x 6 m. To search the entire focal area at 1cm resolution, implying $V_u = 1cm$, requires 2.4×10^7fe's. The following steps are required to get a value for each of the grid points, $\vec{x}$:

1. Obtain the $M(= 24)$ Euclidean distances, $d_i^n(\vec{x})$, from $\vec{x}$ to each microphone. Cost: 3 mults, 5 adds, 1 square root($\approx 12ops$). Cost: 20ops/mic or 480ops/fe.
2. Determine $Q(= 276)$ TDOA's. $\tau_{ij}^n(\vec{x}) = (d_i^n(\vec{x}) - d_j^n(\vec{x}))C$ where C is the inverse off the speed of sound. Cost: 3ops/pair or 728ops/fe.
3. Sum up the PHAT values. Requires a multiplication, addition and truncation to discretize each TDOA value, a memory access and one more addition for the sum itself. Cost: 5ops/pair or 1380ops/fe.

Thus we need $480 + 728 + 1380 = 2588ops/fe$ which implies a cost for grid-search,

$$\begin{aligned} \text{Grid-Search Cost} &= 24x10^6(2588) \\ &= 63,113mo/frame \end{aligned} \qquad (7)$$

Note that the signal-processing cost is tiny in comparison to that of the grid search. In our real time system, we ideally would like to make a decision each 25.6ms, implying a *2.46TF* machine! Therefore the grid-search method is not a practical way of doing SRP-PHAT in real-time.

Using SRC requires the same signal processing, but significantly reduces the number of fe's needed to find the global optimum. As will be shown experimentally, the number of fe's varies with source position in a typical room. For SRC-I and SRC-II, it also depends on the number of frames discarded by the limit on the number of fe's. The number of frames discarded for our data is small; 1 of 105, 2 of 114, 4 of 87 and 4 of 69 for sources 1-4 respectively.

SRC does require a few small additional computations. These are 1) determine each random point, about 21ops/fe and 2) sort to get the best $N_{i+1} - Gi$ points, which has negligible cost. Neither of these additions really affects the computational load appreciably.

5. EXPERIMENTAL RESULTS

The system that we used in our experiments has been described in [6]. A human talker, approximately facing the locator microphones, repeated the first four seconds of the "rainbow passage" from four locations as shown in Figure 3 with the distances and SNR's indicated. It was shown in [6] that SRP-PHAT with grid search significantly outperformed a current real-time(less expensive) algorithm, (LEMSalg) especially under low SNR conditions. However, the SRP-PHAT functional does have some erroneous global maxima for some frames when the noise is very high.

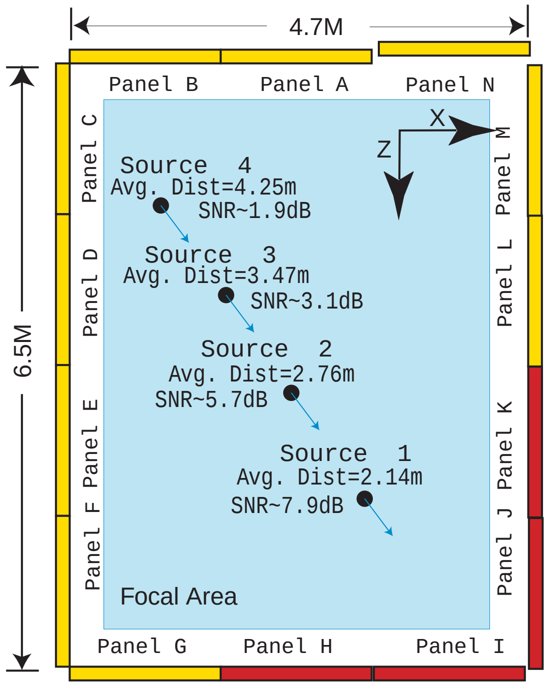

Fig. 3. Top View of the Array, Showing Source Locations and Panels (Locator uses Microphones on Panels H, I, J, K)

Results are given in Table 2 for accuracy and the average number of fe's used for LEMSalg and grid search as well as for SRC-I, SRC-II and SRC-III. Performances for LEMSalg and grid search are absolute and show overall correctness. For the grid search, as SNR decreases there are more frames in which the global maximum is not correctly placed. Performances of the SRC algorithms are **relative to the performance of the grid search**, i.e., Performance will be listed as 100% if the SRC implementation achieved the global maxima everywhere grid-search did.

