INVESTIGATING ARTISTIC POTENTIAL OF THE DREAM INTERFACE: THE AURAL PAINTING

Ivica Ico Bukvic
Virginia Tech
Music, DISIS, CCTAD

Denis Gracanin
Virginia Tech
CS, CHCI

Francis Quek
Virginia Tech
CS, CHCI

ABSTRACT

Discrete REconfigurable Aural Matrix (DREAM) is a multi-speaker array technology designed for sonifying spatial visual data using human anterior discrete spatial aural perception potential. DREAM treats each individual speaker as an aural counterpart to a pixel or a pixel cluster of an anterior visual display surface, such as an LCD screen. The pilot study was conducted to assess DREAM's ability to sonify geometric shapes ranging from simple static objects to more complex layered compositions and consequently to ascertain its potential as a complementing technology in a number of interaction scenarios, most notably as a foundation for the arguably novel art genre, the *aural painting*. Apart from spawning new creative and research vectors, the preliminary study has also yielded promising results with 73% users being capable of perceiving geometric shape, 78% shape location, and 56% shape size, thus warranting additional studies with larger speaker arrays and additional applied scenarios.

1. INTRODUCTION

Imagine an Art gallery. There is a grand opening of a new exhibition tonight. In one of the rooms strangely there are no paintings. Rather, a large frame inconspicuously covered with metal mesh is occupying most of the room's longest wall. As visitors silently enter the space, some move into a preferred place which they believe would give them the best perception of the artwork, while others continue to walk around the room in hope to gain greater understanding through different perspectives. All are busy studying the sound collage that emanates from the mysterious wall. The aural painting accurately portrays nuances of many spatially dependent entities that populate its invisible, yet perceptible world. As visitors' ears become sensitized to these details, a story emerges keeping visitors' interest piqued for an unusually long time. At the exit corridor, there is a chatter ensuing from many concurrent discussions where visitors are comparing notes and sharing ideas as to the artwork's story and the message.

2. BACKGROUND

Even though the human aural perception mechanism offers a unique ability to perceive and place aural stimuli emanating from any direction, studies show that our preference to face the source of our aural attention manifests itself even among visually impaired [13], suggesting that our anterior aural perception offers a much greater perceptive resolution than other angles.

From a historical perspective, experimentation with speaker arrays beyond the traditional stereoscopic systems reach as far as the beginnings of first electro-acoustic studios. Stockhausen in his pioneering work *Gesang der Jünglinge* (1955) [3] explored the diffusion using four speakers. Breakthrough studies by Ruff and Perret [14] showed that humans were not only capable of recognizing shapes using a 10x10 equidistantly spaced speaker matrix in conjunction with sound vectors, but that their success rate was considerably above chance. Although one would expect that such revolutionary findings would elicit a considerable amount of interest, the research in the area of large resolution two-dimensional speaker arrays until very recently can be best described as sporadic bursts. Arguably driven by the McGurk effect [11] despite its scope being limited primarily to English-speaking cultures [15], the Western society has turned its focus onto the visual domain in pursuit of complex spatial solutions.

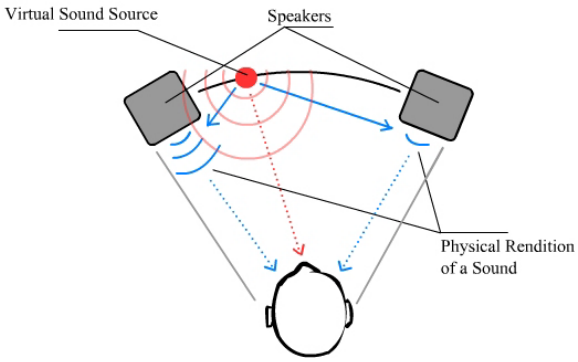

Figure 1. Virtual sound source using amplitude panning model

Existing sound spatialization technologies can be summarized into discrete (e.g. x.1 surround sound) and virtual (e.g. headphones). While virtual solutions offer a high degree of spatial perception accuracy, the inherent disconnect in respect to the listener's head orientation resulting in the spatialized content unaffected by changes in posture, limits the listener's ability to concurrently interact with the immediate surroundings,

thus resulting in a higher cognitive load. Virtual solutions [10] are also limited to a single–user interaction making them inadequate for multiuser and/or communal settings.

