A SPECTRAL EXPERIENCE: SELF CONVOLUTION AND FACE TRACKING

Federico Camara Halac
NYU
New York, USA
fch226@nyu.edu

ABSTRACT

"Hearing The Self: A Spectral Experience" (aka HALLY) is an interactive, immersive, multimedia, and robotic installation, simulating the process by which the human brain perceives the world. This paper explores the role of both sound and image in the definition of the self this installation brings forth. We briefly explore previous approaches to image sonification, and propose that through video-based convolution new conceptualizations of the self can arise. Further, this expression of the self is neither centered on the human participant nor on the socially constructed notions of the self, but on nonhuman aspects such as the reflection and capture of light, or the technological array of the installation as such. The participant's exploration within this spectrality results in an uncanny and playful experience.

Author Keywords

Interactive, immersive, multimedia, installation, sonification, spatialization, nonhuman self

1. INTRODUCTION

"Spectrality is nonhumans, including the 'nonhuman' aspect of ourselves." Timothy Morton, (p.54) [11]

HALLY[1] consists of a pitch-black room surrounded with speakers, two screens on the front, a camera-hacked mic stand, a sustain pedal, and a delineated square on the floor delimiting the capturing area. The experience begins when the participant enters the room and her face is detected in the capturing area, triggering a random scene.

This installation simultaneously addresses that the mechanical process by which the human brain perceives the world, i.e. the Inverse Fourier Transform (IFT) component of perception, even when perceiving one's own self, is only a part of a complex process shaped by many external stimuli, one of the strongest being societal. However, as we attempt to show, nonhuman agency is essential to the definition of self proposed here.

The remainder of this document is concerned with showing previous experiments on image sonification, the mechanics of the installation, the aesthetic opportunities that were explored, and a brief discussion on the type of questions that emerged from this experience of the self.

2. IMAGE SONIFICATION

In opposition to (data) visualization, the temporal nature of sound favors at least three musical qualities. First, the detection of *patterns* and *trends*, understood as salient information from multidimensional data[2, 13], generally approached from a *parameter-mapping* perspective[8, 12, 21], and ultimately arriving at *raster sonification*[20]. Second, the construction of a sequential *narrative of emotion* (i.e., physical properties, visceral information), that is available through sound [15]. Third, more recent research in spatialization has led to sound installations, with variable degrees of interactivity, which convey spatial data sonification[19].

On one hand, previous work on the conversion of images to sound patterns aimed towards a hearing aid for the sight impaired[10]. Although this is the first portable, real-time system that translates image information, its usefulness depends on its scientific accuracy. On the other hand, composers have tended towards *musical* data sonification, i.e. without the need for conveying useful information[2, 13]. To a certain extent, the mathematical preservation of information can be understood as a mid-term in this conjunction[18].

Thus, in our present installation we are conveying a scientifically useless, but mathematically equivalent sonification of a bi-dimensional array containing a constantly updated, gray-scale image, in order to bring into experience the physicality of spatialized sound.

3. MECHANICS

From a mechanical point of view, it has been suggested that the human brain is a machine that performs an IFT through which it constructs a geometric image from correlations of reflections of light.[3] (Visual) perception can be therefore visualized with an image sensor and a Fast Fourier Transform (FFT) computation of the image data.[6] Furthermore, image data (i.e., a large array of values) can filter a given sound signal of the same length. In Pure Data[16] (Pd), if the block (the DSP array) size matches the length of the image array, a video-based convolution is possible in real time. Thus, by filtering uniform noise with the 2d FFT of an image stream we arrive at a sonification of the mechanics of the human brain's perception within the context of an interactive, real time installation.

