Active Binaural Sound Localization Techniques, Experiments and Comparisons

Greg L. Reid

A thesis submitted to the Faculty of Graduate Studies
in partial fulfilment of the requirements
for the degree of

Master of Science

Graduate Programme in Computer Science
York University
North York, Ontario

March 25, 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.

0-612-39225-2
Active Binaural Sound Localization Techniques, Experiments and Comparisons

by Gregory L. Reid

a thesis submitted to the Faculty of Graduate Studies of York University in partial fulfillment of the requirements for the degree of

Master of Science

Permission has been granted to the Library of York University to lend or sell copies of this thesis, to the National Library of Canada to microfilm this thesis and to lend or sell copies of the film, and to University Microfilms to publish an abstract of this thesis.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.
Abstract

Estimating the direction of arrival of sound in three dimensional space is typically performed by generalized time-delay processing on a set of signals from an array of omnidirectional microphones. This requires specialized multichannel A/D hardware. This work is motivated by the desire to only use standard two-channel audio A/D hardware and portable equipment. To estimate direction of arrival of sound, the pose of the microphones is made variable by mounting them on one or more computer-controlled pan-and-tilt units. The thesis will describe a number of approaches for solving the sound direction of arrival estimation problem within these constraints and equipment. The first uses a directional microphone with two rotational degrees of freedom. The second uses a combination of directional and omnidirectional microphones and a third uses two omnidirectional microphones on a fixed baseline which has two rotational degrees of freedom. Implementation of the third algorithm is also described along with the processing techniques used to perform experiments. Experimental results done in untreated, normally reverberant environments demonstrate the feasibility of the approach.
# Contents

1 Introduction ............................................. 1

2 Methodology and Algorithms ................................ 7
   2.1 Directional Cues .................................... 8
      2.1.1 Discrete ITD Cues .............................. 8
      2.1.2 Discrete IID Cues .............................. 13
   2.2 Sound Localization Algorithms ....................... 18
      2.2.1 Active directional microphone .................. 20
      2.2.2 Active directional and a fixed omnidirectional microphone pair .................. 22
      2.2.3 Active omnidirectional microphone pair ...... 24

3 Signal Processing for ITDs ................................ 30
   3.1 Techniques and Tests ................................ 31
   3.2 Processing for impulsive sounds .................... 38
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3 Processing for speech</td>
<td>39</td>
</tr>
<tr>
<td>4 Experimental Results</td>
<td>42</td>
</tr>
<tr>
<td>4.1 Basic Setup</td>
<td>42</td>
</tr>
<tr>
<td>4.1.1 Audio Specifications</td>
<td>43</td>
</tr>
<tr>
<td>4.1.2 Room Specifications</td>
<td>47</td>
</tr>
<tr>
<td>4.2 Impulsive Source</td>
<td>51</td>
</tr>
<tr>
<td>4.2.1 Parameter Settings</td>
<td>51</td>
</tr>
<tr>
<td>4.3 Speech Source</td>
<td>53</td>
</tr>
<tr>
<td>4.3.1 Parameter Settings</td>
<td>53</td>
</tr>
<tr>
<td>4.4 Results</td>
<td>55</td>
</tr>
<tr>
<td>5 Discussion</td>
<td>58</td>
</tr>
<tr>
<td>5.1 Future Work</td>
<td>61</td>
</tr>
<tr>
<td>A Technical Specifications</td>
<td>62</td>
</tr>
<tr>
<td>A.1 Power PC laptop</td>
<td>62</td>
</tr>
<tr>
<td>A.2 Omnidirectional tie clip microphone</td>
<td>63</td>
</tr>
<tr>
<td>A.3 Electret condenser directional microphone</td>
<td>63</td>
</tr>
<tr>
<td>A.4 Pan-Tilt Unit (PTU)</td>
<td>64</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Much of the previous work in sound source localization have used arrays of omnidirectional microphones with 'beam-forming' [6] and generalized time-delay techniques [1] or spectral estimation techniques which mimic the human auditory system [11]. These approaches often require special purpose multichannel A/D hardware which generate significant amounts of signal data and require intensive computation. The work presented here poses the question of how well the direction of arrival of a sound can be determined through the use of only one or two microphones and the simplest physical-based directional information. Sounds from a pair of microphones can easily be captured on standard A/D hardware found on most conventional computers. To solve an underconstrained problem and to obtain the best possible attenuation of background noise (interference), the microphones are mounted on computer-controlled pan-and-
Figure 1.1: Left: A directional microphone mounted on a computer controlled pan-tilt unit (PTU). Right: A pair of omnidirectional microphones spatially separated also mounted on a PTU. The PTUs allow the microphones' position and orientation to be manipulated creating an 'active' system.

tilt units (PTU) and their orientation in space can be changed at will (see figure 1.1).

The use of active microphones achieves with physics what microphone arrays must achieve with massive computation. This approach has been called "Active Audition" [9], the auditory equivalent of "Active Vision" [7], which include techniques in which cameras are mounted on PTUs and verging stereo is used for tracking visual targets. This work investigates the computational principles underlying the Active Audition approach. The objective is to develop and evaluate the performance of algorithms for source direction determination using an active audition system.

In human auditory perception, it is believed that there are three basic cues from
which most, if not all, sound localization is derived. They are summarized from [10, 2] below:

1. ITD - Interaural Time Differences are phase changes or time delays in the signals caused by the path length difference between the two ears. It is the primary horizontal cue for humans in lower frequencies (below 1KHz).

2. IID (or ILD) - Interaural Intensity (or Level) Differences are due to the directionality of the human ear and the shadowing effect of the head between the ears. This is the primary horizontal cue in higher frequencies (above 4KHz), which correspond to wavelengths smaller than the size of the ear.

3. Spectral Cues - These are an effect of a human’s presence in the sound field. The spectral characteristics of a perceived sound are affected by the presence of ones outer ears, head and torso. For humans, these cues extend our perception into the vertical plane.

Other work in detecting the location of sound events have involved such things as microphone arrays [1, 3] and neural networks [4, 11]. These techniques are well proven in many applications but are often restricted to single purpose use both by their design and hardware that they employ.

Microphone arrays are the most common apparatus used for doing sound source localization. In this method an array of microphones are placed with known spatial
distances between them. Sound direction can then be inferred from ITD cues obtained by correlating signals from pairs of microphones. To achieve localization in 3-space, the microphone array would have to consist of at least 4 microphones arranged in a nonplanar 3-dimensional orientation (illustrated on right of figure 1.2). Since 4 or more input channels are beyond the capabilities of audio hardware available from off-the-shelf personal computers, specialized A/D boards must be used along with DSP processors programmed to do the necessary computation at the hardware level. This makes these systems costly and difficult to program as useful tools such as low-level C language libraries tend to be limited and unique to each hardware implementation. In [1] and [3] large spatial separation between microphones and a larger number of microphones (16) are used to steer a camera towards a speaker in a non-anechoic conference room setting. The system performs well in typical conference rooms with good accuracy in both sound direction and location. However the system is non-portable as it depends on the placement of the microphones within the room which can occupy a fair amount of space as the microphones are as much as 0.5 m apart.

