Learning-Based Three Dimensional Sound Localization Using a Compact Non-Coplanar Array of Microphones

Kamen Y. Guentchev and **John J. Weng**
Department of Computer Science
3115 Engineering Building
Michigan State University
East Lansing, MI 48824
guentche@cps.msu.edu
weng@cps.msu.edu

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. Having robotics as a major application, the approach involves 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 3D location of the sound source. Results from the experiments showed under $\pm 3°$ in average angular error and less than $\pm 20\%$ in average radial distance error.

Introduction[1]

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 (Capel 1978) (Carr 1966). Unfortunately acoustical 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 (MacCabe 1994). Numerous studies have attempted to determine the mechanisms used by humans to achieve dimensional hearing (Carr 1966) (Hartmann 1990) (Hartmann 1989). 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) (MacCabe 1994) (Wightman 1992) (Yost 1987). Other cues relate to the spectral variations caused by diffractions at the head and pinnae (Blauert 1969). For sounds with longer duration, cognitive processes start playing an important role, including dynamic head adjustments, high-level reasoning, etc. (Yost 1987).

The problem of 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 (Brandstein 1997a) (Brandstein & Silverman 1997) (Brandstein & Silverman 1995) (Rabinkin 1996). In other studies a human model has been followed to some degree resulting in constraints in applicability and limited accuracy (Martin 1995). 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) (Bub & Weibel 1995) (Rabinkin 1996) or the help of a non-acoustical modality has been used (e.g. vision)(Bub & Weibel 1995).

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 3D 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 (Hartmann 1990). Sometimes it can be convincingly wrong (e.g. Franssen effect) (Hartmann 1989). 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 (Wightman 1992). The clear geometrical representation of the problem makes it the favorite feature to be used when approaching such a task by a machine setup (Brandstein 1997b) (Brandstein 1997a) (Brandstein & Silverman 1995) (Bub & Weibel 1995) (Chan & Plant 1978) (Ianiello 1982) (Knapp 1976) (Rabinkin 1996). However, if we return to the physical aspect of the problem, it is clear that even three sensors (noncollinear) are not sufficient by themselves to establish unambiguously the three dimensional location of the sound source: obviously there are two symmetrical solutions on each side of the plane, on which the sensors lay. It is then reasonable to increase the number to four microphones, to localize arbitrarily placed sound sources.

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 (MacCabe 1994) (Martin 1995). Another reason is the lack of an explicit and stable relationship between ILD and source location which will allow for a simple algorithmic solution to be derived (MacCabe 1994). The learning approach used in this study does not have such limitations and it benefits from the added cues.

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 (Hobbs 1992) (Martin 1995) (Rabinkin 1996), or an approximation model for which a closed-form solution can be derived (Brandstein 1997a) (Brandstein & Silverman 1997). 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, this work employs a learning based approach. It consists of a training phase and a performance phase. In the training phase, sounds from known 3D 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 3D 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 noncoplanar sensor array for full 3D 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. 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 (Weng 1996a) (Weng 1996b). It was found that the results produced by SHOSLIF had identical precision with that of the linear search, while its performance time was nearly 5 times faster.

Theoretical problem and sensor structure

Required by versatile applications such as the dimensional hearing of a mobile robot, we cannot use room-oriented solutions (Brandstein & Silverman 1997) (Rabinkin 1996), 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 dimensions 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 is different and for the case of a source size comparable to the source-detector distance, the problem is outside the scope of this work. The same applies to the case of distinct sources of comparable intensity. To determine the minimum number of sensors and their optimal placement, we need to look into the geometrical aspects of the problem.

