University of Kentucky
UKnowledge

Theses and Dissertations--Electrical and Computer Engineering

Electrical and Computer Engineering

2017

CONSTANT FALSE ALARM RATE PERFORMANCE OF SOUND SOURCE DETECTION WITH TIME DELAY OF ARRIVAL ALGORITHM

Xipeng Wang
University of Kentucky, zywxp0430@gmail.com

Digital Object Identifier: https://doi.org/10.13023/ETD.2017.343

Xipeng Wang, Student

Dr. Kevin D. Donohue, Major Professor

Dr. Cai-Cheng Lu, Director of Graduate Studies

CONSTANT FALSE ALARM RATE PERFORMANCE
OF SOUND SOURCE DETECTION WITH
TIME DELAY OF ARRIVAL ALGORITHM

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in
Electrical Engineering in the College of Engineering
at the University of Kentucky

By

Xipeng Wang

Lexington, Kentucky

Director: Dr. Kevin Donohue, Professor of University of Kentucky

Lexington, Kentucky

2017

Abstract of thesis

CONSTANT FALSE ALARM RATE PERFORMANCE OF SOUND SOURCE DETECTION WITH TIME DELAY OF ARRIVAL ALGORITHM

Time Delay of Arrival (TDOA) based algorithms and Steered Response Power (SRP) based algorithms are two most commonly used methods for sound source detection and localization. SRP is more robust under high reverberation and multi-target conditions, while TDOA is less computationally intensive. This thesis introduces a modified TDOA algorithm, TDOA delay table search (TDOA-DTS), that has more stable performance than the original TDOA, and requires only 4% of the SRP computation load for a 3-dimensional space of a typical room. A 2-step adaptive thresholding procedure based on a Weibull noise peak distributions for the cross-correlations and a binomial distribution for combing potential peaks over all microphone pairs for the final detection. The first threshold limits the potential target peaks in the microphone pair cross-correlations with a user-defined false-alarm (FA) rates. The initial false-positive peak rate can be set to a higher level than desired for the final FA target rate so that high accuracy is not required of the probability distribution model (where model errors do not impact FA rates as they work for threshold set deep into the tail of the curve). The final FA rate can be lowered to the actual desired value using an M out of N (MON) rule on significant correlation peaks from different microphone pairs associated is a point in the space of interest. The algorithm is tested with simulated and real recorded data to verify resulting FA rates are consistent with the user-defined rates down to 10^{-6}.

KEYWORDS: Sound Source Localization, Multiple Targets Detection, Time Delay of Arrival, Constant False Alarm Rate, M out of N Rule

Xipeng Wang

07/29/2017

CONSTANT FALSE ALARM RATE PERFORMANCE
OF SOUND SOURCE DETECTION WITH
TIME DELAY OF ARRIVAL ALGORITHM

By

Xipeng Wang

Kevin D. Donohue
Director of Thesis

Cai-Cheng Lu
Director of Graduate Studies

07/29/2017

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction and literature review

1.1 Introduction

Near field sound source localization is a widely-studied project. It’s used in speech enhancement [1-2], intelligent interaction [3-6], and noise cancellation [7-9]. Two most common approaches are Time Delay of Arrival (TDOA) and Steered Response Power (SRP). SRP algorithm is proved to have a more robust performance in noise and reverberant environment. The performance analysis done by DiBiase in 2001 clearly demonstrated stable detection results in reverberant/noise environments [10]. However, in many applications the SRP system must scan through the whole room space with relative small spatial increments and compute the resulting power from the beamformed signal at each grid point in the space. This can require significant amount of processing time or a massively parallel processing scheme [11] for most typical rom sizes. On the other hand, TDOA only requires cross-correlation peak results from several microphone pairs to find the targets delay of arrival (DOA), which are then used to determine intersection points corresponding to the lots of points implied by each microphone pair’s DOA. This procedure requires less computations than SRP over large spaces, but is less robust. Therefore, the purpose for this thesis is to develop and test a more efficient way to perform the sound source detection and localization with little performance loss compared to SRP.

1.2 Literature review

Many studies have been done to improve two detection algorithms recent years. Work has been done by Hoang in 2012 on SRP algorithm [12]. This article points out that when doing the phase generalization for all cross-correlations of unique microphone pair signals, the correlation value could be poor because the mismatch of two microphone signals. The Figure 1.1 shows a mismatch of temporal events. This mismatch, when repeated in many microphone pairs, will lead to a low value of the SRP algorithm [13]. A pre-alignment enhancement was proposed in the article. The enhancement algorithm pre-aligns the microphone signals to a hypothesized location by shifting the signals. The experimental results show the improved method has a better performance under the 10 talkers experiment compared to the normal SRP algorithm. The limitation of this improvement is that the high computation load of SRP algorithm still exist and the need for stable thresholding algorithms to detect significant peaks as actual targets.

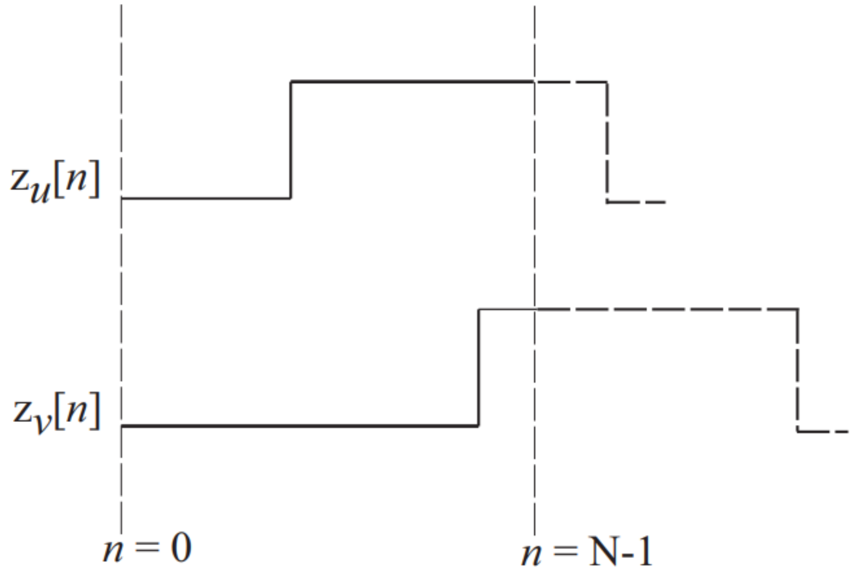


Figure 1.1 Mismatch of temporal events between microphone signals z_u and z_v. N denotes the time-frame size.

In 2013, Firoozabadi did some improvements on original TDOA algorithm [14]. Instead of using original signal received by microphone, this method first decomposes the signal into different frequency sub bands. For every sub band, the Generalized Cross-Correlation (GCC) is calculated. Then this algorithm find the M highest peaks (DOA candidates) in each GCC result, where M denotes the maximum possible targets. With all DOA candidates in each sub band, the histogram for the DOA value is computed. At last, applying an averaging method on the achieved histogram for all sub bands. Because locations with a target should have more power over a broad frequency range, than those without a target, the DOA values for target locations show more consistent DOA (peaks in the histogram) than the noise locations. Therefore, the M highest peaks of averaged histograms are considered as the targets DOAs. The results in this paper shows that the sub-band processing increased the percentage of correction at about 11-17 percent. However, this method uses the M highest value as the targets, which means this algorithm will also give out detection results even with no target presented.

Jamali-Rad and Leus also proposed improvements on the TDOA in 2013 [15]. In the TDOA algorithm, after finding possible targets DOA peaks, the DOA values will be used in a hyperbola formula. All the points on this hyperbola have the same DOA for a given microphone pair. With two or more hyperbolas crossing with each other, the targets locations will be settled. One major problem is that there may be two or more targets on the same hyperbola, which can result in missing a detection. The method proposed in this article discretizes the space into a special grid. Points on this grid will not have similar DOA time for any of the microphone pairs. Then for each point on this grid the combination of DOA values for all microphone pairs will be unique and the targets can be detected by comparing the estimated DOA values with all grid points. The DOA values come from the maximum peaks locations in the cross-correlations results. The most closer points are the target locations. This article solves the problem that two target points may fall on the same hyperbola curves, but it still uses the K max peaks detection algorithm like Firoozabadi's article talked above.

To speed up the SRP algorithm, Yook proposed a space cluster method in 2016 [16]. The article points out that when two candidate locations are close enough, SRP value of these points will have indistinguishable power difference. Therefore, the space can be clustered into several subspaces, candidate locations in these spaces will have similar output power. Then the first stage of this method will only exam the representative locations in each subspace and the subspace containing a maximum power will be passed into the next step. The second level search will be performed to find the target location in the chosen subspace. This space clustering method reduces 61.8% computational cost comparing to the none clustering one. However, the method is proposed under the far field condition and using the maximum value as the result.

In 2014, Donohue and Griffioen studied on a new computational strategy to accelerate the SRP procedure [11]. This new strategy divides the received signals into 0.26s subintervals, which are much larger than the typical analysis window (less than 50ms). Then for every subinterval, the first analysis window is tested over space points for active targets. The subinterval will only be tested throughout when the first analysis window finds at least one active sources. Apart from the time reduction, space reduction is also applied. After the algorithm detected a target, a spherical subspace will be defined while the radius being calculated based on the expected moving speed of the source. Then the detection procedure only runs through the spherical sub space for the next subinterval. To further accelerate the SRP algorithm, a parallelization is applied. The field of interest is divided into several regions and each region will be separately run. The experiment results show that the accelerated algorithm is about 45 times faster than the original one. Therefore, the new computational strategy shows a huge speed enhancement. However, one obvious problem for the acceleration is that when a target is not presented or detectable during the first analysis window, then the target will be lost for the entire 0.26s detection.

Another approach for detecting multiple targets with the SRP algorithms by Donohue, *et. al.* in 2011 [17] focuses on detecting significant peaks in the SRP image based on a constant false-alarm rate threshold. Many sound source detection algorithms, such as [14, 15], use the maximum value finding the target peaks, but the maximum value method must set the possible number of targets before detection. It is unknown for most real-world cases, therefore this article proposed a self-adjusted thresholding algorithm. In the SRP detection, the field of interest is divided into small spatial increments, where sound source locations will result in a higher power than the locations where no source is present. Therefore, the article points out that the threshold value for a considered peak can be estimated by the peak values around the tested peak. The surrounding region is assumed to contain only noise peaks, and by choosing an appropriate noise distribution and tolerable false alarm rate, a threshold can be estimated from the power computed from the surrounding noise peaks. Because this thresholding method is based on adjusting the power in the distribution to maintain the FA rate over all tested peaks, it is called Constant False Alarm Rate (CFAR) threshold. This method performs well during the experiments running in this article. However, the drawback is that the CFAR threshold become sensitive to the noise distribution modeling errors for very low false alarm rates.

These articles presented above all have very interesting aspects and provide a general view of recent research done on the multiple sound source detection and localization problems. Hoang's research optimizes the SRP algorithm on the accuracy. However, it doesn't deal with the computation load problem for SRP. Firoozabadi's method focuses on solving the far field sound source localization and has a drawback that it must assume the exact number of targets beforehand (i.e. M possible targets during the peaks finding progress). This assumption is applied even if no target is present, and thus, this algorithm will consistently result in a false-alarms when applied to dynamic acoustic scenes, such as those found in application like smart room or surveillance/monitoring. Jamali's study provide a new way to applying TDOA method by using space grid points, but its solution still relies on the assumption of exact number of targets before the detection. The next two articles [11, 16] focused on the speed improvement for SRP algorithm. They both use subsections to reduce the computation load. Although their results show a huge acceleration, Yook's method still has the maximum value problem when detecting targets and Dobohue's algorithm may lose some targets. Sayed's study proposed a new way to detect significant peaks with constant false alarm rate. This CFAR threshold will not cause the same problem as the maximum value method, but it will be unstable under low false alarm conditions due to modeling errors for the assumed noise distribution.

