

# 3-D Sound Localization from a Compact Non-Coplanar Array of Microphones Using Tree-Based Learning

Juyang Weng and Kamen Y. Guentchev

Department of Computer Science and Engineering  
Michigan State University  
East Lansing, MI 48824

**Keywords:** 3-D sound hearing, 3-D sound localization, supervised learning, classification trees, regression trees, principal component analysis.

### **Abstract**

One of the various human sensory capabilities is to identify the direction of perceived sounds. The goal of this work is to study sound source localization in three dimensions using some of the most important cues the human uses. In an attempt to satisfy the requirements of portability and miniaturization in robotics, this approach employs a compact sensor structure that can be placed on a mobile platform. The objective is to estimate the relative sound source position in three dimensional space without imposing excessive restrictions on its spatio-temporal characteristics and the environment structure. Two types of features are considered, interaural time and level differences. Their relative effectiveness for localization is studied, as well as a practical way of using these complementary parameters. A two-stage procedure was used. In the training stage, sound samples are produced from points with known coordinates and then are stored. In the recognition stage, unknown sounds are processed by the trained system to estimate the 3-D location of the sound source. Results from the experiments showed under  $\pm 3^\circ$  in average angular error and less than  $\pm 20\%$  in average radial distance error.

PACS classification number: PACS 43.58 (Acoustical Measurement and Instrumentation)

# 1 Introduction

A sound produced by a point-source generates acoustic waves with spherical symmetry, assuming uniform density of the surrounding air and absence of obstacles or other sounds. It is known that the location of the source can be established by detecting the front of the propagating wave and computing the center of the sphere [7] [8]. Unfortunately acoustic waves are not clearly distinguishable objects and such a task is not trivial in real environments even if real-life sources could be approximated by points [17]. Numerous studies have attempted to determine the mechanisms used by humans to achieve dimensional hearing [8] [12] [13]. Most phenomena have been reasonably explained in principle, although many aspects of human dimensional hearing need further study. It is known that two of the most important cues used by humans are the interaural differences: in time and level (ITD, ILD) [17] [21] [22]. Other cues relate to the spectral variations caused by diffractions at the head and pinnae [1]. For sounds with longer duration, cognitive processes start playing an important role, including dynamic head adjustments, high-level reasoning, etc. [22]. However, the computational steps in the use of spectral variations and other information by the human cognitive process are still not well understood. The purpose of the work presented here is sound localization by machines. We use only two low-level cues, ITD and ILD, in the work presented here.

## 1.1 Sound localization by machine

Sound localization can be used in many different applications: robot hearing, human-machine interfaces, monitoring devices, handicappers' aids, etc., where other means fail for different reasons. The obvious importance of building sound localization devices has prompted numerous efforts in the research community and a variety of techniques has been developed. Driven by concrete application needs, sensor setups of different implementations have seldom attempted to follow the human model. The number, size and placement of the sensors in such devices follow the specific needs of the task and are optimized for accuracy, stability, ease of use, etc. For example, a number of microphone subarrays have been placed on the walls with a goal to pick up the location of a speaker in a room [5] [4] [3] [19]. In other studies a human model has been followed to some degree resulting in constraints in applicability and limited accuracy [18]. A significant amount of work has been devoted to devices with a limited functionality (e.g., constrained to localization in a single half-plane while still using large sensor structures) [6] [19] or the help of a non-acoustical modality has been used (e.g. vision)[6].

In contrast to large, fixed sensor arrays for special situations and environments, this work concentrates on

a compact, mobile sensor array that is suited for a mobile robot to localize 3-D sound sources with moderate accuracy. It can be positioned arbitrarily in space while being capable of identifying the relative position of an arbitrarily located sound source. It is necessary to point out that the human three dimensional sound localization capabilities, while amazingly accurate in some instances, often have very serious limitations. The precision depends on various characteristics of the perceived sound: spectral contents, envelope variability as a function of time, volume level, reverberation and echo, etc. It can be disappointingly low and in some instances totally inconclusive [12]. Sometimes it can be convincingly wrong (e.g., Franssen effect) [13]. One major difference between human and engineering setup is the number of sensors available.

Most authors distinguish a single parameter as the most significant factor for dimensional sound localization. It is the interaural time difference (ITD) of the sound as perceived by two sensors. Numerous studies report the ITD as the main cue in human dimensional hearing [21]. The clear geometrical representation of the problem makes it a favorite feature to be used when approaching such a task by a machine setup [2] [5] [3] [6] [10] [15] [16] [19]. Another cue known to have notable importance in human dimensional hearing is the interaural level differences (ILD). Surprisingly ILD have seldom been used in actual system implementations because they are believed to have unfavorable frequency dependence and unreliability [17] [18]. Another reason is the lack of an explicit and stable relationship between ILD and source location which would otherwise allow for a simple algorithmic solution to be derived [17]. The learning approach used in this study does not have such limitations and it benefits from the added cues, both ITD and ILD.

Finally, the processing of the extracted features is one of the dominating factors for the success of a localization procedure. Most works determine the ITD and then use either an iterative search algorithm to minimize a certain objective function [14] [18] [19], or an approximation model for which a closed-form solution can be derived [5] [4]. The former is relatively slow and thus, it may not reach real time speed. The latter introduces model errors and cannot use more feature types for better accuracy.

To use both interaural time differences (ITD) and interaural level differences (ILD) while effectively dealing with the complex nonlinear relationships among these feature measurements and the solution, our work employs a learning based approach. It consists of a training phase and a performance phase. In the training phase, sounds from known 3-D positions are generated for training the system, during which a fast retrieval tree is built. In the performance phase, the system approximates the solution by retrieving the top match cases from the retrieval tree. This flexible framework allows for the use of more than one type of feature, and to deal with the 3-D localization problem without imposing unrealistic assumptions about the

environment, despite the compactness of the sensor structure. As far as we know, this work is the first to use a compact non-coplanar sensor array for full 3-D sound localization.

In order to objectively evaluate the performance of the system, initially a linear search algorithm was used when searching for the nearest neighbors in the 12-dimensional input space. The obtained results were used to evaluate the correctness and the performance of the SHOSLIF procedure<sup>1</sup>. SHOSLIF achieves a high speed of retrieval due to its logarithmic time complexity  $O(\log(n))$ , where  $n$  is the number of cases learned and stored as necessary [20]. It was found that the results produced by SHOSLIF had identical precision with that of the linear search, while its performance speed was nearly 5 times faster.

## 1.2 Related experimental work

Extensive work on sound localization by microphone arrays has been performed by Brandstein et al. with a significant effort in the theory of speech based sound localization and a series of publications [2][5][4][3]. Novel methods for estimating ITD and using it for the localization are presented, among which are a pitch-based approach to time delay estimation and a closed-form location estimator. The implementations concentrate on room oriented solutions - microphone arrays placed in room walls. The sound localization is applied in three dimensions, with one of the dimensions (the vertical axis) having a lesser span than the other two (the height of the room is much smaller than its width or length). The reported accuracy of localization is very high - the space resolution is in the order of centimeters. The placement of the sensors relative to the sound source in Brandstein's experiments is different from our choice. So is their use of multiple microphone arrays, compared to a single array in our work. This makes it difficult to compare the accuracy of both experiments. In the case of room-oriented placement it is impossible to define a direction or distance from the arrays since the arrays surround the source. An absolute location measure is used instead. Furthermore, the application domain of the room-oriented implementation is constrained to indoor use, while a free standing compact array can be transported and used in more diverse environments.

A model for 3-D sound localization that uses both ITD and ILD is presented by Martin [18]. The work simulates the human auditory system and uses a binaural setup. Some assumptions and approximations of the real-life environment are made and as a consequence the system can only estimate angular direction but not distance. The experiments are performed in an idealized environment (e.g. noise is eliminated) and an angular accuracy of around 5° is achieved. All computations were performed offline. In our work estimating

---

<sup>1</sup>SHOSLIF stands for Self-organizing Hierarchical Optimal Subspace Learning and Inference Framework.

the distance to the source is shown to be a challenging problem and it is the parameter we measure with the least accuracy. In our experiment ambient noise is always present and the computations were performed in real time.

One of the applications with hybrid methodologies is the work of Bub et al. [6]. They use a second modality (vision) to improve the performance of the localization device by refining the location estimate, provided by the sound localization. Their approach involves the use of ITD only as determined by a linear array of 15 microphones for redundancy. This setup limits the localization domain to a single horizontal 2-D half-plane in front of the array. With acoustic means only, their system is estimated to achieve an angular accuracy of around  $5^\circ$  and under 10% in distance with background noise only. This error is estimated from a single test point in 2-D space in front of the array. No statistical estimates of the accuracy were presented. The authors indicate that competing noise can significantly degrade the accuracy of localization.

Another work with a practical implementation is by Rabinkin et al. [19]. Their system uses two subarrays of 4 coplanar microphones each, placed on room walls. It employs a DSP to estimate ITD by using a cross-power spectrum phase algorithm. Then a space search algorithm minimizes the error between estimated and computed delays to produce a location estimate. The performance with the algorithms being run offline was reported as a percentage of deviation from a preset bound of  $6^\circ$  and was reported as being generally lower than 20%. The performance from the online tests was reported as being from “moderately well” to “quite poor.”

