Fast Time Difference of Arrival Estimation for Sound Source Localization using Partial Cross Correlation

Mariam Yiwere* · Eun Joo Rhee**

Abstract

This paper presents a fast Time Difference of Arrival (TDOA) estimation for sound source localization. TDOA is the time difference between the arrival times of a signal at two sensors. We propose a partial cross correlation method to increase the speed of TDOA estimation for sound source localization. We do this by predicting which part of the cross correlation function contains the required TDOA value with the help of the signal energies, and then we compute the cross correlation function in that direction only. Experiments show approximately 50% reduction in the cross correlation computation time thereby increasing the speed of TDOA computation. This makes it very relevant for real world surveillance.

Keywords : Sound Source Localization, Time Difference of Arrival, Partial Cross Correlation

Received: 2015. 08. 31. Revised: 2015. 09. 16. Final Acceptance: 2015. 09. 17.

* Department of Computer Engineering, Hanbat National University, e-mail: myiwere@yahoo.co.uk
** Corresponding Author, Department of Computer Engineering, Hanbat National University, e-mail: ejrhee@hanbat.ac.kr

1. Introduction

Sound source localization is the process of estimating the direction from which a received sound originates. It is very useful in various fields including video conferencing, robot-human interaction, video surveillance, etc. Incorporating sound source localization into a video surveillance system [Jacek and Randy, 2013] will improve the system's usefulness and effectiveness by enabling it to capture events which occur outside the camera's field of view. The major techniques for implementing microphone array based sound source localization include the directional method based on high resolution spectral estimation [Lobos et al., 2006], the controllable beamforming method based on the biggest output power [Jean-Marc et al., 2004], and the Time Difference of Arrival (TDOA) method [Knapp et al., 1976].

The TDOA method is more applicable because it is not limited to only narrowband signals and it is relatively fast enough for real-time applications. The TDOA value is generally computed by using cross correlation, which can be implemented in either time or frequency domain. The speed of time domain cross correlation is slow as compared to that of frequency domain, but it is very straightforward to implement. Due to its low speed, it cannot be used to develop efficient sound source localization for video surveillance. On the other hand, the speed of frequency domain cross correlation [Knapp et al., 1976; Hong and Miao, 2010; Bo Qin et al., 2008; Zhou Lin et al., 2015] is faster but it is more complex to implement. It requires the computation of fast fourier transforms, complex conjugate multiplications and inverse fast fourier transform. To incorporate sound source localization in video surveillance, the localization method needs to have reasonable speed and accuracy. If the speed of time domain cross correlation can be increased, it can be used for efficient sound source localization to increase the effectiveness of video surveillance systems, therefore this study focuses on speeding up the time domain cross correlation for TDOA estimation.

For sound source localization, Halim et al. [Halim et al., 2011] proposed two methods in their research. Their first method is based on time domain cross correlation. After obtaining the correlation signal, they apply a band pass elliptic filter to it before searching for the maximum coefficient. Although this filter attenuates the noise samples, it does not eliminate them. This means that the filter adds extra computation time to the TDOA estimation process and reduces its speed. Also, Murray et al. [John et al., 2004] implemented the time domain cross correlation in their work. They computed all possible (2N-1) coefficients before searching for the maximum value. With this method, too many unnecessary computations are made and it can slow down the sound source localization, so it is not appropriate for a surveillance system. Instead, Bert et al. [2012] computed only a limited number of cross correlation coefficients in order to reduce the computation time. They determine the limited range based on the resolution for their microphone array. Even though the number of computations is reduced, there are still unnecessary correlation coefficients being computed because only one correct value is required. To speed up the TDOA estimation in

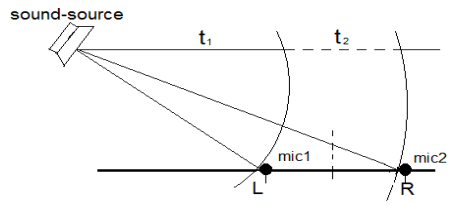
time domain by reducing the number of computations, we suggest a partial cross correlation method.