Discrete solutions [2], despite ongoing improvements in spatialization algorithms, point to the trend of growing a number of required speakers in order to deliver a greater immersion (Figure 1). Despite the much welcomed increase in speaker numbers, the anterior spatialization area remains sparse, commonly being populated by a single center and two side speakers (e.g. x.1 standard). When coupled with large contemporary display surfaces, such as HDTVs, where visual images can span across several square feet, it quickly becomes apparent that the existing anterior speaker cluster is simply inadequate for generating a corresponding level of immersion.

3. EXPERIMENT DESIGN

In order to assess the extents of human anterior aural perception and consequently explore practical uses of the gathered data, we developed Discrete REconfigurable Aural Matrix (DREAM), a tightly-packed 24-speaker 6x4 (1m x 1.1m) aural matrix interface (Figure 2).

DREAM's unique approach to sonification stems from the fact that we treat individual speakers as pixels and the cumulative surface as an aural counterpart to the traditional display technology. Just like visual content, the resulting sonified shapes are subjected to antialiasing that was managed through variations in sound amplitude.

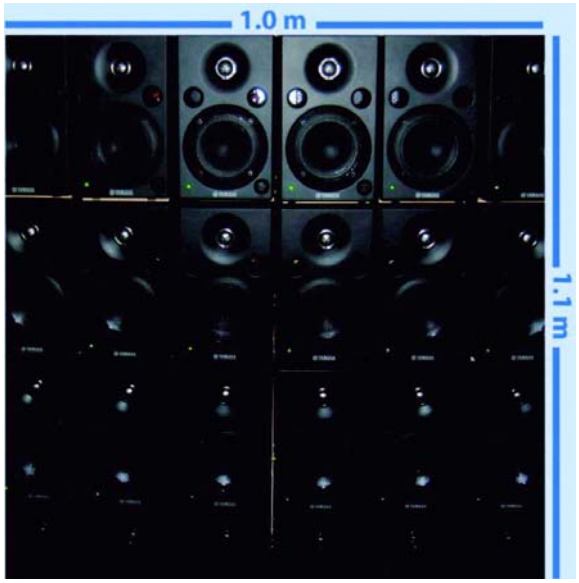

Figure 2. DREAM prototype

DREAM as a technology may appear as a rather common-sense solution that has seen ample research. Yet, while there has been considerable research done with speaker arrays [5, 6, 17], history shows that this specific area of research remains conspicuously sparse. This may be in part attributed to the ongoing efforts to economize technology for the delivery of optimal aural experience. Thus, during research slumps the interest in larger speaker arrays may have been supplanted by efforts to improve headphone-based perception with technologies such as HRTF and ambisonics, and its commercially-oriented counterparts SRS and QSound. Indeed, a number of intriguing research projects point towards the use of environments which rely predominantly upon virtual sound sources. Industry, burned by the commercial flop of early quad systems [7] has taken a rather precarious approach to adopting x.1 Dolby Surround standard which has since seen little commercial competition. The contemporary multimedia technology seems to be driven by the irony of incessant race towards bigger and better displays and in contrast ongoing attempts at delivering audio content with as few speakers as possible.

The last decade of the 20th century has given birth to the Wave Field Synthesis [16]. Its discovery and the slowly expanding x.1 surround standard boosted the research of complex speaker arrays. Driven primarily by the entertainment industry and academic community's interest in new ways of sound diffusion, this momentum resulted in formidable contraptions, including the Berlin Technical University's 840-speaker array. Yet, despite the momentum and significant accomplishments in areas of acoustics and aural spatialization, very few studies involving these technologies have shown a particular interest in the following:

- Significantly lowering reliance on virtual aural sources in order to bolster pinpoint accuracy and consequently the level of immersion (e.g. sources that emanate between two or more speakers through the use of various amplitude-based panning algorithms [12] (Figure 1), as is in essence the case with x.1 systems) in favor of discrete or at least near-discrete capable speaker arrays;
- Underexploited implications of the human inherent preference to face aural stimuli [1, 8, 12];
- Exploring the applied potential of a relatively high level or frontal discrete aural spatial resolution and consequently accuracy that is inversely proportional to an increasing azimuth [9];
- Use of speakers as a seamless flat canvas designed primarily for discrete projection of sound with minimal reliance upon virtual sound sources (speaker's function is akin to that of a pixel on a screen);

To measure the basic capabilities and accuracy of the system, we constructed a framework using Max/MSP [4] software in conjunction with a non-real-time SVG-to-DREAM vertex format translator that allowed us to

create a broad range of aural shapes, from single frequency point sources to multiple, dynamically rendered, moving shapes using broadband noise and timbrally rich sound samples. For this purpose we used following shapes: line, triangle, rectangle, circle (also used as a point), and ellipse. To prevent test subjects from discretely focusing on speakers, a thin screen of mosquito netting was placed in front of the array that occluded the speakers while having minimal effect on the outputted sound.

Subjects sat one meter away from the speaker array facing them head-on, with their ears aligned vertically and horizontally with the center of the array. The speaker array was located in a room that provided adequate acoustic insulation so as to minimize possible perception errors due to reverberation.

Prior to testing, all subjects were given a chance to acquaint themselves with the user interface and were given a list of possible shapes they will encounter in conducted tests. For each aural scene, the subject used a Wiimote as a pointing device with a cursor projected onto the mesh in front of speakers (Figure 3).

Figure 3. Human interaction with Wiimote and video projection on mesh surface

Users were able to record the shape, location, and size of the sound they perceived by pressing a button on the Wiimote using a visual feedback like that of a "paint trail" (Figure 4). In all of the tests, sources which were positioned or were moving between discrete speakers were antialiased using sine/cosine amplitude panning model and distributed among adjacent speakers.

Although we acknowledged a possibility of complex phase interactions among speakers in the tightly packed speaker array, we have not made any measurements of their impact on our study beyond noting their potentially detrimental effect. The resulting data was analyzed

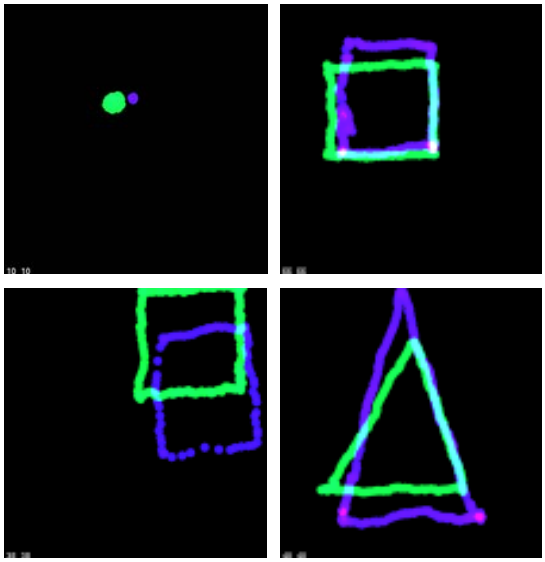

Figure 4: Example data from a test subject, green denoting the rendered shape and blue showing user response.

based upon benchmarks established by the existing research.

The study consisted of the following experiments:

1. Testing subject's hearing ability: In order to isolate any potential hearing deficiencies that may affect the data. Since this study had no facilities to accommodate potential deficiencies, we focused on subjects of various age groups, professions, and cultural/ethnic backgrounds whose hearing ability did not exhibit any notable impairments.
2. Locating sound using fixed discrete and virtual point sources: We used these tests to acclimatize subjects to the system as well as study best possible timbre choices for the purpose of sound localization. Our preliminary data supports existing research suggesting that timbraly rich sources, such as filtered white noise offer much better chance of perception than a single frequency.
3. Ability to track a sound vector: To affirm the existing research data we used five geometric shapes (line, triangle, rectangle, circle and ellipse) and presented their perimeter in a form of a vector using filtered white noise (1KHz bandpass with a Q of 1.4). While subjects were able to perceive shapes relative location, the shape and size recognition remained elusive.
4. Ability to recognize shapes using sonification of its surface: To build a potential vocabulary for populating the space with multiple objects, shapes were rendered using filtered white noise using a low-pass filter (1KHz with a Q of 0.6) that emanated from speakers covered by the shapes surface. This test exhibited diminished success rates

in terms of shape and size, while retaining comparable localization accuracy when compared to the previous experiment.

