Exploration of Sonic Interaction Design in Virtual Space

Thesis

Submitted in Partial Fulfillment of
the Requirements for
the Degree of

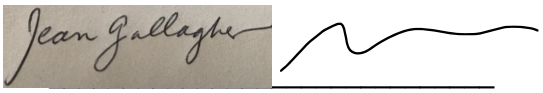
MASTER OF SCIENCE (INTEGRATED DIGITAL MEDIA)
at the
NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Ben Crystal

May 2022

Approved:



Department Chair Signature

May 12, 2022

Date

University ID#: N10562852

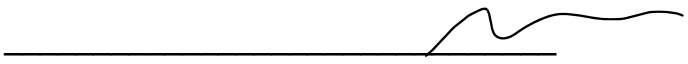
NetID#: bc3099

Approved by the Guidance Committee

Integrated Digital Media

Dan Taeyoung, Thesis Instructor
dan.taeyoung@nyu.edu
Integrated Digital Media
NYU Tandon School of Engineering

May / 12 / 2022



R. Luke DuBois, Faculty Advisor
dubois@nyu.edu
Integrated Digital Media
NYU Tandon School of Engineering

May / 12 / 2022

Vita

Ben Crystal was born in Summit, New Jersey on March 31, 1997. They attended Bayberry Elementary, Valley View Middle, and Watchung Hills Regional Highschool for their early education. They were admitted to the University of Vermont, where they studied electrical engineering with a focus in digital signal processing, and received their bachelors in science in May, 2019. Then, they attended New York University's Integrated Digital Media program, where they focused on interaction design in virtual reality and music production and combined the two to create this thesis, which led to receiving a Masters in Engineering in May, 2022.

Abstract

Immersion is one of the most important factors in creating a somatic, corporeal experience. An intuitive connection between the physical and virtual self leads to a more intuitive program or product, while a novel experience that is not intuitive or immediately gratifying can spoil immersion. This process is deeply integrated in how they co-produce learning and reasoning for users of any program or product with audible feedback or features. In immersive experiences, the rate of improving sonic affordances often have taken the back seat to visuals. Currently, sonic expression in the commercial virtual reality (VR) landscape is very limited. Most audio manipulation is predetermined, heavily quantized in time, pitch, and timbre, and limits users' ability to have a truly corporeal experience in the virtual space.

I propose an exploration of the embodiment of sound through human-computer interaction in a virtual space to create an intuitive and deliberate sonic environment and experience. This thesis is an attempt to bridge the gap between thought (mind) and action (body) in a virtual sonic landscape through an exploration of interaction design. After researching, I created several interaction schemes for a collaborative music creating experience and evaluated how much of the sonic landscape players experience through a variety of quantitative and qualitative metrics.

Table of Contents

1 Introduction

1.1 Motivation

My dream as a designer and engineer is to create methods for people to express themselves in ways they never have before. As someone who grew up immensely shy, I was afraid of taking up space and having my voice heard both physically and metaphorically. Through the creation of music, I have learned to not only have the ability to communicate more freely and clearly with others, but I have also learned how to share that experience, collaborate, and inspire other aspects of my and my peers' lives. Immersion is one of the most important psychological factors for learning, creating, and enjoying an experience. With immersion comes an intense level of focus, potentially leading to a full state of engagement where extraneous distractions cease to exist. I hope through my research and prototyping process to develop a collaborative immersive audio experience that will engross and assist in bonding users of all levels of VR and music creating expertise.

This thesis plan addresses this goal through both a qualitative and quantitative analysis of users' experiences in a series of sonic landscapes controlled entirely by their interactions. If the level of immersion is substantial, the techniques implemented in this integration could be used to further explore tools for collaborative music production, educational, therapeutic, and commercial entertainment tools and experiences in extended reality.

The VR aspect as a whole is a placeholder for the affordances that AR will offer that are not available yet. I want to enhance reality, rather than to hide from or escape it. As mentioned within the discussions about gesture and body language in storytelling, facial expressions, eye contact, and the overall "vibe" of a group during a performance is what really makes it enjoyable to performers and audience alike. In a collaborative musical tool, relating to the people being collaborated with is key. So many people are hindered from their ability to create by both physical and mental limitations, and to give people that power of control, to create, and to connect at any level with a skill floor higher than the ceiling makes my heart flutter and gives me meaning. I want people to be able to express themselves and connect like never before. The goal is to foster connection, self exploration, and creativity in my users.

The final prototype of this thesis is an immersive vocal performance tool that allows vocalists to express aspects of self through a variety of audio effects controlled by their body language. It addresses the question, "how can we control an audio performance in a way that is meaningful and easily understood by both performer and audience?"

1.2 Prior Art

The following sections will discuss some of the software and hardware that have been used to digitally control a virtual soundscape.

1.2.1 Brief History of Digital Sonic Interaction Design

There are a plethora of reasons as to why designers would want to enable their users the affordances of controlling a soundscape. On the more expected extreme, we have musicians and music producers who craft sonic experiences for their audiences, who would benefit from allowing their users to "season their listening experience to taste". A recent example of this experience is Kanye West's *Donda STEM Player* (seen in Figure 1 below), which is a tactile device that allows its users to mix (deliberately combine multiple recording tracks into a cohesive sonic product) their listening experience themselves in real time. On the other extreme, we have music therapy and the physical and mental health benefits that come with processes like meditating to music. In the middle lies educational experiences, art exhibits, games, and other forms of entertainment.

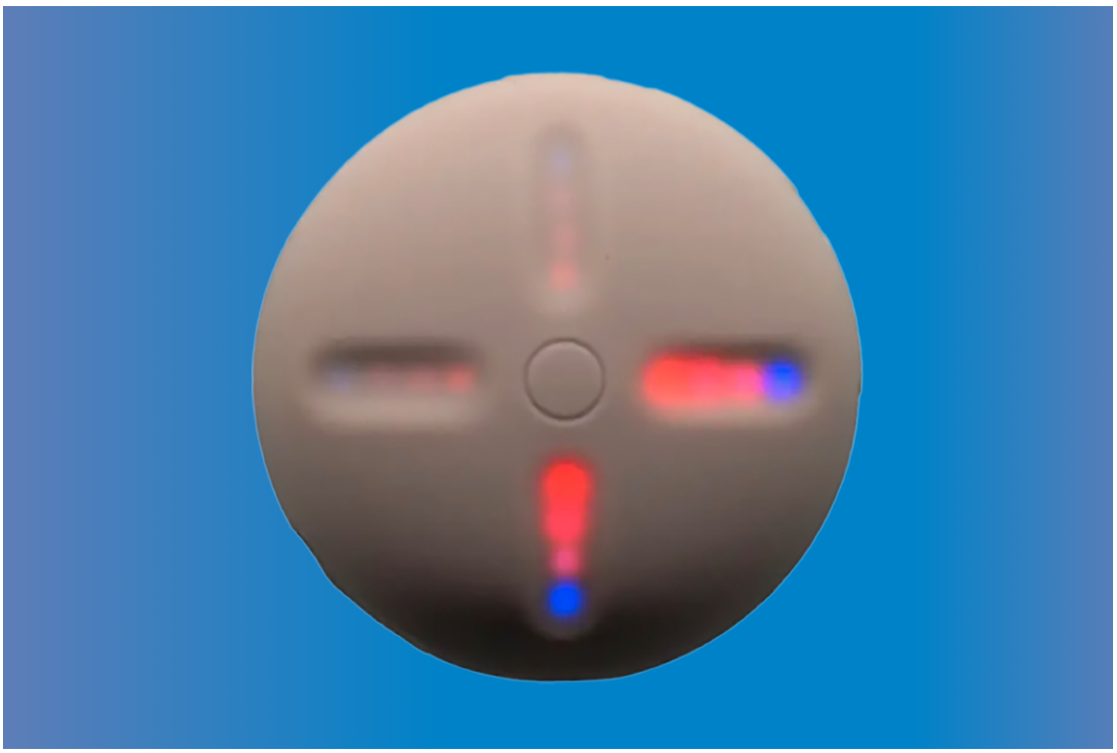


Figure 1: Donda STEM Player

To understand the landscape of synthesizers and digital instruments in the commercial landscape, we must look back at the great schism of synthesizer design in the 1960s in America between the East Coast designers (e.g. Robert Moog) and the West Coast designers (e.g. Donald Buchla). It becomes clear that the East Coast designers were building their instruments for a familiar western practice, with UI centered around the standard piano's keyboard layout, where note temperament scales proportionately to that which is most prevalent in Western music. Contrarily, the West Coast designers wanted to approach audio synthesis from a countercultural, alternative perspective, in an attempt to remove preconceived biases towards western musical structure. Not only was this designated in terms of the parts of the interface that triggered notes (if there were haptic methods of triggering notes in the first place), but it was also embedded in the terminology and routing methods that these designers put forth. Moog's systems approached audio synthesis more with terminology referring to musical structure and notation, while Buchla approached signal synthesis from more of an engineering perspective. Figures 2 and 3 display prominent synthesizers developed by Buchla and Moog, respectively.

Figure 2: Buchla 200e Synthesizer

Figure 3: Moog System 55

For over 2 decades, Buchla and Moog, alongside other analog synth designers and their companies like ARP, Korg, Roland, Obberheim, expanded upon the affordances of analog synthesizers by creating unique ways of molding sounds, expanding upon the amount of polyphony available within singular modules, all while attempting to make the devices more accessible to the public. Some of these companies, like ARP, offered their synthesizers as DIY kits, where consumers would purchase components with pre-cut PCBs that they could assemble themselves for a fraction of the price of their competitors. The lower entry fee brought these types of devices to the DIY scene, as well making these devices less of a stretch to be considered in school systems.

In 1983, Yamaha introduced the world to the DX7 synthesizer (seen in Figure 4)– the first successful full-sized digital synthesizer to reach commercial markets, and one of the most iconic sound providers ever created. The integrated circuit chips in the DX7 were mass produced and did not need to be as scrupulously matched as say transistors in analog hardware, it cost a significant amount less for a ready-out-the-box piano than even the DIY kits of the past, which led many of the aforementioned synth companies to succumb to either fulfilling niches in the market or to go out of business entirely. This synthesizer became the sound of the 80s, notably used by artists from Brian Eno to Whitney Houston to Billy Ocean.

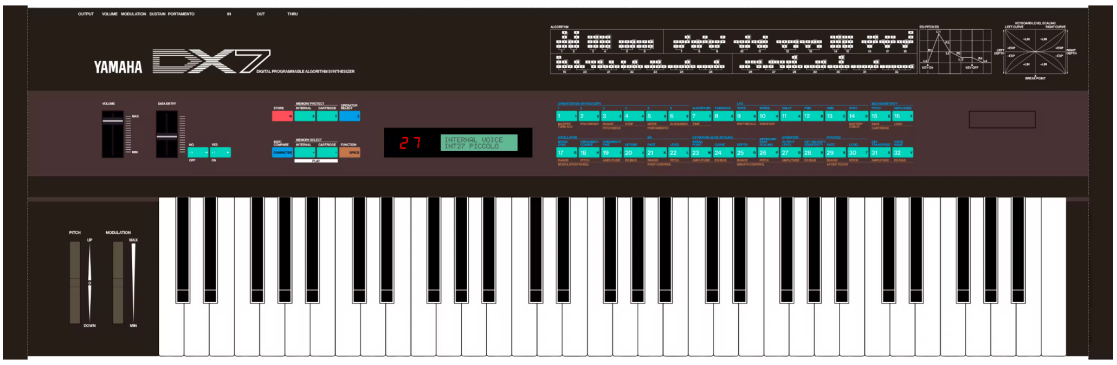


Figure 4: Yamaha DX7

Naturally, covering the entire history of electronic and sampled music is an impossibly large task, but another tangential, crucial evolution in the digital musical landscape was the rediscovery of the Roland TR-808, seen below in Figure 5. This drum sequencer, released in 1980, was originally designed to be a sort of automated backing track,

almost a more robust metronome with simplistic, programmable drums, for the at-home or practicing musician. In the following years, many music producers found that the "Kick" drum synthesizer could be tweaked to create deep sub-bass frequencies that hit the hearts of the audience and produced a range of frequencies that was rare, if not impossible, to produce externally, especially in such a straightforward manner. As such, the "808" became a staple in the early days of hip-hop and electronic music, and is often still used and sampled today.

Figure 5: Roland TR-808

The final "standalone" evolution that will be discussed in this section is the introduction in 1981 of the Akai MPC2000, seen in Figure 6, which was the first device that allows producers to sample audio and easily play it back at the push of a button. Other sampling devices of the era, such as the Mellotron, had keys that would trigger tape reals or other analog audio sources to play on command. However, the "MPC"'s digital setup allowed audio to be sampled, swapped out, and played more easily, all in a device that could be transported in a backpack for a fraction of the cost.

Figure 6: Akai MPC2000 Digital Sampler

These devices were some of the most fundamental and foundational tools to the digital and virtual music production industry that is in place today.

1.2.2 MIDI Controllers

Unlike hardware synthesizers, most digital audio software accepts MIDI (Musical Instrument Digital Interface) as input. MIDI interfaces do not actually produce sound themselves, but rather transmit instructions in the form of up to 3 bytes of data at a time to a virtual software to decipher and synthesize from. The first byte represents the channel, or type of message that the other byte or two will be understood as. Some channels inherently

represent musical concepts, like pitch or velocity, while others are simply general channels called CCs (Control Change), which can be easily mapped to either custom effects or whatever a performer wants. The physical interfaces can take a wide variety of shapes and forms. The most common types of standard MIDI interfaces simply contain interactions like those in the controller in Figure 7 below. Here, you can see velocity sensitive keys, dials that can be mapped to parameters, velocity sensitive drum pads, and 2 modulation wheels that are generally mapped to pitch bending or tamral changes.

Figure 7: Alesis VI25 MIDI Controller

People make MIDI controllers that look pretty far from a piano though. I have converted my computer mouse, an accelerometer paired to an arduino, and both a Nintendo Gamecube and Nintendo Switch to MIDI controllers that trigger melodies and audio effects in VSTs (Virtual Studio Technologies). There are also companies that create MIDI controllers out of, or that look like, guitars through very interesting mediums (e.g. the Yamaha GI-10 and the Artiphon Instrument 1). However, what I find fascinating are interfaces that allow for tambral changes even after a note has been triggered.

When a MIDI controller has polyphonic aftertouch, it is capable of using the amount of pressure applied to each individual key or pad after they've initially been triggered to

control parameters of the soundscape, even if multiple have been played to produce a chord. This allows breath to be surgically added and brings life to a digital instrument. One prominent MIDI interface with polyphonic aftertouch can be seen in Figure 8.

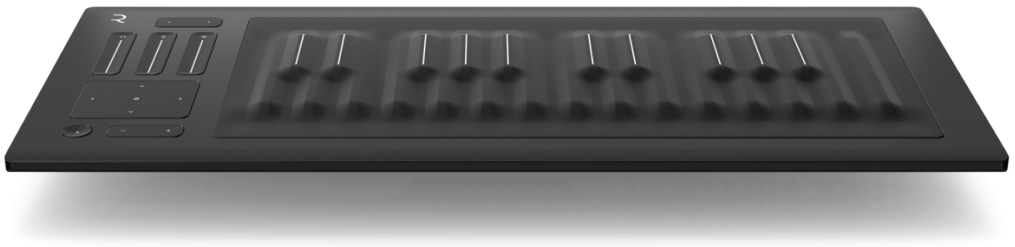


Figure 8: Roli Seaboard Rise

1.2.3 Proprietary Virtual Sonic Interfaces

I will briefly discuss three proprietary virtual sonic interfaces, along with their affordances and limitations that I seek to overcome.

First is the virtual reality game called Electronauts, seen in Figure 9 below. This is a multiplayer experience where you and up to 1 other collaborator manipulate a soundscape based on a combination of DJ turntables and MIDI controller-esque interactions (e.g. pressing buttons, pushing sliders, and turning dials). Users can improvise over pre-playing background tracks on a predetermined scale and toggle a handful of instruments to play said melody. Every interaction is heavily quantized to the right key and beat, so it's virtually impossible to make a "mistake". The user(s) are confined to their DJ booths, so there is no freedom to explore or get close enough to other players to really facilitate interaction. As such, there is very finite use of the affordances of 3 dimensional space; the only use of depth in Electronauts are the sound grenades which trigger audio effects indistinguishably regardless of speed, distance, direction, obstruction, etc. which are essentially a visual illusion, but still provide some sense of corporeal control. Although I want the space I

develop to implicitly bias users as minimally as possible in their interactions, I also want it to be purposefully designed and explored.

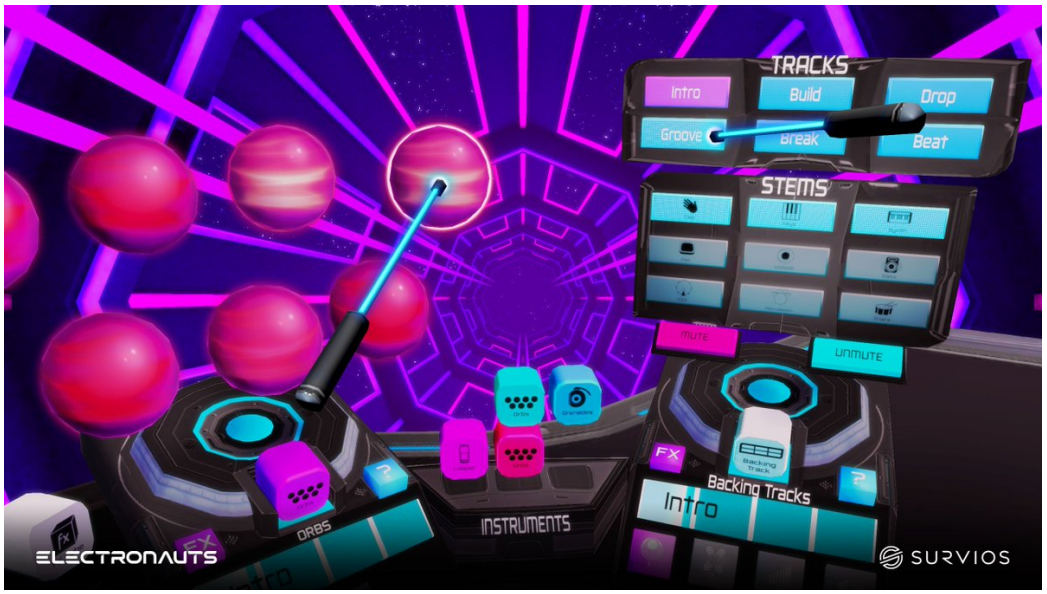


Figure 9: The Virtual Reality Game Electronauts

Second is the virtual experience called SynthVR, displayed in Figure 10. This program allows users to explore and manipulate modular synthesizer modules that can be patched together like the physical parallel. Users can create their own modular synthesizer patches for a fraction of the price without taking up any physical space. However, the software is essentially a 3 dimensional translation of 2 dimensional interfaces that doesn't take advantage of many of the affordances that virtual space allows. Although the interactions themselves didn't take much advantage of the 3D space, it offered full spatial audio with the speaker placement, as well as "anti-gravity patch cords", which does make the mess of cables innate to modular synthesis a little easier to manage. Additionally, there is no way to collaborate with other players (although this could be added quite easily), no form of intuitive interaction (extensive knowledge of audio synthesis is required to fully understand how to maneuver the space), there are no external ways to control the sound (objects in the environment are the only controllers, not external MIDI nor the player(s) themselves), and it is computationally heavy to the point where users need to be attached to a computer to run it.

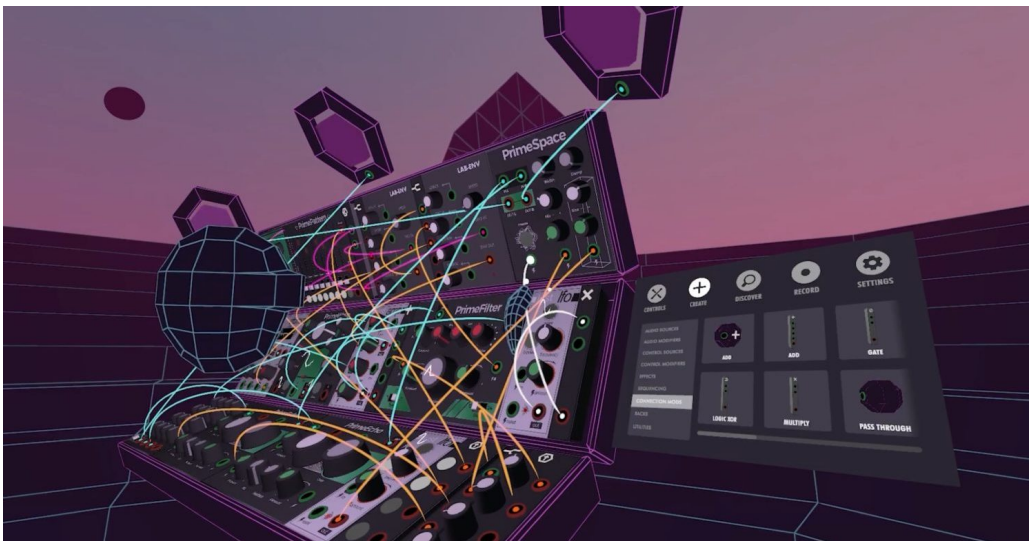


Figure 10: Modular Synthesizer in SynthVR

Finally, we will look at the virtual interface called Modulia Studio, seen in Figure 11. Modulia Studio is a dedicated interface for the digital audio workstation (DAW) called Ableton Live, which allows for some freedom of design, but is literally a 2 dimensional interface in 3D and currently utilizes almost no affordances of the 3 dimensional space. It is essentially a system where the only immediately perceivable benefits over using a mouse and keyboard are the lack of distractions from being in the headset immersed in the space and the fact that users have access to both controllers at the same time, but for most able-bodied users, I personally feel that it is not worth losing hotkey shortcuts for.

