Computation Saving in a SRP-PHAT Sound Source Locator Variant

M. Seifipour *, S. Seyedtabaii **
* Elec. Eng. Department, Shahed University , mohammadseyfipor@gmail.com
** Elec. Eng. Department, Shahed University, **stabaii@shahed.ac.ir**

Abstract: *The well-known Steered Response Power–Phase Transform (SRP-PHAT) speech source locator enjoys robustness and renders accurate results in high reverberation and medium noise situations. However, it suffers from high computation burden, which debars it from many applications. Various arrangements have been developed during the recent years to lessen the calculations and facilitate its implementation. In this paper, a new hybrid scheme is suggested and compared its performance and costs with several well-known contestant algorithms. The simulation results confirm that the algorithm performance is highly promising and noticeable.*

Keywords: Sound localization, Microphone array, SRP-PHAT.

1. Introduction

One of the main complications in video conferencing, remote security surveillance, advanced human–computer interaction, gunshot positioning, video-gaming and robotic applications is the accurate determination of sound source location. In video conferencing, by localizing the sound source, it is possible to aim a pan-tilt-zoom camera to the estimated position for capturing a high-resolution image or face to be monitored and or picking high quality speech by microphone array beam forming.

Sound source locators are loosely divided into three categories. 1) Time difference of arrival (TDOA): an indirect two-step approach that first, the TDOA between microphone pairs are estimated and then, the estimate of the source position based on the geometry of the microphone array and the estimated delays is optimally computed. The two-stage algorithms are fast, however, in noisy environments as the estimation of TDOA deteriorates, the second-stage search fails in rendering accurate results [1, 2]. TDOA based methods are the most commonly used ones in practice 2) Spectral estimation: It is a one stage high-resolution spectral estimation method [3, 4], which also handles multiple-source cases [3]. 3) SRP family: It is a one stage steered beamformer, which its idea is rooted in antenna array design & processing for RADAR. In the Steered Response Power–Phase Transform (SRP-PHAT) algorithm, a candidate source maximizing the output of a steered delay-and-sum beamformer is sought.

Generally, the steered-beamformer strategies are recommended for critical applications where robustness is important [3]. The method can accurately find the sound source location in low noise environments, under relatively heavy reverberations [5]. The algorithm drawback, however, is high computation costs [5, 2]. Taking into account the main requirements of acoustic localization and tracking algorithms where robustness to acoustic disturbances and low computation are immediate requirements, the algorithm fails in supporting the latter.

In this respect, an improved SRP-PHAT algorithm based on principal eigenvector has been suggested in [6] where sound source location is estimated from the principal eigenvector computed from the frequency-domain correlation matrix. In another attempt, in [7], stochastic region contraction (SRC) is proposed to combat the computation cost problem. A saving of nearly three orders of magnitude was achieved. No degradation in location-estimation performance was observed for "four" different locations and different SNR's versus original SRP-PHAT. Coarse-to-fine region contraction (CFRC), also based on the concept of region contraction, is introduced in [8]. It is shown that CFRC has some computational advantages while the performance and the average computational saving achieved by SRC are still maintained. An improved SRP–PHAT is presented in [9] that it reduces a two-dimension searching space into a couple of one-dimension spaces by using an orthogonal linear array.

A different strategy has been put forward in [10] where, instead of evaluating the SRP function at fine spatial grid nodes, the surrounding volume around nodes of a coarse spatial grid is searched by the Generalized Cross Correlation (GCC) lag space corresponding to the volume surrounding each point of a coarse grid. By this technique, it is claimed that while the computation is highly lowered, the accuracy remains as if a relatively fine grid has been adopted.

In this paper, a new hybrid algorithm is presented. The

978-1-4673-5634-3/13/$31.00 ©2013 IEEE

method initially estimates the source location region by the spherical intersection TDOA that is then followed by the modified SRP-PHAT over a coarse grid for a fine source locating. Not searching the entire space and using a coarse grid means reduction in the cost. What remains is that how the quality is affected by the computation reduction. The method is tried for locating a single speech source and its performance compared with the results of some the well-known methods such as stochastic region contraction (SRC-I), the original SRP-PHAT and a version of Hybrid SRP-PHAT. The outcomes indicate that the new set up outperforms the contestants in computation burden while maintains desirably the accuracy.