Another approach is to more closely simulate the human hearing system by focusing on spectral cues for 3-dimensional sound source localization. In [11], researchers model the spectral characteristics from which humans derive directional information. Using this model, it is argued that one can then predict such spectral cues for artificial systems and enable source localization through similar processing as humans. This is
Figure 1.2: Diagrams demonstrating previous approaches. On the left, two directional microphones are used with a trained neural network to locate a sound source (figure from [4]). On the right, an array of microphones localizes sound by analyzing the time differences.

done by means of a neural network where the system would 'learn' its spectral cues to localize sounds. However since spectral cues depend on the effect of the listener's presence in the sound field, for this technique to work effectively on a wide range of frequencies, the system would have to share the same combinations of large mass (human head) and small intricacies (earlobes) of humans. These affect low and high frequencies respectively.

Estimating the direction of a sound source from signals received at two fixed directional microphones has been addressed and tested in simulation mode only in [4]. In that work, the microphones are fixed in space, both pointing forward with a slight difference in the elevation (illustrated on left of figure 1.2). The central problem addressed is how to represent the nonlinear mapping from signal features to
source direction which is solved by an artificial neural network. The inputs to the neural network are estimates of both the ITD and the IID at a number of distinct frequencies. The measure for the ITD is the phase difference at the two microphones. The measure for the IID is the intensity ratio (in dB) at the two microphones at distinct frequencies. The conclusion from the work is that for the realistic case of noisy input during training, the accuracy of localization is tolerable only in the central region of the training space namely for a source near zero azimuth and elevation.

This work will present three alternative algorithms for performing sound source localization which use simple directional cues, one or two microphones and an active system enabling the orientation of the microphones to change. One of these algorithms has been implemented to demonstrate the feasibility of such an approach. The results are discussed and compared with results of previous work where applicable.
Chapter 2

Methodology and Algorithms

This chapter presents methods and algorithms used to perform sound source localization. First the ideas of auditory cues introduced from human perception must be established as measurable and/or computable values suitable for a computer application. There are restrictions on this information due to the transformation into the computer's discretized world. Finally three algorithms are presented which use discrete equivalents to ITD cues, IID cues and a computer-controlled movable apparatus.
Figure 2.1: The diagram on the left shows a sound source emitting towards a pair of microphones. A far field assumption is made so that the incoming signal is considered a plane wave when it strikes each of the microphones. The diagram on the right shows the exaggerated signal responses for the microphones and the directional cues obtained from them.

2.1 Directional Cues

2.1.1 Discrete ITD Cues

The Interaural Time Difference directional cues introduced in section 1 are described in psychological terms rather than mathematical ones. In general this is because the interpretation of these cues by humans are different than that of a computer. Sound travels as a continuous wave in the air arriving at each microphone at slightly different moments in time (figure 2.1). Thus the ITD normally represents the time shift, \( t \), in seconds between the arrival of the sound signal at each microphone. Sound recorded by a computer is sampled at discrete intervals in time and therefore in this case ITD represents the "time lag" to the nearest sampled interval. Since the ITD is measured
as a discrete number of samples, it can only be estimated with a resolution of one sampling interval. In two-dimensions, the ITD is related to the angle of incidence $\alpha$ and is calculated by simple geometric constraints (see figure 2.2):

\[
\sin(\alpha) = \frac{ct}{d} = \frac{cn}{fd}
\]  

(2.1)

where $c$ is the speed of sound, $n$ is the ITD measured in samples, $f$ is the sampling frequency and $d$ is the length of the baseline between the microphones.

The side effect of this result is that discrete ITD calculations will create a discrete number of discernible angles and limit the angular resolution of the algorithm. The angular resolution from one ITD value to the next is given by:

\[
\alpha = \sin^{-1}\left(\frac{c}{fd}\right)
\]  

(2.2)

and the range of discrete ITD values can be given by:

\[
\pm n = \pm \frac{fd}{c}
\]  

(2.3)

The discrete nature of these values will be a possible source of error for the algorithms that use them. It is possible to use interpolation to improve the accuracy of these cues but this was not done for this work.
Figure 2.2: On the left, the basic ITD calculation is shown graphically as explained in section 2.1.1. On the right, an artist rendition of a pair of microphones mounted on a PTU. A technique for solving sound source direction with this sort of apparatus is given in section 2.2.3.
Far-field assumption

Equation 2.1 makes the assumption that the sound source is a large enough distance away so that the direction of arrival of the sound is the same at both microphones. This is only true for a source at an infinite distance away. The actual calculation would depend on the distance to the sound source and this is not generally known.

The general case is given in figure 2.3. A computed ITD value will correspond to the path length difference between \(r_1\) and \(r_2\):

\[
ITD = \frac{(r_2 - r_1)}{c} \tag{2.4}
\]

\(r_1\) and \(r_2\) can be related to the actual angle to the sound source as taken from the centre of the baseline by using the cosine law:

\[
r_1 = \sqrt{r^2 + \left(\frac{b}{2}\right)^2 - 2r\left(\frac{b}{2}\right)\cos(\gamma)}
= \sqrt{r^2 + \frac{b^2}{4} - rb\cos(\gamma)} \tag{2.5}
\]

\[
r_2 = \sqrt{r^2 + \left(\frac{b}{2}\right)^2 - 2r\left(\frac{b}{2}\right)\cos(180 - \gamma)}
= \sqrt{r^2 + \frac{b^2}{4} + rb\cos(\gamma)} \tag{2.6}
\]
Figure 2.3: The first figure (top) illustrates the near-field situation with respect to two microphones listening to the same source. The second figure (bottom) plots the error created as a result of using a far-field approximation in near-field cases. The error is a function of both the angle to the sound source and the ratio of the distance to the source and the distance separating the two microphones.
where $\gamma$ is the complement angle to $\alpha$. The error between the actual angle and the approximate angle using the far-field assumption is given by:

$$\text{Error} = |\alpha_{\text{actual}} - \alpha_{\text{approximate}}|$$

$$= |\alpha_{\text{actual}} - \arcsin\left(\frac{r_2 - r_1}{b}\right)|$$  \hspace{1cm} (2.7)

This error is a function of both the actual source direction, $\alpha$, and the ratio of $\frac{r}{b}$. Figure 2.3 shows the effective error for a number of values of $\frac{r}{b}$ over the full range of $\alpha$. It is noted that for values of $\frac{r}{b}$ greater than 3 that this error is less than 0.1°.

### 2.1.2 Discrete IID Cues

For Interaural Intensity Difference cues, the relevant discretization comes not in the time domain but with signal intensity. Signal information captured by A/D hardware is stored as integer values in a range, $r$, dependent on the number of bits, $n$, used to encode the information:

$$r = 2^n$$  \hspace{1cm} (2.8)

Errors will arise when the signal power is either very high exceeding the range of values or very low where round off error becomes more noticeable with small changes in signal intensity.
IID information is normally calculated as ratios of sound intensities taken from two microphones. This way the exact intensity level of the sound source need not be known but is assumed in the acceptable range of the A/D conversion.

