Multi-Sense Artificial Awareness

by

Parham Aarabi

A thesis submitted in conformity with the requirements for the Degree
of Master of Applied Science
Graduate Department of Electrical and Computer Engineering
University of Toronto

© Copyright by Parham Aarabi 1999
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.
Multi-Sense Artificial Awareness

Parham Aarabi

Master of Applied Science, 1999

Graduate Department of Electrical and Computer Engineering

University of Toronto

Abstract

In the process of exploring multi-sense artificial awareness systems, a multi-camera vision algorithm and an iterative spatial probability based sound localization system were developed. The multi-camera vision algorithm uses spatial probability maps to combine the results of the individual cameras in a simple and efficient manner.

The sound localization system uses an iterative technique that increases the accuracy of the localization while maintaining a high localization rate. It can correctly localize objects in noisy environments with a signal to noise ratio as low as 0 dB. The sound localization system also uses spatial probability maps for the integration of the results of multiple microphones.

Spatial probability maps also serve as the basis for the integration of vision and sound localization. An integrated vision and sound localization (IVSL) system was implemented and it was discovered that the integration of the two senses results in increased accuracy and robustness.
Two years ago, I set upon a journey to explore artificial awareness. During this time, I have benefited from the help and advice of many people. However, any success that may have been achieved in this time can be almost entirely attributed to the guidance of my supervisor Professor Safwat Zaky. I am also grateful for my family’s encouragement and support, especially in the sound localization testing process. During this journey, I have also benefited from the advice of Professor Zukotynski, whose support in the past four years has been and continues to shape my future. I would also like to thank the current members of the GB253 lab who have provided me with some of the most memorable and interesting moments of my life, and for enabling me to develop my Squash skills.
Table of Contents

CHAPTER 1 INTRODUCTION ........................................................................ 1
  1.1 Motivation .................................................................................................. 1
  1.2 Previous Artificial Awareness Projects ................................................... 2
  1.3 Multi-Sense Artificial Awareness Arrays ............................................... 3
  1.4 Thesis Contribution .................................................................................. 4
  1.5 Thesis Outline .......................................................................................... 4

CHAPTER 2 MULTIPLE-CAMERA OBJECT LOCALIZATION ..................... 6
  2.1 Introduction .............................................................................................. 6
  2.2 Previous Vision Systems ........................................................................ 7
  2.3 The Proposed Multiple-Camera Vision Algorithm ................................. 8
    2.3.1 Preliminary Image Decomposition and Spatial Probability Maps ..... 8
    2.3.2 Formation of Spatial Probability Maps .......................................... 9
    2.3.3 Application of Spatial Probability Maps to Multi-Camera Vision ... 12
  2.4 Intelligent False Object Removal ........................................................... 13
    2.4.1 The IFOR Algorithm ....................................................................... 14
    2.4.2 Analysis of the IFOR algorithm ...................................................... 17
    2.4.3 Complexity of IFOR ....................................................................... 18
    2.4.4 Global Peak Selection ..................................................................... 20
    2.4.5 Ambiguities ...................................................................................... 21
  2.5 Analysis of the Discussed Algorithm ....................................................... 22
  2.6 Summary .................................................................................................. 23

CHAPTER 3 SOUND LOCALIZATION IN THE CONTEXT OF ARTIFICIAL
  AWARENESS ............................................................................................... 25
  3.1 Introduction .............................................................................................. 25
  3.2 Overview of Basic Sound Localization ..................................................... 27
    3.2.1 Sound waves and sound triangulation ............................................ 27
    3.2.2 Cross Correlation Based ITD Estimation ....................................... 31
    3.2.3 Frequency Decomposition Based ITD Estimation ......................... 36
    3.2.4 Steered-Beamformer based Sound Localization ............................ 38
  3.3 Previous Implementations ........................................................................ 40
1.1 Motivation

Over the past twenty years, there have been numerous advancements in the field of personal computing which have made computers one of the most essential tools of the modern world. However, while a great amount of effort has been placed in improving the speed and capability of computers, their potential for high-level user interaction and environmental awareness has not been sufficiently explored.

These potentials mainly depend upon the ability of computers to be aware of their surroundings. This ability, which can be called artificial awareness, would allow computers to know the location of the events and objects in the environment and to be aware of the characteristics of those events and objects. Ideally, artificial awareness can be achieved by the integration of fields such as vision, sound localization, and sound recognition. However, the type of sense used in an artificial awareness system can include GPS localization, Internet data monitoring, and many other types of senses.

Artificial awareness of the environment and its occupants can serve a variety of
applications. One such application is the protection of homes and offices against intruders. Unlike modern security systems, artificially aware systems can alert the occupants while giving them detailed information about the number, location, and direction of movements of the intruders. This information is vital to the safety of the house occupants and can save their lives by allowing them to make the correct choices in a state of confusion and panic. Another application is in teleconferencing, where an artificial awareness system can automatically switch the main image to that of the current speaker. Artificial awareness may also lead to the development of new applications that use multi-modal perception to enhance the interaction mechanism between humans and computers.

1.2 Previous Artificial Awareness Projects

In recent years, there have been several environmental awareness projects that have explored the components and algorithms needed ([1], [2], [3], [4], [5], and [6]). For the most part, the goal of these projects has been to transform computers from a lifeless appliance into an interactive and aware element of the environment. Such attempts have usually consisted of significant alterations to the environment. The problem with some of these alterations, however, has been the fact that they are difficult and time consuming to configure and install. Another limitation with the explored solutions has been that they are location specific, meaning that everything is configured for specific elements of a single environment such as window locations and lighting [7].

An interesting approach to artificial awareness is currently under development at
Stanford University, where the main goal is for the computers to be aware of other artificial and biological systems. This project, called the Interactive Workspaces, attempts to create a knowledge base that is common to all artificial systems in an environment, including Personal Digital Assistants, video projectors, and computer terminals.

More traditional approaches exist at MIT’s Media and AI laboratories. Projects such as the Smart Rooms [8] and the Intelligent Environment [9] attempt to integrate multiple sensors in order to know the status of the occupants of the environment.

1.3 Multi-Sense Artificial Awareness Arrays

Multi-Sense Artificial Awareness Arrays, or MSAAAs in short, provide one method of implementing an artificial awareness system. The idea behind these arrays is to use numerous relatively inexpensive awareness modules to accomplish the same level or even a further level of awareness than would be obtained with more expensive single component systems, such as AI based vision. Also, as the “Multi-Sense” portion of the name suggests, MSAAAs are not limited to a single type of sense but can be employed using a variety of senses.

There are several requirements which need to be met before MSAAAs are fully realizable. These include the development of a multiple-sense and multiple-source data integration algorithm that is both computationally simple and general in terms of its applicability. Also, in order to develop a prototype multiple-sense awareness system,
several basic senses need to be selected in order to serve as a basis for the rest of the awareness array. Finally, in order to analyze the benefits of a MSAAA system, a prototype system needs to be implemented and tested. The work discussed in this thesis attempts to address most of these requirements.

1.4 Thesis Contribution

The major contributions of this thesis are in the fields of environmental awareness and sound localization. The main contributions in the field of environmental awareness is the introduction of the idea of the spatial probability map as a means for the integration of location information obtained using different senses. An intelligent false object removal algorithm was developed for identifying false objects erroneously detected by certain types of awareness systems. The contributions in the field of sound localization include the development and implementation of an iterative sound localization algorithm with immunity to background sounds, based on the idea of the spatial probability map.

1.5 Thesis Outline

These contributions are discussed in the following chapters as follows: In Chapter 2, a detailed description of algorithms related to multiple-camera object localization is given. Chapter 3 focuses on the implemented sound localization system including the iterative spatial probability sound localization algorithm and the analysis of the implemented system. In Chapter 4, the algorithms that integrate vision and sound localization as well as the results of the integrated system are discussed. The final
concluding chapter will describe possible applications of the developed systems and suggest future directions for work in this field.
2.1 Introduction

The most important sense for any awareness system, be it biological or artificial, is vision. For humans, vision allows objects to be located, identified, and tracked, and is often used to aid other senses. Vision also serves as a means of extracting information from scenes and environments and as a means of communication. It would certainly be fair to say that vision is central to the interaction, navigation, and awareness of humans and most other biological creatures.

In the artificial realm, vision has been studied in depth for many years ([10], [7]). Algorithms have been developed which enable scene objects to be located, tracked, identified, and analyzed. While these algorithms have not yet reached the level of their biological counterparts, they have achieved a great deal of success in certain situations [10]. These include object identification, localization, and characterization.

In many artificial awareness systems, cameras are used in conjunction with other senses ([5], [7], [9]). In the majority of these cases, images from 2-4 medium-resolution cameras are sent to a processing station where they are analyzed and integrated with the
results of other senses. Hence, the requirements for these projects include a high-bandwidth camera-computer network and one or more computers capable of handling real-time video processing and vision analysis.

This chapter proposes an object localization approach using multiple-camera vision. The proposed algorithm can be applied in a variety of situations with any number of cameras and can be easily integrated with other perceptual results. A spatial probability map is proposed as a means for integrating the results of multiple cameras.

2.2 Previous Vision Systems

Object localization can be achieved by using a single camera. For example, we humans can still locate objects in a scene by using a single eye. Such ability requires knowledge about all of the objects in the environment and is often computationally expensive.

Another common approach is stereo vision. Stereo vision consists of 1 or more pairs of cameras whose images are combined to produce depth information about the objects in the environment. Unlike monocular vision, detailed image understanding and decomposition is not required for depth information when using a multi-camera system. Such a system was implemented in [7], where two cameras were focused onto a grid-marked floor. When an object was placed anywhere in the grid, the cameras would notice an obstruction of the grid and would be able to find the position of the object based
on the location of the obstruction. Using stereo vision instead of monocular vision, object localization can be accomplished more accurately and robustly without any significant increase in computational overhead.

2.3 The Proposed Multiple-Camera Vision Algorithm

Multiple-camera vision consists of the integration and processing of images obtained by a large array of cameras placed at different locations in an environment. In contrast to other vision systems, multi-camera vision does not require medium or high resolution images, high bandwidth interconnections, or powerful computing platforms. The reason for this is that with multiple camera systems, very simple direct image to object location mappings exist that do not benefit greatly from an increase in resolution and do not require extensive computations, as will be discussed in the following sections.

2.3.1 Preliminary Image Decomposition and Spatial Probability Maps

The first step in the multi-camera object localization algorithm involves taking two images, one being an initial background image with no objects present and the second being an updated image which can include any number of objects. In order to identify the objects in the image, each pixel of the updated image is subtracted from the corresponding pixel in the background image. In this way, areas of large intensity differences between the two images would remain while areas with similar intensities will be removed. This procedure is called background segmentation. ([7],[11]).

Next, the extracted objects from the updated image are converted to a two-
dimensional probable location image of the scene, as illustrated in Figure 2.1. This image, which is called the spatial probability map (SPM), is a graphical indication of all of the probable locations of objects in the environment. An observed object (Figure 2.1a) can be a small object close to the camera or a very large object far away (Figure 2.1b). The ambiguity in object size and position is inherent to any system that uses monocular vision, because the distance from the image plane to the center of the camera lens is unknown. This leads to the spatial probability map (SPM) shown in Figure 2.1c.