In section 2, some sound source location algorithms are briefly discussed. In section 3 the new arrangement is elaborated. Section 4 presents the simulation results and lastly conclusion comes in section 5.

2. Sound Source Location Algorithms

2.1 The SRP-PHAT (SRP-P)

Consider an array of M microphones each receiving signal x_i from a single source located at $q_s=[x,y,z]$. The estimation of delay by cross-correlation is a well-known practice

$$R_{ik}(\tau)=\int X_i(\omega).X_k^*(\omega).e^{j\omega\tau}d\omega$$

where $X_i(\omega)$ is the Fourier transform of $x_i(t)$. Based on the SRP-PHAT algorithm, the following equation

$$P_n(q)=\sum_{i=1}^{M}\sum_{k=1}^{M}\int W_{ik}(\omega).X_i(\omega).X_k^*(\omega).e^{j\omega\tau_{ik}(q)}$$

$$W_{ik}=\frac{1}{\left|X_i(\omega).X_k^*(\omega)\right|}$$

expressing the normalized power of the received signals by M microphones is maximized [11] at $\tau_{ik}(q)=\tau_{ik}(q_s)$ where q_s is the source point. W_{ik} is a special weighting function named "phase transform" and $\tau_{ik}(q)$ is the direct path time difference of arrival of the source signal at q to the microphone pair located at q_i and q_k.

$$\tau_{ik}(q)=\frac{\|q_s-q_i\|-\|q_s-q_k\|}{v}$$

where v is the sound speed.

The equation is indeed can be viewed as the summation of cross-correlation of all pairs of the microphones, which resembles the delay-and-sum beamformer.

$$P=\sum_{i=1}^{M}\sum_{k=1}^{M}R_{ik}(\tau_{ik})$$

In other word, the SRP-PHAT algorithm calculates the power of all grid points in the search space, with the aim of finding the source location q_s that provides the maximum power value,

$$q_s=\arg\max_{q}P_n(q)$$

The SRP function for a frame of length T of signals is expressed as follows:

$$P_n(q)=\int_{nT}^{(n+1)T}\left|\sum_{i=1}^{M}w_i x_i(t-\tau_i(q))\right|^2 dt$$

It has been realized that the method spots accurately the sound source location in low noise environments, even under relatively heavy reverberations [5].

2.2 The modified SRP-PHAT (MSRP-P) [10]

In this method, instead of evaluating the SRP functional at discrete fine positions of a spatial grid, it is accumulated over the Generalized Cross Correlation (GCC) lag space corresponding to the volume surrounding each point of a coarse grid as it is visualized by the following equation

$$P_n(q)=\sum_{i=1}^{M}\sum_{k=i+1}^{M}\sum_{\tau=L_{ik1}(q)}^{L_{ik}(q)}R_{ik}(\tau) \quad (1)$$

Due to the selection of a coarse grid, the computation cost is low. However, for maintaining accuracy, a special provision has to be incorporated. Equation (1) shows that for each node (q) a summation of cross-correlation from neighboring points is also observed. This is to compensate for not having a fine grid. The "q and ik" dependent summation boarder parameter L is calculated once for each and every grid point based on $\tau_{ik}(q)$ gradient:

$$L_{ik1}(q)=\tau_{ik}(q)-\|\nabla\tau_{ik}(q)\|*d$$
$$L_{ik2}(q)=\tau_{ik}(q)+\|\nabla\tau_{ik}(q)\|*d$$

where $\nabla\tau_{ik}(q)$ is the gradient of delay function $\tau_{ik}(q)$ in a cubic volume around the point q and d is the distance from q to the cubic boarder along the gradient.

