A NEW REGION SEARCH METHOD BASED ON DOA ESTIMATION FOR SPEECH SOURCE LOCALIZATION BY SRP-PHAT METHOD

Ali Dehghan Firoozabadi, Hamid Reza Abutalebi

Speech Processing Research Laboratory (SPRL), Electrical and Computer Eng. Dept., Yazd University, Yazd, Iran
phone: + (98) 351-8122396, fax: + (98) 351-8200144, email: habutalebi@yazduni.ac.ir
web: pweb.yazduni.ac.ir/engineering/elec/sprl

ABSTRACT

Steered Response Power-PHAse Transform (SRP-PHAT) method has been already proposed and investigated for the sound source localization. Grid search methods can be used to find global maximum of SRP, but they are so computationally expensive that can not be used in real-time applications. In this paper, we have proposed a SRP-based localization method which works in cascade with a DOA estimation module; i.e. first the direction of speaker is recognized by one of the DOA estimation methods; after that, we bound the search region to a space fragment around estimated direction of speaker; then we use SRP-PHAT algorithm computations and volume contraction methods (such as SRC and CFRC) on this fragmentized regions and decrease computational costs to a large extent. By use of the data collected from different (speaker) scenarios, we demonstrate the accuracy and speed gained by proposed method.

1. INTRODUCTION

Speech source localization is one of the main topics in meeting room processes. The primary aim is to obtain three-dimensional location of the speech source. The methods of speech source localization are divided in two categories: 1) one-stage methods, and 2) two-stage ones. TDOA is a two-stage method to localize the source; this is one of the fast methods for speech source localizing. However, its performance drastically degrades in reverberant situations. A family of one-stage methods, called SRP, has been already proposed and investigated in the literature [1,2,3] that are more robust than two-stage algorithms. Nevertheless, the SRP method has to search among lots of local minimums and maximums which results in a large computational complexity. Different region contraction methods have been introduced for SRP-PHAT method, that we name SRC (Stochastic Region Contraction) and CFRC (Coarse-to-Fine Region Contraction) here as the most famous ones [4,5,6]. These methods try to decrease computational cost of SRP-PHAT method by use of an iterative process.

In this paper, we use DOA estimation methods to fragment the region and bound the search region for SRP-PHAT method. After that, by applying SRC and CFRC methods on these fragmentized regions, we try to decrease SRP-PHAT computational costs. In other words, we use a DOA estimation method with a few calculations to firstly determine speaker direction; then, considering this direction, we establish a proper portion of space around that direction and search on this portion.

We show that the proposed method not only decreases computational complexity but also increases the accuracy of SRP-PHAT localization. Our experiments in different positions of speech source show the high yield of proposed method in increasing the speed of computations and decrease of errors. In this paper we have done our simulations by supposing Far-field SRP-PHAT method [7] for DOA estimation.

2. APPLYING SRP-PHAT FOR SPEECH SOURCE LOCALIZATION

For the n-th time frame, SRP, $P_n(\vec{x})$, is a definite-value-function based on three-dimensional vector $\vec{x}$, that is obtained from the output of a directed Delay and Sum Beamformer. It is supposed that biggest maximum in $P_n(\vec{x})$, even in high noisy a reverberant condition, will happen in $x_S^{(n)}(k)$ which contains of K point sources. Biggest maximum will be as $\hat{x}_S^{(n)}(k)$, that for a one point source is $\hat{x}_S^{(n)}(1)$ that which is equal to:

$$\hat{x}_S^{(n)}(1) = \arg\max_{\vec{x}} P_n(\vec{x}) \qquad (1)$$

If $m_i(t)$ is the signal from the *i*-th microphone (in a system with M microphones), for a typical frame (with length of *T*), the SRP is defined as:

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

w_i is the weight and $\tau(\vec{x}, i)$ is the time in which signal travel from the source (x) to the *i*-th microphone. SRP can be calculated by adding up GCCs of all possible microphones. By expanding Equation (2), using frequency-dependent weights $W_l^*(\omega)$, and also considering Parseval theory, we have [4]:

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

By the combination of weight functions, we reach to the following effective weighting (filter) function for the SRP:

$$\Psi_{kl}(\omega) = W_k(\omega) W_l^*(\omega) \tag{4}$$

One of the most important weighting functions is the Phase Transform (PHAT) that is defined as:

$$\Psi_{kl}(\omega) \equiv \frac{1}{\left| M_k(\omega) M_l^*(\omega) \right|} \tag{5}$$