3.1 [pix_fft2]

[1]This installation is a collaboration between Lucia Dora Simonelli (ICTP) in Trieste, Italy, Matias Gonzalo Delgadino (Imperial College) in London, UK, and Federico Camara Halac (NYU) in New York, USA. It was premiered in the Xuhui Art Museum, Shanghai, China, during the International Computer Music Conference, October 2017. The source code is available here: `https://fdch.github.io/specexp/`. A short video is available here: `https://vimeo.com/241401699`

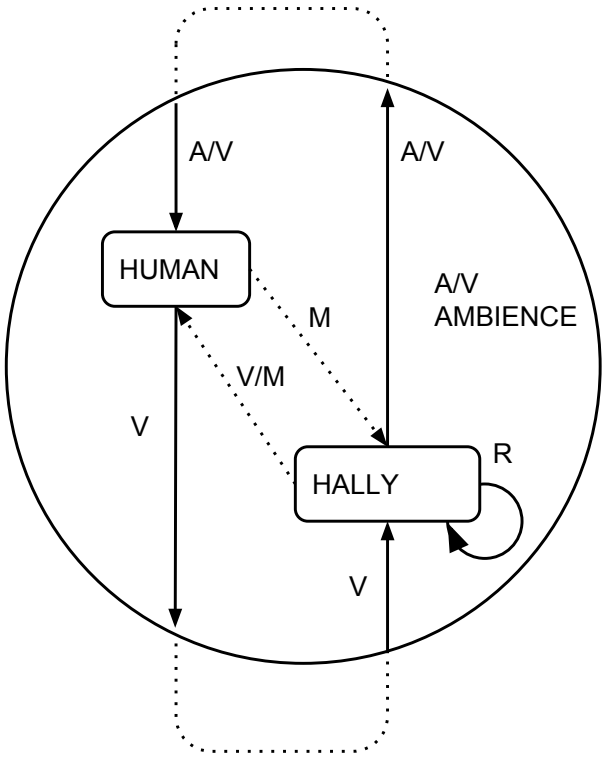

Figure 1: Adapted from Di Scipio's triangular recursive ecosystem connection [17]. A: audio; V: visual; M: motility; R: randomness

The two-dimensional FFT computation of the image array was implemented in a custom external for Gem[4], using the FFTW[2] library. This approach, as opposed to the already existing pix_opencv_dft, was taken namely because of speed -the FFTW algorithms are faster[3] than the native OpenCV Discrete Fourier Transform[4] (DFT) routine- and flexibility regarding future development. Speed in FFTW is dependent upon fixed memory allocation, which became a integral part of the installation in the sense that the computed area of the image was always of the same size (128x128), and remained as such throughout the three days of the performance. Further development of a more embracing Pd object wrapping FFTW would be of interest, particularly in dealing with multi-dimensional FFT computation.

3.2 Face Tracking and repositioning

Antoine Villeret's pix_opencv library[5] includes another library called FaceTracker[7] (from here on FT) which performs face detection and tracking by deformable model fitting (i.e., a face mesh). Given that the size of the image was of 640x480, we could adjust the center position of the face to a box of 128x128 pixels. Thus, by repositioning within the larger image according to the position of the participant within the capturing area, we obtained a responsive surveillance system that constantly looks and follows around the participant's face. Furthermore, the *scenes* that occur throughout the installation are triggered by a sudden face detection. Therefore, when idle, HALLY is computing dark frames looking for shapes to match the FT's mesh. When a mesh is detected, HALLY triggers a random scene.

The weakness of this approach is first the extensive computation that is constantly required to find a face. The second weakness is that, though rarely, dim light conditions and possible occlusions to the face, as well as figures that may resemble a face (e.g. a speaker in the back of the room) make human faces pass undetected. On the one hand, the expensive computations had no effect on the performance of the installation, neither on the Laptop Computer (LC) where FT was called from pix_open_cv (i.e., from Pd), nor on the RPI 3, where the FT demo was running, modified to only send the resulting mesh via TCP/IP through Ethernet to the LC. On the other hand, the problem of the undetected face raised other issues, such as the meaning of the face itself, in relation to the self as such. This last philosophical question we leave to the reader to wonder, much in the same way the participant is left with a similar questioning, regarding the difference of her own self to that of the other.

