# Sound Localization-Based Navigational User Interfaces

Arezou Keshavarz
University of Toronto
Artificial Perception Laboratory
10 Kings College Road
Toronto, Ontario, Canada, M5S 3G4
arezou.keshavarz@utoronto.ca

Parham Aarabi
University of Toronto
Artificial Perception Laboratory
10 Kings College Road
Toronto, Ontario, Canada, M5S 3G4
parham@ecf.utoronto.ca

## Abstract

*In this paper, we propose and compare three navigational user interfaces that are based on acoustic information. The user can navigate through a web page or change slides in a presentation by changing its spatial location. To this end, we employ an array of 24 microphones to record speech signals and localize the speaker. The proposed interfaces are: 1) The Parallax System, which behaves as if the user is looking out of a window, 2) The Corner System, which resizes each image based on the proximity of the speaker to the corners of the environment, and 3) The Tiled System, which divides the environment into a number of tiles and loads the image corresponding to the tile on which the user is standing on. A set of experiments were performed on 15 participants, which showed that the interaction time required for the Tiled System and the Corner System is far less than that required for the Parallax System. The participants were also asked to rank the ease of use of each of the interfaces; it was observed that the users are more comfortable with an interface that requires minimal interaction time.*

## 1. Introduction

The localization of a sound source in space, acheived by an array of microphones, has been extensively studied in the past [4, 2, 1]. This research has resulted in systems that can localize multiple sound sources in real-time and with high accuracy (i.e. errors lower than 10cm) [5, 2, 6]. While most of the prior work in this area has dealt with small array for source tracking and speech enhancement applications, recent work has focused on large-aperture microphone arrays for tracking sound sources in large environments.

In this paper we utilize the University of Toronto's Artificial Perception Lab microphone array to perform real-time sound localization as the basis of a user interface for web-page and presentation navigation applications. In other words, we use real-time information about the location of an acoustic sound source to allow a user to navigate a website, change slides in a presentation, or perform rudimentary navigation in software applications.

Systems that utilize the spatial location of a user as an input control for user interfaces have been developped before. For example, the work of [11] describes the principles for designing an interactive ambient display that can be used collaboratively in public areas to convey personal and public information to multiple users simultaneously. The system estimates the level of interest of the passerby in retrieving information from the display by analyzing the persons orientation, attention and spatial position, and displays relevant information on the screen accordingly. As the authors suggest, this system is designed for short-duration usage, and is most useful for displaying information such as messages, meetings and weather conditions.

In [12] an Acoustic Cue Editor (ACE) was developed that used non-speech sound cues as feedback to control a robot arm in a tele-robotic scenario. For instance, sound was used to indicate that one of the pre-defined gestures, such as single-finger point or fist was recognized. These cues would assist the operator in controlling the tele-robot more efficiently.

The work by [9] developed an audio browsing environment called Dynamic Soundscape. This system maps the temporal navigation of audio data into a spatial interface. Using this user interface, users can browse audio information by remembering the spatial location of the moving sound sources that play multiple segments of a single audio file rather than fast-forwarding and rewinding.

In [7] a haptic interface that supports variability in the movement styles of the users is used. As the author sug-

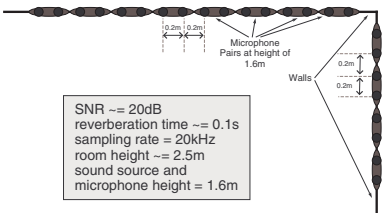


**Figure 1. Environmental setup of the microphone array**

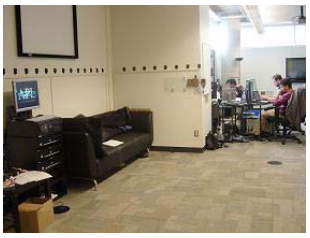

**Figure 2. A photo of the physical settings of the microphone array**

gests, this interface is used to perceive information from the human hand movement patterns. The emphasis of this user interface is not on the precision of perceptions, but on making numerous, coarse distinctions using the haptic sense simultaneously. In other words, they are interested in retrieving as much data from an object simultaneously using the haptic sense as possible.