Pay attention that collective factors for calculating $P_n(\vec{x})$, will establish a symmetric matrix with constant energy on its diagonal. The part of this function which changes in relation with X will be defined in the form of $P_n'(\vec{x})$. It means that:

$$P_{\mathrm{n}}'(\vec{x}) = \sum_{\mathrm{k=1}}^{\mathrm{M}} \sum_{\mathrm{l=K+1}}^{\mathrm{M}} W_{\mathrm{k}}(\omega) W_{\mathrm{l}}^*(\omega) M_{\mathrm{k}}(\omega) M_{\mathrm{l}}^*(\omega) e^{\mathrm{j}\omega(\tau(\vec{x},\mathrm{l}) - \tau(\vec{x},\mathrm{k}))} d\omega \tag{6}$$

In reverberant conditions, PHAT has shown good performance in the case of both TDOA- and SRP-based methods. In fact in SRP-PHAT method we compute the amount of SRP-PHAT function for all of the space bins in the room. This is done by grid search methods. The location which makes this function (equation (3)) maximum, will be selected as the speech source location [4].

3. DOA ESTIMATION BY FAR-FIELD SRP-PHAT METHOD

Different methods have been introduced for specifying the direction of speaker, such as: 1) Far-field SRP-PHAT, 2) root-SRP-PHAT [8] and 3) DOA estimation based on TDOA vector [9]. Here, we suppose that the Far-field SRP-PHAT method has been used for DOA estimation. Actually, the simulation results are based on applying Far-field SRP-PHAT method in specifying direction of speaker.

In figure.1 you can see a profile from the room used for assessment of far-field SRP-PHAT method.

In Figure 2, the wave propagation and the microphone array scheme have been drawn.

The cross power density spectrum for two time signals, $x_l(t)$ and $x_k(t)$ is computed as:

$$\Gamma_{lk}(\omega) = F\left(E\left[x_l(t) x_k^*(t+\tau)\right]\right) \tag{7}$$

where F(.) and E[.] are the Fourier transform and expectation operator, respectively.

Under Far-field assumption, we can say that source signal enters the microphone array as a plane waves. Based on this assumption, TDOA between these two microphones will be:

$$\hat{\tau}_{lk} = (l - k)\hat{\tau} \tag{8}$$

where $\hat{\tau}$ is TDOA for two adjacent microphones. The optimum value for $\hat{\tau}$ is specified as follows:

$$\hat{\tau}_{opt} = \arg\max_{\hat{\tau}} C_{lk}(\hat{\tau}) = \arg\max_{\hat{\tau}} \left(\frac{1}{2\pi} \sum_{l=1}^{M} \sum_{k=1}^{M} \int_{-\infty}^{+\infty} \frac{\Gamma_{lk}(\omega)}{\left|\Gamma_{lk}(\omega)\right|} e^{j\omega\hat{\tau}(l-k)} d\omega \right) \tag{9}$$

The input angle of a linear array, θ, is computed as [7]:

$$\theta = \arccos\left(\frac{c}{dF_s} \hat{\tau}_{opt} \right) \tag{10}$$

where c, d, and F_s are sound propagation velocity, distance between microphones and sampling frequency, respectively.

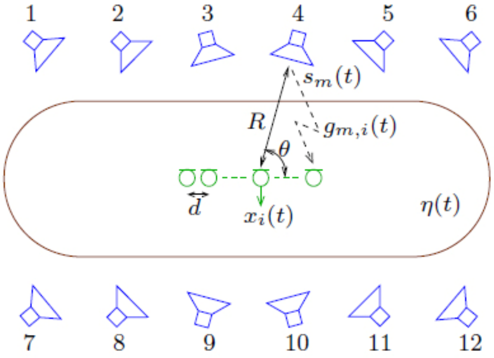

Figure 1 - Conference room with 12 speakers and a linear microphone array [7].

The above algorithm estimates the direction of the speaker (not its 3-D location). This algorithm can be implemented with a low computational cost.

4. PROPOSED METHOD FOR DECREASING THE COMPUTATIONAL COST OF SRP-PHAT

In this research, we have proposed a hybrid sound source localization method as follows: instead the search on the entire region by SRP-PHAT method, we firstly apply the DOA estimator (explained in section 3) to specify direction of sound wave. The estimated DOA estimator specifies the direction of speech source with some value of error. Actually, the Far-field SRP-PHAT method is used to determine the direction of sound arrival, with a low level of computational cost. To take the DOA errors into account, we consider the error bounds and limit the search region as a fragment of entire region.

