Active Stereo Sound Localization

Greg L. Reid
Evangelos Milios

Technical Report CS-1999-09

December 15, 1999

Department of Computer Science
4700 Keele Street North York, Ontario M3J 1P3 Canada

Active Stereo Sound Localization

Greg L. Reid, Evangelos Milios
Department of Computer Science,
York University,
Toronto, Canada, M3J 1P3
eem@cs.dal.ca

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, and careful arrangement of the microphones into an array. This work is motivated by the desire to instead only use standard two-channel audio A/D hardware and portable equipment. To estimate direction of arrival of persistent sound, the pose of the microphones is made variable by mounting them on one or more computer-controlled pan-and-tilt units. In this report, we describe the signal processing and control algorithm of a device with two omnidirectional microphones on a fixed baseline and two rotational degrees of freedom. Experimental results with real data are reported with both impulsive and speech sounds in an untreated, normally reverberant indoor environment. We further discuss two more approaches, one using a directional microphone with two rotational degrees of freedom and another using a combination of a directional and an omnidirectional microphone.

1 Introduction

In human auditory perception, it is believed that there are three basic cues from which most sound localization is derived [12, 2]. Interaural Time Difference is the primary horizontal cue for humans in lower frequencies (below 1KHz). Interaural Intensity (or Level) Difference is the primary horizontal cue in higher frequencies (above 4KHz), which correspond to wavelengths smaller than the size of the ear. Spectral Cues are due to the fact that the spectral characteristics of a perceived sound are affected by the presence of ones outer ears, head and torso. Spectral cues extend our perception into the vertical plane.

Many of existing implementations of sound source localization have used arrays of omnidirectional microphones with beamforming [6] and generalized time-delay techniques [1, 3]. These approaches often require special purpose multichannel A/D hardware which generate significant amounts of signal data and require intensive computation.

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 normally reverberant conference room setting. This system performs speaker localization in three-dimensional space, not simply direction of arrival estimation. 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 using two microphones is to more closely simulate the human auditory system by focusing on spectral cues for 3-dimensional sound source localization [13] and modelling the spectral characteristics from which humans derive directional information. This is done by means of a neural network where the system would ’learn’ its spectral cues to localize sounds.

Estimating the direction of a sound source from signals received at two fixed directional microphones has been addressed and tested only in simulation mode in [4]. In that work, the microphones are fixed in space, both pointing forward with a slight difference in the elevation. 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 time delay and the intensity different at a number of distinct frequencies. The measure for the time delay is the phase difference at the two microphones. The measure for the intensity difference 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

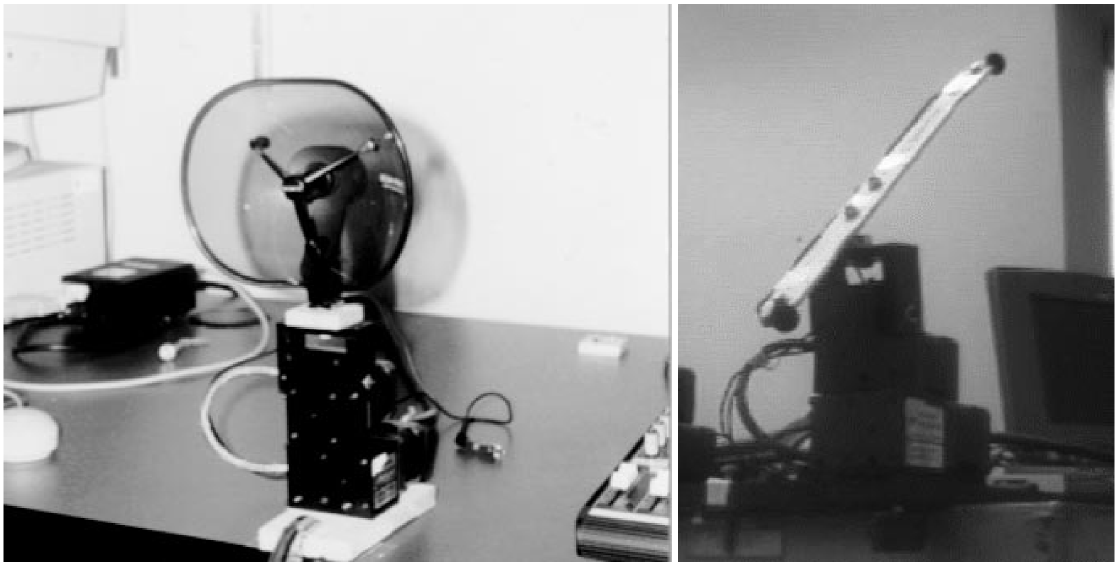


Figure 1: Left: A directional microphone mounted on a computer controlled pan-tilt unit (PTU). Right: A pair of omnidirectional microphones spatially separated also mounted on a PTU. The PTUs allow the microphones' position and orientation to be manipulated creating an 'active' system.

namely for a source near zero azimuth and elevation.

Our work attempts to replace the functionality of an array of microphones with two microphones mounted on a computer-controlled pan-tilt unit, as shown in Figure 1. The use of active microphones achieves with physical motion what microphone arrays must achieve with massive data collection and computation. We call our approach "Active Audition" [10], the auditory equivalent of "Active Vision" [8], where cameras are mounted on PTUs and verging stereo is used for tracking visual targets. Here we investigate 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.

Section 2 describes the proposed methodology and localization principle. Section 3 presents the signal processing required for estimating interaural time differences (time delays) from both impulsive and speech signal data. Section 3 presents experimental results with both impulsive and speech sounds. Section 4 discusses the feasibility and properties of the proposed method.

2 Active Sound Source Localization

We now present the geometry of the active audition approach for sound source localization. The essence of the approach is to locate the direction of arrival of sound as the intersection of two cones sharing the same vertex. We first review the relation between direction of arrival and time delay estimated at two omnidirectional microphones. Then we describe the geometry of direction of arrival estimation in three-dimensional space when the pose of the two microphones is computer-controllable.

Time delay estimation In the two-dimensional version of the direction of arrival estimation, the time delay between the signals from two omnidirectional microphones is related to the angle of incidence α and is calculated by simple geometric constraints.

$$sin(\alpha) = \frac{ct}{d} = \frac{cn}{fd} \tag{1}$$

where c is the speed of sound, t is the time delay in seconds, n is the time delay in samples, f is the sampling frequency and d is the length of the baseline between the microphones. To achieve subsample accuracy in the estimation of time delay t, it is possible to perform quadratic interpolation on the three correlation values at $n-1, n, n+1$ centred at the maximum of the correlation n, but this was not done in this work.

Far-field assumption Equation 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 general case is given in figure 2. Time delay corresponds to the path length difference between r_1 and r_2 :

$$TimeDelay = (r_2 - r_1)/c \tag{2}$$

where r_1 and r_2 are related to the angle γ towards the sound source as taken from the centre of the baseline by using the cosine law :

$$\begin{aligned} r_1 &= \sqrt{r^2 + (\frac{b}{2})^2 - 2r(\frac{b}{2})\cos(\gamma)} \\ &= \sqrt{r^2 + \frac{b^2}{4} - rb\cos(\gamma)} \end{aligned} \tag{3}$$

