Audio Textures

Opening The Inner Eye Through Music

New Mexico Supercomputing Challenge

Final Report

April 3, 2013

Team 80

New Mexico School For The Arts

Team Member:

Mohit Dubey

Teacher:

Acacia McCombs

Mentor:

Stephen Guerin

Table Of Contents

Executive Summary:

The purpose of this project is to apply the Gibsonian approach towards perception to the problem of blind navigation. Using NetLogo and its various extensions, we built a model that translates the visual world into an aural world based upon a practical mapping that analyzes the individual aspects of each patch, or pixel, of an image and produces a note based upon those characteristics. In doing so, I faced many challenges that caused us to implement certain techniques that would keep our transformation from sight to sound as simple as possible such as edge detection, background color reduction, and scale interpretation. Next, I explored different ways to make our model applicable and testable using real world and computer based methods, including designing and coding an entirely new program that tests the effectiveness of our original model in a video game style manner. Finally, using the results from our experiments I created a final product that can provide a basic, successful guidance to someone without sight through music.

Introduction:

Gibsonian Approach To Perception

Around the time of World War II, the American government became very interested in the scientific properties of human perception. Although their original intent was to calculate how warplane pilots could most effectively drop bombs on targets while flying at high speeds, many of the researchers in this field discovered very important things, universal to science. First of all, they realized that they could not conduct experiments in a dark laboratory while attempting to understand how a person perceives depth and distance while flying. Tests had to be done out in the real world. Most importantly, these researchers discovered that in order to identify objects in space, we must first identify the background. In terms of the airplane pilot, the ground and the horizon are far more important to how a bomber sees than the airspace in which he is flying.

As J.J. Gibson says in his preface to The Perception Of The Visual World, "We perceive a world whose fundamental variables are special and temporal – a world which extends and endures" [1]. Therefore, in order to understand how we relate to space and time we must look at the most basic rules of human perception. Now, according to our understanding there are only five senses: sight, touch, smell, taste, and hearing which we use to gather information from our surroundings. However, as you know by being a human who can read this paper, most of that information is relayed to you through your sense of sight. Now we face a curious dilemma; how do these ocular lenses on the front of our face region, which simply create a two-dimensional "retinal image", provide a full picture of reality? Gibson and his followers, called ecological psychologists, propose that we actual perceive the world as a "flow field", where, instead of visualizing the world as a series of flat projections, we understand what environs us through the relative motion of objects in space. For example, as you move towards a door, you see that the door is getting bigger as objects around you pass to the side. The door is not moving, but because you notice its expansion, you know that you are getting closer to it. As Kafka said "Just as motion for the physicist can be specified only in relation to a chosen coordinate system, so is a phenomenal motion relative to a phenomenal framework" [2].

Gibson's theory is based around the concept of "optical flow", or relative motion of objects in a certain field due to the motion of the observer. Such motion can be approximated mathematically by stating that the sum of the partial derivatives of brightness with respect to each spatial dimension is equal to the partial derivative of brightness with respect to time. For two-dimensional space this is represented by:

$$(E_x, E_y) \cdot (u, v) = -E_t.$$ [3] Where (E_x, E_y) represents the brightness derivative with respect to the subscripted axis and (u, v) represents the derivatives of the spatial dimensions with respect to time. Optical flow has many uses and is currently being applied in the field of robotics to help create depth perception and improve obstacle avoidance. The understanding of how objects move in an optical "flow field" relative to how an observer moves is the fundamental focus of the ecological psychology approach to perception.