In our work, we achieve speed improvement by using the signal energy levels [Halim et al., 2011] to skew the time domain cross correlation computation into only one direction instead of computing it for both positive and negative lag values, as in the traditional method [Halim et al., 2011; John et al., 2004]. Firstly, we determine the relevant lag range [Bert et al., 2012] by computing the maximum and minimum possible delays. Secondly, we estimate the energies of the channel 1 and channel 2 components of the captured signal. The signal energies are used to predict the direction in which the cross correlation should be computed. After that, we compute only a limited number of cross correlation coefficients in the predicted direction, and then we compute the azimuth of the sound source based on the TDOA value obtained. The partial cross correlation achieves approximately 50% reduction in the computation time of the traditional time domain cross correlation, which makes the TDOA estimation 2 times faster.

The organization of this paper is as follows. The TDOA method is described in chapter 2. Chapter 3 describes the partial cross correlation, 4 describes the azimuth computation, 5 presents our experiments and discussion and the conclusion is presented in chapter 6.

2. Time Difference of Arrival

The Time Difference of Arrival (TDOA) value is the time it takes for sound to arrive at the second sensor once the first sensor has detected the sound. It is based on the principle of interaural time difference (ITD) used by mammals [David and Benedikt, 2003]. <Figure 1> illustrates the interaural time difference. The sound signal takes a time t_1 to reach the left mic, L, and it takes an extra time t_2 for the same signal to reach the right mic, R. This is because the left mic is closer to the sound source.



〈Figure 1〉 Illustration of Interaural Time Difference (ITD)

The time domain TDOA is computed by implementing the cross correlation function which is used to compare two signals for maximum similarity. The function takes as inputs two signals in the form of vectors or arrays, x and y, captured from the left and right microphones, X and Y. These signals are slid across each other, element by element, and a sum of the products of the coinciding elements is recorded for each slide or shift position (lag value). The cross correlation function [John et al., 2004], CrossCorr(x, y)(j), is shown in equation 1. The offset of the maximum correlation is taken as TDOA value, see equation 2.

$$\text{CrossCorr}(x,\ y)(j) = \sum_{k=0}^{n-1} x(k+j) \times y(k) \qquad (1)$$

$$-N < j < N$$

$$TDOA = \underset{j}{\arg max}\ Cross\,Corr(x, y)(j) \qquad (2)$$

3. Partial Cross Correlation

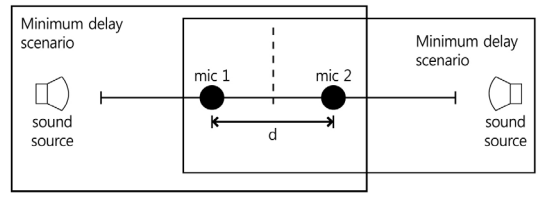
In order to speed up the TDOA estimation time, we propose this partial cross correlation method which eliminates most of the computations in the traditional time domain cross correlation. In this method, we consider only the relevant range of delay values by checking the expected resolution. We also use the energies of the signals to predict which part of the correlation function will contain the TDOA value. By doing so, we are able to compute just a few relevant correlation coefficients, which saves a lot of computation time.

3.1 Relevant Lag Range

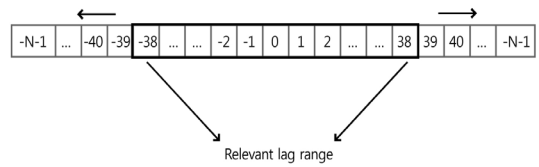
Considering our microphone array set-up and the expected resolution, only a limited number of cross correlation coefficients are required to be computed [Bert et al., 2012], from which only one offset will be taken as TDOA value. Assuming that the distance between two microphones is *d*, the velocity of sound is *v*, and the sampling rate is *f*, we can easily check that the maximum delay (τ) in samples that can be estimated is df/v, and the minimal delay is -df/v as shown in equation 3. <Figure 2(a)> shows the scenarios for maximum and minimum delay values, and <Figure 2(b)> shows the computed relevant lag/delay range.

$$\text{Range} = [min\tau,\ max\tau] = \left[\frac{-df}{v}, \frac{df}{v}\right] \qquad (3)$$

With our microphones set 30cm apart, a sampling rate of 44.1 kHz and the assumed velocity of sound being 343m/s, the maximum delay in samples that can be estimated is 38 and the minimal delay is -38.



