EXPERIMENTS ON SPATIAL GESTURES IN BINAURAL SOUND DISPLAY

Pei Xiang, David Camargo, Miller Puckette

Center for Research in Computing and the Arts (CRCA)
California Institute for Telecommunications and Information Technology Cal-(IT)^2
University of California, San Diego,
La Jolla, CA 92093 USA
{pxiang, dcamargo, msp}@ucsd.edu

ABSTRACT

This paper describes our explorations of sonic display with HRTF filters. While the quality of HRTF filters usually strongly affects the accuracy of sound source localization, an awareness of the spatial trajectory patterns of the source sound materials in such binaural sound displays is particularly promoted. In the experiments, with different choices of sound materials and their spatial gestures, we tend to maximize the effectiveness of the display while minimizing the disturbance that could result from users' constant exposure to sound over headphone. We also experimented a prototype system that could enhance the perception of static environments for the visually impaired. Hardware and realtime software (Pd) implementations are made to estimate the feasibility of this system.

1. INTRODUCTION

Research in assistive devices for the visually impaired has a history of about 40 years. Since hearing has the broadest band for acquiring information after vision [1], audio display naturally has a significant influence on such designs. Our hearing mechanism allows us to locate sound with reasonable resolution in space, so 3D sound spatialization takes an active part in many sound designs. Spatialization with loudspeaker arrays produce convincing perceptions as they have localized air vibrations in space. However spatial hearing over headphones remains a very practical solution, since they occupy less physical space and their sound is usually individualized and private to the user. Many interfaces for blind computer users such as the "GUIB" project [2], "SPUI-B" Interfaces [3] and the Virtual Audio Reality (VAR) system [4] utilize HRTF (Head Related Transfer Function) filters and room acoustics to locate virtual sound objects in the binaural 3D space. The user can then interact with the virtual objects with audio feedbacks. Although real-time rendering systems for binaural sound continue to be implemented in different approaches, most of the sonification work done so far is still working with static point sources or a small number of slow motion sonic objects. We propose to design multiple rapid moving sound sources with complex spatial gestures, so that new perceptual effects may come into play. This is because we normally locate sound better when they are moving than when they are static [5], and it is easier to perceive a trajectory than a point.

Most visually impaired people rely on listening in their everyday life. Their ears are under constant training in identifying sonically active objects. Through these skills, they manage to navigate through busy city areas, to cross streets, to avoid vehicles or people passing by, etc. As our ear is a trainable device, most visually impaired should have developed sufficient skills for listening to the sonic world. Thus, sonification becomes more meaningful when it displays vision-only information, for example static environments like a room with furniture, a hallway with multiple entrances and walls, and concrete obstacles on a pavement. These situations mainly involve physical objects that don't make noise and most of them are surfaces instead of point sources. In our prototype design we use a sonar sensor to scan the surrounding for approximate geometric shapes and then sonify the surfaces with multiple moving sound sources which are individually filtered by HRTF in real time. The design criteria are:

- *source uncorrelation*: the sound sources should be uncorrelated to each other so that they are differentiated perceptually in space.
- *focused but relaxing sound*: the sound materials should be "focused" sources which are easy to spatialize. They should not create stress if listened for a long time.
- *texture created by trajectories*: trajectories of the sound sources should be complex enough so that they are perceived altogether as a texture. We don't want the user to focus too much on particular trajectories. [6]

This paper will introduce the prototype system structure and implementation details about the real-time HRTF rendering, then present two experiments with multiple moving sound sources.