Figure 11: Ableton Live Virtual Mapping through Modulia Studio

I've included a link in the references to www.*historyofsynths.com*, a website that portrays a series of infographics and audiographics of many crucial steps in the design evolution of synthesizer interactions and user interfaces.

1.3 Overview of the Thesis

Chapter 2, *Literature Review*, contains my personal technical and musical takeaways from surrounding literature in the field that will be referenced in throughout the rest of this thesis, as well as throughout the design process of my project overall.

Chapter 3/4, *Planning/ Methods*, walks the reader through my thought processes of the initial prototypes, the findings, the failures, etc.

Chapter 5, *"This is the Project"*, depicts the actual embodiment of how the program interacts and describes the users' experience as they go through with using the product.

Chapter 6, *Conclusion and Future Work*, discusses my reflections, testing and results, and personal evaluation of my research. The success of whether or not the interface lives up to the requirements/ objectives set forth will be addressed, as well as a description and analysis of tests proving or denying such are completed. I will discuss shortcomings of this project, what could be done to solve these issues, and what may be potential next steps in furthering the implementation of virtual sonic embodiment in live musical performances.

Chapter 7, *Technical Documentation*, shows documentation of the Max MSP patches, Unity files, etc. that were created to sculpt the experience designed throughout this thesis.

Chapter 8, *References and Appendices*, shows where I got all of this wonderful information that was not gathered from my own experiences.

2 Literature Review

The following sections will discuss some of the key readings that have shaped my approach in designing my thesis by improving my general understanding of the fields of virtual interaction design and sonic interactions.

2.1 How Bodies Matter: Five Themes for Interaction Design by Klemmer, Hartmann, and Takayama

Hartmann and Takayama use this piece to discuss five key components to consider when designing an embodied experiment to make the experience as empirical as possible. The first two of these steps (*Thinking through Doing* and *Performance*) are categorized under the idea of "Corporeality", or just relating to having a physical presence to embody, while the latter three (*Visibility*, *Risk*, and *Thick Practice*) are external, and viewed under the category of "Social Affordances".

Thinking through Doing basically discusses how thought (mind) and action (body) are deeply integrated, and how they co-produce learning and reasoning. The authors note that "physical interaction in the world facilitates cognitive development" (Klemmer, Harmann, &

Takayama, 2006, p. 2) in infants, and continues to be valuable throughout life. Learning through doing, rather than passively absorbing, has been proven to be a successful way to increase understanding of systems and to be more mentally engaged with an experience, which has been tested extensively with formidable results through applications like the Montessori school systems. Many of the tools developed for educational systems like these map intangible processes to physical interfaces, which provide "natural" mappings that are more familiar to users. It's a parallel take to skeuomorphic design, whereas aspects of previously successful interfaces are incorporated in newer products to make users more comfortable learning something entirely new. However in this case, the new design leans on the predecessor for familiarity is not necessarily a redesign. They also discuss the value of gesture when communicating to both plan speech production and also to help convey messages that are difficult to verbalize.

Performance is essentially about how the majority of tools that humans use do not take advantage of the vast potential of people's fine motor controls and senses. Hartmann and Takayama note that the designer's goal should be to create systems where "the intimate incorporation of an artifact into bodily practice to the point where people perceive that artifact as an extension of themselves; they act through it rather than on it" (Klemmer, Harmann, & Takayama, 2006, p. 4). Physical actions can be both faster and more nuanced than symbolic recognition.

Visibility refers to the role of artifacts in collaboration and cooperation. As mentioned later in *The Craftsman* by Richard Sennett, the process of learning a craft historically was done in a workshop space, where experts directly engage with novices and get them to participate and interact in a community of practice. "The studio model of education employs work practice transparency as a pedagogical technique, affording peer learning, discussion, and "constant critique of work in progress" (Klemmer, Harmann, & Takayama, 2006, p. 5). Visibility in a craftspace enhances coordination, and an openly communicative space not only fosters knowledge transfer, but also facilitates exploration and discovery with new perspectives.

Risk essentially states that the nature of a corporeal existence means that one's experience living is shaped by knowledge of their vulnerability. The concept of risk is the sole reason why decisions are made and why some options are valuable over others. In a digital or virtual space, most of the risks of the physical world (e.g. simulator games or training facilities, the ability to undo/ copy/ paste/ edit countless times) are negated. Surely we can take advantage of the affordances that low risk allows us in a learning or creative environment, but it is important that uncertainty creates deliberation and engagement as well. It will be important to find that balance, especially in a social setting. The authors refer to the lack of risk in a distanced, virtual collaboration as having the fault for lack of opportunities to build trust. On one hand, the lack of social context clues can lead an individual towards a hateful place, but on the other hand, the lack of risks in non-face-to-face interactions can help other individuals feel at ease to be themselves without judgement and can create online communities.

Thick Practice is based on the idea that the designing of any new system both offers new affordances based on the preconceived notions and systems of the past (similar to the skeuomorphic-adjacent conversation earlier). On the contrary, it also eliminates some of the previously available functionality. The development of a system that is immersive enough to *be* part of the real world rather than trying to simulate it, or rather the pursuit of digital verisimilitude (the appearance of being true or real), is more difficult than it might seem, but chasing embodiment interaction can be a more direct and prudent path. Only through embodiment can we maintain the nuance of manipulation and interaction design of the real world equivalent.

2.2 The Sonification Handbook by Hermann, Hunt, and Neuhoff

In *The Sonification Handbook*, Hermann, Hunt, and Neuhoff explore and explain ideas behind how non-verbal sounds can be used as a means for communication.

The authors note that sound designers who create the user interfaces of auditory displays need to create meaningful, powerful, flexible, and "aesthetic" interactions to be as effective as possible. Though they mostly discuss the sound design choices for non-musical

devices, I think the methodology applied to such is fundamentally even more applicable. They refer to a "parametric" model "as one that has a (relatively) few variable parameters that can be manipulated to change the interaction, sound, and perception. A highly parametric model of sound is the technique known as Linear Predictive Coding (LPC, representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model), which uses just a few numbers representing the spectral shape, and the (usually voice) source, to represent thousands of PCM [(pulse code modulated)] samples." (Hermann, Hunt, & Neuhoff, 2011, p. 198) Taking this into account, for the nonverbal aspect of the design I am creating at hand, I recognize that it would be beneficial to create highly parametric controls for the experience where users can gesturally change moods organically rather than solely tweaking parameters linearly. After all, altering single dials or aspects of a controller at a time to linearly alter individual parameters would potentially not be taking advantage of the affordances that a 3 dimensional space has to offer.

In a potential iteration of my project, I would like to refer back to this book if I want to get into the depths of sound design, as it discusses in detail sound synthesis for audio displays, as well as visual representations of audio signals. Although I do not necessarily want to facilitate an "educational experience" from the perspective of portraying the calculations behind signal processing, providing visual representations of the sounds and sonic landscape can help link the visible to the audible in order to help users understand how their movements and actions in a space alter the soundscape. This is similar to that of an oscilloscope (as can be seen in Figure 12), where many engineers, sound designers, and visual artists alike create visual, graphical, real-time representations of the electrical signals that their systems are moderating.

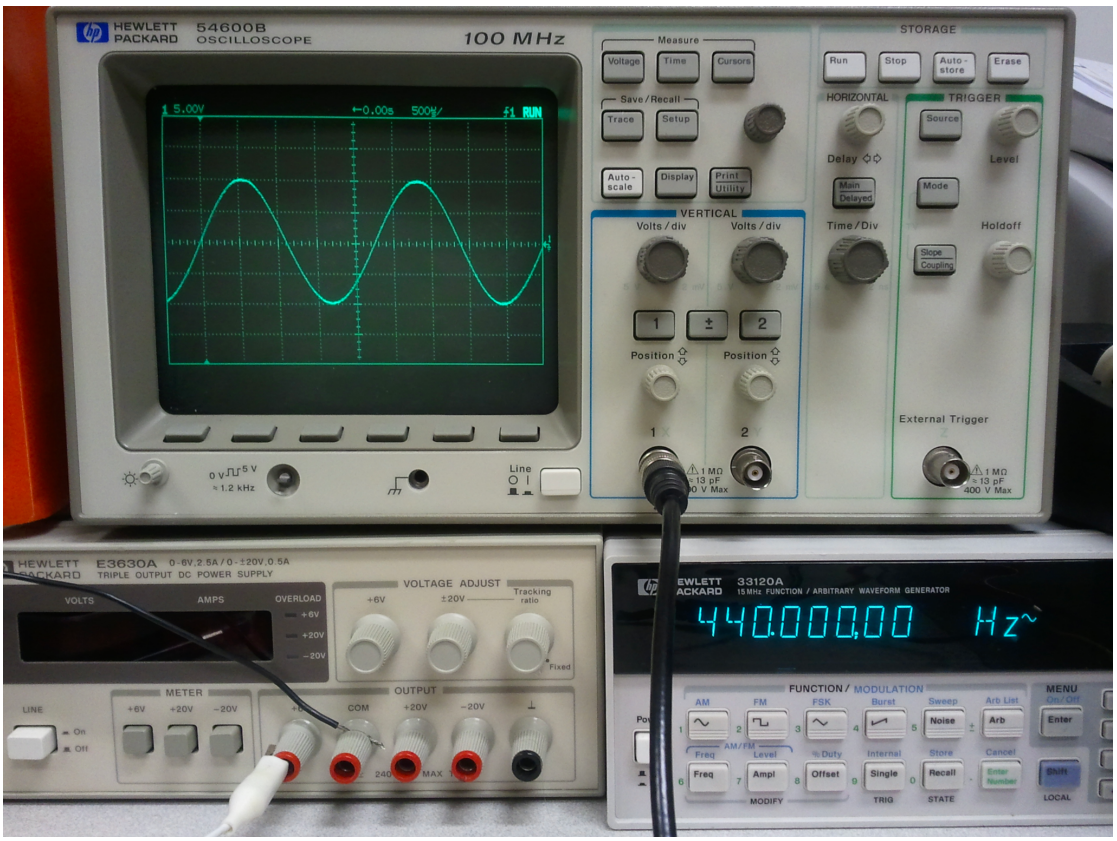


Figure 12: Oscilloscope displaying a 440 Hz Sine Wave

2.3 Soma Literate Design by Stephen Jon Neely

Soma Literate Design is a discussion of the inseparable integration of soma (the body) with techna (techniques and technology), particularly in interaction design. In relation to the limitations of spoken language to communicate refined meaning, Lee et. al. "believe improved somatic empathy (through heightened body consciousness) could improve our ideation not only in movement based interaction but in any interaction that deeply engages our body." (Neely, 2019, p. 141)

Through chapter 3 of this book, Neely gives examples of hypothetical situations and some of his research experiments that tested such. He discusses eurythmics and other music-adjacent ideologies, but one that fascinates me is his approach to his "Haptic Enviro-Sensing Metronome". This project was proposed to help those in need in a variety of

situations understand through haptic vibrations have a rich experience and understanding of the situation they're in. Interestingly enough, I proposed a design for a haptic attachment to the classic white cane that gave vibrational feedback where the rate was proportional to the height of overhead objects, kind of similar to that of some of his approaches. I have personally always found syncing "hot/ cold" to rate, from systems like these to metal detectors to range detection sensors warning reversing cars of obstacles, to be a very rich type of interaction.

He had described a "rich experience" as the equation:

(body + time) * (cohesive) = (rich)

Without having a body experience something for an extension of time, there is no experience. Without the interactions within an experience being embodied gestalt gestures, or without the interactions being symbolic representations to the user of the type of information transmission they are accomplishing, the interaction will not be meaningful. For example, we could consider how intuitive pinching for zooming feels for touch screen smart device users. This is because it is representative of taking whatever is available on the screen in between the users fingers and scaling it larger or smaller to fit between the new distance of their fingers. If the scale sensitivity/ ratio was far higher and a small increase in pinch size brought the users' image from a full page of text to a single letter, it would not only be most likely too sensitive to be usable, but also jarring for the user. If a user did want to zoom in or out rapidly, beyond the scope of what one pinch represents, they can flick their fingers open or closed and remove them from the screen quickly before those two contacted points settle into the last two places their fingers were on the screen. This gestalt gesture feels right and as such has been adopted as an industry standard by a plethora of devices over a variety of platforms.

Creating a rich experience is one of the most crucial details to get right in my design. For this to happen, the mapping between interactions and the sounds they create need to make sense. For example, let's consider controlling a low pass filter, seen in Figure 13, through movement in a virtual space. A low pass filter maintains all of the frequencies of sounds that

go through it below the “Cutoff Frequency” threshold, emphasizing those within the “Resonance” if there is any resonance, and “rolls off”, or scales down, all of the higher frequencies above the cutoff. The effect can be quite noticeable even to the untrained ear if used on harmonically rich signals. Would this customary 2 dimensional graphical representation be the best model to base 3 dimensional controls on? As a designer, this is my visual basis for how to even consider thinking of controls like this in the first place, where some sort of vertical movement represented raising the power of the resonance frequencies while lateral movements would extend and retract the cutoff frequency. However, particularly to users who are not familiar with this graphical representation of a low pass filter in particular, would this mapping make sense?

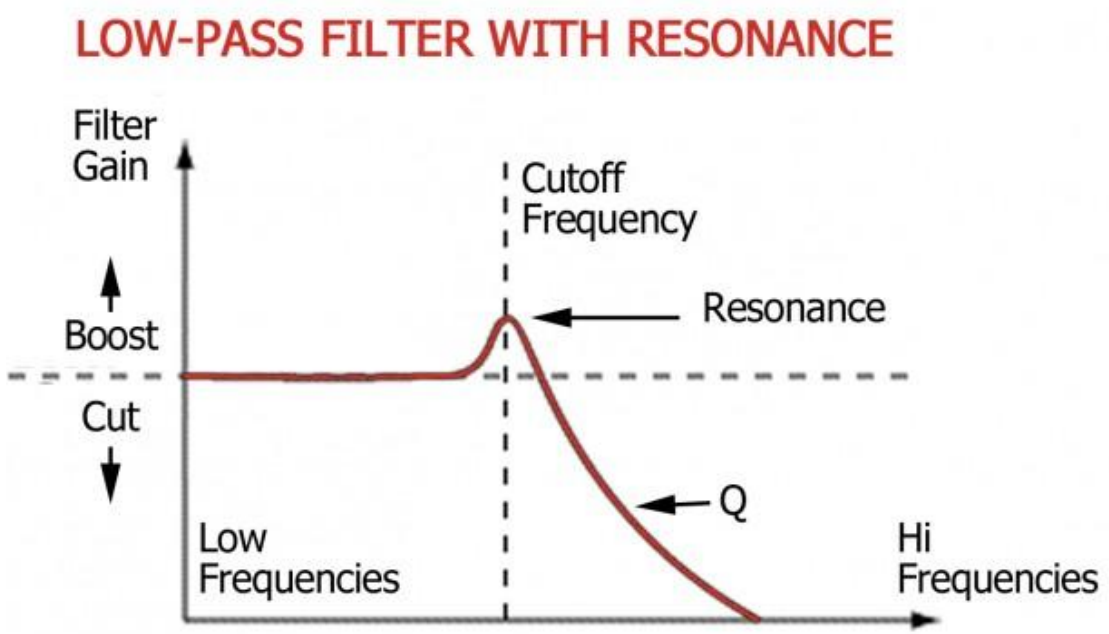


Figure 13: Standard Low Pass Filter Representation

I suppose it would be impossible to clearly define a gesture as *the best* representation of the interactions at hand, but I will create a variety of cohesive mappings that test what people of all levels of knowledge about music production and virtual reality find most intuitive and deliberate.

In chapter 4, Neely digs into how to become “soma-literate”. “Soma Literacy” is a recognition of the value of analyzing and taking advantage of the affordances of our bodies in

both embodied experiences and within interaction design as a whole. What aspects of our bodies need to be accounted for to make designed interactions meaningful?

Neely notes literacy is to be keen to what those who are illiterate do not know. Figure 14 shows Neely's 45 concepts of Soma Literacy:

Figure 14: 45 Universal Principles of Temporal Design

These principles provide a framework for what makes good design possible. What I find particularly interesting is that many of these are musical features already. Neely refers

briefly to music throughout this book, but temporality is truly one of the most distinguishable aspects between visual designs and audible ones. Even starting with the first term, *Accent*, makes me think about the differences between how I perceive an accent in music (usually a particularly emphasized volume or duration of a note to make it stand out) and an accent in still, visual art (for example a complementary color as an accent, sparsely used in a piece, to add emphasized focus to that point). How can I incorporate audio accents, and potentially also visual accents, to make as rich of an interaction as possible?

Some of these interaction guidelines will be more applicable than others (for example thinking about the *beat* and *tempo* of interactions to help users feel like their movements are more of a dance that can be performed relatively in sync with the sonic landscape will most likely be more applicable than finding a *weakest link*, whereas there are parts of an interaction that are clearly less valuable and are closer to failure points or more difficult to use simply because they were prioritized at the end as being less necessary or safer points of malfunctioning). However, even those less prioritized still have immense value and will be a useful evaluation tool for the success of my designs. I will evaluate each interaction against these 45 terms, and I will also evaluate in the user testing phase in reference to these 45 terms to see how well others think each mapping or environment successfully creates rich interactions.

2.4 Too Many Notes: Computers, Complexity and Culture in Voyager by George E. Lewis

The Voyager was a reactive computer embedded in a full sized grand piano that would analyze up to two performers' improvisations in real time and would use them to guide an automated collaborative composition. The process was nonhierarchical, in that multiple streams of music could be generated simultaneously off the performers' input without any streams taking precedence over one another.

Lewis argues that musical computer programs, as with any UX design, inherently represent the ideas and biases of their creators. As an example, one can look at the ways in which the Voyager's responsive interactions "reveal characteristics of the community of

thought and culture that produced them." (Lewis, 33). If we refer back to the conversation surrounding biases in the design of East Coast vs. West Coast synthesizers in the United States, I am personally interested in creating a ubiquitous and equitably interoperable sonic experience. The Voyager, on the other hand, explicitly embodied African-American cultural practices in its development and self-definition of "composition". Below, in Figure 15, you can see the general framework for how the computer was instructed to respond to the improvisation of the player.

```
:ap setphrasebehavior ( -- )
    ::ap" general phrasing " ( task recurs at intervals of 5000-7000 ms )
    5000 time-advance 11 irnd 200 * 5000 + to cycle

    begin
          ::ev
          bodymusic 0=              \ in this version this red light is always zero
                if calcork          \ set up new group of players, including number and position in space
                else allplayersoff  \ turn off all groups and start over with a new group.
                then
\ set up how system will follow input; set MIDI timbres
          setfollowbehavior         setreplies            setvoxbehavior

\ set melody algorithms, pitchsets, reverb and chorus type
          setwavebehavior           setscalebehavior      setreverbbehavior       setchorusbehavior

          computer-solo?            \ if no one is playing, I have a solo

\ set volume and velocity, microtonal tonic transposition
                if setvelbehavior   setvolbehavior        settonicbehavior

\ set octave, interval range, duration range
                setoctbehavior      setintbehavio         setwidbehavior   setlegatobehavior

\ set length of notes
                bodymusic 0=        \ in this version this red light is always zero
                         if setrestbehavior \ set up average degree of silence
                         then

\ set portamento, whether or not to follow tempo, and tempo ranges
                setportabehavior    settempofollow        setspdbehavior
                then

          ;;ev
    cycle time-advance
    again
    ;;ap
;ap
```

Figure 15: The Voyager's *setphrasebehaviour* Script in Pseudocode

Lewis noted that he created over a dozen aperiodic, asynchronous melody lines that could be set to play through a variety of voices (individual note generators that could be played at the same time). The range of randomization for a variety of parameters could be

adjusted to taste by the performers, including "approximately 150 microtonally specified pitch sets, and choices of volume range, microtonal transposition, tactus (or "beat"), tempo, probability of playing a note, spacing between notes, interval width range and MIDI-related ornamentation such as chorusing, reverb and portamento, and how such parameters as tessitura and tempo can change over time." (Lewis, 2000, p. 35) This is the finite level of control that I want to be able to incorporate into my designs. To me, musicality is not the "rhythm and pitch" being played– it is everything in between and built off that which gives that music breathe and life.

I am yet to figure out how to actively anti-bias my designs. However, I would like to create a system or experience that is so adaptable that either any genre or culture-specific tastes and musical languages could be adapted with ease, or a system that either encapsulates all or no innate bias towards pre-existing genres or cultures. The Voyager creates a kind of rapidfire dialogue between the performer(s) and the algorithm, which evolves, ebbs, and flows to create a full conversation over the course of a piece. To create this dialogue not between one or two performers and a machine, but rather between two performers *through* the machine, would require foresite into how individuals react to and want to shape sounds based on how they're coming at them.

With this framework of the dance of musical conversation between two performers through a cybernetic musical translator, one can inquire about how humans would react over a sustained communicative experience, rather than solely instantaneously. If one is in a soundscape that they have the freedom to manipulate, and the sound coming at them becomes progressively amplified, dissonant, and distorted, there are an infinite number of ways that they *could* reactively want to change the direction of said progression. However, would they be more likely to try to continue to expand and dig into that distortion? Would they try to cull the noise, or dig into the rage they hold that the sound may empower? What kinds of gestures would they perform to try to reduce or boost the harshness of their environment? Would these sounds coming from another player rather than from an omniscient "soundtrack provider" change the way that they interpret them? I believe it is all contextual, but these are questions that I look forward to further exploring and analyzing as I create and begin user testing my experience.