The work in this thesis addresses the problem of multiple target detection using a significant peak detection approach on the microphone pair cross-correlations. It addresses the modeling errors of the noise peak probability density function by setting a higher false-alarm rate threshold for the cross-correlation peaks, which keeps the threshold value out of the tail of the distribution where small modeling errors result in large variation in the threshold value. Then using an M out of N rule on all significant cross-correlation peaks associated with each point in space. This new approach will also cost less computation load comparing to SRP without sacrificing performance. The algorithms are tested with simulated data in terms maintaining the desired false-alarm rate and compared to that of the SRP as lower FA rates are specified.

The thesis describes two classic sound source localization algorithms in Chapter 2, and the novel method proposed in this thesis is also described. Chapter 3 presents details on the CFAR threshold process, analyzes its drawbacks, and proposes a new double threshold algorithm using M out of N rule. Chapter 4 shows the simulation results of the algorithms and the experiments with real recordings to verify the simulation results.

Chapter 2 Sound source localization algorithms

2.1 Introduction

This chapter introduces two widely used sound source localization algorithms TDOA and SRP. Section 2.2 presents the equations for TDOA and analyzes the disadvantages that makes it not suitable for multiple targets detection. Then Section 2.3 presents SRP-based algorithms and points out the high computation load problem. At last, Section 2.4 proposed a new algorithm based on TDOA with less calculation cost and better multiple targets detection ability.

2.2 TDOA-based sound source localization approaches

A time delay of arrival (TDOA) localization approach is a two-step procedure [18]. The first step is getting TDOA estimates from signals received by each possible microphone pairs. The second step generates hyperbolic curves corresponding to the sets of possible locations that resulted in the measured delay. Curves from multiple pairs corresponding to the same source intersect at the target location.

Let $u_i(t)$ denotes the i^{th} sound source located at $\mathbf{s}_{\mathrm{i}}$ and total K microphones. Then the signal received by kth microphone located at $\boldsymbol{m_k}$ can be represented as

$$v_k(t) = u_i(t) * h_{ik}(t) + \sum_{j=1}^{J} n_j(t) * h_{jk}(t)\,, k = 1,2,3, \dots , K \qquad (2.1)$$

where J represents the total number of noise sources (sources not located at $\mathbf{s}_{\mathrm{i}}$), $n_j(t)$ is the j^{th} noise source and $h_{ik}(t)$ denotes the impulse responses for the propagation path between s_i and m_k, and can include the multipath effects. Similarly, $h_{jk}(t)$ is the impulse response for the acoustic path between n_j and m_k. Also, "$*$" stands for convolution operator.

In a real room or small space, the reverberation must be considered. Therefore, the impulse function $h(t)$ should take both direct and reflected paths into consideration:

$$h_{ik}(t) = a_{ik,0}\left(t - \tau_{ik,0}\right) + \sum_{n=1}^{\infty} a_{ik,n}\left(t - \tau_{ik,n}\right) \qquad (2.2)$$

where $a_{ik,0}(t)$ represents the direct path between the i^{th} source and k^{th} microphone, $a_{ik,n}(t)$ represents the nth reflection, and $\tau_{ik,n}$ represents the delay time of this path. The larger n means longer path, which also means the power of the signal is weaker. The simulation used in this study takes in to consideration all these effects for a rectangular room [19, 20]

To estimate TDOA, a most common used way is generalized cross-correlation (GCC) [18]. GCC is calculated based on two filtered microphone signals using the Fourier transform (fast Fourier transform in coding) to make the computation more effective. The signal received by k^{th} microphone can be represented using the equation 2.1. The received signal is denoted as $y_k(t)$, and its Fourier transform given by:

$$Y_k(\omega) = \mathcal{F}\big(y_k(t)\big) = H_k(\omega)V_k(\omega) \qquad (2.3)$$

where $V_k(\omega)$ is the Fourier transform of the source signal $v_k(t)$ and $H_k(\omega)$ is the Fourier transform of the room impulse response (transfer function). The same process applied on the signal received by l^{th} microphone is denoted as:

$$Y_l(\omega) = \mathcal{F}\big(y_l(t)\big) = H_l(\omega)V_l(\omega) \qquad (2.4)$$

The two signals are cross-correlated to result in:

$$c_{kl}(\tau) = \mathcal{F}^{-1}\{C_{kl}(\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} C_{kl}(\omega)e^{j\omega\tau}\, d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} Y_k(\omega)Y_l^*(\omega)e^{j\omega\tau}\, d\omega \qquad (2.5)$$

where superscript $'*'$ represents the complex conjugate operator. The cross-correlation result $c_{kl}(\tau)$ is a function of delay times, τ, between the microphone pair. Substitute in equation 2.3 and 2.4 to obtain:

$$c_{kl}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} H_k(\omega)H_l^*(\omega)V_k(\omega)V_l^*(\omega)e^{j\omega\tau}\, d\omega \qquad (2.6)$$

To enhance the peak detection in the cross-correlation function $c_{kl}(\tau)$, many filters have been proposed. One is the generalized cross-correlation (GCC) filter. Ideally, the GCC requires knowledge of the room transfer function. The filter in this case is given by:

$$\Psi_{kl}(\omega) = \frac{1}{\left|H_k(\omega)H_l^*(\omega)\right|} \qquad (2.7)$$

Since these are difficult to estimate and change for every point in the room, other approaches are typically used. One widely used filter is Phase Transform (PHAT) filter [18, 21]. The PHAT filter is designed to a special value to emphasize the local maximum peaks, and is implement by normalizing all spectral magnitudes. The PHAT weighting function is defined as:

$$\Psi_{kl}(\omega) = \frac{1}{\left|V_k(\omega)V_l^*(\omega)\right|} \qquad (2.8)$$

The filter is applied to cross-correlation as:

$$c_{kl}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_{kl}(\omega)H_k(\omega)H_l^*(\omega)V_k(\omega)V_l^*(\omega)e^{j\omega\tau}\, d\omega \qquad (2.9)$$

After applying the PHAT filter, the room transfer function will be eliminated in ideal case. This filter can also be applied in SRP algorithm and it will be stated in next chapter.

Next step is finding the peaks generated by targets in GCC function $c_{kl}(\tau)$. Assume a noise free and reverberation free condition, the $v_k(t)$ and $v_l(t)$ are just two time shifted results for the original signal $u_i(t)$. Therefore, the $c_{kl}(\tau)$ function will have a significant peak at time τ related to TDOA of two microphones k and l. In the single sound source case, the TDOA between microphone pairs can be detected by finding the maximum value of GCC function. However, with two or more sources present (or even no source present), the detection problem changes into finding local significant maxima of GCC function. A practical way is using a Constant False Alarm Rate (CFAR) approach [17], which will be described in detail later.

Each detected peak must be related to peaks from other microphone pair cross-correlations corresponding the same position in space to determine the likelihood that an actual target is presented. For each delay time, corresponding to the detected peak, a curve is drawn using the equation below. Assuming the position of the microphones are $\mathbf{m_1}(x_1, y_1, z_1)$ and $\mathbf{m_2}(x_2, y_2, z_2)$, the relationship between the delay time and source position is given by:

$$\sqrt{(x-x_1)^2+(y-y_1)^2+(z-z_1)^2}-\sqrt{(x-x_2)^2+(y-y_2)^2+(z-z_2)^2}=d_t*c \qquad (2.10)$$

where the d_t represents delay time and c is the speed of sound. These equations are non-simplified hyperbola formulas. Since for the sound source target, only one side of the hyperbola curve should be taken. Therefore, the sign of $\sqrt{(x-x_1)^2+(y-y_1)^2+(z-z_1)^2}-\sqrt{(x-x_2)^2+(y-y_2)^2+(z-z_2)^2}$ is used to determine which side should be taken as valid. With two or more curves drawn, intersections indicate the possible locations of sound source target(s). A two-dimensional example has been shown in Figure 2.1. The delay time obtained by cross-correlation function of A_1 and A_2 generates the curve H_{12}; delay time obtained by cross-correlation function of A_2 and A_3 generates the curve H_{23}. The two curves intersected at “Event”, where will be the detected target position.

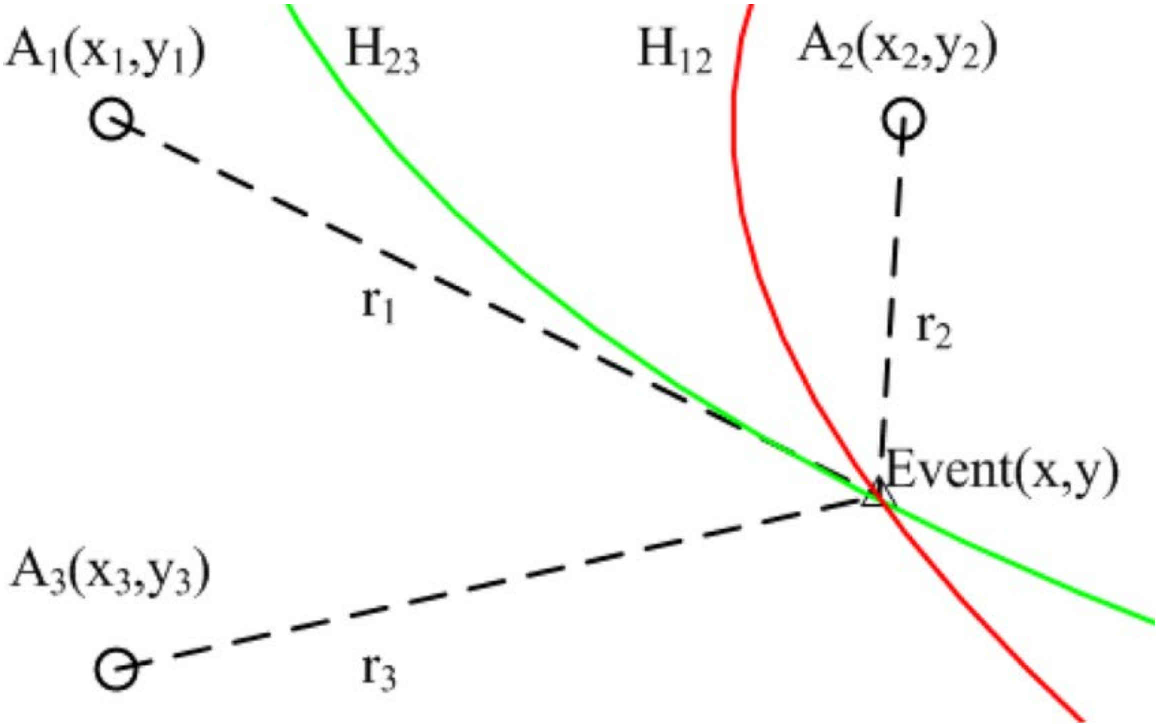


Figure 2.1 A two-dimensional hyperbola curves detection with A_1, A_2 and A_3 microphones, and the detected target position is at Event. Adapted from [22]

In the real detections, a three-dimensional room is the common case. The detection become difficult in three-dimensional space, because the intersections for a single target are typically contained in small three-dimensional volumes. With three or more hyperboloids crossing each other, they won't meet at the same point because the noise or measure error [23]. This also leads to another problem that the different intersections are false detections or actual targets. Consequently, the simple TDOA implementation is not suitable for the multi-targets detection and localization.

2.3 SRP-based sound source localization approaches

Steered Response Power (SRP) algorithm relies on a delay and sum beamforming method to compute acoustic power over grid points in a spatial region of interest. For a certain sound sources located in the region of interest, the SRP algorithm uses the delay times corresponding to each grid point to shift the signals received by microphones and add time-shifted signals together to obtain a power estimate for a sound source emanating from that position. If the position corresponds to a real sound source, the sum of shifted signals obtained should have a significantly higher power value than positions corresponding to no source. Therefore, the places that have significant power peaks are treated as the sound sources in multiple targets SRP localization [11,24]. Figure 2.2 shows the structure of a delay-and-sum system.

