Active Stereo Sound Localization*

Greg L. Reid

Department of Computer Science, York University, Toronto, Canada, M3J 1P3

Evangelos Milios

Faculty of Computer Science, Dalhousie University, Halifax, Canada, B3H 1W5†

(Dated: May 20, 2002)

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 a fixed 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 position of the microphones is made variable by mounting them on one or more computer-controlled pan-and-tilt units. In this paper, 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.

PACS numbers: 43.28.Tc, 43.60.Gk

*A condensed description of this work was presented at the European Signal Processing Conference, September 8 - 11, 1998, Vol. IV, pp. 2353-2356.

†Electronic address: eem@cs.dal.ca

I. INTRODUCTION

In human auditory perception, it is believed that there are three basic cues from which most sound localization is derived[1,2]. Interaural Time Difference is the primary horizontal cue for humans at lower frequencies (below 1KHz). Interaural Intensity (or Level) Difference is the primary horizontal cue at 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 one's outer ears, head and torso. Spectral cues extend our perception into the vertical plane.

Many existing implementations of sound source localization have used arrays of omnidirectional microphones with beamforming[3] and generalized time-delay techniques[4,5]. These approaches often require special purpose multichannel A/D hardware which generate significant amounts of signal data and require intensive computation. Large spatial separation between microphones and a larger number of microphones (16) are used[4,5] 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. Rabinkin, et al[4] report 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 our experiments. This corresponds to an angular resolution of 5.7^o in the sound direction at an average 3 m distance which is comparable to that achieved in our work. Rabinkin, et al also performed tests in larger reverberant rooms like auditoriums and found that their errors increased significantly. However their 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 reported by Zakarauskas and Cynader[6]. The authors simulate the human auditory system by focusing on spectral cues for 3-dimensional sound source localization 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. Simulation results show remarkably good accuracy of better 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 applica-

tions involving broadband sounds. In comparison, the work reported in this paper maintains nearly the same error bounds for both broadband and voice sources although more signal processing is required in the voice testing to maintain reliable time delay values.

Estimating the direction of a sound source from signals received at two fixed directional microphones has been addressed and tested only in simulation mode by Datum, et al[7]. In their work, the microphones are fixed in space, both pointing forward with a slight difference in 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 difference 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. Datum, et al report on performance of their method in simulation. An average position error of less than 10 cm can be reasonably inferred from their figures. At distance of 1 m between source and sensors, this corresponds to an angular error of 5.7^o, and puts those results very close to both[4] and the errors achieved in our 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 5.7^o. 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 that adapts the orientation of the microphone baseline to the direction of incidence of the incoming sound signal.

Our work intends to replace the functionality of an array of microphones with two microphones mounted on a computer-controlled pan-tilt unit, as shown in Fig. 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"[8], the auditory equivalent of "Active Vision"[9], 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. 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

are discussed in a technical report[10].

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

II. ACTIVE 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 position of the two microphones is computer-controllable.

A. 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 (Fig. 2) and is calculated by simple geometric constraints.

$$sin(\alpha) = \frac{ct}{b} = \frac{cn}{fb} \tag{1}$$

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 b 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.

B. 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 approximately the same at both microphones.

This is strictly true for a source at an infinite distance away.

The general case is given in Fig. 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 :

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

$$\begin{aligned} r_2 &= \sqrt{r^2 + (\frac{b}{2})^2 - 2r(\frac{b}{2})\cos(\pi - \gamma)} \\ &= \sqrt{r^2 + \frac{b^2}{4} + rb\cos(\gamma)} \end{aligned} \tag{4}$$

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 this error is less than 0.1^o.

C. 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 positions 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 required are obtained then we have 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 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,

$$s_x b_{ix} + s_y b_{iy} + s_z b_{iz} = \cos\gamma_i = \cos(\pi/2 - \alpha_i) = \sin\alpha_i \tag{7}$$

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 position. 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.

1 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 **b** vectors not be coplanar. To ensure this, baseline control can alternate between changing the azimuth, θ, and elevation, ϕ components of the baseline orientation.

2 Nonlinear solution

For each orientation of the baseline there is a different instantiation of equation 7. 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 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[11]. In order to assure convergence, an initial solution is required which is near the correct solution. The simplest way to ensure convergence is to first calculate the linear solution of a set of three linear equations obtained by instantiating equation 7 for three different orientations of the baseline, and then to refine the linear solution by using the nonlinear approach after inclusion of the nonlinear constraint $s_x^2 + s_y^2 + s_z^2 = 1$.

III. 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[12,13]. 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). A high pass linear-phase FIR filter was designed using the McClellan-Parks algorithm[14] with 185 taps. The upper and lower edges of the two bands as a fraction of the sampling frequency were: band 1 [0.000, 0.005], and band 2 [0.015, 0.5], or, in terms of frequencies in Hz, band 1 [0Hz, 110.25Hz], band 2 [330.75Hz, 11025Hz]. Desired responses were 0 and 1 and weights were 10 and 1 for bands 1 and 2 respectively.