2.5 The Craftsman by Richard Sennett

Sennett uses *The Craftsman* to discuss the history of craft expertise. Sennett often refers to Hannah Arendt, who notes that speech and action are how humans communicate and acknowledge each other as real. Of course, we can see the matter that exists to compose our peers, but it is the speech and action that actually makes one alive. She refers to the ideas of *animal laborans* to refer to humans whose speech is so confined to routine that they may progress and explore their landscape freely and with enjoyment while losing out on humanity— the big picture. For example, there are great minds that lose sight of the humanity in other through a combination of lack of drive for social empathy and hyperfixated absorption into a task so deep that you block out the real world, think Oppenheimer with the Atomic Bomb or Eichmann making gas chambers more efficient. Contrarily, the *homo faber* refers to humans as makers, who see the value of their creations within the system at large and create to bond, rather than to create. "Thus, in her view, we human beings live in two dimensions; in one we make things; in this condition we are amoral, absorbed in a task. We also harbor another, higher way of life in which we stop producing and start discussion and judging together. Whereas Animal laborans is fixated in the question "How?" Homo faber asks "Why?"" (Sennett, 2008, 7)

I hope to create a virtual space as a *homo faber* for both the *animal laborans* and *homo faber*. If one wants to grind away, learning how to manipulate the tool to their advantage as a tool to communicate with others, as a *homo faber* in need of a method to express themselves physically and sonically with no means to do so, I would be pleased. If someone wants to experience my creation as an *animal laboran*, they are free to do so, but I can only hope and attempt to create a system with which the negative repercussions of their actions have no actual consequences on others in the space.

This is part of why I wanted to create a virtual landscape at the same time. Having unknowingly and unwillingly worked on projects for the military in the past, I solely hope to create tools for bonding, or at least for open and transparent communication, rather than for destruction. Although it is impossible to envision every application that any technology can and will be used for before its final iteration, I can only hope and imagine that the maximum

suffering one can inflict on others would be some temporary hearing damage. All jokes aside, if there is a way for users to convey emotional messages to one another through sonic control and have the nuances of them be understood, that would be exceedingly exciting.

A "doer" reaches an expert level of skill when a task can be done exceptionally well with minimal effort or friction in the creative process. This level of competency in context can be referred to as the flow state. The first handful of times in my life that I'd experienced the bliss of utter focus and understanding of my body as an extension of my mind rather than something to battle with were occasions when I was completely engrossed in the energy and creative process while making music. Similarly, several of these occasions were reached when exploring my musical tools collaboratively, oftentimes with people of a far superior skill level who could keep me grounded and guide me in the right direction. When I lived in California, I had a few friends and teachers who I would play music with that could adjust their playing so effortlessly as to make my mistakes seem not only intentional, but *powerful*. That is a sensation that has yet to occur in many of the other crafts that I learn. To be honest, in a digital era with such easy editing capabilities and so little actually created in "real time", that mystique is not only less valuable but also vanishing overall. The "happy accident" that sparks a masterpiece seems to relegate itself to the physical, analog, acoustic world, but that is a conversation for another day.

There was also a traditional social hierarchy in spaces dedicated to a specific craft. In medieval workshops (and often in trades today), a workshop was dedicated to both learning and producing. A community developed around apprentices working alongside the experts to learn hands-on how to efficiently grow in their craft. The knowledge is more widely available than ever for people to teach themselves "DIY" style, but the nuance of mastery and style is lost when the apprentice is no longer spinning the glass over an open fire while the resident artisan is using the bathroom. Amateurs are not learning the trade secrets that wouldn't even be thought to be taught verbally or in a lecture-esque setting, and certainly not getting constant feedback from resident experts on their own craftsmanship as well.

I think there is often a misconception that flow is only achievable through mastery of a craft, whereas I view it as passing a threshold of confident proficiency in a particular moment, or being an expert at that moment where your creativity flows and your production is effortless. Sure, this comes much more easily and frequently the better one is at a skill.

However, people can leverage access to a true master (or a masterfully designed tool, perhaps?) to experience that bliss far earlier on in their journey towards craft expertise than they would have themselves. This dip and foresite into what lies ahead sparks joy, self discovery, motivation, meaning, and inspiration into the "apprentice" in a way that would be impossible without collaboration.

Sennett also discusses the notion of "material culture". Not to be confused with materialism, material culture is a collection of experiments and stories of how people learn about themselves through what they make. Material culture is a lens to explore not only the physical and digital tools available to a group, but also the values that they put into their creative process. This can be applied beyond tangible goods to the tools surrounding music production and the audio space itself. For example, if you were to look at European Gregorian chants that are over a millenia old, you could learn a lot about the culture and why they were composed the way they were. These songs were often commissioned by churches and religious organizations, which explains the religious connotation of lyrics. As such, these songs would be sung primarily in large churches and chapels. These spaces would be incredibly reverberant, which meant that notes would innately blur into one another. In order to make a song that would sound reasonable and powerful to an audience without getting muddled by the space it inhabited, the songs would often have a four part harmony maximum (also relevant as these pieces were almost exclusively solely sung, as opposed to played on a piano or other polyphonic instrument) where all of the notes were incredibly harmonic, small steps apart, extended, and open. There would also be no designation of where the singers arranged themselves in the space, as if they were loud enough, the audience should not have been able to distinguish a singer slightly off to their right from the reverberations they were producing on the left. The music was used and produced to fill the space at hand.

Speaking of which, Sennett also delves into the nuances of the hand. This breaks down into both the ideas of tying muscle memory with visual coordination (essentially reducing the division between mind and hand), as well as simply appreciating the finite details that our hands can discern. As for the muscle memory aspect, experts of any physical craft will have put hundreds, thousands, or even dozens of thousands of hours into their craft,

deliberately repeating the same actions so that they become effortless. Expert musicians of all instruments practice "licks" to the point where they become as fundamental to an improvised melody line as speaking a syllable within a word is to you or I. To reach the level of comfort and understanding to be able to utilize and combine the affordances of a medium without batting an eye is to be truly proficient. At this point, muscle memory becomes an endless database that the creative mind sorts through effortlessly.

As for the fine motor skills in our hands, This is one of the affordances of the physical world that we inherently lose in the virtual world unless our perspective is shifted. Referring back to the concept of *Performance* in *How Bodies Matter: Five Themes for Interaction Design*, we can consider it much more easily believable, as well as intuitive, when our bodies' motions and actions map as closely as possible to the reactions we assign them to when they are embodied. One interesting framework for exception is the idea of action sensitivity and scaling. For example, for some computer users, an inch of travel distance for their mouse could represent anywhere from their cursor traversing a handful of pixels to the entire screen. In the musical domain, I could incorporate controls that let individuals toggle between grand gestural pitch bending of a melodic line (whereas they have the freedom to bend pitches up and down an entire octave) or "fine tuned" control (whereas they can acutely control cents of a semitone).

I aim to design the learning curve of my space to get users to the level of muscle memory where they are able to effortlessly create deliberately as quickly as possible, while still allowing for a skill ceiling that allows those truly acutely in tune with their musicality and the finite controls of their hands and being to be able to transcend expertise into mastery.

3 Final Prototype: The Gesturally Controlled Singing Space

The current iteration of the thesis is designed to be a performance tool for vocalists. Combining the feedback received during my midterm critique with my personal experience and information from my interviews with Bora Yoon and other vocalists, I realized that this would be the most usable and engaging direction to explore. I will discuss the features of my

prototype here, and the developmental process behind my design decisions below in Section 4.

The user is equipped with a VR headset and controllers, wireless lavalier microphone, and over-ear headphones to mitigate external noise. Images depicting user testing and a loose flowchart following the signal path of the equipment at hand can be seen below in Figures 16 through 18..

Figures 16 (Top) and 17 (Bottom) Depicting Users Wearing VR Headset, Controllers, and Headphones with a Shure SM58 and a Sennheiser ME 2 Lavalier Mic

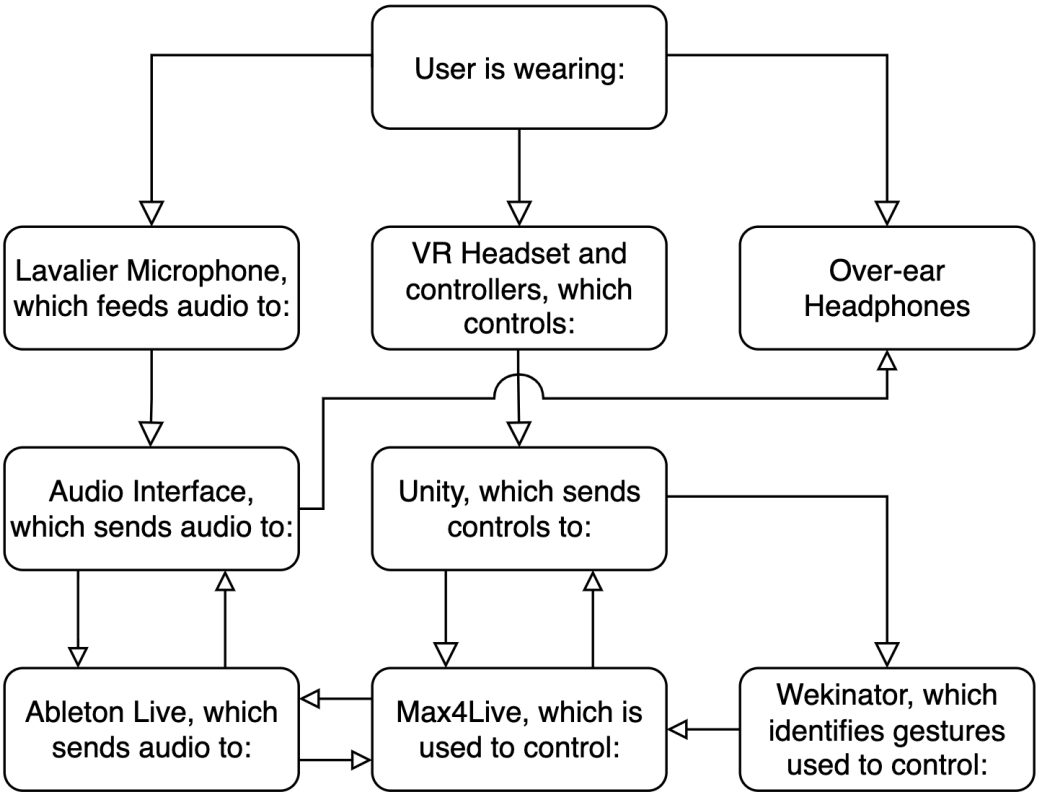


Figure 18: Equipment and Software Flowchart

The environment itself is rather empty as to not instill biases over how the performer should be relating to their music. The notion of an outer space environment has been kept throughout all iterations, as it is a blank canvas and allows for whatever picture the artist wants to paint to be delivered. All that is available is a platform to walk on in a bound, flat environment, and 4 panels that follow the player that can be triggered to mute and unmute various backing tracks so that an artist can navigate and practice or perform over whatever instrumentation they desire, as can be seen in Figure 19.

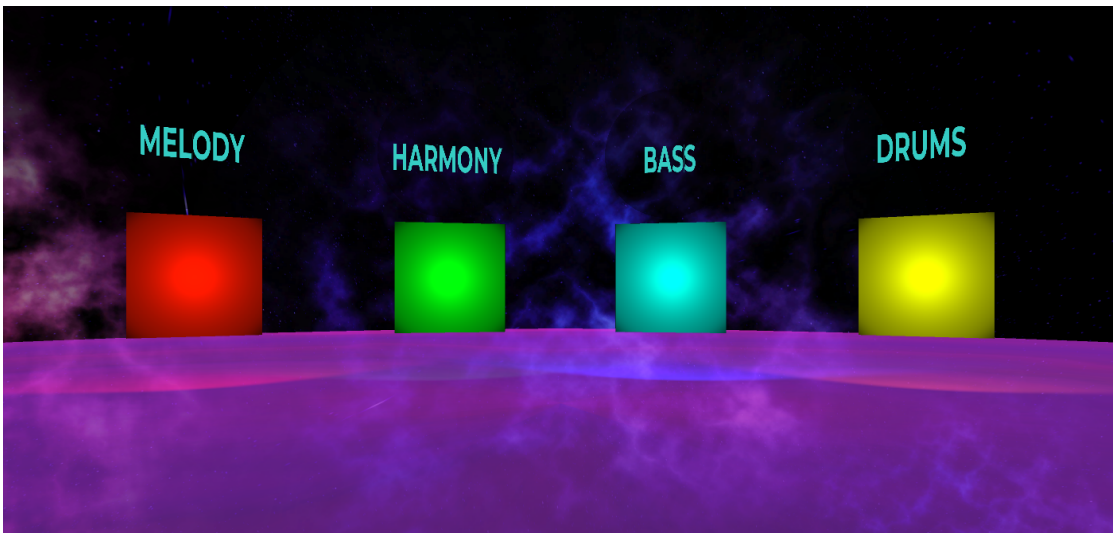


Figure 19: Gesturally Controlled Singing Space Environment

In Unity, there is a menu accessed with the secondary "Y" button on the left controller that allows players to access lyrics of their song in a clear HUD (heads up display), toggle automatically generated vocal harmonies on and off (as a method to override gestural controls), calibrate the vocal harmonies to a desired key and scale by singing the root note of the scale of the song (so minimal music theory knowledge is required), override the tempo of the song, and reset the background music and all audio effects and their settings to a default state. The vocal harmonies are almost entirely composed of the 7th chords or simple major or minor triads with the root doubled at an octave for simplicity. Figures 20 through 22 below depict the various menu screens and lyric HUD.

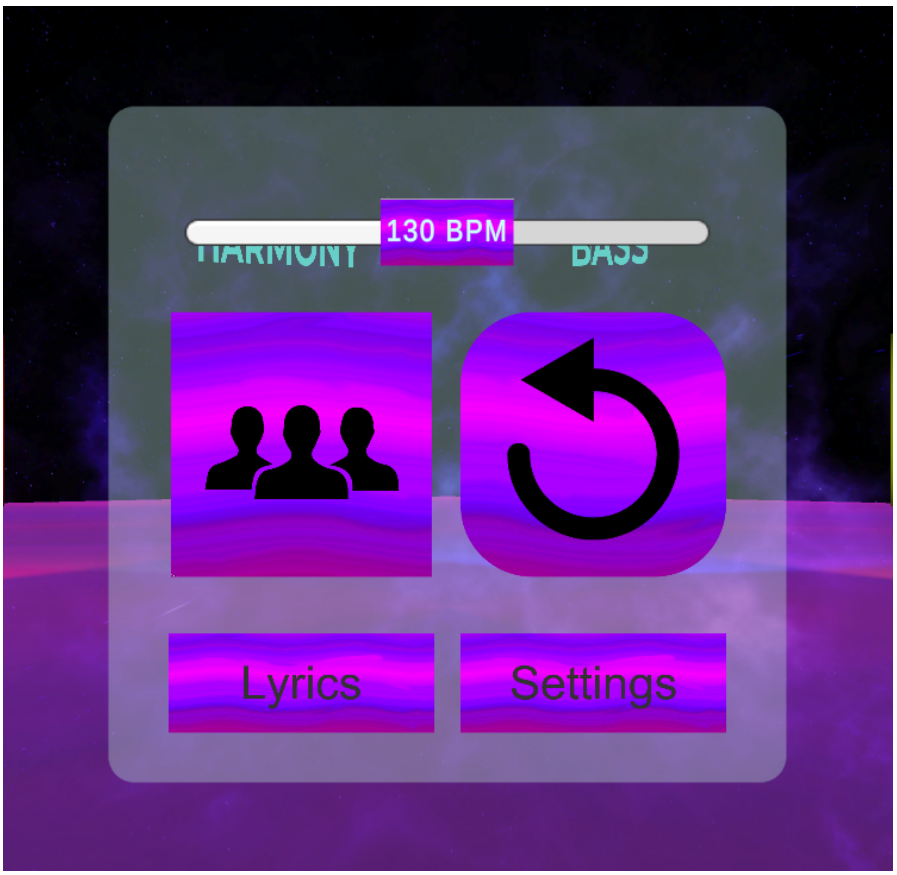


Figure 20: Menu Default Screen

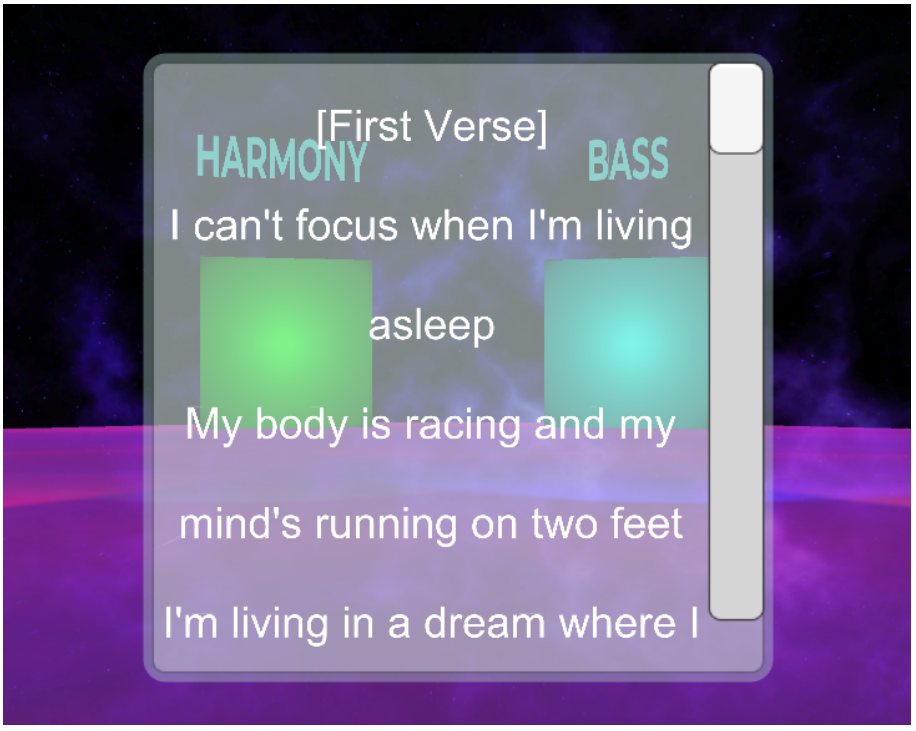


Figure 21: Lyric Display

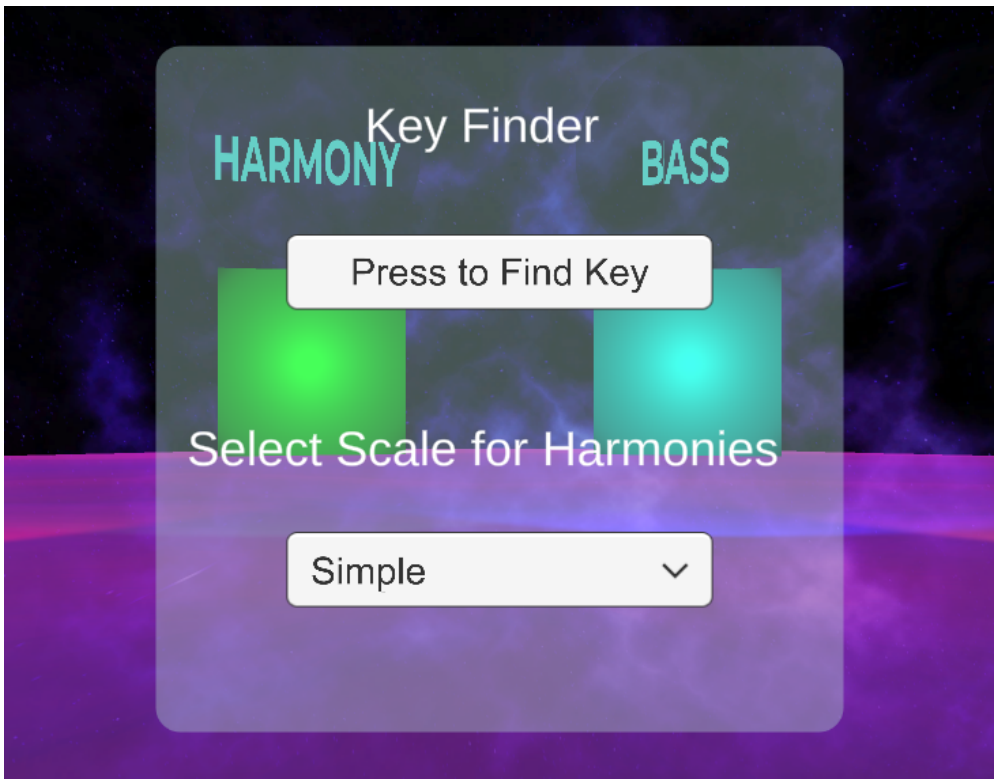
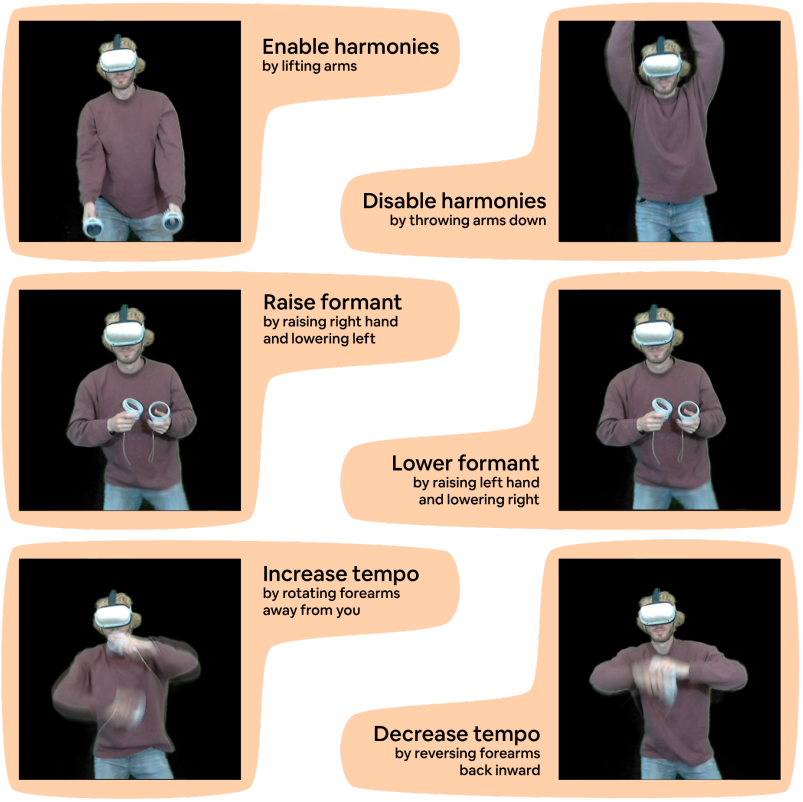
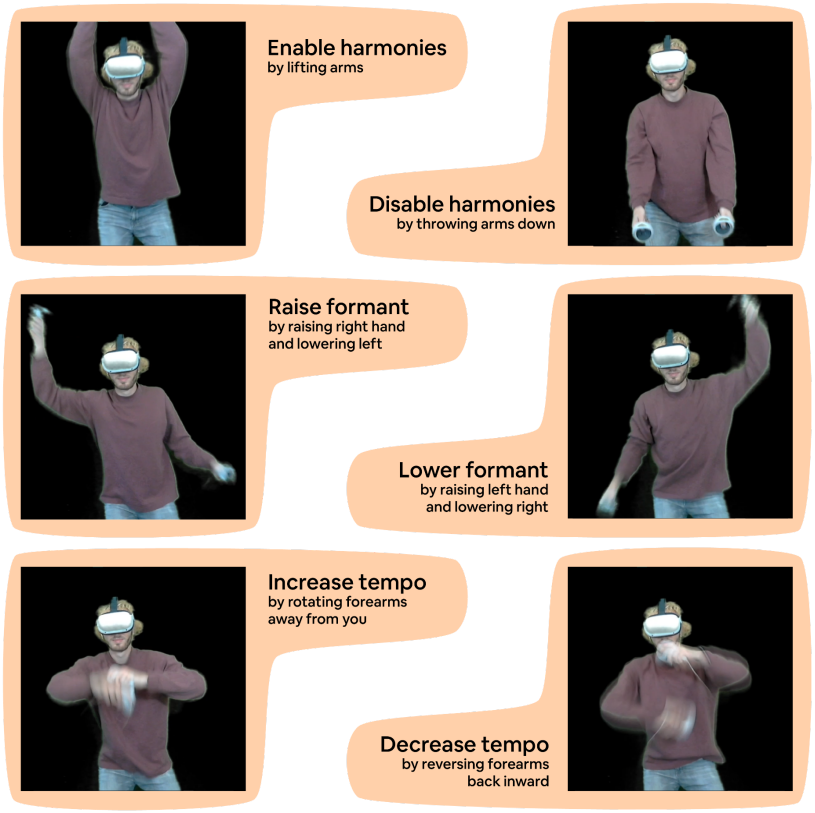