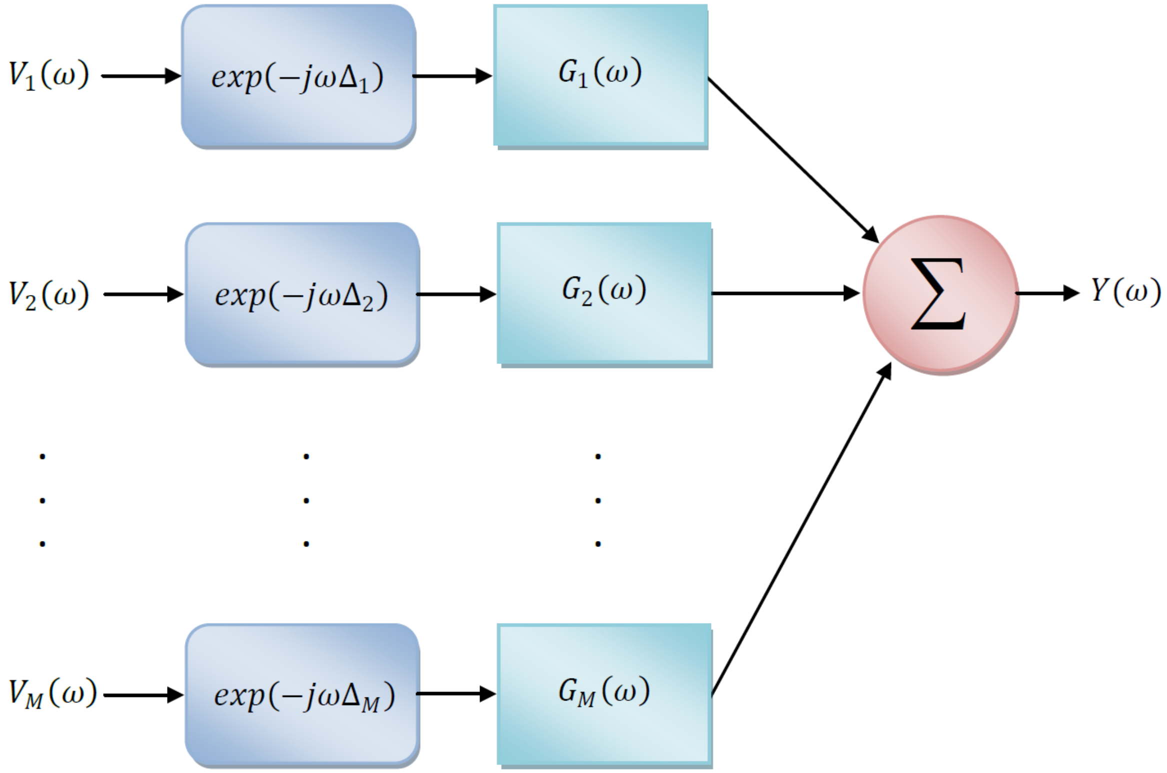


Figure 2.2 A delay-and-sum system structure with $G_M(\omega)$ (impulse response of filter in frequency domain) applied

In last section, the Equation 2.1 shows that microphones signals are filtered versions of shifted source signal, which is to say that microphone signals are time aligned when properly shifted by propagation delays. The shift time are represented as Δ_m for the m^{th} microphone. Let $y(t)$ represents the steered response, then:

$$y(t) = \sum_{m=1}^{M} g_m(t) * v_m(t - \Delta_m) \qquad (2.11)$$

where M denotes the total number of microphones, $g_m(t)$ is the impulse response of the filter applied on m^{th} microphone signal and $v_m(t - \Delta_m)$ is the shifted signal. The $' * '$ represents the convolution operator. The filter applied here is a part of PHAT weighting function.

Then take Fourier transform of the steered response:

$$Y(\omega) = \sum_{m=1}^{M} G_m(\omega) V_m(\omega) e^{-j\omega\Delta_m} \qquad (2.12)$$

where $G_m(\omega)$ and $V_m(\omega)$ are Fourier transform of the filter applied and the mth signal. With the frequency domain equation, the SRP can be written as:

$$P = \int_{-\infty}^{+\infty} Y(\omega) Y^*(\omega)\, d\omega$$

$$= \int_{-\infty}^{+\infty} \left(\sum_{p=1}^{M} G_p(\omega) V_p(\omega) e^{-j\omega\Delta_p} \right) \left(\sum_{q=1}^{M} G_q^*(\omega) V_q^*(\omega) e^{j\omega\Delta_q} \right) \qquad (2.13)$$

where P denotes the steered response power. This equation can be rewritten as:

$$P = \sum_{p=1}^{M} \sum_{q=1}^{M} \int_{-\infty}^{+\infty} G_p(\omega) G_q^*(\omega) V_p(\omega) V_q^*(\omega) e^{j\omega(\Delta_q - \Delta_p)} d\omega \qquad (2.14)$$

where, like equation 2.6, $G_p(\omega) G_q^*(\omega)$ is called weighting function and usually written as $\Psi_{pq}(\omega)$. Much research has been done on the weighting function, and the SRP-PHAT has often been cited as demonstrating the most robust performance [25]. Further research on the PHAT weighting function shows that partial PHAT weighting (PHAT $-$ β) will be superior to the original PHAT weighting [19, 21]. The partial PHAT weighting function is defined as:

$$\Psi_{pq}(\omega) = \frac{1}{\left(|V_p(\omega)||V_q^*(\omega)|\right)^{\beta}} \qquad (2.15)$$

where β is the partial weighting parameter and it’s a real number between 0 and 1. The experiments indicate that a better target detection results can be obtained by setting partial weighting parameter close to 1 [19, 21]. The PHAT weighting function also can be applied in equation 2.2 to improve TDOA algorithm.

Comparing to TDOA-based algorithm, SRP method spends a much longer time scanning the space and computing a power value for each grid point, but more than one target can be detected at the same time. The position where sound sources are located will show significant peaks. It's worth mentioning that in the equation above, the autocorrelation pairs are included in the power computation. The autocorrelation pairs will have no effect on the results, except to keep the power values positive, so usually these pairs are subtraction out to improve detection performance [21].

2.4 TDOA delay table search sound source localization

The TDOA algorithm described above is complicated by the 3-D hyperbolic curves that need to be used to find intersection points. The false detections will have huge impact on performance if not properly controlled. Therefore, to reduce the complexity of looking for intersections and create a system for detecting targets based on a constant false alarm thresholding procedure, a new algorithm called TDOA Delay Table Search (TDOA-DTS) is developed. The new algorithm will have two steps. First step is the same as TDOA-based algorithm, except that peaks in the cross-correlation are selected based on a constant false alarm threshold. So, the result is a set of peaks detected over the set of all microphone pairs, which correspond to delay times from potential sound source. For the next step, instead of generating hyperbolic curves based on the delays, a delay time look-up table is used that relates spatial grid points to peak positions in the microphone pair cross-correlation functions. So, for each point in space a set of significant peaks (potentially empty) are used to determine if a target is present at that position. The rules for doing this and maintaining a constant false alarm rate will be described later.

The method to compute TDOA has been stated in detail in the section 2.2, so let us start from the delay time look up table. Assume a sample point in the field of interest (FOI) $\mathbf{p}_k(x_k, y_k, z_k)$, then the distance between this sample point and i^{th} microphone $\mathbf{m}_i(x_i, y_i, z_i)$ can be represented as:

$$d_{ki} = \sqrt{(x_i - x_k)^2 + (y_i - y_k)^2 + (z_i - z_k)^2} \qquad (2.16)$$

and for all N microphones, a distance matrix for sample point $\mathbf{p}_k$ is:

$$\mathbf{D}_k = [d_{k1}, d_{k2}, \ldots, d_{kN}] \qquad (2.17)$$

Furthermore, let $\Delta_{k(i,j)}$ denotes $d_{ki} - d_{kj}$, distance difference of i^{th} mic and j^{th} mic to sample point $\mathbf{p}_k$:

$$\mathbf{t}_k = \frac{\Delta_k}{c} = \frac{\left[\Delta_{k(1,2)}, \Delta_{k(1,3)}, \dots, \Delta_{k(i,j)}, \dots, \Delta_{k(N-1,N)}\right]}{c} \qquad (2.18)$$

where $i \in [1, N-1]$ and with a certain i value $j \in [i+1, N]$. c is the sound speed. The numerical meaning of this $\mathbf{t}_k$ matrix is that it represents the time difference of arrival between all possible pairs of microphones from a source to the sample point. Then for all the possible points in the FOI, a complete **T** matrix can be generated:

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1 \\ \dots \\ \mathbf{t}_M \end{bmatrix} \qquad (2.19)$$

where M denotes the total number of possible points in FOI. This **T** matrix is called the delay table where each row corresponds to a grid point in the scan space and each column is the corresponding time lag in the cross-correlation of microphone pairs.

Therefore, for each grid point in the FOI the corresponding significant peak locations in cross-correlations can be counted. If significant peaks are detected for a given point, a secondary threshold can be applied to determine if peaks are indicative of a target being present. This is explained in the next chapter.

The Figure 2.3 shows a sample room. As the equations shown above, for each grid point in this sample room, the TDOA for all possible pairs of microphones are computed. The Table 2.1 presents part of delay table for the room in Figure 2.3. Assume that the cross-correlation function of microphone 1 and 2 has one peak detected and the corresponding delay time of this peak is $-0.0084\ s$. By searching the delay table, the TDOA for grid point 3 has the closest value, so it will be marked as a possible target position.

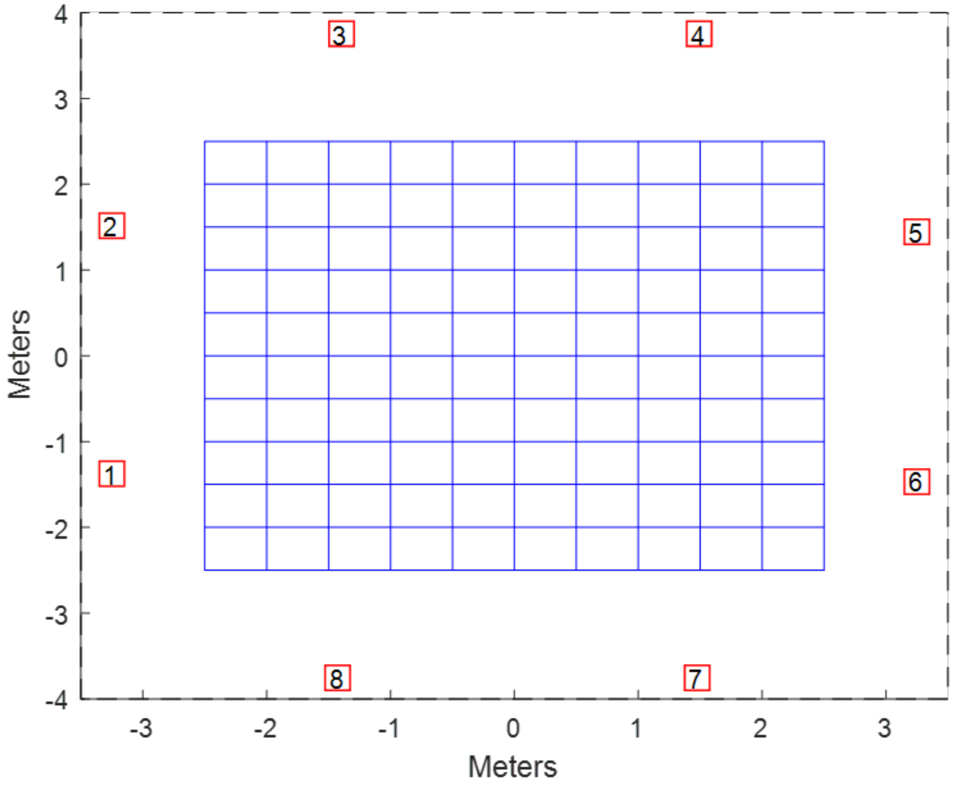


Figure 2.3 A sample room divided by grids. The red squares are the microphone positions.

Table 2.1 Part of delay table based on the Figure 2.3. Columns are all grid points and rows are all possible microphone pairs. The TDOA values are in second.