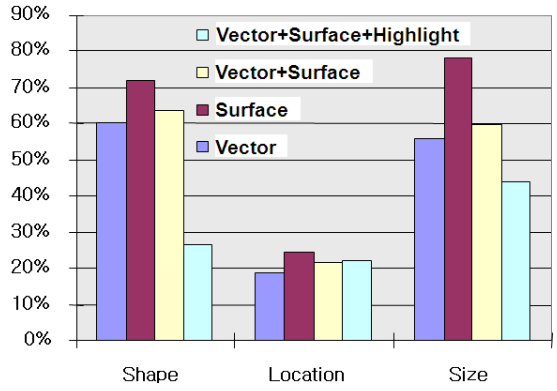

Figure 6. Subject perception error rates for shape, location, and size

5. Basic geometric shape recognition with static and dynamic sound: Same shapes were rendered using a combination of methods found in experiments 3 and 4. Their combined effect had brought perception efficiency back to the levels found in experiment 3 while retaining the presence of the aural surface layer.

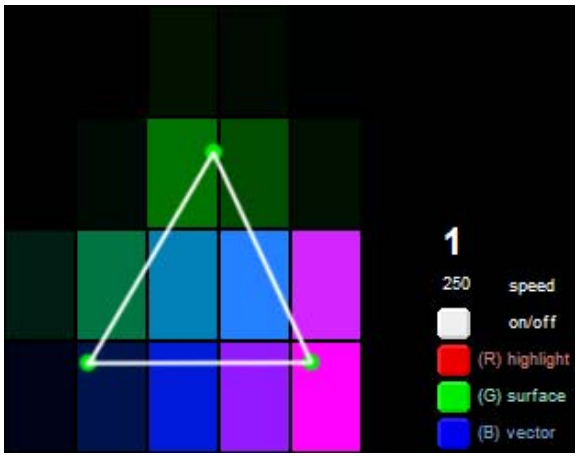

Figure 5. Layered audio rendering based on shape segmentation

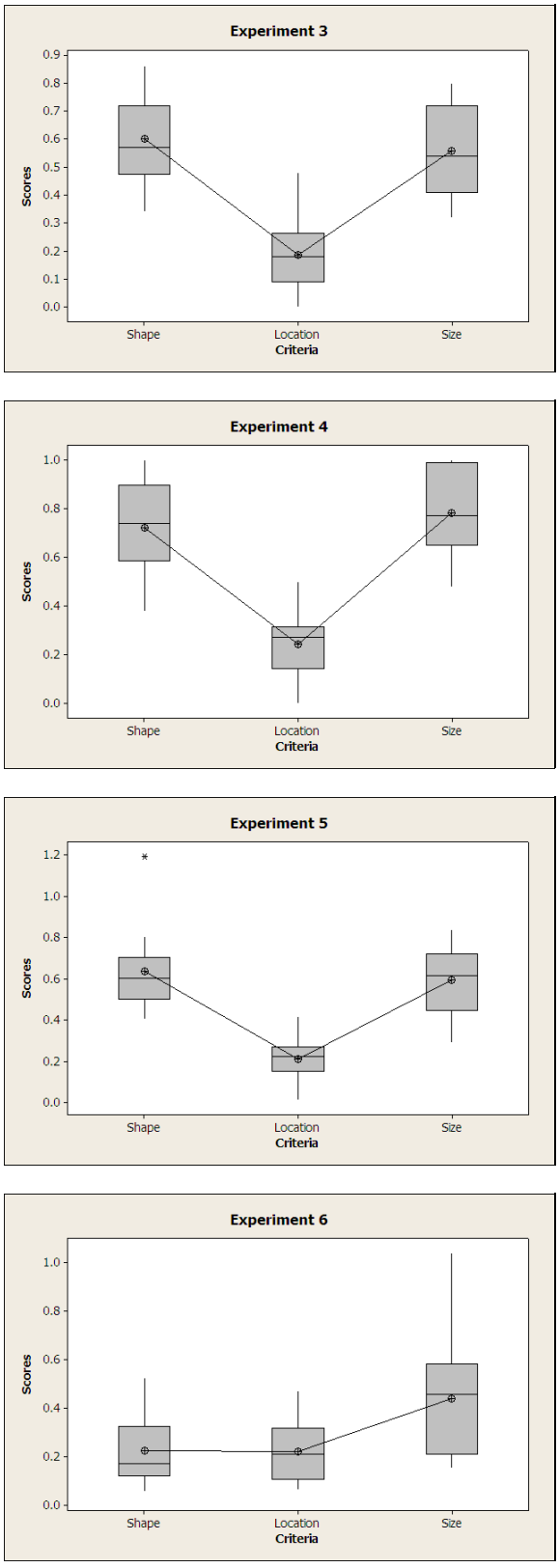

Figure 7. Perception differences for different source sounds.

6. Layered basic shape recognition using surface, vector, and vertex highlight: In order to further reinforce accurate perception of shape, location, and size, the two-layered approach used in experiment 5 was complemented with a vertex highlight method that was timed in conjunction with the vectorization of shapes perimeter (Figure 5). A high-pass filter (500Hz with a Q of 0.6) was used for this purpose with a sharp attack envelope. Shapes without vertices (e.g. circle and ellipse) were presented as 16-point polygons. Test subjects were made aware of this exception prior to taking the test.

4. PRELIMINARY RESULTS

We conducted experiments with 16 participants. The preliminary study has provided us with encouraging data regarding this methods potential as an augmenting technology for the purpose sonification, navigation, and ultimately interaction with complex spatially oriented content and multi-user interfaces. In order to score the user's accuracy, we developed scores based on three criteria, geometric shape, location, and size.

Criteria	Scores
Shape	0: Accurate Less than 0.5: same shape, but different orientation (e.g. ellipse ↔ circle, vertical rectangle ↔ horizontal rectangle, …) 0.5: similar shape (e.g. rectangle ↔ circle, rectangle ↔ similar triangle, …) 1: wrong