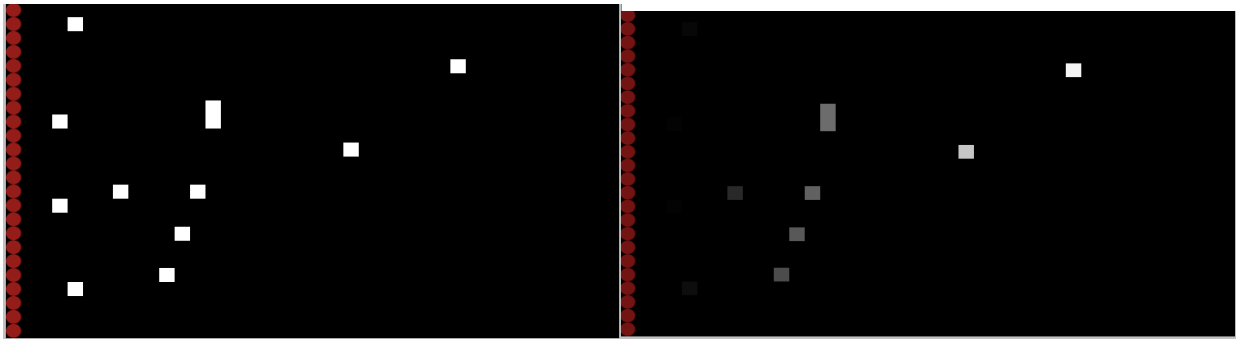
Our Goal Of Blind Navigation

This approach to cognitive perception, dubbed Gibsonian after its author, is very useful when applied to the problem of blind navigation. Optical flow is an excellent means of approximating where objects are in your vicinity and how they are moving relative to you. However, without sight there is still no way to know where these objects are and whether you are bound to collide with them or not. Our project aims to use a mapping from images of 3-dimensional space to 4-dimensional sound in order to provide that exact information. By using Gibson's techniques to interpret the information, then translating them into aural cues, we hope to create a program that can provide assistance in blind navigation.

Computation:

What Is The Best Mapping From Image To Sound?

Our project begins with this simple question, which we can answer effectively with the idea of a "player piano". If we consider the NetLogo world, composed of patches (our chosen dimensions are 40 x 24), as a sort of musical score that is played from left to right over time by a line of turtles, or players, then we have a fine solution. Now we have solved that the x-dimension of our image will be the musical aspect of time, but music consists of four distinct elements: duration, pitch, intensity, and timbre. How, now, shall we distribute the other three variables? We know that the pitches of musical notes have a certain pleasant range, with certain notes being too high and certain notes being too low. This could easily be fixed to the y-dimension of our image, therefore providing a constant harmony to the image (when every patch is played). Next, we can analyze the characteristics of each patch to discover means of translating our last two rudiments of music. In Netlogo, each patch has a specific "pcolor" on a scale from 0 to 120 which can, in turn, be expressed as a combination of a specific hue, saturation, and brightness or amounts of red, green, and blue. For the sake of simplicity we have chosen to express the timbre of each converted note through the patches' "pcolor" and the intensity through their brightness. Therefore, patches that are completely black are not played at all and different instruments perform different colored patches. For example:



The image on the left would be played at a constant volume, while this image on the right would have an increasing volume because the patches slowly become lighter, but both would be played by the same instrument because they are all the same "pcolor".

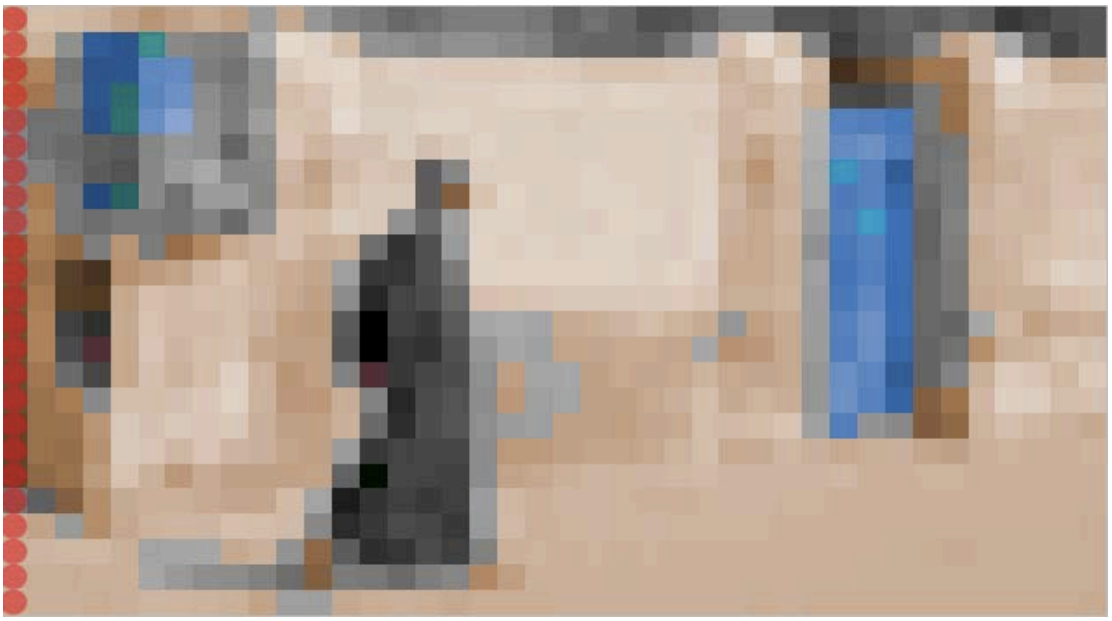
How do we know that this is the best mapping, though? Couldn't we have just as easily called the y-axis time or had our red "players" start from a corner of the image and move out? Yes, we could, but we would end up with nearly the same results: the dimension that we consider as time is arbitrary. How about the other variables? Well, suppose we made patches with higher "pcolor" values have higher pitch and then have the y-axis determine the timbre of the sound. We would end up with a constant wash of timbre and a completely random sounding mess of pitches that would jump all over the place each time the turtles would take a "step". At least by mapping the y-coordinate of a patch, the image has the possibility of having a specific melody and the color to timbre translation just makes sense aesthetically (when you listen to symphonic music, the conductor sounds as though he or she is painting with the different instruments of the orchestra).