The work by [10] describes a proactive user interface for a shopping assistant, which offers assistance and tailored suggestions based on the products that the user has selected. This interface makes use of Radio Frequency Identification (RFID) sensors to monitor the shopper's actions and provides suggestions by trying to infer the goals of the user. The plan recognizer communicates with the user through a Tablet PC screen mounted on the shopping cart.

Here, at the Artificial Perception Laboratory, we have developed user interfaces for browsing web-pages or navigating slides in presentations. The underlying basis of this interface is the Sound Localization system, which uses acoustic data to estimate the location of the speaker. The user can navigate the web-page or change slides in the presentation by changing its spatial location while talking. For instance, the user can move to the next slide by moving to the right, or moving to the previous slide by moving to the left.

## 2 REAL-TIME SPEAKER LOCALIZATION

Sound source localization has been extensively studied before [4, 2]. In this paper, the same approach as [3] is utilized for real-time robust sound localization.

The speech signal generated by the speaker is recorded by an array of 24 microphones. Figure 1 illustrates the setup of the environment, and figure 2 shows the physical settings of the microphone array in the room.

The localization system utilizes a modified version of the SRP-PHAT algorithm [5] which uses Time Delay of Arrival (TDOA) histograms [2]. With the modified algorithm, TDOA histograms are computed for different pairs of microphones by taking repeated Phase Transforms (PHAT) [8, 5, 2] for several consecutive 20ms time-segments, with the maximizing time-delay for each PHAT being incorporated into the histogram. The maximizing PHAT time-delay $\tau$ between microphones $n$ and $m$ for the signal of time segment $k$ is defined as:

$$\tau_{n,m,k} = \arg\max_{\beta} \int_{-\infty}^{\infty} \frac{X_{m,k}(\omega)\overline{X_{n,k}(\omega)}}{\left|X_{m,k}(\omega)\overline{X_{n,k}(\omega)}\right|} e^{-j\omega\beta} d\omega \quad (1)$$

where $X_{m,k}(\omega)$ and $X_{n,k}(\omega)$ are the Fourier Transforms of the $k$th 20ms signal segment recorded from the $m$th and $n$th microphones, respectively. In practice, each 20ms signal is obtained by sampling the continuous-time microphone signal and then windowing the samples by half-overlapped Hanning windows (and assigning an integer index number $k$ to each segment with $k = 0, 1, 2, ...$). The frequency representation of the finite-duration and discrete-time signal is obtained by performing a Fast Fourier Transform (FFT), resulting in discrete frequency components. As a result, the integral of equation 1 is in practice a summation over the discrete FFT frequencies.

Assuming that for a given localization a total of $K$ time-segments are available, then for microphones $m$ and $n$ the TDOA histogram can be defined as follows:

$$h_{m,n}(\tau) = \mathrm{hist}([\tau_{m,n,0}\, \tau_{m,n,1}\, \tau_{m,n,2}\, \ldots\, \tau_{m,n,K}], \tau) \quad (2)$$

where the $\mathrm{hist}(\mathbf{t}, \tau)$ function is a histogram operator (i.e. counting the number of TDOA estimates that fall within a finite set of preset bins) for the TDOA vector $\mathbf{t}$ and bin center $\tau$.

Now, a given location $\mathbf{x}$ has a set of TDOAs corresponding to each microphone pair. We can precalculate the TDOA $\Omega_{m,n}(\mathbf{x})$ between microphones $m$ and $n$ corresponding to position $\mathbf{x}$ using $\Omega_{m,n}(\mathbf{x}) = (\|\mathbf{x_m} - \mathbf{x}\| - \|\mathbf{x_n} - \mathbf{x}\|)/\nu$, where $\mathbf{x_m}$ and $\mathbf{x_n}$ are the spatial locations of the $m$th and $n$th microphones, respectively, and $\nu$ is the speed of sound in air (approximately 345m/s).