Location	0: Perfect match Between 0 and 1: how close it is to the original location. (partial overlapping). 1: there is no overlapping.
Size	-1: less than half of the original size Between -1 and 0: smaller than the original size, but not greater than half 0: perfect match Between 0 and 1: bigger than the original size, but less than twice the original size 1: more than twice of the original size *For statistical analysis, we convert any negative number to a positive number, e.g. -0.5 is converted to 0.5, because its absolute value is only considered for the analysis.

Table 1. The scores for criteria

The shape score is based on the similarity of the presented shape and the shape constructed by the user. The value ranges from 0 (exact match) to 1 (mismatch). Lower values (< 0.5) indicate the same shape but with different orientation.

The location score is based on the area of overlap between the presented shape and the constructed shape. The value ranges from 0 (total overlap) to 1 (no overlap).

The size score is expressed as the ratio between the size of the user-constructed shape and the presented shape. The value ranges from -1 (the ratio less than 0.5) over 0 (the ratio is 1) to 1 (the ratio is more than 2). Table 1 describes scores for each criterion.

The preliminary study has yielded promising results: 73% accuracy in recognizing geometric shape, 78% location, and 56% shape size in experiment 6 (Figure 6).

One interesting result is illustrated in Figure 7. In experiments 3, 4, and 5, subject perception error rates were significantly ($F(2, 45) = 3.20$, $p < .05$) lower for 'location' than for the other two parameters. Likewise, experiment 6 shows that there was a significant difference ($p < .05$) between 'size' and the other two parameters.