![Figure 2.1](image_url) - The conversion of a scene image to a spatial probability map

### 2.3.2 Formation of Spatial Probability Maps

In a 2-D top-view of the environment, each SPM has the following relation to the coordinates of the detected object in images (Refer to Figure 2.2):

\[
Q(x,y) = S_{\text{norm}}(x,y,x_\circ,y_\circ,I,\theta)
\]  

(2.1)

Where \(Q(x,y)\) is the SPM image intensity at point \((x,y)\), \((x_\circ, y_\circ)\) is the location of the
camera, \( I \) is the scene image retrieved by the camera, \( \theta \) is the angle of rotation of the camera, and \( S_{\text{trace}} \) is the function generating the camera traces similar to the one shown in Figure 2.1 c).

\[
\begin{align*}
\text{Figure 2.2- Transformation of the top-view spatial location of an object to a location in the image plane} \\
\end{align*}
\]

\( S_{\text{trace}} \) can be derived from Figure 2.2 as follows:

Consider Figure 2.2a where a rotated camera is placed at \((x_c, y_c)\) in the environment. This system can be converted to that in Figure 2.2b with the following translation and rotation:

\[
P^1 = (P - P_c) \cdot R(-\theta) = [X \quad Y]
\]

(2.2)

Where \( P^1 \) is the position vector of the rotated and translated coordinate system depicted in the center image which corresponds to the regular position vector \( P \). \( P_c \) is the position vector of the camera location and \( R \) is the rotation matrix. As illustrated in Figure 2.2c, the point on the image plane of the modified coordinate system can be computed as follows:
\[ V(X,Y) = \frac{D \cdot Y}{X} \] (2.3)

Where \( D \) is the virtual lens to image plane distance in terms of scene image pixels specific to the camera, \( X \) and \( Y \) are as defined by equation 2.2, and \( V \) is the image plane location of the 2-D space coordinate (Note that \( V \) will be the horizontal coordinate in the image retrieved by the camera). From ([12]), the following definition of \( R \) is known:

\[
R(\theta) = \begin{bmatrix}
\cos(\theta) & \sin(\theta) \\
-\sin(\theta) & \cos(\theta)
\end{bmatrix}
\] (2.4)

Hence, if \( P = [x \ y] \), then the following relation between the elements of \( P^I \) and \( P \) can be determined:

\[
P^I = \begin{bmatrix} X & Y \end{bmatrix} = [x - x_c \ y - y_c] \cdot R(-\theta)
\] (2.5)

The image plane coordinate, can be determined from the above expressions as follows:

\[
V(x,y,x_c,y_c,I,\theta) = \frac{Y \cdot D}{X} = \frac{D \cdot [(x - x_c) \cdot \cos(\theta) + (y - y_c) \cdot \sin(\theta)]}{-(x - x_c) \cdot \sin(\theta) + (y - y_c) \cdot \cos(\theta)}
\] (2.6)

With this definition of \( V \), we can finally derive \( S_{trace} \) as illustrated below:
\[
S_{\text{trace}}(x,y,x_c,y_c,I,\theta) = \begin{cases} 
-1 & \text{if } V(x,y,x_c,y_c,I,\theta) \notin O_b \text{ for } b = 1, 2, \ldots, N \\
0 & \text{if } (x,y) \text{ is outside the field of view of the camera} \\
1 & \text{otherwise}
\end{cases} 
\]

Where \(O_b\) consists of the set of horizontal pixels associated with the \(b\)th object out of a total of \(N\) objects that are found in the current scene image. The reason that \(S_{\text{trace}}\) has a value of 0 outside the field of view of the camera is to insure that the probability of an object in those regions is not affected when combining multiple SPMs from different cameras.

2.3.3 Application of Spatial Probability Maps to Multi-Camera Vision

A trace from a single camera places constraints on the probable object position, but does not pinpoint the location. In order to obtain a more accurate estimate of the object position, further constraints, which can be obtained from different cameras, are required. One method of merging these constraints is by the point-wise addition of each of the SPMs obtained for each camera, as illustrated by Equation 2.8. The result of this operation will be a grayscale image that has a local intensity peak at the location corresponding to the object's position. In the case of multiple objects, multiple peaks would be expected.

\[
L(x,y) = \sum_i Q_i(x,y) = \sum_i S_{\text{trace}}(x,y,x_i,c,y_i,c,I_i,\theta_i) 
\]
Where $L$ is the merged SPM image and $Q_i$ is the SPM associated with the $i$th camera. This method is a direct mapping of object coordinates in an image to their possible positions in the environment. All that is required is the location and direction of the cameras, and from there, object locations in the environment can be determined based on each of the images that are obtained by the different cameras.

### 2.4 Intelligent False Object Removal

When several SPMs are merged according to Equation 2.8, the peaks of the merged SPM are used as an indication of the location of the objects in the environment. However, in the presence of multiple objects, it is possible for there to exist SPM peaks that do not correspond to any real object. Figure 2.3 illustrates the formation of a false object in a scene with two cameras and two real objects. The shaded regions in this figure correspond to the local maxima of the SPM and the lines define the exterior of the object traces produced from the images retrieved by each camera.

![Figure 2.3 - The formation of false objects](image)
The uncertainty in the validity of the identified peaks requires a solution that can differentiate between the false objects and the real ones. As a result, the Intelligent False Object Removal (IFOR) algorithm was developed.

2.4.1 The IFOR Algorithm

The IFOR algorithm analyzes the scene after the SPM peaks have been identified and after all possible object locations have been enumerated. This enumeration consists of labeling all of the SPM peaks numerically. Then, for each individual trace, the set of objects containing the numerical identifiers whose corresponding objects reside inside that trace is formed, as shown in the example below:

Figure 2.4 illustrates a simulated scene with 3 cameras and 2 real objects. Also shown are the numerical object sets associated with each trace. In this scene, there are a total of 8 identified objects with 6 being false.
In order to resolve the nature of the enumerated objects, the trace sets are analyzed by the IFOR algorithm. The first step in this process is to select a candidate object from each trace set and to combine all of the candidates to produce the candidate set $L$. The objects that were not selected as candidates are combined to produce the remainder set $R$. It should be noted that it is possible for an object to appear several times in the candidate set or the remainder set. Returning to the example of Figure 2.4, one possible candidate set is $\{4,6,1,4,1,3\}$, which consists of the first member of each of the trace sets. The remainder set corresponding to this candidate set is hence $\{2,3,7,5,2,5,7,8,4,6,5,8\}$.

The candidate set contains one possible set of real objects. In our example, the selected
candidate set proposed objects 1, 3, 4, and 6 to be real. In order to analyze the validity of each of the possible candidate sets, a heuristic approach was developed. This approach is based upon an experimental observation that real objects are seen by most of the cameras. The heuristic contains two weighing parameters that are analyzed for each of the members of the candidate set and summed at the end. The first parameter, $w_{i,C}$ is defined by the number of times that an object in the candidate set appears in the same set. For example, in our selection of the candidate set $\{4, 6, 1, 4, 1, 3\}$, the first object, which is 4, appears twice. Hence the value of $w_{i,C}$ for that element is 2. This procedure is carried out for each of the elements of the candidate set, even if an object is repeated. For the example candidate set, the total value for the first parameter is $2+1+2+2+2+1=10$.

The second parameter, $w_{i,R}$, is defined by the number of times that an object appears in the remainder set. Returning to our example, with the selected candidate set, the value of the $w_{i,R}$ parameter for the first element of the candidate set, which is object 4, will be 1. The total value for the second parameter is $1+1+0+1+0+1=4$.

The overall heuristic value consists of the weighted addition of the two parameters, as shown below:

$$h(L, R) = \sum_{\text{all objects } i \in C} b \cdot w_{i,C} + w_{i,R}$$

(2.9)

The $b$ parameter scales $w_{i,C}$ in order to give it more significance than $w_{i,R}$. It was experimentally determined that for $b \geq 3$ the candidate set with the highest heuristic value would entirely consist of real objects. Hence, the IFOR algorithm would consider all possible candidate sets and would select the set with the highest heuristic value. All of
the objects that appeared in that candidate set would be labeled as real and any other object that does not appear would be labeled as false.

With b=3, the heuristic value for the example candidate set would be 3*10+4=34. Considering the candidate set \( L_x = \{4,5,5,4,4,5\} \), the heuristic value for this will be 3*18+0=54. After considering all possible candidate sets, it is found that the \( L_x \) candidate set has the highest heuristic value, hence the objects 4 and 5 (the only ones that appear in \( L_x \)) are labeled as real and all other objects are labeled as invalid, which is correct.

2.4.2 Analysis of the IFOR algorithm

When the IFOR algorithm was fully developed, it was tested on several scenes that contained many objects, similar to the scene in Figure 2.5. In all of the cases, the program successfully identified the real objects.

![Figure 2.5 – A complicated scene that was correctly analyzed by the IFOR algorithm.](image)
Note that in this image, all local maxima are used as possible objects.

Figure 2.6 illustrates the situation where two real objects appear in a single trace. In this case, the IFOR system was able to correctly identify both objects as real ones.

![Figure 2.6 - An hidden object case that IFOR correctly resolved](image)

### 2.4.3 Complexity of IFOR

In order to evaluate the applicability of the IFOR algorithm, its complexity needs to be determined. This can be achieved as follows:

Assuming that we have $C$ cameras and $D$ objects with no hidden objects or missed objects\(^1\), we can derive an equation about the number of objects in the environment and the complexity of finding the real ones. With 1 camera and $D$ objects we have 0 possible objects identified because there will be no trace intersections or local SPM maxima. When we add another camera, for every camera trace there will be $(D-1)$ false objects identified, which works out to $D*(D-1)$ since there are $D$ camera traces. The third camera produces an extra $2*(D-1)*D$ false objects, since each of its $D$ traces will

---

\(^1\) By assuming that there are no hidden or missed objects, we are accounting for the most complex scenario. Any hidden or missed objects would decrease the computational requirements of the algorithm.
intersect \((D-1)\) traces of the first and \((D-1)\) traces of the second camera. Eventually, the final camera, camera number \(C\), will produce an addition of \(D^*(C-1)^*(D-1)\) false objects. Hence, the total number of objects \(X\) produced is:

\[
X = (D-1) \cdot (1 + 2 + 3 + \ldots + (C-1)) \cdot D + D = (D-1)^* C^* (C-1) \cdot \frac{D}{2} + D \tag{2.11}
\]

Which is composed of \((D-1)^* C^* (C-1)^* D\) false objects and \(D\) real ones.

Each camera trace will intersect \((C-1)^*(D-1)\) other traces and also 1 real object. This represents the number of elements in each IFOR validation equation. The total number of equations is consequently equal to the number of traces per camera times the number of cameras, which is \(D^* C\).

With a normal depth first search, the number of solutions that have to be searched and hence the complexity of finding a solution is:

\[
(1+(C-1)^*(D-1))^{D^* C} \tag{2.12}
\]

Table 2.1 illustrates the rapid growth of this function. The number of solutions that need to be analyzed for complex environments with many cameras and objects is not practical. For example, for ten simulated 4 camera and 3 object environments, the average execution time of the IFOR algorithm exceeded 4 hours when executed on a Cyrix M2-200 based computer.
<table>
<thead>
<tr>
<th>of Cameras</th>
<th>of Objects</th>
<th>Number of Solutions Analyzed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>729</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>729</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1.95 *10^6</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1.38 *10^10</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>6.55 *10^4</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1.38 *10^10</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1.00 *10^10</td>
</tr>
</tbody>
</table>

Table 2.1 – The complexity and execution time of the IFOR algorithm

2.4.4 Global Peak Selection

In a stand-alone multi-camera vision system, it is not necessary to allow all local maxima to be real objects. Hence, by selecting only the global peaks of the SPM, the search time is greatly reduced. It may still possible for some of the global maxima to correspond with false objects, however, the likelihood of these false objects is very small. Since their occurrence is dependent primarily on the placement of the cameras and the location of the objects, it becomes very difficult to quantitatively analyze the effects of global SPM peak selection for the general case. However, for the ten simulated 4 camera and 3 object scenes that took on average 4 hours with a local SPM maxima selection algorithm, the global SPM maxima selection algorithm execution time was decreased to an average of 80us on a Cyrix M2-200 based computer. This modification allows the IFOR algorithm to remain practical in most situations.