(a) Scenarios for min and max delay



(b) Relevant lag range

〈Figure 2〉 Illustration of Relevant Lag Range

3.2 Signal Energy Estimation

After we determine the relevant lag range, we further reduce the amount of computations by computing only one half of the lag range. The reason is that, there are still too many computations since the number of multiplications required to compute each single correlation coefficient is directly proportional to the length of the signals. Moreover, the required TDOA value is either positive or negative in most cases, which means that it is not necessary to compute all the correlation points within the relevant lag range. A large amount of unnecessary multiplications can be ignored if we know which part of the correlation function contains the TDOA value.

In view of this, we estimate the signal energies [Halim et al., 2011] to predict which part of the function contains the TDOA, and then we compute the cross correlation into that direction, that is, either for positive or negative lags only. <Figure 3> shows a flow diagram of

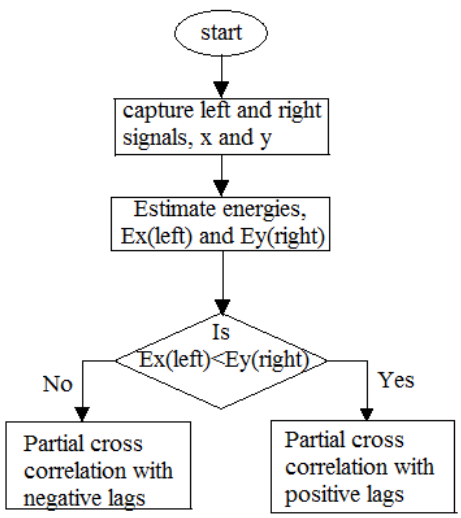
the direction prediction process. The amplitudes of vibration caused by the signal on the receiving sensor varies according to how close or farther away the sound source is, so the microphone that is closer to the sound source is expected to have a higher energy. It is noted that, the microphones used for sound capture must be identical in order to get reliable comparison between energy estimates.

We estimate the signal energy levels, $E_{x(left)}$ for signal x and $E_{y(right)}$ for signal y by taking a sum of the absolute values of their samples [Halim et al., 2011] as shown in equations 4 and 5.

$$\text{Energy of x: } E_{x(left)} = \sum_{i=0}^{N-1} |x_i| \quad (4)$$

$$\text{Energy of y: } E_{y(right)} = \sum_{i=0}^{N-1} |y_i| \quad (5)$$

where N is the signal length and x_i and y_i are the samples of signals x and y respectively.



〈Figure 3〉 Flow Diagram for Direction Prediction

3.3 Partial Cross Correlation using Energy Estimation

As mentioned above, we eliminate majority of the traditional cross correlation computations by computing the limits of the relevant lag range which includes positive and negative lags. Using the distance between our microphones, our sampling rate and the assumed velocity of sound, we determine the minimum and maximum delay in samples as described in section 3.1.

Next, we predict which part of the cross correlation function will include the TDOA value by using the signal energy levels as described in section 3.2. When the energy level of signal x is greater than that of signal y, that is, $E_{x(left)} > E_{y(right)}$, the negative lag values are used in the partial cross correlation. Similarly, when the energy level of signal y is greater than that of signal x, that is, $E_{x(left)} < E_{y(right)}$, the positive lag values are used as shown in equation 6. We name this method the Partial Cross Correlation.

$$\text{CrossCorr(x, y)(j)} = \begin{cases} \sum_{k=0}^{N-1-j} x_{(k+j)} \cdot y_{(k)}, & E_{x(left)} < E_{y(right)} \\ \sum_{k=0}^{N-1-j} x_{(k)} \cdot y_{(k-j)}, & E_{x(left)} > E_{y(right)} \\ 0, & else \end{cases}$$

$$if \quad E_{x(left)} < E_{y(right)}, \quad 0 \le j \le \max\tau$$

$$if \quad E_{x(left)} > E_{y(right)}, \quad \min\tau \le j \le 0 \quad (6)$$

where $max\tau$ and $min\tau$ are df/v and -df/v, and $E_{x(left)}$ and $E_{y(right)}$ are the energies of signals x and y respectively.