2. THE PROTOTYPE

2.1. Structure

The system begins with a sonar ranging module that collects distance information from a transducer to a BrainStem chip, then with the chip's *TEA program*, the sensor data is passed to Pure Data (Pd)[1], a real-time software by Miller Puckette. Real-time rendering of binaural sound is then handled and choreography of the trajectories are designed. Finally the sound mixture is passed to the user. (Figure 1)

2.2. Hardware

Similar to systems like the Miniguide [7] and Sonic Pathfinder [8], we use a relatively crude solution for prototyping purposes. We choose basic modules including the Senscomp 7000 Transducer,

[1] http://crca.ucsd.edu/~ msp/software.html

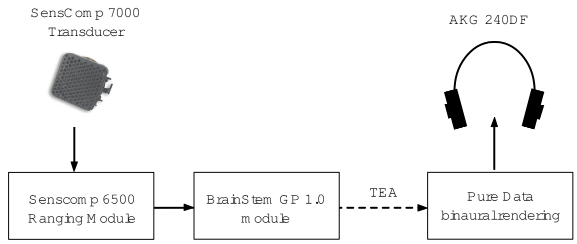

Figure 1: *System block diagram*

Senscomp 6500 Ranging Module, and Brainstem GP 1.0 Module from acroname.com [9]. The sensing is initiated by first creating sonic pings of roughly 16 high-to-low transitions between +200v and -200v. These transitions are fed to the transducer at around 50 kHz. Once the ranging module "sees" enough cycles of the reflected signal, it changes it's ECHO output to reflect the received reflected signal or echo, and distance is calculated as the product of sound speed and time delay of the pings. Figure 2 and Figure 3 gives the Two-echo timing diagram.

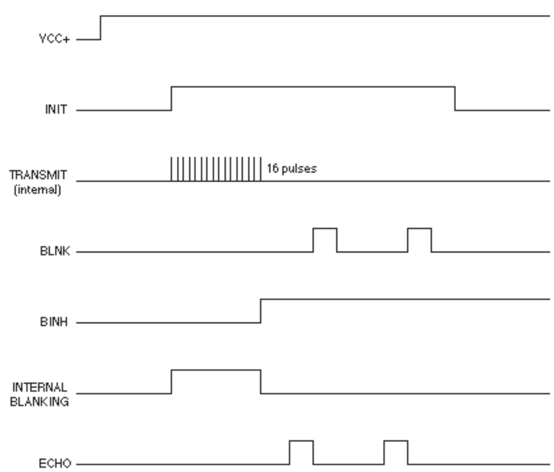

Figure 2: *Two-echo timing diagram of the Senscomp 6500 Ranging Module [9]*

For a distance of about 10 meters, it takes approximately 30ms to complete one distance measure. If mounted on a rotating head, it can perform 16 measures in a rotating proximity scan in about half a second. 0.5 second is a reasonably quick update of the pictures of the surroundings. A plot generated with one cycle of proximity scan is quite similar to the one in Figure 3 where notches and peaks can be later interpreted as objects with sound display.

The interface between the sonar data and Pd is done through the BrainStem module's *TEA program*. It is C based, and through socket programming, it passes data to Pd with UDP.

2.3. Binaural rendering

Real-time binaural sound production can be traced back to the hardware *Convolvotron* by Crystal River Engineering [10]. Recently software engines also becomes available [11]. To provide an easily re-useable software implementation for binaural spatial-

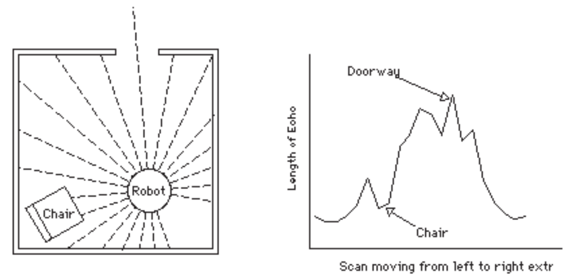

Figure 3: *Sonar proximity scan example [9]*

ization with interpolated HRTFs, we've developed a Pd external *earplug˜*. Instead of using several chosen HRTFs to make static virtual loudspeakers [11], *earplug˜* interpolates all available 368 (722 if mirror left and right ear) measured impulse responses on a spherical surface with -40 to 90 degree elevation and 360 degree azimuth. HRTFs are obtained from the KEMAR data sets [12]. Listening tests and error analysis have shown that shorten HRTFs up to 128 points yields a satisfactory localization accuracy [13] [14], thus the 128 point HRTF set of the KEMAR measurements is chosen here. In the data sets, the HRTF measurement points are not evenly distributed on the spherical surface, therefore a linear interpolation is chosen to save computational cost.