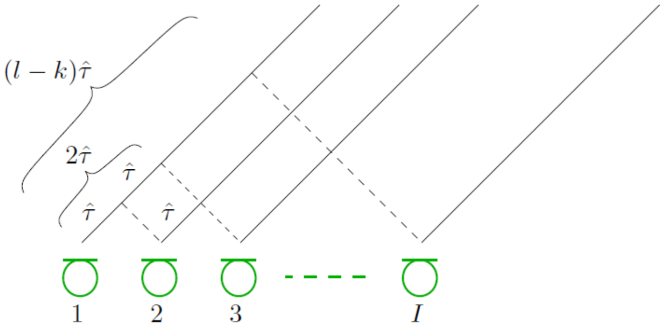

Figure 2 - A representation of plane waves entering a uniform microphone array [7].

Let θ to be the estimated DOA for the reference microphone and θ_e to be the bound on the estimation error. Thus, we fragmentize a region of space which is limited between

$\theta-\theta_e$ and $\theta+\theta_e$. Since the DOA estimation method has determined the direction of source with a definite error, by extending the search region to a section limited to the angles [$\theta-\theta_e$, $\theta+\theta_e$], we expect to cover the source location with a high probability.

Limiting the search region results in high reduction of computational costs. For example, suppose that algorithm reports the source is located in direction of 30^0 toward reference microphone array axis (with $\pm 10^0$ error bound). As a result, fragmentized region is limited to the angles of 20^0 and 40^0 toward array direction. So, it is actually $\left(\frac{40^0-20^0}{360^0}=\frac{1}{18}\right)$ times of the whole region that should be searched.

It means that this method is much faster than already-proposed contraction methods, including SRC and CFRC; furthermore, we can also apply these two methods on the extracted fragment and decrease the computational cost much more.

The high reduction in the computational complexity of region search has been attained at the cost of an extra DOA module. Considering that the DOA estimation methods, especially Far-field SRP-PHAT method, have a very low computational cost, we can neglect their complexity in comparison with 3-D SRP-PHAT localization method. Although the errors of DOA estimation can results in the missing of 3-D location of speech source, however, by considering the proper value for the error bound, the probability of this event will be minimized.

In general, we claim that the proposed method highly decreases the computational cost of 3-D speech source localization method, while preserving the system accuracy at the satisfactory level.

5. SIMULATION

In our simulations, we suppose that the speech source is located in a $4.5m \times 4.5m \times 3.5m$ room. It is assumed that the source signal is degraded by both Gaussian white noise and reverberation. We have used the IMAGE algorithm [10] to simulate environment reverberation. An L-shaped microphone array (see Figure 3) has been considered in the middle of the room. It consists of 10 microphones with 30cm distance between adjacent microphones.

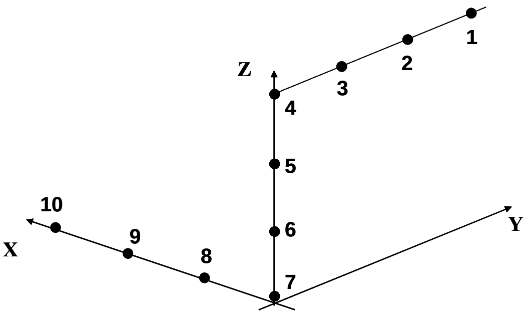

Figure 3 - L-shaped microphone array.

We do our experiments on data in 3 different scenarios. In the first scenario, we supposed that reverberation is the dominant degradation factor. In this scenario, reverberation time is supposed to be 500ms and SNR for the reference microphone is considered 20dB. This kind of environment is called reverberant environment. In the second scenario, the noise is considered as the main degradation factor. In this situation, reverberation time is supposed to be 200ms and reference microphone SNR is considered 5dB. This environment is called noisy one. In the third scenario, both noise and reverberation have considerable effect. In this condition, reverberation time is 500ms and we will have 5dB for reference microphone SNR. We call this environment a noisy-reverberant environment. Hamming window, with 60ms length and 50% overlap, was used in simulations.

Also, the experiments have been done in 3 different positions for the speaker. In the first situation, speaker is located in front of the array, with a 1.5m distance. In this situation, speaker position is (350,200,180)cm. In the second situation the location of speaker is (200,350,180)cm. In the last situation, the speaker's location is considered at the corner of the room with the position equal to (430, 20,180)cm. The positions of speaker (in these three situations) have been shown in figures (4) to (6), respectively.

Results obtained for specifying direction of speaker in 3 different speaker situations are reported as follows:

1- For the speaker in the 1st location, the true DOA (measured for the microphone#1) is 59^0 (referred to the upper branch of L-shaped array).