Parameters

From the front of the spherical acoustic wave, the two main parameters that can be estimated 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 proportional to the difference of the distances between each of the detectors and the sound source:

$$\text{ITD} \sim r_1 - r_2, \tag{1}$$

where r_i is the distance between the sound source and the i-th microphone, $i = 1, 2$, and $\sim$ indicates proportional (Fig. 1). Also, since the amplitude of the sound

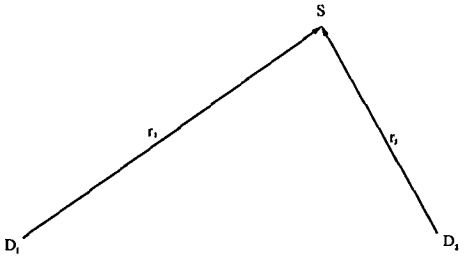

Figure 1: Two detectors, D_1 and D_2, and the sound source S.

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:

$$\text{ILD} \sim \frac{r_2^2}{r_1^2} \tag{2}$$

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

Number and placement of detectors

In order to determine the minimum number of detectors necessary we will first have to consider the geometric representation. For each couple of microphone detectors, we can derive the locus of the source points corresponding in three dimensional space to a given measured value of ITD or ILD. From equation (1) the locus is represented by a hyperboloid of two sheets of revolution with foci D_1 and D_2. Depending on the sign of the difference, one of the sheets contains the source location. Then from equation (2), for the matching ILD, it is less obvious but it can be shown the locus is a sphere (Guentchev 1997) (Albert 1966) (Bell 1918) (Sommerville 1929). The intersection of these surfaces, defined by a number of detector couples, will determine the solution. It is clear that, apart from some special cases, with three couples the intersection is two points located symmetrically on both sides of the plane passing through the three detectors. If four non-coplanar detectors are used the intersection is a single point. Since there are no restrictions on the source-detector placement, an equidistant structure seems reasonable. In the case of four sensors this suggests a tetrahedron (Fig. 2). A mobile robot requires that the structure of the sensor array be compact, while accuracy consideration requires a large array. In the experiment, an equal-side tetrahedron with a 20cm side was used.

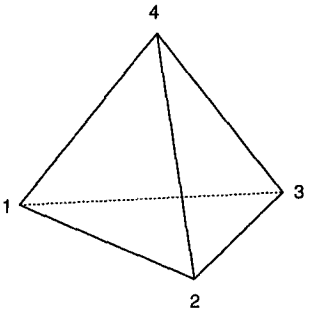

Figure 2: Placement of the 4 microphones on the array.

Methodology

As outlined above, efficient sound localization can be performed after having extracted the necessary differential features. It can be shown (Guentchev 1997) that the minimum number of detectors required to obtain unambiguously a solution in three dimensional space is four and that it is unique. In order to fully solve the problem of three dimensional sound localization two main steps need to be performed. First the chosen features need to be extracted from the acquired signal. Then the actual sound location is estimated using those features.

Feature Extraction

As discussed, the two features considered in this work are ITD and ILD. Their extraction is based on the fact that the signal detected by the different sensors bears

a significant degree of coherency when produced by a single dominating source.

Using the appropriate hardware, the acoustic signal can be first converted to electrical signal (microphones) and then to digital form (analog to digital converter board). The digitized data is a sequence of values representing the intensity 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 sample. For instance, the amplitude of the signal's envelope can vary over time, e.g. with speech this corresponds to 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. 3). 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.

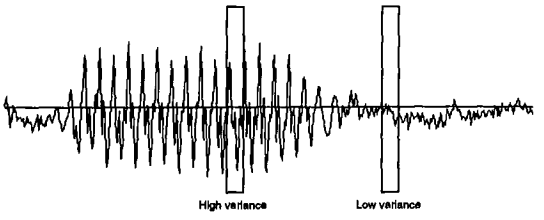

Figure 3: Preselection according to signal variance

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 (Bub & Weibel 1995). We find the peak of the cross-correlation function varying across a range of possible time-delays (3). When $j = 0$, k varies from 0 to $n-1$ and when $k = 0$, j varies from 0 to $n-1$:

$$R_{max} = max_{\substack{j=0,\ 0\le k<n \\ k=0,\ 0\le j<n}} (R_{i,j}), \qquad (3)$$

where

$$R_{i,j} = \frac{\sum_{i=0}^{N-1}(X_{i+j} - \overline{X_j})(Y_{i+k} - \overline{Y_k})}{\sqrt{\sum_{i=0}^{N-1}(X_{i+j} - \overline{X_j})^2}\sqrt{\sum_{i=0}^{N-1}(Y_{i+k} - \overline{Y_k})^2}}$$

and where R_{max} is the maximum cross-correlation, X_i and $Y_i, i = 0, \ldots, N$ are the samples from both channels, $N + n$ being the total number of samples in the window and n the number of samples corresponding to half of the maximum possible time delay.