Changes in IID information are, in general, caused by two things: a) change in the environment surrounding the listener and the source and/or b) change in the orientation of the listener or source with respect to each other. In this work, the environment will be considered fixed and constant eliminating the first problem. If a fixed source is used and is considered to be in the far-field then the latter condition is only true if the listener has some directional characteristics to their hearing. Here this is accomplished with the use of directional microphones.

Directional microphones are those which exhibit an intensity response that varies with the sound wave's direction of arrival. The directivity pattern represents the intensity of the received sound signal as a function of the direction of arrival to the microphone. In general they can be specified as:

$$I_{\text{dir}} = f(\alpha)I_{\text{src}}$$ (2.9)

These patterns are typically cardioid in shape [5] and are parameterized by frequency. Figure 2.4 shows the measured 2D directivity pattern for the directional microphone used in this work. The microphone is equipped with a parabolic reflec-
Figure 2.4: The measured directivity pattern for a SONY directional microphone. Note that it is approximated by a 2nd order cardioid in the range of -90 to 90 degrees. The side lobes are not reliable, partly due to the shadowing effect of the parabolic reflector, as they correspond to direction of arrival behind the reflector.
tor (see appendix A). The side lobes of the pattern are not reliable and depend on the acoustic properties of the surrounding space and the position and orientation of the microphone within it. The main lobe within ±45° of the direction of maximum response has been experimentally determined to be stable with respect to changes in the environment surrounding the microphone.

**Far-field assumption**

Equation 2.9 makes the assumption that the sound source is a large enough distance away such that the signal energy loss due to the path length difference is considered insignificant.

The geometry is the same as the ITD case and given in figure 2.3. The path length, $r_1$ and $r_2$, are computed from equations 2.6 and 2.6. If the path length difference is taken into account then equation 2.9 is rewritten as:

$$I_1 = f_1(\alpha_1) \frac{E_{src}}{(r_1)^2}$$  \hspace{1cm} (2.10)

and

$$I_2 = f_2(\alpha_2) \frac{E_{src}}{(r_2)^2}$$  \hspace{1cm} (2.11)

where $E_{src}$ is the energy emitting from the source. The IID would then be
Figure 2.5: The top figure shows the change in $\frac{(r_2)^2}{(r_1)^2}$ that is introduced by making a far-field assumption. It is a function of both the angle to the sound source and the ratio of the distance to the source and the distance separating the two microphones.
The error introduced by making the far-field assumption depends on how \( \frac{(r_2)^2}{(r_1)^2} \) varies from the value of 1 and how \( \alpha_1 \) and \( \alpha_2 \) differ from each other. This error is a function of both the actual source direction, \( \alpha \) as measured from the centre of the baseline, and the ratio of \( \frac{r_2}{r_1} \). Figures 2.5 and 2.6 show the effect on \( \frac{(r_2)^2}{(r_1)^2} \) with a number of values of \( \frac{r_2}{r_1} \) over the full range of \( \alpha \). It is noted that for values of \( \frac{r_2}{r_1} \) greater than 8 that this error is less than 0.1.

### 2.2 Sound Localization Algorithms

The following restrictions have been proposed for solving the sound localization problem:

1. one or two microphones

2. simple physical based IID and ITD directional cues

3. an active system to change orientation of microphone(s)

Three algorithms are described that meet these restrictions and their attributes are discussed.
Figure 2.6: This figure shows a slice of 2.5 for $\frac{r}{\delta} = 8$. It shows that the ratio $\frac{(r_2)}{(r_1)^2}$ only changes 0.1 from the expected value of 1.0 if a far-field is assumed.
2.2.1 Active directional microphone

For this method, it is assumed a sound source has fixed position in space and is persistent over time. This allows enough time to change the pose of the microphone and to take more than one measurement. With enough measurements, one can fit the pattern shown in figure 2.4 and estimate the direction of arrival of sound as the orientation corresponding to the peak of the fitted pattern. Thus the task is split into two subproblems. The first is to select a number of measurements that are guaranteed to correspond to the main lobe of the pattern and are fairly high above the values at the tails. The second is to fit the pattern to them and estimate the direction of arrival of sound.

The selection problem is simplified by assuming that the direction of arrival of the source is between -70 and 70 degrees. In this range there is practically no sidelobe present and the directivity pattern is unimodal. By applying a small number of steps (2-3) of the golden section search in one dimension [8], sufficient measurements are collected on the main lobe and high enough above the tails to do the fitting.

Solving the fitting problem with the measurements is done by viewing the 2D directivity pattern as a function \( f(\phi) \) where \( \phi \) is the direction of arrival. Furthermore, assume that the set of measurements extracted is \( I = \{ I_1, I_2, \ldots, I_N \} \) corresponding to arrival angles \( \phi_1, \phi_2, \ldots, \phi_N \) respectively. The problem of fitting \( f(\phi) \) to the set \( I \) involves estimating a scale \( s \) and a translation \( \delta \) that will induce the best fit of \( f \) to
I. For a perfect fit at translation $\delta$, the following will be true:

$$\frac{f(\phi_1 - \delta)}{I_1} = \frac{f(\phi_2 - \delta)}{I_2} = \ldots = \frac{f(\phi_N - \delta)}{I_N} = s$$  \hspace{1cm} (2.13)

Therefore, one can formulate the estimation of translation $\delta$ and implicitly the scale factor $s$ as a one dimensional search for the minimum of the following function of $\delta$:

$$\min_\delta \sum_{i\neq j} \left( \frac{f(\phi_i - \delta)}{I_i} - \frac{f(\phi_j - \delta)}{I_j} \right)^2 \hspace{1cm} (2.14)$$

This method requires only that the directivity pattern of the directional microphone be known a priori, i.e. measured during a calibration stage. Extension of this approach to 3D sound direction determination involves searches in two dimensions in both azimuth and elevation.

**Notes and Comments**

This algorithm represents the simplest of the three being a modified searching technique. However the time required to perform the search limits the sound source type to something that is persistent over that interval. The algorithm is also dependent on the consistency of the microphone’s directional pattern. Little tolerance to noise or variation in signal intensity in the received signal would be allowed and will be its
greatest source of error.

2.2.2 Active directional and a fixed omnidirectional microphone pair

This method uses an active directional microphone, a fixed omnidirectional microphone and IID$s$ in the form of ratios to perform localization. The intensity of sound $I_{\text{dir}}$, at a single frequency, at the directional microphone depends on the direction of arrival $\alpha$ and is proportional to the intensity of the source $I_{\text{src}}$. Restating equation 2.9:

$$I_{\text{dir}} = f(\alpha)I_{\text{src}} \quad (2.15)$$

The intensity of sound at the omnidirectional microphone $I_{\text{omni}}$, on the other hand, is only proportional to the intensity of the source:

$$I_{\text{omni}} = kI_{\text{src}} \quad (2.16)$$

where $k$ is a constant, independent of $\alpha$. Therefore, the intensity ratio between the directional and the omnidirectional microphone at a single frequency is only dependent on the direction of arrival because the intensity of the source has been
cancelled out:

\[
\frac{I_{\text{dir}}}{I_{\text{omni}}} = \frac{f(\alpha)}{k}
\]  

(2.17)

Equivalently, the intensity ratio as a function of direction of arrival is the same as the directivity pattern of the directional microphone to within a constant scale factor. The pattern \( \frac{f(\alpha)}{k} \) must be estimated. Calibration is carried out by measuring \( \frac{I_{\text{dir}}}{I_{\text{omni}}} \) for a number of uniformly spaced directions of arrival \( \alpha \).