$$r_2 = \sqrt{r^2 + (\frac{b}{2})^2 - 2r(\frac{b}{2})\cos(\pi - \gamma)}$$

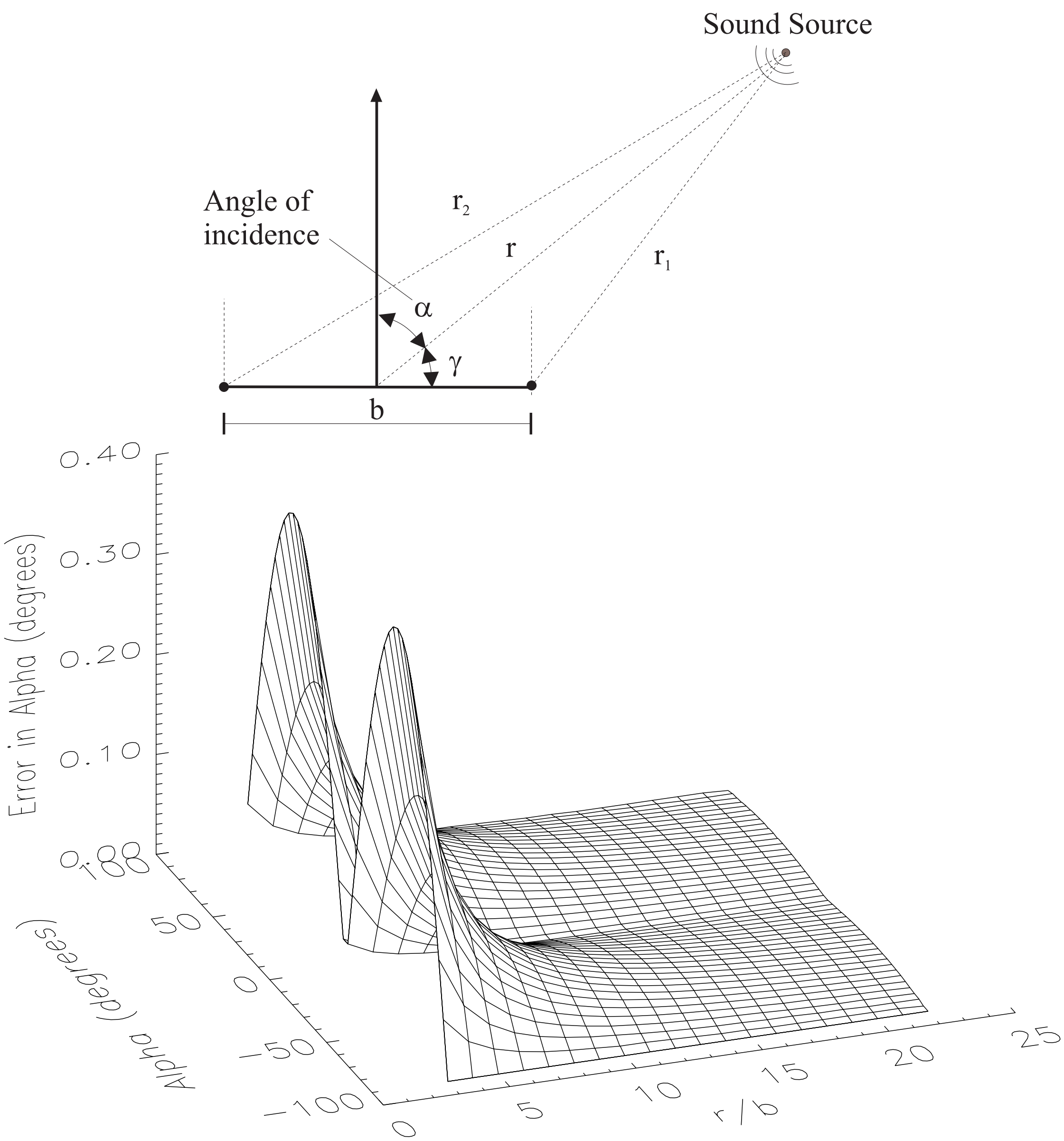


Figure 2: The first figure (top) illustrates the near-field situation with respect to two microphones listening to the same source. The second figure (bottom) plots the error created as a result of using a far-field approximation in near-field cases. The error is a function of both the angle to the sound source and the ratio of the distance to the source and the distance separating the two microphones.

$$= \sqrt{r^2 + \frac{b^2}{4} + rb\cos(\gamma)} \tag{4}$$

where $\alpha = \frac{\pi}{2} - \gamma$. The error between the actual angle α and its approximation using the far-field assumption is given by :

$$\begin{aligned} Error(\alpha, \frac{r}{b}) &= |\alpha - \alpha_{approximate}| \\ &= |\alpha - arcsin(\frac{r_2 - r_1}{b})| \end{aligned} \tag{5}$$

This error is a function of both the actual source direction, α, and the ratio of $\frac{r}{b}$. Figure 2 shows the effective error for a number of values of $\frac{r}{b}$ over the full range of α. It is noted that for values of $\frac{r}{b}$ greater than 3 that this error is less than 0.1^o.

Active omnidirectional microphone pair This method uses two omnidirectional microphones forming a baseline and relies on time delay information to compute angles of incidence (directions of arrival) with respect to two different poses of the microphone pair.

The intuition behind this method is the following. A single angle of incidence measurement from a single orientation 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 (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 3 and 4 illustrate the concept.

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 α between the normal of the baseline and any line of the cone that contains its vertex is given by equation 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 it is an overdetermined problem and the solution is found by satisfying a least squares criterion.

Consider the unknown source direction as a unit vector $\mathbf{s}$ 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 $\mathbf{s}$ then applies for a particular orientation i of the baseline $\mathbf{b_i}$ and a

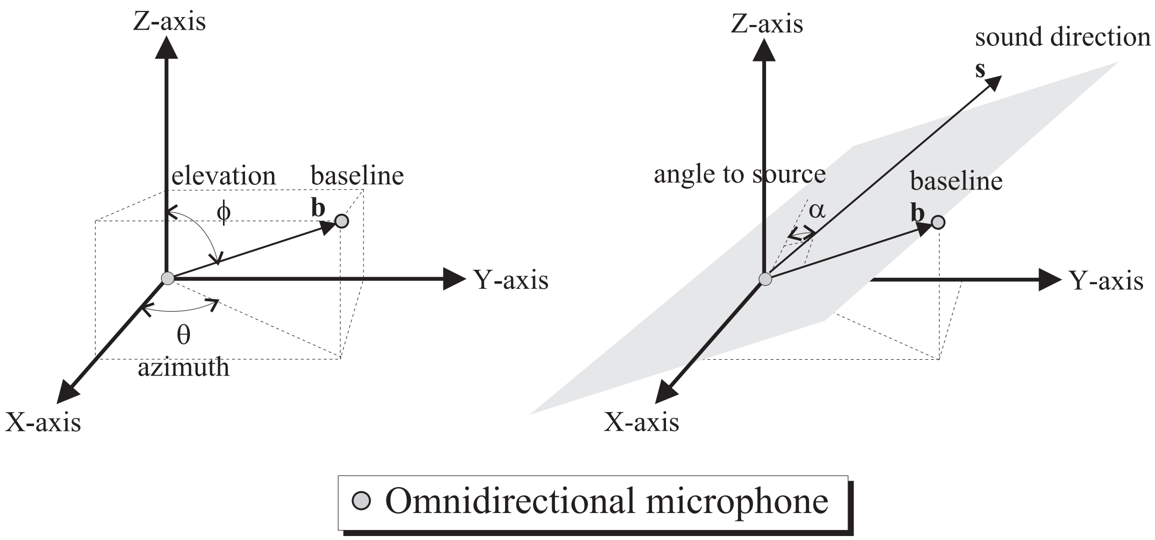


Figure 3: Two microphones form a baseline, $\mathbf{b}$ with an orientation (θ, ϕ) in 3D space. The direction to a sound event is given by $\mathbf{s}$. The simple two-dimensional solution yields the angle, α, between $\mathbf{b}$ and $\mathbf{s}$ on the plane that they form. This is the basis to the three-dimensional solution.

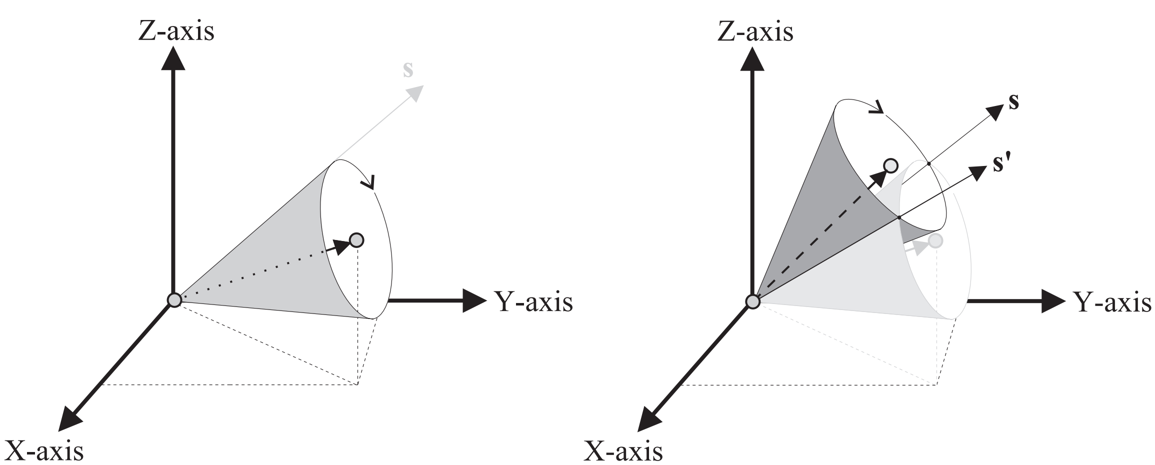


Figure 4: Two different poses of the baseline yield two solution cones, which intersect at two lines denoted by $\mathbf{s}$ and $\mathbf{s}'$, representing possible directions of arrival.

direction of arrival at angle γ_i with respect to baseline (unit) vector $\mathbf{b_i}$:

$$\mathbf{s} \cdot \mathbf{b_i} = \cos\gamma_i \tag{6}$$

or equivalently,

$$s_x b_{ix} + s_y b_{iy} + s_z b_{iz} = \cos\gamma_i = \cos(90 - \alpha_i) = \sin\alpha_i \tag{7}$$

where quantities b_{ix}, b_{iy}, b_{iz} are the cartesian coordinates of a unit vector with azimuth and elevation given by (θ_i, ϕ_i) respectively. Azimuth and elevation of the microphone baseline are controlled by the motors of the pan-and-tilt unit. Angle $\gamma_i = \frac{\pi}{2} - \alpha_i$ represents the direction of arrival with respect to the selected baseline pose. Using equation 1, γ can be computed from the measured time delay of the microphones. Combining 1 and 7 becomes :

$$s_x \ \cos\theta \ \cos\phi + s_y \ \sin\theta \ \cos\phi + s_z \ \sin\phi = \sin\alpha_i = \frac{cn}{fd} \tag{8}$$

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 and s_z. Since equation 8 is linear, this can be solved by obtaining three linear equations. Before solving this linear system of equations using one of the standard methods, it is necessary to ensure that the equations are consistent and yield a unique solution. Equivalently, we require that the three corresponding $\mathbf{b}$ vectors are not coplanar. To prevent this, baseline control can alternate between changing the azimuth, θ, and elevation, ϕ components of the baseline orientation.