A number of other publications [1][9][10][12][13][14][15][16][22] treat theoretical aspects of auditory localization and the various methodologies used in practical sound source localization. However, there is no experimental work or actual implementations reported. The problem of full three-dimensional sound source localization with a compact mobile structure, as examined by this paper, has not been studied in any existing works that we are aware of.

## 2 Theoretical Issues and Sensor Structure

This work targets versatile applications such as the dimensional hearing of a mobile robot. For this reason we cannot use room-oriented solutions [4] [19], which typically use a large intersensor distance, with all the sensors fixed in the room. In our case the sound source will necessarily be located outside of the sensor structure. Furthermore the distance to the source will generally be significantly larger than the span of the

sensor structure. Most of the sound sources that are of interest for the purposes of sound localization are compact enough to be assumed point-sources. If the source cannot be approximated by a point then the problem of localizing that source is different from what we address here and thus is outside the scope of this work. The same applies to the case of multiple sources with comparable sound intensity. To determine the minimum number of sensors and their optimal placement, we need to look into the geometrical aspects of the problem.

## 2.1 Sensory measurements

From the front of the spherical acoustic wave, the two main parameters of measurements for each pair of sensors are ITD and ILD. Assuming that the speed of sound is constant, which is true only for uniform media (density, temperature, chemical and physical contents, etc.), ITD is equal to the difference of the distances between each of the detectors and the sound source, divided by the speed of sound. For simplicity we leave the constant out of the equation:

$$\text{ITD} \sim r_1 - r_2, \quad (1)$$

where  $r_i$  is the distance between the sound source and the  $i$ -th microphone,  $i = 1, 2$ , and  $\sim$  indicates proportionality (Fig. 1). The constant (the speed of sound in air) is irrelevant to the localization and is left

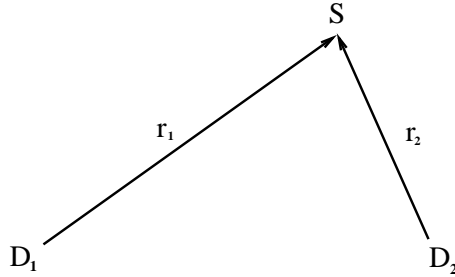


Figure 1: Two detectors,  $D_1$  and  $D_2$ , and the sound source  $S$ .

out. Also, since the amplitude of the sound wave, or the intensity of the sound, varies inversely with the square of the distance, the ILD is proportional to the difference between the inverse values of the square of the distance. However if we take the difference, we will be confronted with high-order terms in the equations which will lead to unnecessary complication of the computations. A much simpler form is provided by the ratio of the two values:

$$\text{ILD} \sim \frac{r_2^2}{r_1^2} \quad (2)$$

Both parameters in (1) and (2) can be estimated from the signals, detected by a pair of microphones.

## 2.2 Uniqueness

The major theoretical issues that need to be discussed here are (1) whether there is a unique solution to the 3-D location of the sound source, (2) how many sensors are minimally required for a unique solution, and (3) the reliability of the solution. In this section, we consider the first two issues that are related to uniqueness. The third issue will be discussed in Sec. 2.3.

It is known that spatial hearing in humans involves spectral analysis, which is beyond the scope of this paper. For the uniqueness issues related to our experimental system, we will consider three cases here: using ITD only, using ILD only and using both. Further implications are necessary to make the theoretical uniqueness problem tractable. We assume that the sound source and every sensor are all compact enough so that we can consider each as a point source. Further, we also simplify the environment. We assume that the environment is filled with uniform still air, free of objects other than the sound source and microphones whose volume is negligible. Although the above assumptions are generally not exactly true in a realistic application environment, the results from the analysis will give us insight into the important issue of uniqueness. A proper system design may help to approximately satisfy these assumptions. For example, we use only light supporting material for our microphone array to minimize its sound absorption and reflection. It is worth noting that our actual sound localization method is based on learning and it does not use the mathematical relations discussed in this section.

A pure algebraic analysis of this complex problem seems neither tractable nor intuitive either. In the following analysis we choose to use an analysis approach that is largely geometric in nature. Consequently, our analysis considers general placements of sensors and it excludes various degenerate cases that defeat our original purpose, such as various coplanar four-detector placements.

### 2.2.1 Using ITD only

First, consider a sound detector pair which consists of detectors  $D_1$  and  $D_2$ . As shown in Fig. 2, without loss of generality, we position our 3-D coordinate system in such a way so that  $D_1$  is at  $(c, 0, 0)$  and sensor  $D_2$  is at  $(-c, 0, 0)$ . By definition, a hyperbola (in 2-D) is the set of points for which the difference of the distances

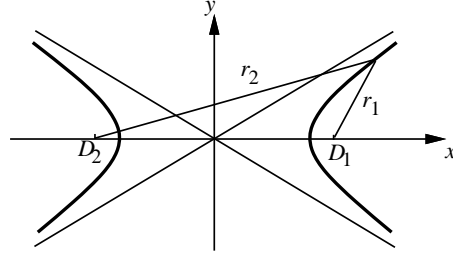


Figure 2: From two detectors, a given ITD determines one sheet of a hyperbola.

from two fixed points  $D_1$  and  $D_2$  is constant, as shown in Fig. 2. In 3-D, such points lie in a rotational hyperboloid whose intersection with  $xy$ -plan is shown in Fig. 2, where the axis of symmetry is the  $x$ -axis.

Suppose that the distances between a point  $X = (x, y, z)$  and  $D_1$  and  $D_2$  are denoted by  $r_1$  and  $r_2$ , respectively. The ITD gives  $r_2 - r_1 = \text{ITD} = 2a$ , where we define variable  $a$  is equal to a half of the ITD. If  $a$  is positive, only the right sheet in Fig. 2 can contain the sound source. Otherwise, only the left sheet can. If  $a = 0$ , the two sheets of the hyperboloid are degenerated into a plane  $x = 0$  in 3-D.

Next, consider three noncollinear detectors,  $D_1$ ,  $D_2$  and  $D_3$ , as shown in Fig. 3. The plane in which these three detectors lie is called *the detector plane* of detector 1-2-3. The hyperboloid determined by the ITD of detectors  $i$  and  $j$  is called hyperboloid  $i - j$ . Thus, hyperboloid  $i - j$  is the same as hyperboloid  $j - i$ . In Fig. 3, hyperboloid 1-2 and hyperboloid 1-3 intersect in 3-D to give a 3-D curve  $C_{123}$ . The third hyperboloid 2-3 (which is a plane in Fig. 3) does not give any additional constraint, since all the three sensors have been considered already by the first two hyperboloids 1-2 and 1-3. We define the *solution set* for the sound source as the set of all the possible 3-D locations of the sound source. Therefore, we have the following observation:

**Observation 1** *All the ITD measures from 3 noncollinear sound detectors generally define an infinite number of solutions for the 3-D location of sound source and the solution set is a 3-D curve.*

In Fig. 3, the hyperboloid 2-3 is a plane, which is a special case for a hyperboloid. Thus, the solution set is in a plane that is perpendicular to the detector plane 1-2-3.

Suppose that we add another sound detector  $D_4$  above the detector plane 1-2-3 (e.g., above  $D_1$  when the page is horizontally placed). Note that  $D_4$  cannot be coplanar with the other three nonlinear detectors to avoid degeneracy. The detectors  $D_1$ ,  $D_4$  result in a single sheet of hyperboloid 1-4. The intersection of the curve  $C_{123}$  and the single sheet of hyperboloid 1-4 give two 3-D points, one is the true sound source and the other is a spurious sound source. If the true sound source  $A$  in Fig. 3 is coplanar with  $D_1$ ,  $D_2$  and  $D_3$ , the

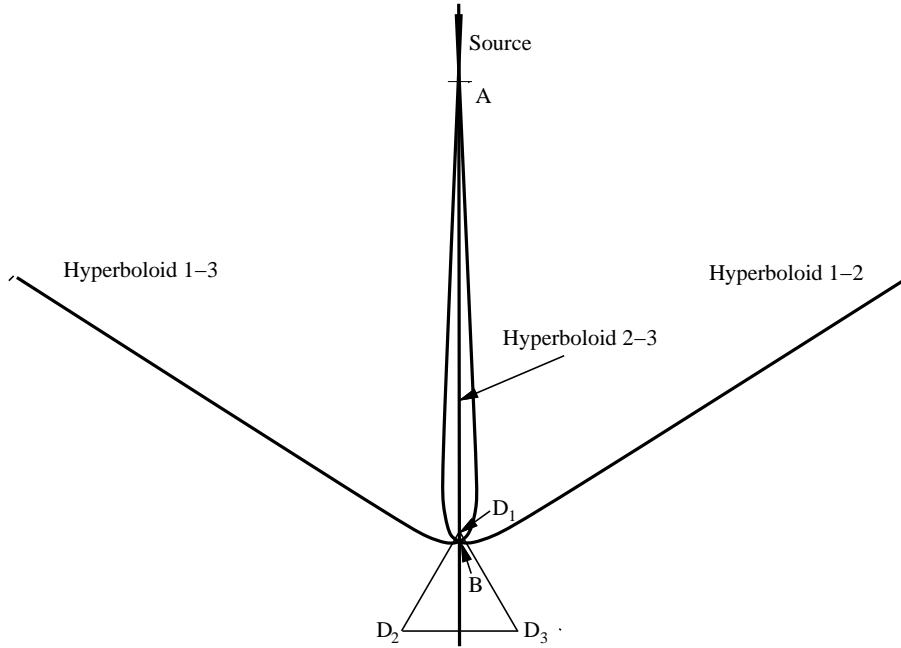


Figure 3: From three detectors, all possible ITD measurements determine a solution curve  $C_{123}$  in the 3-D space, which is the intersection of any two hyperboloids. In the above situation, the plane in which the curve  $C_{123}$  lies is orthogonal to the page.