2. *Signal Level Test.* This involves a test to discriminate between the presence of a sound event to be located and "silence". Only if the average signal level and absolute peak signal level within the 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 Domain 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 ranges over the time window of the input signals. Correlation index i is calculated over a much smaller range ($\pm$ 20 samples) corresponding to the expected range of delays for the baseline of 30cm, which corresponds to a propagation delay of 20 samples between the two microphones. The time window used contains 4000 samples, or 0.18s. From Fig. 6 we see that reflections off the floor or ceiling travel over a path of about 3.4 m, reflections off side walls and wall behind the speaker travel over a path of about 6.5 m. Given that the direct path is 2.1 m, the above propagation paths correspond to delays of 3.8 ms (85 samples) and 13 ms (288 samples) for floor/ceiling and side/back wall reflections respectively. This implies that the time window includes reflected signals. The strongest reflected signals are the ones off the floor and ceiling, which arrive from the same azimuth as the direct path signal, but from different elevations. As a result, we would expect higher repeatability and lower variance in the azimuth estimates. This is confirmed by the experimental results presented later. The value of i that maximizes $c(i)$ corresponds to the time delay estimate for the direct path signal. 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 are characterized by a sudden large intensity change which quickly decays into background noise. Examples include a hand clap or a slamming door. The frequency spectrum is broadband. Since these events are very short in time duration, the

entire signal is often captured within one sampling window. The sudden large intensity peaks are easy to detect by the peak-based Signal Level test described earlier.

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 low intensity. So in the case of speech it is desirable not only to eliminate sample data which does not contain sound, but also data which is less likely to produce reliable time delay estimates. This is accomplished by the signal level test.

IV. EXPERIMENTAL RESULTS

Two sets of experiments are described in this section. The first uses an impulsive sound and the second experiment uses speech. The experiments apply a listening apparatus controlled by computer to locate the direction of arrival of sound in three dimensions. The sound source is a speaker, which is placed in a fixed position and the computer is asked to estimate its position 25 consecutive times, while the speaker plays back continuous speech. The experiment is repeated for impulsive sounds without changing the position of the speaker, while the speaker plays back repeated hand claps.

The source is then moved and the experiment is repeated for both speech and hand clap sound in the new position. For each source position, the difference between the estimates of azimuth and elevation obtained from the speech and hand clap sound, and the standard deviation of the multiple estimates give an indication of the consistency of the algorithm.

A. Experimental setting

The environment for the following experiments is an ordinary rectangular room about 6 m wide, 7 m long and 3 m high. The centre of the room has been cleared of furniture and the listening apparatus is placed on a cart at a distance of 2.1 m (7 ft) from the area where the sound source is be located. The room is carpeted, has standard ceiling tiles, windows, a white board, wall-mounted 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 positioning of apparatus and sound source are shown in Fig. 6.

The listening apparatus for these experiments consists of a pair of two Genexxa 3303003 electret condenser microphones (with diameter of approximately 8 mm) mounted at either end of a wooden rod forming a baseline b of length 0.3 m. The assembly is mounted on a Pan-Tilt Unit (PTU) allowing its orientation (b_θ, b_ϕ) to be controlled by computer.

In order for the far-field assumption to be applicable, the distance r to the sound source must be sufficiently larger than the length of b (equation 5). We selected a distance of $r = 2 \ \ m$ which corresponds to $\frac{r}{b} = 6.7$.

Figure 2 shows the error introduced by making the far-field assumption with these parameters. It is less than 0.05^o for the worst case, which is acceptable for this application.

Sound signal 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 $f =$ 22,050 Hz and using 16-bit resolution. Using equation 1 and setting the time delay value to the equivalent of $n = 1$ gives the best resolution that can be expected from this listening apparatus:

$$\alpha_{min} = \sin^{-1}(\frac{c}{fb}) \approx 3^o \tag{13}$$

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 $\alpha = 90^o$ to the baseline and apply equation 1 :

$$\pm n = \pm\frac{fd}{b} = \approx \pm 19 \tag{14}$$

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. Table I 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.

Table I demonstrates that the theoretical resolution of approximately 3^o or equivalent maximum error due to angle quantization of $\pm 1.5^o$ is accurate only at time delay of 1 but

remains reasonably close to that value up to about time delay of 14 (46.4^o) where after the error begins to diverge. This is reflected in the values of maximum error due to angle quantization. Therefore, it would be desirable to orient the microphone baseline so as to obtain time delay values inside the -14 to 14 range wherever possible. This is accomplished using a simple algorithm 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 II. 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 set of equations. In 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.

The need for mechanical steering of the microphone pair makes our method suitable in cases where the sound source is either stationary in space or moves much slower than the time constants involved in the mechanical movements of the pan-and-tilt unit. Fixed arrays have an advantage where rapid movement of the sound source is involved.

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 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 precision of these estimates.

B. Impulsive Source

For the impulsive sound experiment a recording of a single hand clap was used and repeated at one second intervals. The sound event is the same to that used in Fig. 7. Using the rationale described in section III, the following tests and thresholds are chosen:

1. *Filtering* : The filter used was described earlier in section III. The cutoff frequeny of 200 Hz is used to attenuate line noise and environmental noise such as overhead ventilation fans.

2. *Peak level* : A signal level threshold equal to 9.54 dB above background noise, while signal level peaks were often about 14 dB above background noise.