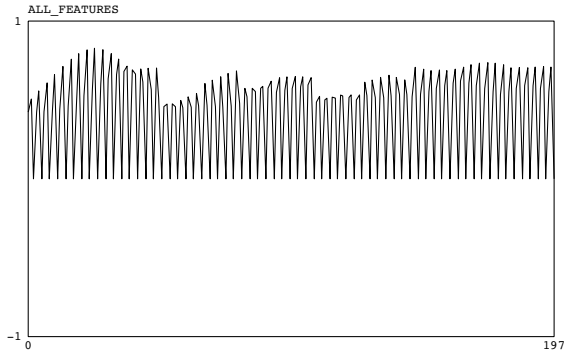

Figure 2: Array holding the 198 points of the face mesh, as determined by the FaceTracker library.

3.3 Mimicking tilt gesture

This is an inherently physical aspect of HALLY: the two image sensors are aligned so that they resemble a face with two eyes. Given that both sensors are screwed to the plastic part of a microphone stand which adjusts the angle of the top microphone's arm (removed for the purposes of this installation), HALLY's face has the ability to tilt +15 or -15 degrees in the z-axis. With the aid of a string attached to a servo motor at the bottom of the microphone stand, this tilt is performed automatically with an Arduino (UNO) that is receiving impulses via USB from the LC and interpreting them as -15 or +15 degrees of tilting. In turn, these impulses are triggered by analyzing the angle of the four points assigned to the nose in the FT's face mesh.

The above results in a mirrored responsive gesture on the part of HALLY: if the participant's nose is tilted to the right, HALLY will tilt to the left, and vice-versa. However, if the participant's face is exactly aligned at 0 degrees, then HALLY will turn randomly left and right in a feedback loop: the tilting rotates the image sensor by +15, so it sends an impulse to adjust -15, and then it has to rotate back to +15 degrees, etc. This feedback loop is ongoing until the participant mimics the gesture as well, therefore exiting the loop.[6] The sonic repercussions were audible, since the en-

[2] `http://fftw.org`
[3] `http://www.fftw.org/speed`
[4] `https://docs.opencv.org/2.4/modules/core/doc/operations_on_arrays.html\#dft`
[5] `https://github.com/avilleret/pix\_opencv`
[6] The macrame string that provided this mechanic motion

tire image captured was rotated (by a small amount), and therefore the pixel values shifted places changing the shape of the convolution. This mimicking gesture is a brief but uncanny quality that HALLY brings forth, not only because it resembles a human gesture, but also because it places the non-human and the human in the same feedback loop, therefore providing both with the same resonance in listening[7] the self.

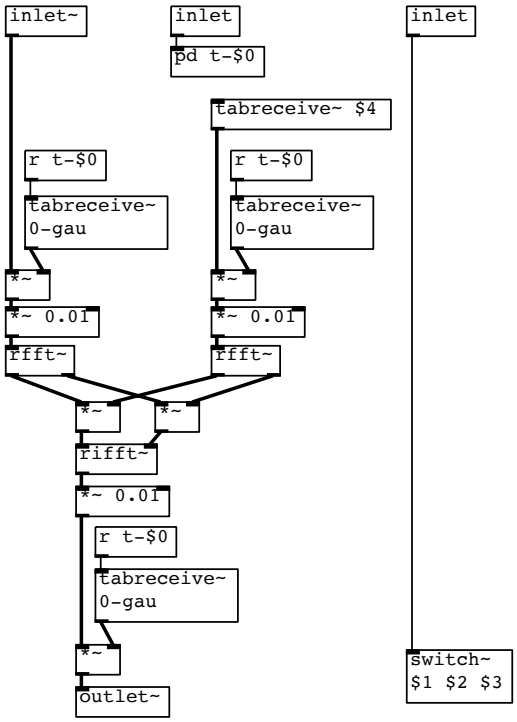

Figure 3: Abstraction dedicated to perform convolution. The first three arguments are 16384 16 0.25 (those of [switch~]: blocksize, overlap, up/down-sampling), and the fourth is the name of the array holding the image.