Figure 22: Auto Harmony Keyfinder and Scale Selector

Gestural controls are accessed to navigate the interactions within this environment using body motions. Performers can increase and decrease the tempo of the song by spinning their controllers over one another in front of their chest, either forwards to increase or reversed to decrease the tempo. They can also use a shrugging motion with either the left arm raised and the right lowered to raise formant shifting on their voice, or vice versa to lower it. Additionally, the vocal harmonies can be triggered outside of the 2 dimensional menu by either raising both arms rapidly to invite the harmonies in, or by throwing both arms down to disperse the automated backup vocalists. The first and last frames of gifs displaying these gestures are visible below in Figures 23 and 24.


Enable harmonies
by lifting arms
Disable harmonies
by throwing arms down
Raise formant
by raising right hand
and lowering left
Lower formant
by raising left hand
and lowering right
Increase tempo
by rotating forearms
away from you
Decrease tempo
by reversing forearms
back inward



Figures 23 and 24: First (Top) and Last (Bottom) Frames of Onboarding Gif

Additionally, an introductory video showcasing the initial features can be found in the resources at the end of this thesis, along with my website containing updated iterations of this project.

4 Methods/ Prototyping Process/ Planning/ Ideation and Interviews

This project approached several iterations with the goal of progressively narrowing the scope to a useful space.

4.1 Prototyping Process

The following subsections will walk through the development and implementation of four different approaches to creating musical tools in virtual reality. The key mediums that

all of these iterations utilize are combinations of Unity as the control interface with either Max4Live as a conversion interface and Ableton Live as the audio engine, or a standalone copy of Max MSP as both the converter and the audio engine. The way that information is carried between these programs, and others that are discussed later, is via OSC (open sound control). This is a networking protocol generally used on the same network to control multimedia experiences (e.g. audio or visual triggers in live performances). Essentially, you send a constant stream of messages through a designatable port using UDP (user datagram protocol) through your local network. It is relatively light, fast, and accurate, and although UDP is unidirectional and not as reliable as the bidirectional, error-checking TCP (transmission control protocol), it has been more than sufficiently effective for these prototypes.

4.1.1 Space Jam VR

The first iteration was based on my "Space Jam VR" project, whereas the goal was to create essentially fully functional control of Ableton Live in Unity via Max4Live. The environment was based on an outer space scene, revolving around objects fit with visual feedback of what was happening in Ableton. For example, rings around the planets labelled "Kick and Snare", "Hats", and "Guitars" would come into existence and vanish based on whether the player triggered their respective audio clips to play or pause. The large, central black hole orb was synced to grow in discrete intervals every measure, and every 8 measures would play an exploding animation and return to its original, smaller size to give the performer a sense of where in the music's temporal space they were. The animations on the intergalactic dancers also synced up to the beat of the song being played, which gave users visual feedback of the groove and made the space and project feel more fun and bright, granted their dancing was very genre specific. The scene is visible in Figure 25 below.

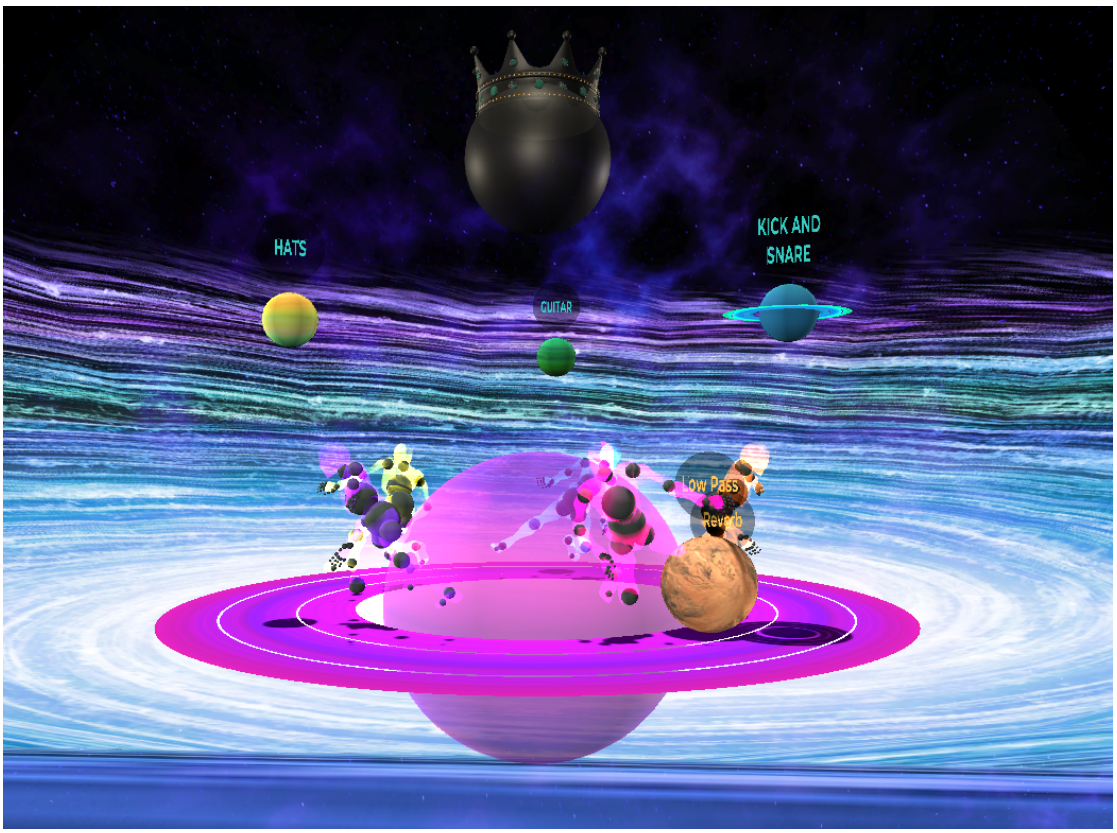


Figure 25: Space Jam VR Scene

On top of the ability to mix, the space provided control over the master track's tempo through a slider within a 2D menu, as well as perhaps the only unique utilization of VR controls in the space– the manipulation of reverb, delay, and a low pass filter on the master track of the space. Each audio effect was paired with a unique "moon" object that could be thrown around the galaxy, and both its lateral and horizontal distances from the "origin of the universe", or the larger ringed planet in the center of the scene, would control different parameters. The lateral distance, which I specifically designed to be direction-agnostic (whether the moons were thrown laterally any direction outward from the center) would control what I viewed as the "primary" functionality of the audio effect. The vertical distance of each moon would control the modifier of said primary function. The primary functions basically controlled how much the effect was happening in general, and the secondary functions represented how powerful or distinguishable that feature was, which can all be seen in Table 1 below.

	Primary (Lateral) Function	Secondary (Vertical) Function

Reverb	Wet/ Dry	Decay Time
Delay	Wet/ Dry	Feedback
Low Pass Filter	Cutoff Frequency	Resonance

Table 1

4.1.2 Embodied Synthesizer

The next phase of prototyping explored the opposite direction. Rather than being the director of the full sonic landscape scene, this was an attempt at synthesizer embodiment, whereas the foundational idea was that each player in the space had complete control over the various parameters of a synthesizer. One key motivation behind this shift was to fully take advantage of the affordances of having three separate controllers, each with 6 degrees of freedom that could be tracked. In the previous environment, the interactions were more or less "point-and-click", making them no more useful than using a computer mouse for live control other than moving the moons for audio effects around the space. In fact, with the rotating planets not always being in view of the player, it limits the player and makes it more difficult to do what traditional interfaces offer, and I wanted to simplify my approach.

In theory, I wanted to link every specification on a synthesizer using the VST Serum to a player's motions. In practice, I narrowed it down to some of the more prominent features. Players controlled a single duophonic synthesizer, whereas each controller was representative of one of the available voices. The controller dictated the processing on said voice, whereas the lateral distance that the controllers were from the player's headset represented the pitch being played. This idea was based on a theremin (Figure 26), an instrument that the player interacts with by manipulating the orientation of their hands around the electromagnetic field of two antennas that control the pitch, volume, and timbre of a single voiced oscillator. However, many musical instruments have a linear layout of pitch, where interacting with the instrument further in one direction increases the pitch being produced and further in the other decreases it (e.g. pianos and western synthesizers, guitars and other stringed instruments, etc.).

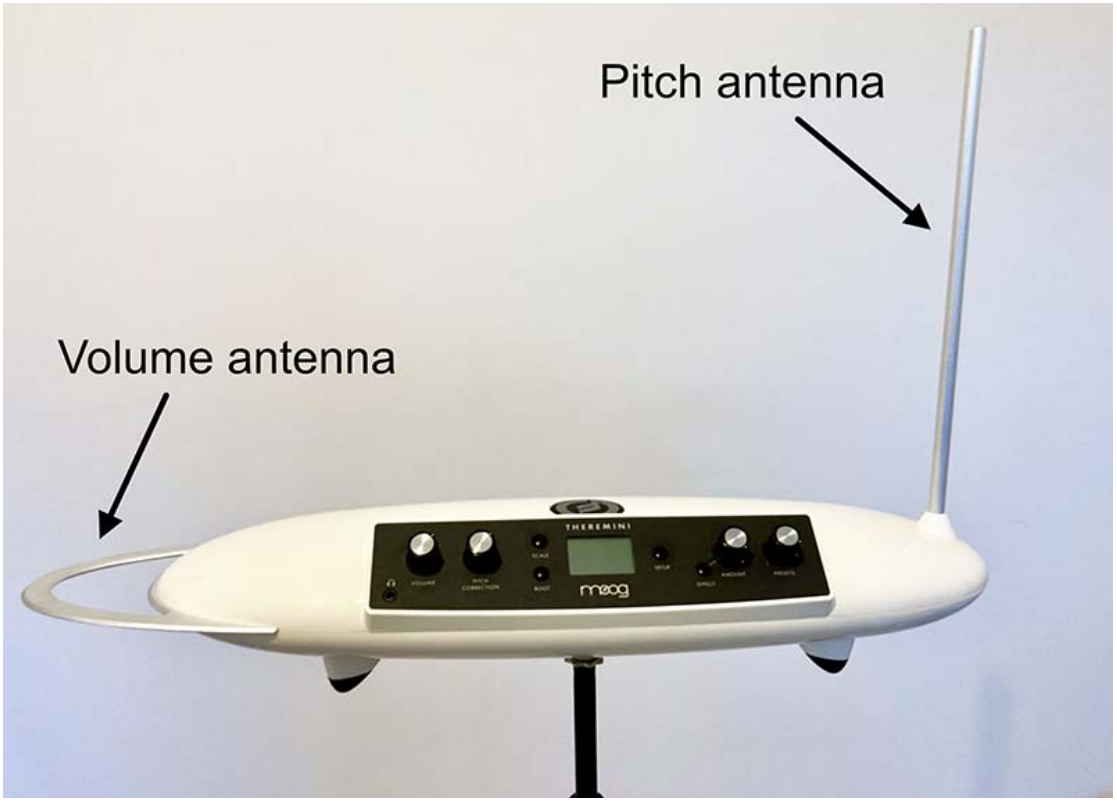


Figure 26: Theremin Diagram

This phase more than any other revealed how many preconceived assumptions I had of what the layperson would find an intuitive interface for synthesizer control. Low pass filtering each voice was represented via motion based on a standard graphical EQ, whereas the height of each controller represented resonance and the further forward from the player's headset their hands were, the lower the cutoff frequency. Not only are the concepts surrounding low pass filtering not inherently intuitive to people who have not learned about music production before, but also the standard visualization technique representing a low pass filter is not the only representation. I realized that a lot of the biases that I had of control over a performance space were based on my understanding of the tools that had already been developed and were standard to the industry, rather than of the human connection to the audio effects I was trying to designate.

Additionally, the Unity environment that contained this performance space was rather empty– it was simply the first environment with all of the planetary objects and dancers removed. There was no real sense of directionality, as I wanted the controls to be entirely intrinsic. For testing purposes, there was a simple UI that displayed the positional coordinates of each controller and the headset, but it was rather difficult to make sense of. I

realized that in the next iteration I would want to guide users visually to pair their actions with real time confirmations of their levels or states.

At this point, I was still heavily considering making this experience multi-user, and as such I envisioned the distances between players to have profound effects on the music as well (e.g. tempo control). One other unique characteristic of this space was that it was the only "standalone" iteration that I developed, where the sound being produced was entirely generated in Max MSP, rather than Max4Live. This would have allowed the program to be a lot lighter and more accessible, as it could run standalone and generate the sound natively without needing a centralized server running an expensive program like Ableton Live in the background to do the audio processing. However, there were many difficulties with this approach, as I had to design my own synthesizers from scratch rather than using a pre-existing interface and preset sounds in a DAW.

The Max patch that this scene is based on is composed of 7 different sections, which can be seen below in Figure 27.

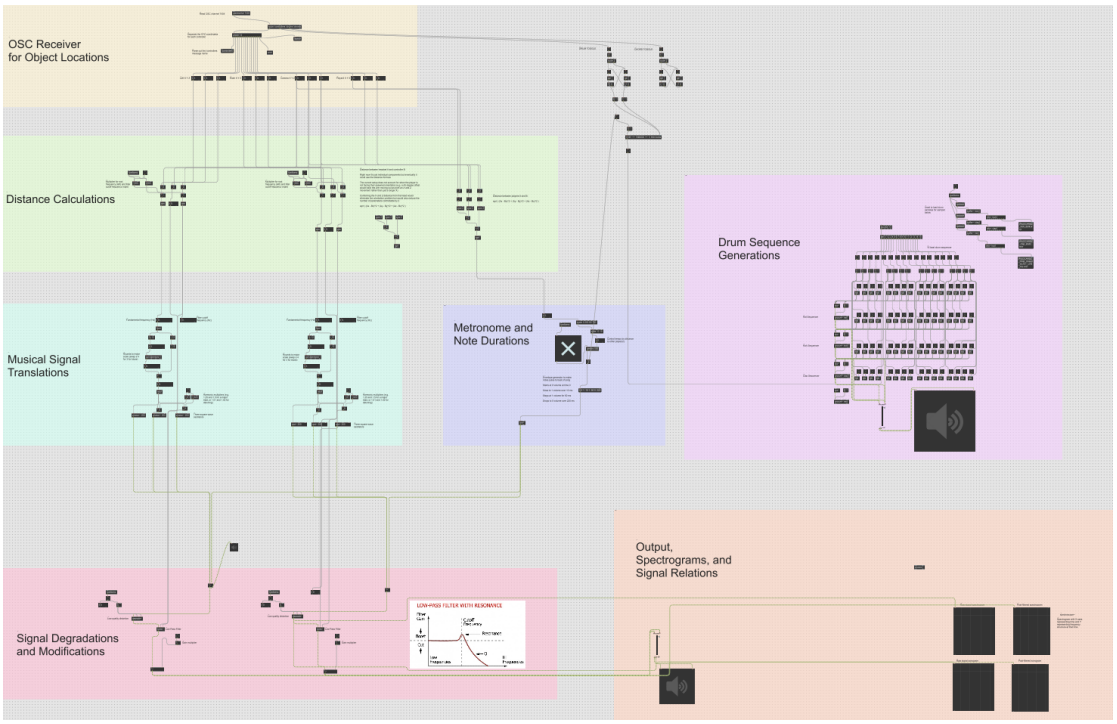


Figure 27: Synth Embodiment Max Patch

The patch begins with "OSC Receiver for Object Locations", viewable in Figure 28 below, which is in charge of receiving and parsing the positional coordinates of both

controllers and the headset, as well as simulated coordinates of a hypothetical second player whose movement was randomly generated in Unity, and separate information about whether or not the player(s) have toggled drum and chords loops to play in the background.

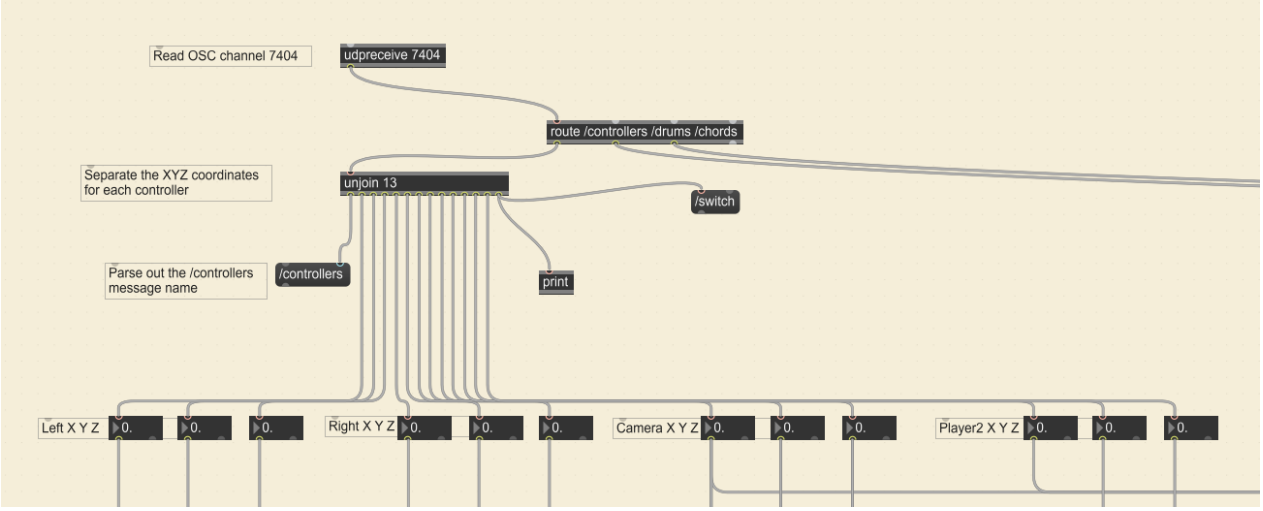
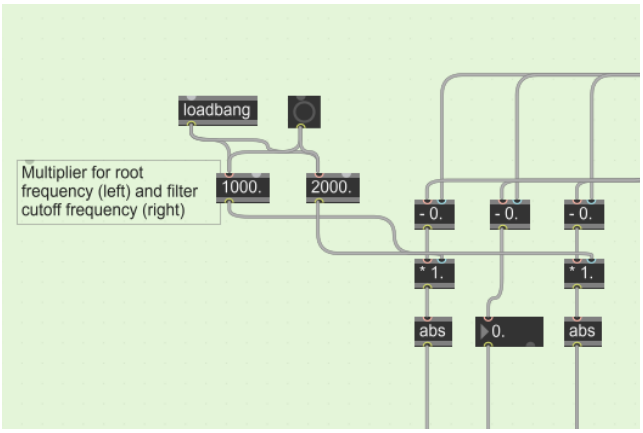
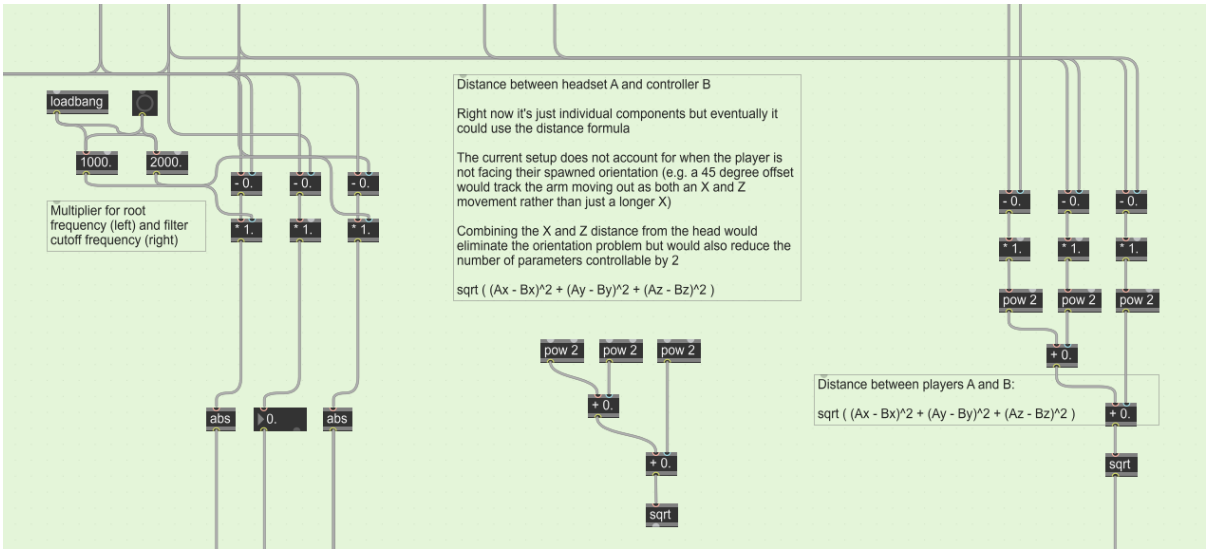


Figure 28: OSC Receiver for Object Locations

This then feeds into "Distance Calculations" seen below in Figures 29 and 30, which is where the "heavy lifting" of this patch was done. This is where the calculations of the distances of the controllers from the player and the players from each other are performed. In later iterations, I realized that it would significantly reduce CPU usage to do these calculations while in Unity to reduce the amount of information being sent over OSC if there were not many calculations being done overall. However, if I were to extend the Synth Embodiment prototype, there would likely be many more features and parameters being controlled, so it would be possible that doing the calculations in Max would be lighter than in Unity and sending over dozens of results.