3. *Correlation domain* : The correlation domain 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 above a specified multiple of the average of the correlation signal, for it to be acceptable towards estimating the time delay. It has been determined experimentally that a ratio of 20 succeeds in differentiating between correct and false peaks. If the correlation signal level is below 20 times the average, then the primary peak is considered unreliable and the time delay estimate derived from it is rejected. Strong peaks due to reflected signals may lead to a high average, and therefore rejection of the primary peak.

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 over a time interval of 0.9 s and the median value was considered the final time delay estimate.

C. 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. Using the rationale described in section III, the following tests and thresholds are chosen:

1. *Filtering* : Same as in the impulsive case.

2. *Peak level* : A signal level at 1.58 dB the background noise level was chosen. The reason such a low signal level was chosen is because, compared to the impulsive case, speech is considerably closer in both signal level and in frequency to the background noise level in the room.

3. *Correlation domain* : 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 multiple 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 chosen for this test.

5. *Multiple Time Delay Estimates* : Even after the application of the previous tests, the system did not produce good time delay estimates as often as in the impulsive case. In order to ensure good time delay estimates, the median of seven consecutive time delay estimates over a time interval of 1.26 s was taken as the final estimate.

D. Results

The results for both experiments are shown in Fig. 8. Each "cross" is the result of 25 measurements of the direction of arrival of the sound source in a fixed location. The cross is centered at the average position of those measurements with its width and height illustrating the standard deviation of the azimuth and elevation estimates respectively. In general what is seen is a somewhat smaller standard deviation in azimuth than in elevation: the standard deviations of the azimuth and elevation estimates from the hand clap experiments are 2.55^o and 3.61^o, and from the speech experiments are 2.53^o and 3.65^o respectively.

Figure 9 shows the average estimated positions for the impulsive and speech source experiments overlaid on top of each other. There is good agreement between the impulsive and speech source estimates: the average distances between the impulsive and speech source

estimates are 0.85^o and 1.43^o along the azimuth and elevation dimension respectively. The standard deviations of the distance between the impulsive and speech source estimates are 0.68^o and 1.04^o along the azimuth and elevation dimension respectively.

V. 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 pan-and-tilt unit, whose orientation 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 under constrained 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 a signal processing and control algorithm associated with our experimental design. Experimental results are reported with both impulsive and speech sounds in an untreated, normally reverberant indoor environment, resulting in about 4^o accuracy in both azimuth and elevation.

REFERENCES

[1] 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.

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

[3] 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.

[4] D. V. Rabinkin, 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] G. Reid and E. Milios. Active stereo sound localization. Technical Report CS-1999-09, Department of Computer Science, York University, 1999.

[11] 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.

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

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

[14] P. Embree and B. Kimble. *C Language Algorithms for Digital Signal Processing*. Prentice Hall, Englewood Cliffs, 1991.

LIST OF FIGURES

CONTENTS

Time Delay in integer units	**Angle**	**Max Error**	**Time Delay** in integer units	**Angle**	**Max Error**
0	0.0^o	1.48^o	10	31.1^o	1.76^o
1	3.0^o	1.49^o	11	34.7^o	1.84^o
2	5.9^o	1.49^o	12	38.3^o	1.94^o
3	8.9^o	1.51^o	13	42.2^o	2.07^o
2	11.9^o	1.52^o	14	46.4^o	2.24^o
5	15.0^o	1.55^o	15	50.9^o	2.48^o
6	18.1^o	1.57^o	16	55.8^o	2.85^o
7	21.2^o	1.61^o	17	61.5^o	3.41^o
8	24.4^o	1.65^o	18	68.5^o	5.34^o
9	27.7^o	1.70^o	19	79.2^o	5.34^o