3.4 Video-based convolution

This technique was used to generate the main sound source for the installation. It consists simply on filtered noise, where noise is Pd's [noise~] (which will be mentioned later) and the filter is the array of pixels where the box with the participant's face is. The convolution takes place in the complex realm, as a multiplication of real an imaginary values of both the noise and the image array (Figure 3). A compromise needed to be made between frame resolution and audio computation: the image data filter dimensions in pixels was set to 128x128, and the audio block size, to 16384 samples, respectively.

On the one hand, the benefits of cropping a squared image were used in two aspects: the 2d FFT computation of the image using [pix_fft2], and the face detection could be adjusted to the square and repositioned within the larger 640x480 image. Both of these aspects where explained above. On the other hand, the 16384 sample block of audio computation results in a fundamental pitch that is constant throughout the experience. This constant pitch was used as a performance opportunity in two ways: the first one is controlled by the participant and the second is controlled by HALLY.

This fundamental pitch is a result of the 2^n space in which memory is more efficiently handled in fast computation loops. This is to say that, in this case, pitch is determined by Pd's frequency of audio computation. Crucial to this pitch, however, is the sample rate at which it is played. Since the [switch~] object allows to control these parameters, they were set so that (1) 16 overlaps (i.e. superimpositions or imbrications of the data) were performed in order to smooth the iterative quality of the rather large block-size; and (2) the convolution patch was down-sampled by a factor of 0.25, thus lowering the pitch by 2 octaves. The total latency of each block was therefore of 92.8798 milliseconds, the resulting size of each overlapping section was of 1024 samples, and the sample rate of the convolution window remained at 11025 Hz (Figure 4).

The resulting pitch is of $1/1024/44100 = 43.0664Hz$, which is about a low F on a grand piano. This provided a more relaxed, but slow computational effort, which was suitable for the low-pitched quality of the drone-sound. Moreover, the 92.8798 milliseconds are suitable for a quiet and slower performance, since it lays within the limits of a middle-range keyboard instrument damping mechanism (as noted by [14] from[1]). Furthermore, in order to counteract the response speed (and since there was enough computational space) we set the speed of the Gem window (including video capture and overall image processing) to 60 frames per second (fps). Therefore, a rapid response was present visually, preparing the participant for the listening experience by providing the ear with extra time to allow expectation to grow.

In this way, the spectrum of this fundamental pitch was *played* by the participant directly. As she moved and interacted, tilted, got closer or farther, blocked or covered her face or HALLY's eyes, and in many other ways, the participant performed a spectral experience with HALLY.

3.5 Spatialized granular synthesis

This is the second way the pitch was used, as controlled by HALLY. Random partials of this spectrum were used to determine the amplitude of granular sounds. These grains were understood as saliences out of the drone-like sound.

A polyphony of maximum 32 grains were synthesized in the following way:

- Step 1 : Random pitch f selected from 32^2 space
- Step 2 : A pixel value $0 > g >= 0.3$ from the image array was selected at $(ftom(f) * k) - rand(1000)$, where $k = 44100/16384 * 64$.
- Step 3 : g determined the amplitude of randomly selected envelopes from a custom selection of tables developed for [5]
- Step 4 : A lookup table was read by [phasor~] at frequency f, where the table was the array containing the face mesh (Figure 2)

was hardly ever properly adjusted, so it often happened that it detached. Notwithstanding, this un-working of the installation was left as such, since we considered it a very small but delicate attention on our side to aid HALLY, and it constituted a break into the reality of the technologies involved.

[7]This installation's first part of the name "Hearing the Self..." originally came from the theme of the ICMC where it was performed. The hearing of the self remains as such only if you pass by the installation space, hearing a low drone or a very active and loud sound space; if you go through the curtains, though -only when entering- , the listening of the self takes place.