spurious sound source is point  $B$  in Fig. 3. From Fig. 3, we can see that typically one of the solution point is outside (or far from) the detector array and the other solution point is inside (or near) the detector array. We can view geometrically why there are only two solution points. In Fig. 3, suppose the ITD between detectors 1 and 4 are zero (by e.g., adjusting the position of  $D_4$  slightly), the one sheet of the hyperboloid 1-4 degenerates into a plane that goes through source  $A$  and the mid-point of the line segment connecting  $D_1$  and  $D_4$ . The intersection of the planar hyperboloid 1-4 with the closed curve  $C_{123}$  gives two points. We cannot expect that other hyperboloids formed by  $D_4$  with other sensors can reduce the multiplicity of the solutions, as we saw in the 3-detector case. Thus, we have the following observation:

**Observation 2** *All the ITD measures from 4 non-coplanar sound detectors generally define two possible solutions for the 3-D location of the sound source.*

We have used a geometric method to reach this result. In fact, a direct algebraic proof of this uniqueness involves intersection of three quadrics in 3-D and as far as we know there has been no known general closed-form solution for such a problem. Even if a closed-form were available there would probably be no sufficient

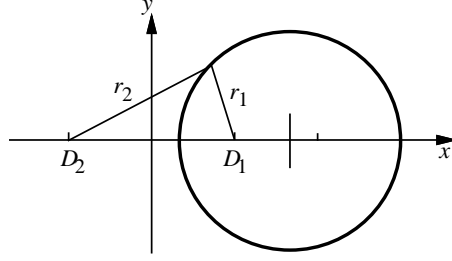


Figure 4: From two detectors, a given ILD determines a sphere.

space in this article to fully present it.

### 2.2.2 Using ILD only

We first determine the solution set from a single ILD measurement. Without loss of generality, we position the coordinate system and scale its axes so that the detectors  $D_1$  and  $D_2$  are at  $(1, 0, 0)$  and  $(-1, 0, 0)$ , respectively. Since ILD is proportional to  $r_2^2/r_1^2$ , let

$$\frac{r_2^2}{r_1^2} = \frac{(x-1)^2 + y^2 + z^2}{(x+1)^2 + y^2 + z^2} = \alpha$$

where  $\alpha > 0$ , with an exception  $\alpha = 1$  when the surface becomes a plane  $x = 0$ . The above equation can be rewritten as

$$(x-c)^2 + y^2 + z^2 = r^2$$

where

$$c = \frac{1+\alpha}{1-\alpha} \quad \text{and} \quad r^2 = \left(\frac{1+\alpha}{1-\alpha}\right)^2 - 1$$

In other words, given an ILD, the solution set is a sphere centered at  $(c, 0, 0)$  with a radius  $r$ , as shown in Fig. 4. Using three ILDs from three detectors, the solution set is the intersection of two spheres, sphere 1-2 and sphere 1-3, which gives a circle as shown in Fig. 5. Therefore, we have the following observation.

**Observation 3** *All the ILD measures from 3 noncollinear sound detectors generally define an infinite number of solutions for the 3-D location of sound source and the solution set is a 3-D circle.*

Suppose that a fourth detector  $D_4$  is added. The sphere 1-4 determined by detectors 1 and 4 intersects the circular solution set from detector 1, 2, and 3 at two points, one being the true location of the sound source and the other being the spurious one. As we can see, the fourth detector  $D_4$  should not be coplanar

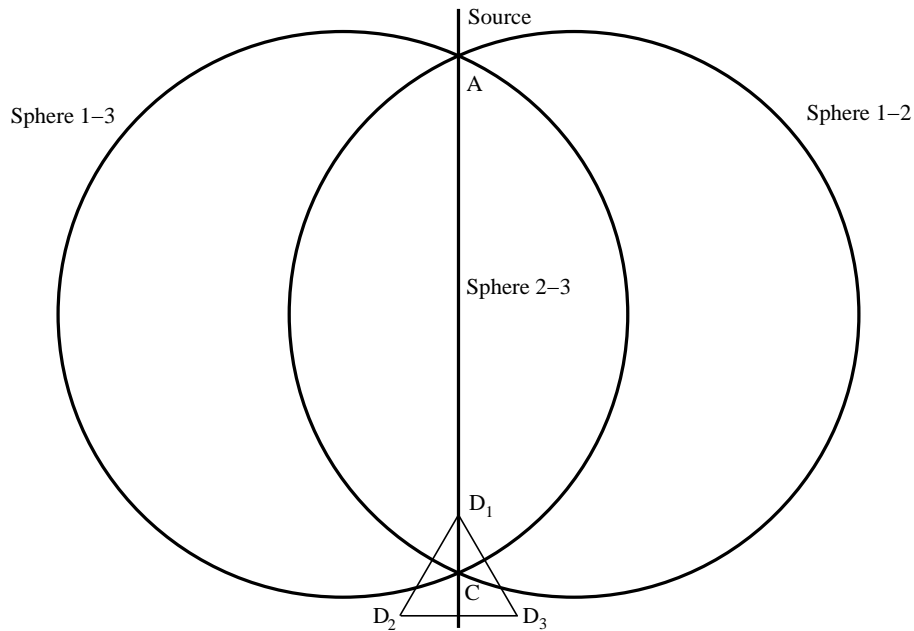


Figure 5: From three detectors, all possible ILD measurements determine a solution circle in the 3-D space, which is the intersection of any two spheres. In the above situation, the plane in which the solution circle lies is orthogonal to the page.

with the first three detectors. Otherwise, it does not reduce the degrees of freedom of the solution. This leads to the following observation:

**Observation 4** *All the ILD measures from 4 non-coplanar sound detectors generally define two possible solutions for the 3-D location of the sound source.*

### 2.2.3 Using ITD and ILD jointly

Suppose that both ITD and ILD are used from 4 non-coplanar sound detector array, as shown in Fig. 6. The spurious solution  $B$  from the ITDs is typically not the same as the spurious one  $C$  from the ILDs, as illustrated in Fig. 6. Therefore, we have the following observation:

**Observation 5** *All the ITD and ILD measures from 4 non-coplanar sound detectors generally define a unique solution for the 3-D location of the sound source.*

A question remains as to whether ITD and ILD measures from 3 noncollinear detectors are sufficient to give a unique solution. The answer is no. Consider the plane  $P$  in which the 3 detectors lie. Starting with

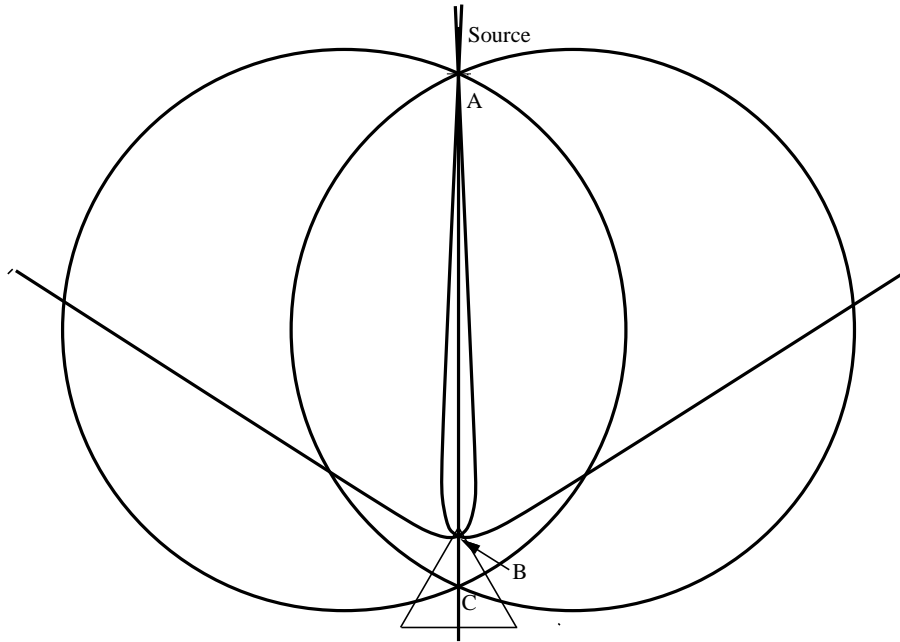


Figure 6: When ITD and ILD measurements are both used.

the true 3-D position of the sound source  $A$ , mirroring  $A$  with respect to the plane  $P$  as a mirror produces a superfluous point  $A'$  on the opposite side of the mirror. This superfluous point  $A'$  gives exactly the same ITD and ILD measures as  $A$ . Thus, we have the following observation:

**Observation 6** *All the ITD and ILD measures from any 3 sound detectors do not define a unique solution for the 3-D location of the sound source.*

The above reasoning is also applicable to any number of coplanar sensor arrays. Using a similar reasoning using the plane in which the sensors lie as the mirror plane, we get the following observation.