When spatializing a mono source with azimuth and elevation control, the point on the spherical surface determined by the controller values at the beginning of each signal block (Pd's default is 64 samples, approximately 1.5 milliseconds) is regarded as the sound source location for this entire block. The HRTF at this point is computed by linearly interpolating 4 points where the measured data is located. As an example, in Figure 4, P is the location point for the current signal block and a, b, c, and d are points with measured HRTFs, forming an enclosure around P. In the data sets, data points are actually in horizontal "rings" so that a and b turn out to have the same latitude, and so does c and d. Latitudes associated with each point are denoted by X, and longitudes by Y. Suppose HRTFs at these points are denoted by I_P, I_a, I_b, I_c, and I_d, then

$$\begin{aligned} I_P =& \frac{Y_P - Y_{ab}}{Y_{cd} - Y_{ab}} \cdot \left(\frac{X_P - X_a}{X_b - X_a} \cdot I_a - \frac{X_b - X_P}{X_b - X_a} \cdot I_b\right) + \\ & \frac{Y_{cd} - Y_P}{Y_{cd} - Y_{ab}} \cdot \left(\frac{X_P - X_c}{X_d - X_c} \cdot I_c - \frac{X_d - X_P}{X_d - X_c} \cdot I_d\right) \end{aligned} \tag{1}$$

Each sample in the current block is computed by the time domain convolution with a 128-tap filter. The filter is again a result of linear interpolation between two HRTFs. Let I_{P_0} be the HRTF of the previous block, and I_P the current one, a sample with index k in the current block is obtained by

$$\begin{aligned} y(k) =& \sum_{n=0}^{127} x(k-n) \cdot \\ & \left(\frac{k}{blocksize} \cdot I_P(n) + \frac{blocksize - k}{blocksize} \cdot I_{P_0}(n)\right) \end{aligned} \tag{2}$$

where

$$k = 0, 1, \cdots, blocksize - 1$$

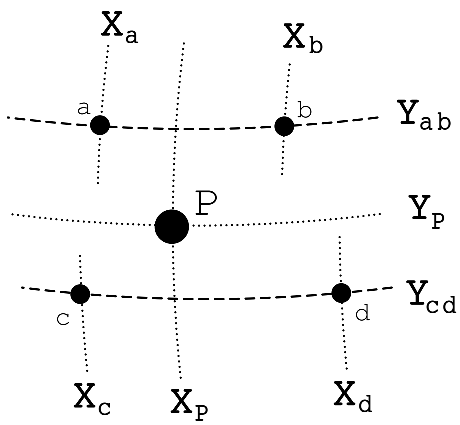

Figure 4: *HRTF Interpolation*

x denotes the input signal and y the filtered sequence.

In practice, this binaural rendering method gives convincing binaural effects and is capable of smoothly handling very rapid movements of sound. However, time domain convolution is computationally expensive. This prevents us from running many *earplug~* s on one computer. This external is currently downloadable.[2]

3. APPROACH

We experimented with two approaches, one with random organization of the spatial trajectories and another one with predictable gestures but more sound sources.

3.1. Random scan

Two kinds of sound sources were tested. One is synthesized FM sound with granular envelops to benefit the spatialization. The other one is filtered sample of water drops in a cave from the BBC sound library. It is a more natural and broad band sound. Mono sources are individually rendered by *earplug~* , so that the sounds "scan" within a circular track around the listener.