2- For the speaker in the 2nd location, the true DOA (measured for the microphone#1) is 0^0 (referred to the upper branch of L-shaped array).

3- For the source in the 3rd situation, the true DOA (measured for the microphone#1) is 111^0 (referred to the upper branch of L-shaped array).

We considered $\pm 15^0$ error bound in simulations. It is just for increasing the possibility of locating source in the fragmentized region. In this way we will have a specific region fragment for every 3 different positions of source. We will do the same computations with these fragments as what was done with entire region in SRP-PHAT method, i.e. we will apply grid search, SRC method and CFRC method on these fragments to find the maximum in each case.

To examine the algorithm complexity, we consider the number of functional evaluation (fe) for the algorithms. Actually, it has a direct relation to the complexity of the algorithm.

We used 10cm resolution in grid search method. Also number of primary search points for SRC-I, SRC-II and SRC-III methods are:

$$\left(\frac{30^0}{360^0}\right)\times\left(\frac{450}{10}\right)\times\left(\frac{450}{10}\right)\times\left(\frac{350}{10}\right)\approx 5906\,.$$

This number for the CFRC method is 5906 for the first stage and $5906\times\frac{750}{3000}=1746$ for next stages.

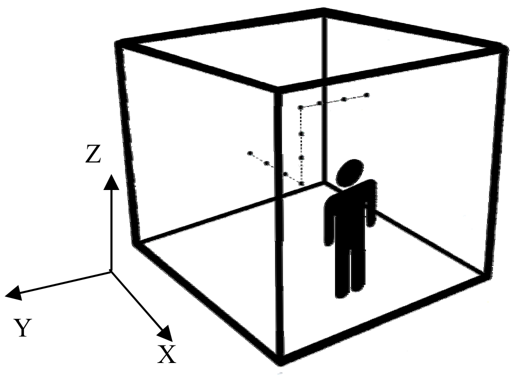

Figure 4 - The standing situation of speaker in front of array (first position)

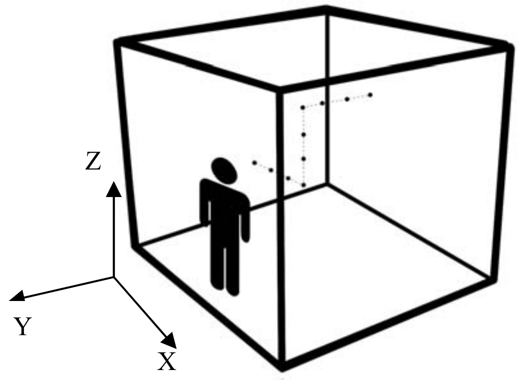

Figure 5 - The standing situation of speaker beside the array (second position)

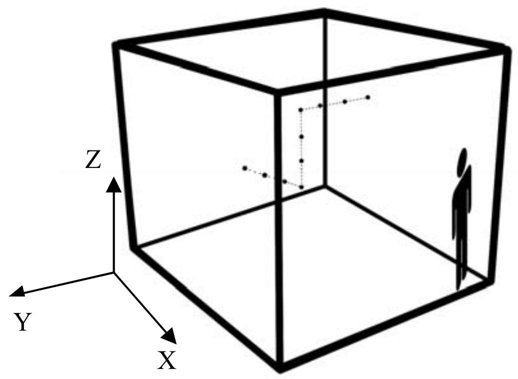

Figure 6 - The standing situation of speaker on the room corner (third position)

In the case of random-point-based contraction methods (i.e., SRC and CFRC), the program has been repeated 100 times and the average results have been reported. This is to acquire a normal distribution for the system error.

We have also evaluated the method accuracy based on localization Mean square Error (MSE). Tables 1, 2 and 3 present number of functional evaluations (fe) and MSE values obtained for different scenarios, applying grid search, SRC, and CFRC methods on fragmentized regions.

Table 1 shows the results for the 1st location of speaker. If the search is done on the entire region, the functional evaluation (fe) is about 70875 for grid search, 12000 for SRC-I, 14000 for SRC-II, 20000 for SRC-III, and about 17000 for CFRC method. These amounts are decreased considerably by use of proposed method; for example, our proposed method has decreased functional evaluation (fe) of algorithm to 5472 (from 70875) in the grid search, and to about 1000 (from 12000) in SRC-I. Also, it has decreased to about 1200 (from 14000) in SRC-II, to about 1400 (from 20000) in SRC-III, and to 1350 (from 17000) in CFRC method.