For a directivity pattern like the one shown in figure 2.4, given an intensity ratio greater than the level of the side lobes, two candidate directions of arrival are obtained in the 2D version of the problem, or a cone of candidate directions in the 3D version of the problem. To resolve the ambiguity arising from a single measurement, the pose of the directional microphone is changed yielding additional cones of possible directions. Unless the directivity pattern of the directional microphone is rotationally symmetric about the axis of maximum response, the cones of possible solutions will not be circular and their intersection will have to be computed numerically, as opposed to analytically.

Notes and Comments

Compared to the previous algorithm, this one’s use of a second microphone and relative intensities will produce more consistent IID values by making it less susceptible to noise and variation in the input signal. It will also require less positions to narrow
down the source which will put less restrictions of the duration of the sound however it still must be persistent. Future enhancements to this algorithm could incorporate ITD information however the effects of reflection inside the directional microphone would have to be taken into account.

This algorithm, like the previous one, remains dependent on the consistency of the IID values which it uses for computation. In the process of completing this work, efforts were made to implement these algorithms and were met with only limited success, mostly with simulated data. The primary reason for this was the lack of consistency in the IID values. It is believed that the main fault lies in the equipment available which was insufficient for this kind of application. As a result, focus was turned to ITDs as the primary source of directional information in the next algorithm.

2.2.3 Active omnidirectional microphone pair

This method uses two omnidirectional microphones forming a baseline and relies on ITD information in the form of angles of incidence to make sound source localization estimates. The intuition behind this method is the following. A single angle of incidence measurement from a single position of the microphone baseline constrains the source direction to be on a right circular cone. This cone has its vertex at a fixed reference point (for example at the midpoint of the baseline) and its axis of symmetry is the baseline itself. A single rotation of the baseline about a horizontal or vertical
axis through its midpoint yields another cone on which the source direction should lie. Figures 2.7 and 2.8 show the concept in steps.

For a single baseline position, the solution cone is defined as follows. Its vertex is the reference point (the midpoint of the baseline), its axis of symmetry is the unit vector along the baseline, and its angle \( \alpha \) between the normal of the baseline and any line of the cone that contains its vertex is given by equation 2.1. Solving for the source direction is a geometric problem of finding the intersection between two cones. More generally, if time delay measurements from more than the minimum number of baselines are obtained then this is an overdetermined problem and the solution is found satisfying a least squares criterion.
Figure 2.8: Two different poses of the baseline yield two solution cones, which intersect at two lines denoted by \( s \) and \( s' \), representing possible directions of arrival.

Consider the unknown source direction as a unit vector \( s \) with its start at the reference point and pointing towards the sound source. This vector is the unique solution and is independent of the orientation of the baseline. The following constraint on \( s \) then applies for a particular orientation \( i \) of the baseline \( b_i \) and a direction of arrival at angle \( \gamma_i \) with respect to baseline (unit) vector \( b_i \):

\[
s \cdot b_i = \cos \gamma_i
\]  
(2.18)

or equivalently,

\[
s_x b_{ix} + s_y b_{iy} + s_z b_{iz} = \cos \gamma_i = \cos(90 - \alpha_i) = \sin \alpha_i
\]

(2.19)

where quantities \( b_{ix}, b_{iy}, b_{iz} \) are the cartesian coordinates of a unit vector with azimuth
and elevation given by \((\theta_i, \phi_i)\) respectively. Azimuth and elevation are set by the user by controlling the motors of the pan-and-tilt unit. Angle \(\gamma_i\) represents the direction of arrival with respect to the selected baseline pose and is the compliment of \(\alpha\), the angle of incidence. Using equation 2.1, \(\gamma\) can therefore be computed from the measured ITD of the microphones. Combining 2.1 and 2.19 becomes:

\[
s_x \cdot \cos \theta \cdot \cos \phi + s_y \cdot \sin \theta \cdot \cos \phi + s_z \cdot \sin \phi = \sin \alpha_i = \frac{cn}{fd}
\]  
(2.20)

which is a linear equation with three unknowns, \(s_x, s_y, s_z\).

**Linear solution**

To find a unique solution for the sound source direction requires solving for the unknown variables \(s_x, s_y, s_z\). Since equation 2.20 is linear, this can be solved by obtaining three sets of equations and performing a Gauss-Jordan Elimination or another such method. However with any G-J elimination it is necessary to ensure that the equations are not inconsistent and unable to yield a unique solution. This will occur if the three corresponding \(b\) vectors to the equations are coplanar. To ensure this will not happen, orientations of the baseline must be chosen well. The simplest scheme is to alternate between changing the azimuth, \(\theta\), and elevation, \(\phi\) components of the baseline orientation. The result will be that each motion will be at a right-angle and therefore no three consecutive orientations will be coplanar.
Nonlinear solution

For each orientation of the baseline there is a different version of equation 2.19. There is also an implicit nonlinear constraint that $s_x^2 + s_y^2 + s_z^2 = 1$ since $s$ and $b$ are unit vectors. To compute $s$, a least squares approach can be used.

Rewriting equation 2.19 gives:

$$f_i(s_x, s_y, s_z) = s_x b_{ix} + s_y b_{iy} + s_z b_{iz} - c_i = 0$$  \hspace{1cm} (2.21)

where $c_i = \cos\gamma_i$ and

$$f_{\text{nonlinear}}(s_x, s_y, s_z) = s_x^2 + s_y^2 + s_z^2 - 1 = 0$$  \hspace{1cm} (2.22)

The solution can be obtained by solving the following minimization problem in $s_x$, $s_y$, and $s_z$,

$$\min_{s_x, s_y, s_z} \left( f_{\text{nonlinear}}(s_x, s_y, s_z) + \sum_i f_i(s_x, s_y, s_z) \right)$$  \hspace{1cm} (2.23)

Established algorithms can then be used to solve this using a least squares approach [8]. This approach takes an initial solution $s_0$ then using the minimization function (equation 2.23) and its derivative, calculates a residual and applies it to the solution. The process is then repeated. The algorithm will either converge onto a unique solution when the residual becomes small enough that it does not significantly

28
change the last solution in the iteration or it will diverge and terminate after some given number of iterations. Whether the algorithm finds a unique solution will therefore depend on the behaviour of the minimization function and the initial solution which is chosen. In order to assure convergence, an initial solution is required which is near the correct solution. The simplest way to ensure this is to calculate the linear solution of three equations and then use the nonlinear approach to refine the solution.