There are situations where the local peaks may be useful, however. One example is when the confidence in the camera results is not high, in which case an object is either
falsely detected by the camera or a real object is not detected at all. In these cases, it becomes possible for non-global local SPM maxima to be associated with a real object. In order to identify the false objects in such a case, either all local and global peaks have to be taken into account, which results in extensive computational requirements for IFOR, or other types of senses can be utilized to aid the localization process.

2.4.5 Ambiguities

While IFOR is designed to deal with multi-camera environments, it is possible to obtain ambiguous results in certain situations. The simplest of these ambiguities occurs in a 2 camera and 2 object environment. In Figure 2.8, for example, there are two equally valid solutions. Either objects 1 and 2 or objects 3 and 4 are real. Without additional information, it is impossible to determine which of the objects are real. Again, the addition of other types of senses, such as sound localization, can be used to resolve ambiguities. The integration of vision and sound localization is discussed in Chapter 4.

Figure 2.7 – An ambiguous situation with two cameras and two objects
2.5 Analysis of the Discussed Algorithm

In comparison to vision algorithms which attempt to localize objects by decomposing and understanding the scene, the SPM+IFOR approach has the advantage of requiring much less computations and being much simpler. Of course, the level of detail of information obtained from the two is quite different. With our approach, no information is known about the type and characteristics of objects being localized, whereas traditional vision algorithms do reveal more information about the nature of the objects.

In comparison to other multi-camera vision algorithms, the SPM+IFOR approach is advantageous because it does not require a training phase or any modifications to the environment such as the grid that was employed in [7]. Also, our approach has lower computational complexity and as a result can be implemented in a less expensive manner. The low computational overhead in our algorithm arises from the ease in which the results of each individual camera are integrated with those of other cameras because only point-wise additions of the SPMs are needed for integration.

Another advantage of this system over other multi-camera systems ([7], [8]) is the reduction in required bandwidth between the cameras and the host computers which are typically in charge of processing the camera information. Currently, almost every vision system consists of a camera that sends captured images to a host computer where the images are decomposed and analyzed. With the discussed multi-camera object localization system, image decomposition can be done locally inside each camera with
only the image locations of the objects being sent to a host PC for further processing. The results of the image decomposition only involve image coordinates and do not require a large amount of bandwidth to send to the central computer. In order to illustrate this point, we shall consider an example:

Assuming that a camera takes 30 images every second with an 8-bit resolution at a size of 1024 by 768 pixels, for all previous systems 189 Mbits of data needs to be transferred to the host computer every second. With the SPM+IFOR approach, the image can be decomposed by background segmentation and only the coordinates of the discovered objects are sent to the host computer. Each set of coordinates will consist of four 10-bit parameters which define the bounding box of the object in the image. Even in the unlikely event that there are 20 objects present in the room (and all are seen by the camera), the total amount of information that would be sent to the host computer would be equal\(^2\) to 24 Kbits/second. This represents an approximately 8000-fold improvement in the amount of bandwidth required for object localization.

2.6 Summary

In this chapter, a multi-camera object localization algorithm was proposed. This algorithm consisted of the background segmentation of an image, the production of spatial probability maps for each camera, which are 2-dimensional probability distribution of the positions of objects, and the point-wise addition of the SPMs obtained from all of the cameras. One problem that occurred with this algorithm was the

\(^2\) This is derived by multiplying 20 objects/image x 40 bits/object x 30 images/second
formation of SPM peaks in the final image that did not correspond with real objects in the environment.

In order to deal with these false objects, the Intelligent False Object Removal algorithm was developed. Based on a heuristic analysis function, this algorithm allowed false objects to be removed in all unambiguous situations in a simple fashion. Also, the complexity of the IFOR algorithm was calculated and an SPM global peak selection enhancement was made to it in order to reduce its execution time. In the final section of this chapter, the benefits of this algorithm over other traditional vision algorithms were discussed.
3.1 Introduction

Sound Localization is the ability to estimate the position of a sound source in the environment. For an artificial awareness system, the ability to localize sound is of extreme importance. The information gathered by such a subsystem can be used to obtain a more complete picture of the environment. In many situations, other senses alone cannot provide enough information about the environment and its occupants. Sound localization provides detailed position information about the vocally active objects in the environment. Furthermore, it allows the rest of the system to become aware of some of the behaviors associated with the objects in the environment such as the level of participation of an individual in a conversation.

Typically, in human communication and interactions, sound is used as a means of focusing attention. For example, at gatherings, social occasions, meetings, and conferences, at any given time there is usually one speaker while others listen to the speech. An artificial sound localization system could allow the attention of the awareness
system to focus on the speaking occupants of the environment thereby minimizing the amount of processing and bandwidth required for analyzing the situation.

Sound localization would also be able to aid the awareness system in tracking sound producing objects where senses such as vision may fail. This can occur in cases where the environment is too dark such as at night or when there are not enough cameras placed to cover the entire scene. Also, sound localization systems can be more cost effective than cameras in terms of object detection because they cost much less and require less processing.

Another potentially useful application of sound localization is to enable system modules to locate each other. This last possibility provides a means of designing a completely self-locating and self-configuring artificial awareness system. For example, when a new module is added to the system, it would emit a sound pulse that the sound localization system would detect and convert to a spatial position. Thus, it would no longer be required for the location of module to be manually measured, resulting in a significant decrease in the installation time.

The inspiration behind sound localization stems from the sound locating ability of humans and other biological creatures ([20],[21]). Not only can we humans communicate through this sense but we can also locate speaking objects in the environment. In fact, we can localize sound when there are several people or objects producing a large amount of noise (e.g. listening to a conversation in a crowded and
noisy room). Such an ability would be the ideal for any artificial sound localization system.

3.2 Overview of Basic Sound Localization

Sound localization is based upon measuring the differences in the time of arrival of sound at several locations. Based on these time differences, the exact position of the sound source can be computed. In an idealized situation sound localization is straightforward. However, in the presence of background noises and reflections, it becomes extremely hard to determine differences in the time of arrival. In order to deal with practical situations, it is helpful to initially review some of the sound localization techniques for the ideal case.

3.2.1 Sound waves and sound triangulation

When sound is emitted from a sound source, it propagates as a spherical wave centered at the location of the source. The distance traveled by sound from the source to an observer is given by:

\[ d = (t_i - t_0) \cdot v \]  \hspace{1cm} (3.1)

Where \( t_0 \) is the time of emission of the sound wave by the source, \( t_i \) is the time of its arrival at the observation point, and \( v \) is the speed of sound in the air (which is typically about 345 m/s at 25°C ([12])).
Consider the sound source and the two observers illustrated in Figure 3.1. For the source at location \((x,y)\), the distances \(d_1\) and \(d_2\) are given by:

\[
d_1 = \sqrt{(x)^2 + (y-a)^2} \tag{3.2}
\]
\[
d_2 = \sqrt{(x)^2 + (y+a)^2} \tag{3.3}
\]

After a few simple manipulations, the relations described below can be derived:

\[
y = \frac{d_2^2 - d_1^2}{4a} \tag{3.4}
\]
\[ x = \sqrt{d_1^2 - \left( \frac{d_2^2 - d_1^2}{4a} - a \right)^2} \] (3.5)

The problem in determining the distances \( d_1 \) and \( d_2 \) is that the time of origination of the sound source, \( t_0 \), is not known. One parameter that can be measured is the time difference \( t_2 - t_1 \), which is known as the interaural time difference (ITD). It is related to the observer to source distances as follows:

\[
\text{ITD} = t_2 - t_1 = \frac{d_2 - d_1}{v} \tag{3.6}
\]

The knowledge of the ITD parameter alone does not pinpoint the location of the sound source. Instead, it constrains the possible location of the sound source to a hyperbola. From equations 3.6, 3.3, and 3.2, we have:

\[
\sqrt{x^2 + (y + a)^2} - \sqrt{x^2 + (y - a)^2} = d_2 - d_1 = v \times \text{ITD} \tag{3.7}
\]

After rearranging, we obtain:

\[
\frac{4}{(v \times \text{ITD})^2} y^2 - \frac{4}{4a^2 - (v \times \text{ITD})^2} x^2 = 1 \tag{3.8}
\]

Equation 3.8 describes one of two hyperbolas depending on the sign of ITD, as illustrated in Figure 3.2. For \( \text{ITD} = 0 \), the two hyperbolas collapse to the line \( y = 0 \).
In order to estimate the location of a sound source in a 2-D environment, at least two microphone pairs are required. With two microphone pairs, we would obtain two hyperbolas whose intersection corresponds to the location of the sound source, as shown in Figure 3.3. This simple triangulation calculation is the first step in many sound localization systems ([13], [14], [15]).
In order to ensure the existence of two microphone pairs, a two dimensional sound localization system must consist of at least three microphones. Figure 3.3 illustrates one possible microphone configuration. Such a system needs to estimate the ITDs of the two microphone pairs. Finally, source localization can occur using Equation 3.8 by finding the intersection of the two hyperbolas that are associated with each ITD, as shown in Figure 3.3.

Another useful piece of information, other than the ITD, is the level difference between the sound signals received by the two microphones. This level difference, known as the interaural level difference (ILD), has been shown in certain cases to be useful in aiding sound localization ([16], [13]).

3.2.2 Cross Correlation Based ITD Estimation

A key step in sound localization is the estimation of the ITD parameter between two microphones. This would be relatively simple if the sound signals obtained by each
of the microphones were identical time-shifted versions of one another and if both signals had a recognizable start, as shown in Figure 3.4.

![ITD Estimation using Signal Starting Points](image)

Figure 3.4 – Estimation of the ITD using signal start times

In Figure 3.4, the difference in the start times of the signals obtained by each microphone is ITD. In most situations, however, there is not a clear start for the sound signals that are received by each of the microphones and a start-time ITD estimation method cannot be used.

In the absence of a well-defined starting point, an estimate of the ITD can be obtained by cross correlating the signals received by the two microphones. In the absence of noise, the location of the peak of the cross correlation curve corresponds with the true ITD ([13], [5]).
Consider the two time-shifted signals shown in Figure 3.5, which are given by:

\[ m_1(t) = a \cdot x(t) \]  
\[ m_2(t) = b \cdot x(t - \tau) \]

The cross correlation between the two signals can be stated as:

\[ R_{12}(u) = \int_{-\infty}^{\infty} m_1(t)m_2(t - u)dt \]

In the frequency domain, we can state the cross-power spectral density \( S_{12}(w) \), which is the Fourier transform of the cross correlation function \( R_{12}(w) \), as:

\[ S_{12}(w) = ab \cdot X(w) \cdot X^*(w) \cdot e^{-j\omega \tau} = ab|X(w)|^2 \cdot e^{-j\omega \tau} \]  
\[ (3.12) \]

Where \( X(w) \) is the Fourier transform of the original sound signal \( x(t) \). Note that \( |X(w)|^2 \) is the Fourier transform of the autocorrelation function \( R_x(u) \) of the sound signal \( x(t) \). We can transform back to the time domain and restate equation 3.12 as:

\[ R_{12}(u) = \int_{-\infty}^{\infty} ab \cdot \delta(t - \tau) * R_x(u - t) \cdot dt = ab \cdot R_x(u - \tau) \]

\[ (3.13) \]
Hence, by cross correlating the two time shifted signals, we obtain the shifted autocorrelation function of the sound signal. Figure 3.6 illustrates the result of the cross correlation of the sound signals in Figure 3.5.

![Cross Correlation of a Signal with a time shifted version of itself](image)

**Figure 3.6** – The cross correlation of a sound signal with its ideal time shifted version

The cross correlation function ideally has a peak whose index corresponds to the shift between the two input signals. The correlation magnitude is an indication of the degree of correlation between the two signals. Due to the fact that the signals in Figure 3.5 are identical, their resulting cross correlation peak (Figure 3.6) has a main peak of 1.