In order to compute the likelihood of a speaker at position $\mathbf{x}$, we sum up the histogram values at the TDOAs corresponding to $\mathbf{x}$ using $\psi(\mathbf{x}) = \sum_m \sum_n h_{m,n}(\Omega_{m,n}(\mathbf{x}))$,

**Figure 3. An example of the Sound Localization system**

**Figure 4. Screen-shot of the Visual Sound Localization Systems in use. The figure on the left shows the SLF map projected on a staic background image. The figure on the right shows the SLF map projected on the background image while the foreground image is streamed from a live web cam.**

where $\psi(\mathbf{x})$ is a spatial likelihood function (SLF) representing the likelihood of a speaker at each point in space. By normalizing the SLF according to $f(\mathbf{x}) = \psi(\mathbf{x})/\sum_{\mathbf{u}} \psi(\mathbf{u})$ we obtain a pseudo probability distribution representative of the probability of the speech source being at location $\mathbf{x}$ given the entire data collected from all microphones for each localization.

For the described sound localization process, twenty 100 millisecond samples were used to generate the histogram. Also, only microphone pairs that were 60cm apart or less were used to form pairs. It was experimentally determined that for greater inter-microphone distances, the localization accuracy improvements would not be significant. Furthermore, it was experimentally determined that the accuracy of the proposed sound localization algorithm was slightly better than the SRP-PHAT technique, and as a result the former algorithm was chosen for implementation. Figure 3 shows a demo of the sound localization system.

Moreover, the system can be enhanced by taking the sound localization system to a multi-sensor level. For instance, by combining acoustic and visual information, a more realistic view of the room can be displayed on the screen. As figure 4 shows, the SLF is projected on the floor in the background image of the room using a Projection Matrix (P-Matrix) that was developed in the calibration process. This was done by measuring the 3D real-world coordinates and the 2D image coordinates of six distinct points. In particular, these six points could not all lie on the same plane in the 3D space. The 3D space coordinates and the 2D image coordinates provided enough information to solve for the P-Matrix. As a result, the location of the speaker relative to other objects in the room can be viewed on the screen [Figure 4] .

This can be taken one step further by streaming live images using a web cam, as opposed to using a static background image. Then, the difference of the background image and the live image is placed on top of the SLF map to re-create the physical situation of the room on the screen [Figure 4].

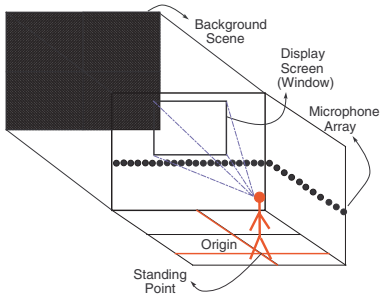


**Figure 5. The behavior of the Parallax System is analogical to looking out of a window**

## 3 DESCRIPTION OF USER INTERFACES

### 3.1 The Parallax System

The *Parallax System* simulates a parallax view of the loaded image by assuming that the image is located 2 meters behind the display screen. In other words, the *Parallax System* behaves like a window from which the speaker can see the image as a background scene [Figure 5]. Using the same analogy, as the speaker moves to the front, a wider view of the outside scene can be seen through the window, whereas moving back will narrow down the range of view of the speaker [Figure 6].

### 3.2 The Corner System

In the case where there are multiple images loaded on the screen, the user can select one of them using the *Corner System*. This is done by measuring the distance of the speaker from the corners of the room. The size of the displayed image is related to the distance of the speaker to the closest corner by a decaying exponential function. The function

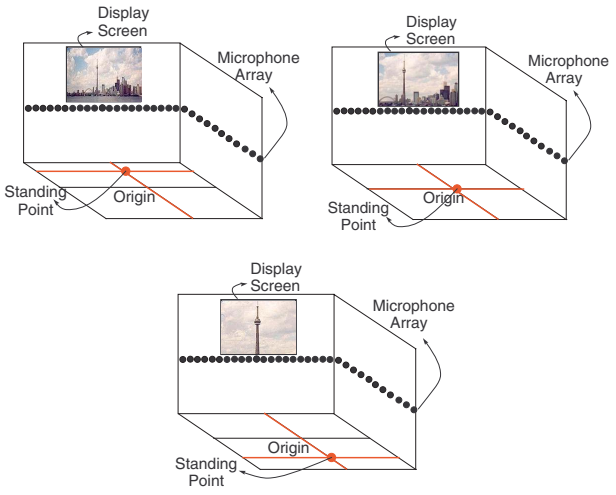


**Figure 6. An example of the Parallax System: Moving back will show a broader view of the image (top left); Moving to the front slightly narrows the range of view of the user (top right). Moving more to the front towards the screen will show a very constrained view of the image (bottom).**