4. Azimuth Computation

The azimuth calculation [John et al., 2004] is dependent on the Time Difference of Arrival value. It is basically a computation of the angle of incidence of the sound received at the two microphones. After the TDOA value is obtained by taking the offset of the maximum correlation value, the following variables are employed to get the value of the angle. The first variable is the time increment used for sampling the signals. This variable, delta △, is determined from the sampling rate used in recording sound. In our system, we used a sampling rate of 44.1 kHz, that is 44,100 samples per second and our delta value is shown in equation 7.

$$\text{Delta} = \frac{1}{44{,}100} = 2.2676\times10^{-5} \qquad (7)$$

Other variables are the velocity of sound, delay time and the distance d between the two microphones. The velocity of sound, v, is assumed to be $343m/s$ and the delay time is the delay in samples multiplied by delta, see equation 8. Azimuth (angle) value θ, is derived based on the trigonometry function arcsine, as shown in equation 9.

$$t = \Delta \times \tau \qquad (8)$$

$$\theta = arcsine\frac{vt}{d} \qquad (9)$$

5. Experiment and Discussion

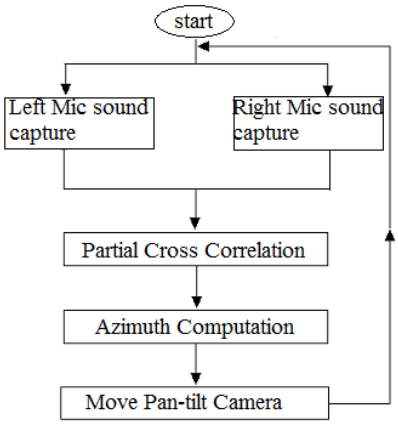
This sound source localization method is implemented on an Intel core PC in C++ using Portaudio library for real-time audio capture. The set-up for our system includes two dynamic cardioid microphones set at a distance of 30cm apart and a pan-tilt camera. The microphones are connected to an M-Audio Mobilepre USB multichannel audio interface which takes two microphone inputs through its channel 1 and channel 2 XLR ports, see <Figure 4>.

(a) Multichannel audio interface (b) Two cardioid microphones (c) Pan-tilt camera

〈Figure 4〉 Set-up for Experiments

The experimental data we used were stereo audio signals captured in real-time by using our microphone array. A person stands or sits at a distance of at least one meter from the microphone set-up and either speaks out loud or clap their hands to generate sound. We capture sound for a short period of time and the signal energy levels are estimated using equations 4 and 5 stated in section 3.2. We then estimate the direction of the sound source in the time domain by using both the traditional cross correlation function and the proposed partial cross correlation for TDOA computation and compared their effect on the speed of the TDOA estimation.

<Figure 5> shows the flow diagram of the sound source localization. Stereo sound is captured by the use of our microphone array and the two components of the signal are passed to the cross correlation function for TDOA estimation. After that, the azimuth of the sound source is calculated and the value of the angle is used to move the pan-tilt camera.



〈Figure 5〉 Flow Diagram for Sound Source Localization

A significant improvement in the speed of the cross correlation method is seen in the proposed partial cross correlation method. There is about a 50% reduction in the computation time and the speed of the TDOA estimation is 2 times as fast as the traditional cross correlation. Furthermore, our partial cross correlation shows reliable accuracy in the angle tests, which is comparable to that of the traditional cross correlation method.

The results of our experiments are shown in tables 1, 2, 3 and 4. <Table 1> presents a comparison of the computation time spent by the traditional cross correlation using the relevant lag range and the proposed partial cross correlation methods. Some results of the angle tests performed with the partial cross correlation method are shown in <Table 2>, and the statistics of these results are presented in <Table 3>. <Table 2> shows the angle tests performed in ten different scenarios for ten different angles. For each scenario corresponding to a particular angle, we tested the system 12 times to confirm its accuracy.

<Table 4> shows the performance of our energy-based direction prediction. It shows the direction predictions made in 10 random scenarios, in comparison with their computed delay values. Each delay value is the offset of the maximum correlation value for the two signals using the cross correlation. It shows how much one signal lags behind the other in terms of samples. Negative and positive delay values mean sound originates from the left and the right sides respectively. By comparing the predicted direction with the cross correlation delay values, we confirmed the accuracy of the energy-based direction prediction.