TABLE I: Discernible angles for Time Delays in integer units for our experimental setup. The maximum error due to angle quantization is also shown.

```
// 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 II: The orientation change algorithm. The term *random* refers to a function which would produce a uniformly distributed random number between 0 and 1.

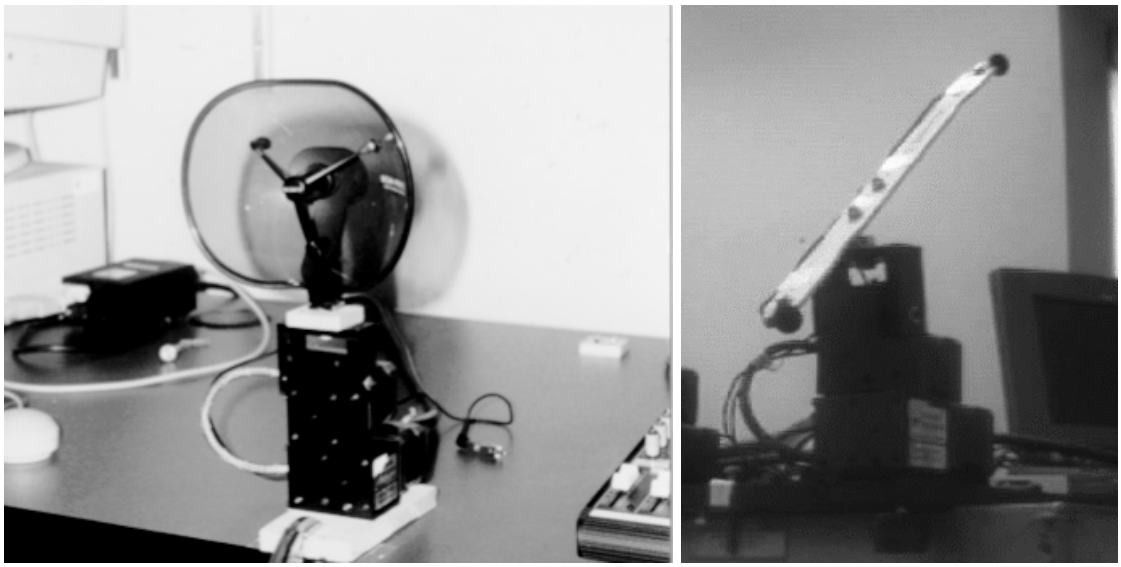

FIG. 1: Left: A directional microphone mounted on a computer controlled pan-tilt unit (PTU). Right: A pair of omnidirectional microphones spatially separated also mounted on a PTU. The PTUs allow the microphones' position and orientation to be manipulated creating an 'active' system.