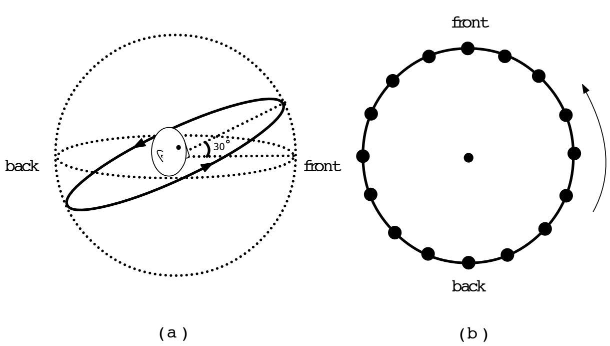

Figure 5: *Rotation planes*

Since front-back errors are very prominent in non-individual binaural filters [11], we choose not to locate the circular track in the horizontal plane. Empirical experience of the authors reveal that sound localization doesn't have very good results in the area where usually our eye covers, i.e. the frontal ear-level space. Binaural filtered sounds in that area are often perceived as inside the head or behind. We also find that frontal sound image is sufficiently convincing with higher elevation, and rear sounds with lowered elevation produces satisfying results. Thus, we tilt the track 30 degrees as is shown in Figure 5(a). A larger tilting angle like 40 degrees is not chosen because it deviates the sound image too much from the reality. In this way, all frontal images are slightly elevated and the rear images lowered. This elevation cue is a simple way to resolve front-back confusion, with some initial training to the user. Head-tracking for the binaural display is included, since we choose to obtain the information of the proximity with the sonar sensor mounted on the user's head (headphone).

During the scan, each sound source chooses a new location on the circular track after each time step (a random time interval, less than one second), then smoothly move to the new location along the shortest path on the track. Physical object distance is represented by amplitudes. During the movement, all sound sources refer to the same sensor input plot, like the one shown in Figure 3 (right), so that peaks represent farther physical objects and notches are the closer ones. In this way, each sound source gets amplified in places where there's an adjacent object, and attenuated when it scan through an open area. In order to create enough amplitude differences, the amplification curve follows an inverse quartic law, which means

$$linear\ amplitude = \frac{1}{(distance)^4} \cdot original\ signal \qquad (3)$$

This is a curve that close to exponential yet computationally efficient. After superimposing multiple sound sources in this way, it becomes hard to perceive their trajectory individually, but the amplification along the track accurately reflects the sonar data. The sonic display gives the impression of a continuous surface.

As the filter is computationally expensive, we can run two at once on an AMD XP2600+ PC. To simulate the effect of simultaneously running several of them, we run two filters a time and then mix the sound files. Sound examples of simulated simple scenes are available online.[3] In comparison, The FM sound is easier to localize, but prolonged exposure is unpleasant. The water sample is more soothing but it compromises a little in localization when compared with synthetic sounds.

3.2. Rotating plane

This approach uses only the water sample, and the movements are along the same circular track as in the previous approach. 16 similar but uncorrelated samples are arranged on the 30-degree tilted plane in a "ring topology" [15], as shown in Figure 5(b). By saying uncorrelated, we choose non-overlap time segments of the original sample, loop them, and assign them to the 16 spatial positions.

This plane then rotates with a constant speed, maintaining the relative position of the 16 samples, and each of them gets amplified the same way as described in the random scan approach. each sound source is also processed by a time-varying bandpass filter, in order to increase their perceptual differences and add complexity to the overall sound mixture.

[2] http://crca.ucsd.edu/~ pxiang/research.htm

[3] http://crca.ucsd.edu/~ pxiang/research/sonar_hrtf.htm

Sound examples are provided at the same link including simulations of walking through a space with static objects. In this approach, spatial patterns are much more regular, since every scanning source is rotating with a constant speed. At the same time, this regularity can be utilized to detect object width. With narrower objects which are interpreted in the sensor plot as narrow notches, sharp pulses can be heard, while for wider objects, less clear envelops can be heard, because the amplification envelop of moving sound sources at the detected object location overlaps each other creating a net envelop close to a constant gain.

4. DISCUSSION AND FUTURE WORKS

What we have done so far is an initial exploration into using multiple moving sources to display continues physical surfaces. Also, our system is a prototype design, thus many details can be improved in the future.