2.3 Spherical Intersection TDOA (SX)

It is well known that by using three pairs of microphones the source location can be determined. By generalized cross correlation algorithm, TDOA of each pair of microphones is calculated and a hyperboloid is formed. The intersection of three hyperboloids, then, exposes the source. Intersecting the hyperboloids is done numerically since its closed form solution is difficult to find [12]. Moreover, the intersection of hyperboloids is very sensitive to the hyperboloid parameters, TDOA estimation. Another solution to the problem is intersecting three spheres which are less sensitive to the sphere parameters and there is a closed form solution for it as has been suggested in [12] which is expressed as follows,

$$s=0.5M^{-1}(\Delta-2R_s d)$$

where s is the source position, M is the matrix of microphones coordinates, R_s is the source radios and d is the delay vector with respect to the origin microphone (the forth microphone). The equation can be solved in a two-step calculation, which has been detailed in [12]. The method requires four microphones where one of them is located at origin (0, 0, 0). SX is not considered to be robust; however, it is very fast and can generate a set of candidate locations for later fine search [12].

2.4 Hybrid Algorithm (HA) [13]

Hybrid Localization algorithm does not indifferently examine all nodes of a fine grid as the conventional SRP–PHAT does. Instead, the algorithm evaluates the time delays using generalized cross correlation algorithm to collect much smaller candidate points based on SX-TDOA. Then, SRP-PHAT examines the points for picking the most suitable one as the speech source. By this provision, the overall computation drops substantially. It is reported that the algorithm accomplishes the task with little degradation of accuracy in a low noise and reverberation area.

3 New Hybrid Source Localization

The hybrid method that is suggested here, first marks the most probable region of source, instead of points, and then exploits the area for accurate source location. For each of the two steps some schemes may be exploited. Those that are used here are SX-TDOA [13] and the modified SRP-PHAT [10].

3.1 Best Time Delay Selection

Highlighting the most probable region of source location using TDOA in a noisy and reverberant environment involves a major difficulty. Cross correlation between the received signals of each pair of microphones does not necessarily yield a single peak, obtaining multiple peaks is common as shown in Fig. 1. Meaning that, there exist huge ambiguities in estimating the difference delay and subsequently the locus of points with equal time delay differences.

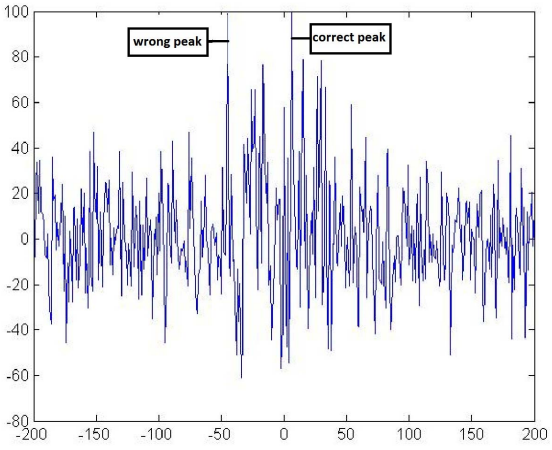

Fig. 1:Cross correlation of a pair of microphone signals

On the other hand, it is noticed that not all of the peaks point to a location within the room space. To exclude those points while maintaining diversity for later fine source locating search, a strategy is derived here. Based on that, each microphone pair introduces a set of delays, which their corresponding cross correlation amplitude exceeds around 70% of the peak. The level is sufficient in low noise cases even under high reverberations. The number of chosen peaks is limited to three (four). As a result, each microphone triplet presents 27 (64) candidate points in the search space. For a seven microphone sound source location set-up, there are 20 microphone triplets that all are examined the same way to form a set of at least 540 (1280) probable points.

3.2 Fine location spotting

Afterwards, the points located outside of the room are excluded. Then, the most populated area is searched. Different from HA, Not only the exact points are examined but their neighborhood is also evaluated by the modified SRP-PHAT algorithm in a search for pinpointing the most desired one as the speech source location.

4. Simulations

For the test purposes, a room of 8m by 6m by 3 m is considered. The rectangular room has uniform reflective boundaries. Seven omni-directional microphones are placed on seven nodes of its coarse grid (1m distance) as shown in Fig. (2).