One consequence is that the basic shape recognition was noticeably better when the three-layered surface, vector, and vertex highlight approach was used, but the accurate size recognition had showed only marginal improvement. From this result, we suspect that the lower size accuracy was likely affected by the low resolution of the speaker array [5], i.e. the limited 6x4 resolution of the prototype, preventing us from ascertaining extents of human anterior perception. We also suspect that the test scores of latter experiments may have been affected by participants' fatigue as the entire test took approximately an hour to complete. The last experiment's considerably increase in size deviation is certainly suggestive of this hypothesis.

Although additional research is required to further refine the layered approach to shape sonification including its spatial composition and the filtering of the spectral content, it is encouraging to observe that despite the layering that has resulted in a more complex aural image, the overall ability to perceive the three parameters either improved (e.g. shape and size in experiments 3-6) or remained consistently good (e.g. location across all the tests). In this respect even the marginal improvement in the shape size found in the sixth experiment has some test subjects scoring as low as 0.2.

5. PONG WITH COMPOSITE AURAL SHAPES

As an exercise, DREAM was also coupled with a simple Pong game [18] in which two paddles were controlled with Wiimote (Figure 8). Paddles and the ball were invisible to users. Although no scientific measurements

were taken to assess the success of this exercise, playing the game offered the research team many hours of fun.

The Pong example also suggests that the system has a potential at delivering concurrent multiple shapes in a perceptible and efficient fashion which would allow it to be used as a platform for delivery of complex spatial data in a variety of scenarios, most notably as an assistive technology for visualization of images for the blind.

We also observed that there was a variation among test subjects in terms of their ability to interact with the system. While a majority of subjects performed well in our tests, a small contingent of subjects showed either a consistent lack of ability to recognize any of the parameters (shape, location, and size) or performed exceptionally well. This tri-modal distribution warrants further study in personal variations.

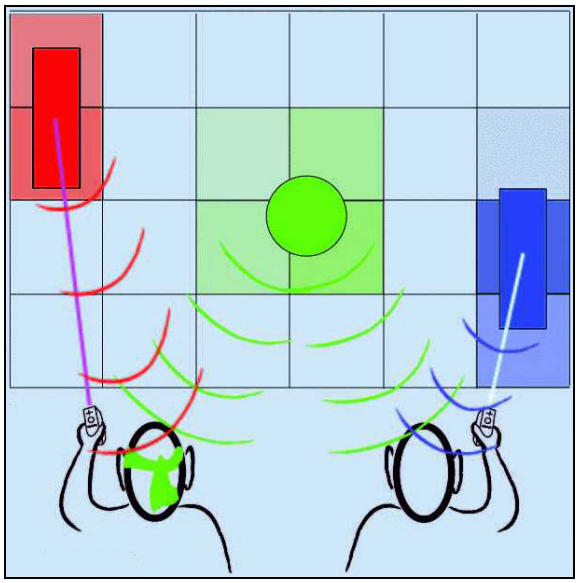

Figure 8. Aural Pong

6. THE AURAL PAINTING

As can be observed from the aforesaid Pong scenario, DREAM offers a platform for exploration of a number of applied scenarios, including art and entertainment. As a first step towards uncovering its full potential, we designed three additional tests that were informally presented to subjects. The tests consisted of three concurrent shapes, each "painted" using a concrete, natural sound, thus resulting in cumulative aural image conveying a particular story or a mood.

The choice of a natural sound in this case seems both practical and logical as humans are inherently drawn to and clearly influenced by familiar sounds. This critical psychoacoustic trait affords us an easy way to attach meaning to otherwise abstract geometric shapes. By spatially layering shapes we are in effect "painting" an image whose properties in many ways correspond to that of a visual painting:

- Every painted area can be simplified into a surface of a particular shape
- As the shapes overlap, so does their aural content. Thus with a careful choice of sounds, one could sonify interaction between the two colors and/or textures.
- The strength and prominence of the shape or its relative position to other shapes can be punctuated using attenuation and filtering of the sound.