Figures 29 and 30: Controller to Headset Distances Calculation

This then feeds into “Musical Signal Translations”. This is where the two distinct voices are generated (in this case, a triangle wave and a square wave). The distances are converted to frequencies, which are then converted into MIDI notes that fit standard western musical scales. This way, the distances of the controllers are quantized to recognizable distances that have pre-fit relations to one another. There is an embedded table that rounds the chromatic notes to a scale of choice as well to further narrow down the input and ensure ease of consonance and control, which can be seen below in FIgure 31.

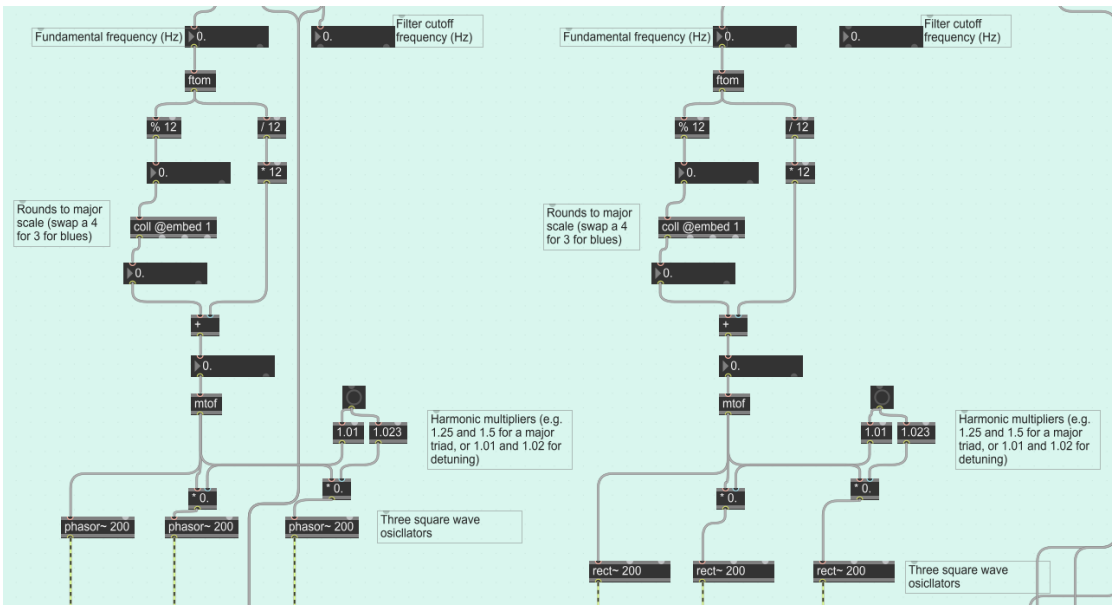


Figure 31: Musical Signal Translations

The “Metronome and Note Durations” section would dictate the tempo of the song, whether or not the synths were playing, and the shape of the volume envelope of each synth. This was one of the features I was least satisfied with, because it would essentially treat the player as the controller of a constantly running arpeggiator and didn’t leave room for rhythmic nuance or control. Additionally, the ADSR (attack, decay, sustain, and release) envelope controlling the volume of each instance of the synths was constant. I envisioned mapping these parametrically to the rotational positions of each controller, but I moved on from this idea before that was implemented.

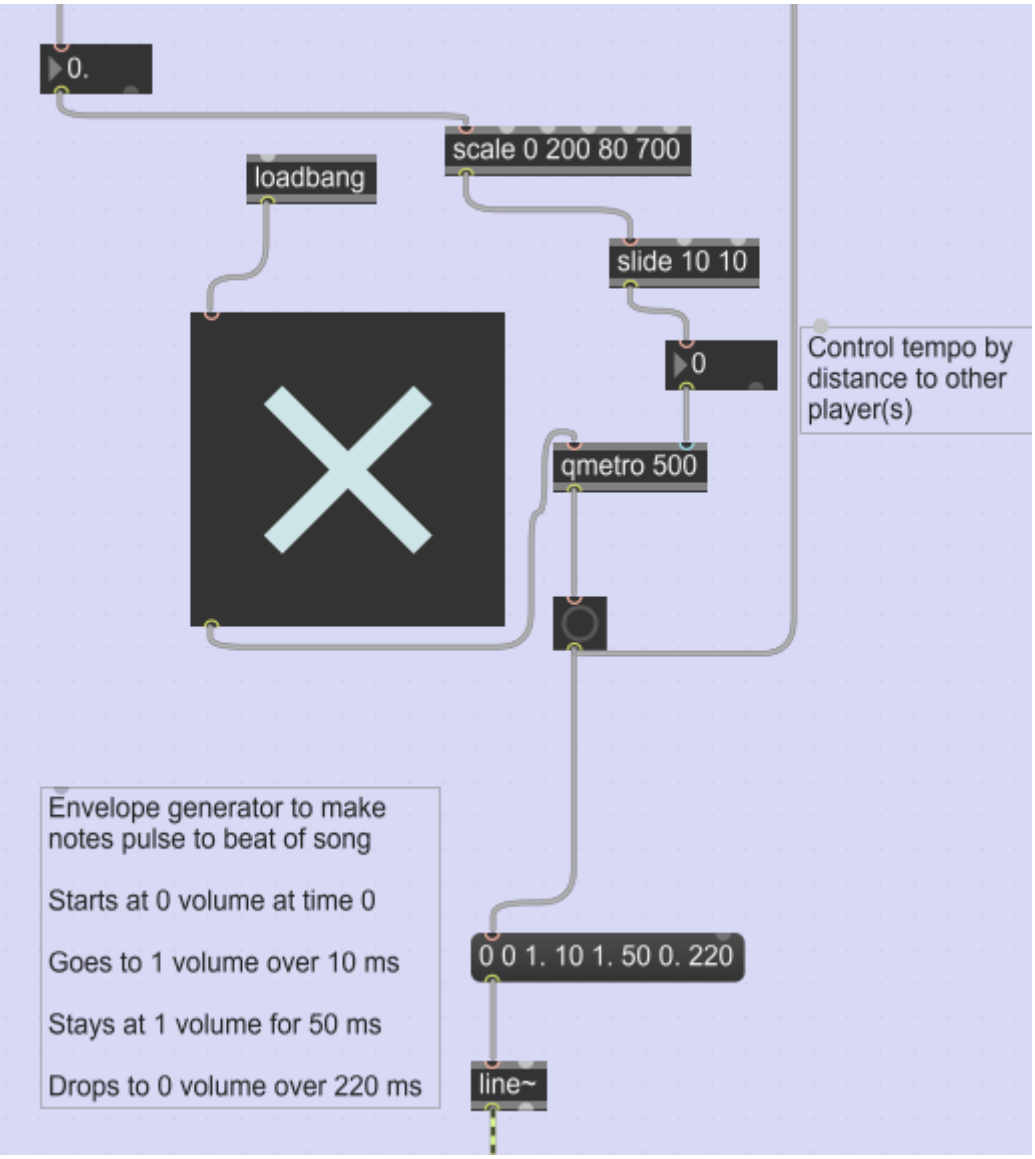


Figure 32: Metronome and Note Durations

Off of the signal path was a separate section called “Drum Sequence Generation”, which is where I would load in preset patterns for drums to play in the background to give a grander musical scheme to the scene.

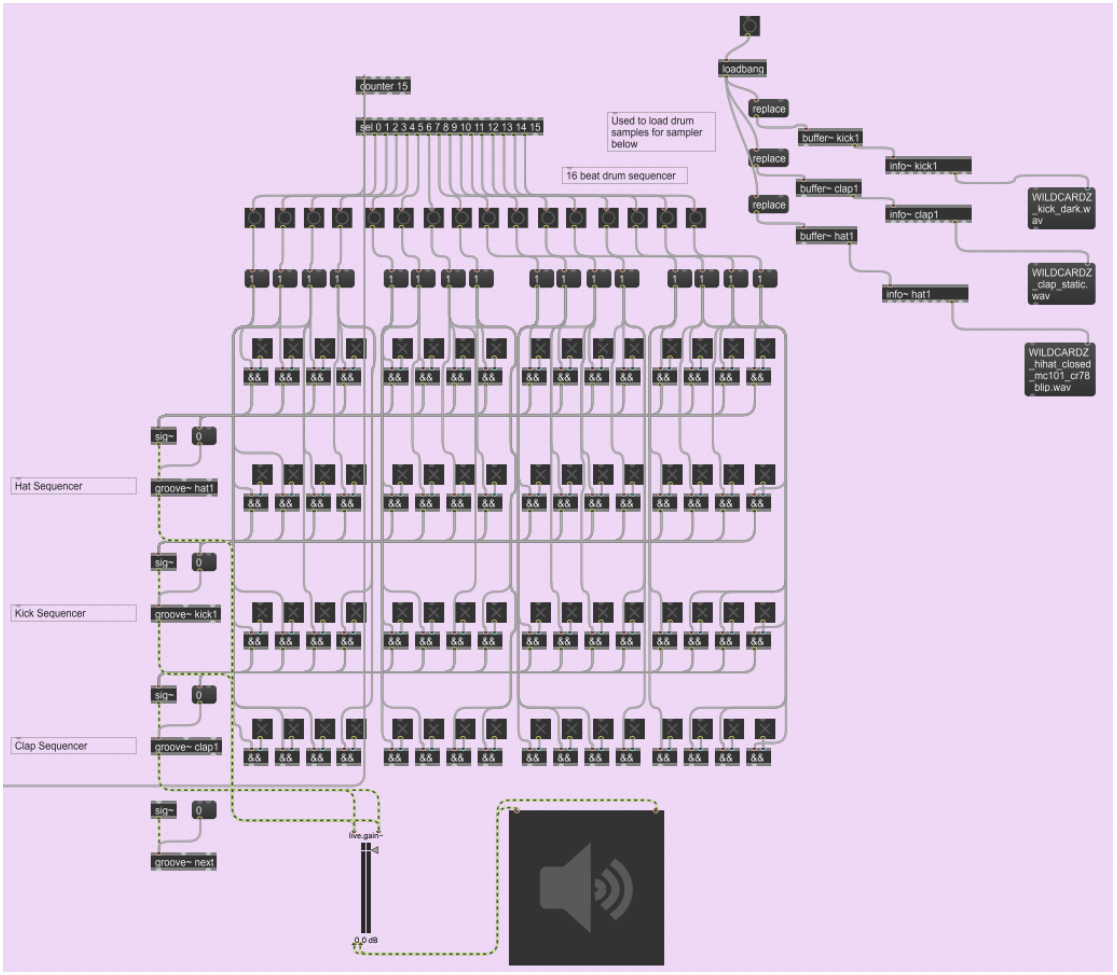


Figure 33: Drum Sequence Generations

Back onto the main signal path is “Signal Degradations and Modifications”. In this section, similarly to the multiple detuned voices in the “Musical Signal Translations” section, the signal is brought through a degradation distortion object, which reduces the sampling rate and bit depth of the signal to make it richer and show to make the effects of the low pass filters more prevalent in the mix.

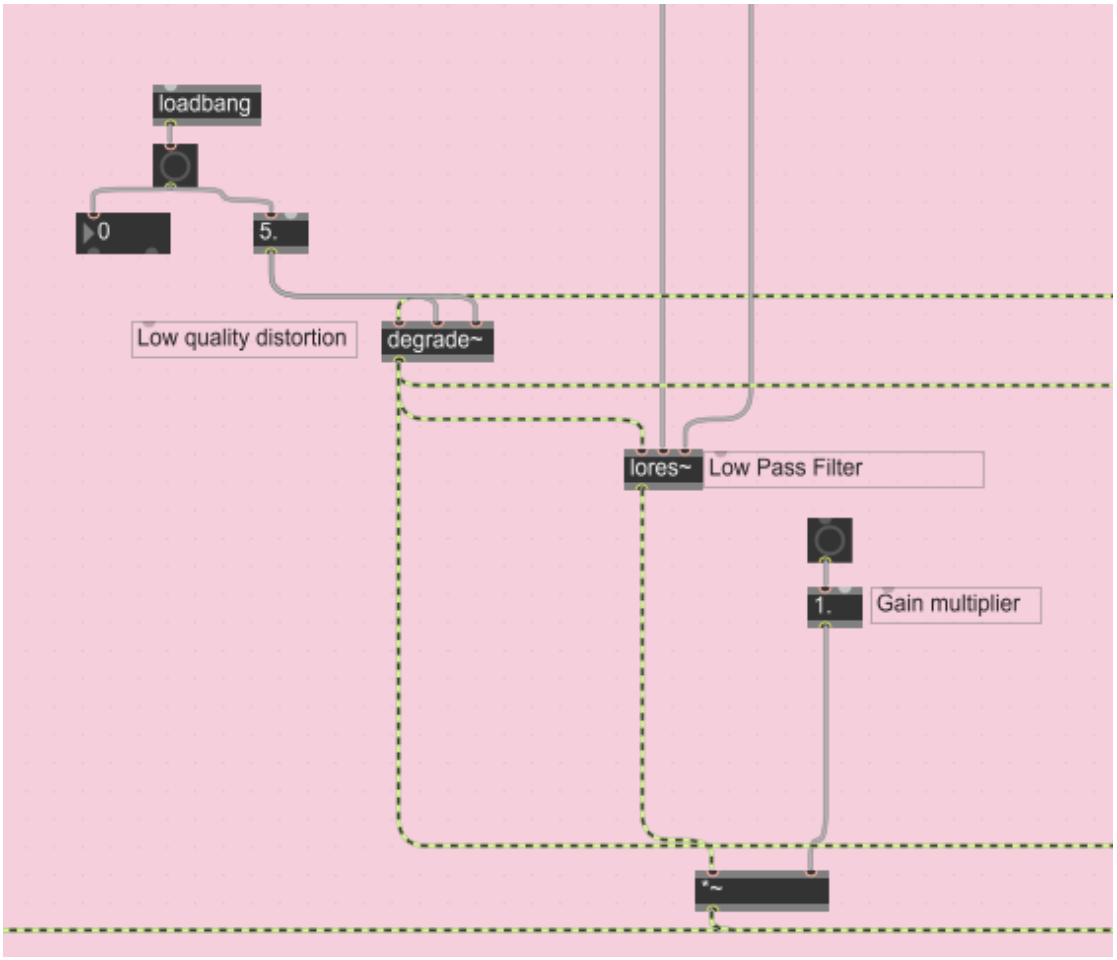


Figure 34: Signal Degradations and Modifications

Finally, the signals make it to the “Output, Spectrogram, and Signal Relations” section. This feeds the audio to a speaker output, as well as into spectroscopes that allow me to analyze what was being processed and heard at the output.

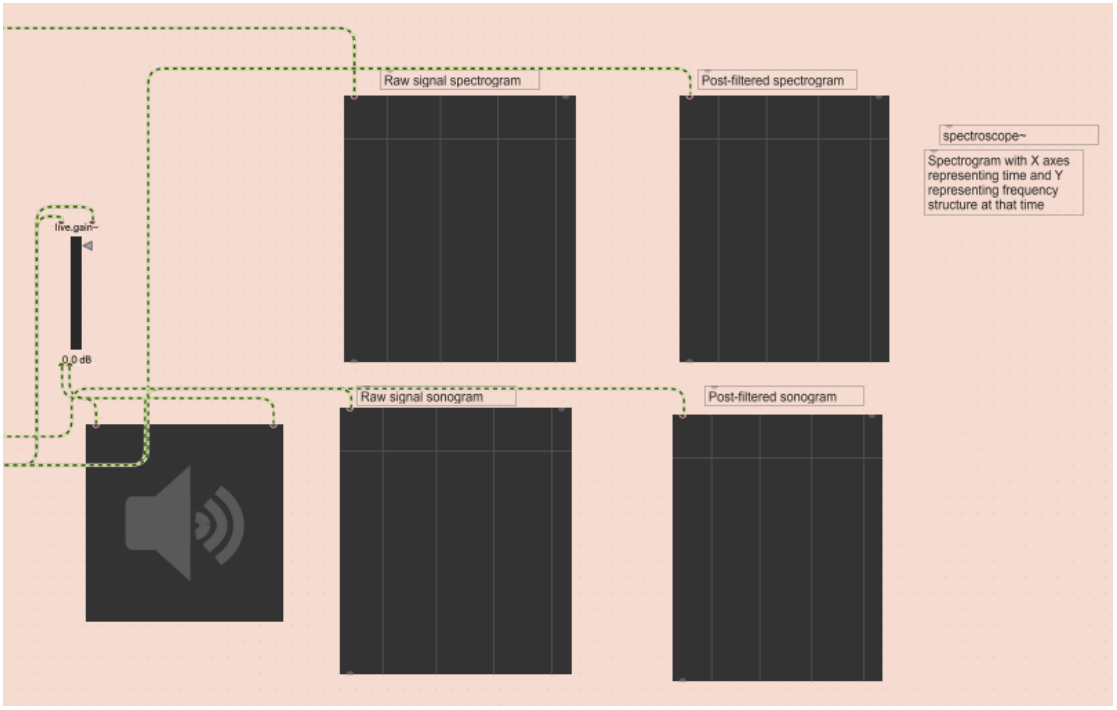


Figure 35: Output, Spectrograms, and Signal Relations

4.1.3 Manual Vocal Performance Space (MVPS)

After working on the embodied synth space, I revisited what inspired me about the the Space Jam environment– in order to get the level of functionality that I want (especially within a reasonable amount of complexity or time), it would require a more powerful host computer running a copy of Ableton Live, rather than a single Max MSP patch. It would be much less transportable, and more expensive, but it would allow me to probably create the most engrossing experience of the options I've explored so far. It was at this point that I realized it would be beneficial to shift towards the singing space. There were a few driving factors behind this shift– first of all, with the time constraints of the thesis, I realized that creating a collaborative space would be at best an afterthought, and it would be more meaningful to focus on individual users at a time. Additionally, Dan Taeyoung introduced me to the design framework of thinking about who I want to design for by posing the question: would I rather create a specialty tool that could be enjoyed/ utilized by a few people for hundreds of hours, or a generalist tool that hundreds of people would utilize for a few minutes? This sparked a paradigm shift in my goal for my thesis where I was no longer creating a novel showcase of VR for musical embodiment for the layperson, and rather

creating a useful performance tool that could become part of another performer, or even my own, repertoire.

Wanting to utilize the unique affordances of VR in particular in its current state, I needed to consider what types of musicians would be able to perform with their hands occupied by controllers, which I narrowed down to vocalists and, again, people in the producer/ DJ space. After experimenting with vocal plugins and plights myself, as well as having the majority of critiques of my own music be vocal-centric, I realized that this was the path that I wanted to pursue.

The first singing space was again based on a more finite, manually controlled interaction schema. This scene revived control of the sonic landscape at large with segmented panels representing the different instrumental tracks that could be toggled to play or pause at a quantized time interval with the click of a trigger. This allowed users to vamp, or improvise over a backing track of undesignated length as long as they saw fit, in a custom environment that was as stripped back as they pleased. Additionally, I introduced vocal harmonies as the main control schema in this environment. The menu-based, 2-dimensional interactions did not utilize the affordances of virtual reality, and were, to be quite honest, not engaging, easy, or fun to interact with. As such, I tried to limit their presence by only being relevant to settings that would be changed before a performance and not altered again. The Max4Live patch can be seen below in Figure 36.