**Notes and Comments**

This algorithm is notably more complicated than the previous two both in understanding and in computation although it is still relatively simple when compared to algorithms using spectral information. The main advantage over the IID algorithms is that it no longer depends on the sound being persistent but instead it can be a varying signal, like speech, or a repeating impulse. However not unlike the other algorithms, its discretized ITDs are its largest source of error. Incorrect ITDs will incur errors which are inherently non-linear as suggested by equation 2.1. These errors were not as severe as in the IID case so greater success was found implementing this sort of algorithm although some additional signal processing was used which ideally would not be required. In future chapters this signal processing will be discussed as well as the implementation and results from the experiments.
Chapter 3

Signal Processing for ITDs

In the previous chapter, techniques for solving for sound direction of arrival from angle of incidences were discussed. Different characteristics of the sound source as well as the surrounding environment will affect the reliability of the incident angles by influencing the calculation of $n$, the ‘time lag’ between two input signal measured in samples (see section 2.1.1). Reliable angles are a necessity for the algorithms to function correctly. This chapter deals with the signal processing techniques used to ensure that reliable angles are supplied to the sound localization algorithms by making sure that reliable values of $n$ can be produced. It will also examine how these techniques are applied to impulsive sound sources and speech.

In section 2.2.3 it was demonstrated how an interaural time difference (ITD), $n$, between two microphones can be used to derive directional information. ITDs are
calculated by collecting, in stereo, some *window* of the sound event by specifying a time interval, correlating the two channels of data and examining the peak of the resulting function (see figure 3.1). If the signals do not represent a specific sound event then their resulting correlation will be meaningless (see figure 3.2). However without establishing some parameters about the sound source it is difficult to determine if a correlation function is producing useful information. In addition, the correlation function can be computationally expensive to produce. Indication that a sampled window may not include a sound event would avoid unnecessary work so methods to distinguish samples become beneficial.

### 3.1 Techniques and Tests

By making reasonable assumptions about the sound source and the environment, tests can be used to control the reliability and robustness of signal data used in estimating angle of incidence.

Combined with other signal processing techniques, they create a group of methods used to regulate directional information to the algorithms. Figure 3.3 shows a summary of the processing techniques and how they are applied to the input signals to produce reliable ITD values.

The five techniques are categorized as:
Figure 3.1: The three plots show two artificially generated signals and their resulting correlation. The signals are damped sine waves at a frequency corresponding to about 450 Hz at a 22kHz sampling rate (the sampling rate which will be used in future experiments) with some random noise added corresponding to about a 10dB S/N. The second channel has been time delayed by 10 samples to simulate the ITD effect expected in future experiments (see Chapter 4 for details). The resulting correlation has a single maximum corresponding to the correct ITD.
Figure 3.2: The three plots show two artificial random signals and their correlation. The resulting correlation shows many peaks although no meaningful ITD information is present in the signal. Note that the correlation values here are several orders of magnitude different from those seen in figure 3.1. Well chosen tests would eliminate these types of samples.
Figure 3.3: This flow chart demonstrates the processing performed on the raw recorded signals in order to retrieve reliable ITD values.
1. **Filtering Technique**: Filtering can serve to produce better ITD values in two ways. First, filtering the known background noise better increases the chance of more clearly hearing the sound event. Second, the use of a low-pass filter (LPF) centred at some frequency, \( f_t \), would eliminate sound frequencies higher than \( f_t \) that could create erroneous peaks in the correlation function due to full wavelength phase shifts. However this may not be desirable if the target sound is characterized by frequencies above \( f_t \) in which case filtering would remove most of the signal power. The choice of \( f_t \) will be based on other tests discussed later.

2. **Signal Level Test**: For a sampled window of signal to produce an acceptable estimate of the angle of incidence, it must first contain a significant portion of the sound event which is to be localized. This is done by considering the average power and absolute peak power of the sampled signal compared to an estimate of the power of the background noise. An appropriate threshold value will then differentiate between samples that contain the sound event and those that correspond to relative silence. This threshold will determine initially which windows will be kept for further analysis or discarded.

3. **Correlation Limit Technique**: Standard correlation method involves iterating the calculation over the total length of the sampled signals resulting in a
correlation function with the same length of the original signals. The correlation function, $c$, is computed by:

$$c(i) = \sum_t s_r(t + i)s_l(t)$$

where $s_r$ and $s_l$ are the right and left signal channels respectively and $t$ and $i$ are over the time interval of the input signals.

The two-dimensional ITD calculation (equation 2.1) can result, however, in only a finite number of discernible angles between the microphone baseline and the sound source. The number of time delays corresponding to these angles is generally much smaller than the length of the sampled windows. Therefore correlating only over the suitable ITD values can significantly decrease the total time needed. In addition, erroneous peaks in the correlation function can appear from full wavelength phase shifts in the sound. Limiting the length of the correlation function will eliminate these peaks below a threshold frequency since the signals are never shifted a full wavelength with respect to each other. This threshold frequency is given by:

$$f_t = f_s/n_c$$
where $f_t$ is the threshold frequency, $f_s$ is the sampling frequency of the sampled data and $n_c$ is the length of the correlation function which is the total number of discernible ITD values (see section 2.1.1).

4. **ITD Level Test**: For good directional information, the resulting correlation function from the sampled signals is expected to have strong positive peaks. The primary peak should represent the position of the ITD while secondary peaks represent full wavelength phase shifts of higher frequency components of the sound event. However due to noise, secondary peaks could become greater than the primary peak of interest and the algorithm will not be able to discern between them. To avoid this problem it is observed that the proper ITD should be represented by a maximum peak which is considerably greater than both the largest secondary peak as well as the average of the correlation function. Depending on the sound source, threshold percentages can be used to ignore erroneous correlation functions.

5. **Multiple ITD Test**: Despite all the previous tests put in place to eliminate incorrect directional information, choices in thresholds and the nature of the sound source may still allow erroneous ITD values to come from the process described this far. Since small errors in an ITD can lead to large errors in the result (from equation 2.1), a final check on the result can be performed by cal-
culating the ITD on a number of consecutive sampled windows of signal and discard the outliers by looking for the most common answer among them, commonly referred to as clustering. This will considerably hamper the performance of the overall system since it will require the sound event to be either persistent or repeating such that the desired number of windows can be recorded and processed in sequence. This may be necessary, however, if any single ITD value cannot be considered nearly 100% reliable.

3.2 Processing for impulsive sounds

Impulsive sounds are characterized by a sudden large intensity change which quickly tapers into background noise. Examples include a hand clap or a slamming door. The frequency response is typically broadband with many high frequency components. Since these events are very short in time duration, the entire signal is often captured within a sampling window. The sudden large intensity peaks makes them easily detectable by the peak-based Signal Level test described earlier. The strong signal also lends itself well to the correlation calculation, however its broadband nature will introduce more high frequency components and thus secondary peaks will appear in the correlation function (see figure 3.4). However, since much of the sound energy lies in the higher frequencies, using a low-pass filter centered at $f_t$, as suggested earlier by the filtering technique, may eliminate too much of the signal to produce.
good directional information. Instead, a high-pass filter used to eliminate system and background noise only would be more effective.