	Mic 1 and 2	**Mic 1 and 3**	**…**	**Mic 6 and 8**	**Mic 7 and 8**
Grid Point 1	-0.0083	-0.0152	…	0.0134	0.0080
Grid Point 2	-0.0083	-0.0152	…	0.0133	0.0080
Grid Point 3	-0.0084	-0.0153	…	0.0133	0.0079
…	…	…	…	…	…

The next chapter explain how this table is used for a secondary threshold procedure to improve the consistency of the experimental FA rates.

Chapter 3 Thresholding process

3.1 Introduction

This chapter focuses on the thresholding process for peaks detection in localization algorithms. First in section 3.2, the concept of false alarm rate is introduced. It is an important parameter when computing the threshold. Then Section 3.3 presents the CFAR thresholding method and states the drawbacks caused by the low false alarm rate and noise distribution shape parameter. Last Section 3.4 talks about the M out of N rule which can be applied after the first CFAR threshold to robustly reduce the false-alarm rate to much lower value.

3.2 False alarm rate

Before stepping into the thresholding methods, an important term should be introduced: false alarm rate. For this application, a false alarm occurs when a noise peak is falsely detected as a true sound source. Figure 3.1 shows example probability distributions of expected noise peaks and true source peaks, the false alarms probability in this case for a given threshold is determined by the area indicated in green on the right side of threshold. The false alarm rate is the false alarm probability.

CFAR detectors will use the false alarm rate given by the user to determine the threshold based on the noise peak distribution model. Therefore, the accuracy of false alarm rate for a given threshold depends on how well the distribution models the actual noise distribution.

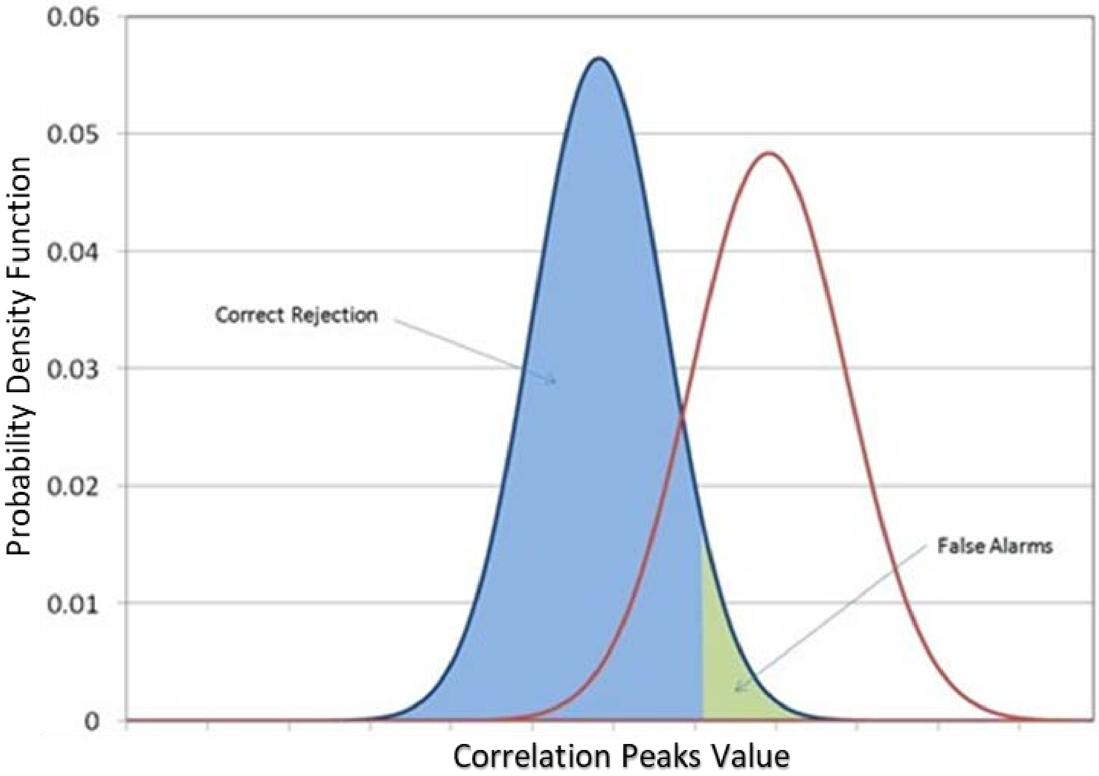


Figure 3.1: Blue line denotes the noise distribution probability density function (PDF) and red line is the PDF of target distribution.

From Figure 3.1, it is obvious that as the false alarm rate lowers, the threshold is closer to the tail of the curve. In this case, the shape parameter differences will have direct impact on the tail shape of the curve and the threshold. As the shape parameter is unknown in most conditions, it’s value will not be accurate [26]. Therefore, the empirical false alarm will be directly affected by mismatch in the assumed and actual shape parameter.

3.3 CFAR

For the algorithm stated above, the accuracy of significant peak detection is critical to performance. The TDOA-based algorithm determines the sound source locations using the different delay times from the sound signals arriving at different microphones. The delay times are obtained from the significant peaks of cross-correlation of different microphones. If there is only one target source, the detection algorithm will be simply selecting a maximum peak. However, with two or more targets, the algorithm should be able to determine which peaks are significant (i.e. likely targets) among all peaks. Similar problem happened in the SRP algorithm. The SRP algorithm will output a power image based on the SRP estimated. The targets are located at the significant peaks in the power image. For these two cases, it's not effective using a constant threshold, because the variations in signal and noise powers impact the detectors. Therefore, a self-adjusted constant false-alarm rate (CFAR) threshold algorithm can be applied to deal with this problem.

CFAR has been used so that multiple targets can be detected based on a specified false alarm rate. A CFAR threshold is estimated based on a probabilistic model of the noise-only distribution, and this threshold comes from the local data to maintain a fixed false alarm rate adapted to changing noise levels. In other words, a CFAR threshold is an adaptive threshold that follows changes of local noise power. The CFAR approaches are commonly used in the radar system and others where large amount of noise-only data is available [27, 28] (i.e. target samples are sparse in time/space relive to the noise samples).

The purpose of CFAR detector is to detect the weakest true target signal, while maintaining an expected FA rate. This means finding the lowest threshold that results in the desired/tolerable FA rate. Since the noise-only distribution can change, especially due to variations in power, and since the occurrence of a target in space and time is much rarer than the occurrence of noise, local sections of the cross-correlation peaks can be used to updated the probability model and compute thresholds using data local to the peak under test. In the TDOA-based algorithm, the CFAR detector is applied on the cross-correlation results. For each peak tested, a guard band is applied to ensure the energy from the true target does not dilate into the surrounding neighborhood used to estimate the noise peak distribution power. Figure 3.2 shows an example of a guard band. It ignores the small part near the test peak. This requirement affects the left cutoff of the guard band. Another assumption should be made is that the noise parts chosen by guard band have a symmetric distribution. This assumption influences the right cutoff of the guard band

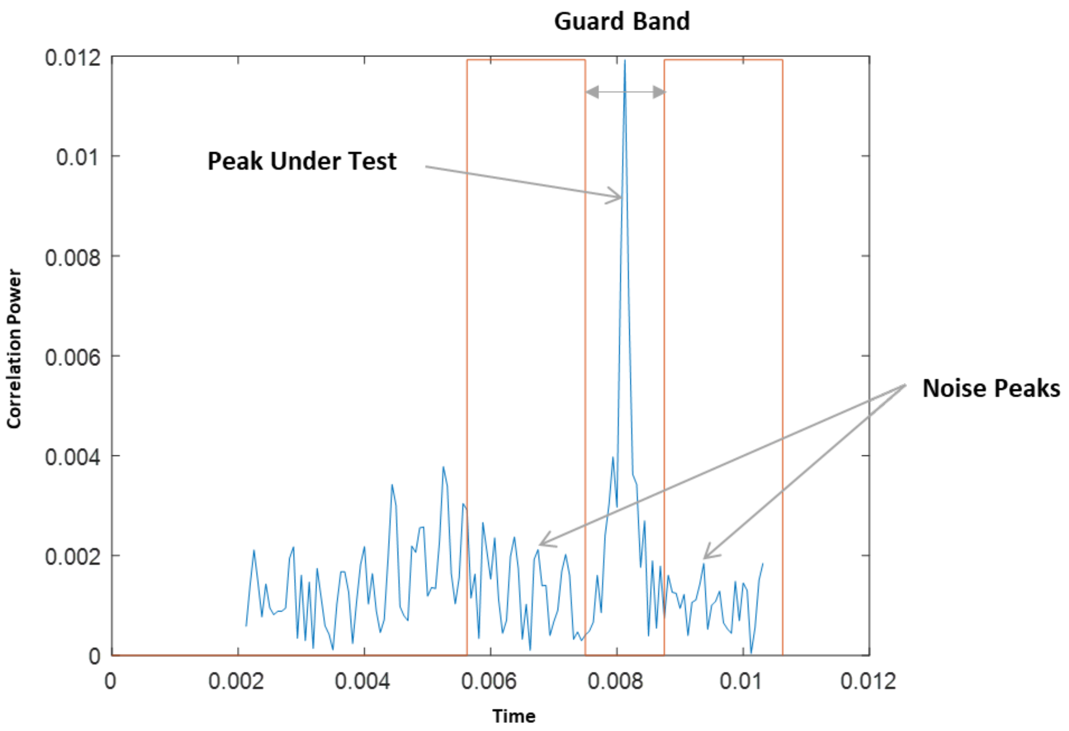


Figure 3.2 An example for a guard band applied on a part of signal. Guard band is shown in orange line and signal is blue line.

Apart from choosing a reasonable part using guard band, it's also important to choose a suitable noise distribution since the threshold is estimated based on a parametric estimate of an assumed distribution. One useful distribution is the Weibull distribution [26]. The false alarm rate and it corresponding threshold can be related through the Weibull distribution to result in:

$$P_{fa} = \exp\left(\frac{T}{a}\right)^{b} \quad , when\ T \geq 0. \qquad (3.1)$$

where a and b are the scale and shape parameters. A critical advantage of the Weibull distribution is that with different shape parameters, the distribution can be parametrically adjusted between the exponential distribution and the Rayleigh distribution. Assuming a given shape parameter b, the scale parameter is computed from the peaks value obtained by guard band near the testing peak:

$$a = \left(\frac{1}{\|N\|}\sum_{i=0}^{N}|H_i|^b\right)^{\frac{1}{b}} \qquad (3.2)$$

where H_i is the peak height of the i^{th} peak in guard band. N is the total number of peaks in this area. This band area avoids the impact of large value close to a targeting peak on the mean calculation. For a false alarm (FA) rate provided by user, the threshold is calculated from the inverse of Equation 3.1:

$$T = a[-ln(P_{fa})]^{1/b} \qquad (3.3)$$

where P_{fa} is the desired FA probability. The threshold established above is used to decide whether a peak is significant. The user can control the detection results using different FA rates, where lower rates lead to less detection sensitivity, but less peaks to process for each scan.

The curves in Figure 3.3 where generates from simulations of noise detections different shape parameters, b, from 1.0 to 2.0. The x-axis is the desired false alarm rate set by the CFAR detectors using equation 3.3, and y-axis is the empirical detected false alarm rate from the simulator over the set false alarm rate. It's obvious that the detection accuracy becomes less with lower set false alarm rates. The number of false alarms detected is around 10^2 times difference at a 10^{-4} false alarm rate for 2 of the shape parameters, while it's only 10 times at a 10^{-1} false alarm rate for all shape parameters. Therefore, it shows that shape parameter inaccuracy will be a problem for CFAR algorithm at lower false alarm conditions, and, for this reason, the MON filter is applied to make a stable double-threshold detector.