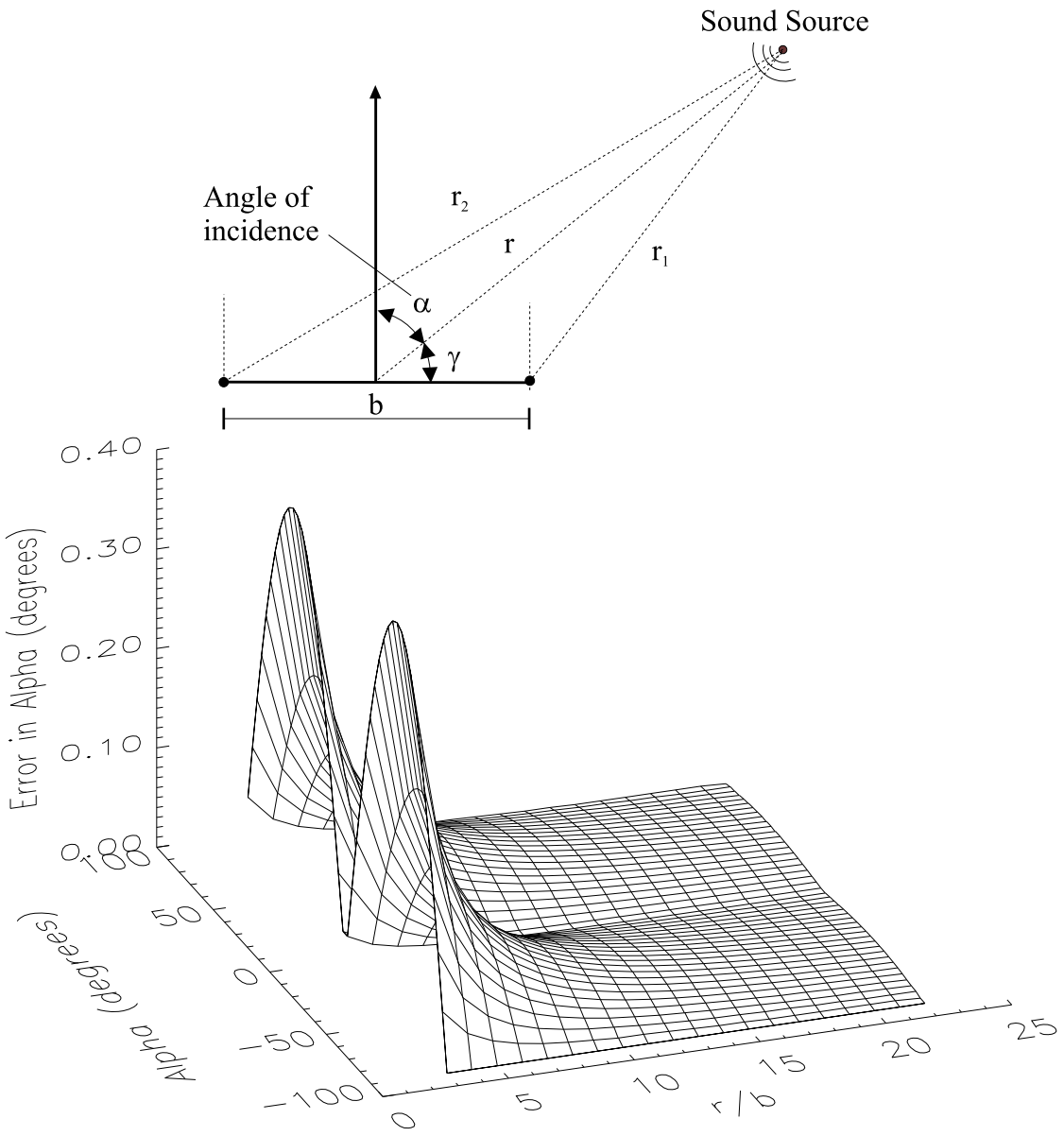

FIG. 2: The top figure illustrates the near-field situation with respect to two microphones listening to the same source. The bottom figure plots the error created as a result of using a far-field approximation in near-field cases. The error is a function of both the angle to the sound source and the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

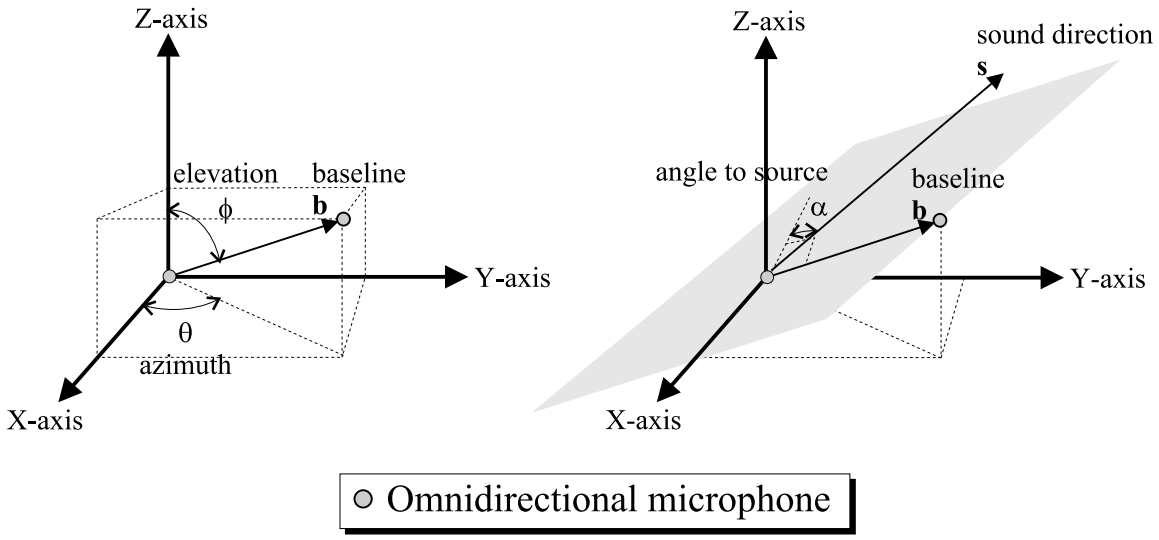

FIG. 3: Two microphones form a baseline, **b** with an orientation (θ, ϕ) in 3D space. The direction to a sound event is given by **s**. The simple two-dimensional solution