**Observation 7** *All the ITD and ILD measures from any number of coplanar sound detectors do not define a unique solution for the 3-D location of the sound source.*

We have reached our conclusions about uniqueness using geometric reasoning, which is more intuitive, more informative and easier to understand for our geometric problem than a purely algebraic approach. We will further see these advantages of geometric reasoning in the following section when we investigate the stability of these solutions. We have tried an algebraic approach but we were not successful to reach a solution, partially due to the fact that a closed form algebraic solution for the intersection of general quadrics does not exist. In other words, we can geometrically reason about these solutions but we cannot express the

solutions in an algebraically closed form. As far as we know, our uniqueness analysis presented here is the first complete one for 3-D that covers both ITD and ILD.

### 2.3 Solution stability

The above discussion about how the solutions are determined is also very useful for providing insight into the stability of the solution from ITD and ILD features.

We are interested in determining the relative 3-D position of the sound source from an array of non-coplanar microphones. The relative position can be determined by the displacement vector from the center of the detector array to the position of the sound source. This displacement vector can be specified by the 3-D *direction* of the vector and the length of the vector. The direction tells in which direction the sound comes from. The length tells the *distance* from the detector array to the sound source.

In Fig. 3, we can see that the 3-D location of the sound source is determined by three surfaces. With measurement errors, the position of three surfaces has also certain amount of error. We can imagine that the amount of error in the position of each surface is represented by the thickness of the surface. In other words, the hyperboloid surfaces in Fig. 3 have a certain degree of thickness. Further, the thickness is not constant for each hyperboloid. The farther the surface patch is away from the detector array, the thicker the part of the surface is. As shown in Fig. 3, the 3-D location of the sound source  $A$  is determined by the intersection of three surfaces that are almost parallel and vertical (assuming that the diagram is a top view when the page is horizontal). Therefore, ITDs give a bullet-like uncertainty region with the long axis of the bullet pointing toward the center of the detector array, as shown in Fig. 3. In other words, the distance in the solution is less reliable than that for direction.

A similar reliability analysis can also be applied to ILD. In Fig. 5, we can see that the 3-D location of the sound source is determined by the intersection of three spheres. If only sphere 1-2 and sphere 1-3 are used, the uncertainty region after their intersection will be like a disk, whose rotational axis is aligned with the line of sight from the center of the detector array to the sound source. The additional intersection with sphere 2-3 will probably reduce the width of this uncertainty disk by a minor amount, because the thickness of the sphere 2-3 is large at this long distance. Thus, ILDs give a disk-like uncertainty region with the normal of the disk pointing toward the center of the detector array. Comparing the bullet and disk shapes of the uncertainty regions of ITD and ILD, respectively, it appears that ITD is relatively better for direction estimate and ILD is relatively better for distance estimate. This observation has been confirmed by our

experimental data.

Since each ITD or ILD is separately estimated by comparing the sound signals from two microphones, each of ITD and ILD from every pair of detectors is potentially useful in providing over-determination for combating noise. This is why we will use all the possible ITD and ILD measures in the estimation of 3-D sound location in our system.

## 2.4 Number and placement of detectors

From the above analysis, we know that if four non-coplanar detectors are used the intersection is unique. A mobile robot requires that the structure of the sensor array be compact, while accuracy consideration requires a large array. Thus, an equidistant structure seems reasonable. In the case of four sensors this suggests a tetrahedron (Fig. 7). In our experiment, an equal-side tetrahedron with a 20cm side was used. At each apex of tetrahedron is a miniature microphone. Light carton paper was used to support the microphone array, without causing significant sound absorption.

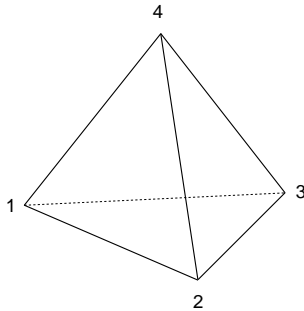


Figure 7: Placement of the 4 microphones on the array.

## 3 Method

As outlined above, efficient sound localization is possible after having extracted the necessary ITD and ILD measures. We have shown in Sec. 2.2 that the minimum number of detectors required to obtain unambiguously a solution in 3-D space is four and that it is unique in general. In order to fully solve the sound localization problem, two main steps need to be performed. Step 1, extract ITD and ILD measures from the acquired signal. Step 2, estimate the actual sound location using those feature measures.

### 3.1 Feature extraction

As discussed, the two features considered in this work are ITD and ILD. For successfully detecting ITD, we like to avoid narrow-band acoustic signal such as a tone of a single frequency, since such signals tend to be nearly periodic in time. ITD will have severe a ambiguity problem when it is computed from a periodic signal. In practice, we use a human voiced sentence as the sound source, since it is typically the case for our robot applications. A microphone signal from such human verbal sounds contains low frequencies as well as higher order harmonics and thus has a relatively wide frequency band.

Using the appropriate hardware, the acoustic signal can be first converted to an electrical signal (by microphones) and then to a digital form (analog-to-digital converter board). The digitized data is a sequence of values representing the temporal waves of the sound, as picked by the respective detector for a determined period of time. A window of sufficient duration is used to define the searchable domain. Some preprocessing is applied to ensure satisfactory quality of the sound sample. For instance, the amplitude of the signal’s envelope can vary over time. With speech this corresponds to, e.g., accents and pauses within and between words. These sound blanks contain little useful information and using them can degrade the quality of the estimates. In order to avoid this problem, the window is divided into smaller intervals in which the variation of the signal is evaluated (Fig. 8). This preprocessing selects only signals with high variance for feature extraction. A measure of the “efficiency” of the sample is returned by the procedure as the percentage of used subframes in the whole window.

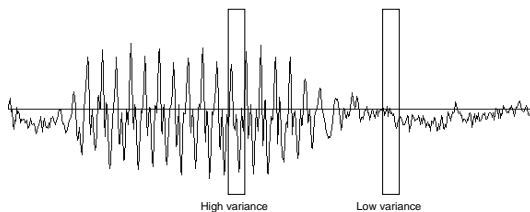


Figure 8: Preselection according to signal variance

The next phase involves the use of a cross-correlation procedure to determine the shift between the sampled signals at each of the sensor couples. This gives a direct measure for the ITD [6]. We find the peak of the cross-correlation function varying across a range of possible time-delays. Suppose that  $\{x_i, x_{i+1}, \dots, x_{i+N+n-1}\}$  is the digital signal from the first microphone channel and  $\{y_i, y_{i+1}, \dots, y_{i+N+n-1}\}$  from the other channel, where  $N$  is the length of the segments, in number of samples, used for correlation

and  $n$  is half of the maximum possible time delay. We can define the normalized cross correlation at time  $i$  with shift variable for the first microphone channel  $j$  and shift variable for the other channel  $k$  as

$$R_{j,k} = \frac{\sum_{i=0}^{N-1} (x_{i+j} - \bar{x}_j)(y_{i+k} - \bar{y}_k)}{\sqrt{\sum_{i=0}^{N-1} (x_{i+j} - \bar{x}_j)^2} \sqrt{\sum_{i=0}^{N-1} (y_{i+k} - \bar{y}_k)^2}} \quad (3)$$

where  $\bar{x}_j = \sum_{i=0}^{N-1} x_{i+j}/N$  and  $\bar{y}_k = \sum_{i=0}^{N-1} y_{i+k}/N$ . Using the above we can find the maximum correlation by successively trying shifts in each direction:  $R_x$  to cover the case when the source is closer to the first sensor and  $R_y$  - when the source is closer to the other sensor:

$$R_x = \max\{R_{j,k} \mid j = 0, 0 \leq k < n\} \quad (4)$$

$$R_y = \max\{R_{j,k} \mid k = 0, 0 \leq j < n\} \quad (5)$$

Then,  $R_{max}$  is the maximum cross-correlation:

$$R_{max} = \max\{R_x, R_y\}. \quad (6)$$

The sign of the time-shift  $T$ , is determined by how  $R_x$  and  $R_y$  in (6) compare in magnitude. The absolute value of  $T$  is proportional to the value of  $j$  or  $k$  that maximizes the value in Eqs. (4) or (5), depending on the sign of  $T$ . More precisely it is the product of that number and the digitization interval (the inverse of the sampling frequency). The value at the maximum is selected and is returned along with a parameter reflecting the sharpness of the correlation curve at the peak (Fig. 9). A combination of those two parameters is used as a quality estimate for the ITD (“score”). A high value of  $R_{max}$  is an indication of good similarity between the signals from both channels. However, a high value of  $R_{max}$  does not mean that the ITD is reliable. The second parameter, however, discriminates between a wide, flat correlation curve and a narrow, sharp one. The latter tends to give a more accurate ITD.