The time-shift T will be proportional to the value of j or k that maximizes the value of (3), 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. 4). 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 coherence of the signals from both channels, which is true for a clear and compact sound source but also is true for an even isotropic noise. The second parameter, however, discriminates accurately between those two cases and helps select only the appropriate matches: a wide, flat correlation curve would indicate a predominance of noise, and a narrow, sharp curve - a clear, distinguishable sound source.

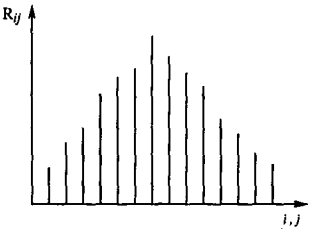

Figure 4: Typical cross-correlation curve with good sharpness

We should note a very useful side effect, which is closely related to the precedence effect in human auditory perception (Hartmann 1990) (Hartmann 1989). In case there are reflections of the incoming sound from incident surfaces (echoes, reverberation (Champagne & Stephenne 1996)) a secondary maximum will appear in the correlation curve. However, with the presented approach, it will be ignored because the coherency of that signal will be significantly lower and thus it will correlate less well (lower and wider peak).

Using the thus 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. The value for microphone pair 1 and 2 is shown in equation 4.

$$\mathrm{ILD}_{12} = \int_{t_i}^{t_j} \frac{S_1(t)}{S_2(t+T)}\, dt \qquad (4)$$

where $S_1(t)$ is the signal picked from microphone 1 and $S_2(t+T)$ is the signal picked from microphone 2, T is the previously determined time shift and $t_i - t_j$ is the length of the sample window.

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 will also suggest when a sample can be reliably used for localization and when it should be discarded as useless.

Source localization

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 these features uniquely define a solution and thus we have a direct correspondence between the three dimensional coordinates of the sound source and the extracted ITD and ILD values. We should also note the extreme complexity of the actual mapping function. It is appropriate then to use learning as an efficient way of approximating complex high-dimensional functions (Weng 1996a). Being able to explicitly model the function would significanlty increase the accuracy of prediction. However, in the current case, when considering all the presented variables, it is very difficult to establish the form of such a model. Furthermore, the model might strongly depend on the training environment and hence the generality of the performance application domain will be seriously limited.

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: azimuth, elevation and radial distance. Thus the mapping is of the form $[x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{26}] \rightarrow [y_a, y_e, y_r]$. The polar representation in output space is used because it is known that the radial distance is a very unreliable estimate.

The goal of the SHOSLIF recursive partition tree (RPT) is to find the top k nearest neighbors in $O(log(n))$ time, where n is the number of training samples learned and stored. The SHOSLIF scheme (Weng 1996a) (Weng 1996b) automatically generates the RPT, which recursively partitions the input space X into smaller and smaller cells. Each leaf node contains exactly one training sample (x_i, y_i), where $x_i \epsilon X$ and $y_i \epsilon Y$. The partition is based on automatically computed principal component subspace or linear discriminant subspace at each internal node of the RPT. A distance-based weighting according to (5) is applied to the y_i vectors of the top-k matched leaf nodes to give an interpolated output vector $y \epsilon Y$, given input vector $x \epsilon X$.

$$y = \frac{1}{\sum_{i=1}^{k} w_i} \sum_{i=1}^{k} w_i y_i \tag{5}$$

where w is the weighting function and y_0 is the nearest neighbor:

$$w_i = \alpha^{-\frac{\|y - y_i\|}{\|y - y_0\| + \epsilon}} \tag{6}$$

where $\alpha \geq 1$.