The parameters for SRC were determined experimentally using the focal volume of the room as 4mx1mx6m or $V_{room} = 24m^3$ and, while for each source location V_{peak} is different, the worst case makes the ratio $\frac{V_{peak}}{V_{room}} \approx 0.005$. This implies from Table 1 that a value of $J_0 = 3000$ will err by missing the peak volume less than 0.1% of the time. Frames of 102.4ms, advancing each 25.6ms within the speech were used for testing, and an estimate was considered an error if it were either off by more than 5cm in x or z or 10cm in y, the vertical dimension.

6. CONCLUSION

We have verified here that SRP-PHAT is superior, especially under higher noise conditions, to a less costly, real-time, two-stage location-estimation algorithm. We have also shown that we can reduce its large computational cost by more than two orders of magnitude by using Stochastic Region Contraction (SRC), thereby making the use of SRP-PHAT practical for a real-time use – about 40 estimates per second using a 15GF machine.

Algorithm **SNR**	**Source 1** **7.9dB**		**Source 2** **5.7dB**		**Source 3** **3.47dB**		**Source 4** **1.9dB**	
	% Corr.	**# fe's**	**% Corr.**	**# fe's**	**%Corr.**	**# fe's**	**% Corr**	**# fe's**
LEMSalg	99.1	–	96.1	–	35.9	–	43.1	–
Grid Search	100	2.4×10^7	96.6	2.4×10^7	87.8	2.4×10^7	67.3	2.4×10^7
SRC - I	100	46,646	100	49,536	100	63,262	100	72,531
SRC - II	100	50,580	99.1	50,597	97.5	74,553	98.4	61,489
SRC - III	100	14,370	99.2	144,000	100	573,000	92.2	585,000

Table 2. Performance of SRP-PHAT estimates for grid search and three SRC parameterizations for four different locations

It was clear that performance varied significantly with the SNR of the direct-wave signal at the set of microphones. Under our worst-case conditions, $SNR \approx 1.9dB$, it was best to use the SRC-I algorithm to get the full accuracy of SRP-PHAT with an computational advantage of 333:1. If the conditions are less noisy such as our best case of $SNR \approx 7.9dB$, then using SRC-III gives full accuracy with a computational advantage of 1670:1, or only needing about a 1.5GF machine for 40 frames per second in real-time.

7. REFERENCES

[1] P. Svaizer, M. Omologo, and M. Matassoni. Acoustic source location in a three-dimensional space using cross power spectrum phase. In *Proceedings of ICASSP-97*, pages I–231, I–234, Munich Germany, April 22-25 1997.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. Robust localization in reverberent rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Techniques and Applications*, pages 157–180. Springer-Verlag, 2001.

[3] J. Benesty. Adaptive eigenvalue decomposition for passive acoustic source localization. *J.Acoust.Soc.Amer.*, 107(1):384–391, 2000.

[4] P. Aarabi. The fusion of distributed microphone arrays for sound localization. *EURASIP Journal of Applied Signal Processing(Special Issue on Sensor Networks)*, 2003(4):338–347, 2003.

[5] J. Chen, J. Benesty, and Y. Huang. Time delay estimation in room acoustic environments: an overview. *EURASIP Journal on Applied Signal Processing*, 2006(26503):1–19, 2006.

[6] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson III. Performance of real-time source-location estimators for a large-aperture microphone array. *IEEE Transactions of Speech and Audio Processing*, 13(4):593–606, July 2005.

[7] S. T. Birchfield. A unifying framework for acoustic localization. In *Proceedings of European Signal Processing Conference(EUSIPCO)*, Vienna, Austria, September 2004.

[8] J. Dibiase. *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberent Environments using Microphone Arrays*. PhD thesis, Brown University, Providence, RI, May 2000.

[9] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-24(4):320–327, August 1976.

[10] M. Berger and H. F. Silverman. Microphone array optimization by stochastic region contraction (SRC). *IEEE Transactions on Signal Processing*, 39(11):2377–2386, November 1991.