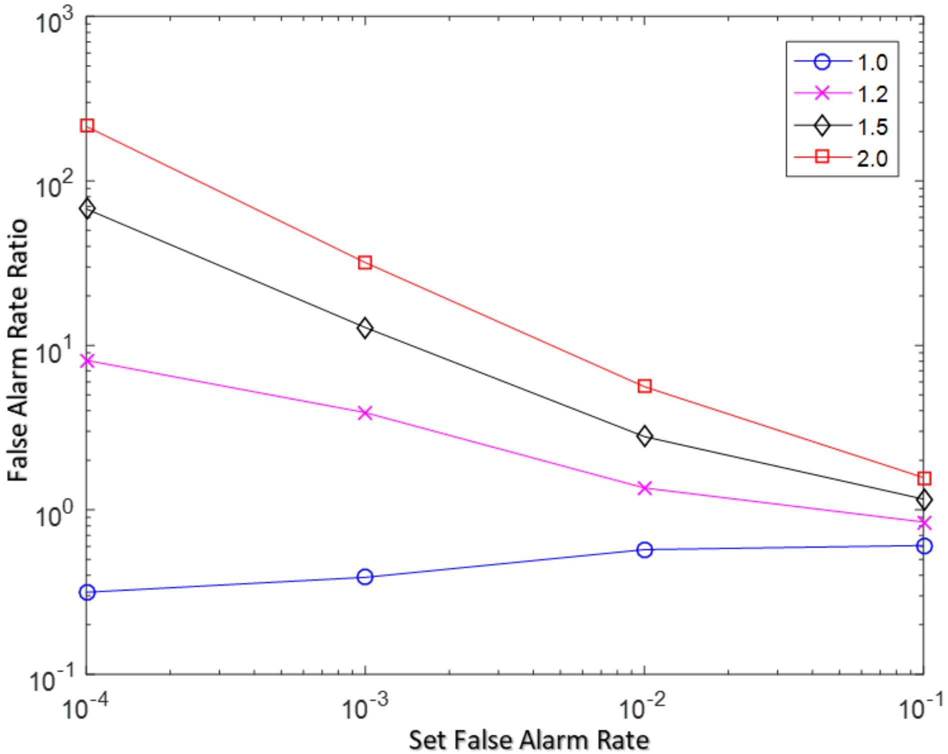


Figure 3.3 This figure shows empirical false alarm vs true false alarm applied with different shape parameters feeding into CFAR filter from 1.0 to 2.0.

3.4 M Out of N (MON) Rule Statistical Filtering

M out of N detection was commonly used in radar system in 1950s and 60s [29]. Under the assumption of noise only case, let p_N represent the probability of a peak value crossing the threshold, n represent the number of cross-correlation pairs and m represent the number of peaks associated with a given location crossing the threshold. This rule is modeled by the binomial distribution. So, the equation for relating the M out of N rule to a resulting FA rate is given by [29]:

$$P_{fa} = \sum_{i=m}^{n} \binom{n}{i} p_N^i (1 - p_N)^{n-i} \qquad (3.4)$$

where the $\binom{n}{i}$ part stands for the binomial coefficient, which means the number of combinations of n things taken i at a time. To use this in setting a desired FA rate, a low FA rate on the cross-correlation peaks is selected first at a relatively high value to ensure a consistent FA rate robust to mismatches in the distribution model. Then the m value can be selected the lower the FA rate to a more reasonable value. Suppose that we have $p_N = 10^{-1}$ (the FA rate used in equation 3) and $n = 28$. The m value ranges from 1 to 28, integer values only. The P_{fa} that closest to the 10^{-4} (desired false alarm rate) will be chosen and then use the look up table based on equation 3.4 to find out the m value. The m value obtained here denotes the minimum number of detections for a given false alarm rate, in this case 10^{-4}.

As we mentioned above, an appropriate false alarm rate (or threshold) will be able to detect most targets with few or no false detections. However, the CFAR detectors become less accurate with lower false alarm rates, so a MON filter will be applied as a secondary threshold. The CFAR-MON filter will not suffer from the shape parameter differences, as a high enough false alarm rate (10^{-1} or 10^{-2}) is applied for the CFAR part and it shows a stable detection for different shape parameter in Figure 3.3. For the MON part, it’s only associated with microphone pairs and data size, so the results should be much more stable than CFAR filter under lower false alarm rate cases. The primary source of error in this test will be the assumed p_N which is the FA rate on the first threshold test.

Chapter 4 Simulation

4.1 Introduction

This chapter demonstrates the performance of TDOA-DTS algorithm compared to SRP algorithm and the stable detection performance of CFAR-MON double thresholding process. The Section 4.2 introduces the room setup for simulations. With the same room set up, it makes sure the room size, sound speed and noise power level are stay the same. Then the target sound sources, noise sources and their positions are same for a comparative study. The reflection rate of the wall and the number of noise sources are variable. Section 4.3 compares SRP and TDOA-DTS results under same setups. Section 4.4 shows the CFAR-MON dual-threshold performance and Section 4.5 presents the computational load comparison between TDOA-DTS and SRP. Last section, Section 4.6, is experiment verification which uses the real data collected from lab.

4.2 Room setup

To test the algorithm and verify computational and CFAR performance, a simulation room has been set up. The room has a size of $7\times8\times3.5m$. Coordinates of two opposite corner points for this room are $[-3.5, -4, 0]$ and $[3.5, 4, 3.5]$. As the Figure 4.1 shows a resulting SRP image plane with eight microphones placed around the area of interest. The color-bar denotes the SRP values. Two black circles are the position of two sound sources. The signal model includes reverberation as presented in chapter 1. The reverberation is simulated using the image method [23, 20]. To limit the computational load of the simulation, multiple room reflections were computed until the reflected energy dropped 60 dB from the first reflection. In some simulation, one or two coherent noise sources will be put outside the area of interest. The power of noise sources will be controlled by a given signal-noise-ratio (SNR). Typically, it is set to -30. The system noise is assumed as white noise and its SNR is set to -30 either.

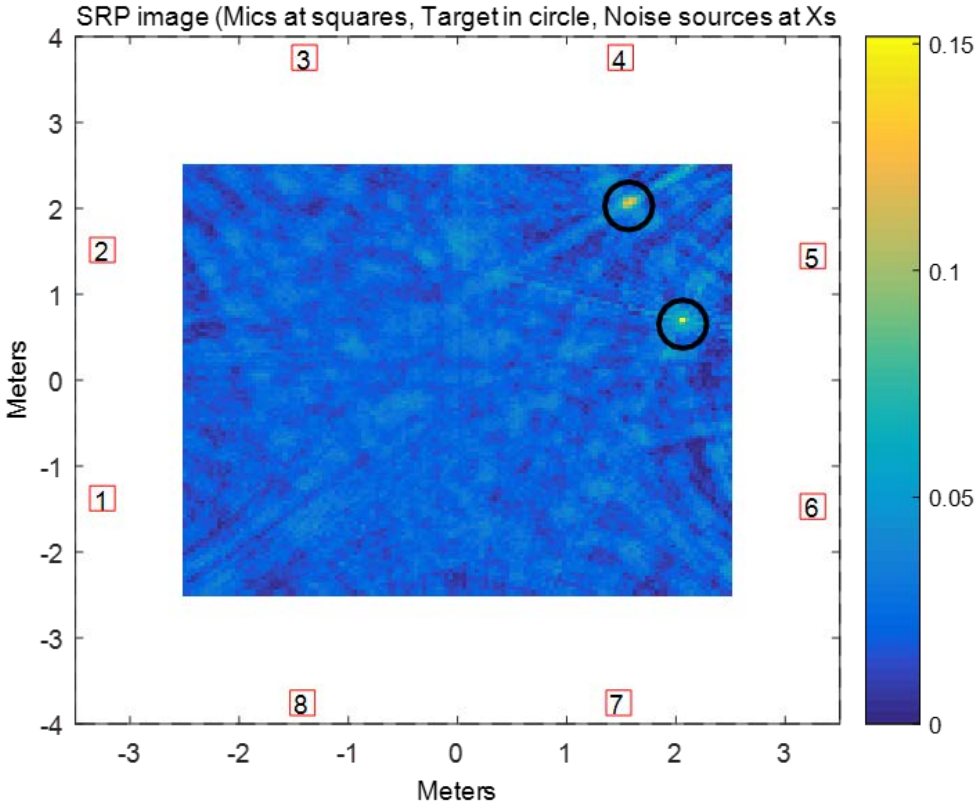


Figure 4.1 One example of the simulation room. Reflection coefficients for four walls are 0.5 and 0.4 for floor and ceiling. System noise SNR is -30dB. No noise target presented.

Table 4.1 This table shows the test parameters and which detection rule is using.

Figure Number	Reflection Coefficients (walls, floor and ceiling)	Number of Noise Sources	Detection Rule
4.2	0.5, 0.4, 0.4	1	(a) SRP (b) TDOA-DTS
4.3	0.5, 0.4, 0.4	2	(a) SRP (b) TDOA-DTS
4.4	0.7, 0.5, 0.5	1	(a) SRP (b) TDOA-DTS
4.5	0.7, 0.5, 0.5	2	(a) SRP (b) TDOA-DTS
4.6	0.9, 0.7, 0.7	1	(a) SRP (b) TDOA-DTS
4.7	0.9, 0.7, 0.7	2	(a) SRP (b) TDOA-DTS

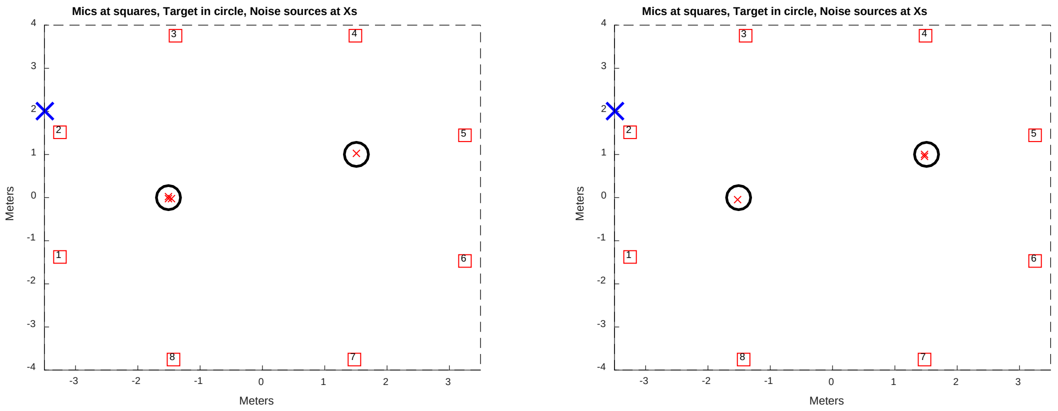
Other parameters: CFAR false alarm rate is 10^{-1}, MON rule false alarm rate is 10^{-1}, coherent noise sources SNR -30, signal positions are $(-1.5, 0, 1.5)$ and $(1.5, 1, 1.5)$, coherent noise sources for one is $(-3.5, 2, 1.5)$, for two noise sources are $(-3.5, 2, 1.5)$ and $(2, -4, 1.5)$.

4.3 SRP vs TDOA-DTS

The simulation is using the room setup described above, but with different number of noise sources and different reflection rate. Noise sources are simulating the noise coming from windows, vents and other sources outside the FOI. Therefore, they are put outside the FOI and won’t scanned by the SRP system. The simulation is using the parameters shown in Table 4.1. SRP algorithm uses CFAR threshold with Weibull distribution shape parameter equaling to 1.26 [26]. Because the detection is operated on a 2D image, the guard band is all segments between 2 segments away from tested segment to 5 segments. While the TDOA-DTS algorithm uses CFAR-MON threshold with shape parameter equaling to 2.5 (Figure 4.8 in Section 4.4). The guard band is 5 to 55 sampling points away from the tested peak. The two sound sources are human voices and the noise sources are randomly generated white noise.