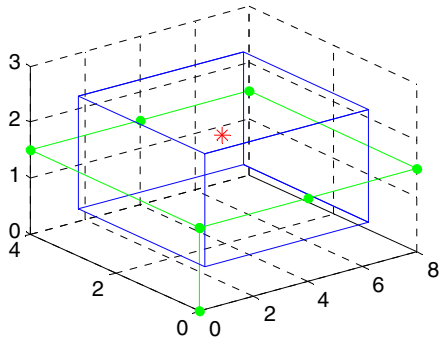

Fig. 2: Location of microphones and the single speech source

The received signals are assumed to be contaminated by white noise and marred by reverberations. The 20db and 5db SNR's cases are examined.

The reverberation is generated based on the image method [14]. The parameters of the reverberation model impulse response are the reverberation 60db decaying time T_{60} and the reflection coefficient, β. Two different values of 100 are 500ms for T_{60} are assigned. The reflection coefficient, β, is determined by Eyeing's formula [6]:

$$\beta = \exp\left(-13.82 \Big/ \left[v\left(L_1^{-1} + L_2^{-1} + L_3^{-1}\right) T_{60}\right]\right)$$

where L_i's are the room dimensions and v is 342m/s, the speed of sound.

The single source of sound is a piece of male English speech located in a point in the room (not necessarily on

the grid nodes), sampled at 8 kHz. The speech segment is framed to the packets of 2000 samples and windowed using Hamming. The room search box is 6m by 3m by 2m.

The performance index is the root mean squares of the location estimation error

$$E_{RMS} = \sqrt{\frac{1}{N_{iter}} \sum_{iter=1}^{N_{iter}} \sum_{j=1}^{3} \left(\hat{s}_j - s_j\right)^2}$$

where s_j is the exact and $s\square_I$ is the estimated location of the source. Ten random points, N_{iter}, are selected and for each point, four times the algorithm are executed.

Table 1 and Fig.3 show the performance of SRC-I [7], SRP-PHAT (0.1m space grid) [3], Hybrid [13] and the new hybrid.

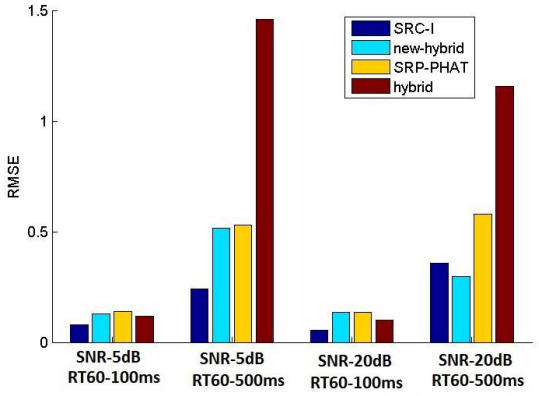

Fig. 3: Performance index of SRP-PHAT variants in locating a speech source.

Table 1: Comparison of the performance of SRP-PHAT variants in locating a speech source

RMSE	SNR=5dB RT60=0.1s	SNR=5dB RT60=0.5s	SNR=20dB RT60=0.1s	SNR=20dB RT60=0.5s
SRC	0.0783	0.3572	0.0535	0.2424
New Hybrid	0.1283	0.5170	0.1371	0.2976
SRP-PHAT	0.1403	0.5304	0.1339	0.5792
Hybrid	0.1188	1.4613	0.0994	1.1553

There are fluctuations in the performance merit based on the parameters of reverberation and noise. To make a judgment, the average performance depicted in Table 3 may be considered. The table figures indicate that the new hybrid is in the second place after SRC, while its computation cost is very much lower. The new hybrid consumes just 1/10 of what SRC needs.

On the other hand, Table 2 and Fig. 4 exhibit the computation cost of the implementation of the algorithms. Except in SRP-PHAT, the cost of operation of the algorithms is not fixed. Hence, for a matter of comparison, the average computation cost depicted in Table 3 is surveyed which reveals that the new hybrid is in the second place again. However, this time, behind the HA that is overran by the new hybrid from viewpoint of performance.

Comparing the four methods with respect to both the computation cost and the performance index, places at top the new proposed hybrid method, especially for real time applications and implementation on a low cost hardware.