Nonlinear solution For each orientation of the baseline there is a different version of equation 7. There is also an implicit nonlinear constraint that $s_x^2 + s_y^2 + s_z^2 = 1$ since $\mathbf{s}$ and $\mathbf{b}$ are unit vectors. To compute $\mathbf{s}$, a least squares approach can be used.

Rewriting equation 7 gives:

$$f_i(s_x, s_y, s_z) = s_x b_{ix} + s_y b_{iy} + s_z b_{iz} - c_i = 0 \tag{9}$$

where $c_i = cos\gamma_i$ and

$$f_{nonlinear}(s_x, s_y, s_z) = s_x^2 + s_y^2 + s_z^2 - 1 = 0 \tag{10}$$

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

$$\min_{s_x, s_y, s_z} (f_{nonlinear}(s_x, s_y, s_z) + \lambda \sum_i f_i(s_x, s_y, s_z)) \tag{11}$$

Weight λ was chosen equal to 1. Iterative Non-linear optimization algorithms can then be used to solve this problem [9]. 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.

3 Signal Processing for Time Delay estimation

We now describe the signal processing techniques for reliable time delay estimation. A time delay is estimated by correlating the two channels of a window of stereo sound data from the two microphones, and looking for the peak of the correlation function. The location of the peak corresponds to the time delay estimate (interpolation to achieve subsample accuracy was not used). Before correlation, filtering is carried out to reduce noice, and a signal level test is performed to check for the presence of a genuine sound event. The peak in the correlation function must be strong for it to be used for time delay estimation. The correlation level test is carried out for this purpose. A conservative threshold is chosen to reduce the likelihood of a false peak being used. The field of optimal time delay estimation has a long history and it is fairly advanced [7, 11]. In this work we have followed a rather basic approach to the problem. In future work, we plan to use more sophisticated techniques from the literature. Figure 5 shows a summary of our approach.

The five steps are the following:

1. **Filtering**. This involves high-pass filtering to eliminate low-frequency interference (for example due to ventilation fans), and low-pass filtering to eliminate high-frequency noise. Filtering depends on the expected sound source (impulsive or speech).

2. **Signal Level Test**. This involves a test to discriminate between the presence of a sound event to be localized and ”silence”. Only if the average power and absolute peak power of the signal within the

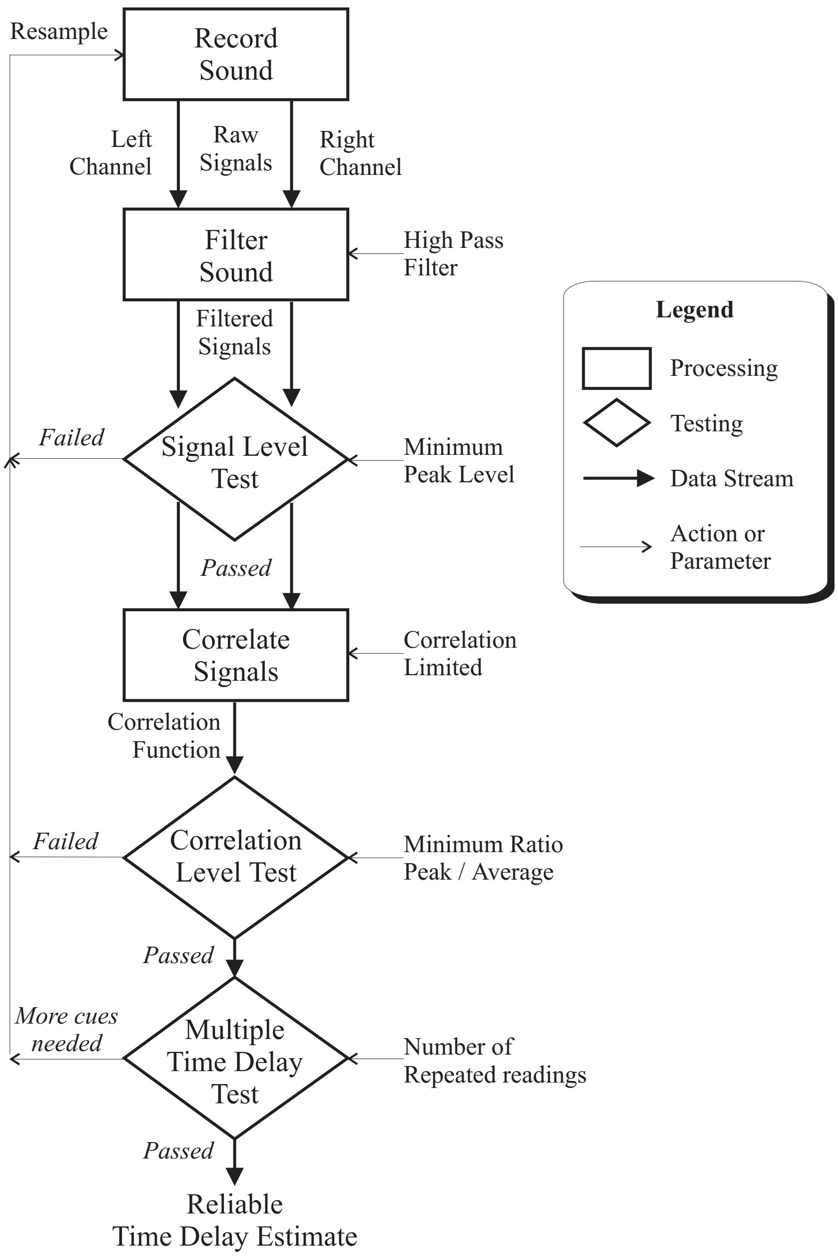


Figure 5: Signal processing for time delay estimation.

recorded window is significantly larger than estimates of the same quantities for background sound will the window be used to estimate a time delay. Otherwise it will be discarded.

3. **Correlation Range Limit**. The correlation function, c, is given by :

$$c(i) = \sum_t s_r(t+i)s_l(t) \tag{12}$$

   where s_r and s_l are the right and left signal channels respectively and t and i are over the time window of the input signals (with appropriate assumptions about the boundaries). The value of i that maximizes $c(i)$ corresponds to the time delay estimate. The interval over which i can vary for time delay estimation is much smaller than the duration of the input signals. As a result, it is sufficient to use the direct formula above for computing the correlation function over the interval of interest. Limiting the interval over which $c(i)$ is computed has the effect of eliminating ghost peaks in the correlation function that are due to shifts equal to multiples of the period of a periodic signal.

4. **Correlation Peak Level Test** : For reliable time delay estimation, the correlation function should have strong positive peaks. To reduce the likelihood of false peaks, we require that the maximum peak be considerably greater than both the largest secondary peak as well as the average of the correlation function.

5. **Multiple Time Delay Test**: A final check on the result is performed by clustering the time delay from a number of consecutive signal time windows (with fixed baseline orientation) and discarding the outliers.

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 broadband. Since these events are very short in time duration, the entire signal is often captured within a sampling window. The sudden large intensity peaks are easy to detect detectable by the peak-based Signal Level test described earlier.

Speech 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 within a sampling window. Some of these sounds will be more difficult to estimate time delay from, for example unvoiced 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 time delay estimates.

4 Experimental Results

This section describes our implementation of the algorithm that uses an active omnidirectional microphone pair. 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.

The experiments consist of a listening apparatus controlled by computer attempting to locate the direction of arrival of sound in three dimensions. The sound source is a speaker, which is placed in position and the computer is asked to estimate its position 25 consecutive times. 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.

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. The assembly is mounted on a Pan-Tilt Unit (PTU) allowing its orientation (b_θ, b_ϕ) to be controlled by the computer. For exact specifications see Appendix B.

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 (equation 5). We selected a distance of $r = 2m$ which corresponds to $\frac{r}{b} = 6.7$.