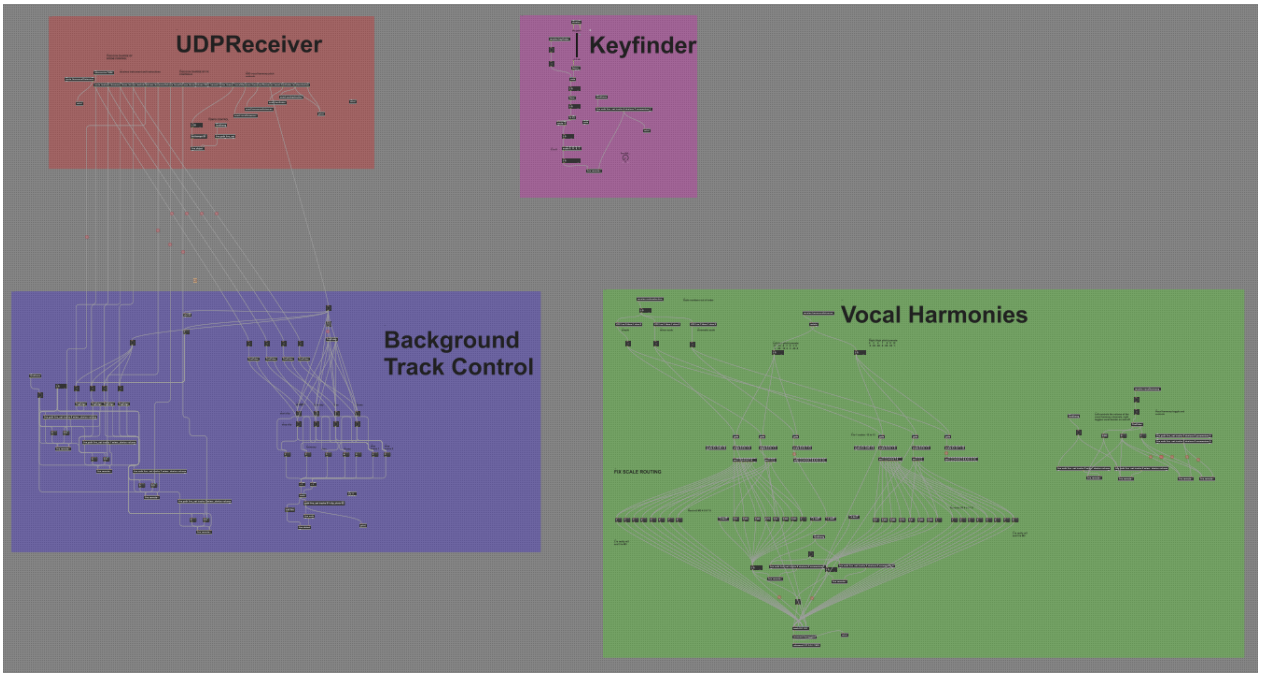


Figure 36: MVPS Max Patch

The "UDP Receiver" section parses the information coming through OSC to their respective parameters. There are several unused message routings that I wasn't sure whether I wanted to incorporate at the time but were left in as placeholders.

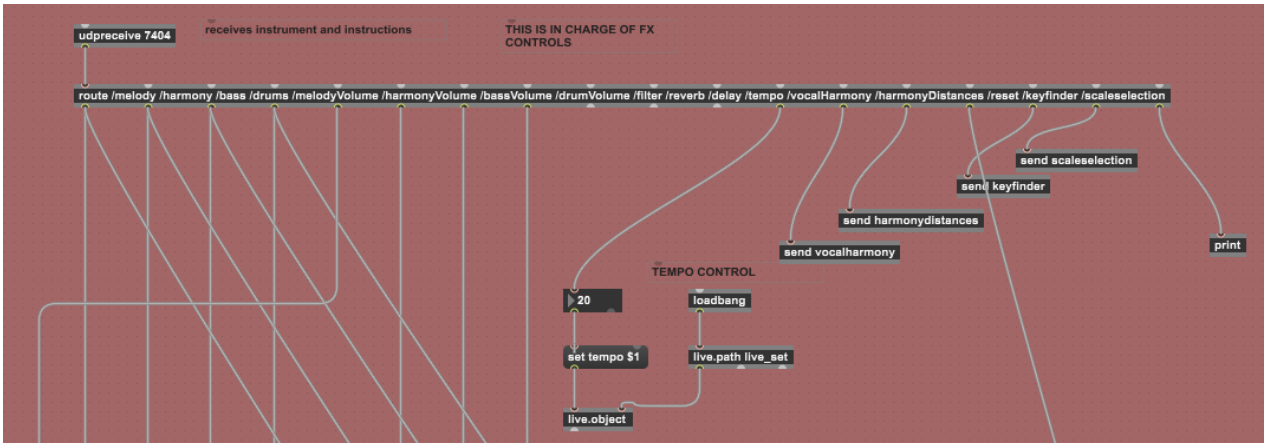


Figure 37: MVPS UDP Receiver

The "Background Track Control" section controls triggering the premade audio clips as well as altering their volumes. The clip triggering allowed players to ensure that each instrument was synced and quantized to the same cycle, which reduces room for error. The volume portion allows for more instantaneous control of reintroducing and removing instruments, as it doesn't wait for the first beat of the next measure for the control to come into action.

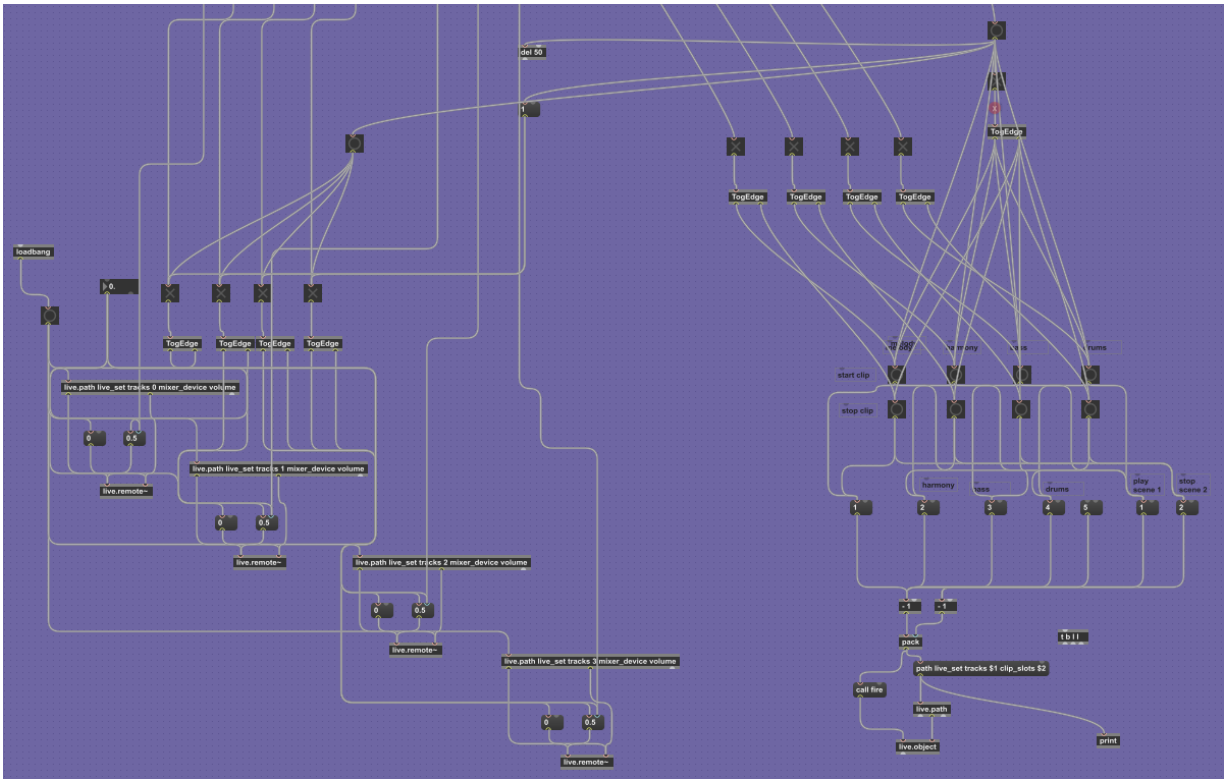


Figure 38: MVPS Background Track Control

At this point, the project at large was still being created for the virtual singer at large, and I wanted the tool to be as easy to use and adaptable as possible. I wanted users to not have to think too much about the music theory behind their decisions with certain automations of how the Max patch controlled their audio, and so I incorporated an automatic scale-selecting device for both the automatic tuning of the main vocal, as well as a basis for the scaling of the vocal harmonies. The key-finding Max Patch can be viewed in Figure 39 below.

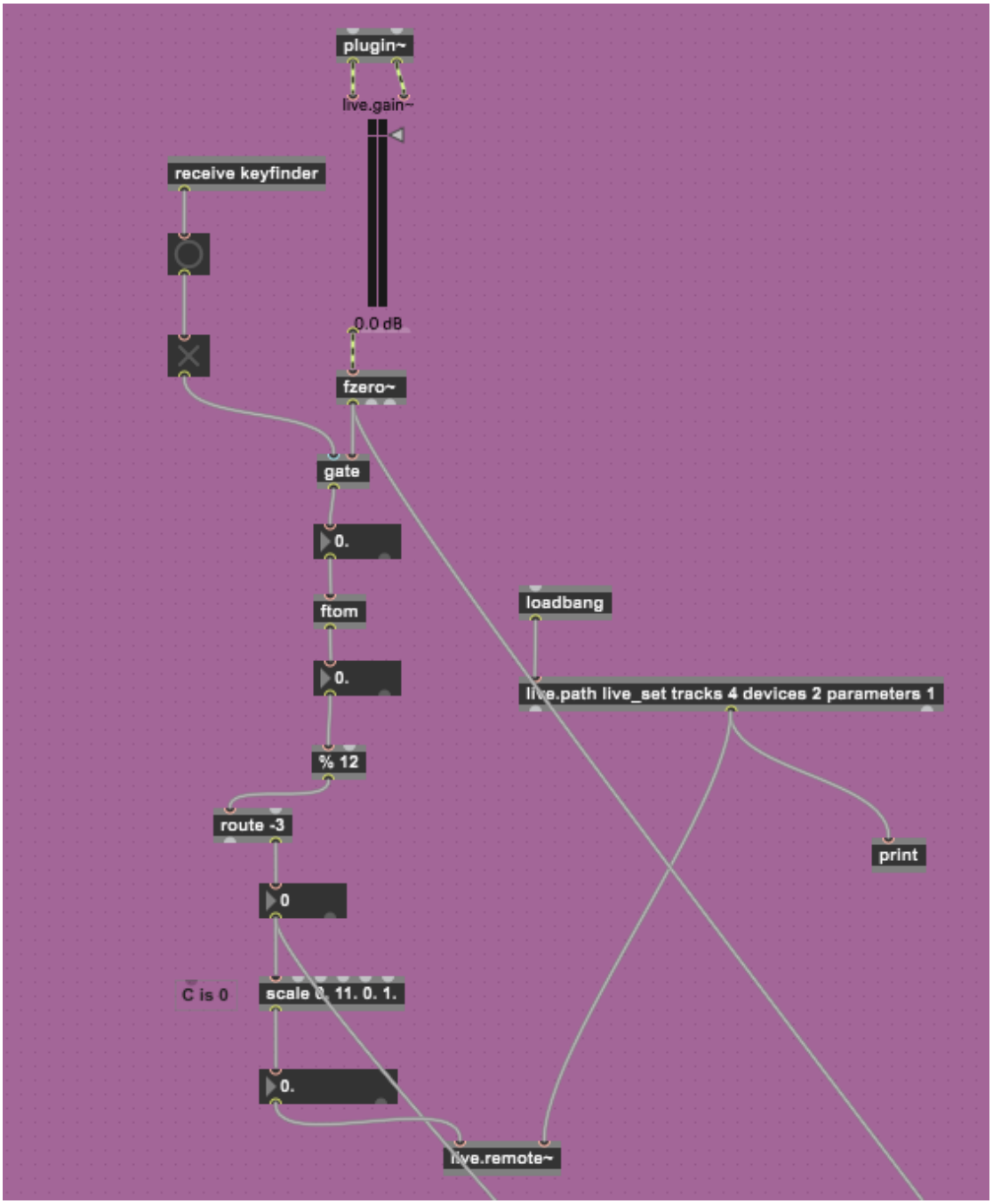


Figure 39: MVPS Keyfinder

Additionally, one of the only other menu-based interactions that exists in the space is the ability to choose the complexity of the scale structures that the vocal harmonies can be tied to. As you can see below in Figure 40, there are three levels of complexity that are synchronous between both hands. I decided to bring a more classical game-like UI structure into the space with these slider UI elements. They are representative of the distance between each controller and the player's headset, while each segment of the rainbow pastel color bars are representative of whichever musical structure was selected. I chose to incorporate only

three settings in this schema: a “Simple” mode, a “Harmonic Minor” mode, and a “Chromatic” mode. “Simple” mode contains just the root (unison), perfect fifth, and octave above the fundamental pitch that the performer is singing. The “Chromatic” setting allowed users to create vocal harmonies on every semitone up an octave from what the user sings. I have found that the “Harmonic Minor” setting is what would perhaps be most useful as a “goldilocks” setting, whereas the technical skill required to hit exactly the right note on both hands in the “Chromatic” setting is very high, while the rigidity of only 2 non-unison harmonies per voice in the “Simple” settings is very limiting. It was easier to create very interesting movements of chord progressions using this setting than in the others, but it was still rather difficult at large. I wanted this scene to be manually controlled and focused primarily on this single interaction, which is why it is the solely displayed visualization. In Figures 40 and 41 below you can see the sliders in context and the layout of the three available harmony modes.

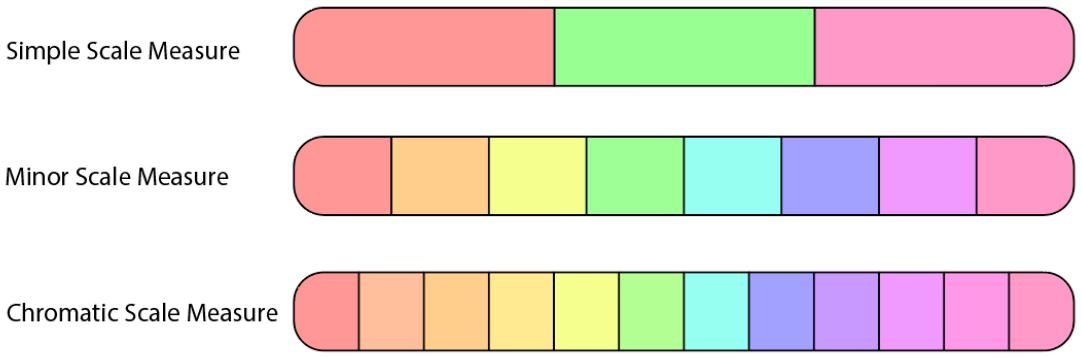


Figure 40: Harmonic Scale Degree Visualizer

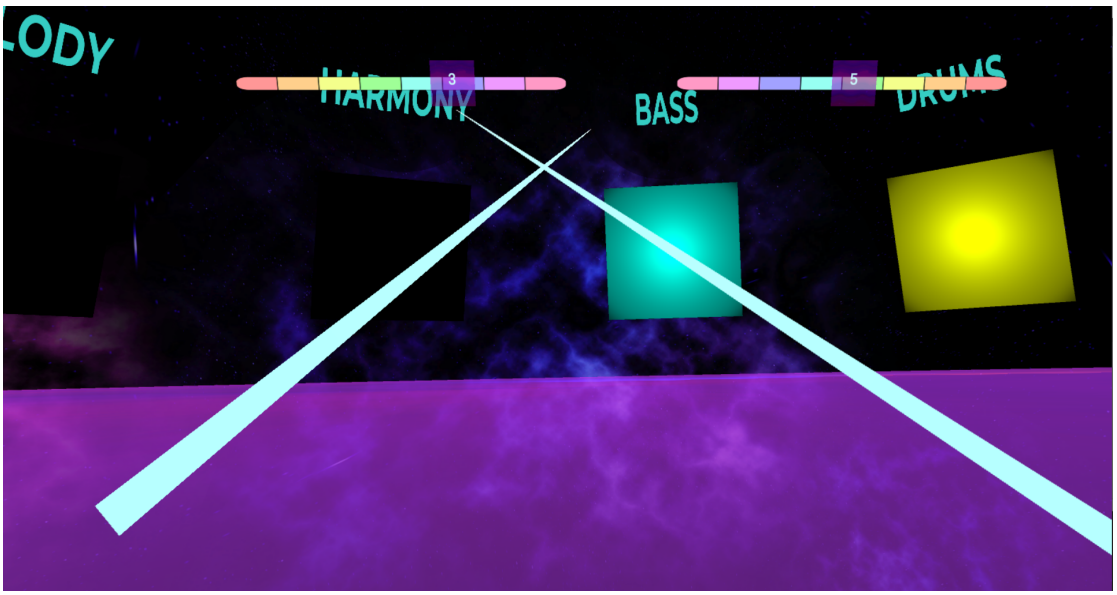


Figure 41: Manually Controlled Vocal Harmony Space

The next image depicts the harmonic calculations section, which distributes the controller's distances to their respective scale, and translates to the parameter value representing the note needing to be produced, and simultaneously changing the formants as well. The vocal harmonies are each produced on an individual Ableton Live track using a Waves Vocal Bender plugin.

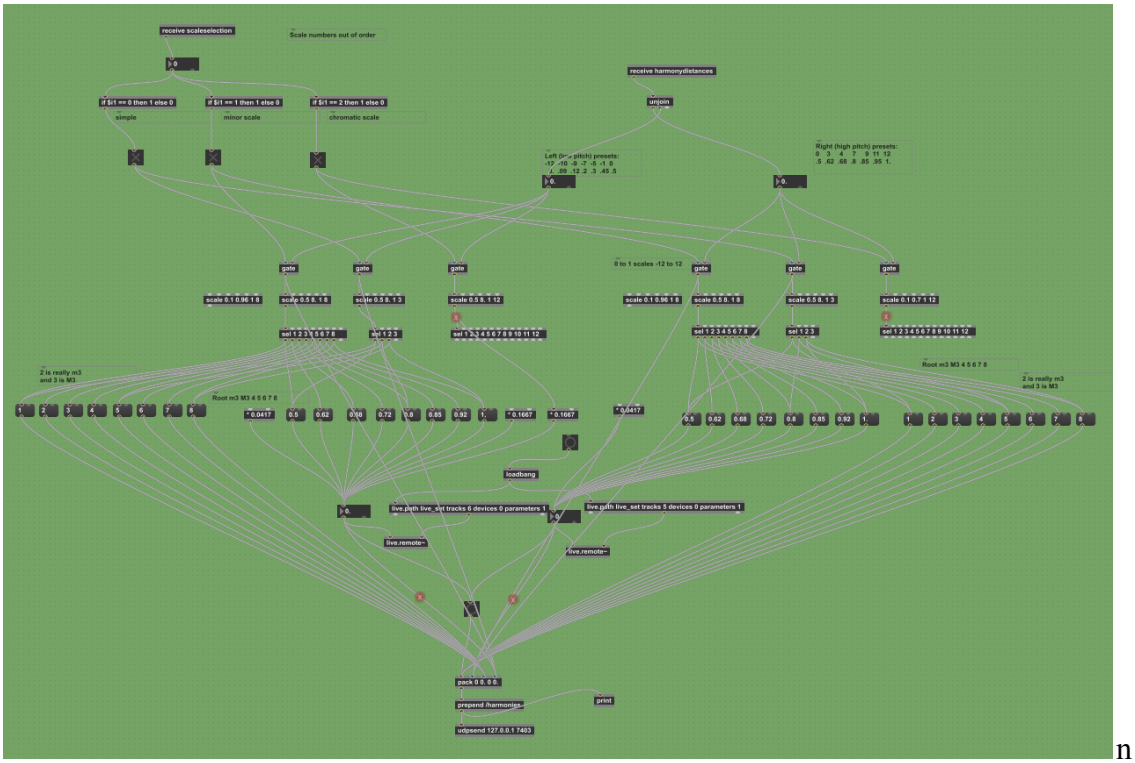


Figure 42: MVPS Vocal Harmony Calculations

Additionally, the following figure controls whether or not the vocal harmonies are present at all or muted based on player menu input from Unity.