While this approach clearly cannot reproduce every nuance of a visual painting, this shortcoming is supplanted by a set of unique opportunities:

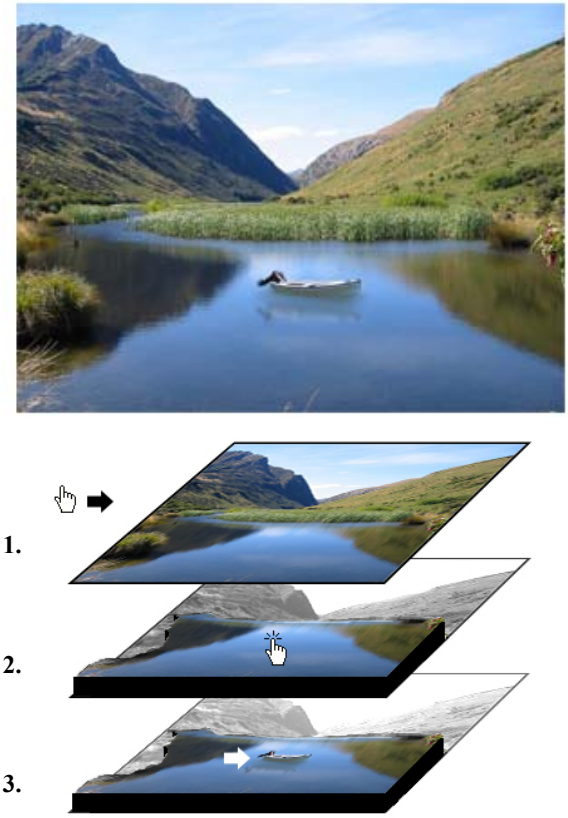

Figure 9. Layered approach to handling detail

- Akin to the dynamic version of its visual counterpart (e.g. movie), the aural painting can also change over time, offering a quantifiable and more importantly controllable temporal aspect. Complex imagery could thus cycle objects in and out of focus so as to allow listener to perceive nuances that may be otherwise lost in the overall collage.
- With the use of an adequate navigation interface (e.g. Wiimote), user can focus on a particular shape

in order to bring it to focus by increasing its amplitude and/or attenuating other shapes represented on the aural canvas. Once the shape is emphasized, additional layers can be brought out for added detail (Figure 9).

The layered approach using shape segmentation presented in the preliminary study can be applied to just about any sound whose spectral composition is rich enough to offer minimum necessary amount of content per layer. Some observable exceptions do exist, however. For instance, stemming primarily from the fact that for a sound to be perceived it requires the context of time, if a natural sound has a tendency to do filter sweeps akin to pinna-induced filtering that helps us place sound sources vertically, it can inherently provide user with misleading image of a sound that ascends and/or descends.

With the aforesaid considerations in mind, the three shapes in these tests consisted of a sound of water, wind, and a seagull. The first test positioned the three shapes in a way that is evocative of natural occurrence: the water was a rectangle at the bottom half of the DREAM canvas, air at the top half, and the seagull was positioned in the top-right corner. The second and third test had utilized the same aural shapes, but their position was less "natural" (e.g. seagull was emanating from the water or the horizon was vertical).

While we did not gather user perception data nor conducted any kind of a questionnaire in association with these tests (primarily because at the time we lacked the conclusive data that would affirm the effectiveness of the layered approach to sonifying visual content) users have met the experience with pronounced curiosity and enthusiasm.

Based on their feedback, one could clearly envision use of the newfound genre in an Art gallery, a vehicle for interactive storytelling and theatre, even as a décor in reasonably quiet public or personal spaces (e.g. homes). Likewise, although very promising, the entertainment potential beyond the aforesaid Pong scenario remains largely unexplored.

7. FUTURE WORK

We are unquestionably excited by the findings and are continuing to pursue newly identified areas of study with vigour. There are a number of lingering questions that require our immediate attention. Answering them will affirm the importance and value of the DREAM interface.

Despite the obvious advantages over existing technologies in communal scenarios (e.g. where two or more users are to concurrently perceive and/or interact with the content, or where user's aural perception should not be limited through the use of headphones, as is the preference of the visually impaired), there is an obvious need for a comparative study to quantify its advantage over virtual spatialization systems that utilize considerably smaller number of speakers. We anticipate conducting such a study in the summer of 2008.