Figure 6 shows the error introduced by making the assumption with these parameters. It is less than 0.05^o 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 1 and setting the time delay value to the equivalent of $n = 1$ gives the minimum theoretical resolution that can be expected from this listening apparatus:

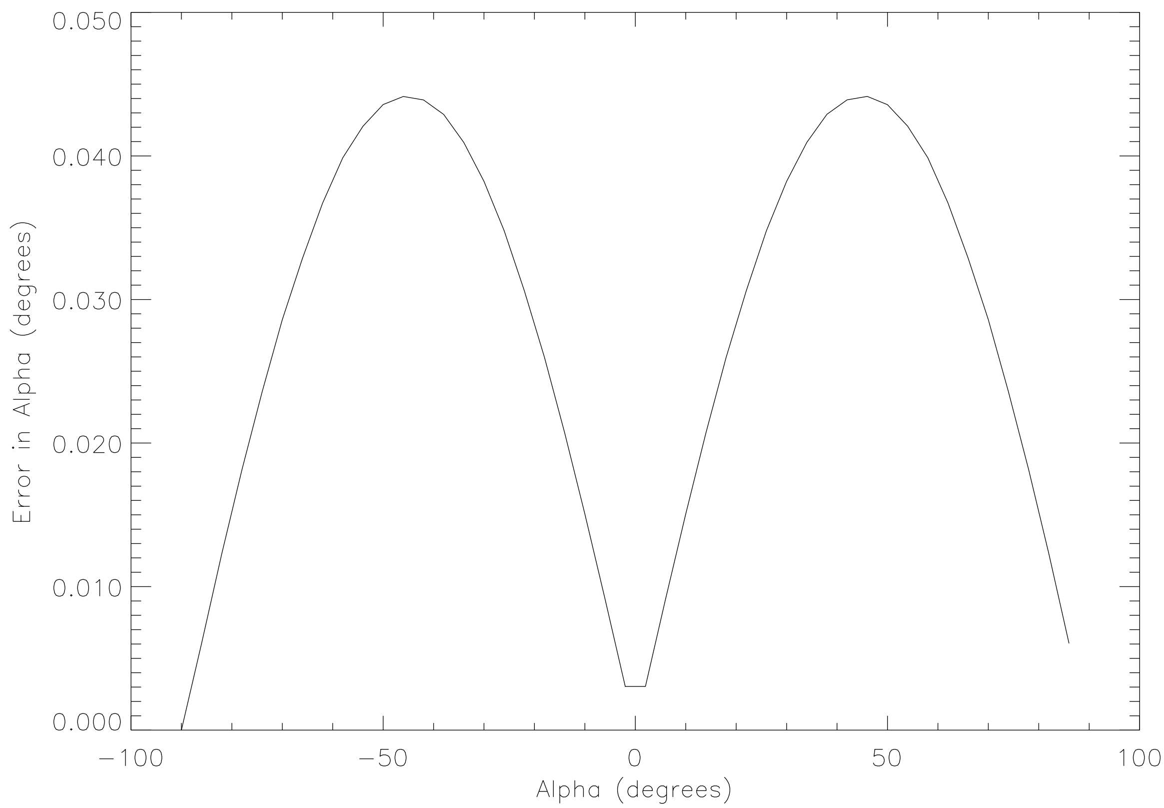


Figure 6: This figure show a slice of Figure 2 at the value of $\frac{r}{b} = 6.7$ which will be used in experiments for this work. It represents the error introduced by making a far-field assumption in the algorithm. The error for this experiment is relatively small, less than 0.05^o, so the far-field assumption can be used.

$$\alpha_{min} = \sin^{-1}(\frac{c}{fd}) = \sin^{-1}(\frac{342m/s}{22050Hz * 0.3m}) \approx 3^o \quad (13)$$

The number of discernible angles can be then be calculated by determining the range of time delay values for this setup. To find this, consider a sound source at α = 90 degrees to the baseline and apply it to equation 1 :

$$\pm n = \pm\frac{fd}{c} = \pm\frac{22050Hz * 0.3m}{342m/s} \approx \pm 19 \quad (14)$$

The value of n must be an integer since it represents the time delay as a number of finite samples (assuming no interpolation to achieve subsample resolution in estimating n). The number of possible time delays is then the range of [-19, 19] which is 39 discrete values. Equation 1 is used to map all the possible time delays in integer units to their corresponding angles. The following table shows the discernible angles for α which can be achieved for time delay values of 0 to 19. The table is symmetric for the negative time delay values.

Time Delay in integer units	**Angle**	**Time Delay** in integer units	**Angle**
0	0.0^o	10	31.1^o
1	3.0^o	11	34.7^o
2	5.9^o	12	38.3^o
3	8.9^o	13	42.2^o
2	11.9^o	14	46.4^o
5	15.0^o	15	50.9^o
6	18.1^o	16	55.8^o
7	21.2^o	17	61.5^o
8	24.4^o	18	68.5^o
9	27.7^o	19	79.2^o

Table 1: Discernible angles for Time Delays in integer units for our experimental setup.

As table 1 demonstrates, the theoretical resolution is accurate only at time delay = 1 but remains reasonably close to that value up to about time delay = 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 time delay 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 within this range. The ability of the active approach to adapt the geometry of the sensing apparatus to the direction of arrival is an important advantage over the fixed array approaches.

The algorithm used for orientation control is given in Table 2. It uses the last time delay 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. 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, $\mathbf{b_i}$, cannot be coplanar, which would create an underdetermined solution set. For these experiments the algorithm is applied to azimuth (pan) movements. This is due to the PTU elevation (tilt) range being too limited, so two specific elevations are alternated.