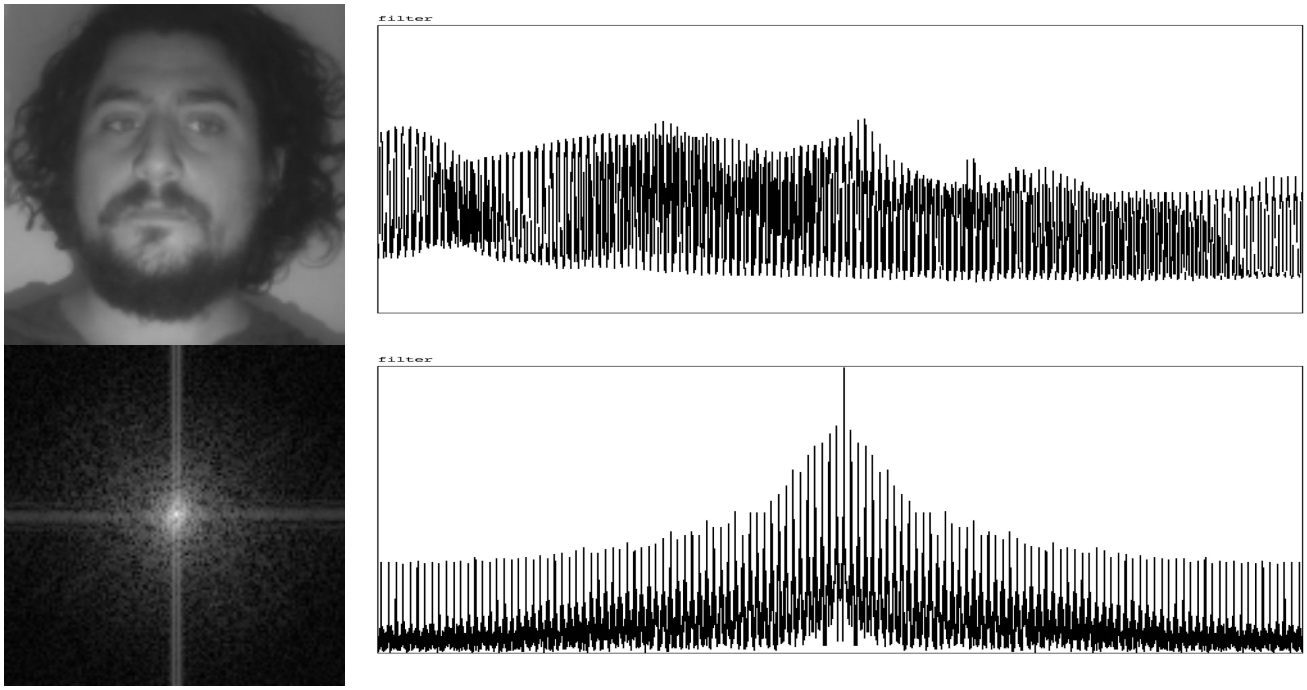

Figure 4: Left-Top: Cropped box with face. Center-Top: face array as real valued filter. Right-Top: filtered noise with real values. Left-Bottom: 2d FFT of the above cropped box with face. Center-Bottom: face array as complex valued filter. Right-Bottom: filtered noise with complex values.

- Step 5 : Finally, a random speaker assignment determined the localization, and randomized ranges from 200-5200 milliseconds determined the length of the grains.