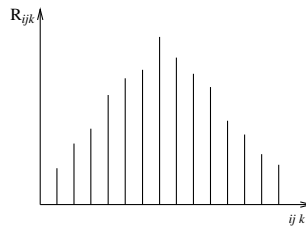


Figure 9: Typical cross-correlation curve with good sharpness

We should note a very useful side effect, which is closely related to the precedence effect in human auditory perception [12] [13]. In case there are reflections of the incoming sound from incident surfaces

(echoes, reverberation of broad spectrum sounds [9]) a secondary maximum will appear in the correlation curve. However, with the presented approach, it will be ignored because the similarity of that signal will be significantly lower and thus it will correlate less well (lower and wider peak).

Using the obtained information for the ITD, it is possible to evaluate the ILD by computing an integral value of the signal energy from the shift-adjusted signal. To estimate the signal energy we can use the variation in time of the voltage output of each microphone  $S(t)$ . The value for microphone pair 1 and 2 is shown in Equation (7).

$$\text{ILD}_{12} = \frac{\int_{t_i}^{t_j} |S_1(t)| dt}{\int_{t_i}^{t_j} |S_2(t+T)| dt} \quad (7)$$

where  $S_1(t)$  is a measure of the signal picked by microphone 1,  $S_2(t+T)$  - by microphone 2,  $T$  is the previously determined time shift and  $t_j - t_i$  is the length of the sample window. We should note that the sample window is chosen such that it is much larger than the maximum possible time difference between any pair of microphones (500ms vs 0.5ms).

The estimates for ITD and ILD are considered reliable only if the efficiency and score of the sample are satisfactory, i.e., above a predefined threshold. Thus the described procedure not only extracts the needed features but also information about whether the feature can be used for localization or it should be discarded as useless.

### 3.2 Tree-based learning

Once the ITD and ILD are extracted from the signal picked up by the detector array, the next step is to perform the actual sound source localization. The discussed disadvantages of the currently available methods can be avoided to a large extent by taking a learning-based approach. As stressed above in Sec. 2.2, these features uniquely define a solution and thus we have a direct correspondence between the extracted ITD and ILD values and the 3-D coordinates of the sound source. Of course, in the presence of distortions, echoes and noise of various natures, all the ITD and ILD measures are not consistent and we like to use over-determination to combat those effects that have not been modeled by our theoretical exposition.

In the current case the input feature space  $X$  is 12 dimensional: 6 for ITD and 6 for ILD (one for each combination of detector pairs). The output space  $Y$  is 3-dimensional:  $(y_a, y_e, y_r)$  representing azimuth angle, elevation angle and radial distance, respectively. This polar representation in output space is used because we know from the above reliability analysis that the uncertainty regions for ITD and ILD are both elongated in the radial direction. The polar representation can better isolate the errors in the resulting components.

Thus the mapping to be computed is a mapping  $f : X \mapsto Y$ , where  $X$  is the 12 dimensional input space and  $Y$  is the 3-dimensional output space.

We should note the extreme complexity of this actual mapping. The closed-form solution has not been found even in an ideal situation (e.g., no echo is considered). The existing methods have used an approximate closed-form solution or an iterative search procedure to minimize a nonlinear objective function. The former type suffers from the model error and the latter type either requires time in an exhaustive search or does not give a guarantee of global minima in the nonlinear objective function to be minimized (the practical situations may vary). Further, in a real situation, the model used should take into account the working environment, such as the material of the walls and objects in the environment as well as their acoustic properties. Such a complete model is extremely difficult to construct and even if it is constructed, the generality of the method is very limited since the model is only applicable to a particular known environment.

We use a learning based method by constructing a regression tree from a set of training samples  $t = \{(x_i, y_i) \mid y_i = f(x_i), x_i \in X, y_i \in Y, i = 1, 2, \dots, n\}$ . Although the system trained in one environment can be, in principle, only used in the same environment, our method can be used to train the system in virtually any environment. If the accuracy requirement is not very high, the system can be trained using a training set that uses training data collected in a variety of environments.

If the output space  $Y$  consists of a list of discrete class labels  $\{l_1, l_2, \dots, l_n\}$ , the problem of constructing  $f$  is called a classification problem. If the output space  $Y$  is a numerical space, the problem is called a regression problem. Therefore, our problem here is a regression problem. Further, the number of samples  $n$  is typically large, which is an important issue for real time application. After input vector  $x$  is computed, the corresponding estimated output vector  $y' \approx y = f(x)$  must be computed very quickly. Thus, it is not possible to search all the possible training pairs in  $L$ . The goal of our regression tree method is to organize the training samples by a tree data structure so that top  $k > 0$  best matched samples  $(x_i, y_i)$  in the training set  $L$  can be found quickly without having to examine all the training samples in  $L$ .

We present our SHOSLIF recursive partition tree (RPT) which we used to approximate the mapping  $f$ . Fig. 10 illustrates how the training samples are organized by the RPT. The input space  $X$  is depicted as a 2-D space in Fig. 10 (a), where the location of each number pair  $i, j$  represents the  $x$ -part of a training sample  $(x, y)$ . The number  $i$  corresponds to level in the tree shown in Fig. 10(b) and  $j$  denotes the index of the nodes at this level. The entire space  $X$  is represented by the root of the tree. The children of the root further divide the space of the root into smaller cells. The grandchildren of each node further divides the

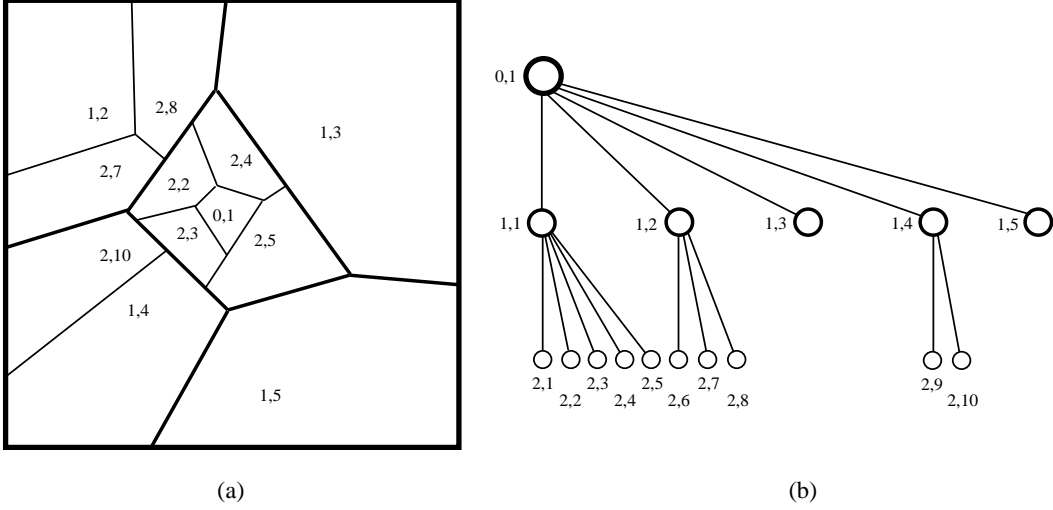


Figure 10: A 2-D illustration of a hierarchical Dirichlet partition and the corresponding recursive partition tree (RPT). (a) The partition, where the label indicates the center of a cell. The label of the child to which its parent’s center belongs is not shown due to the lack of space. (b) The corresponding recursive partition tree.

space of its parent into smaller cells. Such recursive division is carried on until the resulting cell contains only one training sample and the corresponding node is called a leaf node. In the experiment, we set the number of children to be two, and thus the tree used for the experiment was a binary tree in that each non-leaf node has exactly two children (a special case of that in Fig. 10).

How do we divide the cell of each node? We certainly do not want to divide the space in such a way that one node has all the samples in the parent and the other has no sample at all. Better, we want to divide the cell across the direction in which the samples  $x_i$ ’s spread wide. This is important since the input samples might lie in only a subspace  $X'$  of the input space  $X$  and the dimensionality of  $X'$  may be significantly smaller than that of  $X$ . Suppose  $S = \{x_1, x_2, \dots, x_m\}$  are the samples in a node  $A$  ( $A$  is the root of the tree to start with). We compute the first principle component vector  $v \in X$  from the samples in  $S$ , which is the eigenvector of the sample covariance matrix of  $S$  corresponding to the largest eigenvalue [11]. In other words, along vector  $v$ , the samples in  $S$  has the greatest variance than any other direction. Then, we compute the projection  $a_i$  of  $x_i$  onto  $v$ .  $a_i = \langle x_i, v \rangle$  where  $\langle \cdot, \cdot \rangle$  denotes the vector inner product. The mean of  $a_i$ ,  $\bar{a} = \sum_{i=1}^m a_i / m$ , is used to compute the position of the partition hyperplane along vector  $v$ . Every sample in  $S$  is assigned to one of the two children in the following way: For each  $x_i$ , if its projection  $a_i$  is less than  $\bar{a}$ ,  $x_i$  is assigned to the left child. Otherwise,  $x_i$  is assigned to the right child. In other words, a hyperplane

whose normal is  $v$  and whose position is at the centroid of the projections of  $S$  along  $v$ , divides the samples in  $S$  to the two children. Such a division process for the root results in two children. The samples assigned to each child is further subdivided, recursively, until the resulting child has only one sample.