Experimental setup and results

In order to test the methodology, an experimental setup was used to perform a number of tests. A set of four identical Lavalier microphones is placed at the tips of a solid tetrahedron with 20cm side (Fig. 2). The signal from the microphones is amplified by four modular microphone preamplifiers to bring the signal level in range. It is then supplied to an analog-to-digital converter board mounted in a personal computer. The software is designed to visualize, train, recognize and run various sound localization related tasks. Samples were taken from various points with known 3D coordinates, some were used for training and others for testing. The results were analyzed with linear search and the performance of SHOSLIF was evaluated.

Experiment and results

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 of the Department of Computer Science, which is hardly suitable for high precision acoustic experiments. The test space is located in the middle of the laboratory, in between cubicles with computers and reflecting, and absorbing surfaces of irregular shape. The number of devices producing strong noise of different frequencies and levels is above 20. Often laboratory members would speak softly in the background while samples are being taken for training or retrieval. This highly problematic environment was close to the real world environments, in which a typical sound localization device would be intended to work.

Experiment At the training stage a continuous sound, originally produced by a human speaker uttering a short sentence, is reproduced using a hand held tape recorder, from a set of previously defined locations (Fig. 11). Without 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 thus defined 700 points were

used for training. They were selected to simulate an uniform angular density and ten samples were taken from each of those points. The approximate angular density of the training points was around 15°. Thus the angular span of the training area was about 180° in azimuth and a little less than that in elevation. The site was the PRIP Laboratory where the acoustics is very challenging and the background noise was significant. At this time a room with better acoustic properties (e.g. anechoic chamber) was not available but as it was mentioned earlier, this type of environment is close to the one in which an actual sound localization device would be exposed to.

At the recognition stage the same device produced the original sound, however the performance of the system was tested with other sources, as well. A human speaker would produce variable results. One important observation was that the quality of the sound, which directly influences the reliability and the accuracy of recognition, depended on the amount of useful information (ratio of sounds vs pauses) but mostly on the compactness of the source - the aperture of the mouth.

The system can be started to collect 0.5s samples and provide location estimates based on each of them or to store them for further analysis. Test samples from 79 points on the grid, but different from the ones used for training, were taken. They were analyzed offline with the specially designed software. 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 three dimensional 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 employed algorithm uses two parameters for fine tuning. One is the relative weight of the two extracted features: ITD and ILD. Because of the importance of the ITD, only the ILD was multiplied by a variable factor, called *scaling on ILD*, thus increasing its weight (the original values being very low numbers) as needed. This allows us to estimate the relative influence of those two parameters on the accuracy of the results. A low value of this parameter would mean neglecting the ILD (a value of zero means only ITD are used), while a higher value indicates a predominance of the ILD. Their relative weight is practically equal for a value of *scaling on ILD* of around 13. The other parameter is the weight coefficient α in the interpolation of the retrieved nearest neighbors (6). A low value of α would indicate that all considered nearest neighbors are almost equally weighted (for $\alpha = 1$ we have averaging) while a big value of alpha emphasizes the role of the nearest neighbor.