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

// Apply direction change alternately to pan or tilt
   if (nextPan)
       panPos += changeDirection;
   else
   {// Alternate between two tilt values, -10 and -40 deg.
       if (tiltPos <= -30.0)
           tiltPos = -10.0;
       else
           tiltPos = -40.0;
   }
   nextPan = !nextPan;
```

Table 2: The orientation change algorithm. The term $random$ refers to a function which would produce a uniformly distributed random number between 0 and 1.

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 7.

The experiment requires measurements to be taken with the sound source at different locations in 3-dimensional space. 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 $\theta = -40^o$ to 40^o in 10^o increments at elevation $\phi = 0^o$. Likewise, the heights were taken from $\phi = -30^o$ to 20^o in 10^o increments with $\theta = 0^o$. 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 θ will 'stretch' with elevations off of zero. The final planned sound source positions are shown in figure 8.

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 speech) before being moved to the next position. In this way, the proximity between the direction estimates from the impulsive and speech sounds is an indication of the accuracy of these estimates.

4.1 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 9.

Parameter Settings Using the rationale described in the previous section, the following tests and thresholds are chosen:

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 then 10.0 and initial peaks of the sound were often greater then 50.0.

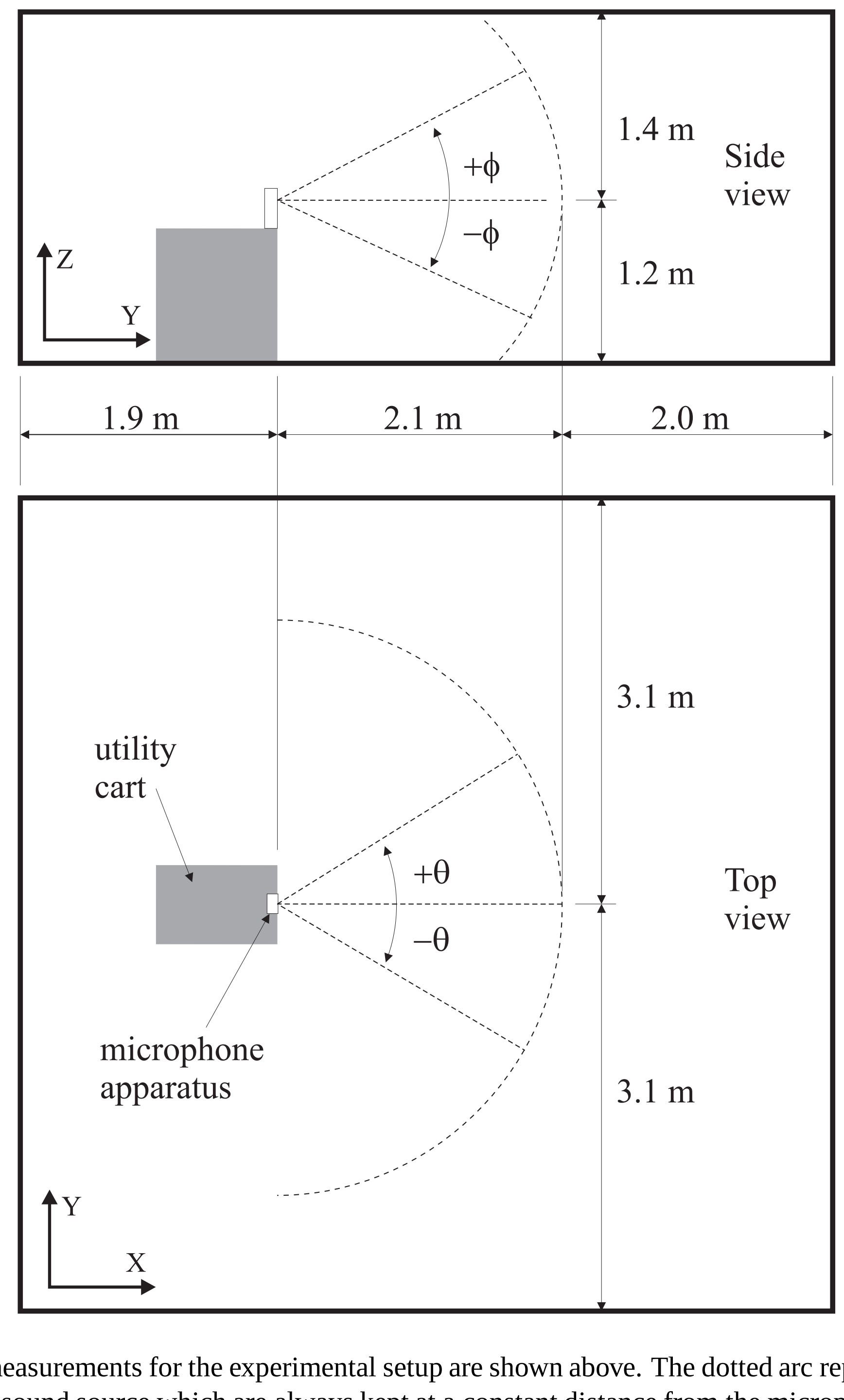


Figure 7: The measurements for the experimental setup are shown above. The dotted arc represents possible positions of the sound source which are always kept at a constant distance from the microphone apparatus.

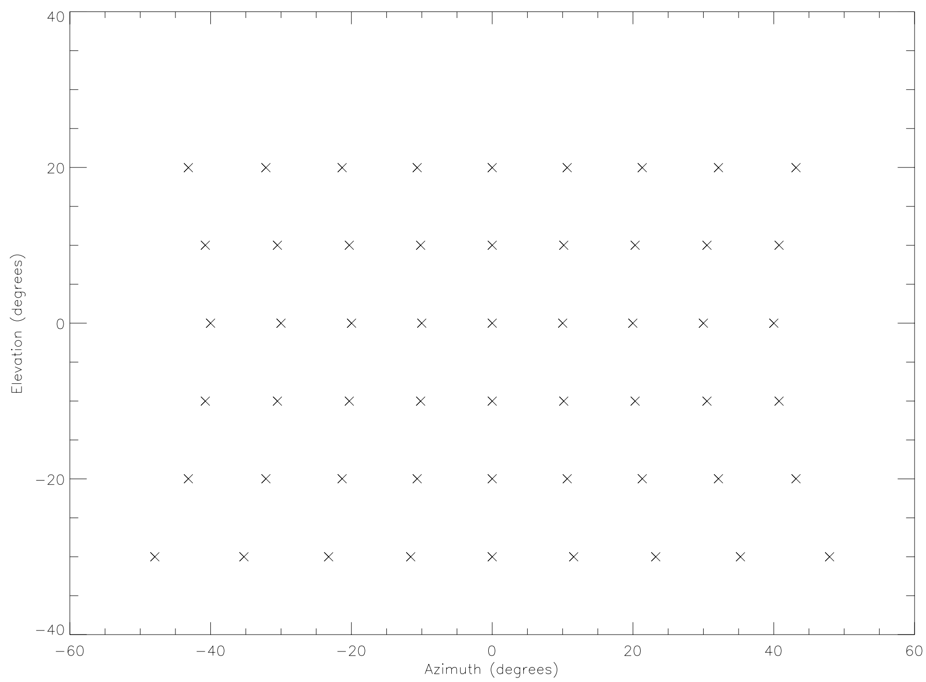


Figure 8: This graph shows the ideal sound source locations to be used in experiments. Note the stretching effect on the azimuth due to elevation.

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.

4. **Correlation signal level** : The primary peak in the correlation signal as a function of time delay 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 correct and false time delay estimates, which are created by correlating just the tails of the sound signal. If the correlation window only contains the tail end of an impulsive sound event, then the correlation strength at the correct delay value will be too low and the probability of a false time delay estimate muhc higher.

5. **Multiple Time Delay Estimates** : While most often the above techniques produced the correct time delay estimates, multiple readings were required to ensure good estimates were produced. Five consecutive time delay estimates were obtained and the median value was considered the correct time delay estimate.

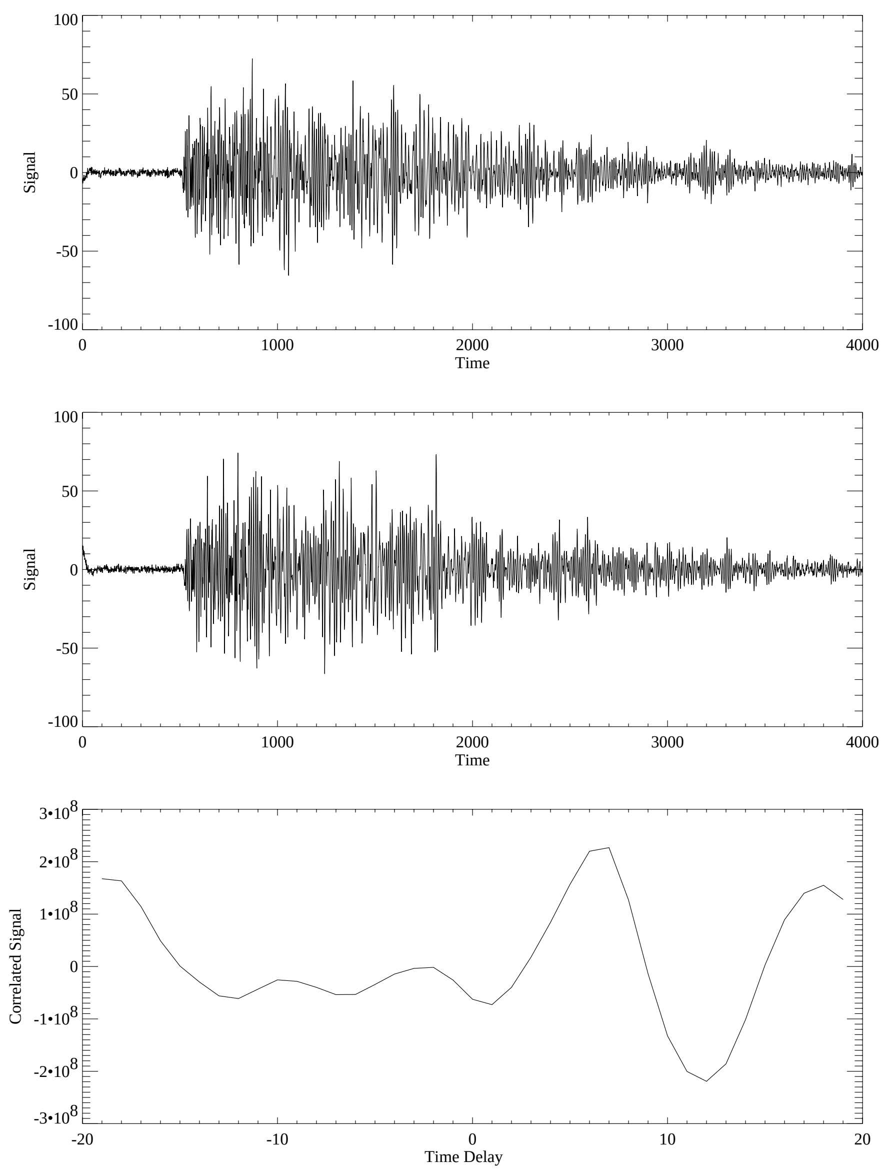


Figure 9: The three plots show two channels of sampled data from the experimental setup (see chapter 4) and their correlation. The sound recorded is of a single clap, a good example of an impulsive sound (high peaks and short duration). Since impulsive sounds have strong high frequency content, the resulting correlation has several smaller peaks. The strongest peak represents the correct time delay for the given experimental setup (7 samples).

4.2 Speech Source

The speech sample was recorded from a female subject reading a short passage of text. The sound sample was then played back in a continuous loop. The same passage was also read by a male reader and similar results were obtained.

Parameter Settings Using the rationale described in the previous section, the following tests and thresholds are chosen:

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 20% above the background noise level was chosen.

3. **Correlation limit** : Same as in the impulsive case.

4. **Correlation signal 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 than impulsive sounds, they are lower in magnitude and therefore result in a lower ratio of peak to average. A ratio of 5.0 was a reasonable value for this test.

5. **Multiple Time Delay Estimates** : With all of these techniques in place, the system did not produce good time delay estimates 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 time delay estimates, seven consecutive samples were used and their median was taken.

4.3 Results

The results for both experiments are shown in figure 10. 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 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^o at all positions.