3.3 Processing for speech

In the impulsive case, tests are used to determine if the sound event occurred within the sampled data. In the case of speech, the sound event will likely have occurred over several consecutive sampling windows. Speech is comprised of many different kinds of sounds, some or all of which could appear in any given sample. Some of these sounds will be more difficult to retrieve directional information from, for example from soft sounds or whispering due to their overall lack in intensity. So in the case of speech it is desirable not only to eliminate sample data which does not contain sound, but also those which are less likely to produce reliable ITD values.

While speech typically does not have the peak strength of impulsive sounds, their longer duration over the interval will result in greater average power which can be used as a Signal Level Test. Speech also has lower frequency components than impulsive sounds creating fewer erroneous peaks and more selective filtering may be used to enhance the signal. Figure 3.5 shows an example of speech data and the typical peaks in the correlation as a function of ITD.
Figure 3.4: The three plots show two channels of sampled data from the experimental setup (see chapter 4) and their correlation. The sound recorded is of a single clap, a good example of an impulsive sound (high peaks and short duration). Since impulsive sounds have many high frequency components, the resulting correlation has several smaller peaks. The strongest peak represents the correct ITD for the given experimental setup (6 or 7 samples). The secondary peak could have been mistaken for the correct ITD had the noise acted less favourably.
Figure 3.5: The three plots show two channels of sampled data from the experimental setup (see chapter 4) and their correlation. The sound recorded is of a person reading a quote from a book. As with figure 3.4, the correct ITD for the experimental setup was between 6 and 7 samples and the correlation has its largest peak in that range. The signal has fewer and broader humps due to its fewer high frequency components however the signals peaks are not as strong as that of impulsive sounds making them more prone to error due to random noise.
Chapter 4

Experimental Results

This chapter describes an implementation of the algorithm described in section 2.2.3. It describes the equipment and room setup used along with the restrictions that they will impose on the expected results. Two sets of experiments are described in this section. The first uses an impulsive sound and the second experiment uses speech. How the techniques and tests from chapter 3 are applied to each of these cases are discussed and finally the results are presented.

4.1 Basic Setup

The experiments consist of a listening apparatus controlled by computer attempting to locate a sound source in 3-space. The sound source, played through a speaker, is placed in position and the computer is asked to estimate its position 25 times.
consecutively. The source is then moved and the process is repeated. For each source position the average and mean deviation of the multiple estimates give an indication of the accuracy and error bounds of the algorithm.

4.1.1 Audio Specifications

The listening apparatus for these experiments consists of a pair of tie-clip omnidirectional microphones mounted at either end of a wooden rod at \( d = 0.3 \) m apart forming the baseline \( b \). It is affixed atop a Pan-Tilt Unit (PTU) allowing its orientation \( (b_\theta, b_\phi) \) to be controlled by the computer. For exact specifications see Appendix A.

In order to use the far-field assumption the distance to the sound source must be chosen relative to the length of \( b \) to keep the error introduced by the assumption small (see section 2.1.1). A distance of \( r = 2m \) which corresponds to \( \frac{r}{b} = 6.66667 \). Figure 4.1 shows the error introduced by making the assumption with these parameters. It is less than 0.05° for the worst case which is acceptable for this application.

Sound collection is done in stereo through a conventional A/D sound board on a Macintosh Powerbook 520 (upgraded to a PowerPC processor) at a sampling rate of \( f = 22050Hz \) and using 16-bit resolution. Using equation 2.1 and setting the ITD value to \( n = 1 \) gives the minimum theoretical resolution that can be expected from this listening apparatus:

43
Figure 4.1: This figure show a slice of Figure 2.3 at the value of $\zeta = 6.66667$ which will be used in experiments for this work. It represents the error introduced by making a far-field assumption in the algorithm. The error for this experiment is relatively small, less than 0.05°, so the far-field assumption can be used.
\[ \alpha_{\min} = \sin^{-1}\left(\frac{c}{fd}\right) = \sin^{-1}\left(\frac{342m/s}{22050Hz \times 0.3m}\right) \approx 3^\circ \] (4.1)

The number of discernible angles can be then be calculated by determining the range of ITD values for this setup. To find this, consider a sound source at \( \alpha = 90 \) degrees to the baseline and apply it to equation 2.1:

\[ \pm n = \pm \frac{fd}{c} = \pm \frac{22050Hz \times 0.3m}{342m/s} \approx \pm 19 \] (4.2)

The value of \( n \) must be an integer since it represents the ITD as a number of finite samples. The number of possible ITDs is then the range of \([-19, 19]\) which is 39 discrete values. Equation 2.1 is used to map all the possible ITD values to their corresponding angles. The following table shows the discernible angles for \( \alpha \) which can be achieved for ITD values of 0 to 19. The table is symmetric for the negative ITD values.

As table 4.1 demonstrates, the theoretical resolution is accurate only at ITD = 1 but remains reasonably close to that value up to about ITD = 14 (45 degrees) where after it begins to diverge. Since this resolution will govern the accuracy and error bounds of the final result, it would be desirable to obtain ITD values inside the -14 to 14 range wherever possible. To accomplish this a simple algorithm can be used to steer the orientation of the microphones towards the best orientation as it takes ITD
Table 4.1: ITD values with discernable angles for the experimental setup.

<table>
<thead>
<tr>
<th>ITD</th>
<th>Angle</th>
<th>ITD</th>
<th>Angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0°</td>
<td>10</td>
<td>31.1°</td>
</tr>
<tr>
<td>1</td>
<td>3.0°</td>
<td>11</td>
<td>34.7°</td>
</tr>
<tr>
<td>2</td>
<td>5.9°</td>
<td>12</td>
<td>38.3°</td>
</tr>
<tr>
<td>3</td>
<td>8.9°</td>
<td>13</td>
<td>42.2°</td>
</tr>
<tr>
<td>2</td>
<td>11.9°</td>
<td>14</td>
<td>46.4°</td>
</tr>
<tr>
<td>5</td>
<td>15.0°</td>
<td>15</td>
<td>50.9°</td>
</tr>
<tr>
<td>6</td>
<td>18.1°</td>
<td>16</td>
<td>55.8°</td>
</tr>
<tr>
<td>7</td>
<td>21.2°</td>
<td>17</td>
<td>61.5°</td>
</tr>
<tr>
<td>8</td>
<td>24.4°</td>
<td>18</td>
<td>68.5°</td>
</tr>
<tr>
<td>9</td>
<td>27.7°</td>
<td>19</td>
<td>79.2°</td>
</tr>
</tbody>
</table>

readings. This ability to favour lower, more accurate ITD values is a big advantage over the fixed array approaches.