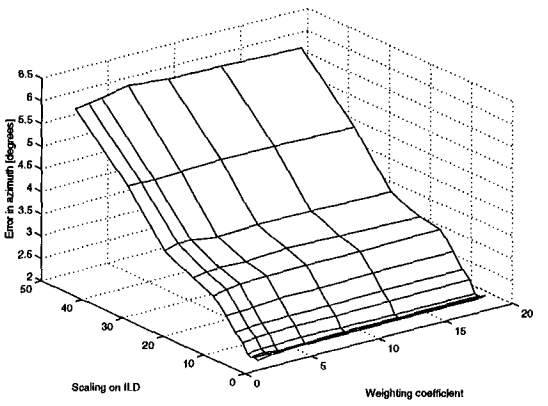

Figure 5: Distribution of error values for Azimuth

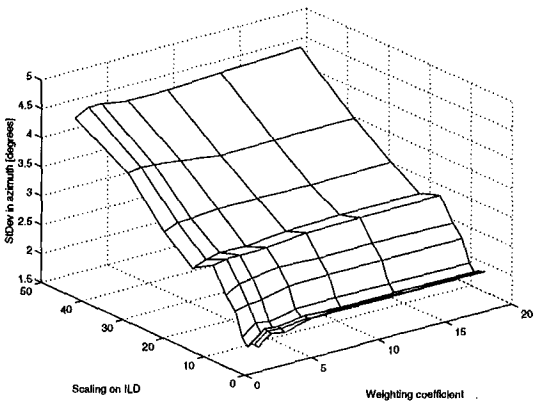

Figure 6: Distribution of standard deviation for Azimuth

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 abandoned at this stage.

Results The results obtained in this manner were used to study the above mentioned relations. A number of plots is used to show some observed trends. Fig. 5 shows how the relative weighing between ITD and ILD affects the accuracy of estimation of the azimuth, Fig. 7 - of the elevation and Fig. 9 - of the dis-