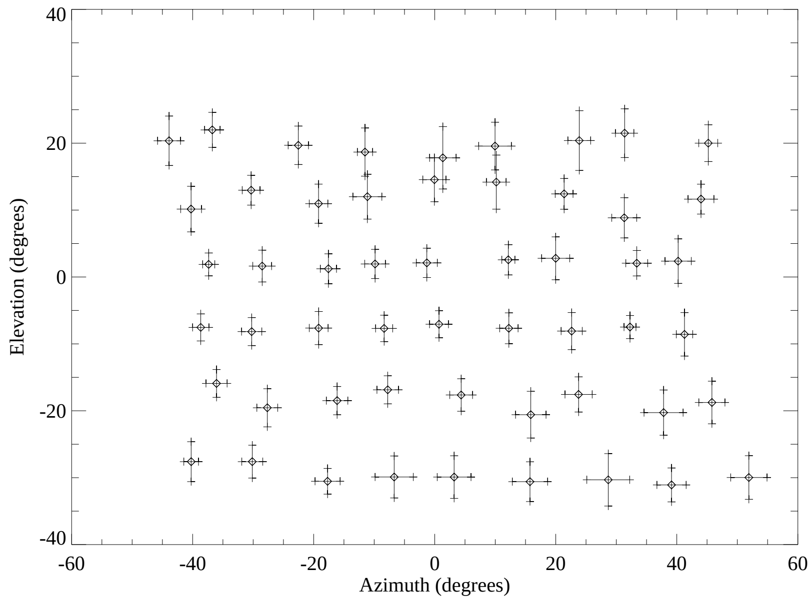
Impulsive Source			Voice Source		
Elevation	**AMD Az**	**AMD El**	**Elevation**	**AMD Az**	**AMD El**
20^o	1.8^o	3.5^o	20^o	1.5^o	2.4^o
10^o	1.8^o	3.0^o	10^o	1.5^o	2.8^o
0^o	1.6^o	2.4^o	0^o	1.8^o	2.8^o
-10^o	1.5^o	2.4^o	-10^o	1.8^o	2.9^o
-20^o	2.1^o	2.7^o	-20^o	2.4^o	3.4^o
-30^o	2.5^o	2.9^o	-30^o	2.3^o	2.6^o
Overall	1.9^o	2.8^o	**Overall**	1.8^o	2.8^o

Table 3: This table shows the average of the AMD for all the positions at a given elevation as well as over all positions.

Figure 11 shows the average estimated positions for each experiment overlaid on top of one another. Although the general pattern in the same as the planned positions (figure 8), many of these estimates are not exactly where they were planned to be. However, as discussed previously 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.

5 Discussion

We have presented techniques for estimating the direction of arrival of sound in three dimensional space using only two omnidirectional microphones on a fixed baseline mounted on a pant-and-tilt unit, which is actively controlled to optimize the geometry of incidence of sound and collect multiple angle of incidence measurements that can be combined to adequately constrain the direction of arrival. Our approach is similar in spirit to active vision, whereby cameras are mounted on computer-controlled pan-and-tilt units and the geometry is adjusted to resolve computer vision problems that are underconstrained in the fixed geometry case. Our approach does not require specialized multichannel A/D hardware, and can be implemented on off-the-shelf computing platforms that are equipped with the standard stereo sound input. In this article, we describe the signal processing and control algorithm of our experimental design. Experimental results are reported with both impulsive and speech sounds in an untreated, normally reverberant indoor environment,



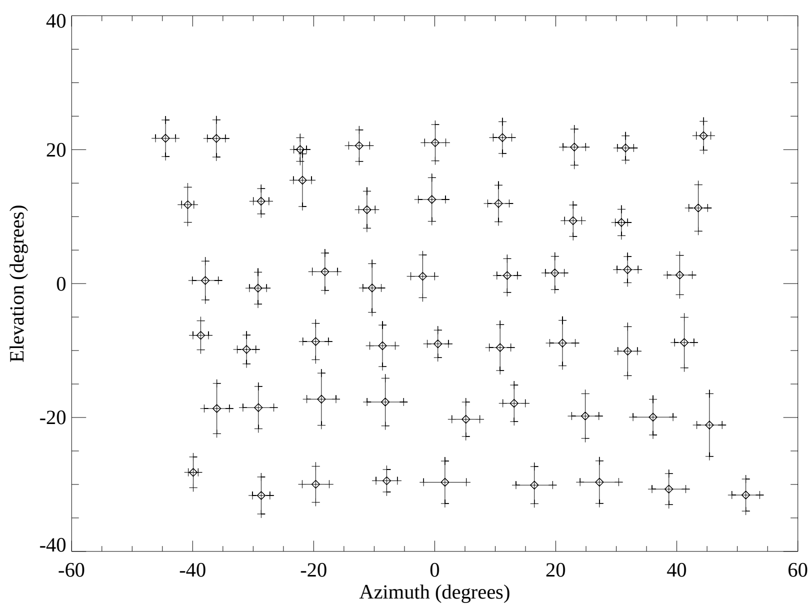


Figure 10: The graphs show the estimated sound direction of arrival in both azimuth and elevation for an impulsive source (top) and a speech source (bottom). Each position is computed from 25 determinations of the source in the same position. The diamond represents the average position, size of cross represents the absolute mean deviation in both azimuth and elevation.

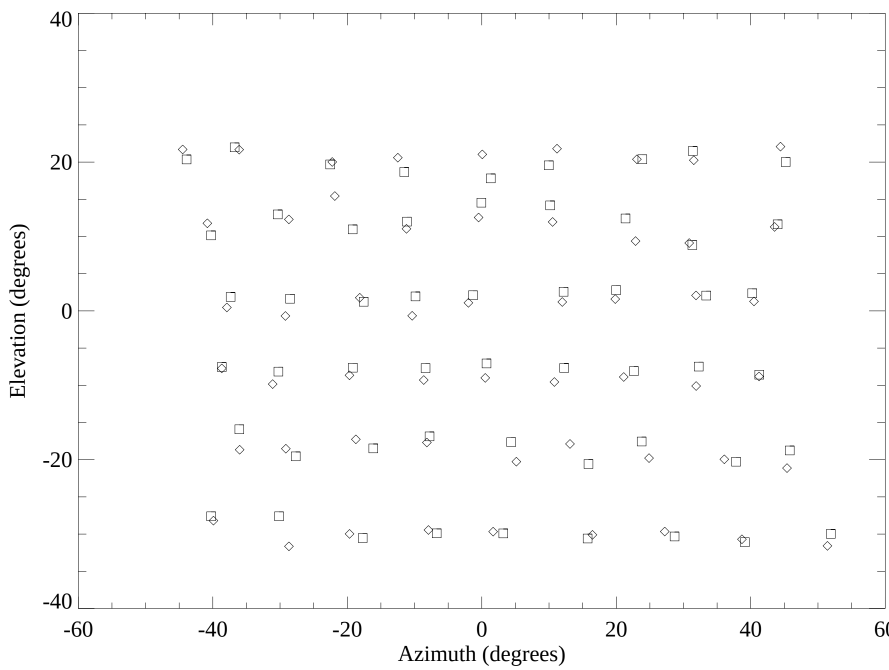


Figure 11: The graph has the average estimated positions of the experimental data from figure 10 overlaid on top of one another. The squares and diamonds are the average estimated source locations for the impulsive and speech sources respectively.

resulting in about 4^o accuracy in both azimuth and elevation.

We further discuss in the appendix two more approaches, one using a directional microphone with two rotational degrees of freedom and another using a combination of a directional and an omnidirectional microphone.

Similar work done in [1] on real date using a spatially distributed set of microphones reports a 30 cm accuracy in the estimate of a sound source position in three dimensional space for similar size and style rooms as used in these experiments. This corresponds to an angular resolution of 6^o 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.

Implementation of the work in [4] using two directional microphones was only performed in simulation but still allows for some comparisons to this work. An average error of less than 10 cm can be reasonably inferred from their figures. At distance of 1 m between source and sensors 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 changes greatly by position and by the amount of noise added to the training sessions. Localization of sources near zero incidence angle can be estimated most accurately, but for other incidence angles the error is much worse than 10cm. Our system, in a similar azimuth range, did not show the same degradation at larger angles which was naturally expected due to our active approach.

Spectral estimation is used in [13], whose results are also given from simulations, show remarkably good accuracy at less than 1^o for broadband sounds. However for tests involving human voice their results were relatively poor with $15^o - 30^o$ error. This limits their approach to applications involving broadband sounds. 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 time delay values.

The immediate extension of this work would be to complete the implementation of the algorithms presented in the appendix that use directional microphones and signal intensity. While an attempt was made for this work, signal intensities proved to be unstable with our experimental setup, and therefore made it impossible to experimentally demonstrate these methods.