After the RPT has been constructed, it is used as follows. Given a measurement vector  $x \in X$ , we use the RPT to estimate the approximate output  $y' \approx y = f(x)$  by querying the RPT. Starting from the root, decide to which child  $x$  belongs using the way in which the samples are assigned to the children. Only that child is further explored recursively. Finally, a leaf node will be reached which stores one sample  $(x_i, y_i)$ .  $y_i$  can be used as the estimate for  $f(x)$ . It is worth noting that this tree retrieval process may not always give the best matched  $x_i$  for all possible  $x$ . Thus, instead of exploring one path down the tree, we explore  $k > 1$  parallel passes by keeping the best  $k$  nodes at each tree level for further exploration. Finally,  $k$  leaf nodes have been reached.  $y_i$  vectors of these top- $k$  matched leaf nodes are used to give an interpolated output vector  $y \in Y$ , given input vector  $x \in X$ .

$$y = \frac{1}{\sum_{i=1}^k w_i} \sum_{i=1}^k w_i y_i \quad (8)$$

where  $w_i$  is the weighting function and  $y_i$  is the output part of the  $i$ -th nearest neighbor of  $x_i$ . The weights are determined in such a way that the nearest neighbor has the highest weight while others have less weight, depending on each distance from  $x$ :

$$w_i = \alpha^{-\frac{\|x - x_i\|}{\|x - x_0\| + \epsilon}} \quad (9)$$

where  $x_0$  is the nearest  $x_i$  from  $x$ ,  $\alpha > 1$  is a decreasing factor and  $\epsilon > 0$  is a small constant to avoid denominator from going to zero. For example, when the distance  $\|x - x_i\|$  is twice as large as that from the nearest neighbor:  $\|x - x_0\|$ , the corresponding weight is decreased by a factor  $\alpha^{-1}$ . This way, the RPT can give a reasonably interpolated output from  $k$  near samples even if the nearest neighbor was not among the  $k$  leaf nodes provided by the RPT. The time complexity of computing the estimated output from the RPT is roughly the number of levels in the tree, which is roughly  $O(k \log_2(n)) = O(k \log_2(n))$  where  $n$  is the number of samples in the entire training set  $T$ . For more detail about SHOSLIF RPT and its performance advantages over other general function approximators such as the feedforward neural networks and the radial function networks, the reader is referred to [20].

## 4 Experiments

In order to test the methodology, an experimental setup was used to perform a number of tests. A set of four identical Lavalier microphones was placed at the tips of a solid tetrahedron with 20cm side (Fig. 7). The signal from the microphones was amplified by four modular microphone preamplifiers to bring the power of the signal level to a required range. It was then supplied to an analog-to-digital converter board mounted in a personal computer. The software was designed to visualize, train, recognize and run various sound localization related tasks. Samples were taken from various points with known 3-D coordinates, some were used for training and others for testing. The results were compared with linear search and the performance of SHOSLIF was evaluated.

### 4.1 Experimental environment

The dedicated hardware was built from off-the-shelf consumer and industrial quality items. All experiments were held in the Pattern Recognition and Image Processing laboratory in the Department of Computer Science at Michigan State University, which is hardly suitable for high precision acoustic experiments. As shown in Fig. 11), the test space was located in the middle of the laboratory, in between cubicles with



Figure 11: The experimental setup. The compact sensor array is shown at the top of the pole standing on the edge of the center table. The intersection points on the grid on the floor mark the control points in the work space.

computers and reflecting, and absorbing surfaces of irregular shape. The number of devices producing weak to strong noise of different frequencies and levels was above 20. Often laboratory members would speak softly in the background while samples were being captured for training or retrieval. A room with better acoustic properties (e.g., anechoic chamber) is not consistent with our goal. The laboratory environment such as the one above was close to the one in which our actual sound localization device would be exposed

to in our intended applications.

## 4.2 Experimental setup

At the training stage a continuous sound, originally produced by a human speaker uttering a short sentence, was reproduced using a hand held tape recorder, from a set of previously defined locations (Fig. 12). Without



Figure 12: Training the system. The compact sensor array is shown at the upper left corner of the figure. A tape recorder plays the sentence at a predefined height right above a floor grid point. A weight is used to guarantee vertical displacement between the grid point and the tape recorder.

significant loss of generality, the span of the training grid was set to an arbitrary section of  $3 \times 3 \times 2.1$  meters, with the microphone array in the middle of one of the sides. The density is linear in Cartesian coordinates with a granularity of 0.3m. However, only 237 of the defined 700 points were used for training. They were selected to simulate uniform angular density and ten samples were taken from each of those points. The approximate angular density of the training points was around  $15^\circ$ . Thus the angular span of the training area was about  $180^\circ$  in azimuth and a little less than that in elevation.

At the performance test stage, we used the same sound source. Other voices were also used to investigate the variation of accuracy. We were able to produce a similar accuracy using sentences of similar style of speaking. We have also observed that the quality of the sound directly influences the reliability and thus the accuracy of location estimate. The quality includes, e.g., ratio of sounds over pauses and the compactness of the sound source. In reality, the sound source is not exactly a point source.

The system can be operated to collect 0.5s samples and provide location estimates based on each of them

or to store them for further analysis. The test samples were obtained from 79 positions the grid, but their locations are different from the ones used for training. They were used for computing estimation errors offline using another specially designed evaluation program. Estimates for the location of each of the test points were thus produced and recorded. They were compared to the known actual values of the 3-D coordinates and the error was computed as the difference between the actual and estimated value for the angles, and the ratio of that difference to the actual value, for the radial distance.

The algorithm used to compute the coordinates of the sound source uses two parameters for fine tuning its performance. One is the relative weight of the two input arguments: ITD and ILD. Because of the greater magnitude of the ITD, the ILD was multiplied by a variable factor, called *scaling on ILD*, thus increasing its relative weight as needed. This allows us to estimate the relative significance of those two parameters on the accuracy of the final results. A low value of this parameter would mean neglecting the ILD (a value of zero means only ITD are used), while a very high value indicates a predominance of the ILD. Their relative weight is practically equal when the value of *scaling on ILD* is around 13. The other parameter is the weight coefficient  $\alpha$  used in the interpolation of the retrieved nearest neighbors as defined in Eq. (9). A low value of  $\alpha$  would indicate that all considered nearest neighbors are almost equally weighted when estimating the solution (for  $\alpha = 1$  we have averaging) while a big value of alpha emphasizes the role of the nearest neighbor.

It is known that ITD and ILD are frequency dependent, e.g., ITD uses predominantly the low frequencies, while higher frequencies are the only ones that can be used for estimating the ILD. A preliminary signal filtering can be employed to leave only the useful frequencies when determining each of those two parameters. The actual response of those two filters can be another subject for fine tuning. However, the real-time implementation requirements for this project impose serious limitations on the amount of preprocessing that can be performed and thus spectral analysis of the signal is not adopted.

### 4.3 Results

The results obtained in this manner were used to study the above mentioned relations. A number of plots are used to show actual values and some observed trends. Fig. 13 shows how the relative weighing between ITD and ILD affects the accuracy of estimation of the azimuth, Fig. 15 for the elevation and Fig. 17 for the distance. The respective standard deviations are shown in Figs. 14, 16 and 18, respectively. The horizontal axes are the scaling on ILD — the coefficient by which ILD is multiplied when considered for estimating the nearest neighbors using interpolation using distance-based weighting (see Eq. 8). The range on the axes was

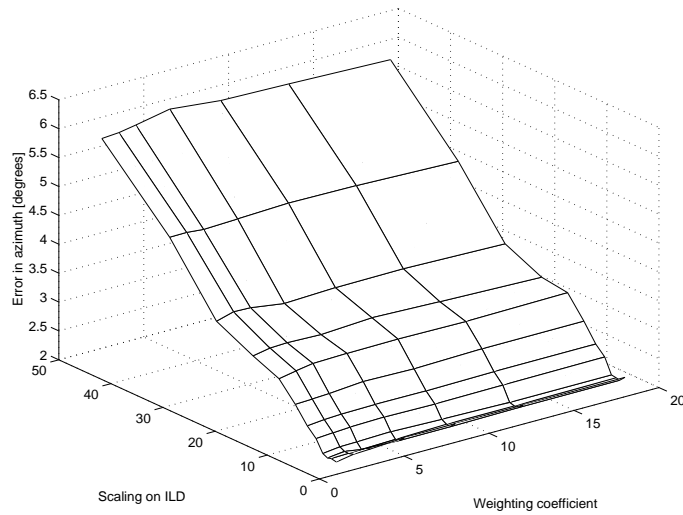


Figure 13: Distribution of error values for Azimuth

chosen equal for all the plots for compatibility. The direction of the axis for the scaling on ILD was inverted in Figs. 17 and 18 so the surface can face the viewer. The distance plot shows a trend of a descending error value but its further extent was not followed because the standard deviation is increasing for those same values, rendering the low error values unreliable. In these trials a number of KNN=7 (nearest neighbors) was used. The values of ILD are theoretically unbound hence it is impossible to get a correct number for the balance of relative weights of ITD and ILD but an empirically estimated value of scaling on ILD of around 13, for which their weight is approximately equal, was found.