3.6 Did someone say Noise?

Given that noise is the main source of *otherness* that we are taking for granted in this installation, it begs the question, at this point, to wonder about the current state of noise within Pure Data. Pd's [noise~] and [random] objects contain somewhat different pseudo-random number generators (PRNG), for audio and control rate respectively:[8]

3.6.1 [noise~]

The first term of the algorithm in [noise~] computes the bitwise difference between *val* and the Mersenne Prime 2^{31}, then subtracts 2^{30} and divides it from the result. The result is outputted at sample rate, i.e., on every iteration of the audio loop. The second term simply multiplies the previous result by $2^{32}+1$, adding to it a *magic* number 382842987, finally storing the value in memory for the next iteration.

```
[...]
static int init = 307;
val = (init *= 1319);
[...]
((val & 0x7fffffff) - 0x40000000) / 0x40000000;
val = val * 435898247 + 382842987;
[...]
```

3.6.2 [random]

In the case of [random], the use of double precision instead of float precision above entails more accuracy, at the possible cost of speed. Further, like in the previous case, a seed is assigned on each instance, but here there is a seed method that enables the input of a seed (aka *randval*, aka $x->x\_state$). We will discuss this further down. Moreover, the *range* variable adds a hard upper boundary that cuts the output at $range-1$. The two first terms resemble [noise~]'s algorithm, without the bitwise computation, with all numbers being *magical* except the second term, where the seed and the range are divided by 2^{30}.

```
x->x_state = randval * 472940017 + 832416023;
nval = range * randval / 4294967296;
if (nval >= range) nval = range-1;
```

These methods can be understood as improvements on the native C function called with rand(), since it uses the Mersenne Prime, proven to be a more reliable source of randomness [9]. However, neither the nature of the magic number, nor the reliability of the PRNG have been determined.[9] The following code was simplified for readability purposes:

The flexibility of this type of PRNG is its ability to use a seed, which guarantees the a unique sequence of values for each seed. The point here is that the possibility for repetition, of two random sequences speaks of a different attitude towards art. In the audio version, as we just saw, the seed is fixed for every instance, with the comment in line 469: *"seed each instance differently. Once in a blue moon two threads could grab the same seed value. We can live with that."*. In fact, this means that two different instances of [noise~] created simultaneously will render the same randomness.

[8]The code snippets presented here are from `https://github.com/pure-data/pure-data/blob/master/src/d_osc.c` and `https://github.com/pure-data/pure-data/blob/master/src/x_misc.c`. Both are simplified for readability, to the point that they can be understood as pseudo code.

[9]The authors of this paper would like to thank, not criticize, the Pd community for this minor imprecision in the statistical processes involved. It is in these still unexplained or accidental properties where there is space for creativity and further exploration, and more importantly, for community to emerge.

In the control rate version, the overuse of magic numbers comes with the comment in line 44: *"this is strictly home-brew and untested"*. These lines are curiously at odds with the shift to double precision, while simultaneously in resonance with the more relaxed approach towards an artistic generation of random data, rather than strictly statistical, let alone cryptographic uses of randomness.

3.7 Discussion

There is still more to be said about the visual aspect of the installation (such as the automated image web search for the keyword *face* each time the installation starts, which was truncated by governmental regulations in Shanghai, China), and about other kinds of interaction by a piano sustain pedal present just below the microphone stand. Similarly, we have only touched key concepts throughout this paper (sonification, immersion, robotics, tracking, spatialization, noise, feedback, etc). We will have to limit ourselves on these topics, and focus on the task at hand.

Now it is time to address the second half of the title of the installation (and the first part of this paper's): the "spectral experience" part. As mentioned in the introduction, our intention with this installation focused on sonifying the mechanics of brain's IFT-based perception. The constant use of FFT-s throughout the filtering or convolution seems to address this focus, given that perception according to [3] happens in the complex realm, i.e., the spectrum. However, more recent inquiries into the spectral, as understood by Timothy Morton in [11]:

" ... spectrality is the flavor of the symbiotic real, where everything is what it is, yet nothing coincides exactly with itself." (P.54). The uncanniness of this flavor is what resounds with HALLY. What we termed A/V Ambience in Figure 1, can be an instance of Morton's *"symbiotic real"*.