Since the cross correlation function is dependent on the sound signal’s autocorrelation function, it often has a smooth peak accompanied by several secondary peaks. In the presence of background sounds and reflections, it becomes possible for secondary peaks to be elevated to become the main peak of the cross correlation curve.
thereby resulting in false ITD results. This problem can be solved by the introduction of a cross correlation frequency whitening filter. By introducing the adaptive filter:

\[ H(w) = \frac{1}{ab|X(w)|^2} \]  

Equation 3.12 can be rearranged to:

\[ e^{j\omega r} = S_{12}(w) \cdot H(w) \]  

and can be represented in the time domain as:

\[ \delta(u - \tau) = R_{12}(u) * h(u) \]  

Equations 3.15 and 3.16 illustrate that if the cross correlation \( R_{12}(u) \) is filtered by \( H(w) \), then the result will be a delta function positioned at \( \tau \) for the case where the two sound signals are identical time-shifted versions of one another. This approach, which is called whitening of the cross correlation spectrum, allows the time-shift between the two signals to be easily found, and it can be shown to be the optimal approach for estimating \( \tau ([5]) \).

In practice, however, small differences in the retrieved signals and background sounds can greatly distort the results of the ideal cross correlation filter, as shown in Figure 3.7. In this figure, the time-delayed signals of Figure 3.5 are cross correlated and shifted according to Equation 3.14. While the correct time-shift peak is much sharper than the peak of the original cross correlation function shown in Figure 3.6, it also is accompanied by several secondary peaks. The reason for the secondary peaks is that the
original sound signals are not completely identical, because they are composed of the finite-window segments of two identical time-shifted signals.

![Cross Correlation Filtered by H(w)](image)

**Cross Correlation Filtered by H(w)**

Figure 3.7 – The filtered cross correlation of a sound signal with its ideal time shifted version

As a result, various other filters ([5],[17]) have been introduced. One of the more successful non-optimal filters is discussed in Section 3.3.3.

### 3.2.3 Frequency Decomposition Based ITD Estimation

An alternative estimation method is to use a frequency decomposition of the sound signals to estimate a correct value for the ITD ([14], [15]). This technique can be described as follows:
Assuming that the sound signal emitted from the source is $a \cdot m(t-t_0)$ and that the sound signal arriving at the first and second observation points are $b \cdot m(t-t_1)$ and $c \cdot m(t-t_2)$ respectively. These signals can be represented in the frequency domain as defined below:

$$M_0(w) = a \cdot M(w)e^{-j\omega_0}$$  \hspace{1cm} (3.17)

$$M_1(w) = b \cdot M(w)e^{-j\omega_1}$$  \hspace{1cm} (3.18)

$$M_2(w) = c \cdot M(w)e^{-j\omega_2}$$  \hspace{1cm} (3.19)

Where $M_0(w)$, $M_1(w)$, and $M_2(w)$ are the respective frequency representations of the sound signals at the source, the first observation point, and the second observation point. From equations 3.10 and 3.11 we obtain:

$$\frac{M_1(w)}{M_2(w)} = \frac{b}{c}e^{j\omega(t_2-t_1)} = \frac{b}{c}e^{j\omega(\tau_{TD})}$$  \hspace{1cm} (3.20)

Thus, if $P(w)$ be the phase of the ratio of the Fourier transform of the first sound signal to the Fourier transform of the second sound signal, we have:

$$P(w) = \angle\left(\frac{M_1(w)}{M_2(w)}\right) = w \cdot ITD$$  \hspace{1cm} (3.21)

With this approach robust estimation techniques can be applied to obtain an estimate for the ITD by using the relation:
\[
ITD = \frac{P(w)}{w}
\]  

(3.22)

This approach, while computationally expensive, does have several advantages including small signal duration requirements and accuracy ([14]). Section 3.8 compares the various sound localization algorithms.

3.2.4 Steered-Beamformer based Sound Localization

Another sound localization approach that has been employed is the steered-beamformer approach ([18]). This method does not rely on specific ITD estimates but rather obtains its localization results directly from the sound signals. This is done by computationally steering the microphone array to a certain location and analyzing the relative power of sound at that position ([18], [19]).

Figure 3.8 – Possible placement of microphones and source for a steered-beamformer sound localization system

As an example, consider the triple-microphone situation shown in Figure 3.8. The localization system first selects a spatial location \( D \) for analysis and calculates the
distances $DA$, $DB$, and $DC$. Assuming that the microphones are paired as $\{A,B\}$ and $(B,C)$, the time difference of arrival $\tau$ associated with each pair is calculated as follows:

For the $\{A,B\}$ pair:

$$\tau_{AB} = \frac{\text{Distance Differential}}{\text{Speed of Sound}} = \frac{DB - DA}{v} \quad (3.23)$$

and for the $\{B,C\}$ pair:

$$\tau_{BC} = \frac{\text{Distance Differential}}{\text{Speed of Sound}} = \frac{DC - DB}{v} \quad (3.24)$$

To obtain a sound power estimate for the selected location, and assuming that the retrieved sound signals for microphones $A$, $B$, and $C$ are $m_A(t)$, $m_B(t)$, and $m_C(t)$, respectively, the following estimation algorithm is used:

$$\text{Power}(D) = \int m_A(t) \cdot m_B(t - \tau_{AB}) \cdot dt + \int m_B(t) \cdot m_C(t - \tau_{BC}) \cdot dt \quad (3.25)$$

By searching several spatial locations, the position yielding peak sound power is located and is assumed to be the sound source position.

The problem with the steered-beamformer approach is that in the presence of noise, the power distribution over space has many local maxima that make the peak search difficult. Also, most beamformer-based systems are sensitive to the start position of the search and involve relatively complex computations which in some cases prohibit
their use for real-time implementations ([19]). The advantages of the steered-beamformer over other algorithms include more scalability, because the power function can incorporate any number of microphone pairs, and the ability to find multiple sound sources, because each source will appear as a local maxima.

3.3 Previous Implementations

Many different sound localization systems have been implemented in the past ([13], [14], [5], [18]). In all of these systems, sound signals of a fixed duration are collected and analyzed. The actual details of the analysis vary from system to system. Some of the more successful sound localization approaches are described below:

3.3.1 Learning-based Approach

One method of translating multi-source sound signals to a spatial location is by using a learning algorithm. In this approach, the sound signals for each microphone pair are cross correlated and the resultant peak location, the peak correlation, and other relevant parameters are entered into an artificially intelligent program which learns the relationship between its input parameters and the corresponding sound source location.

A successful learning-based three-dimensional sound localization system, which was developed at Michigan State University, is described in ([13]). In this system, a non-coplanar array of four microphones is used in a two-stage sound localization procedure. In the first stage the algorithm is trained by obtaining ILD and ITD results for a speaker
at a known spatial location. In the second stage, ITD estimates between the four microphone pairs are passed as arguments to the learning algorithm that estimates the spatial location of the sound source. In both stages, 0.5 s sound windows are used. Also, ITD estimates are obtained with the unfiltered cross correlation method. This learning-based system was tested with an estimated Signal to Noise Ratio (SNR) of 20 dB.

Such a procedure is effective in dealing with permanent phenomena such as background sounds and sound reflections. However, an initial learning period is needed for every new environment. Also, this system cannot easily deal with secondary speakers or unexpected sounds. Furthermore, this algorithm has a large memory and computer resource requirement, which makes it impractical for many situations.

3.3.2 The ITD to Location Approach using Frequency Decomposition

Another successful sound localization implementation developed at Brown University is that of [19]. Here a 10-element bilinear array of microphones placed at 0.25m intervals in two parallel rows has been realized. Sound frames with a duration of 25.6 ms sampled at 20 kHz are collected from each microphone and analyzed only if they contained speech.

The analysis follows the frequency decomposition approach. The analysis of each frame either results in a location estimate or is deemed invalid. The SNR in the test environment was estimated to be somewhere between 5 dB and 30 dB depending on the location and orientation of the speaker.
3.3.3 The Direct ITD to Location Approach using Cross Correlation Filtering

An implementation using the cross correlation technique described in Section 3.2 is that of [5]. Here, a sound localization system was employed for guiding a camera to the active speakers in a room. The sound localization system used a sub-optimal filtered cross correlation approach. The sub-optimal filter improves performance in the presence of noise and reflection.

The optimal filter attempts to whiten the frequency spectrum of the cross correlation. This is accomplished by amplifying the frequencies that have small magnitudes and by attenuating the frequencies that have large magnitudes. However, in most speech spectrums, frequencies with large magnitudes are usually associated with the main speaker and those with smaller magnitudes are more likely to belong to background sounds. Hence, a side effect of the optimal filter is to significantly increase the SNR of the filtered cross correlation. A solution to this problem is proposed by [5], where the optimal filter is modified so that frequencies with larger magnitudes are weighed more heavily than those with less magnitudes. [5] proposes the following class of filters:

\[
H(w) = \frac{1}{(|M_1(w)| \cdot |M_2(w)|)^\rho}
\]

(3.26)

Here, the exponential coefficient \(\rho\) is used to change the weight associated with signals of different magnitudes. A \(\rho\) of 1 yields the original optimal filter, and a \(\rho\) of 0 corresponds to unfiltered cross correlation. A value between 0 and 1 gives higher weight to
frequencies with larger magnitudes. According to [5], a $\rho$ of 0.75 was found to produce the best results for a variety of different environments.

3.4 Description of the Proposed Sound Localization System

The problems common to most previously implemented sound localization systems are the sensitivity to the SNR and the high computational demand. In order to develop the sense of sound localization for artificial awareness systems with particular attention to low SNR robustness and computational simplicity, a two-dimensional sound localization system was implemented. The system consists of 3 microphones placed in a linear fashion at 0.5m intervals as shown in Figure 3.9. The linear array is placed at a height of 1.6m in order for it to be coplanar with the environment speakers.
The test environment is composed of a 3m by 4m room with the microphones placed on the front wall, as shown in Figure 3.10. Also, there are two computers near the rear of the room whose fans are the main source of background noise in the environment.

![Figure 3.10 — Top view of the test environment](image)

In order to minimize interference, each microphone module has an on-board amplifier, low-pass filter, and an 8-bit analog-to-digital converter (ADC). Figure 3.11 shows an actual picture of a microphone module along with all of its main components.
Each module is responsible for the initial analog processing of the sound signals, digitization, and transmission to a host computer via a data interface module (Figure 3.12). All ADC control signals which include a clock and a chip select line are provided by the host computer.
In the analog processing stage, the low-pass filter is used to remove high frequency noises that are produced by the power supply, with a cut-off frequency of 10 kHz. The filter also behaves as an anti-aliasing filter. The 8-bit ADC has a sampling rate of 24 kHz. After the sound signals are digitized and adjusted so that their mean is 0, they are transferred to the host computer where their variance is analyzed every 10 samples according to Equation 3.28 below. A sudden increase in the variance, as illustrated in Figure 3.13, is used to initiate the sound window collection process.

![Speech Onset Detection](image)

**Figure 3.13** – The selection of 6000-sample sound bursts for localization analysis

If \( m(k) \) is the sound sample acquired from the ADC with an absolute mean \( \bar{m} \) and \( N \) is the number of samples in the analysis window, then the associated variance is given by:

\[
V(N) = \sum_{k=0}^{N-1} \left( m(k) - \bar{m} \right)^2
\]

(3.28)
Sound windows are collected in sequential 256-sample (10.67 ms) windows whose variance is again analyzed using Equation 3.28. Each window is processed if its variance is greater than a predetermined variance associated with the ambient sounds in the environment. The background sound variance is obtained during the initialization of the sound localization program. This approach assumes that the background sounds do not increase significantly after the system is started. Processing is done using an iterative spatial probability algorithm that will be discussed in Section 3.6.