In addition to a considerable reduction in computational costs, our proposed method has decreased the localization error to some extent. In applying SRC and CFRC methods on the entire region, some of the evaluated points have a clear difference with the true locations; however, in the proposed method we do not see such a clear difference again, because points' amounts has been limited to the fragmentized parts of region. So, this method not only decreases computations of SRP-PHAT method to a large extent, but also increases the localization accuracy. Tables 2 and 3 show results obtained for 2nd and 3rd positions of source, respectively. These two tables clearly show the superiority of our proposed method, too. As it is shown, new search method (based on DOA estimation) significantly decreases computational cost and also increases SRP-PHAT localization accuracy.

6. CONCLUSION

SRP-PHAT method is a one-stage method in speech source localization which has a large amount of computation. This disadvantage prevents us to use this method in real-time systems. Methods such as SRC and CFRC decrease SRP-PHAT method computation cost, but they are still far from real-time applications.

In this research, we proposed, implemented and examined a new region search method which is based on space fragmentizing by DOA estimation. It decreases the computational complexity of the SRP-PHAT to a large extent. The combination of this method with SRC-I decreases functional evaluation (fe) from 70875 (for complete grid search) to about 1000. This improvement is obvious in different scenarios and different speaker positions. Moreover, new proposed method not only decreases computational costs considerably, but also increases localization accuracy in SRP-PHAT method. The increase of accuracy can be justified by the limiting of search field.

REFERENCES

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

[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] 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.

[4] 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. of ICASSP 2007*, vol. 1, Honolulu, Hawaii, Apr. 2007, pp. 121-124.

[5] H. Do, and H. F. Silverman, “A fast microphone array SRP-PHAT source location implementation using coarse to fine region contraction (CFRC),” In *Proc. WASPAA*, pp. 295–298, 2007.

[6] 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.

[7] A. Johansson, N. Grbic, and S. Nordholm. “Speaker localisation using the far-field SRP-PHAT in conference telephony,” In *Proc. ICICS*, Kaohsiung, Taiwan ROC, 2002.

[8] A. Johansson, and S. Nordholm, “Robust acoustic direction of arrival estimation using Root-SRP-PHAT, a real-time implementation,” In *Proc. ICASSP*, vol. 4, pp. 933-936, March 2005.

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

[10] J. Allen, D.Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

Table 1. Results (functional evaluation (fe) and localization MSE (in *cm*)) for different methods in the case of speaker located at position 1.

Position 1	Full Search Method										Proposed Search Method									
	Grid search		SRC-I		SRC-II		SRC-III		CFRC		Grid search		SRC-I		SRC-II		SRC-III		CFRC	
	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE
Scenario 1	70875	46	12155	46	14396	38	20154	49	17487	41	5472	43	1068	41	1202	38	1475	47	1407	36
Scenario 2	70875	55	12251	42	14410	54	20131	50	17414	39	5472	49	1041	37	1174	48	1440	41	1391	38
Scenario 3	70875	80	12202	82	14415	84	20206	69	17452	80	5472	76	1079	73	1196	82	1443	68	1399	76

Table 2. Results (functional evaluation (fe) and localization MSE (in *cm*)) for different methods in the case of speaker located at position 2.

Position 2	Full Search Method										Proposed Search Method									
	Grid search		SRC-I		SRC-II		SRC-III		CFRC		Grid search		SRC-I		SRC-II		SRC-III		CFRC	
	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE
Scenario 1	70875	41	12003	43	14775	47	20196	52	17044	31	5472	41	1101	29	1195	35	1621	41	1341	30
Scenario 2	70875	45	11955	51	14727	51	20145	47	17007	57	5472	54	1091	50	1209	55	1625	46	1365	47
Scenario 3	70875	73	11981	84	14699	94	20114	65	17049	70	5472	83	1095	79	1218	84	1674	62	1393	64

Table 3. Results (functional evaluation (fe) and localization MSE (in *cm*)) for different methods in the case of speaker located at position 3.

Position 3	Full Search Method										Proposed Search Method									
	Grid search		SRC-I		SRC-II		SRC-III		CFRC		Grid search		SRC-I		SRC-II		SRC-III		CFRC	
	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE	fe	MSE
Scenario 1	70875	76	13321	76	15295	71	20609	64	17220	79	5472	66	1032	68	1104	65	1571	62	1319	71
Scenario 2	70875	87	13302	91	15361	93	20625	80	17254	81	5472	89	1039	81	1096	88	1584	74	1332	73
Scenario 3	70875	96	13295	109	15364	105	20639	113	17259	93	5472	110	1030	98	1118	101	1596	102	1346	92