yields the angle, α, between **b** and **s** on the plane that they form. This is the basis for the three-dimensional solution.

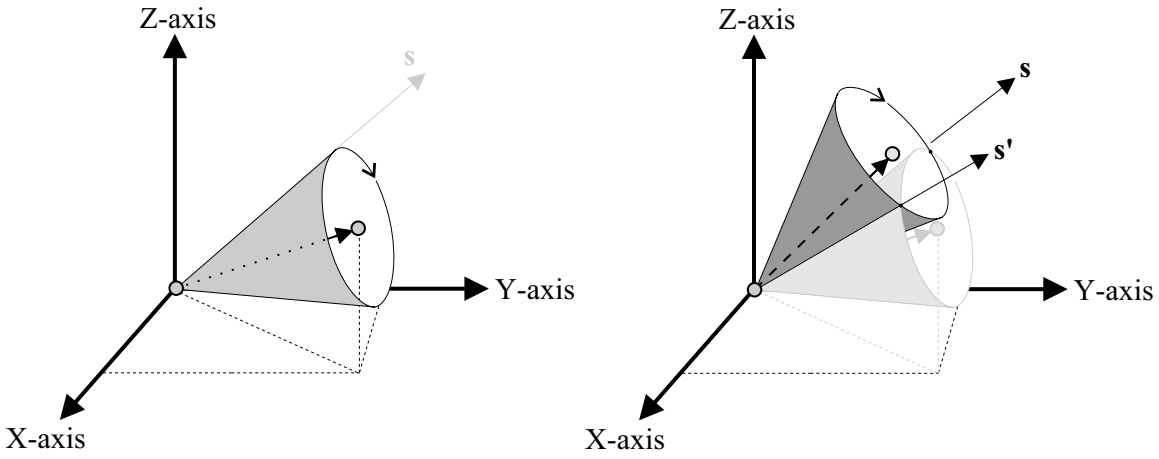

FIG. 4: Two different positions of the baseline yield two solution cones, which intersect at two lines denoted by **s** and **s′**, representing possible directions of arrival.

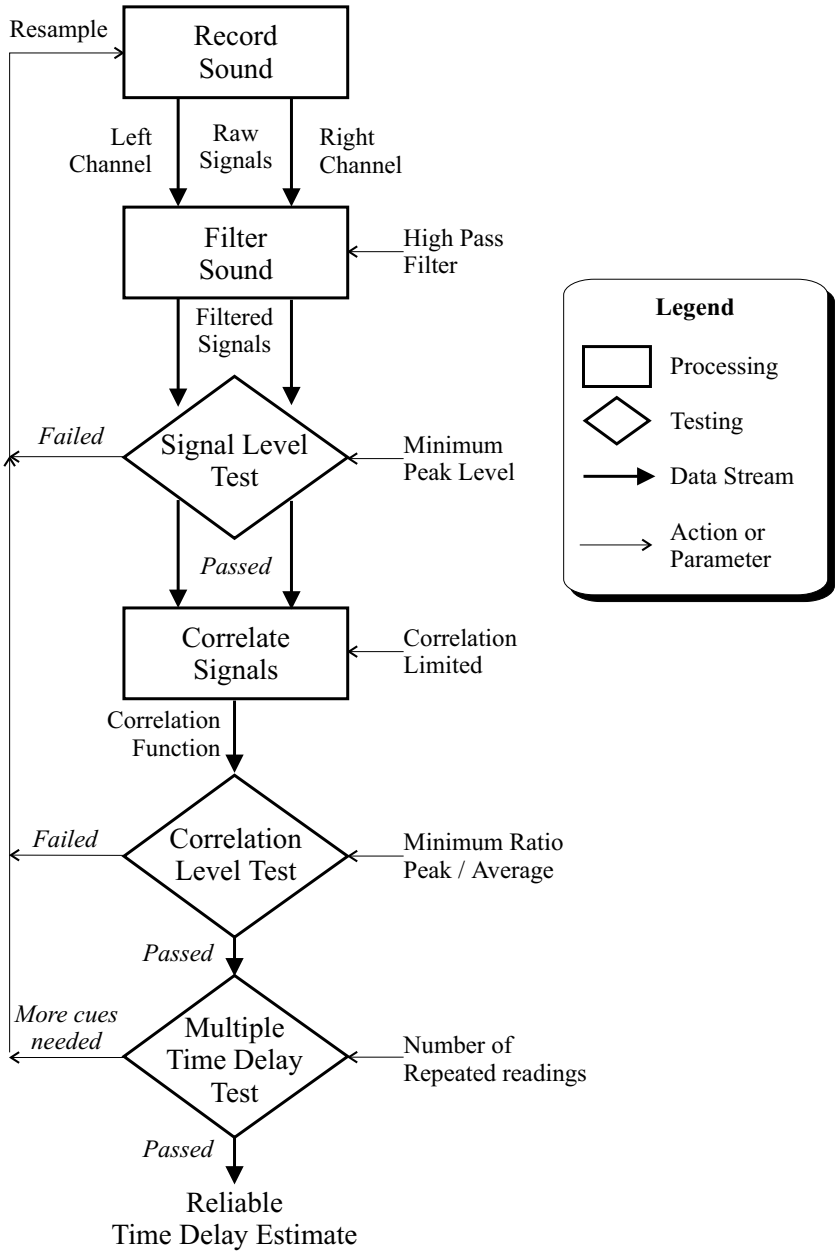

FIG. 5: Signal processing for time delay estimation.

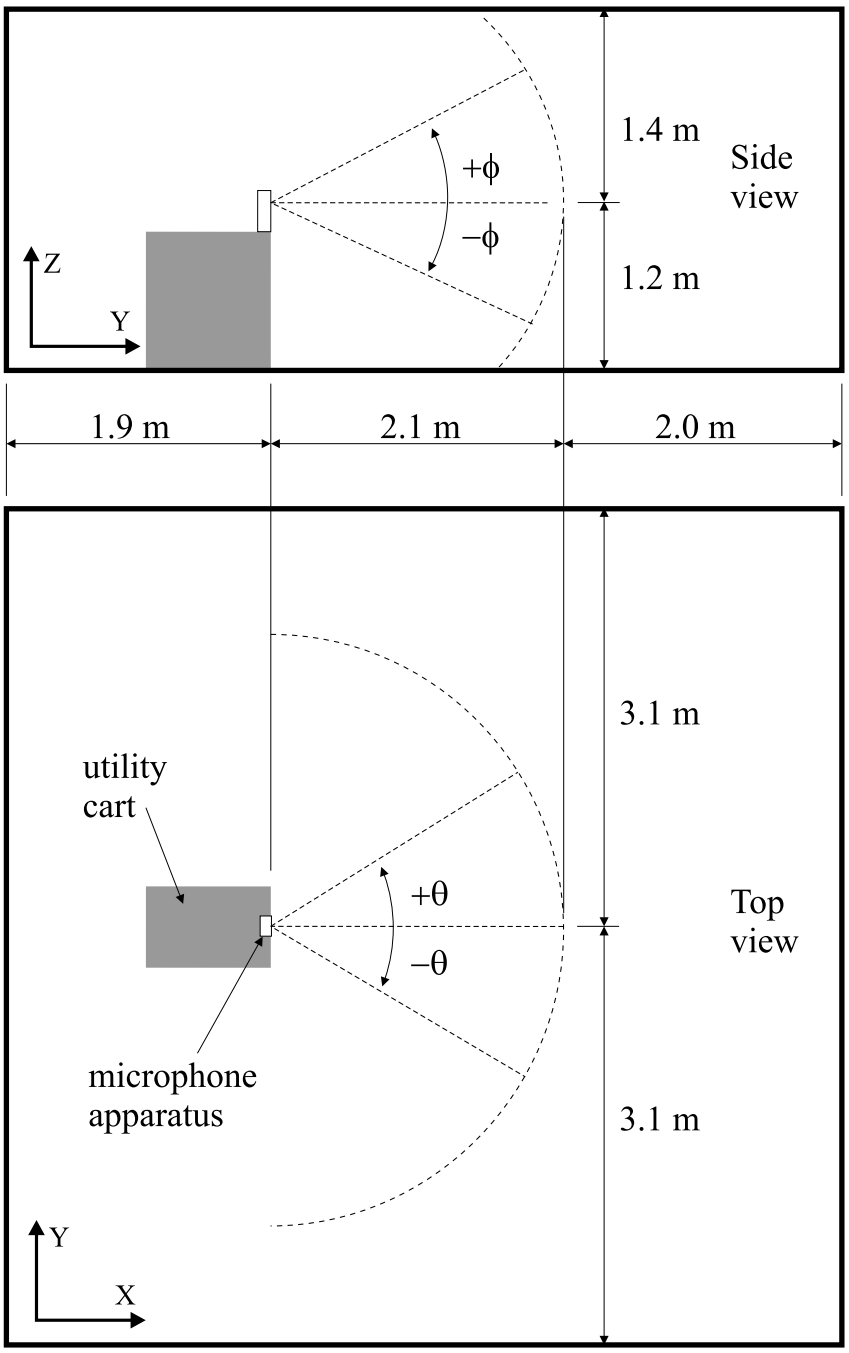

FIG. 6: The measurements for the experimental setup are shown above. The dotted arc represents possible positions of the sound source which are always kept at a constant distance from the microphone apparatus.