The sound window collection process continues until there are enough windows to insure a correct localization result (The details of this will be discussed in Section 3.6.3). Sometimes, especially in high intensity noise situations, it is possible for sound windows to be collected from multiple speech bursts.

The sound localization system requires real-time data collection and control. For proper operation, the host computer was equipped with the MS-DOS operating system. This ensured that there would be no context switching or unexpected delays during the sampling of the sound signals. To interact with the user and to illustrate the results of the localization process graphically, a second computer was used.

The processing performed by the first computer was also implemented with Field Programmable Gate Arrays to test their suitability for this application. However, this resulted in increased implementation complexity. The increase in speed was not of any major benefit, because the original PC-based system was capable of real-time sound
localization. Hence, the final sound localization system relied on dedicated sound localization computers.

3.5 Practical Considerations in Sound Localization

For a sound localization system to function in a real environment, there are many issues including background sounds and reflections which need to be addressed. This section presents some of the difficulties encountered in practical sound localization systems. In the presence of secondary sound sources, such as the speech of individuals other than the main speaker or the presence of background noise sources such as computer fans, the cross correlation results can be disturbed greatly. These difficulties have led to the development of the iterative algorithm described in Section 3.6.

3.5.1 The Effects of Background Noise

Let $n(t)$ and $n(t + \xi)$ represent the noise sound signals received by the first and second microphones, respectively. Equations 3.9 and 3.10 can be modified to:

$$m_1(t) = a \cdot x(t) + c \cdot n(t)$$  \hfill (3.29)

$$m_2(t) = b \cdot x(t - \tau) + d \cdot n(t - \xi)$$  \hfill (3.30)

Equations 3.29 and 3.30 illustrate the incorporation of a new sound source into the sound signals for each microphone. In contrast to Equation 3.12, the cross-power
spectral density between the two microphone signals is now given by (Using Equation 3.11):

\[
S_{12}(w) = \left( e^{-j\omega \tau} \cdot \left( ab + \frac{bc \cdot N(w)}{X(w)} \right) + e^{-j\omega \tau} \cdot \left( \frac{ad \cdot N'(w)}{X'(w)} + \frac{cd \cdot |N(w)|^2}{|X(w)|^2} \right) \right) |X(w)|^2 \tag{3.31}
\]

Equation 3.31 incorporates the effects of the presence of a secondary sound source on the cross-power spectral density function. These effects are dependent on the spectrums of the main and secondary speakers and cannot be easily analyzed. The term with the 'ab' coefficient represents the cross-power spectral density component that is associated with the main sound signals. The 'bc' and 'ad' terms result from the joint presence of the main sound and the noise source.

In some cases, the effects of the background sound sources are sufficient to cause the peak of the cross correlation curve to no longer correspond with the main speaker. An example is shown in Figure 3.14. The results in this figure correspond to a situation with a main-speaker signal-to-noise ratio of approximately 4 dB. Even with this relatively high SNR, the main sound and noise combined to produce a cross correlation peak which does not correspond to either the noise source ITD or the speaker location ITD.
Figure 3.14 – The cross correlation of two sound signals in the presence of a noise source

By filtering the cross correlation curve by the optimal filter of Equation 3.14, the effects of the noise source are reduced, as shown in Figure 3.15. However, the peaks that are not associated with the main sound source are still very large and can in certain situations, such as increased background sound intensity or reduced main speaker intensity, overtake the correct cross correlation peak. Figure 3.15 gives the filtered version of the cross correlation curve illustrated in Figure 3.14.
3.5.2 The Effects of Sound Reflections

The next problem that can occur in a sound localization system is reflections in the environment of both the noise and the sound. This is a difficult effect to model, because the strength of the reflected signals depends on the location and direction of the sound source as well as the nature of the reflecting surfaces. For example, if a person speaks directly towards the microphone array the reflection will be insignificant whereas if he speaks towards one of the side walls the strength of the reflected sound may be greater than the actual straight speaker to array sound signal.

For the implemented system, a limitation was placed on the angle of speech of the speaker in order to insure that the reflections did not significantly disrupt the localization process. This limitation consists of the speaker facing the array when speaking. Small
deviations to this angle of speech, as will be shown later in Section 3.7.2, did not cause any significant disruptions.

3.6 The Iterative Spatial Probability Algorithm

In order to overcome some of the problems described in the previous section, the iterative spatial probability (ISP) algorithm was developed. This algorithm attempts to address problems such as accuracy in low SNR situations and sensitivity to reflections. Basically, the ISP algorithm consists of a pseudo beam-forming sound localization algorithm that dynamically adjusts the number of sound samples used for analysis in order to yield a more accurate location estimate.

3.6.1 Overview of the ISP Algorithm

In the system developed by the author, which was described in Section 3.4, 256 element sound windows are continuously acquired for each of three microphones, and the samples are sent as input to the ISP routines. Here, the three cross correlation functions obtained from the three possible combinations of microphone pairs are computed\(^3\). This can be accomplished as illustrated below:

\[
X_{cor}(m) = \sum_{c} s_{i,1}(c) \cdot s_{i,3}(c-m) \quad \text{with } i = 1, 2, 3, \ldots
\]  

(3.32)

Where \(s_{i,1}(c)\) and \(s_{i,2}(c)\) are the sound samples received by the ADCs of the first and second microphones of the \(i\)th microphone pair, respectively. After the cross correlations

\(^3\) Note that while only two cross correlations are necessary, the third one, which may seem redundant, is also computed, for reasons which will be discussed in Section 3.6.
are computed, they are filtered using the sub-optimal filter introduced by [5] which was discussed in Section 3.3.3.

This process is repeated \( N \) times using \( N \) successive sound windows and each time the index of the main peaks of the three filtered cross correlation functions are incorporated into three histograms. The histograms record the number of times a cross correlation peak occurred at a given sample index (Refer to Figure 3.13). The reason for using histograms is to obtain an approximation of the probability density function (pdf) [22] of the cross correlation peaks. In fact, as \( N \to \infty \), each histogram will approach the cross correlation peak pdf of the corresponding microphone pair. We can define the filtered cross correlation function \( Y_{cor_{i,z}}(m) \) as:

\[
Y_{cor_{i,z}}(m) = \sum_{\beta} h_{z}(\beta) \cdot X_{cor_{i,z}}(m - \beta)
\]  

(3.33)

Where \( h_{z}(\beta) \) is the impulse response of the sub-optimal filter associated with the \( z \)th sound window stated in Equation 3.26, and \( X_{cor_{i,z}}(m) \) is the cross correlation function of the \( i \)th microphone pair and associated with the \( z \)th sound window defined by Equation 3.32.

The histogram function \( H_{i}(m) \) is given by:

\[
H_{i}(m) = \sum_{z=1}^{N} T_{peak}(Y_{cor_{i,z}}(m))
\]

(3.34)

Where \( T_{peak}(x(t)) \) has a value of 1 at \( t_0 \) if and only if \( x(t_0) \) is the maximum of the function \( x(t) \) and 0 otherwise, \( N \) is the number of iterations or sound windows used in the ISP algorithm. Figure 3.16 shows an example of a cross correlation peak histogram.
Finally, the histograms are converted to a 2-D spatial probability map (SPM) as described in Section 3.6.2 below. The SPM is then used to estimate the location of the sound source.

3.6.2 Spatial Probability Maps in the Context of Sound Localization

In this thesis, a spatial probability map (SPM) is proposed as a means of integrating the ITD histogram results obtained from each of the microphone pairs. An SPM is a two dimensional image in which the brightness of each image pixel is proportional to the probability that the sound source is at that location. In Figure 3.17, an SPM and the corresponding environment model are illustrated. The point of highest intensity in the SPM corresponds correctly with the location of the speaking object in the environment.
In two dimensions, ITD estimates for at least two microphone pairs allow the sound source to be located by calculating the intersection of two hyperbolas. However, in many situations, several different ITD values may be equally dominant and hence it becomes difficult to select the correct ITD for transformation into a hyperbola. Basically, the SPM is created by overlapping the corresponding hyperbolas of every possible ITD for each microphone pair. The intensity of each hyperbolic curve is adjusted based on the frequency of occurrence of the corresponding ITD.

The frequency of occurrence of an ITD can be obtained from the cross correlation histograms. For example, if $G_i(x,y)$ represents the intensity of the SPM image at coordinates $(x,y)$ due to the $i$th microphone pair, then we have:

$$G_i(x, y) = H_i(ITD(x, y))$$  \hspace{1cm} (3.35)
Where \( H_i(m) \) is the cross-correlation peak histogram for \( \text{ITD}=m \), and \( \text{ITD}(x,y) \) is the inter-arrival delay for location \((x,y)\), as computed by Equation 3.6. Hence, our equation of \( G_i(x,y) \) can be restated\(^4\) as:

\[
G_i(x,y) = H_i\left( k \cdot \sqrt{(x-x_{i,1})^2 + (y-y_{i,1})^2} - \sqrt{(x-x_{i,2})^2 + (y-y_{i,2})^2} \right)
\]  (3.36)

Where \((x_{i,1}, y_{i,1})\) and \((x_{i,2}, y_{i,2})\) are the 2-D spatial coordinates of the first and second microphone of the \(i\)th microphone pair, \(k\) is a SPM resolution and ADC sampling rate dependent parameter used to convert image plane time units to histogram index units, and \(v\) is the speed of sound in air as defined previously. The \(k\) parameter can be defined as follows:

\[
k = \text{SPM Resolution (in meters per pixel)} \times \text{Sampling Rate (in samples per second)} \]  (3.37)

The resolution of all SPMs in this Chapter was 5 cm per pixel and the sampling rate was fixed at 24 kHz; hence the \(k\) parameter had a value of 1200 m-samples/(pixels-seconds). This equation directly relates the spatial coordinates to their intensity given experimental results of the ISP algorithm.

\(^4\) Note that the Histogram index in Equation 3.36 is rounded because the histogram is computed with integral indexes.
In order to incorporate other microphone pairs, we simply add the frequencies of occurrence for each pair; that is:

$$I(x, y) = \sum_i G_i(x, y)$$

(3.38)

Where $I(x,y)$ is the overall intensity of the SPM image composed of layers indexed by $i$ and defined by $G_i(x,y)$. Equations 3.36 and 3.38 can be combined to form the more practical equation that is shown below:

$$I(x, y) = \sum_i H_i \left( k \cdot \frac{\sqrt{(x - x_{i,1})^2 + (y - y_{i,1})^2} - \sqrt{(x - x_{i,2})^2 + (y - y_{i,2})^2}}{v} \right)$$

(3.39)

This equation gives us a direct mapping between the cross correlation peak histograms obtained from the sound signals of all of the microphones and the intensities of the SPM image. When this mapping is conducted for two or more microphone pairs ($i=1,2,3...$), then it becomes likely that there will be a region of high intensity forming in the SPM which is indicative of the main speaker location. Shown below is an example illustrating the formation of the SPM as the number of iterations is increased:
There are several distinct characteristics of SPMs which make them well-suited to sound localization systems. First, they allow the results of any number of microphones positioned anywhere in the environment to be easily incorporated into a single database without the need for any complex computations or extensive memory requirements. Second, SPMs provide a suitable platform for the integration of sound localization with other localization senses, such as vision. The latter point will be addressed in the next chapter.