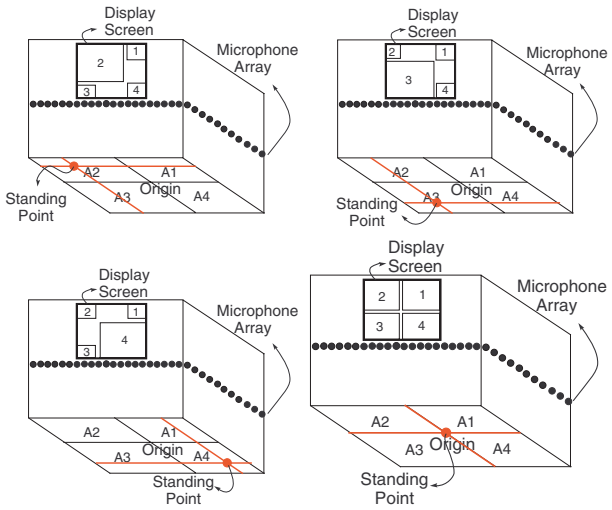


**Figure 7. An example of the Corner System: Approaching the front left corner zooms into the top left image (top left); Approaching the back left corner zooms into the bottom left image (top right); Approaching the back right corner zooms into the bottom right image (bottom right); Standing roughly in the middle of the room causes the four images to have equal size(bottom left).**

behaves such that the image sizes change drastically in the border region, whereas the change in the image sizes decays as the speaker moves away from the border region to the corners.

As an example, assume the case where there are four images loaded on the display. As figure 7 demonstrates, the top left image can be selected when the user is standing close to the front left corner of the room. This interface can be very useful for loading consecutive images in a presentation. Using the *Corner System*, the speaker can move from image to image by moving to different corners of the scene while presenting.

### 3.3 The Tiled System

Using the *Tiled System*, the image is divided into rectangular tiles which represent neighboring sections of the floor. As the speaker walks around the room, the system maps the speaker into the corresponding tile. As a result, the speaker can effectively walk on the image that is projected on the display screen. Figure 8 illustrates the behavior of the *Tiled System* in action. Assume a 2x2 square is originally loaded on the display screen. Depending on the tile on which the speaker is standing on, the corresponding section of the image is displayed on the screen. As an example, if the speaker stands on tile 2, the middle top section of the image is shown on the display screen.

## 4 USABILITY EVALUATION

Fifteen participants were selected to evaluate the usbaility of the three interface systems. Initially, a brief description of the functionality of each user interface was given to all participants. A 2x2 square image, containing the numbers 1-4, was loaded on the display screen.

The participants were asked to select 3 separate number sequences. Sequence one consisted of selecting 1, 2, 3 and 4 consecutively. The other two sequences contained ten randomly selected integers in the range of 1 to 4 (2311243124, 1423214132). The participants were asked to perform the same experiment on all three user interfaces, *The Parallax System*, *The Corner System*, and *The Tiled System* [Figure 9].

The duration of completing each sequence of numbers was measured for each participant. The collected data were used to calculate the average and the standard deviation in selection time corresponding to selecting each sequence using each of the interfaces [Table 1].