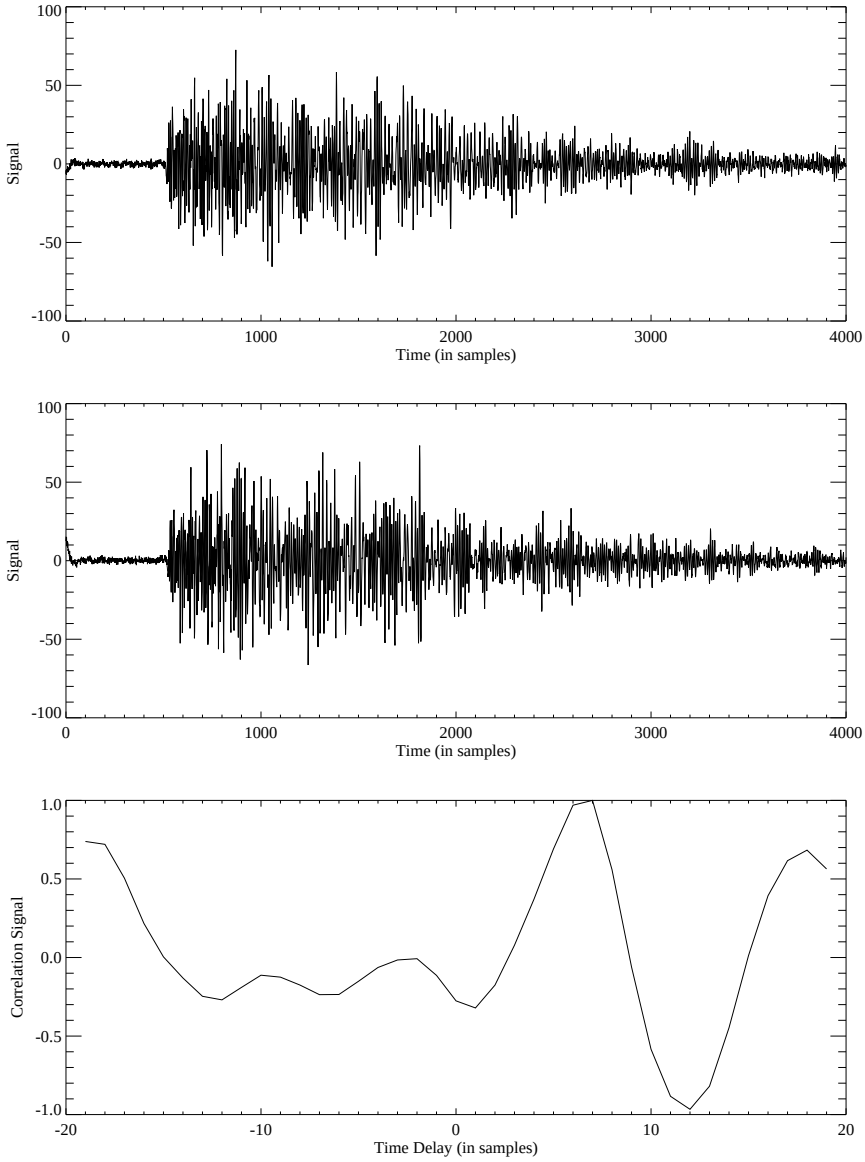

FIG. 7: The three plots show two channels of sampled data from the experimental setup (see section IV) and their correlation. Time and Time delay unit is one sampling period. The sound recorded is of a single hand clap, a good example of an impulsive sound (high peaks and short duration). Due to reflections, the resulting correlation has several smaller peaks. The strongest peak represents the correct time delay for the given experimental setup (7 samples).

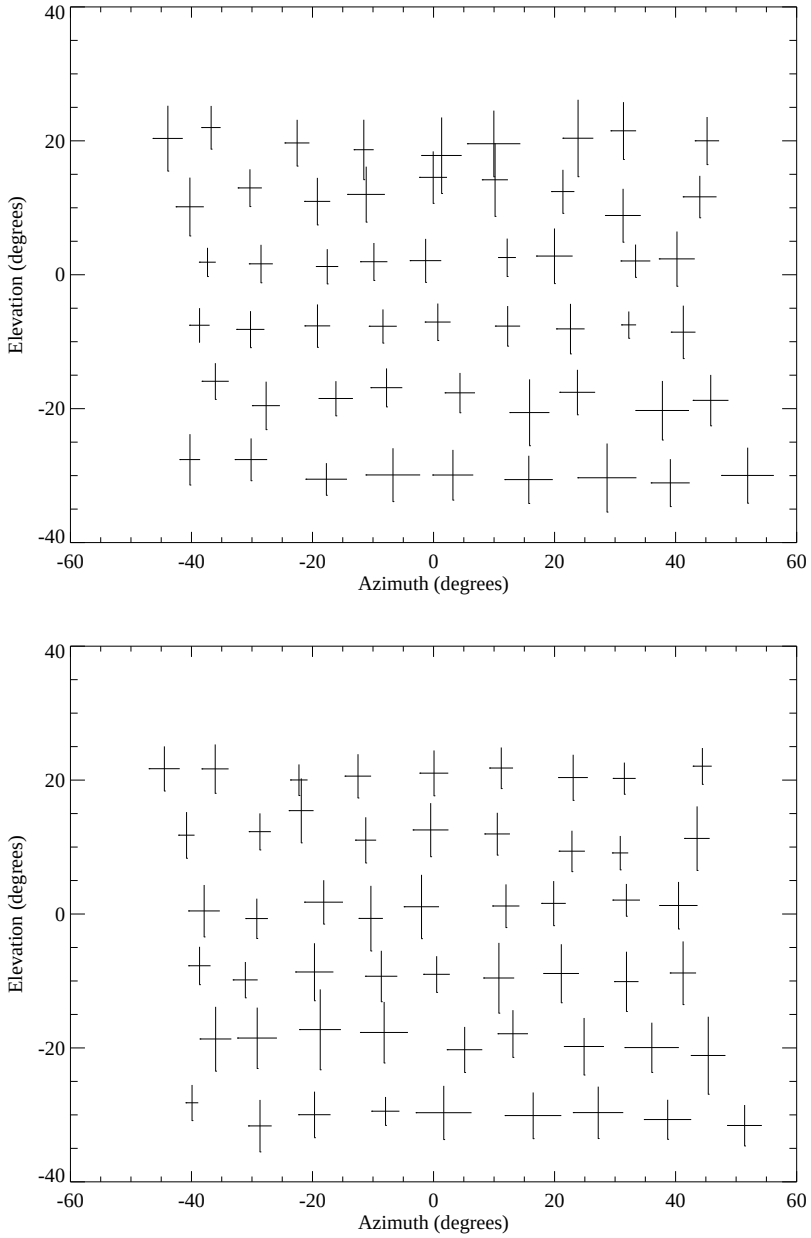