Reducing Entropy Within The Model

Once we developed a working translation from the visual world to an aural world, we set out to make it the as efficient and user-friendly as possible. To begin, we made sure that the music being played was spread out tonally rather than in a chromatic manner, which the NetLogo sound extension does naturally. This means that instead of the notes being merely a half step apart and therefore causing dissonance, we implemented a spacing based upon the major scale using the "tonalizer" value seen below:

```
ask patches with [ pycor >= 0 ][ set tonalizer 0 ]
ask patches with [pycor >= 3][ set tonalizer 1 ]
ask patches with [pycor >= 7][ set tonalizer 2 ]
ask patches with [pycor >= 10][ set tonalizer 3 ]
ask patches with [pycor >= 14][ set tonalizer 4 ]
ask patches with [pycor >= 17][ set tonalizer 5 ]
ask patches with [pycor >= 21][ set tonalizer 6 ]
ask patches with [pycor >= 24][ set tonalizer 7 ]
```

Next, we added a piece of code that would allow for sustained notes when two patches of the same “pcolor” are adjacent to one another. Finally, we included a value called the “backgroundcolor”. Basically, this calculates the most common color in the image and treats it as though it is black and therefore is not played. We did this so that color-rich images wouldn’t end up sounding like the same harmony played over and over with only variations in timbre due to the fact that none of the patches have zero brightness.



This image would filter out the light brown “backgroundcolor”, cleaning up the sound.

Expanding The Model To The Real World

Now we have the potential to expand our model to the real world. Our first step was to use the camera on an Android Smartphone in order to bring images directly into the model from our surroundings (previously we were importing .png files from a folder of images). In order to do so, we had to introduce the “url” and “bitmap” extensions into NetLogo and write a method for the images to be taken from a webcam app running on the Android.

Once this was done, we could start interpreting our world using our Gibsonian method. However, first, we wanted to have a means of understanding what was actually

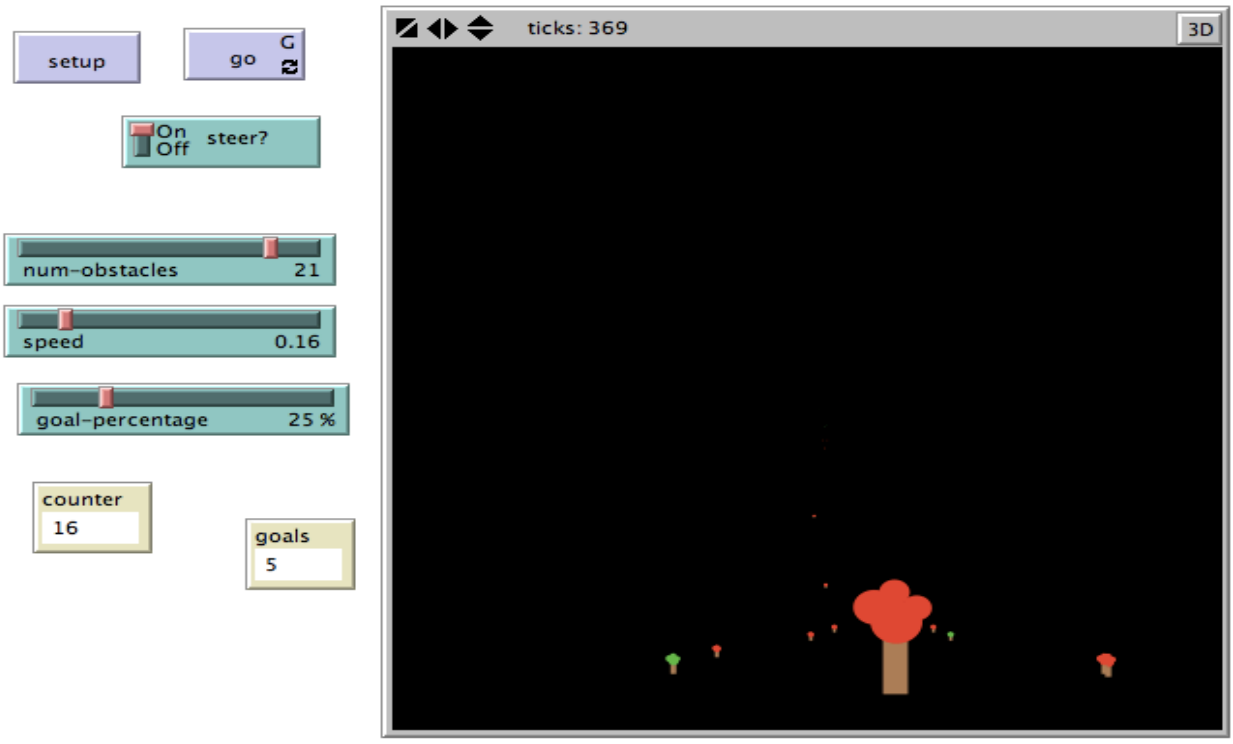
being seen in our optical region, so we introduced a basic form of edge detection using a Sobel filter. A Sobel filter detects edges by changing the brightness of a patch to eight multiplied by the brightness of the patch minus the sum of the brightness of the eight surrounding patches. This process turns all patches that are not along a “Sobel edge” black so that they are not played by the turtles.