We plan to conduct additional studies that measure DREAM's effectiveness using more complex imagery, and consequently its impact as a vehicle for the aforesaid art form as well as entertainment scenarios.

Based on the gathered data, we see DREAM as a foundation for a new breed of assistive technologies with special focus on the visually impaired, as a means of enhancing the consumer audio format, and as a foundation for a new art genre. DREAM is both the framework by which we will conduct the research and a technology (or a vehicle) by which the newfound method of sonification could be applied to a variety of scenarios, including:

- assistive technologies for the disabled (e.g. sonification of imagery for visually impaired);
- patient rehabilitation and therapy (e.g. interactive interface for rehabilitation of a stroke patient by facilitating retraining of the right arm);
- navigation and interaction within complex communal and collaborative environments (e.g. augmenting the audio-visual idiom by supplanting the anterior aural output);
- spatial awareness in virtual interactive environments (e.g. enhancing virtual classrooms);
- augmenting spatially-oriented time-critical interfaces (e.g. navigation interfaces).
- serving as a foundation for augmentation of existing and creation of entirely new artistic genres, such as the *aural painting*.

8. REFERENCES

[1] Abouchacra, K. S., Emanuel, D., Blood, I., and Letowski, T. 1998. Spatial perception of speech in various signal to noise ratios. Ear & Hearing 19, 4, 298–309.

[2] Bronkhorst, A. 1995. Localization of real and virtual sound sources. The Journal of the Acoustical Society of America 98, 5, 2542–2553.

[3] Chadabe, J. 1997. Electric Sound The Past and Promise of Electronic Music. New Jersey, Prentice Hall.

[4] Cycling 74. Max/MSP/Jitter. Retrieved from: http://www.cycling74.com/products/max5.

[5] Dobler, D., and Stampfl, P. 2004. Enhancing threedimensional vision with three-dimensional sound. In ACM SIGGRAPH 2004 Course Notes, no. 25.

[6] Hamasaki, K., Komiyama, S., Hiyama, K., and Okubo, H. 2004. 5.1 and 22.2 multichannel sound productions using an integrated surround sound panning system. In Proceedings of 117th AES Convention.

[7] Kapralos, B., Jenkin, M., and Milos, E. 2002. Auditory perception and spatial (3d) auditory systems. Tech. rep., York University.

[8] Langton, S., and Bruce, V. 1999. Reflexive visual orienting in response to the social attention of others. Visual Cognition 6, 5, 541–567.

[9] Lewalda, J., and Ehrenstein, W. 1998. Auditory-visual spatial integration: A new psychophysical approach using laser pointing to acoustic targets. The Journal of the Acoustical Society of America 104, 3, 1586–1597.

[10] Lumbreras, M., and Sanchez, J. 1999. Interactive 3d sound hyperstories for blind children. In Proceedings in SIGCHI 1999.

[11] McGurk, H., and MacDonald, J. 1976. Hearing lips and seeing voices. Nature, 746-748.

[12] Pulkki, V., 2001. Spatial sound generation and perception by amplitude panning techniques. Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology.

[13] Roman, N., Wang, D., and Brown, G. 2003. Speech segregation based on sound localization. The Journal of the Acoustical Society of America 114, 4, 2236–2252.

[14] Ruff, R.M., and Perret, E.. Auditory spatial pattern perception aided by visual choices. Pyschological Research 38, 369-377.

[15] Sekiyama, K., and Tohkura, Y. 1991. McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. The Journal of the Acoustical Society of America, 90, 4, 1797-1805.

[16] Sporer, T. 2004. Wave field synthesis - generation and reproduction of natural sound environments. In The 7th international Conference on Digital Audio Effects (DAFx'04).

[17] Trueman, D., and Cook, P. R. 1999. BoSSA: The Deconstructed Violin Reconstructed. In proceedings of the International Computer Music Conference.

[18] Winter, D., 2007. Atari pong - the first steps. Retrieved from: http://www.pong-story.com/atpong1.htm.