Other extensions include tracking of a slowly moving sound source, discrimination of multiple sound sources, and localization of a sound source in 3-dimensional space using two active units far apart in space. The availability of a prediction of the direction of arrival allows us to optimize the pose of the microphones to minimize error in all cases.

A Source direction of arrival estimation using signal intensity cues

A different approach to source direction of arrival estimation is based on intensities of signals from directional microphones, whose directivity patterns are known 12. The ratio of signals from two directional microphones in close proximity depends on the direction of arrival of sound to the microphones.

The directivity pattern represents the intensity of the received sound signal I_{dir} as a function of the direction of arrival to the microphone α.

$$I_{dir} = f(\alpha) I_{src} \tag{15}$$

I_{src} is the intensity for angle $\alpha = 0$, therefore $f(0) = 1$.

These patterns are typically cardioid in shape [5] and they depend on frequency. Figure 12 shows a typical 2D directivity pattern for the directional microphone used in this work, measured in a regular reverberant (not anechoic) room. The microphone is equipped with a parabolic reflector (see appendix B). 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 $\pm 45^o$ of the direction of maximum response has been experimentally determined to be stable with respect to changes in the environment surrounding the microphone.

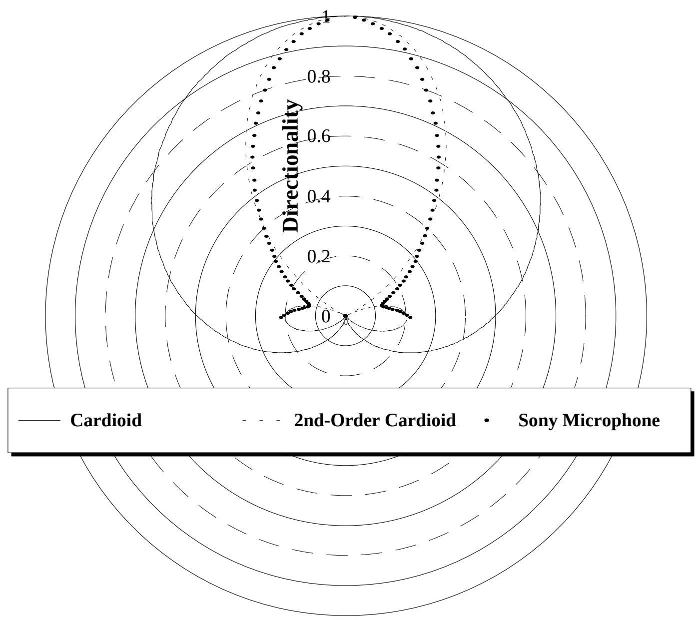
Far-field assumption Equation 15 makes the assumption that the sound source is a large enough distance away such that the signal energy loss due to the path length difference between the two microphones is negligible. We want to examine the error introduced by the far-field assumption.

The geometry is the same as the time delay case and given in figure 2. The path length, r_1 and r_2, are computed from equations 4 and 4. If the path length difference is taken into account then equation 15 is rewritten as:

$$I_1 = f_1(\alpha_1) \frac{E_{src}}{(r_1)^2} \tag{16}$$

$$and$$

$$I_2 = f_2(\alpha_2) \frac{E_{src}}{(r_2)^2} \tag{17}$$



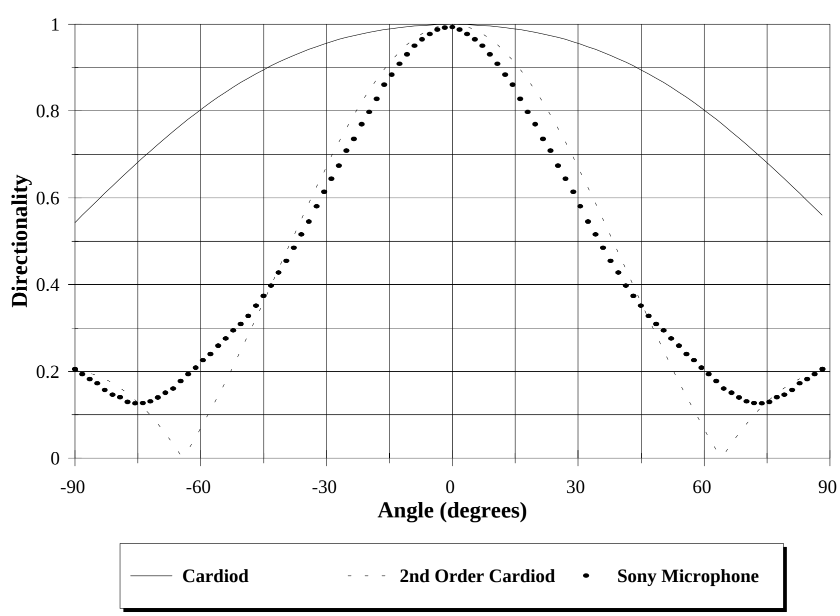


Figure 12: The measured directivity pattern for a SONY directional microphone. Note that it is approximated by a 2nd order cardioid in the range of -90 to 90 degrees. The side lobes are not reliable, partly due to the shadowing effect of the parabolic reflector, as they correspond to direction of arrival behind the reflector.

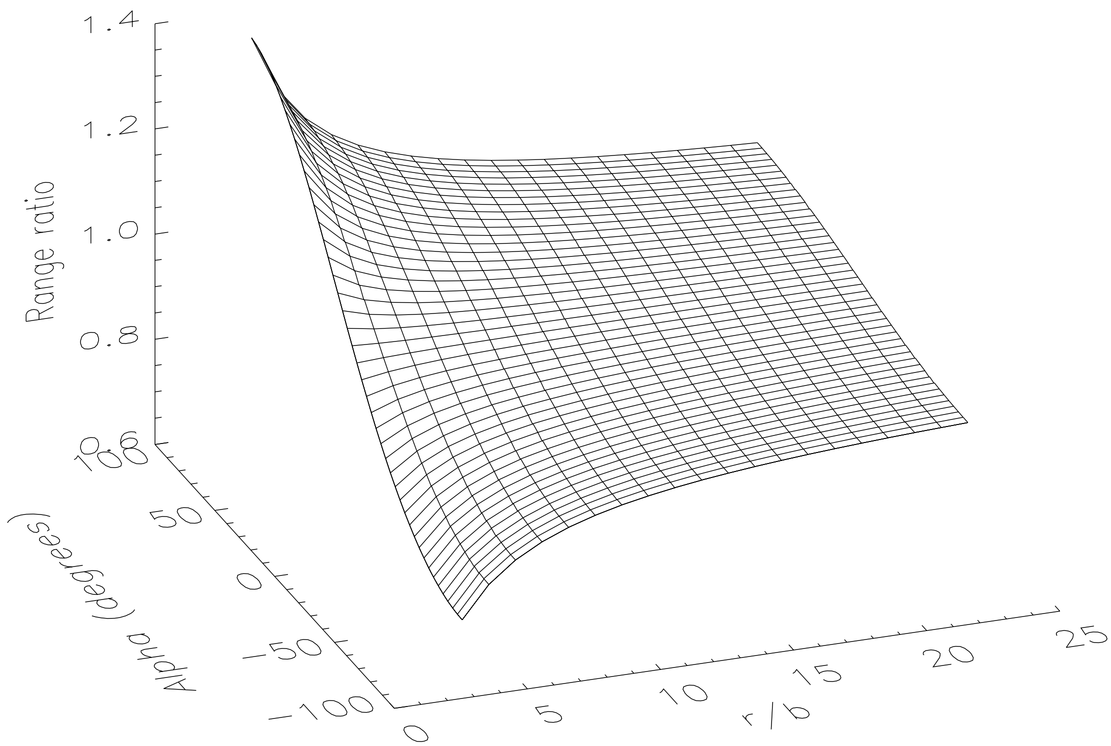


Figure 13: The top figure shows the change in $\frac{(r_2)^2}{(r_1)^2}$ that is introduced by making a far-field assumption. It is a function of both the angle to the sound source and the ratio of the distance to the source and the distance separating the two microphones.

where E_{src} is the energy emanating from the source. The signal intensity ratio would then be defined as :

$$\frac{I_1}{I_2} = \frac{f_1(\alpha_1)}{f_2(\alpha_2)} \frac{(r_2)^2}{(r_1)^2} \tag{18}$$

The error introduced by making the far-field assumption depends on how different $\frac{(r_2)^2}{(r_1)^2}$ is from the value of 1 and how α_1 and α_2 differ from each other. This error is a function of both the actual source direction α as measured from the centre of the baseline, and the ratio of $\frac{r}{b}$. Figures 13 and 14 show the effect on $\frac{(r_2)^2}{(r_1)^2}$ with a number of values of $\frac{r}{b}$ over the full range of α. It is noted that for values of $\frac{r}{b}$ greater than 8 that this error is less than 0.1.

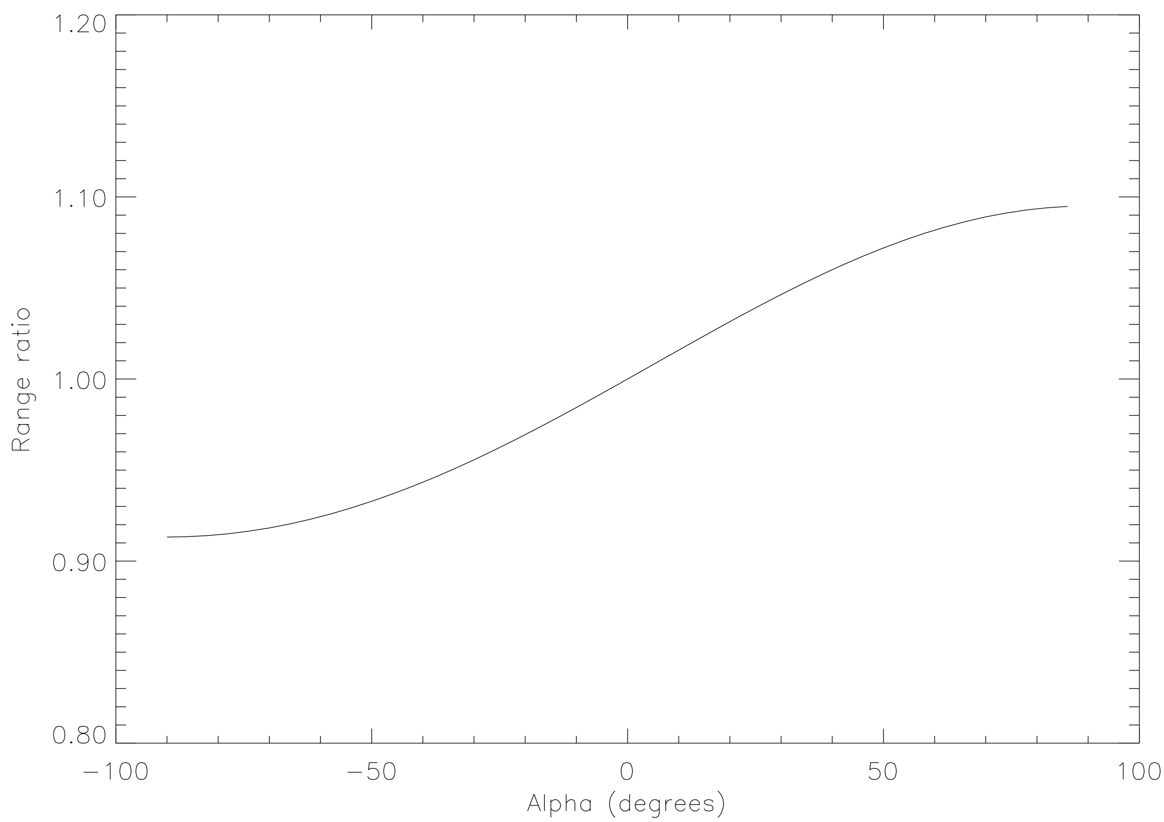


Figure 14: This figure shows a slice of 13 for $\frac{r}{b} = 8$. It shows that the ratio $\frac{(r_2)^2}{(r_1)^2}$ only changes by 0.1 from the expected value of 1.0 if a far-field is assumed.

A.1 A single active directional microphone for direction of arrival estimation in two dimensions

For this method, it is assumed a sound source has fixed position in space. 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 12 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 [9], sufficient measurements are collected on the main lobe and high enough above the sidelobes.

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

$$\frac{f(\phi_1 - \delta)}{I_1} = \frac{f(\phi_2 - \delta)}{I_2} = ... = \frac{f(\phi_N - \delta)}{I_N} = s \tag{19}$$

Therefore, one can formulate the estimation of translation δ and the scale factor s as a one-dimensional search for the minimum of the following function of δ :