From these data plots we can see how the direction (angular) error is low when the relative importance of ITD is high (scaling on ILD is low). The minimum, however, is registered at a non-zero, but nearly zero value of scaling on ILD as shown in Figs. 13 and 15. This means that ILD is useful for reducing angular error, but not much. The contribution of ILD in reducing the error in distance is, however very significant, as shown in Fig. 17. This confirms our error analysis in Sec. 2.3. It becomes clear from those observed trends that when it is necessary to estimate both direction and distance to the sound source, both ITD and ILD should be taken into account. An appropriate amount of contribution from ILD could result in lower error in both direction and distance estimates, especially for distance.

The best precision measured for points located within the sector designated for training but between the grid points used for training, was estimated at around  $\pm 2.2^\circ$  in azimuth,  $\pm 2.8^\circ$  in elevation and  $\pm 19\%$  in distance. The super-resolution is achieved by the KNN interpolation in RPT. The relatively lower accuracy

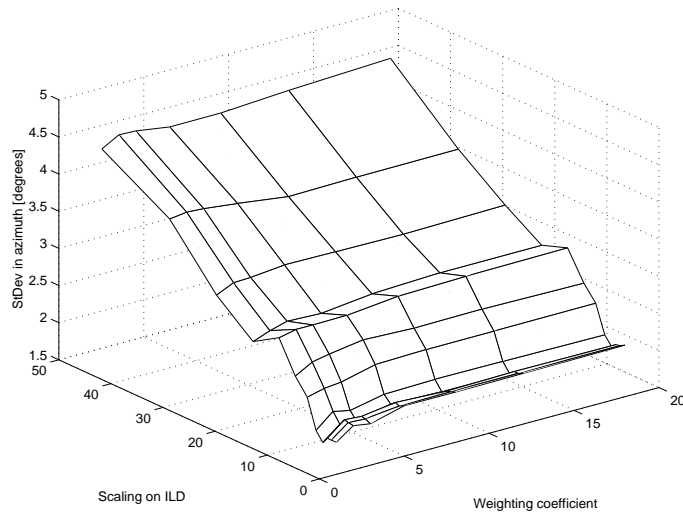


Figure 14: Distribution of standard deviation for Azimuth

in distance is expected for such source-sensor placement. It should also be noted that the specific task the system was tested for — indicating the actual location of the sound — is very difficult for humans in such situations, as well, without the help of other sensing modalities such as vision.

A sample set of error values used to produce a single point in the average errors plots is presented in Table 1.

#### 4.4 RPT performance and speed

To verify the performance of the RPT and its speed, a deterministic linear search algorithm was implemented to find the nearest neighbor in input space from the set of training samples  $L$ . The results obtained in such a linear search method were used to compare the performance of the SHOSLIF procedure. The speed was confirmed to be considerably faster with SHOSLIF. As timed on the test PC, a single retrieval from the tree, with 2370 test samples, took 2.5ms on average for SHOSLIF, versus 15ms with the linear search (see Table 2). The accuracy was comparable to that of the linear search. The timing for the preprocessing indicated an average of 230ms which, although being considerably longer than the retrieval time, is still shorter than the signal scan time of 500ms (single window). This situation indicates that for the current moderate number of training samples collected within the 3-D work space tested, preprocessing takes far more time (230ms) than the retrieval (2.5ms). In other words the RPT algorithm, as implemented, is twice as fast as needed for real time application.

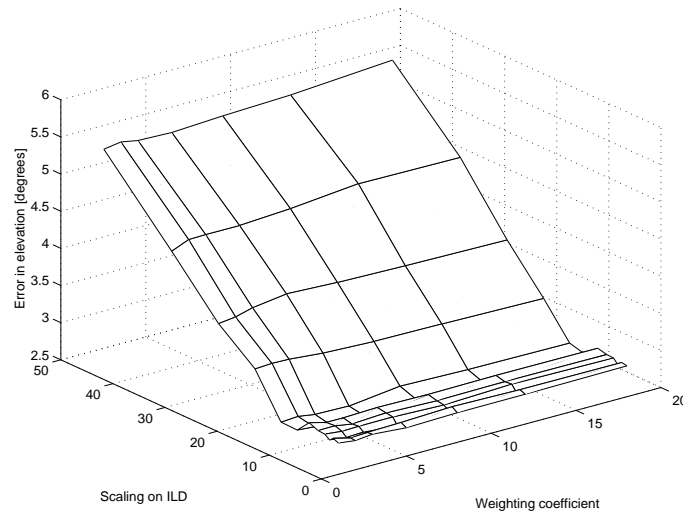


Figure 15: Distribution of error values for Elevation

From the relative time difference between preprocessing and retrieval, one question is in order here. Is it true then that the time gain of SHOSLIF is not important? This is in fact not the case. Consider that the problem becomes larger, in that the number of samples is increased greatly. Such a larger problem arises when, e.g., the work space to be covered is increased, the sample density is increased for better accuracy, or the number of environmental settings is increased for handling more settings. The preprocessing time is somewhat constant, but the retrieval time is not when the number of samples increases. SHOSLIF is able to handle a much larger number of samples without slowing down significantly, due to its logarithmic time complexity. But a linear search method cannot. In other words, SHOSLIF method can “scale up” to larger problems but a linear search cannot.

A graphical user interface has been developed for the program which allows the user to pass all necessary parameters to the program, to select the various options, as well as to view the waveform of the scanned signals.

#### 4.5 Implementation restrictions

As mentioned before, the system performed well despite different unfavorable factors, like background noise, reflections, unreliable sound sources, etc. However, it should be noted that although no exact measurements have been performed, these and some other factors would influence its reliability and accuracy depending on their strength. In most experiments the acoustic noise was kept at a S/N ratio of around 20dB (as estimated

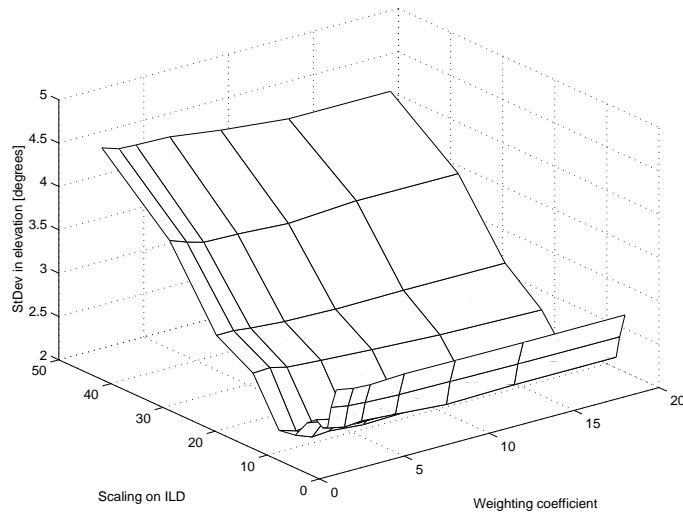


Figure 16: Distribution of standard deviation for Elevation

visually from the displayed waveform) but in real life situations the noise can be as strong as the signal or even stronger. Another problem would be multiple sound sources. In the case of signal reflection, the intensity of the reflected signal would be significantly weaker and thus it will be ignored by the preprocessing routine. However, with secondary sources, the intensity of the sources can be comparable and this might lead to jumping between the two sources and even complete wash-out of the correlation curve and thus incorrect localization.

Most of the experiments were performed with a sound source steadily fixed in space. A moving source would present a challenge for the current implementation. With the current windowing approach, a source movement would be similar to having a source of a larger spatial size (aperture), which would produce a lower signal correlation. The performance with a shortened window has not been studied extensively at this point. In a similar way, an influence on the accuracy of detection was observed when varying the size of the aperture of the sound source. For instance, sounds produced with a wide open mouth would yield a higher error value. An accurate study of this relation needs to be performed in order to determine the correct way of compensating the increase in source size.

One of the typical disadvantages that training presents is the difficulty for a learning system to perform in untrained environments, compared to the environment in which it was originally trained. The sensitivity to environmental changes has not been quantified yet.

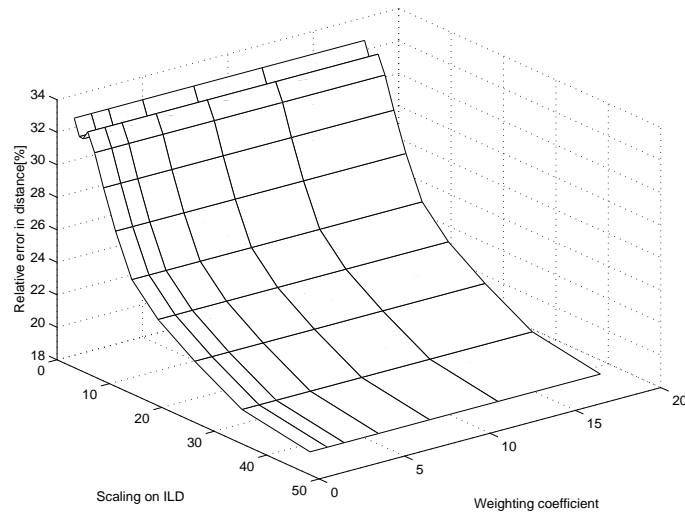


Figure 17: Distribution of error values for Distance. Note that the axis direction for ILD is reverted for better viewing.