〈Table 1〉 Comparison of Computation Time

Cross Correlation Method	Time Spent(ms)
Relevant Lag Range	13.0
Partial Cross Correlation	7.0

〈Table 2〉 Results of Angle Tests

Scenario/Angle	1	2	3	4	5	6	7	8	9	10	11	12
1/+80	80.12	73.59	73.59	73.59	80.12	80.12	80.12	80.12	73.59	73.59	80.12	80.12
2/+60	65.15	65.15	61.82	65.15	61.82	65.15	65.15	65.15	65.15	65.15	65.15	61.82
3/+45	44.42	44.42	44.42	42.38	42.38	42.38	44.42	42.38	44.42	42.38	44.42	44.42
4/+30	27.81	27.82	26.15	26.15	27.81	29.51	29.51	29.51	31.23	31.23	29.51	29.51
5/+10	7.44	8.94	8.94	8.94	7.44	7.44	10.45	10.45	10.45	10.45	8.94	8.94
6/0	0	-1.48	-1.48	-1.48	1.48	0	1.48	1.48	0	0	0	-1.48
7/-15	-16.57	-16.57	-16.57	-16.57	-16.57	-15.02	-16.57	-16.57	-16.6	15.02	-15.02	-18.12
8/-40	-38.47	-38.47	-40.4	-40.4	-40.4	-40.4	-38.47	-38.47	-38.5	-40.4	-42.38	-38.47
9/-50	-46.54	-46.54	-46.54	-48.75	-46.54	-46.54	-46.54	-46.54	-46.5	-48.75	-48.75	-48.75
10/-60	-58.82	-58.82	-56.06	-58.82	-58.82	-58.82	-58.82	-53.84	-58.8	-58.82	-56.06	-53.84

〈Table 3〉 Statistics of the Angle Test Results

Test Scenario	Actual Angle	Average Angle	Confidence (Uncertainty)
1	+80	77.39	±4
2	+60	64.31	±3
3	+45	43.57	±1
4	+30	28.95	±2
5	+10	8.69	±2
6	0	-0.74	±2
7	-15	-16.31	±2
8	-40	-39.43	±1
9	-50	-47.46	±2
10	-60	-57.50	±4

〈Table 4〉 Predicted Direction by Energy Level vs Delay Value by Cross Correlation

Test	E_y(right)	E_x(left)	Predicted Direction	Delay Value
1	45.031647	68.428101	Left	-7
2	63.609283	94.671051	Left	-19
3	31.923920	61.448822	Left	-11
4	43.439117	79.583923	Left	-25
5	37.790649	57.022675	Left	-31
6	96.873901	86.539734	Right	19
7	101.08813	97.115051	Right	27
8	87.131775	82.360647	Right	8
9	93.190979	82.964905	Right	13
10	95.004456	82.597687	Right	38

The reasons for the significant improvement in the computation time of the TDOA estimation are, first, the use of a reduced lag range and second, the use of energy estimate-based direction prediction. Parameters form our experimental set-up are used to check the relevant lag range needed for the cross correlation by computing the maximum and minimum delay values. This step initially eliminates majority of the unnecessary computations as it reduces the number of coefficients to be computed. We do not have to compute 2N-1 correlation coefficients as in the traditional cross correlation as well as the frequency domain implementation of GCC.

Furthermore, we predict which side of the relevant lag range will contain the TDOA value, and then we proceed by computing only the predicted side of the correlation function. We make use of the energy levels of the signals to predict which part of the cross correlation will contain the TDOA value. By computing coefficients for only half of the relevant lag range, we get approximately 50% reduction in the computation time as compared to the traditional cross correlation. This gives a significant increase in speed of the TDOA estimation.

6. Conclusion

This paper describes a fast Time Difference of Arrival (TDOA) for sound source localization using a proposed partial cross correlation in time domain. TDOA estimation in the time domain is not fast enough, especially when large windows of signals are being processed.