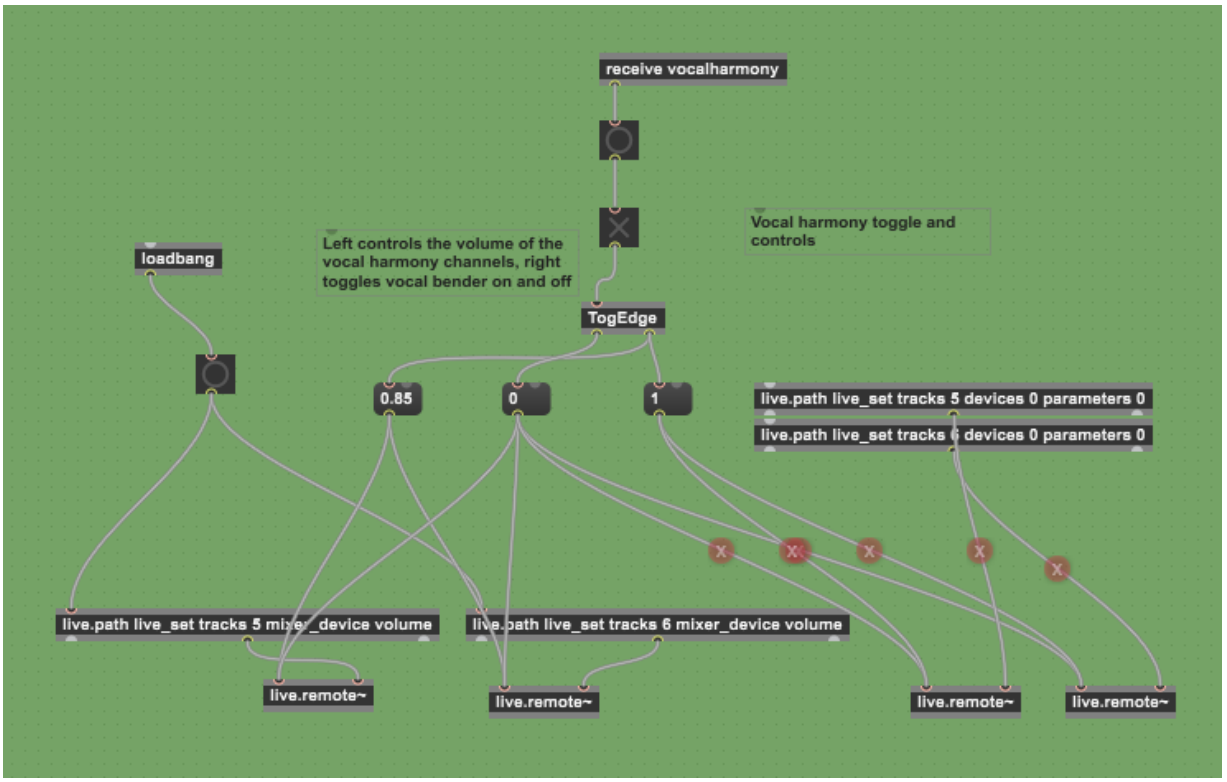


Figure 43: MVPS Vocal Harmony Toggle

4.1.4 Gestural Vocal Performance Space

Finally, this brings us to the gesture-controlled vocalist space. I knew the project to be more of a performance/ creative development tool rather than a production oriented one, but this was the point where I fully recognized that shift. Bora Yoon mentioned that one of the keys to a great performance is in the artistry of a performers' movements. Does their interaction with their instrument, the stage and space, and the audience provoke emotions? What story do their movements tell, and how can an audience perceive and relate to said motions?

If we refer back to Klemmer, Hartman, and Takayama's notions on performance in interaction design, we can consider how vital body language is in conveying nuance and meaning. For an audience to view a change in desire from a vocalist, gazing upon the lips of a stagnant body conveys a very different message than a singer who is leaning into the

audience and painting their lyrics with their actions. This is not to say that idleness is wrong or lesser, but it inherently invokes messages of constraint rather than emphasizing the emotion of what's being otherwise stated. Considering the shift from finite control to more gestural as well, it is a lot more practical to pursue embodiment interaction than digital verisimilitude, whereas there is no need to perfectly simulate reality in a virtual world if the illusion of control is sufficient. In other words, the threshold of embodiment to aim for when designing a virtual interaction space is one where the world teaches the controls and is inherently comprehendible simply by doing and exploring.

One of the key advantages of a mouse and keyboard and why they have stood the test of time is how useful they are for making acute changes across a variety of interfaces. In a DAW environment, you can accurately transcribe MIDI information, navigate every menu, etc. as the interface was designed around these tools. However, as a generalist tool, it is not necessarily optimized for the job of temporal changes or anything where many selections have to occur. In the case of a performance tool, it is important to create an interaction scheme that is "quick and dirty" and functions efficiently rather than focusing explicitly on precision. For example, minute changes in volume in a mix will be less noticeable and worth the effort of adjusting than missing harmonies overall.

With this in mind, we can also refer back to the idea of parametric control. In some of the earlier prototypes (e.g. the synth embodiment iteration), every minute change in hand and head position, orientation, and navigation in the virtual space, pseudo-continuously shaped the overall sound of what it was controlling. With the goal of creating more nuance and conveying a story via bodily gestures, it would be more useful to make grand changes in parameters than finely tuned, individualized ones. As such, several of the gestures have been designed to have multiple features or to change parameters in preset iterations that were specifically selected to sound good to reduce choice paralysis and to make each movement more deliberate and significant.

After considering what level of control over vocal harmonies I wanted, I realized that it would be more user-friendly to automatically generate bulk chords, rather than to have individual note control. After hours of experimenting with the previous method of controlling vocal harmonies, I realized it was an incredibly difficult feat to sing while harmonizing two other singers deliberately with your arms, and interacting with the environment in tandem without accidentally moving the harmonies out of sync was nearly

impossible. This adds the ability for more than 2 “musically correct” harmonies to be generated at a time by not trying each individual harmony with the 2 handheld controllers. It does eliminate the ability for a vocalist to sing one root note and actively improvise the harmonies, but the peace of mind and new access to mobility allows for other types of gestural controls to enter the space without inherently interfering with the vocal controls simultaneously.

As such, one of the first switches I made was replacing the individual note selection structure seen in Figure 42 with an automatic harmony generator, viewable in Figure 44. The performer first uses the automatic scale selection system implemented in the keyfinder visualization above and selects whether to automatically generate chords in the major or minor key. Then, the fzero~ fundamental frequency finding object from the keyfinder is routed to generate the respective major or minor third, perfect or diminished 5th, and octave, major 7th, or minor 7th, depending on the incoming scale degree.

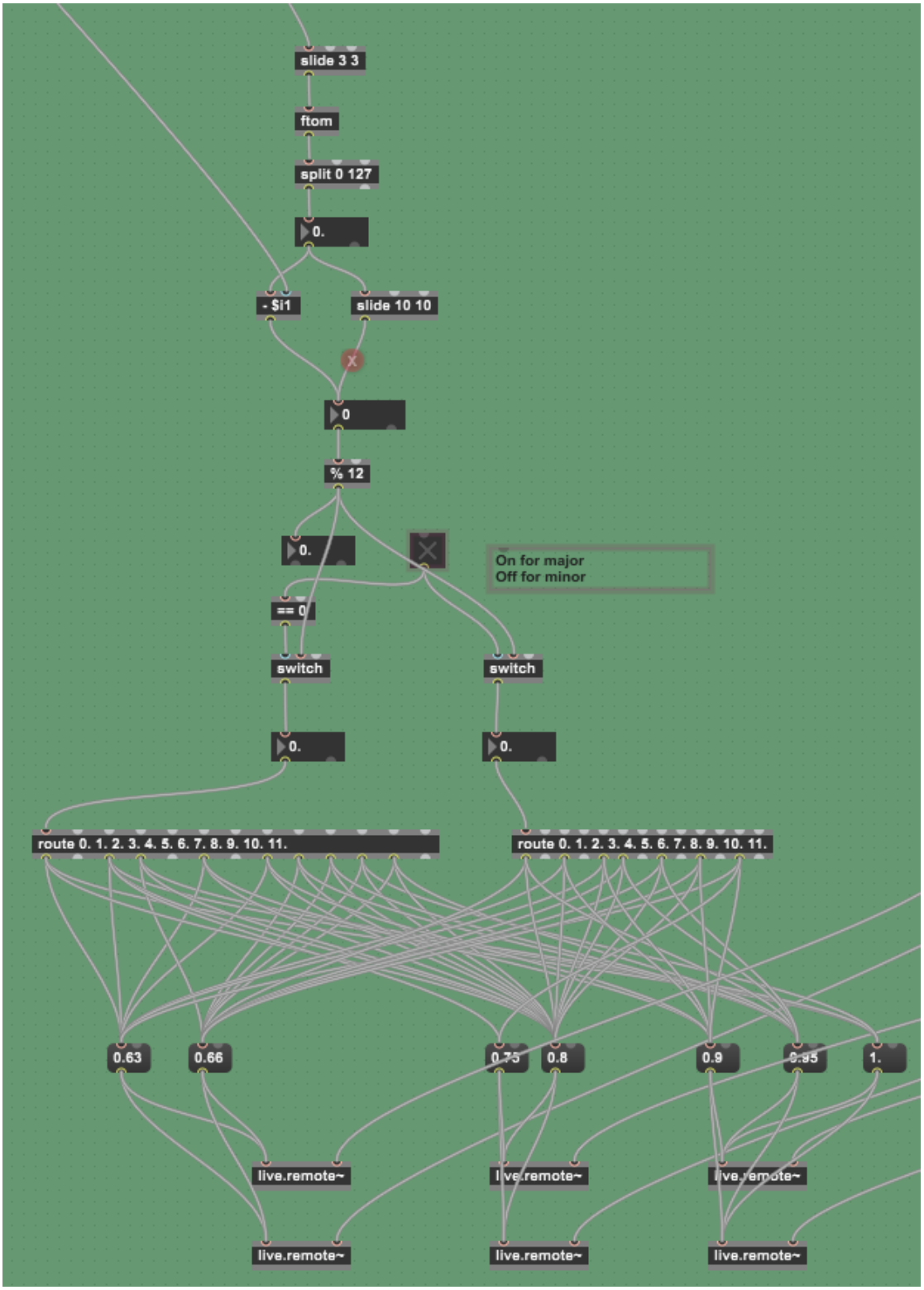


Figure 44: Auto Harmony Generation

Each of these live.remote~ objects controls an instance on one of 3 alternative tracks of the Waves Vocal Bender plugin, seen below in Figures 45 and 46. The tracks labelled "Harmony L/3s", "Harmony R/5s", and "Harmony 7s" are titled to refer to the harmony

being generated on that track, as well as their respective panning to give the auto-generated harmonies a larger sense of space and a stereo effect. These tracks receive audio directly from the "Lead VOX" track post-effects so that they are mixed and auto-tuned to give clear harmonies that are quantized to the same root pitch. Additionally, by being based after the effects chain on the "Lead VOX" channel, additional EQing, compression, and other processing only happens one time, and all that is being calculated on each harmony track is pitch and formant shifting, further reducing CPU usage.

Figure 45: Ableton Live Session View

Figure 46: Waves Vocal Bender Mono Plugin on the 3rd-Generating Track

After deciding to introduce gestural control, I wanted to revisit the Wekinator, which is a simple machine learning tool that classifies inputs in a variety of ways. This program basically acts as a middleman, where it predicts controller movement as a particular gesture you train it on. Using the DTW (dynamic time warping) mode, we can train the neural network on the motion of controllers, rather than just static positions, and create gestural predictions. Essentially, after training on a few examples of each gesture, Wekinator continuously streams a prediction that that gesture was completed in the past few samples of input data. On the bottom is an adjustable threshold slider which lets the user tailor how sensitive the program is in its prediction that a gesture definitively occurred. An instance of Wekinator can be seen below in Figure 47.

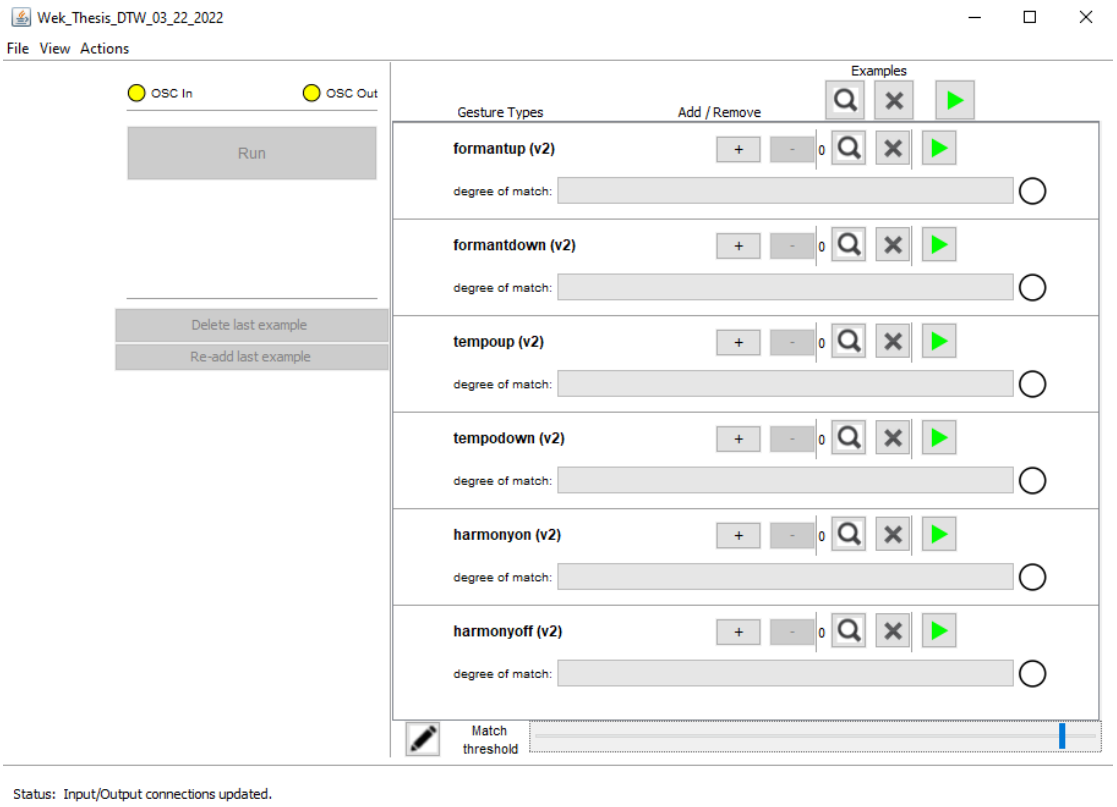


Figure 47: 6 Wekinator Controlled Gestures

I came up with these 6 control gestures by sketching all of the audio effects and functionalities that I wanted to be gesturally controlled and confronting my own assumptions about what those interactions would look like. I wanted many of the controls that previously existed in menu settings to be accessible through motion. For example, adjusting the tempo of the song was previously controlled with a slider on a 2D menu. I have created an interaction where the tempo can be increased by having the user rotate their forearms and controllers away from themselves in front of their body, and decreased by reversing their arms back inward. This action, which resembles the "cha cha real smooth" section of the "Cha Cha Slide" by Mr C the Slide Man, gives a feeling of movement, either progression to a faster pace or reeling the song back into a more relaxed tempo. Additionally, the previous control for adding and removing the backup harmonies were done with a 2D menu button, but I wanted to incorporate it with the expression of "everyone coming together" and "everyone settling down". As such, the motion of rapidly raising both arms triggers the harmonies to engage, and throwing down the arms causes the harmonies to silence.

I also wanted to add a few gestures that were not featured in previous iterations, such as formant shifting or "gender shifting" of the main vocal. At first, I thought about the societally associated body language behind "masculinizing" and "feminizing" the self, but I realized that this was something that I did not want to contribute further to. As a queer individual, there are moments and environments where I want to be seen vastly differently from others and by others, and I wanted to express this perplexity with a gesture that captures my own rapidly changing decisions, acceptance, and unknowance. As such, I decided to use an exaggerated "shrugging" motion, with the raising of formant with one arm higher (currently the left), and the lowering with the other arm (currently the right). There is no reason to have chosen the direction or preference between, but I find this gesture quite natural to the audio effect it is paired with, although at times I think to myself about how much it resembles Steve Martin's King Tut dance.

Additionally, I realized there were certain effects that would be better triggered as a one-off instance, and others that would benefit from being held continuously. For example, I wanted to sync the ability to access speaking mode to bringing the hands close to the mouth, simulating the idea of closing off the voice, whispering, and having a more intimate moment. Once the hands are brought out of tight proximity from the headset, the singing effects chain toggles back on and resumes with the parameters from before. Additionally, I wanted to incorporate the idea of a filter sweep on the master track when the performer stares down at their feet and releases when their gaze heads back to neutral, which I incorporated by looking at the forward rotational parameter in Unity. As such, these controls are navigated with a linear threshold in Max and bypass Wekinator entirely. In Unity, an onboarding screen can be accessed with the "B" button on the left controller, which displays 6 gifs correlating to each of the triggerable interactions. An initial prototype sketch of these motions can be seen below in Figure 48, and Figures 23 and 24 in Section 3 display the first and last frames of the gif respectively.

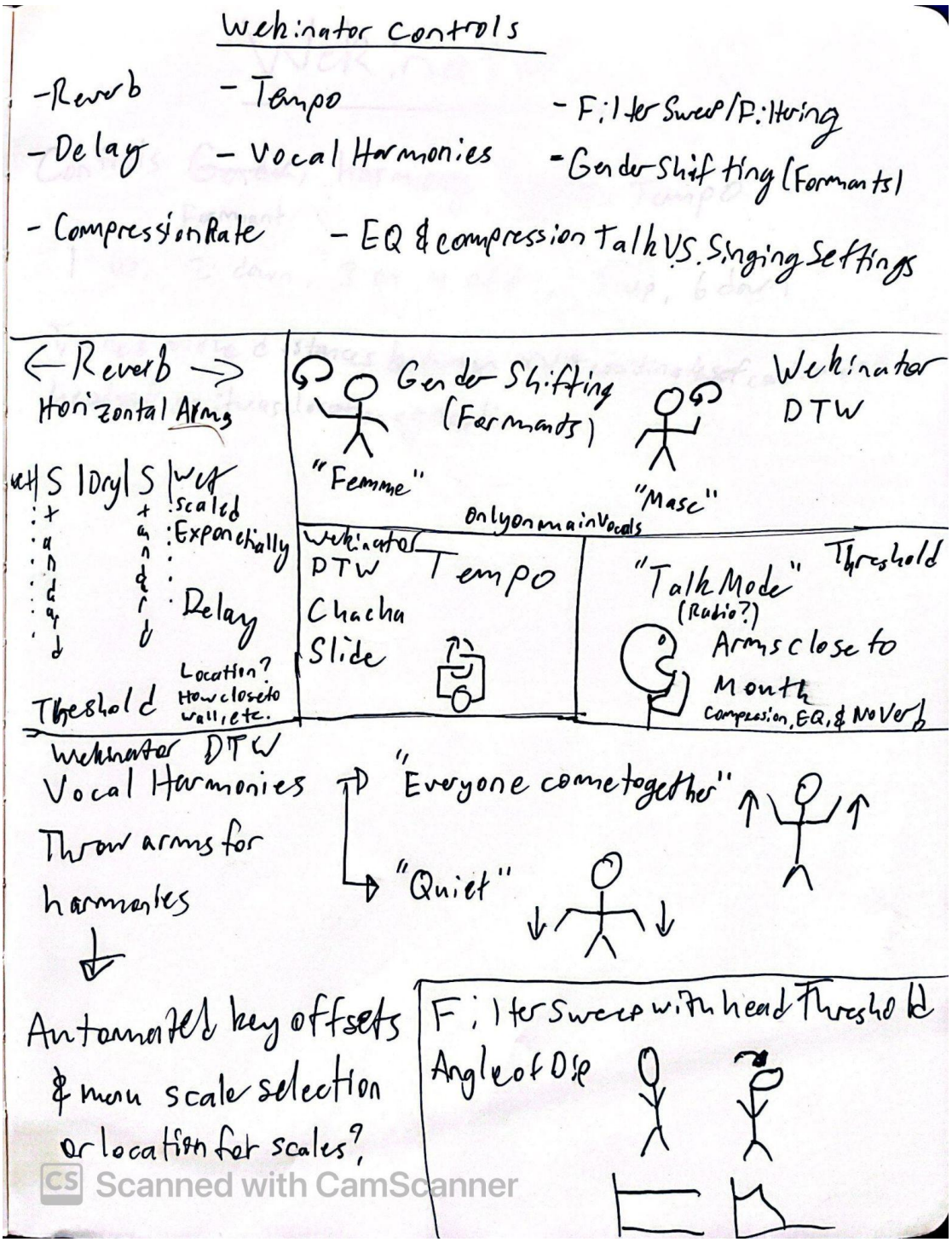


Figure 48: Wekinator and Gestural Control Sketch

Since this environment was primarily centered around the concept of gestures triggering togglable effects or state changes in audio effects, I realized there would be a lot more to keep track of, and wanted to change the UI to reflect this. While singing, one could

hear the transitions of the vocal effects chain in real time. However, if they wanted to implement these changes between lines or while transitioning between segments of a song, there would be no way to tell that it occurred without some external sensory feedback. I implemented a simple text-based representation of the state of each audio effect at the top of the user's field of view. It constantly states whether or not the automatic vocal harmonization is engaged, whether the formant shifting of the main vocal is low, non-active, or high, and the current tempo of the song for frame of reference. The default of this text before receiving updates to the state of the harmony generator, formant shifting, and tempo can be seen below in Figure 49.

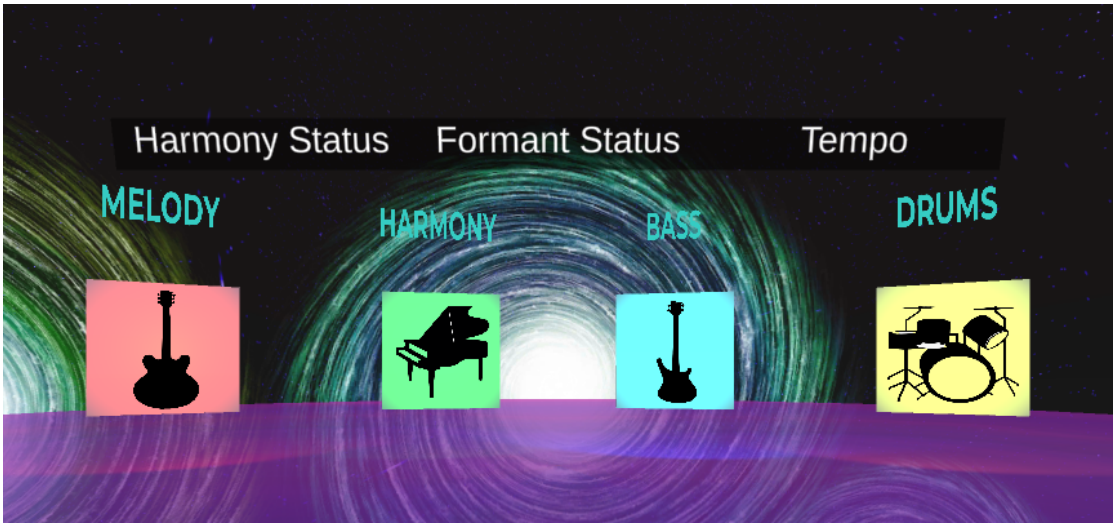


Figure 49: Unity Performance Space with Audio Effect Text GUI

4.2 Ideation through Interviews and Playtesting

4.2.1 Interviews

I interviewed and reached out to several producers and vocalists over the evolution of the project, but the most relevant to the most recent iteration were Bora Yoon, a vocalist and music composition doctoral student at Princeton, and Les Stuck, an immersive sound installation designer at Meow Wolf and studio engineer. I will briefly discuss some of the key takeaways and how they intervened with my design philosophy.