Perceived auditory distance is usually significantly compressed relative to source distance, and using a virtual acoustic display to convey motion of a source is likely to result in very significant perceptual distortion [16]. Also, we used the inverse quartic law for prominent representation of distance, whereas psychoacoustic research usually state inverse square law for familiar sound sources and inverse cubic law for unfamiliar sources [17]. Choices of distance laws and adjustments for distortions has to be resolved through subjective listening tests. Also, the current sound design mainly relies on the authors and some other CRCA researchers' listening experience. Many research indicates auditory compensation of reorganization and reallocation at the cortex level that benefits the hearing of the visually impaired [4], while some others claim no difference in sound localization along the azimuth between the blind and the sighted people [1]. These, together with different opportunities of auditory training cause many dissimilarities between sighted people and the visually impaired. Thus, it is very important to carry out listening tests for visually impaired people in the future.

The system itself can be improved in different ways. As the computing power increase or by rendering with multiple computers, the scanning sound source number can increase, creating more complex patterns. The random scan approach can be altered so that not only a circular track is scanned but also the whole spherical surface. In possible conditions, multiple sonar sensors can be used to speed up the update of the proximity scan and obtain data in more directions. So far, the distance is only interpreted by amplitude variations in sound. More features such as timbral change and sound source switch can be added to increase the perceptual dimension.

5. ACKNOWLEDGEMENT

This research is supported by the Center for Research in Computing and Arts in University of California, San Diego. The authors would like to thank the ICAD reviewers for their remarks.

6. REFERENCES

[1] R.W. Massof, "Auditory assistive devices for the blind," in *International Conference on Auditory Display (ICAD)*, Boston, USA, July 2003, pp 271-275.

[2] K. Crispien and H. Petrie, "The 'GUIB' Spatial Auditory Display–Generation of an Audio-based Interface for Blind Computer Users," in *International Conference on Auditory Display (ICAD)*, Santa Fe, USA, November 1994.

[3] L. Boelke and P. Gorny, "Auditory Direct Manipulation in Acoustical Objects by Blind Computer Users," in *International Conference on Auditory Display (ICAD)*, Santa Fe, USA, November 1994.

[4] C. Frauenberger and M. Noisternig, "3D audio interfaces for the blind," in *International Conference on Auditory Display (ICAD)*, Boston, USA, July 2003, pp 280-283.

[5] J. Blauert,"Spatial Hearing", 2nd ed., MIT Press, Cambridge, MA, 1997

[6] M. Gröhn,"Localization of a moving virtual sound source in a virtual room, the effect of a distracting auditory stimulus," in *International Conference on Auditory Display (ICAD)*, Kyoto, Japan, July, 2002, pp 1-9.

[7] http://www.gdp-research.com.au/ultra.htm

[8] http://www.sonicpathfinder.org/

[9] http://www.acroname.com

[10] http://www-cdr.stanford.edu/DesignSpace/sponsors/Convolvotron.html

[11] M. Noisternig, T. Musil, A. Sontacchi, and R. Höldrich, "A 3D real time rendering engine for binaural sound production," in *International Conference on Auditory Display (ICAD)*, Boston, USA, July 2003, pp 110-113.

[12] W. G. Gardner and K. D. Martin, "HRTF Measurement of a KEMAR," in *J. Acoust. Soc. Am.*, vol. 97, 1995, pp. 3907-3908

[13] A. Sontacchi, M. Noisternig, P. Majdak and R. Höldrich, "An Objective Model of Localisation in Binaural Sound Reproduction Systems," in *Proc. AES 21st Int. Conf.*, St. Petersburg, Russia, 2001 June

[14] A. Sontacchi, P. Majdak, M. Noisternig and R. Höldrich, "Subjective Validation of Perception Properties in Binaural Sound Reproduction Systems" in *Proc. AES 21st Int. Conf.*, St. Petersburg, Russia, 2001 June

[15] K. Crispien, K. Fellbaum, A. Savidis, and C. Stephanidis, "A 3-D Auditory Environment for Hierarchical Navigation in Non-Visual Interaction," in *International Conference on Auditory Display (ICAD)*, Palo Alto, USA, November, 1996.

[16] D. Berque, "Basic and applied research relating to auditory displays for visually impaired people," in *International Conference on Auditory Display (ICAD)*, Boston, USA, July 2003, pp 300-302.

[17] F.R Moore,"Elements of computer music", P T R Prentice Hall, Upper Saddle River, NJ, 1990