3.6.3 Automatic Termination of the Iterative Spatial Probability Algorithm
An important point that arises from Figure 3.18 is the fact that the system quickly locates the expected object. In this case, the actual object is located correctly after approximately 30 iterations. Subsequent iterations add little value. What is required is a metric for evaluating the status of the iterative process. One such metric was found to be the variation in the cross correlation peak histograms. For this purpose, we define the ITD vector set that, for iteration \( z \), consists of the vector of histogram peaks after the cross correlation results at iteration \( z \) are added to the database. The histogram variation is then represented by the square of the distance between the previous ITD vector set (at time \( z-1 \)) and the current ITD vector set (at time \( z \)) as shown below:

\[
\text{Var}(z) = \sum_i \left( \text{ITD}_i(z) - \text{ITD}_i(z-1) \right)^2
\]  

(3.40)

Where \( \text{ITD}_i(z) \) represents the position of the histogram peak obtained after the \( z \)th iteration for the \( i \)th microphone pair and \( \text{Var}(z) \) represents the total localization variation. One problem with using the localization variation as an indication of the progress of the localization is that the system may incorrectly show a small variation in the presence of several consecutive erroneous ITD results. In particular, this could happen in the early stages of the iteration. In order to compensate for this, an exponential decay element is added to Equation 3.40, which resulted in the following equation:

\[
\text{Var}(z) = \frac{\text{Var}(z-1)}{Q} + \sum_i \left( \text{ITD}_i(z) - \text{ITD}_i(z-1) \right)^2
\]  

(3.41)

with \( \text{Var}(0) = G \)
Where $Q$ is the drop-off factor and $G$ is an initial variation used to ensure that a certain minimum number of iterations are performed. These three parameters can be experimentally tuned to ensure the validity of the variance in all environment conditions. In our tests the best results were obtained with $Q=2$ and $G=80$.

Figure 3.19 illustrates the histogram variations corresponding to the results of Figure 3.18. Here, we see a direct correlation between the stabilization in the SPM images of Figure 3.18 and the changes in the value of $\text{Var}(z)$ in Figure 3.19. Above 37 iterations, the variations are very small.

![Histogram Variation Versus Iteration #](image)

Figure 3.19 – The variations in the ISP histograms as the number of iterations increase

The histogram variation can now be used as a means for terminating the iteration process. There, when $\text{Var}(z)$ drops below a predetermined threshold, the iterations stop and the final SPM image is produced. For the SPMs produced in this section, this
threshold was set to 8. While in some cases this variation threshold overestimates the number of required iterations, it was found to yield the most accurate estimate of the iteration status without underestimating the required number of iterations.

Figure 3.20 illustrates the complete steps involved in the ISP sound localization process.

![Figure 3.20 –Steps involved in the ISP algorithm](image)

### 3.7 Redundant Microphone Pair Processing

One of the features of the sound localization system proposed in this thesis that is different from its predecessors is redundant microphone pair processing. In a system with three microphones (Refer to Figure 3.21), two pairs are usually selected and sound localization is conducted based on the results of the two pairs. If the system is linear then other microphone pair combinations would be redundant. However, due to the non linearity of the system, the third microphone pair arrangement can also be useful. In
order to illustrate this fact, we can consider a single wall reflection scenario with an omni-directional sound source that is depicted in Figure 3.21.

![Figure 3.21 - Propagation delay of direct and reflected sound](image)

From the above figure, we can obtain the following relations:

\[
\begin{align*}
    d_0^2 &= x^2 + (y + a)^2 \\
    d_1^2 &= x^2 + y^2 \\
    d_2^2 &= x^2 + (y - a)^2 \\
    r_0^2 &= x^2 + (2p - y + a)^2 \\
    r_1^2 &= x^2 + (2p - y)^2 \\
    r_2^2 &= x^2 + (2p - y - a)^2
\end{align*}
\]

Where \(d_0\), \(d_1\), \(d_2\), \(r_0\), \(r_1\), and \(r_2\) are the microphone to signal source and microphone to reflection distances as illustrated by Figure 3.21. We can relate these distances to the
Signal to Reflected Sound Ratio (SRSR) for each microphone by taking into account propagation attenuation ([13], [19]),

\[ SRSR_i = \frac{I_{D,i}}{I_{R,i}} = \frac{r_i^2}{d_i^2} \]  

(3.48)

By applying the distances of Figure 3.21 to the above relation, we obtain:

\[ SRSR_0 = \frac{x^2 + (2p - y + a)^2}{x^2 + (y + a)^2} \]  

(3.49)

\[ SRSR_1 = \frac{x^2 + (2p - y)^2}{x^2 + y^2} \]  

(3.50)

\[ SRSR_2 = \frac{x^2 + (2p - y - a)^2}{x^2 + (y - a)^2} \]  

(3.51)

These three equations define a relation between spatial coordinates and the SRSR of each microphone. Since in the process of obtaining the cross correlation curves all microphone signals will be normalized, we need to choose two microphone-pairs from the three microphones in such a way that the sound reflection components of the cross correlation curves is minimized. As illustrated in Figure 3.22, there do not exist two microphone pairs that constantly have the highest combined SRSR. Different pairs perform better at different locations in the environment.

Figure 3.22 illustrates the spatial regions for which each different microphone produces the best results (the largest SRSR) results based on Equations 3.49, 3.50, and 3.51. If the sound source is in the gray area at the top left corner of Figure 3.22, microphone pairs [A,B] and [A,C] provide the highest SRSR, the black area is indicative
of the region for which microphone pairs [B,A] and [B,C] are better, and the remaining area, which is white, corresponds to the regions of superiority of microphone pairs [C,A] and [C,B]. These results assume that only the top wall (drawn in dark black) reflects sound.

![Diagram of microphone pairs]

Figure 3.22 – The relation between minimized SRSR and microphone selection. These results are based on the assumption that only the darkened wall reflects sound.

Figure 3.22 indicates that for the example reflection case, the optimal microphone pairs are [B,C] and [A,C], since they perform better for the majority of the environment. For reflections due to the opposite wall of the optimal microphone pairs are [A,B] and [A,C]. Hence, these results suggest that neither of the three possible microphone pairs are redundant and in different environments, based on the placement of walls, direction of the speaker, and placement of the speaker, two of the three pairs will be optimal.

This hypothesis was experimentally tested and the results are illustrated in Figure 3.23.
Figure 3.23 – The dominant cross correlation peak for various microphone structures

This figure illustrates the results of the analysis of the effectiveness of different microphone pair combinations in different situations. For the test environment depicted in Figure 3.10, the best dual microphone pair combination was the [A,B] and [B,C] combination. However, when incorporating the results of all three possible pairs and selecting the best 2 out of three, the success rate almost doubles. As a result, throughout this implementation, all three microphone pairs were utilized in the sound localization process and their result were used in the production of the SPMs. The reason for this process is that when all three pairs are incorporated into an SPM, the probability that at least two will coincide to produce an intensity peak at the location of the sound source is high. This, in effect, automatically selects the best two out of three microphone pairs and uses their results for a correct localization.
3.8 Analysis of the Performance of the Sound Localization System

Overall, the sound localization system performed well as long as the speech was directed towards the microphone array and as long as the background sounds were less intense than the sound produced by the main speaker. Cases in which these measures were not observed, the system was still able to locate the main speaker, but the number of iterations required was much greater.

3.8.1 The effects of SNR on system performance

Figure 3.24 illustrates the relation between the required number of iterations and the signal to noise ratio. These results consist of the average of ten trials per data point. From these results it can be observed that for low SNR values there is a linear dependence between the required number of iterations and the SNR, but for medium and high SNR values, the curve experiences a rapid decrease in the required number of iterations.
Figure 3.24 illustrates the ability of the system to localize the main sound source in the presence of intense background sounds, with SNRs of as little as 0 dB. These results were obtained with the speaker directly facing the microphone array. Figures 3.25 and 3.26 show the SPMs for a high SNR and low SNR situation, respectively.

Figure 3.24 – The relation between SNR and required number of iterations

Figure 3.25 – A correctly localized sound source with only 20 iterations, SNR=8 dB
Figure 3.25 illustrates a correctly localized speaker with a signal to background signal ratio of 8 dB and with the speech directed towards the microphone array. In this case, only 20 iterations where required to produce the final SPM. Figure 3.25a illustrates the final sound localization SPM and Figure 3.25b illustrates the selected peak of the SPM, which correctly identifies the main speaker.

![Figure 3.25](image)

Figure 3.26 – A correctly localized sound source but 140 iterations, SNR=2 dB

Figure 3.26 illustrates sound localization in the presence of a strong background sound which resulted in a signal to background sound ratio of 2 dB. As a result, the localization process takes 140 iterations to complete and the final SPM in Figure 3.26a does not have a clear peak when compared to Figure 3.25a. Figure 3.26b illustrates the correctly identified sound source that corresponds to the peak of Figure 3.26a.

3.8.2 The effects of reflections on system performance

Figures 3.27 and 3.28 illustrate the effects of reflection on the localization process for the case of a single speaker standing in the middle of the room. As shown in Figure 3.27, the required number of iterations to obtain a correct position doubles with a deflection of $20^\circ$ from $0^\circ$ from the horizontal axis. However, there is an extensive jump in number of iterations as the horizontal deflection angle surpasses $45^\circ$. 

68
In Figure 3.28, the effects of vertical angle of speech are analyzed. While in all deviations between 0° to 45° the required number of iterations does not change much, for negative deflection angles (i.e. when speaking towards the floor) the required number of iterations increases significantly.
3.8.3 Analysis of Sound Localization Accuracy

Small errors in the ITD estimate produce large errors in the sound localization accuracy, as illustrated in Figures 3.29 and 3.30. ITD errors can usually be attributed to speaker movement, sound source aperture size, and the sound signal sampling rate. Sampling incurs an ITD error of +/- 0.5 samples.

![Figure 3.29](image1)

Figure 3.29 - Expected localization error for six different positions with an ITD error of a) +/-0.5 samples, b) +/-1.0 samples, and c) +/-1.5 samples

Figure 3.29 illustrates the relation between the localized area and the ITD error for six different spatial positions. A better illustration of the spatial area error due to an ITD error is shown in Figure 3.30, where the intensity of each point is representative of a larger localization area or greater spatial area error.

![Figure 3.30](image2)

Figure 3.30 - Localization error depicted as the intensity of each spatial position with an ITD error of a) +/-0.5 samples, b) +/-1.0 samples, and c) +/-1.5 samples

For the implemented system, error patterns in the experimental results were similar to Figure 3.29. Figure 3.31 illustrates the localization errors with a 10 dB SNR.
In comparison to the estimated error regions of Figure 3.29, the results of Figure 3.31 roughly correspond with a +/- 0.5 to +/- 1.0 sample system error.

The sound localization results illustrated in Figure 3.31 are obtained by overlaying the results of several different trials on a single spatial map. In order to differentiate a location that is selected many times and a location that occurs only a few times, the overlaying procedure has been designed to adjust the intensity of each location according to its frequency. Brighter locations are indicative of more frequent localizations while darker spots are indicative of occasional localizations.

![Figure 3.31 - Overlapped results of several different successful localization experiments with the presence of a 10dB SNR](image)

As shown in Figure 3.32, the localization area error does not significantly change with SNR changes, however, in low SNR situations, the system sometimes mistakes the background noise sources as the main speaker, and hence the erroneous source localizations to the right of the images illustrated in Figure 3.32 d) and e) are obtained.

![Figure 3.32 - Error regions with different SNR](image)
Figure 3.32 – The change in localization accuracy in the presence of a) a 10 dB SNR, b) a 7 dB SNR, c) a 3 dB SNR, d) a 2 dB SNR, and e) a 0.5 dB SNR

In order to assess the accuracy of the above localization results, a localization variation analysis was performed. This analysis involves the averaging of the distances between each non-zero intensity pixel and the true sound source location, as illustrated below:

\[
\text{Error} = \frac{\sum I_i \cdot \sqrt{(x_i - s_x)^2 + (y_i - s_y)^2}}{\sum I_i}
\]  

(3.52)

Where \( I_i \) is the frequency of occurrence of the \( i \)th localization result\(^5\), \((x_i, y_i)\) is the coordinate of the corresponding result, and \((s_x, s_y)\) is the true location of the sound source.