One key phrase that stuck with me throughout was about determining what actually makes for good UI and UX in a design? Bora Yoon noted that in an audio-visual interface,

the key is to induce synesthesia as a form of creating intuition. If the visual design choice stimulates thoughts of a sound, or thoughts of a sound make the user imagine a certain type of interaction or interface, then a successful link has been found. Not only will it be easier to memorize and to get into the flow of using a tool with this sensory syncing, but also the interaction itself will be more meaningful. Additionally, I've thought about gestural control before this meeting, but during this conversation I realized that dancing could be considered a collection of gestures telling a story over time, and that story could conduct music through body expression. This was a pivotal moment in my approach for interaction design.

Instead of thinking of interactions as individualistic, and dance as rhythmic and following the music, I could think of gestures as being theatrical. The nature of grand body movements in theater originated as an accessibility design choice, whereas would help people who were further away from the stage be able to read body language and follow along a story when movements were exaggerated. If we think about theatrical prompts, there are cultural norms in what the body does when attempting to convey certain emotions, and these can be tapped into as a form of emphasis, or in the case of this thesis, as a form of gestural control.

I was also at this time made aware of one of the key audio engineering issues for vocal performers– the way that you EQ and compress speech, rapping, and singing are unique and individualized to the performer. As such, unless the engineer is actively following the performer, generally the audio in a non-rap based performance is set up specifically for singing, and the talking in between or during non-melodic sections of songs are not optimal. I tested gestures to toggle between singing and speaking mode and it worked rather well, but the gestures chosen often conflicted with the gestures for other actions, and I would like to reincorporate it in future iterations.

Les emphasized the idea of incorporating multidimensional, parameterized controllers, and also focusing on the elements of a design process that are not just intellectually/ academically interesting, but that actually have an effect on the listener. Sometimes throughout my design process, I would get lost in what I called "feature hell", where I would get carried away by possibilities and optimizing for nuanced cases, while letting what I thought was actually novel and powerful for everyone involved in the performance flow to the wayside. He used the example of how great art is just one thing that

hits you. Nothing about lying on the beach is complex, yet when kept a rarity, it remains beautiful every time. As such, it would be more valuable to perfect the foundation of my space than to try to build my way out of "feature hell".

Additionally, one note that I found particularly interesting was around how spatialized sound makes VR believable, and even further how the flow of audio currents in a space can create a dreamscape-like environment. Although none of my prototypes incorporated spatialized audio, this did convince me to pursue further use of stereo in several of the prototypes for where the harmonies were sourced from, rather than everything being mono and sourced down the center like the root vocal.

4.2.2 Self Playtesting through Development

The most rapid expansions of growth and directionality throughout my thesis process have been through my own discoveries. Being part of the primary user base for whom this thesis is designed has allowed me to test the affordances of each step as they were being developed to tweak and reevaluate in tandem. The grand shifts in each prototype's landscape were designated after my own personal playtesting experience and newfound comprehension for the affordances that they'd each laid out before me.

For example, with the first prototype, I felt like the amount of control afforded to the player was far too finite and linear– it was a 2D DJ interface laid in a 3D space that was harder to control than a simple mixing board, albeit more visually engaging. The only unique affordances in this landscape were that of the "moons" which controlled audio effects on the master track. These allowed for nonlinear control of audio effects that could be tossed to further roll and bounce around an environment with continuous control of the effects they were synced to. I realized this was the most engaging part of the experience, which is what then led me to the "synth embodiment" design.

Throughout testing the "synth embodiment" segment, there were many decisions that had to be made. I had to figure out how much control I wanted to give the user– did I want to bind them to producing solely semitones in a western scale, or even further bound to specific scale structures? Alternatively, did I want to grant them essentially continuous control over the frequency spectrum so that they could just as easily access notes 37 cents flat of anyone

they'd be collaborating with? There were also decisions needing to be made around the number of voices, how notes were triggered, etc.

After playtesting and playing around with a variety of parameters, I realized that without any UI, there was too much to keep track of in terms of the relationship between the player's headset, controllers, and environment, and as such the design decisions I was making. I played around with each hand's controller modulating a unique voice that could be pitched with lateral movement from the body and filtered standalone. However, to this day I question if my decisions over what movements felt most natural for filtering were based on my actual experience playing around with alternative setups or my preconceived understanding of how graphical EQs are represented and how many interactions with instruments (e.g. pianos and a single string on a guitar) change pitch unidirectionally. Would it be more natural if the summation of outward and upward expansion increased pitch together? In this case, there would be more unique ways to control pitch, but there would hypothetically also be one less interaction available to control other parameters without either layering parameters under the same action or hyperparameterizing ranges that control different parameters rather than totally unique gestures via machine learning.

Before I eventually decided that I needed to make the dynamic switch to a HUD-esque UI centric environment rather than a primarily externally interacted one, I was trying to craft an ideal experience with the limited playground I'd been developing. I had notes triggered semi-randomly on quarter beats with smooth, ambient transitions to chromatic, scale bound tones that could not become out of sync (neither tempo nor key) with the background music playing in the scene. It was indeed quite soothing and engaging, and the trained ear did not struggle to understand what was being controlled by each of their gestures, but it was very limited in its functionality.

The next evolution of the thesis expanded on the need of a simple UI by taking that dance-like control of pitch and combining it with the interview feedback mentioned above to reach into the vocal sphere. From here I realized a plethora of things– the users should know exactly how close they are to thresholds of discrete changes in pitch, and the gestures should not interfere with the singer's ability to interact with a microphone. It seems obvious in retrospect, but when playing around with gestural controls for bringing things close in or

nodding the performer's head with VR glasses on, a dropped mic or even a concussion are just around the corner. In an ideal world the performer would have a wearable microphone around their ear or headset, but this is not always the case in reality. This was where I formally decided to bound users' harmonies to western scale structures, and I generated the visual UI seen in Figures 35 and 36. From here, users could tell how close they were to transitioning their generated vocal harmonies to alternative ratios very clearly. However, after doing extensive playtesting, even in the "simple" mode that only produced up to 3 potential notes per arm, it was not obvious or intuitive how to best use or "play" this interaction schema. I found again that this interaction in combination with a significant amount of lag built up through the pipeline of resampling vocals for harmony generation with the method I was using at the time was best for slow, drone-like music. It was much more peaceful than previous iterations, and allowed the user to be much more intentional with the music they were creating. However, it was almost too much creative freedom, in that the alignment of physical notes on either side required not only a solid knowledgebank of chord theory (or experience with the program), but also multiligament control, whereas the mouth and both hands must be moving minutely with immense accuracy that scale in difficulty depending on the amount of notes the user selected to have access to, as well as independently of one another to fully access the space. Although this monitoring process could've been improved with features like a haptic tick when each arm crossed the threshold into another note being played, I realized it would still require too much focus to be taken away from the meat of the performance– the artist's main vocal and expression surrounding such.

This led to the final shift of gestural controls and automatically generated harmonies. This way, the user could have full access to notes that were bound to be correct and in line with what they were trying to produce, without needing to focus on the physical technicalities of making sure they don't make a mistake. When playing around with this space, it became clear that this was much more engaging and enjoyable, as you could hypothetically hop in, turn on auto-harmonies, and nothing would sound horrible. There were some nuances of this interface schema as well, but more mathematically than anything else. One of the keys was that Wekinator, the program used to recognize the transition to gestures over time, seemed to struggle to train properly when the player was oriented in different directions. When I trained and tested Wekinator while standing in the center of the virtual space and facing the panels

that control each track's volume, the formant and harmony controlling gestures were accurately recognized over 95% of the time. However, whether tested on the forward-trained only or given a training set with dozens of angles worth of samples, it was difficult for Wekinator to parse which controls were being input. Additionally, I mentioned that the formant and harmony controlling gestures were successful, but the minute distinction in direction between the tempo increasing and decreasing gestures made it difficult for Wekinator to parse them. All in all, my experience interacting, discovering, and growing with my thesis allowed me to be confident in presenting it to other users to test.

4.2.3 External Playtesting

First I will discuss the setup for what and how I tested for the successes and weaknesses of my project, and then I will discuss the findings from this process.

The test will be a performance by singers and laypeople using songs that they are familiar with. I originally planned on using a stem-splitting website to separate the chords, bass, drums, melodies, and vocals for any song they please or use their own stems if available and map them into a compatible Ableton project, and would like to revisit this idea in the future. However, for the sake of increasing playtesting efficiency, I produced a track that uses the most common four chord pattern in pop music (the I V vi IV progression) that could be adapted to the lyrics of any of a plethora of songs that participants would be familiar with, and I dragged their preferred lyric set onto its respective screen.

A few participants in the early phase were testing in my apartment, but in general over a dozen users tested in 15 minute increments with a 15 minute Q&A follow up in the XR Lab (Room 248 at 370 Jay Street) over the course of two weeks. The users were hooked up to an Oculus Rift while I held a Shure SM58 and wires to keep them untangled and out of the way. The last few playtesters were able to use a wireless lavalier microphone that did not obstruct any of their motions and more consistently held their volumes being fed into Ableton Live. The playtesters were from a variety of different backgrounds, but the majority were either Integrated Digital Media, Interactive Telecommunication Program, or Music Technology Masters students.

There are a few key questions I asked to qualitatively and quantitatively evaluate the success of the design. I will list the question and then discuss a summation of answers.

First, I monitored and asked if users were able to actively and deliberately control their audio with each of the effects in place (formant shifting, tempo shifting, and harmony generation), as well as if each of these effects were noticeable or overwhelming. In general, people were able to access the effects successfully, with the gesture of lowering the tempo being a little harder to access because of the similarity in motion that the dynamic time warping algorithm had to distinguish from speeding up the tempo. As such, I retrained the tempo adjustments to require additional rotations of the forearms to activate, which gives Wekinator enough time and information to better parse speeding up from slowing down. Additionally, people seemed to love the formant shifting gesture, but I received a lot of feedback that people didn't know what formants were before playing around with it. One key point of feedback that I'm working on implementing is a better system for the UI to be interpreted both faster (non text based) and more intuitively (not requiring users to know specific terminology to understand what level a feature is at).

I asked users if the interactions were meaningful and easy to do in the middle of a performance. One of the most prominent responses to this was that the gestures were easy to activate and people quickly absorbed the motions and what linked to what audio effect. However, if they were not paying close attention to the sound of their voice or actively watching the GUI text bar, they would not know whether an effect was activated or not. A common recommendation was to have some sort of visual paired to more actions like in earlier iterations of the thesis– for example, a confetti particle effect explosion when the tempo increases, or dancing 3D models of the instruments when they're playing and the spotlights turning off when they're disabled rather than simply turning on and off a light over a 2D plane.

I asked people if they were engaged in the experience, and if the movements were theatrical. People seemed to really like the interactions, but often noted that the space felt empty. I also had a lot of users who wanted to try to dance, but they noted that it was impossible to dance with large gestures without accidentally triggering other audio effects. As such, I am working on using a button potentially either on the right controller or embedded in the environment to quickly toggle a locking mechanism to the audio effects that are in place. Additionally, a lot of feedback surrounded the fact that the environment is pretty

much empty other than the interactables, and it is essentially static and non responsive as well. This was a conscious decision in order to create minimal distractions and to allow for easy modularity for different genres without instilling too much of a bias as to how to perform, but I am currently expanding on creating unique feedback for triggering different effects and keeping people aligned in the space to create a more engaging environment.

Finally, I asked for general feedback on anything else that was not previously covered. One of the main points participants brought up here was that the onboarding UI, at least in a demo setting, should be embedded into the environment so that it wasn't constantly blocking the interactables but could still be glanced at easily in the middle of a performance. Additionally, currently the floor of the environment is semi transparent, and this in combination with the unfilled space in the environment might be disorienting to the players. I am in the process of integrating both of these by putting the performer on a stand with a set of tools ahead of them that orients them towards the instrument controlling panels and allows for these visuals to constantly be on display, while giving them a sense of place in the environment. Lastly, I received feedback from several people throughout this process to adapt it into an alternative product as an advanced karaoke machine, and I am in the process of building this as well.

5 Conclusion and Future Work

The iterative design process throughout developing this project has opened my eyes to where interaction design in virtual reality is heading. From my user testing, it became clear that especially for individuals who had limited experience with virtual reality before, it was often easier to use gestures to communicate actions than their digital counterparts. I envision that this will be kept to rather infrequent but crucial actions, or the gestures will become unique and discrete but more minimal than what I've prescribed so as to not tire users more.

I also can imagine the collaborative nature of these tools in performative audio spaces. For example, I was able to connect a second set of headphones to the audio interface so that either I or other user testers could hear each other perform, and oftentimes the secondary person connected would want to dance and sing along and engage in the action. The participatory nature of music is innate, and the gestural meaning behind dance can and will

purvey itself well into the future of interaction design in virtual environments. I think it would be interesting to explore not only expanding on the vocabulary of potential gestures (and cleaning up their consistency, etc), but also introducing multi-user gestural control. This could be perceived as both allowing multiple users to input the same bank of gestures, as well as tag-teamed gestures where performers dance together, either having their avatars make contact and interact or by inputting paired gestures simultaneously to create a magnificent, syncopated effect.

In the short term, coming from user feedback and the impact of my own experiences interacting with the space, I will significantly improve on the visuals of the space. This will be a combination of feedback when interacting with unique tools, a revamped UI to give the user a cleaner and more concise sense of space and control over their affordances, and also just general aesthetic improvements on the space to orient the player and make it a more pleasant environment to engage with.

In one future iteration of the design schema, I would like to tap into what Les Stuck noted in his interview. He mentioned that one of the key reasons for the success of the Meow Wolf project, an immersive art collective and set of exhibitions, was the combination of collaborative interactions between audience members, and the freedom of movement throughout the space. One of the key affordances of VR that I did not tap into is the limitlessness of virtual space. From the perspective of a performing user, the idea of menus and setting changes could potentially occur by navigating into different rooms and dragging the audience along with them. However, I think one of the real tickets to success is in the incorporation of both collaborative and audience-baseds interaction designs in a virtual space like this. As a singing performance tool, the majority of the decisions of how the audio space is routed is from the perspective of the performer. However, it would be an incredible use of the platform to give some degrees of agency to the audience in ways that would not ruin the experience for the performer or other audience members. Perhaps this could be through audience members being able to control the spatialization of vocal harmonies. Throughout human history, sound has been shaped by the environments that they are bound to (e.g. the design of cathedral halls for massive reverb and the symbiotic relationship with how religious worship music is shaped with slow vocal changes). I would love to see how the limitless bounds of virtual vocal environments help the music being performed evolve as well.

I have thought a lot about simulating the movement of a space as a means of modal control. For example, a user could roam from room to room, where each would have an entirely different set of preset values for each audio effect, as well as the visual effects in the space. This would minimize, or even eliminate, the need for 2D menu usage entirely, would cause the visuals of the spaces to be more individually tailored for each song or segment of a song, and could potentially be used to give audiences different amounts of control over the performance at different times depending on an artist's needs.

I played around with visual feedback throughout this project, but there is plenty more room to explore. The UI in the manually controlled singer space is reactive and streams the scale degrees being modulated in real time back as a visual cue, but many of the iterations are lacking in visual feedback. In the current iteration, it would be rather simple to have some sort of visualization the moment that the program recognizes that a harmony or tempo change is registered. I would love to explore what an optimized HUD with maximum information and minimal interference would be. While prototyping I would print coordinate information on the screen, but it was just to make sure that I was aligned properly, and in use it would surely do more harm than good. However, there should be more discreet ways for a performer to verify their changes and keep track of where they are in a set than by testing directly with the microphone.

Another feature I wanted to consider was whether or not the microphone could be a tool in and of itself. Both Luke DuBois and Bora Yoon discussed moving to a standalone mic, which would not hinder the performer from standing stationary either on stage or in a studio environment. I have not been able to acquire one, but I would love to continue testing and see what I can make with the ability to move more freely in my space and how I can blend the usage of multiple microphones at the same time. If the controller was held, there could be proximity and angle sensors added to it either as additional inputs for control or to supplement one of the handheld VR controllers. Alternatively, the actual mic input could be registered (e.g. crossing certain gain thresholds or using speech recognition to trigger cues in the effect processing).

I would also love to consider more intimate control of spatialized panning in a virtual space. Mentioned before was the ability for audience members to control the location of the

audio source of each vocal harmony, but they could also be gesturally conducted to sweep around audiences in their environment (whether it be virtual or a physical space translated from their virtual performance). This could have profound effects on conveying real time narratives through song in a more elegant and engaging way. It would be interesting to see how creatively performers can paint with their words if they could literally relocate their sources in 3D space as if they were painting.

Additionally, I would love to adapt this to both incorporate looping and recording ability, as well as to a larger set of instruments. Looping would give artists the ability to sing exactly what harmonies they want (albeing not in real time) without dealing with any novel interface to play said instruments. It would also allow for performers like beatboxers and acapella artists/ groups to create amazing pieces with this virtual structure. However, currently Max4Live mainly has functionality to control the session view (more improvisational based) of Ableton Live, and there are very few resources for creating this in-DAW tool, although I believe it is possible. Also, although the affordance of free hands in the singing space is crucial for my iteration, it is not impossible to imagine designated MIDI keys on a keyboard or drum kit to trigger different effects in a space, or if and when AR is incorporated, to have the performer be able to use gestural controls to add effects to their keys. One thing mentioned in the interviews I conducted was surrounding the introduction of other physical hardware-based interfaces. The ability to toggle between scenes and presets could be synced to a foot pedal on the ground, or even to a wearable or app for easy access even with a layer of virtual reality between the performer and said interface extension.

In my own future iteration of this space I would love to translate the interface to Unreal Engine. There are more libraries and more easily accessible collaboration tools for expanding this further, which would improve the developmental process greatly. I attempted to make the switch, but was consistently having issues with OSC in their blueprint system crashing Unreal Engine and corrupting the entire project, which hindered my ability to even reference previous blueprints to be remade. With more time, I think this would be a worthwhile transition.

One possible adaptation that I would like to continue working on is as a karaoke tool. From the top of the design process all the way through user testing, people noted that this would make an excellent adaptation to these tools and environments. In most karaoke spaces, the ability for the performer to adjust how their voice is amplified is generally fully restricted,

or with volume control at most. I think the ability to produce real time vocal harmonies, to gesture and be able to change their formant, to be able to see the lyrics no matter where they are facing could create an enjoyable and much more intricate experience that would be engaging enough to become a viable product. Of course, often in karaoke bars, etc, the "customers", or user-base, are often inebriated and do not often have a musical background, so there would need to be a lot of automation using these tools, but that could have modular levels of control. This karaoke system could have presets and plugins baked into the program and run standalone, so people could sing along with audio generated and transferred solely using a standalone headset no matter where they were at. This could also be an incredible songwriting tool, as it would allow artists to freely sing along and try to control their own songs' harmonies and vocal effect chain anywhere, any time.

As for myself, I hope to keep innovating on tools that allow people to express themselves in ways they never thought possible. More specifically, this has been an amazing exploration of some possible use cases of XR to control audio, and it opened up a world of opportunities that could be fleshed out into entire performances, exhibitions, and DAWs. I would love to make music production and performances more engaging and accessible by integrating gestural tools into a plethora of pre existing workflows.

As for the field at large, I have a vision where performers would be able to make careers using VR music platforms as their primary medium, and also envision augmented reality performance tools being integrated with classic concert sets in ways that people have never explored. This includes everything from unique spatialized visualizations to real time live individualized audience participation affecting everyone's unique audio visual perceptions of performances.

We are entering a new age of performance, where audio and visual experiences can be synced more intimately, meaning can be conveyed more deeply, and people can connect from further than ever before, and I am both grateful and excited to be a part of it.

6 References and Resources

6.1 References

Klemmer, S., Hartmann, S., & Takayama, L. (2006). How Bodies Matter: Five Themes for Interaction Design. *DIS '06: Proceedings of the 6th conference on Designing Interactive systems*.

https://dl.acm.org/doi/10.1145/1142405.1142429

Lewis, G. E. (2000). Too Many Notes: Computers, Complexity and Culture in "Voyager." *Leonardo Music Journal*, *10*, 33–39.

http://www.jstor.org/stable/1513376

Hermann, T., Hunt, A., & Neuhoff, J. G. (2011). *The sonification handbook*. Berlin, Germany: Logos Publishing House.

https://sonification.de/handbook/download/TheSonificationHandbook-HermannHunt Neuhoff-2011.pdf

Neely, J. (2019). *Soma literate design.* Pittsburgh, Pennsylvania: Carnegie Mellon University Press.

DOI: 10.13140/RG.2.2.17690.88007

Sennett, R. (2008). *The craftsman*. New Haven, CT: Yale University Press.

DOI: https://doi.org/10.1086/605737

Schon, D. *Designing as Reflective Conversation*.

https://doi.org/10.1016/0950-7051(92)90020-G

McCullough, M. (1998). *Abstracting Craft: The Practiced Digital Hand*. MIT Press.

https://dl.acm.org/doi/10.5555/524384

Cross, N. (1982). *Designerly Ways of Knowing*. *Design Studies, Volume 3 Issue 4, 40-97*.

https://doi.org/10.1016/0142-694X(82)90040-0

6.2 External Resources

Atiagina, A. and Sanders, S. “The History of Synths”. *historyofsynths.com*. 2017.

https://www.historyofsynths.com/

Intructory Demo Video: https://youtu.be/hquJ52XpiNU

My Website: www.bencrystal.me

My Github: https://github.com/bencrystal