Pseudo-code to the algorithm used for these experiments is given in Table 4.2. It uses the last ITD measurement to determine how to change the orientation of the microphones and whether to make a relatively large or small change. A random factor is added so that the same set of orientations are not repeated in cycle and meaningful error bounds can be calculated. Movements are made in azimuth or elevation but not in both and alternate each time an orientation is changed. This is done to ensure that every three consecutive orientation vectors, $b_i$, can never be coplanar which would create an ill-conditioned solution set (see section 2.2.3). For these experiments the pseudo-code only applies the algorithm to azimuth (pan) movements. This is due to the PTU elevation (tilt) being too limited to be effective so instead two reasonable
elevations are alternated.

```c
// Choose next direction
if (ITD <= 14 && ITD >= -14)
  // Small move
  changeDirection = -1*(sign(ITD)*20.0 + 10*random);
else
  // Larger move
  changeDirection = -1*(sign(ITD)*30.0 + 20*random);

// Apply direction change alternately to pan or tilt
if (nextPan)
  panPos += changeDirection;
else
  { // Just rotate between the two
    if (tiltPos <= -30.0)
      tiltPos = -10.0;
    else
      tiltPos = -40.0;
  }
nextPan = !nextPan;
```

Table 4.2: Pseudo-Code for a simple orientation change algorithm. The term `random` refers to a function which would produce a uniformly distributed random number between 0 and 1.

### 4.1.2 Room Specifications

The environment for the following experiments is an ordinary rectangular room about 6 m in width, 7 m long and 3 m high. The centre of the room has been cleared of furniture and the listening apparatus is placed upon a cart at a distance of 2.1 m (7 ft) from the area where the sound source will be located. The room is carpeted with
standard ceiling tiles, windows, white board, book shelves and a small counter in one corner. No special treatment was made to the room, therefore the room exhibits reverberation qualities typical of a conference room. The room arrangement and exact positioning of apparatus and sound source are shown in Figure 4.2.

The experiment requires measurements to be taken with the sound source at different locations in 3-space to demonstrate the abilities on the algorithm. To accomplish this, predetermined distances from the wall and heights from the floor are mapped out in a grid. These distances were calculated by considering source positions at azimuths from $\theta = -40^\circ$ to $40^\circ$ in $10^\circ$ increments at elevation $\phi = 0^\circ$. Likewise, the heights were taken from $\phi = -30^\circ$ to $20^\circ$ in $10^\circ$ increments with $\theta = 0^\circ$. However in order to maintain equal signal levels throughout the experiment, the sound source must always be kept at a constant distance from the listening apparatus. The result is that the actual sound source azimuth $\theta$ will 'stretch' with elevations off of zero. The final planned sound source positions are shown in figure 4.3.

Results shown in future sections will follow this pattern well but not exactly. It is worth noting here that the placement of the sound source could not be measured accurately as it involved using a measuring tape and calculating distances to far walls. In particular, higher elevations were increasingly more difficult to measure the closer the sound source was to the light fixtures. In an attempt to offset this difficulty, the sound source remained in the same position for both experiments (impulsive and
Figure 4.2: The measurements for the experimental setup are shown above. The dotted arc represents possible positions of the sound source which are always kept at a constant distance from the microphone apparatus.
Figure 4.3: This graph shows the planned sound source locations to be used in experiments. Note the stretched effect on the azimuth due to elevation.
speech) before being moved to the next position. In this way, the actual position of
the sound source could be judged by the correspondence between the results of the
two experiments. This aspect will be discussed further in section 4.4.

The two different sound sources were chosen to demonstrate the algorithm’s ability
to work in either end of a range of sound types. Only the test values were different
according to the type of sound, either impulsive or speech, being localized. In the
following sections the parameter settings used in each case are discussed as well as
the results that they produced.

4.2 Impulsive Source

For the impulsive sound experiment a recording of a single clap was used and repeated
at 1 second intervals. The sound event is the same to that used in figure 3.4 although
the intensity differs.

4.2.1 Parameter Settings

Using the recommendations described in section 3.2, the following test values are
chosen for the five techniques outlined in section 3.1. They give the most consistent
ITD values used in the algorithm.
1. **Filtering** : A filter which eliminates frequencies below 200Hz is used. Much of the line noise and environmental noise such as overhead ventilation fans are in this range.

2. **Peak level** : A signal level of 30.0 was chosen given that the average background noise was less than 10.0 and initial peaks of the sound were often greater than 50.0.

3. **Correlation limit** : The correlation limit is dependent upon the geometry of the apparatus and sampling rate used in audio collection. It is therefore independent of the nature of the sound source (see section 3.1).

4. **ITD level** : The primary peak in the correlation signal as a function of ITD must be at least a specified multiple of the average of the correlation signal. It has been determined experimentally that a multiple by a factor of 20 quickly differentiates between good ITD values and bad ones which are created by correlating just the tails of the sound event. If just the tail end of the sound event is captured of an impulsive sound then its overall power level will be too low and the addition of noise will cause the probability of bad ITD values to be much larger.

5. **Multiple ITDs** : While most often the above techniques produced the correct ITD values, multiple readings were required to ensure good values were
produced. Five consecutive readings were done and the median value was considered the correct ITD.

4.3 Speech Source

The speech sample was taken from a female subject reading a short passage of text. The sound sample was then placed in a continuous loop. The same passage was also read by a male reader and similar results could be shown for each with the same set of tests.

4.3.1 Parameter Settings

Section 3.1 and 3.3 describe the tests and techniques involved in speech processing. Below are the values and reasoning behind each test which was applied for this experiment.

1. **Filtering**: A filter which eliminates frequencies below 200Hz is used. Much of the line noise and environmental noise is in this range. It was thought that a more selective band-pass filter would be more effective for this type of sound, however it failed to produce the desired improvements. Therefore the final experiments were done with the same filter as the impulsive sounds.
2. **Peak level**: A signal level of 12.0 was chosen which was just over the background noise level with a level of about 10.0. This was done because speech does not necessarily contain large peaks but a uniform increase in the overall power. A test based on the average signal power may be more appropriate but was not used here in order to have the two experiments run with the same test variables.

3. **Correlation limit**: The correlation limit is dependent upon the geometry of the apparatus and sampling rate used in audio collection. It is therefore independent of the nature of the sound source (see section 3.1).

4. **ITD level**: The primary peak of the correlation function is located and its value is considered as a ratio of the average correlation value. While speech tends to have fewer peaks in general, they are lower in magnitude and therefore result in a lower ratio to the average. A ratio of 5.0 was a reasonable value for this test.

5. **Multiple ITDs**: With all of these techniques in place, the system did not produce good ITD values as often as the impulsive case. This was expected since the speech sample was considerably closer in both power and in frequency to the background noise level in the room. In order to ensure good ITD values for the direction of arrival algorithm (section 2.2.3, seven consecutive samples
were used and their median was taken.

4.4 Results

The results for both experiments are shown in figure 4.4. Each cross is the result of 25 measurements of the sound source in a fixed location. The cross is centred at the average position of those measurements with its width and height illustrating the absolute mean deviation (AMD) of the results in azimuth and elevation respectively. In general what is seen is a somewhat smaller and better AMD in azimuth than in elevation. Table 4.3 shows the average AMD in azimuth and elevation across a single elevation as well as for the overall experiment. This shows the AMD of the estimates are typically below 4° at all positions.