Finally, each participant was asked to rank their level of comfort in using each of the interfaces on a scale of 1 to 5, with 1 corresponding to the most difficult to use interface, and 5 corresponding to the easiest to use interface. Again, the average and the standard deviation of the user-assigned

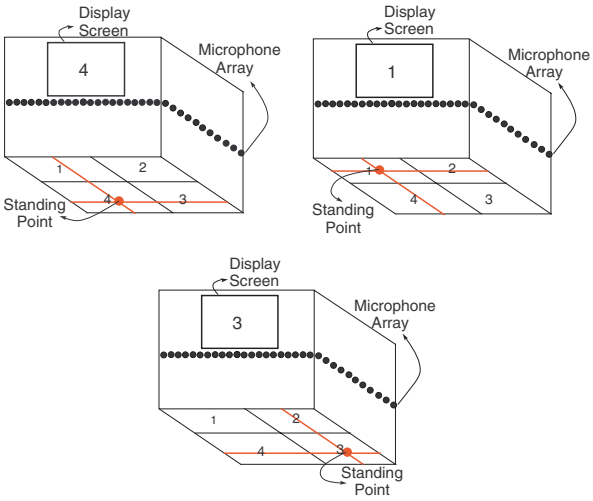


**Figure 8. An example of the Tiled System: Standing on the back left tile displays the bottom left part of the image, number 4 (top left); Standing on the front left tile displays the top left part of the image, number 1 (top right); Standing on the back right tile displays the bottom right part of the image, number 3 (bottom).**

scores to each interface was calculated [Table 2]. A histogram of the user-assigned scores for each of the three systems is shown in figure 11.

## 5 DISCUSSION OF RESULTS

As can be seen in Table 1, the selection time for the same set of numbers using the *Tiled System* is by far less than that of the *Parallax System*. Moreover, although the interaction time with the *Corner System* is less than that of the *Parallax System*, it is still slightly higher than the *Tiled System*. Figure 10 shows the selection time per digit in each of the number sequences, with each set of bars corresponding to one of the interfaces. As demonstrated in figure 10, the average selection time in the *Parallax System* per digit is very high when the participants first use the system. Although this decreases significantly in the subsequent experiments, the selection time per digit is still considerably higher as compared to the other two interface systems.

In terms of interaction speed, the three systems are ranked as follows:

1. The Tiled System
2. The Corner System
3. The Parallax System

Similarly, the user-assigned scores listed in Table 2 are used to rank the three interfaces in terms of ease of use. Using these data, the three interfaces are again ranked in the same order, which proves that the users are more comfortable with an interface that requires minimal interaction time.

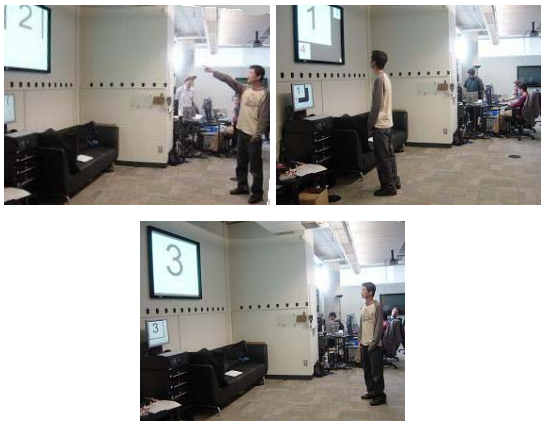


**Figure 9. A participant using the three User Interfaces: The Parallax System (top left), The Corner System (top right), and the Tiled System (bottom)**

| | Sequence | Average (sec) | Standard Deviation (sec) |
|---|---|---|---|
| **Parallax** | 1234 | 91.60 | 61.25 |
| | 2311243124 | 140.13 | 46.71 |
| | 1423214132 | 144.13 | 54.16 |
| **Corner** | 1234 | 27.27 | 5.63 |
| | 2311243124 | 76.67 | 11.95 |
| | 1423214132 | 75.13 | 12.65 |
| **Tiled** | 1234 | 26.07 | 11.70 |
| | 2311243124 | 64.53 | 18.82 |
| | 1423214132 | 67.60 | 14.55 |

**Table 1. Duration of each sequence selection for each interface**

## 6 Conclusion

This work proposes and compares sound localization-based navigational user interfaces. The speech signal generated by the speaker is recorded by an array of 24 microphones. These speech signals are then used to locate the speaker using TDOA histograms. The user can navigate through a website or change slides in a presentation by changing its spatial location.

| | Average | Standard Deviation |
|---|---|---|
| **Parallax** | 1.6 | 0.611010 |
| **Corner** | 3.7 | 0.927361 |
| **Tiled** | 4.1 | 1.098483 |

**Table 2. User-assigned scores for each interface**

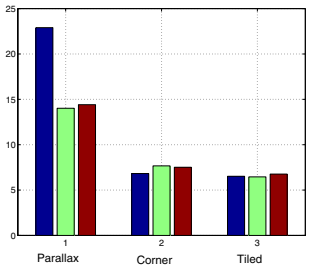


**Figure 10. Selection time (per digit) for each of the interfaces**

Three main interfaces are introduced, each of which makes use of Sound Localization to locate the position of the speaker (*The Parallax System*, *The Corner System*, and *The Tiled System*).