The partial cross correlation method which is implemented in time domain is in two steps. First, the relevant lag range of the cross correlation is determined in order to avoid computing unnecessary coefficients. Secondly, the energy levels of the signals are estimated in order to predict which part of the relevant lag range contains the TDOA value. Specifically, we predict that the TDOA value will be in the direction of the signal with higher energy level. The cross correlation is then implemented partially, in the predicted direction and this drastically reduces the computation time. Our experimental results show that the proposed partial cross

correlation method gives approximately 50% reduction in computation time, which speeds up the TDOA estimation. It also shows reliable accuracy during the angle tests. Its shorter response time for TDOA estimation shows that it can be used in real world surveillance applications for better effectiveness.

In future, specific audio signals such as screaming or gunshot will be detected prior to the TDOA estimation in order to improve the effectiveness of the system.

References

[1] Bert, V. D. Broeck, Bertrand, A., Karsmakers, P., Vanrumste, B., H. Van hamme, and Moonen, M., "Time-Domain GCC-PHAT Sound Source Localization for Small Microphone Arrays", Education and Research Conference (EDERC), 2012 5th European DSP, 2012, pp. 76-80.

[2] Qin, B., Zhang, H., Fu, Q., and Yan, Y., "Subsample Time Delay Estimation via Improved GCC PHAT Algorithm", Proceedings of 9th International Conference on Signal Processing, 2008, pp. 2979-2982.

[3] Knapp, C. H. and Carter, G. C., "The generalized correlation method for estimation of time delay", *IEEE Transactions on ASSP*, Vol. 24, No. 4, 1976, pp. 320-327.

[4] McAlpine, D. and Grothe, B., "Sound localization and delay lines-do mammals fit the model?", *TRENDS in Neuroscience*, Vol. 26, No. 7, 2003, pp. 347-350.

[5] Sayoud, H., Ouamour, S., and Khennouf, S., "Speaker Localization using Stereo-based Sound Source Localization", 2011 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), 2011, pp. 231-234.

[6] Liu, H. and Shen, M., "Continuous Sound Source Localization based on Microphone Array for Mobile Robots", IEEE/RSJ International Conference on Intelligent Robots and Systems, 2010, pp. 4339-4339.

[7] Stachurski, J., Netsch, L., and Cole, R., "Sound Source Localization for Video Surveillance Camera", 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2013, pp. 93-98.

[8] Valin, J.-M., Michaud, F., Hadjou, B., and Rouat, J., "Localization of simultaneous moving sound sources for mobile robot using a frequency domain steered beamformer approach", IEEE International Conference on ICRA, pp. 1033-1038, New Orleans, USA, 2004.

[9] Murray, J. C., Erwin, H., and Wermter, S. "Robotic Sound-Source Localization and Tracking Using Interaural Time Difference and Cross Correlation", AI workshop on Neuro-Botics, Germany, 2004.

[10] Lobos, T., Leonowicz, Z., Rezmer, J., and Schegner, P., "High-resolution spectrum-estimation methods for signal analysis in power systems", *IEEE Transactions on Instrumentation and Measurement*, Vol. 55, No. 1, 2006, pp. 219-225.

[11] Lin, Z., Ziao-Yan, Z., Xu, C., and Zhen-Yang, W., "Binaural Sound Source Localization based on Sub-band SNR Estimation", *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 10, No. 5, 2015, pp. 303-314.

Author Profile

Mariam Yiwere

She received her B.S. degree in Computer Science from the Kwame Nkrumah University of Science and Technology, Kumasi, Ghana in 2012. In 2015, she received her M.S. degree in Computer Engineering from Hanbat National University, Daejeon, Korea. She is currently conducting research in the area of sound source localization in the Artificial Intelligence and Computer Vision Lab in the Graduate School of Information and Communications, Hanbat National University, She is interested in computer vision, digital signal processing and artificial intelligence.

Eun Joo Rhee

He is a Professor of Department of Computer Engineering at College of Information Technology, Hanbat National University, Daejeon, Korea. He has the degree of Ph.D in Electronics Engineering from Chungnam National University. His research interests include in image processing, pattern recognition, computer vision and artificial intelligence. His papers have appeared in IEICE Trans. on Information and Systems, Journal of KIISE, Journal of the Institute of Electronics and Information Engineers, Journal of Information Technology Applications and Management, Journal of Information Technology Application, Journal of the Modern Linguistic Society of Korea, Journal of Korea Multimedia Society, Journal of the Korea Academia-Industrial cooperation Society.