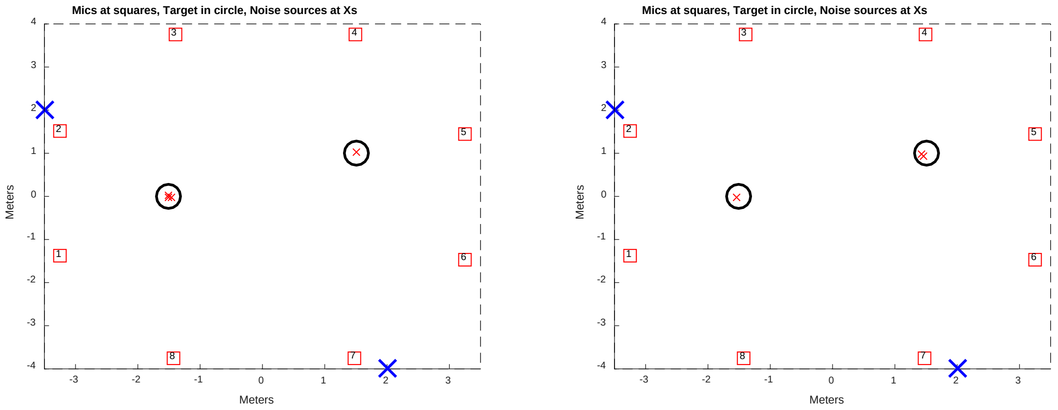
In the simulations, the false alarm rate is set to an extremely low value, which will make no targets detected. Then the false alarm rate will be raised until two sound sources are detected. By comparing the number of false detections, the performance of both algorithms can be evaluated fairly.

The results show that the SRP and TDOA-DTS have similar performance under the low and medium reverberation conditions. However, both algorithms detect much more false detections and make the actual target indistinguishable. Therefore, the TDOA-DTS algorithm shows a similar performance comparing to the SRP algorithm.



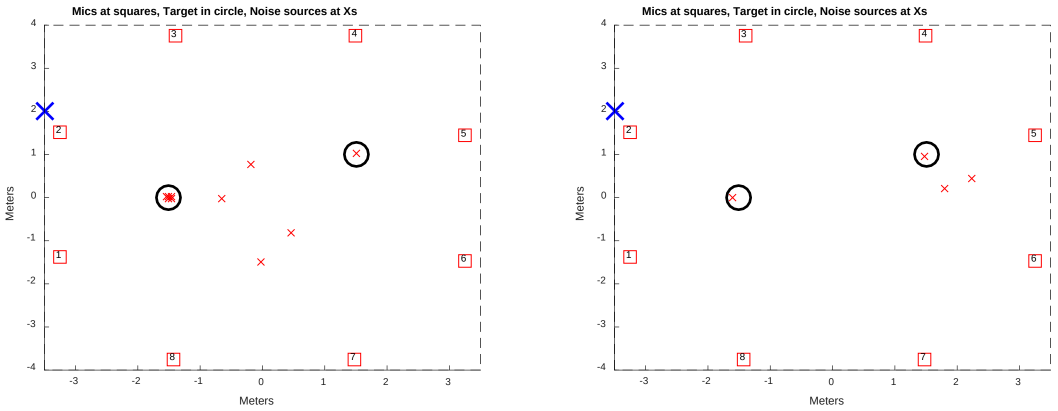
(a) (b)

Figure 4.2 Two simulations with wall, floor and ceiling reflection coefficients set as 0.5, 0.4 and 0.4. Actual positions of two targets are represented by black circles. One noise source at blue X. SRP: No false detection. TDOA-DTS: No false detection



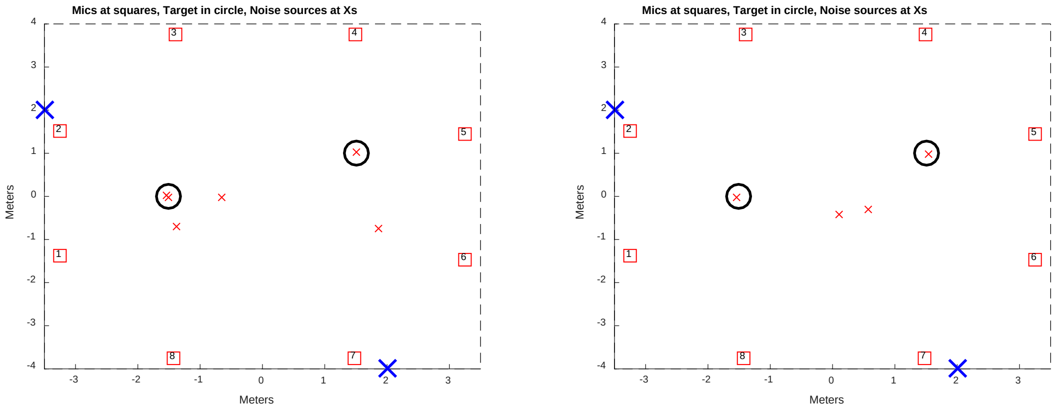
(a) (b)

Figure 4.3 Two simulations with wall, floor and ceiling reflection coefficients set as 0.5, 0.4 and 0.4. Actual positions of two targets are represented by black circles. Two noise sources at blue Xs. SRP: No false detection. TDOA-DTS: No false detection



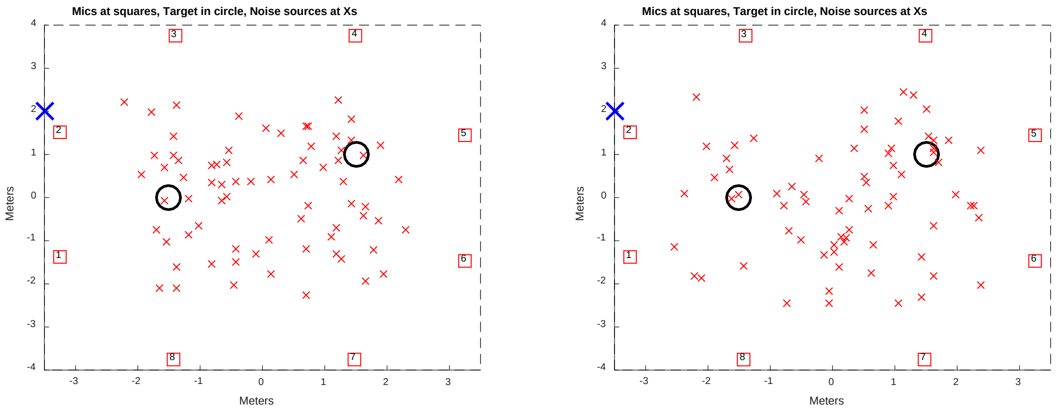
(a) (b)

Figure 4.4 Two simulations with wall, floor and ceiling reflection coefficients set as 0.7, 0.5 and 0.5. Actual positions of two targets are represented by black circles. One noise source at blue X. SRP: 4 false detections. TDOA-DTS: 2 false detections



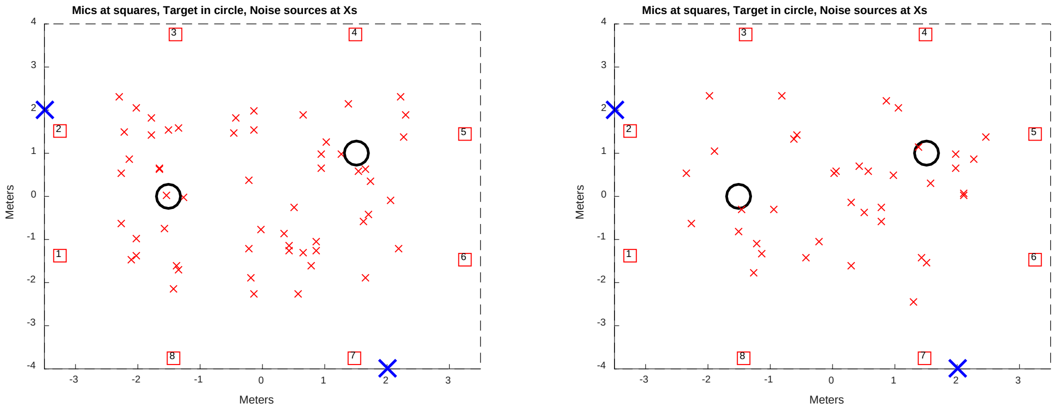
(a) (b)

Figure 4.5 Two simulations with wall, floor and ceiling reflection coefficients set as 0.7, 0.5 and 0.5. Actual positions of two targets are represented by black circles. Two noise sources at blue Xs. SRP: 3 false detections. TDOA-DTS: 2 false detections



(a) (b)

Figure 4.6 Two simulations with wall, floor and ceiling reflection coefficients set as 0.9, 0.7 and 0.7. Actual positions of two targets are represented by black circles. One noise source at blue X. SRP: 68 false detection. TDOA-DTS: 64 false detection



(a) (b)

Figure 4.7 Two simulations with wall, floor and ceiling reflection coefficients set as 0.9, 0.7 and 0.7. Actual positions of two targets are represented by black circles. Two noise sources at blue Xs. SRP: 65 false detection. TDOA-DTS: 49 false detection

4.4 CFAR and MON threshold

As mentioned in chapter 3, CFAR and MON are applied to detect significant peaks. To achieve a better filter, some simulations are performed on these filters. The CFAR filter assumes noise distribution model to estimated threshold. Therefore, the first simulation is running to find a best fit distribution model.

The simulation room is the same used above. To set up a noise only condition, the sound sources are put outside the field of interest (and region in the cross-correlation were excluded from the test). Weibull distribution is a good choice, because it can be interpolated from exponential distribution to Rayleigh distribution by applying different shape parameter.

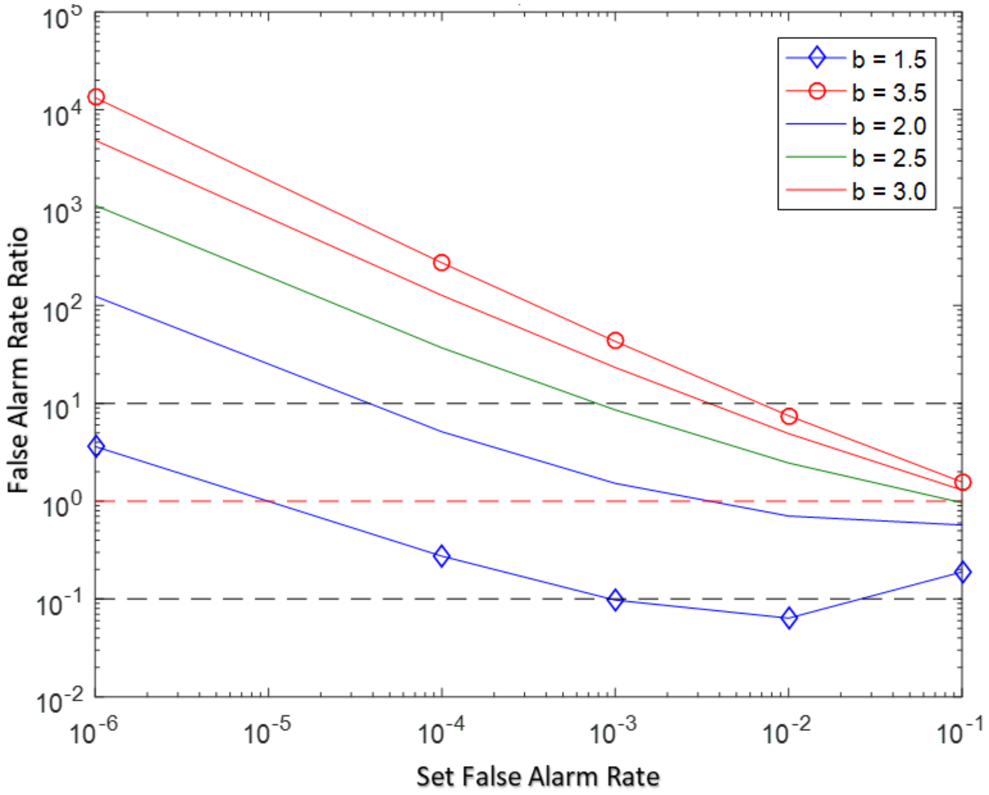


Figure 4.8 The results of CFAR filter with shape parameter "b" from 1.5 to 3.5 and step is 0.5. Total four different false alarm rates are tested. 5000 different test run with more than 20 million peaks checked.