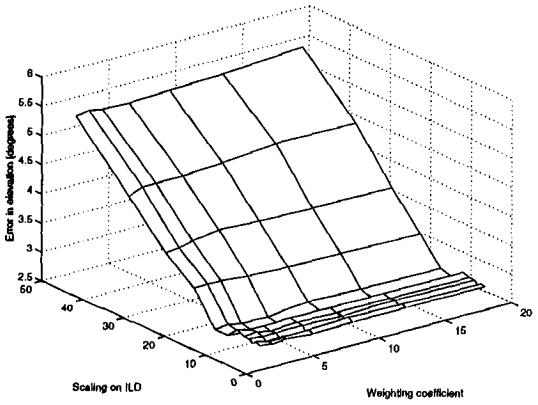

Figure 7: Distribution of error values for Elevation

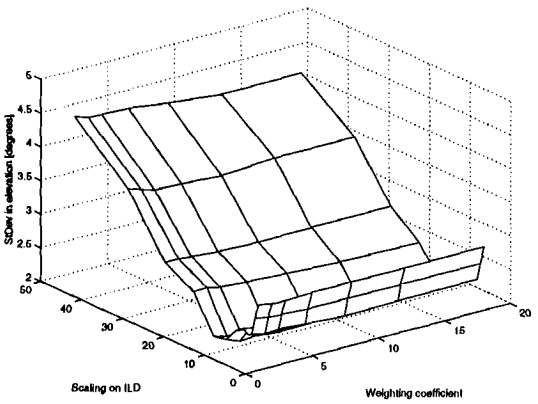

Figure 8: Distribution of standard deviation for Elevation

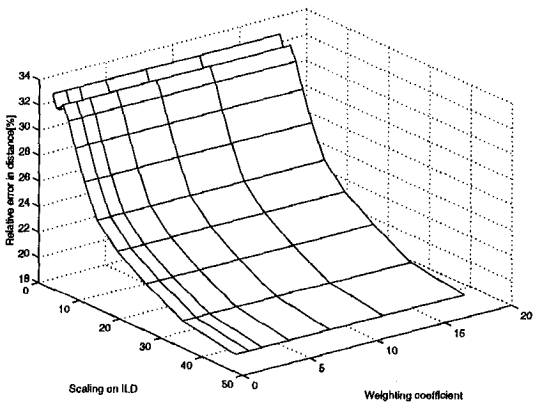

Figure 9: Distribution of error values for Distance

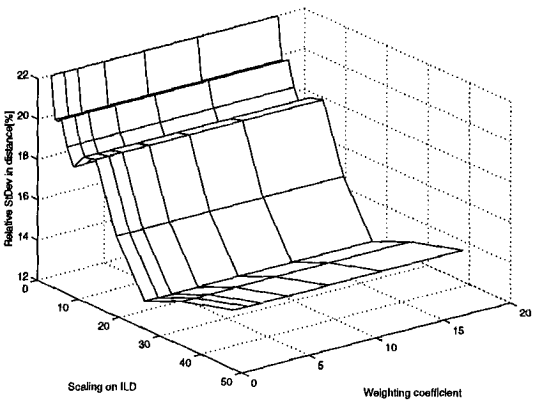

Figure 10: Distribution of standard deviation for Distance

tance. The respective standard deviations are shown on Fig. 6, Fig. 8 and Fig. 10. The horizontal axes are the scaling on ILD - the coefficient by which ILD is multiplied when considered for estimating the nearest neighbors, and the weighting coefficient α, which indicates how much influence the further neighbors have in the interpolation (see equation 5). The range on the axes was chosen equal on all plots for compatibility and the direction of the axis, representing the scaling on ILD, was inverted for the distance plot 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.

We can notice how the angular error is low when the relative importance of ITD is high (scaling on ILD is low). The minimum, however, is registered for a non-zero value of scaling on ILD. We can make the contrary observation for the error in distance. 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. A compromise value of scaling on ILD would ensure aceptable error both on angular and radial estimates.

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 $\pm 2.2°$ in azimuth, $\pm 2.8°$ in elevation and $\pm 19\%$ in distance. The superresolution is achieved by the KNN interpolation while the lower performance 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.

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

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

Program performance and details

For accuracy testing and parameter fine tuning a determinsitic linear search algorithm was implemented to find the nearest neighbors in input space. The results obtained in such a way were used to estimate 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). In other words the algorithm, as implemented, is twice as fast as needed for real time application.

Table 2: Comparative timings of various routines

PreProcessing	Linear Search	SHOSLIF
230ms	15ms	2.5ms

The program was written in C++ and is object oriented with the exception of the C code, adapted from a previous implementation of SHOSLIF. The Graphical User Interface 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.

Figure 11: Training the system

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 visually from the displayed waveform) but in real life situations the S/N ratio can be as high as 0dB (noise is as strong as signal). 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 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 main disadvantages that training presents is the difficulty for a learning system to perform in unknown environments, compared to the environment in which it was originally trained. No exact measures have been taken to establish the actual error values in a new environment but many observations indicate that the system can perform within some reasonable expectations.

Another obvious limitation is the absolute distance from the detectors to the sound source. Because of the physical characteristics of sound propagation in the air, the coherency of the sound waves decreases significantly with distance: non-linear frequency absorption, distortion, fading, are just a few of the factors. A good example is the noise, coming from the engines of airplanes high in the sky - it is very difficult to establish the location of the source, even after turning around several times (another factor intervenes here, too: the relatively slow speed of sound introduces a serious delay, relative to the optical image, i.e. the actual location of the object). For the presented implementation distances of just 5 to 10 meters would already pose a serious problem.

Conclusions

Discussion

A learning-based method for determining the three-dimensional location of a sound source is presented. Two of the most important features used by humans in sound localization are used. Their extraction is based on a fast and efficient algorithm that is capable of not only computing those parameters with satisfactory accuracy but also provides a very 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. Thus the total number of training samples can some times approach tens of thousands. However, the logarithmic retrieval makes the system easily reach real-time response speed.

The 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 made possible by the learning approach, also unique for this range of problems. The employed non-coplanar array with a minimal number of sensors is another distinctive feature of this work.

Future research

Some results suggest several directions for future work. The observed dependence of the "quality" of the sound on the size of its source (aperture) is an important issue because of the typical application domain of this method - human voice. Because of the different size of the mouth of different people and because of the inherent variations in the ways of pronouncing phonemes, the detectability of human sources can suffer severely. The chosen approach in preprocessing of captured sounds could be revised to reflect the expected variations in the source size.

Another direction for further study is the impact of sound source movement over the selected methodology. The relation between the speed of source movement and the discretization interval (the time segment, used to extract the features) can be adjusted so that the source movement remains relatively slow but this approach will inevitably result in loss of accuracy. Other techniques for compensating the source movement can be developed to efficiently handle this problem.

The issue of multiple sound sources also should be studied in more details. With the present approach the presence of a second sound source is ignored but it degrades the quality of localization for the primary source.