## 5 Conclusions

We analyzed the uniqueness and reliability of 3-D sound localization from ITD and ILD. The extraction of ITD and ILD is based on a fast and efficient algorithm that is capable of not only computing those parameters with satisfactory accuracy but also provides a useful means of evaluating the usability of the taken sample. The three dimensional localization is performed by a learning technique. The applicability of the proposed implementation is more general than the majority of the currently available solutions, in that various features can be used without the need to explicitly model the relationship between feature values and the 3D location estimates. The method needs to store a large number of samples (over-learning is avoided by SHOSLIF by only storing samples that are necessary). In order to achieve good accuracy the training density needs to be close to the expected resolution. This can lead to the need of taking samples from hundreds of three dimensional locations, and to ensure stability, several samples from each point need to be taken. However, the logarithmic retrieval enables the system to easily reach real-time response speed.

An originality of this work is in the versatility of its application domain. First the lack of spatial constraints allows for a wide range of applications. The use of a compact sensor array makes it suitable for mobile robots, embedded devices and other human-machine interaction apparati. The simultaneous use of ITD and ILD as related attributes is another advantage because of their complementary character. It is

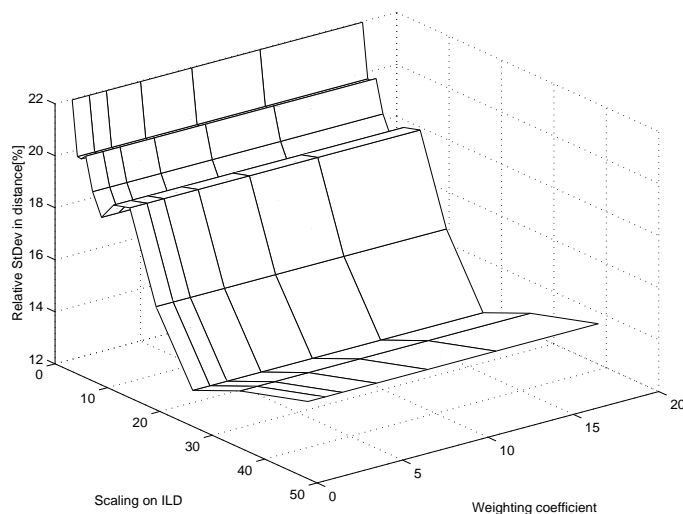


Figure 18: Distribution of standard deviation for Distance

made possible by the learning approach, also unique for this range of problems. We have shown that the use of ILD in addition to ITD does improve the accuracy of direction and distance estimates of the sound source, especially for the distance. The employed non-coplanar array with a minimal number of sensors is another distinctive feature of this work.

## 6 Acknowledgements

The authors would like to thank Shaoyun Chen for making SHOSLIF-N code available for use by this project and having helped the project in many other ways.

## References

- [1] J. Blauert. Sound localization in the median plane. *Acustica*, 22:205–213, 1969.
- [2] M.S. Brandstein. A pitch based approach to time-delay estimation of reverberant speech. In *Proc. 1997 Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, October 19-22, 1997*, 1997.
- [3] M.S. Brandstein, J.E. Adcock, and H.F. Silverman. A practical time-delay estimator for localizing speech sources with a microphone array. *Computer, Speech and Language*, 9:153–169, September 1995.

Table 1: Error values for *Scaling on ILD = 11* and *weighting coefficient = 1* , A=Azimuth [°] , E=Elevation [°] , D=Distance [%]

A	E	D	A	E	D	A	E	D
2.0	6.0	17.45	1.0	1.0	44.25	1.1	0.3	22.16
1.1	6.4	7.70	1.7	5.9	38.35	1.4	1.0	19.54
1.0	1.0	39.46	3.0	2.0	34.02	4.9	0.6	6.95
12.0	0.0	0.00	1.0	4.0	40.24	3.0	3.0	29.82
4.0	2.0	11.82	1.0	4.0	40.24	8.9	3.3	31.68
0.0	10.0	25.37	1.1	8.0	14.55	0.0	2.1	22.08
2.6	3.3	17.22	2.0	2.0	19.14	7.0	2.0	56.67
0.0	0.7	25.35	6.7	0.9	17.04	9.0	1.0	45.56
3.0	4.0	23.95	5.0	2.0	19.79	2.0	3.0	30.86
3.0	4.0	23.95	1.7	3.9	26.79	6.6	4.0	29.94
0.0	1.1	24.45	5.7	1.1	7.72	3.0	2.0	41.13
3.0	0.9	14.07	0.9	4.4	2.83	2.0	3.0	30.86
1.4	2.9	11.51	7.7	2.9	16.54	0.0	0.0	3.89
1.3	0.4	11.64	5.3	3.1	1.98	0.0	0.0	50.34
0.1	2.7	10.45	3.1	1.1	9.45	6.4	5.9	0.61
11.0	3.0	32.09	7.0	0.0	8.99	5.4	1.3	22.39
0.0	5.3	3.38	2.4	5.6	7.70	0.7	0.9	4.52
9.0	1.0	45.56	3.6	3.1	7.05	3.0	1.0	39.36
2.0	9.0	37.70	0.0	5.1	25.36	0.9	4.6	13.76
0.3	5.1	51.18	3.0	0.0	26.72	2.4	2.3	13.80
3.0	2.0	41.13	5.1	8.4	17.51	0.7	4.4	13.60
0.3	1.7	22.14	0.1	1.0	4.81	1.0	4.0	23.59
9.0	4.1	57.37	2.3	2.3	6.60	4.0	0.3	14.68
1.1	2.4	69.98	0.9	3.3	46.99	6.0	3.0	12.55
1.1	4.6	19.53	8.0	3.0	22.56	4.7	2.0	7.40
1.0	1.0	44.25	1.4	0.3	1.99	4.6	1.1	12.31

Table 2: Comparative timings of various routines

Preprocessing	Linear Search	SHOSLIF
230ms	15ms	2.5ms

- [4] 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.
- [5] M.S. Brandstein and H.F. Silverman. A practical methodology for speech source localization with microphone arrays. *Computer, Speech and Language*, 11(2):91–126, April 1997.
- [6] U. Bub, M. Hunke, and A. Weibel. Knowing who to listen to in speech recognition: visually guided beamforming. In *Proceedings of the 1995 ICASSP, Detroit, MI, 1995*.
- [7] V. Capel. *Microphones in action*. Fountain Press, Argus Books Ltd., Hertfordshire, England, 1978.

- [8] H.A. Carr. *An introduction to space perception*. Hafner, New York, 1966.
- [9] B. Champagne, S. Bedard, and A. Stephenne. Performance of time-delay estimation in the presence of room reverberation. *IEEE Transactions on Speech and Audio Processing*, 4(2):148–152, March 1996.
- [10] Y. Chan, R. Hattin, and J. Plant. The least squares estimation of time delay and its use in signal detection. *IEEE Trans. Acoust., Speech, Signal Processing*, 26(3):217–222, 1978.
- [11] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, NY, second edition, 1990.
- [12] W.M. Hartmann. Localization of a source of sound in a room. In *Proc. AES 8th International Conference*, pages 27–32, 1990.
- [13] W.M. Hartmann and B. Rakerd. Localization of Sound in Rooms IV: The Franssen Effect. *J. Acoust. Soc. Am.*, 86(4):1366–1373, October 1989.
- [14] S.L. Hobbs. Asymptotic statistics for location estimates of acoustic signals. *J. Acoust. Soc. Am.*, 91(3):1538–1544, March 1992.
- [15] J. Ianiello. Time delay estimation via cross-correlation in the presence of large estimation errors. *IEEE Trans. Acoust., Speech, Signal Processing*, 30(6):998–1003, 1982.
- [16] C. Knapp and C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust., Speech, Signal Processing*, 24(4):320–327, 1976.
- [17] C.J. MacCabe and D.J. Furlong. Virtual imaging capabilities of surround sound systems. *J. Audio Eng. Soc.*, 42(1/2):38–48, jan/feb 1994.
- [18] K.M. Martin. Estimating azimuth and elevation from interaural differences. In *1995 IEEE Mohonk workshop on Applications of Signal Processing to Acoustics and Audio*, October 1995.
- [19] D.V. Rabinkin et al. A DSP Implementation of Source Location Using Microphone Arrays. In *131st meeting of the Acoustical Society of America, Indianapolis, Indiana, 15 May 1996*, 1996.
- [20] J. Weng and S. Chen. Vision-guided navigation using SHOSLIF. *Neural Networks*, 11:1511–1529, 1998.
- [21] F.L. Wightman and D.J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.*, 91(3):1648–1661, March 1992.

- [22] W.A. Yost and G. Gourevitch. *Directional hearing*. Springer-Verlag, New York, 1987.