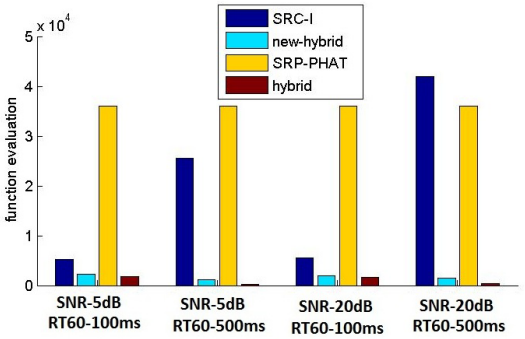

Fig. 4: Computation cost of the SRP-PHAT variant algorithms

Table 2: Computation cost of the SRP-PHAT variant algorithms

Function Evaluation	SNR=5dB RT60=0.1s	SNR=5dB RT60=0.5s	SNR=20dB RT60=0.1s	SNR=20dB RT60=0.5s
SRC	5241	25622	5568	41950
New Hybrid	2213	1142	1932	1430
SRP-PHAT	36000	36000	36000	36000
Hybrid	1725	250	1582	383

Table3: The average performance of the methods

Average Of Experiment	RMSE	FE
SRC	0.1828	19596
new Hybrid	0.2700	1680
SRP-PHAT	0.3459	36000
Hybrid	0.7087	985

5. Conclusion

In this paper the performance of a new set up of hybrid SRP-PHAT for accurate speech source location against SCR-I, SRP-PHAT and Hybrid (HA) are examined. The method enjoys spherical intersection strategy TDOA for region approximation and the modified SRP-PHAT for fine pinpointing of the source location. The computation cost of the method (in average) is 1/10 of SRC, 1/20 of SRP-PHAT (0.1m space grid) and 1.6 of Hybrid [12]. From viewpoint of performance, it outperforms the SRP-PHAT (0.1m space grid) and the hybrid [12]. Nevertheless, it is marginally behind SRC. Taking into account its lower computations, it is exposed as the winner of the competition especially for real time applications.

Acknowledgements

This work has been partially supported by the research department of Shahed University, Tehran, IRAN.

References

[1] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–19, 2006.

[2] J. S. Hu, C. M. Tsai, C. Y. Chan, and Y. J. Chang, "Geometrical Arrangement of Microphone Array for Accuracy Enhancement in Sound Source Localization," *in Proc. 2011 8th Asian Control Conference (ASCC) Kaohsiung, Taiwan, May 15-18*, pp 299-304

[3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in: Proc. 2001

Microphone Arrays: Signal Processing Techniques and Applications. Berlin, Germany: Springer-Verlag,, pp. 157–180.

[4] K. Nakamura, K. Nakadai, F. Asano, and G. Ince, "Intelligent Sound Source Localization and Its Application to Multimodal Human Tracking," *in Proc. 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems September 25-30, San Francisco, CA, USA*, pp. 143-148.

[5] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in low noise reverberative environments?," *in Proc. 2008 ICASSP,* pp.2565-2568.

[6] X. Wan, and Z. Wu, "Improved steered response power method for sound source localization based on principal eigenvector," Applied Acoustics, vol. 71, pp. 1126–1131, 2010.

[7] 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 Proc. 2007 *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, HI, pp. 121-124.

[8] H. Do, and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proc. 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, New Paltz, NY.

[9] W. Cai, S. Wang, and Z. Wu, "Accelerated steered response power method for sound source localization using orthogonal linear array," Applied Acoustics, vol. 71, pp. 134–139, 2010.

[10] M. Cobos, A. Marti, and J. J. Lopez, "A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling," *IEEE SIGNAL PROCESSING LETTERS,* VOL. 18, NO. 1, pp 71-74, JANUARY 2011.

[11] C. H. Knapp, and G. C. Carter, "The generalized correlation method for estimation of time delay," *Trans. Acoust. Speech, Signal Process*., vol. 24, pp. 320–327, 1976.

[12] H. C. Schau, and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of arrival differences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 8, pp. 1223-1225, August 1987.

[13] J Peterson, and C Kyriakakis, "Hybrid algorithm for robust, real-time source localization in reverberant environments," *In: Proceedings 2005 ICASSP, vol. 4, Philadelphia, PA, March 19–23*, p. 1053–1056.

[14] J. B. Allen, and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *J Acoust Soc Am, vol.* 65, pp.943–950, 1979.