<table>
<thead>
<tr>
<th>Elevation</th>
<th>Impulsive Source</th>
<th>Voice Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AMD Az</td>
<td>AMD El</td>
</tr>
<tr>
<td>20.0°</td>
<td>1.8°</td>
<td>3.5°</td>
</tr>
<tr>
<td>10.0°</td>
<td>1.8°</td>
<td>3.0°</td>
</tr>
<tr>
<td>0.0°</td>
<td>1.6°</td>
<td>2.4°</td>
</tr>
<tr>
<td>−10.0°</td>
<td>1.5°</td>
<td>2.4°</td>
</tr>
<tr>
<td>−20.0°</td>
<td>2.1°</td>
<td>2.7°</td>
</tr>
<tr>
<td>−30.0°</td>
<td>2.5°</td>
<td>2.9°</td>
</tr>
<tr>
<td>Overall</td>
<td>1.9°</td>
<td>2.8°</td>
</tr>
</tbody>
</table>

Table 4.3: This table shows the average of the AMD for all the positions at a given elevation as well as over all positions.

Figure 4.5 shows the average estimated positions for each experiment overlaid on
Figure 4.4: The graphs show the estimated sound direction of arrival in both azimuth and elevation for an impulsive source (top) and a speech source (bottom). Each position is computed from 25 determinations of the source in the same position. The diamond represents the average position, cross represents the absolute mean deviation in both azimuth and elevation.
Figure 4.5: The graph has the average estimated positions of the experimental data from figure 4.4 overlaid on top of one another. The squares and diamonds are the average estimated source locations for the impulsive and speech sources respectively. It can be noted is that although the general pattern in the same as the planned positions (figure 4.3), many of these estimates are not exactly where they were planned to be. However, as discussed in section 4.1.2 this was somewhat expected since exact sound source locations were difficult to measure. Areas where the separate experimental results tend to agree with each other are most likely poorly placed sound sources rather than bad estimates. This is the case for most of the positions.
Chapter 5

Discussion

Three algorithms for doing “Active Audition” have been introduced and their limitations and restrictions have been presented. All of these techniques employ no more than two-channel (binaural) A/D input hardware and use computer controlled PTUs to estimate direction of arrival of persistent or repeating distant sound sources.

The technique using two omnidirectional microphones has been implemented and source estimation has been demonstrated for a limited angle range with two different types of sound sources (impulsive and speech). Testing was done with conventional computer hardware and in a room with no special acoustical properties. The resulting 4° resolution in both azimuth and elevation indicates the feasibility of such an approach.

Similar work done in [1] on real test cases claim a 30 cm accuracy in a source’s
position for similar size and style rooms as used in these experiments. This corresponds to an angular resolution of 6° in the sound direction at an average 3 m distance which is comparable to the results achieved for this work. They also performed tests in larger reverberant rooms like auditoriums and found that their errors increased significantly. While no tests were performed in large rooms in this work, the same degradation would be expected since they are both based on similar ITD calculations. While the technique presented here do not allow for the exact source location to be found, an application such as ‘video steering’ would only require directional information if the camera and microphones were mounted in close proximity to each other.

Implementation of the work in [4] was only performed in simulation but still allow for some comparisons to this work. While no statistics of their results were available in the publication, an average error of less than 10 cm can be reasonably approximated from their result figures. A distance from their testing plane of 1 m puts those results very close to both [1] and the errors achieved in this study. However it should be noted that their range of errors change greatly by position and by the amount of noise added to the training sessions. At central locations, it appears to have much better accuracy but further out it is well worse than 10 cm. The system implemented for this work, in a similar azimuth range, did not show the same degradation at larger angles which was the expected result of our “Active” approach. The same benefit would be
expected for the algorithms involving active directional microphones (sections 2.2.1 and 2.2.2).

Spectral estimations used in [11], whose results are also given from simulations, show remarkably good accuracy at less than 1° for broadband sounds. However for tests involving human voice their results were relatively poor with 15° – 30° error. This limits their approach to applications only using these broadband sounds. Since their technique is based on analysis of frequencies at many levels, their result was somewhat expected. In comparison, the work implemented here maintained nearly the same error bounds for both broadband and voice sources although more signal processing was required in the voice testing to maintain reliable ITD values.

In general, it can be said that the algorithm in section 2.2.3 implemented in this work performs comparably to previous designs and has less limitations due to source direction and source type. However due to the limitations of the components used in the implementation, too much repeated signal processing was required to produce reliable ITD information. It took far too long, roughly a minute or more per estimation, to compute the sound direction of arrival to make this approach feasible for real-time applications. Reducing the time required to localize a single sound event and enabling it to perform in near real-time are the issues that have to be overcome.
5.1 Future Work

The immediate extension of this work would be to complete the implementation of the other algorithms presented here that use directional microphones and IIDs. While an attempt was made for this work, intensities proved to be unstable and therefore made it impossible to demonstrate these methods. If these issues can be addressed then the overall feasibility as well as effectiveness with different kinds of sound sources could be better evaluated. Better equipment and more sophisticated analysis should be able to overcome the problems encountered with both the IID and ITD calculations.

Other extensions would include tracking of a slowly moving sound source and discrimination of multiple sound sources. The availability of a prediction of the direction of arrival allows us to optimize the pose of the microphones to minimize error in both cases.

Applications of this work could include combining audio and visual cues for use in video conferencing, enhanced lecture recording or remote tele-operation.
Appendix A

Technical Specifications

A.1 Power PC laptop

Manufacturer: Macintosh Computers Inc.

Model: Powerbook 500 Series

Weight: ≈ 3 kg

Dimensions: 290 x 260 x 250 mm (width/height/depth)

Cost: ≈ $3000

Description: A powerPC was used to control listening apparatus. The PTU device was controlled via serial connections and the built-in stereo sound card was used as the A/D converter for our stereo sound input channels. C-libraries for both the sound card and serial interface are
public domain software.

A.2 Omnidirectional tie clip microphone

Manufacturer : Genexxa

Model : 33-3003

Weight : ≈ 20 g

Dimensions : 17.6 x 8 mm (length/diameter)

Cost : ≈ $50

Description : The omnidirectional microphones used for these experiments were simple tie-clip microphones purchased at a commercial electronics store.

A.3 Electret condenser directional microphone

Manufacturer : Sony

Model : ECM-PB1C

Weight : ≈ 20 g

Dimensions : 174 x 160 x 70 mm (width/height/depth)

Cost : ≈ $50

Description : The directional microphone tested for the work consisted

63
of an omnidirectional microphone supported at the focal point of a small parabolic reflector. This product is sold commercially as an adapter for personal video camera recorders adding directional sensitivity and gain to the audio recording.

A.4 Pan-Tilt Unit (PTU)

Manufacturer: Directed Perception Inc.

Model: PTU-46-17.5

Weight: \( \approx 2 \text{ kg} \)

Dimensions: \( 65 \times 130 \times 90 \text{ mm} \) (width/height/depth)

Cost: \( \approx $3000 \)

Description: The PTU is a mechanical device used to control orientation of the listening apparatus during the experiments. It is connected via serial cable to a computer and is controlled using C-language libraries. The libraries have been written and modified by students of other projects which include Matthew Izatt, Tom Luu, Henry Wong, Willen Straten and Greg Reid.
Bibliography