$$\min_{\delta} \sum_{i \neq j} \left(\frac{f(\phi_i - \delta)}{I_i} - \frac{f(\phi_j - \delta)}{I_j} \right)^2 \tag{20}$$

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, i.e. in both azimuth and elevation.

A.2 An active directional and a fixed omnidirectional microphone pair

This method uses an active directional microphone and a fixed omnidirectional microphone. The intensity of sound I_{dir}, at a single frequency, at the directional microphone depends on the direction of arrival α and is proportional to the intensity of the source I_{src}, as per equation 15 :

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

$$I_{omni} = kI_{src} \tag{21}$$

where k is a constant, independent of α. 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 :

$$\frac{I_{dir}}{I_{omni}} = \frac{f(\alpha)}{k} \tag{22}$$

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 via a calibration procedure, carried out by measuring $\frac{I_{dir}}{I_{omni}}$ for a number of uniformly spaced directions of arrival α.

For a directivity pattern like the one shown in figure 12, 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.

Comments This algorithm, like the previous one, depends on the consistency of the signal intensities it relies on. Unfortunately, we were unable to obtain such consistency on real data with the consumer-quality equipment we used.

B Technical Specifications

B.1 Power PC laptop

Manufacturer : Apple Computer Inc.

Model : Powerbook 500 Series

Weight : ≈ 3 kg

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

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.

B.2 Omnidirectional tie clip microphone

Manufacturer : Genexxa

Model : 33-3003

Weight : ≈ 20 g

Dimensions : 17.6 x 8 mm (length/diameter)

Cost : $\approx US\$30$

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

B.3 Electret condenser directional microphone

Manufacturer : Sony

Model : ECM-PB1C

Weight : ≈ 20 g

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

Cost : $\approx US\$80$

Description : The directional microphone tested for the work consisted 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.

B.4 Pan-Tilt Unit (PTU)

Manufacturer : Directed Perception Inc.

Model : PTU-46-17.5

Weight : ≈ 2 kg

Dimensions : 65 x 130 x 90 mm (width/height/depth)

Cost : $\approx US\$2000$

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 by Matthew Izatt, Tom Luu, Henry Wong, Willen Straten and Greg Reid.

References

[1] D. V. Rabinkin amd R. J. Renomeron, A. Dahl, J. C. French, J. L. Flanagan, and M. H. Bianchi. A DSP implementation of source location using microphone arrays. *Journal of the Acoustical Society of America*, 99(4):2503, April 1996.

[2] Jens Blauert. *Spatial Hearing*. The MIT Press, 1997.

[3] M. S. Brandstein, J. E Adcock, and H. F. Silverman. A closed-form location estimator for use with room environment microphone arrays. *IEEE Transactions on Speech and Audio Processing*, 5(1):45–50, January 1997.

[4] Michael S. Datum, Francescon Palmieri, and Andrew Moiseff. An artifical neural network for sound localization using binaural cues. *Journal of the Acoustical Society of America*, 100(1):372–383, July 1996.

[5] Don Davis and Carolyn Davis. *Sound System Engineering, 2nd edition*. Focal Press, Boston, 1997.

[6] J. Flanagan, J. Johnston, R. Zahn, and G. Elko. Computer-steered microphone arrays for sound transduction in large rooms. *Journal of the Acoustical Society of America*, 78:1508–1518, 1985.

[7] Jian Li and Renbiao Wu. An efficient algorithm for time delay estimation. *IEEE Trans. on Signal Processing*, 46(8):2231–2235, Aug. 1988.

[8] E. Milios, M. Jenkin, and J. Tsotsos. Design and performance of TRISH, a binocular robot head with torsional eye movements. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(1):51–68, February 1993.

[9] W. Press, S. Teukolsky, W. Vetterling, and B. Flannerey. *Numerical Recipes in C: The Art of Scientific Computing (2nd ed.)*. Cambridge Univ. Press, Cambridge, UK, 1992.

[10] G. Reid and E. Milios. Active binaural sound localization. *IX European Signal Processing Conference (EUSIPCO)*, IV:2353–2356, September 1998.

[11] R.J. Vaccaro. The past, present, and the future of underwater acoustic signal processing. *IEEE Signal Processing Magazine*, 15(4):21–51, July 1998.

[12] Frederic L. Wrightman and Doris J. Kistler. Factors affecting the relative salience of sound localization cues. In Robert H. Gilkey and Timothy R. Anderson, editors, *Binaural and Spectral Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates, 1997.

[13] Pierre Zakarauskas and Max S. Cynader. A computational theory of spectral cue localization. *Journal of the Acoustical Society of America*, 94(3):1323–1331, September 1993.