Following Morton's notions of interconnectedness, a new diagram can be drawn in simplification of the previous one (Figure 5). In this last one, agency is more evident in terms of the ongoing loops that are neither aimed towards the center nor the edges, but to the condition of connection. This inevitable link between environment, human and nonhuman is understood by Morton as spectrality itself, which for him *" [is] not spirits in the divine realm, even if that realm has been relocated in the human -that's the concept of Humanity. Spectrality is nonhumans, including the 'nonhuman' aspect of ourselves."*

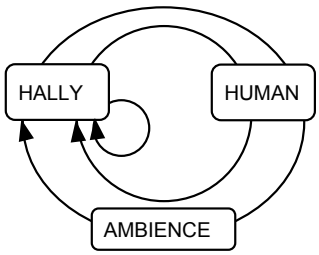

Figure 5: Simplification of Figure 1

4. CONCLUSIONS

We have proposed an interactive, immersive, multimedia, and robotic installation, discussing the role of sound and image in the definition of the self. Sonification techniques were proposed, amongst which video-based convolution was preferred. We have aimed at non-anthropocentric approaches and introduced terms such as ecosystemic or spectrality, that aim to bring the nonhuman agency in play. Further, we have tried to bring this agency through as expression of the nonhuman self, in the anthropomorphizing HALLY and her motility and sight. The participant's exploration within this spectrality results in an uncanny and playful experience, in which she can further question these issues, and bring new aspects into play.

5. REFERENCES

[1] A. Askenfelt and E. V. Jansson. From touch to string vibrations. i: Timing in the grand piano action. 88:52–63, 07 1990.

[2] O. Ben-Tal and J. Berger. Creative aspects of sonification. *Leonardo*, 37(3):229–232, 2005.

[3] A. Connes. The music of shapes. September 2012.

[4] M. Danks. Real-time image and video processing in gem. *ICMC Proceedings*, 1997.

[5] F. C. Halac. fdlib. 2016.

[6] F. C. Halac. pix_fft2. 2017.

[7] S. L. J. M. Saragih and J. F. Cohn. Face alignment through subspace constrained mean-shifts. *International Conference of Computer Vision (ICCV)*, September 2009.

[8] G. Kramer. Some organizing principles for representing data with sound. *SFI Studies in the Sciences of Complexity, Proceedings*, XVIII:197–202, 1994.

[9] M. Matsumoto and T. Nishimura. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, 8(1):3–30, Jan. 1998.

[10] P. B. L. Meijer. An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering*, 39(2):112–121, February 1992.

[11] T. Morton. *Humankind: Solidarity with Non-Human People*. Verso, London, 2017.

[12] B. C. M. D. G. S. Oded Ben-Tal, Jonathan Berger and P. Cook. Sonart : The sonification application research toolbox. *Proceedings of the 2002 International Conference on Auditory Display*, July 2002.

[13] M. D. ODED BEN-TAL and J. BERGER. De natura sonoris: Sonification of complex data. *CCRMA*, 2004.

[14] J. Oliver and M. Jenkins. The silent drum controller: A new percussive gestural interface. *ICMC Proceedings*, 2008.

[15] A. Polli. Atmospherics/weather works: A spatialized meteorological data sonification project. *Leonardo*, 38(1):31–36, 2005.

[16] M. Puckette. Pure data another integrated computer music environment. *ICMC Proceedings*, 1996.

[17] A. D. Scipio. âĂŸsound is the interfaceâĂŹ: from interactive to ecosystemic signal processing. *Organized Sound*, 8(3):269–277, 2003.

[18] J.-B. Thiebaut, J. Bello, and D. Schwarz. How musical are images? from sound representation to image sonification: An eco systemic approach. 08 2007.

[19] G. Weinberg and T. Thatcher. Interctive sonification: Aesthetics, functionality and performance. *Leonardo Music Journal*, 16:9–12, 2006.

[20] J. Yeo, Woon Seung; Berger. Raster scanning: A new approach to image sonification, sound visualization, sound analysis and synthesis. *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06)*,

September 2006.

[21] W. S. Yeo and J. Berger. A framework for designing image sonification methods. *Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display*, 2005.