The error results which correspond to Figure 3.32 are shown below:

Figure 3.33 – The relation between the intensity of background sounds and localization error

\(^5\) It should be noted that the localization result frequency is proportional to the intensity of the spatial plots of the results.
Figure 3.33 illustrates the relation between the localization error and SNR. As expected, errors are high at the low SNR because the system mistakes the background sound as the speaker. In order to analyze this point further, another series of localization results are illustrated in Figure 3.34 and the associated errors are shown in Figure 3.35.

![Figure 3.34 - Overlapped localization results for situations with a) a 10 dB SNR, b) a 3 dB SNR, c) a 2 dB SNR, and d) a 0.5 dB SNR](image)

![Figure 3.35 - The relation between the intensity of background sounds and localization error](image)
Similar to the case depicted in Figure 3.33, as the SNR drops the system remains accurate until the SNR reaches 3 dB below which the system accuracy is greatly reduced. This is again associated with the erroneous localizations that can be found in Figure 3.34 where the system occasionally mistakes the noise source to be the true sound source.

3.9 Comparison of the ISP algorithm with previous approaches

The sound localization system proposed in this thesis differs from previously reported systems ([5], [13], [14], [15]) in two main ways. First, in our system, an iterative procedure of dynamic length is used to ensure the validity of the results, and second, a spatial probability mapping technique which would allow the system to be immune to high intensity background sounds and secondary sound sources is used. The developed spatial probability mapping technique also has the advantage of not limiting the number or location of microphones. However, the latter point has not been explored in this thesis.

Table 3.1 summarizes the performance characteristics of the different sound localization systems. Note that the ISP characteristics in Table 3.1 have been provided for a SNR of 3dB. However, the ISP system can function in 0-3 dB SNRs with a reduced localization rate.
<table>
<thead>
<tr>
<th>Sound Localization Algorithm</th>
<th>SNR (average)</th>
<th>Highest Localization Rate (Hz)</th>
<th>Correct Localization Probability</th>
<th># of Microphones</th>
<th>Analysis Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning-Based [13]</td>
<td>&gt;20</td>
<td>2</td>
<td>N/A</td>
<td>4</td>
<td>Sample Based</td>
</tr>
<tr>
<td>Frequency Decomposition [19]</td>
<td>&gt;18</td>
<td>77</td>
<td>0.43</td>
<td>10</td>
<td>Sub-Sample</td>
</tr>
<tr>
<td>Filtered Cross Correlation [5]</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
<td>8</td>
<td>Sample-Based</td>
</tr>
<tr>
<td>Iterative Spatial Probability</td>
<td>&gt;3</td>
<td>16</td>
<td>0.99</td>
<td>3</td>
<td>Sample-Based</td>
</tr>
</tbody>
</table>

Table 3.1 – Comparison of different sound localization implementations

In comparison with the learning approach of Section 3.3.1 ([13]), the ISP algorithm has the advantage of a shortened configuration time and reduced computer resource requirement. Furthermore, the ISP-based system is capable of functioning in a much lower SNR than the learning algorithm and it has a higher peak localization rate. Both approaches estimate the ITD to the nearest sample number. As will be discussed in Section 3.9.3, this per-sample estimation results in an increase in the ITD error.

The frequency decomposition algorithm of Section 3.3.2 [19] performs much better than the learning-based sound localization algorithm. It is capable of operating in SNRs around 18 dB and has an extremely fast localization rate, which is approximately 77 localizations per second. Due to the nature of the frequency decomposition algorithm, the final ITD results are sub-sampled, meaning that they are much more accurate than the sample-based estimation techniques that are employed by both the learning algorithm and the ISP algorithm.
However, while the ISP-based system does not localize as fast as the frequency decomposition approach, it has a much higher success probability (0.99 compared to 0.43). Furthermore, the ISP-based system can function correctly in the low SNR situations\textsuperscript{6}, whereas for the frequency decomposition case no analysis was conducted at those SNRs. Finally, some consideration should be given to the fact that the ISP-based system obtains its performance characteristic with only 3 microphones instead of the 10 microphones used for the frequency decomposition system.

The filtered cross correlation system of Section 3.3.3 [5] in many ways resembles the initial steps of the ISP algorithm. In both cases, sub-optimal filtered cross correlations are used as means of ITD estimation. However, while no information regarding the correct localization probability or average SNR of this system could be found, it is known that it runs at a constant 2 localizations per second compared to the 16 localizations per second of the ISP algorithm.

3.10 Analysis of System Parameters

3.10.1 Microphone Placement and Structure

One of the main attributes of any sound localization system is the structure and placement of the microphones. There have been numerous studies in the past involving the structure of the microphone arrays ([19],[24]) and the optimality of each different

\textsuperscript{6} In this context, low SNR situations correspond to those where the SNR is less than 5 dB.
configuration. In general, larger microphone distances yield lower localization errors at the cost of a decrease in success ratio. This is due to the fact that large inter-microphone distances make the system more reflection and noise prone since the sound signals observed by one array element could be very different from those observed by another array element. In systems with small inter-microphone distances, the sound signals observed would be roughly the same for all the elements, except an ITD induced time shift. Hence the correlation between the result of each element would be greater.

Based on some of the previous microphone placements ([19], [24]) and several experiments conducted by the author, the microphone array was placed in a linear fashion with 0.5m gaps between the array elements. However, more detailed placement analysis is required in order to find the optimum microphone placements.

3.10.2 Sound Window Duration

Another factor that significantly affects the success of the algorithm is the number of samples or duration of each sampling window whose results are cross correlated and appended to the histogram. If this duration is too small, then it unfairly biases the histogram towards an ITD of 0 or close to it since for higher ITDs there will not be many points to correlate. If the duration of the window is too large, then it would take too long to acquire the necessary samples and go through the required iterations. Also, it becomes possible with large window lengths for the speaker’s movements to affect the sound localization results. A window length of 12.5 ms was used in the compilation of these
results and was found to be adequate in terms of localization duration, accuracy, and immunity to source movement.

3.10.3 The effects of sampling rate on localization results

Generally, the human voice occupies the frequency spectrum from approximately 30 Hz to 5 kHz. Hence, a sampling rate of 10 kHz would theoretically be enough to ensure no aliasing occurs and that the sound signals are properly captured. However, several sound localization techniques, including the ISP algorithm, employ ITD estimation techniques to the nearest sample. For these approaches, the granularity of the samples determines the ITD estimation accuracy. Figure 3.32 illustrates the situation where the peak of the cross correlation curve occurs at an index of 1.5, but since the cross correlation is analyzed at every sample, the peak is selected to have an index of 1, thereby resulting in an ITD error of +/- 0.5 samples [23]. In general, per-sample ITD estimation techniques result in a +/- 0.5 sample ITD error. This error can be reduced by increasing the sampling rate which results in a decrease in the sampling period. As a result, most per-sample sound localization algorithms use ADC sampling rates higher than the minimum 10 kHz, usually in the range from 16 to 24 kHz ([5], [13]). For the ISP algorithm, a sampling rate of 24 kHz was used.
3.11 Summary

In this chapter, the basic theoretical principles and algorithms behind sound localization were discussed. Based on these algorithms, an iterative spatial probability (ISP) based sound localization system was proposed and implemented. The ISP algorithm uses an iterative approach in which measurements are repeated until a desired accuracy is achieved. This leads to good performance in situations with low signal to noise ratios. The ISP algorithm also involves conversion of the cross correlation histograms to a spatial probability map where the intensity of each location is related to the probability of the existence of a sound source at that location. Figure 3.37 illustrates the complete set of steps involved in the sound localization process.
In the latter sections of this chapter, the localization accuracy and performance of the system were analyzed and the system was compared to some of the previous approaches.
4.1 Introduction

Sound localization, vision, and a variety of other senses can be utilized to obtain the location of an object. However, the combination of senses can be a more robust and accurate method of obtaining such information. Sense integration is an ability that is common to biological awareness systems and is effective because it allows the perception systems to be applicable in a greater number of situations than would be possible with a single sense alone.

In each of the previous two chapters, object localization using specific senses was discussed. In this chapter, the system developed in Chapter 3 is integrated with a single-camera object localization system in order to analyze the benefits and problems that are associated with multi-modal perception. The single-camera system that is employed here utilizes the algorithm that was introduced in Chapter 2 with the exception that Intelligent False Object Removal is not used. The reason for this is that with one camera the possibility of the formation of false objects does not exist.
The reason behind using only a single camera instead of the multi-camera system described in Chapter 2 was primarily that the benefits of multi-modal perception could be satisfactorily observed with a single camera. Systems with a higher number of cameras do possess several advantages including coverage of a larger environment and a higher localization accuracy. However, these advantages are not closely related to the integration of multiple senses and hence a single-camera vision system was utilized for the analysis of the integration of vision and sound localization.

4.2 Previous Multi-modal Perception Systems

There have been many attempts in the past to integrate multiple senses as part of an artificial awareness system ([5], [7], [8]). In the system described by [5], sound localization using an array of microphones was used to steer a single camera to point to the source of sound. The sound localization portion of the system discussed in [5], which was explained in detail in Chapter 3, consisted of an array of 8 microphones which localized sound with the filtered cross correlation approach in 3 dimensions. Based on this estimate, a camera that was placed on a mechanical arm would aim at the sound source position and as a result the image of the speaker would be obtained by the camera.

Such a system can be readily used as an automated teleconferencing system. One of the major drawbacks of this implementation was the fact that the camera did not participate in the localization of objects. It was used simply to take consecutive images of the sound source after it had been localized. Also, the system was not tested in
situations with low SNR, hence its performance characteristics in those situations remains unknown.

4.3 Description of Proposed System

In order to implement the integrated vision and sound localization (IVSL) system, two requirements have to be satisfied. First, the control system responsible for the sound localization has to be dedicated to the sound-sampling algorithm in order to assure that no samples are missed. If this requirement is not achieved, the existence of sample "gaps" which distort the result becomes possible. A simple way for this requirement to be achieved is for there to not be any context switching while the program is executing.

Also, the single-camera vision system has to be implemented in Microsoft Windows 95 due to the fact that the available camera interface programs are provided for this operating system only and also because multiple thread execution is required. The IVSL routines also require multiple thread execution since they consist of image decomposition, sound localization and vision integration, and visual report generation programs.

With the above requirements in mind, the IVSL system was implemented on two processing systems; one dedicated to the sound localization system and the other dedicated to the concurrent execution of an image capture routine, a sound and vision merging routine, an image decomposition routine, and a visual report generation...
program. The processing systems are linked with a serial RS232C connection interface which allows the sound localization system (processor 1) to transfer location information to the vision and control system (processor 2) with a Message Passing protocol ([25]). An added benefit of this implementation is the fact that most components are standard off-the-shelf components, which minimizes the cost and complexity of the IVSL system.

The main component of the IVSL system is the sound location algorithm that initiates the detection process, as illustrated in Figure 4.1. The sound localization program continuously obtains new sound samples, computes the corresponding ITD peak histograms, and transmits the histograms to the second processor. These first three steps are the only ones executed by the first processor.

![Diagram](image)

Figure 4.1 – The steps involved in the IVSL program – Note that the first computer starts at step 1 while the second computer starts at step 4.
Concurrent to the sound localization process, scene images are obtained from a single camera and are processed (illustrated by steps 4, 5, and 6). The processing involves the detection of all objects in the image by background segmentation ([11]). Due to the nature of the camera system, the image capture and processing programs have to be executed concurrently and independently. After the processing stage (Step 6), the processing thread blocks itself until a message from the sound localization processor is received (Step 7). This message contains the contents of the ITD histograms that allow the vision processing thread to produce the joint Spatial Probability Map (SPM) of the environment (Step 8). The details of the joint SPM will be discussed in section 4.4.