The simulation is running with eight microphones and two sound sources outside the field of interest. The reflection coefficients are [0.5, 0.5 ,0.5 ,0.5 ,0.4 ,0.4] for four walls, floor and ceiling. The x axis represents the test false alarm rates and the y axis denotes the detected false alarm rate over test false alarm rate ratio, which means the 10^{0} is the ideal results. From the plot, 2.0 and 2.5 would be better choice for the shape parameter. Furthermore, some extra information can be concluded. The CFAR detector is more stable at higher false alarm rates (10^{-1} and 10^{-2}) conditions, so lower false alarm rate will make the detector much more sensitive to the distribution mismatch. The problem of high false alarm rate is the huge amount of detection peaks. It makes the direct use of CFAR results very difficult. Therefore, MON detector will do the second thresholding to achieve the final detection results.

The results of CFAR-MON double threshold filter with the same set up as Figure 4.8. As shown in Figure 4.9, it's obvious that the established false alarm rates are stable when different false alarm rates applied. The CFAR-MON filter will get more stable results under lower false alarm rate (lower than 10^{-4}) comparing to CFAR filter. When the shape parameter equals to 1.5, the CFAR detections are much lower than the expected presented in Figure 4.6. Therefore, after applied the MON detection algorithm, there is no enough data for false alarm rate lower than 10^{-3}.

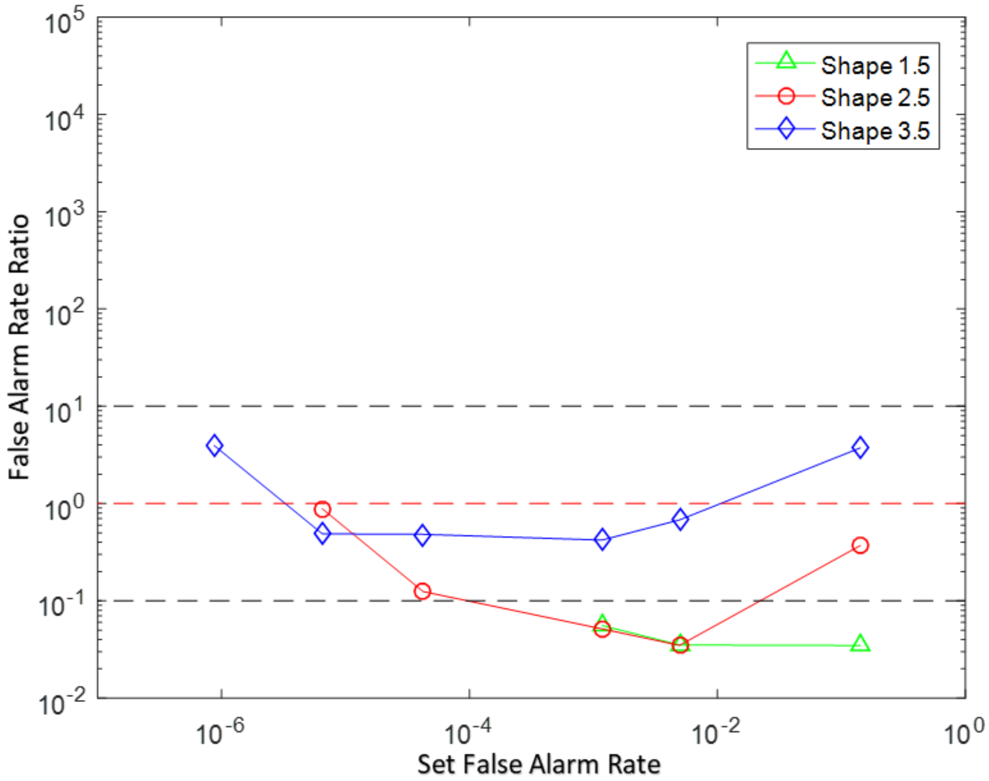


Figure 4.9 The results of CFAR-MON filter with CFAR false alarm rate 0.1 and MON false alarm from 10^{-1} to 10^{-6}. The whiten parameter is 1 in all the simulations. Three different shape parameters are applied.

4.5 Time complexity comparison between SRP and TDOA-DTS algorithms

The purpose of this thesis is using TDOA algorithm to speed up the sound source localization procedure. This part will focus on the time complexity of these two different algorithms. Assume the number of microphones is N and length the sampled microphone time segment is X. The field of interest is divided into small segments and total number of spatial points is represented as S. As the Section 2.3 presented, SRP algorithm will first scan through all S grid points in the field of interest. For a single run of one segment, N signals with length X will be shifted and added together to compute the power. Therefore, the SRP will have total $S * N * X$ real multiplications. The TDOA-DTS algorithm will test cross-correlation results for all possible microphone pairs, so the number of microphone pairs is $\frac{N(N-1)}{2}$. For each pair of X length microphone signals, cross-correlation is computed, and the complexity of TDOA-DTS will be $\frac{N(N-1)}{2} * X^2$. The ratio that TDOA-DTS over SRP is $\frac{(N-1)X}{2S}$. If the field of interest (FOI) is huge enough and makes that $S \gg NX$, the TDOA-DTS will have much more fast speed than the SRP.

In the simulation, the field of interest is $5 \times 5 \times 1.5m$ with $0.04m$ steps, so $S = 126\times126\times76 \approx 1.2\times10^6$. Assuming the $N = 8$ and $X = 6000$, the expected value should be:

$$\frac{(N-1)X}{2S} = \frac{7\times6000}{2\times1.2\times10^6} = 0.0175 \qquad (4.1)$$

Another case, if the simulation just runs in two dimensional $5 \times 5m$ with $0.04m$ steps then $S = 126\times126 = 15876$. Keep other parameters the same, the ratio will be:

$$\frac{(N-1)X}{2S} = \frac{7\times6000}{2\times15876} \approx 1.3228 \qquad (4.2)$$

With same value plugged into the simulations, 500 different runs have been tested. The time is obtained by using tic and toc function in MATLAB. Table 4.2 shows the results. The third column, simulation average, is the average value of TDOA-DTS running time over SRP running time, and fourth column is the standard deviation of all results. It’s clear that the equation mentioned above can predict the results within an order of magnitude.

Table 4.2 Simulation results and equation estimated results comparison with two different number of pixels. The microphone number $N = 8$, and signal length $X = 6000$

Number of Pixels (S)	Estimated Value ($\frac{(N-1)X}{2S}$)	Simulation Average	Simulation Standard Deviation
1.2×10^6	0.0175	0.0383	0.0015
15876	1.3228	1.1331	0.0377

4.6 Simulation verification

Based on the simulations shown above, the CFAR-DTS algorithm performs as expected. So, experiments are set up in the lab and the actual data is being recorded to verify the simulation results. The real data is collected in the Distributed Audio Lab located in Davis Marksbury Building. The lab environment is shown in the Figure 4.10.

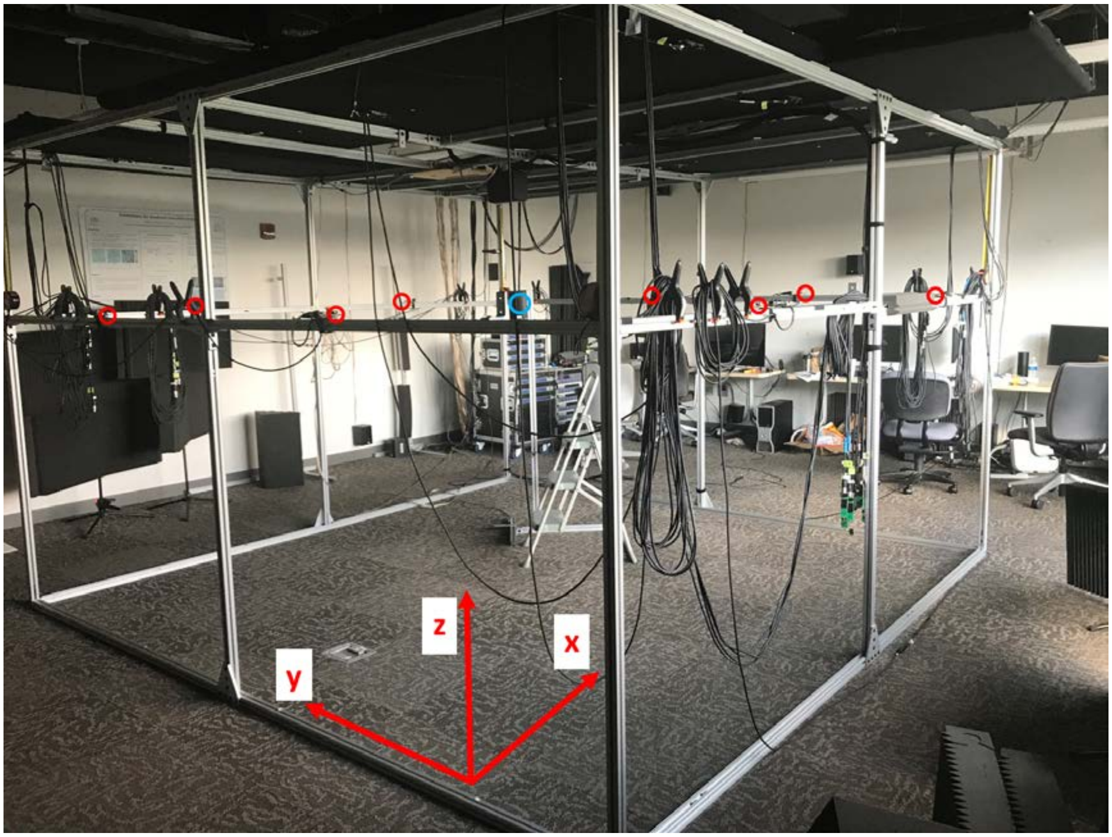


Figure 4.10 The lab environment where the experiment data collected. Red circles are microphones and blue circle is the speaker (sound source).

The frame in the center of the room is the experiment area and its size is 3.659m × 3.659m × 2.29m (including the frame width). The eight microphones are settled on the “wall” of the frame distributed with 1.5m high from the ground. Speaker is also placed at the same height as the microphones. The top-down view of the microphones, speaker and the frame is shown in Figure 4.11.

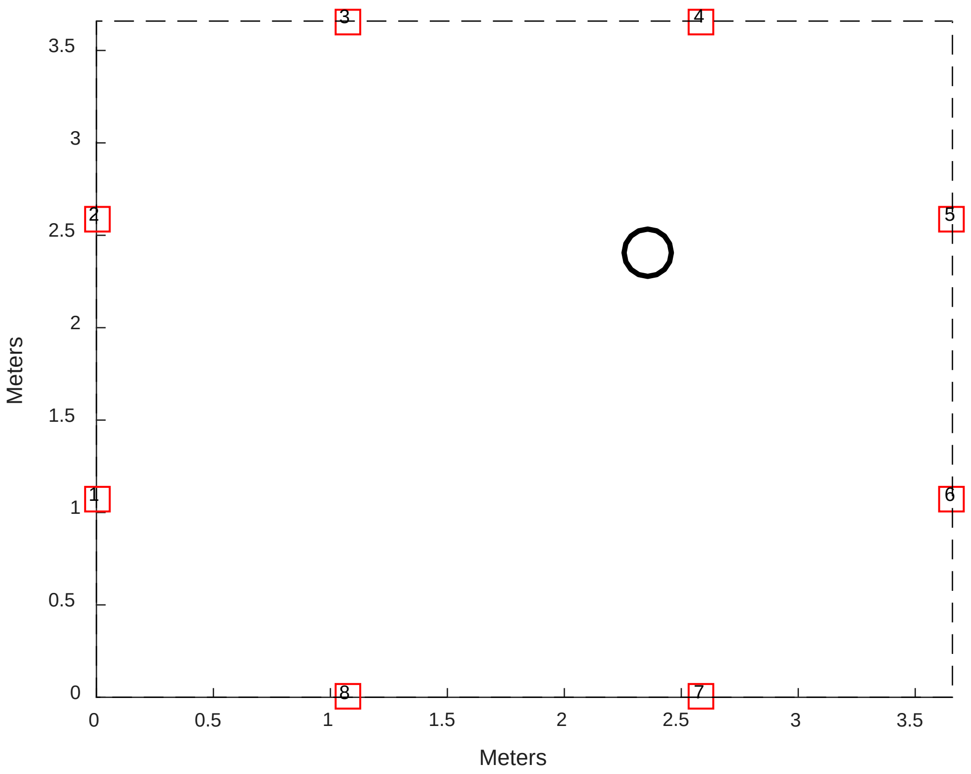


Figure 4.11 Microphones are represented by red squares, sound source is placed at the circle.