FIG. 8: The graphs show the estimated sound direction of arrival in both azimuth and elevation for an impulsive source (top) and a speech source (bottom). Each estimated position is computed from 25 estimates of the source in the same position. The size of the cross represents the standard deviation in azimuth and elevation, and the intersection of the standard deviation lines is at the estimated position.

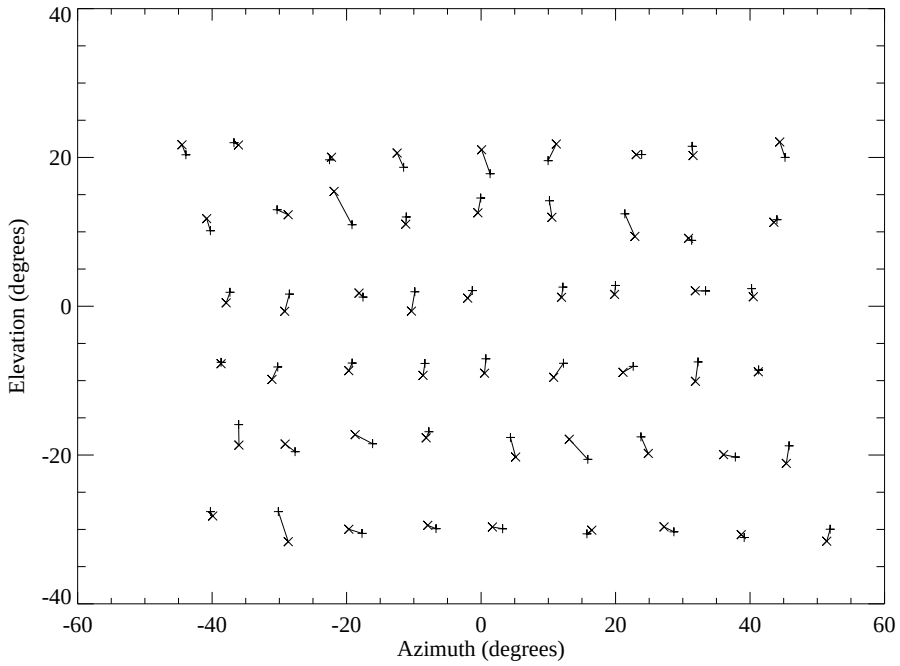

FIG. 9: The graph has the overlaid average estimated positions of the experimental data from figure 8. The estimated positions are '+' and 'x' for hand clap and voice respectively. Hand clap and voice estimates for the same source location are connected with a line.