After the joint SPM is obtained, its peak identifies the location of the speaker and based on a 2-D spatial transformation to image coordinates, the most likely object in the scene image corresponding to the speaker is selected (Step 9). Finally, all IVSL results such as the joint SPM, scene image, speaker location in 2-D spatial coordinates, and identified speaker image are displayed on IVSL status windows (Step 10).

4.4 Integrating the results of sound localization and vision

The integration of the sound localization and vision results are eased by the fact that both senses utilize SPMs as a means of combining multiple-camera or multiple-microphone pair information streams. As a result, the integration of sound localization and vision consists of summing their individual SPMs. This method is effective irrespective of the number of microphone pairs or cameras.
4.4.1 Description of the Integration Process

The sum of the intensities of the sound localization SPM and vision SPM can be states as:

\[ J(x,y) = \beta \cdot L(x,y) + I(x,y) \]  \hspace{1cm} (4.1)

Where \( J(x,y) \) is the joint SPM of the environment, \( L(x,y) \) is the vision SPM, \( I(x,y) \) is the sound localization SPM, and \( \beta \) is an appropriate weight factor. The \( \beta \) factor is used to weigh the vision SPM differently than the sound localization SPM. In our experiments, it was discovered that any high value for \( \beta \) which would weigh the vision results slightly more than the sound localization results would be satisfactory. The relationship between multiple-source and multiple-sense results can be derived from equations 2.8 and 3.38 as shown below:

\[ J(x,y) = \beta \cdot \sum_{i} Q_i(x,y) + \sum_{b} G_b(x,y) \]  \hspace{1cm} (4.2)

Where \( Q_i(x,y) \) is the vision trace due to the \( i \)th camera, and \( G_b(x,y) \) is the sound localization SPM layer due to the \( b \)th microphone pair. What is important in Equation 4.2 is that the joint SPM is independent of the sequence in which the SPM layers are added. Hence, the joint SPM equation can be generalized to the following:

\[ J(x,y) = \sum_{i} f_i \cdot V_i(x,y) \]  \hspace{1cm} (4.3)

Where \( V_i(x,y) \) is the SPM for the \( i \)th layer (it should be noted that this can be an SPM for any type of sense) and \( f_i \) is its corresponding weight coefficient. For the IVSL system, \( f_i = 1 \) for cases where the \( i \)th layer is a sound localization SPM and \( f_i = \beta \) for cases where the \( i \)th layer is a vision SPM.
Figure 4.2 illustrates the processing steps undertaken by the IVSL system. The background and updated scene images are depicted in Figure 4.2a and 4.2b respectively. Figure 4.2c illustrates the SPM obtained from the sound localization system and 4.2d illustrates the integration of the sound localization SPM and vision SPM. Finally Figure 4.2e illustrates the localization of the speaker and f) illustrates the image of the object responsible for the production of sound based on he sound localization results.

4.4.2 IVSL performance in practical situations

Figure 4.3 illustrates the localization of an environment with the presence of two individuals of which one was the active speaker. In this case, the camera observes two objects (Figure 4.3b) and after the integration of the vision and sound localization SPMs (Figure 4.3d), the speaker position is located (Figure 4.3e) and a portion of his image is highlighted in the original camera image (Figure 4.3f).
Figures 4.4 and 4.5 illustrate the situation where two objects simultaneously speak in the environment. In Figure 4.4, the person near the center of the environment speaks for a longer time than the second person, which speaks only occasionally. However, as a result of the occasional secondary speech, the sound localization SPM illustrated in Figure 4.4c is not as definitive as that of Figure 4.3c, although any ambiguity is removed with the aid of the vision results, as illustrated in Figure 4.4d.

![Figure 4.4 - The results of the IVSL system in the presence of two speaking individuals](image)

Figure 4.4 – The results of the IVSL system in the presence of two speaking individuals

Figure 4.5 illustrates another example of the multiple-object multiple-speaker situation, except here the main and secondary speakers speak for the entire duration of the localization process. The reason for the increase in clarity of the sound localization SPM in Figure 4.5c is that the intensity of the correctly localized main speaker is much higher than the second individual.

![Figure 4.5 - The results of the IVSL system with the presence of two speaking individuals](image)

Figure 4.5 – The results of the IVSL system with the presence of two speaking individuals

In some of the above examples, only parts of the speaker's body were selected in the final speaker selection image. This is due to the background segmentation algorithm.
where objects are identified after the system perceives a difference in intensity between several neighboring pixels. Sometimes, parts of objects correspond in intensity with the background image and hence those sections are not identified by the imaging system. Occasionally, this has the effect of breaking objects in two and in the localization process, only the more probable of the two sub-parts is selected and boxed in the speaker selection image.

### 4.5 Effect of Integration on Localization Accuracy

In the IVSL environment, the localization accuracy of the sound localization system was much greater than that of the vision system because of the fact that typical objects such as individuals cannot be pinpointed like sound. Hence, the addition of vision to a sound localization system should not increase localization accuracy. However, vision is unaffected in the presence of several speakers and as a result the presence of background sound sources which can confuse the sound localization system is less of a problem with the integrated IVSL system. As a result, all errors that were attributed to the formation of the false localization results illustrated by Figures 3.28 and 3.30 should be removed with the addition of the vision results.

An experiment was conducted with the speaker and background sounds situated exactly as those in Figure 3.32. Based on the localization results, the overall accuracy was computed according to Equation 3.52 in the presence of four different SNR values. The final localization accuracy results of the IVSL system, which are depicted in Figure 4.6, are plotted alongside the results of Figure 3.33. As can be seen, the localization
accuracy is increased in all cases, especially at the low SNR situations where the sound localization system by itself would occasionally mistake the background sound source as the main speaker resulting in a sudden increase in localization error. With the addition of the vision sense, the sound localization system is no longer capable of detecting the background sound source and hence the accuracy is greatly increased.

![Localization Error vs. SNR](image)

Figure 4.6 – The relation of IVSL localization accuracy to SNR

4.6 Comparison of the IVSL with Previous Systems

The system implemented in [5] and the IVSL system have many similarities. They have an identical end goal, which is to obtain the image of the speaker in an environment. The means by which this goal is reached, however, are quite different. First, the IVSL system benefits from both sound and vision. This means that in cases
where sound localization is not able to correctly locate the speaker, the vision system can aid the localization process, as illustrated in Figure 4.6. Also, the IVSL system was tested in a variety of SNRs, while the implementation in [5] was tested with a single SNR.

Another difference is that the IVSL system does not require a camera aiming procedure. Unlike the implementation in [5], the IVSL has a fixed camera pointing to the environment. The image of the speaker is a subset of the image obtained by this camera.

Overall, the IVSL system offers superior functionality and robustness to background sounds and noises. In terms of accuracy at high SNRs, both of the two systems being compared use a per-sample (refer to Chapter 3) analysis which means that the accuracy at these SNRs is roughly equivalent. At low SNRs, however, the IVSL system can consistently locate the speaker, in contrast to the system of [5].

4.7 Summary

This Chapter described the process of integration of a single camera vision system with the results of a sound localization system. This process involves the weighted addition of the spatial probability maps of each sense. Such a method enabled the results of different senses to be combined in a scalable and computationally simple manner.
The implementation of an integrated vision and sound localization system was also performed and discussed. Overall, the integration of the two senses resulted in a reduction of the localization error and in increased performance. This was more evident in low SNR situations, where the sound localization system’s performance was degraded.

The advantage of the IVSL is that when a specific sense fails to locate an object, the other sense can aid in the localization process. For example, cases with low SNR where speaker localization was not possible with just the sound localization system were correctly performed by the IVSL system. This advantage arises from the fact that, in the IVSL system, the senses of vision and sound localization are completely integrated through the mechanism of the SPM, such that the final decision uses all available information. This is in contrast to most previous systems where each sense operates mainly on its own and integration takes place after all the results of the individual senses have been obtained.
CHAPTER Conclusion

5

5.1 Summary and Conclusion

The goal of this thesis was to design the object localization component of an artificial awareness system. The objective was to design object localization components and algorithms that allow different senses, such as sound and vision, to be integrated into a single system easily and efficiently.

In Chapter 2, a multi-camera object localization algorithm was proposed. This system consisted of the extraction of basic object information from images and the transformation of this information into spatial probability maps. The main problem of such a transformation, namely the formation of false spatial objects, was resolved with the introduction of IFOR, the Intelligent False Object Removal algorithm. Also discussed was the notion of global SPM peak selection that significantly reduced the computational requirements of the IFOR algorithm.
In Chapter 3 the theory of sound localization was discussed and an iterative sound localization algorithm was proposed. Unlike other approaches in the sound localization field, the proposed algorithm allowed the number of times that cross correlation analyses took place in a single localization to vary, resulting in a system that was shown to be robust in the presence of extensive environment reflections and background sounds. The information received from the microphone pairs is analyzed with the aid of the same spatial probability maps proposed in Chapter 2. The effectiveness of the sound localization system in terms of immunity to reflections and background sounds, consistency, and accuracy were analyzed.

Chapter 4 focused on the integration of a single-camera system and the sound localization system. It discussed the integration of the two senses (vision and sound localization) and illustrated the effectiveness of spatial probability maps in handling multiple-sensor and multiple-source situations. The effect of sense integration on localization accuracy was also analyzed, and it was shown that the merging of senses could greatly benefit the overall results of the awareness system. For example, when the sound localization system performance was degraded in low SNR situations, the information from the camera removed all ambiguities and kept the localization error at a relatively constant low level.

In conclusion, the integrated sound localization and vision system demonstrated the ability of multiple-sensor and multiple-source systems to provide artificial awareness in an effective and accurate manner.
5.2 Directions for Future Work

There are several aspects of the developed system that can be extended in future research projects. These include the implementation of a multiple-camera vision system, the extension of the sound localization system, and the enhancement of the IVSL system.

5.2.1 Implementation of a multiple-camera vision system

In Chapter 2, a multiple-camera object localization algorithm was proposed. While parts of this algorithm were employed in the IVSL implementation described in Chapter 4, the full multiple-camera algorithm requires further analysis with the aid of a real-time implementation. The relation between the number of cameras and the accuracy and robustness of the localization system, especially, is a very viable future extension.

5.2.2 Sound Localization Extensions

The developed sound localization system illustrated the ability of an iterative spatial probability algorithm to localize sound in a practical setting with strong background sounds and the possibility of sound reflections. One aspect of this approach that was not analyzed was the application of spatial probability maps to the detection of multiple speakers. In such an analysis, careful consideration should be given to the manner in which cross correlations from different iterations are combined into a single database.
Another direction of extension of the sound localization system is the addition of more microphones in different configurations. Since the location and placement of microphones does not affect the complexity of the ISP algorithm, a thorough analysis of the relation between localization accuracy and the number, location, and arrangement of the microphone arrays using the ISP algorithm may prove to be very valuable.

5.2.3 Enhancement of the IVSL System

The Integrated Vision and Sound Localization was able to accurately localize objects. Using this object localization system, more awareness sub-systems can be implemented. The developed sense integration approach that is defined in Equation 4.3 can be applied not only to sound localization and vision modules but to other types of senses such as motion detectors and proximity sensors.

Other viable directions of future extension for the IVSL system include the addition of self-configuration and location-based sound extraction. Since the system is capable of localizing sound sources, each awareness module which includes all microphones and cameras, can emit a sound based on which they would be recognized. This introduces a framework for the automatic localization of all modules if an initial sound localization system is present. This would greatly simplify the configuration of the artificial awareness system. Also, after a sound source has been localized, the sound emitted from that source may be extracted from other sounds thereby allowing the awareness system to 'listen' to a specific speaker or conversation. This is a task that is commonly performed by humans in crowded environments.
References


[12] Calculus, One and Several Variables, S. Salas and E. Hille, Wiley, 1990