Then by using our Android connection and our Sobel filter, we could begin doing audio translations of optical flow. Originally we recorded movies on the Android, imported them into NetLogo, and then created a second version of the same video with the Sobel filter applied which the model could interpret by simply performing the remaining edges. By knowing the motion of edges, a blindfolded person could navigate towards an object because the edges of that object would get larger and larger as the navigator gets closer resulting in a louder and denser “audio texture”.

Application:

A Way To Test The Model Using Netlogo

In order to test the effectiveness of the model, we created a video game in which a blindfolded player must avoid approaching obstacles in a flow field using our provided visual cues. This "soundskier" model was also coded in NetLogo and uses a slight variation of the mapping we have been using in our original model: distance on the y-axis corresponds to pitch, distance on the z-axis corresponds to loudness, and the color of the instrument corresponds to its timbre. The main difference between the game and the actual model is that, in the game, the objects themselves have corresponding musical notes where as in our model turtles must "play" the patches in order to create the sounds related to the image. In this game, tree shaped objects move towards the player and every certain amount of ticks, or time steps, the player hears how near the trees are. This allowed us to optimize how frequently the objects should provide aural information for navigation.

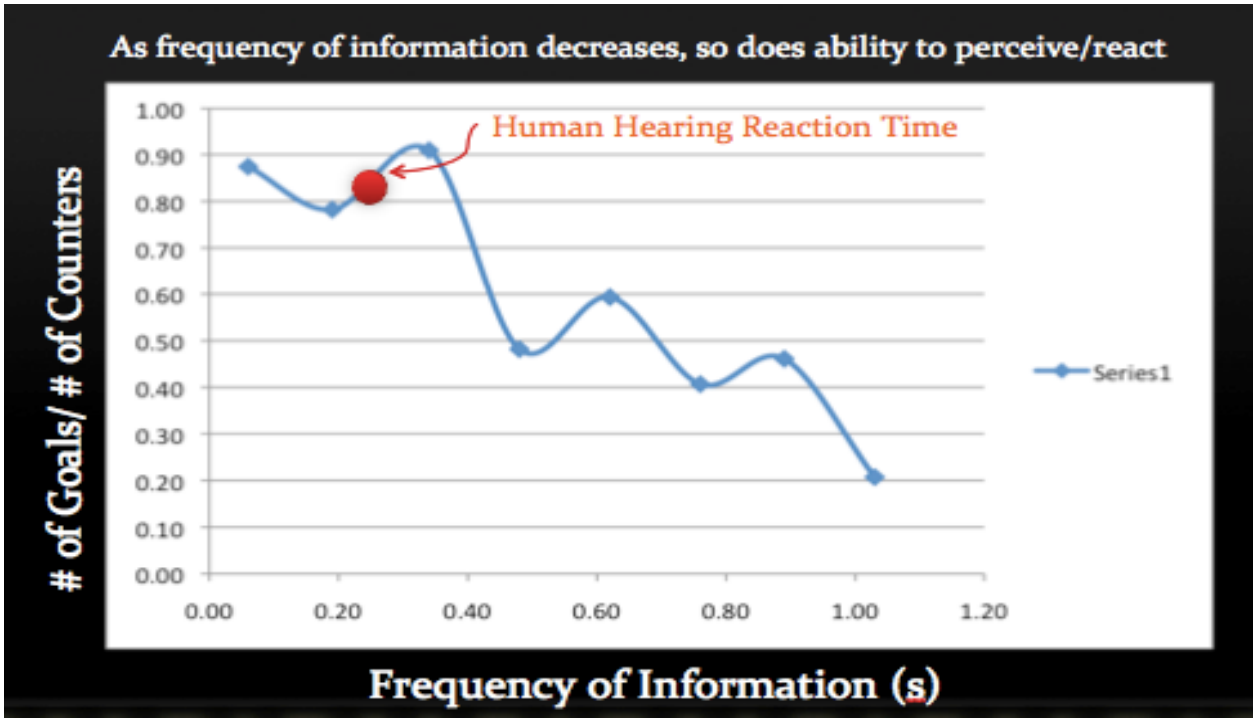


In this image you can see trees approaching the observer, sliders to change the number of trees, speed of the model, and number of trees that are “goals”. Below that are two boxes labeled “counter” and “goals”. Basically, to test the effectiveness of this “soundskier”, we created “goal” trees with one timbre and “counter” trees with another timbre. The job of the player was to try at hit as many goal trees as possible while avoiding the “counter” tress at all costs based solely on the difference in sound between the two.

Results:

First of all, we found that the human ear requires very distinct differences in timbre in order to react to an aural stimulus. We saw this in the difference in the average number of goals per counter ratio taken over multiple players for two different instrument settings: one where goals were "Piccolo" and counters were "Timpani" and the other when goals were "Pizzicato Strings" and counters were "Tremolo Strings". When the two timbres were as different as piccolo and timpani, the players scored about 15% higher on average than when the sounds were more similar (same instrument different technique).

Next, we discovered that expected values for the human aural reaction time were very near to the values we obtained through our model. Again, after averaging over multiple trials we plotted the percentage of goals per counters against the amount of time lapsed between each "sounding of the trees", or the frequency of the information:



As you can tell from the graph, the model is most effective when the frequency of aural information is in the same range as the human hearing reaction time. This proves that we

require a pretty well defined frequency of aural information in order to react to our surroundings and that our model must suit that need.

Finally, we concluded that we could create a model that applies the Gibsonian idea of perception as a function of the relative motion of objects to the problem of blind navigation using visual to aural translation. Using NetLogo and an Android phone we could effectively guide someone to an object using this technique as well as come up with a means of testing its effectiveness through the computer. Audio Textures has been successful in bringing together the arts and sciences in a computer program.

Further Research:

Creating An Android Application

Currently our model runs by important a constant stream of images from an Android webcam enabled Smartphone into our NetLogo model which can then send the audio through a Bluetooth headset and navigate a person using sound. In the future we hope to find a coding language, most likely JavaScript, which will allow us to program our model as an Android app so that the entire process may simply run on a Smartphone. By doing that, we hope to have our model available to people who are actually visually impaired and get their responses about our work.

The Future Of "Audio Textures"

In conclusion, we would like to say that this has been a wonderful project to work on and it doesn't end here. We hope to expand our model to include a wider midi database that would support left-right stereo panning as well as recode our model in order for it to function entire as a Smartphone app. Throughout this year, I have learned an incredible amount about myself as a programmer and worker and have found a passion for bringing aspects of the arts into computer science which I will continue pursuing throughout my life. You can keep in touch with us at www.audiotexture.org, thank you.

Gratitude And References:

None of this would have ever have been possible without the tremendous help of Mr. Stephen Guerin of Redfish who has guided and encouraged me through thick and thin and Ms. Acacia McCombs who has always been there for me when I needed it. Thank you both.

[1] Gibson, James J. *The Perception of the Visual World.* Boston: Houghton Mifflin, 1950. Print.

[2] Greeno, James G. "Gibson's Affordances." *Psychological Review*. American Psychological Association, 1994. Web. 1 Apr. 2012. <http://ftp.idiap.ch/pub/courses/EE-700/material/31-10-2012/gibsonAffordances.pdf>.

[3] Horn, Berthold K. P, and Brian G. Shunk. "Determining Optical Flow." *Artificial Intelligence*. Artificial Intelligence Laboratory, n.d. Web. 1 Apr. 2013. <http://people.csail.mit.edu/bkph/papers/Optical_Flow_OPT.pdf>.