It is also necessary to establish the degradation of performance when the system is put in an unknown environment. When training the system no assumptions are made about the acoustics of the environment and the ability of the system to overcome obstacles like noise, reflections, etc., could be due to the environment specific training. It is interesting to study the relation between the characteristics of the environment and the quality of sound localization.

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

Albert, A. 1966. *Solid Analytic Geometry.* Phoenix books: University of Chicago Press.

Bell, R. 1918. *Coordinate Geometry of three dimensions.* London: Macmillan.

Blauert, J. 1969. Sound localization in the median plane. *Acustica* 22:205–213.

Brandstein, M.S.; Adcock, J., and Silverman, H. 1995. A practical time-delay estimator for localizing speech sources with a microphone array. *Computer, Speech and Language* 9:153–169.

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

Brandstein, M.S.; Silverman, H. 1997a. A practical methodology for speech source localization with microphone arrays. *Computer, Speech and Language* 11(2):91–126.

Brandstein, M. 1997b. 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.*

Bub, U.; Hunke, M., and Weibel, A. 1995. Knowing who to listen to in speech recognition: visually guided beamforming. In *Proceedings of the 1995 ICASSP, Detroit, MI.*

Capel, V. 1978. *Microphones in action.* Hertfordshire, England: Fountain Press, Argus Books Ltd.

Carr, H. 1966. *An introduction to space perception.* New York: Hafner.

Champagne, B.; Bedard, S., and Stephenne, A. 1996. Performance of time-delay estimation in the presence of room reverberation. *IEEE Transactions on Speech and Audio Processing* 4(2):148–152.

Chan, Y.; Hattin, R., and Plant, J. 1978. The least squares estimation of time delay and its use in signal detection. *IEEE Trans. Acoust., Speech, Signal Processing* 26(3):217–222.

Guentchev, K. 1997. Learning-based three dimensional sound localization using a compact noncoplanar array of microphones. Master's thesis, Dept.of Computer Science, Michigan State University.

Hartmann, W.M.; Rakerd, B. 1989. Localization of Sound in Rooms IV: The Franssen Effect. *J. Acoust. Soc. Am.* 86(4):1366–1373.

Hartmann, W. 1990. Localization of a source of sound in a room. In *Proc. AES 8th International Conference*, 27–32.

Hobbs, S. 1992. Asymptotic statistics for location estimates of acoustic signals. *J. Acoust. Soc. Am.* 91(3):1538–1544.

Ianiello, J. 1982. Time delay estimation via cross-correlation in the presence of large estimation errors. *IEEE Trans. Acoust., Speech, Signal Processing* 30(6):998–1003.

Knapp, C.; Carter, C. 1976. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust., Speech, Signal Processing* 24(4):320–327.

MacCabe, C.J.; Furlong, D. 1994. Virtual imaging capabilities of surround sound systems. *J. Audio Eng. Soc.* 42(1/2):38–48.

Martin, K. 1995. Estimating azimuth and elevation from interaural differences. In *1995 IEEE Mohonk workshop on Applications of Signal Processing to Acoustics and Audio.*

Rabinkin, D. e. a. 1996. A DSP Implementation of Source Location Using Microphone Arrays. In *131st meeting of the Acoustical Society of America, Indianapolis, Indiana, 15 May 1996.*

Sommerville, D. 1929. *Analytical Conics.* London: Bell.

Weng, J. 1996a. Cresceptron and SHOSLIF: Toward comprehensive visual learning. In Nayar, S.K.; Poggio, T., ed., *Early Visual Learning.* New York: Oxford University Press. 183–214.

Weng, J.J.; Chen, S. 1996b. Incremental learning for vision-based navigation. In *Proc. International Conference on Pattern Recognition, Vienna, Austria, Aug. 1996*, volume 4, 45–49.

Wightman, F.L.; Kistler, D. 1992. The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.* 91(3):1648–1661.

Yost, W.A.; Gourevitch, G. 1987. *Directional hearing.* New York: Springer-Verlag.