A two second white noise signal is used as the sound source. To comparing the experiment and simulation results, the same setup is used in this simulation. Additionally, the reflection parameters of the simulation are 0.7 for the wall and 0.5 for the ceiling and floor and the SNR is -10 dB. The noise distribution uses Weibull distribution with shape parameter 1.5. The experiment and simulation test one second received signals with 50ms window nonoverlap. The detections that within 0.12m (three times grid pix width) from sound source will be considered as right detection, other detections will be counted as false detections. Figure 4.12 shows the experiment and simulation results. The difference between these two results are less than the 95% confidence limits, so the simulation follows the change of real data experiment with acceptable differences. The high error shown under low false alarm rate conditions is caused by not having enough test peaks. The simulation has more false alarms than experiment in Figure 4.12. The reason is that the target in simulation have same power for all directions, while the experiment is using a speaker focus a fixed direction. Therefore, as shown in Figure 4.13, the simulation has a larger size of target detection, which leads to more false alarms.

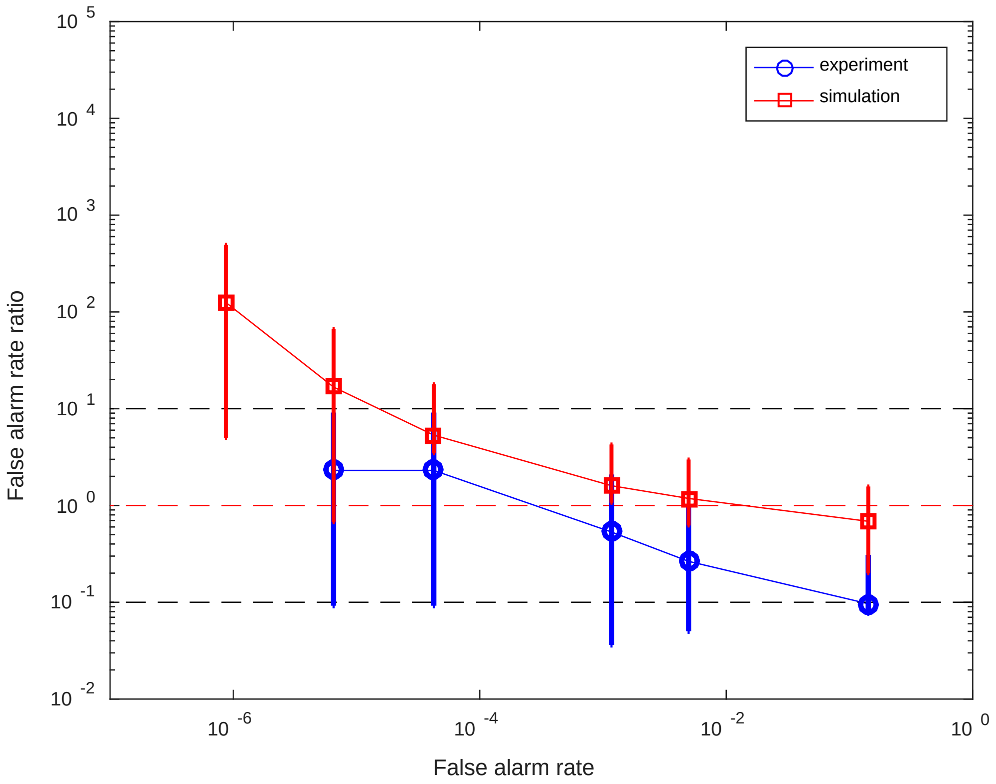
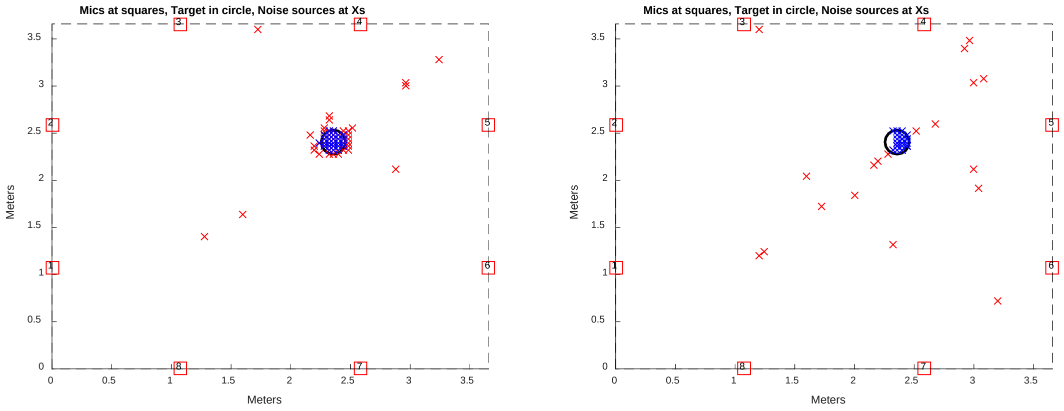


Figure 4.12 The vertical lines are error bar with 95% confidence. x axis is the ideal false alarm rate, and y axis is the ratio that experiment and simulation results over ideal value. Closer to red dotted line means more accurate and black dotted line shows the area within tolerable difference.



(a)

(b)

Figure 4.13 Detection results. The blue Xs denote the right detections and red Xs are false detections. (a) Simulation results, (b) Experiment results.

Chapter 5 Conclusion

Conclusion

This thesis presents two classic solutions to sound source localization and proposes a new algorithm called TDOA-DTS algorithm. The new algorithm uses a Delay Search Table replacing the original hyperbola drawing method. It has similar detection performance and about 26 times faster speed comparing to SRP under the same simulation cases. A double thresholding process, named CFAR-MON, is also introduced to solve the problem caused by CFAR and maximum peaks finding detection. The CFAR-MON proves to be a more stable detection method than the CFAR under different false alarm rate. The simulation results are verified by the experiment with real data recorded in the lab.

Bibliography

[1] Mohammadiha, Nasser, Paris Smaragdis, and Arne Leijon. "Supervised and unsupervised speech enhancement using nonnegative matrix factorization." *IEEE Transactions on Audio, Speech, and Language Processing* 21.10 (2013): 2140-2151.

[2] Valin, Jean-Marc, François Michaud, and Jean Rouat. "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering." *Robotics and Autonomous Systems* 55.3 (2007): 216-228.

[3] Valin, Jean-Marc, et al. "Robust recognition of simultaneous speech by a mobile robot." *IEEE Transactions on Robotics* 23.4 (2007): 742-752.

[4] Kim, Ui-Hyun, Kazuhiro Nakadai, and Hiroshi G. Okuno. "Improved sound source localization in horizontal plane for binaural robot audition." *Applied Intelligence* 42.1 (2015): 63-74.

[5] Nakadai, Kazuhiro, et al. "Design and Implementation of Robot Audition System'HARK'—Open Source Software for Listening to Three Simultaneous Speakers." *Advanced Robotics* 24.5-6 (2010): 739-761.

[6] Nakadai, Kazuhiro, Hiroshi G. Okuno, and Hiroaki Kitano. "Real-time sound source localization and separation for robot audition." *INTERSPEECH*. 2002.

[7] Gur, M. Berke, and Christopher Niezrecki. "A source separation approach to enhancing marine mammal vocalizations." *The Journal of the Acoustical Society of America* 126.6 (2009): 3062-3070.

[8] Andrea, Douglas, and John Kowalski. "Adaptive noise cancellation and speech enhancement system and apparatus therefor." *U.S. Patent* No. 5,251,263. 5 Oct. 1993.

[9] Boll, Steven, and D. Pulsipher. "Suppression of acoustic noise in speech using two microphone adaptive noise cancellation." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.6 (1980): 752-753.

[10] DiBiase, Joseph H., Harvey F. Silverman, and Michael S. Brandstein. "Robust localization in reverberant rooms*." Microphone Arrays. Springer Berlin Heidelberg*, 2001. 157-180.

[11] Donohue, Kevin D., and Paul M. Griffioen. "Computational strategy for accelerating robust sound source detection in dynamic scenes." *SOUTHEASTCON 2014, IEEE*. IEEE, 2014.

[12] Do, Hoang, and Harvey F. Silverman. "Robust cross-correlation-based techniques for detecting and locating simultaneous, multiple sound sources." *Proc. IEEE ICASSP*. 2012.

[13] Carter, G. "Bias in magnitude-squared coherence estimation due to misalignment." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.1 (1980): 97-99.

[14] Firoozabadi, Ali Dehghan, and Hamid Reza Abutalebi. "Localization of multiple simultaneous speakers by combining the information from different subbands." *Electrical Engineering (ICEE), 2013 21st Iranian Conference on*. IEEE, 2013.

[15] Jamali-Rad, Hadi, and Geert Leus. "Sparsity-aware multi-source TDOA localization." *IEEE Transactions on Signal Processing* 61.19 (2013): 4874-4887.

[16] Yook, Dongsuk, Taewoo Lee, and Youngkyu Cho. "Fast sound source localization using two-level search space clustering*." IEEE transactions on cybernetics* 46.1 (2016): 20-26.

[17] Donohue, Kevin D., Sayed M. SaghaianNejadEsfahani, and Jingjing Yu. "Constant false alarm rate sound source detection with distributed microphones." *EURASIP Journal on Advances in Signal Processing* 2011.1 (2011): 1-12.

[18] Ma, Wing-Kin, et al. "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach." *IEEE Transactions on Signal Processing* 54.9 (2006): 3291-3304.

[19] Donohue, Kevin D., Jens Hannemann, and Henry G. Dietz. "Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments." *Signal Processing* 87.7 (2007): 1677-1691.

[20] Allen, Jont B., and David A. Berkley. "Image method for efficiently simulating small-room acoustics." *The Journal of the Acoustical Society of America* 65.4 (1979): 943-950.

[21] Ramamurthy, Anand, Harikrishnan Unnikrishnan, and Kevin D. Donohue. "Experimental performance analysis of sound source detection with SRP PHAT-β." *Southeastcon, 2009. SOUTHEASTCON'09. IEEE*. IEEE, 2009.

[22] Han, Guangjie, et al. "Localization algorithms of underwater wireless sensor networks: A survey." *Sensors* 12.2 (2012): 2026-2061.

[23] Zhang, Li, and XiaoDong Yu. "A Kernel-based TDOA localization algorithm." *Computer Application and System Modeling (ICCASM)*, 2010 International Conference on. Vol. 11. IEEE, 2010.

[24] Ramamurthy, Anand, Harikrishnan Unnikrishnan, and Kevin D. Donohue. "Experimental performance analysis of sound source detection with SRP PHAT-β." *IEEE Southeastcon 2009*. IEEE, 2009.

[25] DiBiase, Joseph H., Harvey F. Silverman, and Michael S. Brandstein. "Robust localization in reverberant rooms." *Microphone Arrays*. Springer Berlin Heidelberg, 2001. 157-180.

[26] Saghaian Nejad Esfahani, Sayed Mahdi. "STATISTICAL MODELS FOR CONSTANT FALSE-ALARM RATE THRESHOLD ESTIMATION IN SOUND SOURCE DETECTION SYSTEMS." (2010).

[27] H. Rohling, “Radar CFAR thresholding in clutter and multiple target situations,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 19, no. 4, pp. 608–621, 1983.

[28] K. D. Donohue and N. M. Bilgutay, “OS characterization for local CFAR detection,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 21, no. 5, pp. 1212–1216, 1991.

[29] Norouzi, Yaser, Maria S. Greco, and Mohammad M. Nayebi. "Performance evaluation of K out of n detector." *Signal Processing Conference, 2006 14th European*. IEEE, 2006.

Vita

Xipeng Wang was born in Taihu, China. He received his B.Sc degree in Electrical Engineering in 2015 from both University of Kentucky and China University of Mining and Technology, Xuzhou, China. He was put into Dean's List in 2013 Fall, 2014 Spring and 2014 Fall for outstanding academic performance.