In each interface, the speaker can navigate through a set of images, or different parts of the same image, by changing the position of the sound source/speaker.

By studying the results of the experiments on 15 participants, it was observed that the *Tiled System* and the *Corner System* required far less interaction time than the *Parallax System*. Moreover, the participants were asked to rank the three interfaces in terms of ease of use; it was again observed that the users are more comfortable with an interface that requires minimal interaction time.

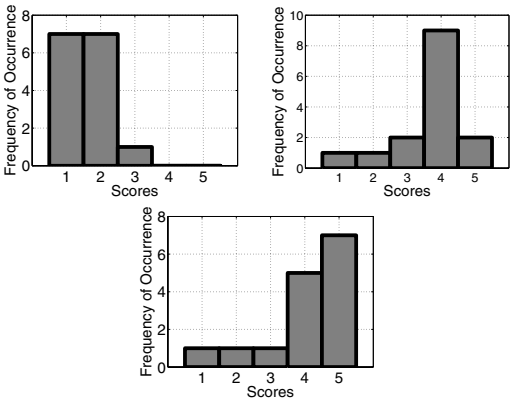


**Figure 11. Histogram of the user-assigned scores for the Parallax System (top left), the Corner System (top right), and the Tiled System (bottom)**

## References


[1] P. Aarabi. Self localizing dynamic microphone arrays. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 32:4:474:484, Nov. 2002.

[2] P. Aarabi. The fusion of distributed microphone arrays for sound localization. *EURASIP Journal on Applied Signal Processing Special Issue on Sensor Networks*, 2003:4:338–347, Mar. 2003.

[3] P. Aarabi, Q. H. Wang, and M. Yeganegi. Integrated displacement tracking and sound localization. In *Proceedings of the 2004 IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), pp. 171-174*, volume 5, page 937:940, May 2004.

[4] M. Brandstein and H. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, May 1997.

[5] J. DiBiase, H. Silverman, and M. Brandstein. Robust localization in reverberant rooms. *M.S. Brandstein and D.B. Ward (eds.), Microphone Arrays:Signal Processing Techniques and Applications*, 2001.

[6] D. Halupka, N. J. Mathai, P. Aarabi, and A. Sheikholeslami. Robust sound localization in 0.18um cmos. *IEEE Transactions on Signal Processing*, 53:6:2243:2250, June 2005.

[7] A. E. Kirkpatrick. Interactive touch: haptic interfaces based upon hand movement patterns. In *CHI '99: CHI '99 extended abstracts on Human factors in computing systems*, pages 59–60, New York, NY, USA, 1999. ACM Press.

[8] C. H. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24(4):320–327, Aug. 1976.

[9] M. Kobayashi and C. Schmandt. Dynamic soundscape: mapping time to space for audio browsing. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 194–201, New York, NY, USA, 1997. ACM Press.

[10] M. Schneider. Towards a transparent proactive user interface for a shopping assistant. In *Workshop on Multi-User and Ubiquitous User Interfaces*, Saarbrucken, Germany, 2004.

[11] D. Vogel and R. Balakrishnan. Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users. In *UIST '04: Proceedings of the 17th annual ACM symposium on User interface software and technology*, pages 137–146, New York, NY, USA, 2004. ACM Press.

[12] E. M. Wenzel, S. S. Fisher, P. K. Stone, and S. H. Foster. A system for three-dimensional acoustic "visualization" in a virtual environment workstation. In *VIS '90: Proceedings of the 1st conference on Visualization '90*, pages 329–337, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.