MIDI.CITI: Designing an Experience-oriented Musical Cityscape

Kunwoo Kim
CCRMA, Stanford University
kunwoo@ccrma.stanford.edu

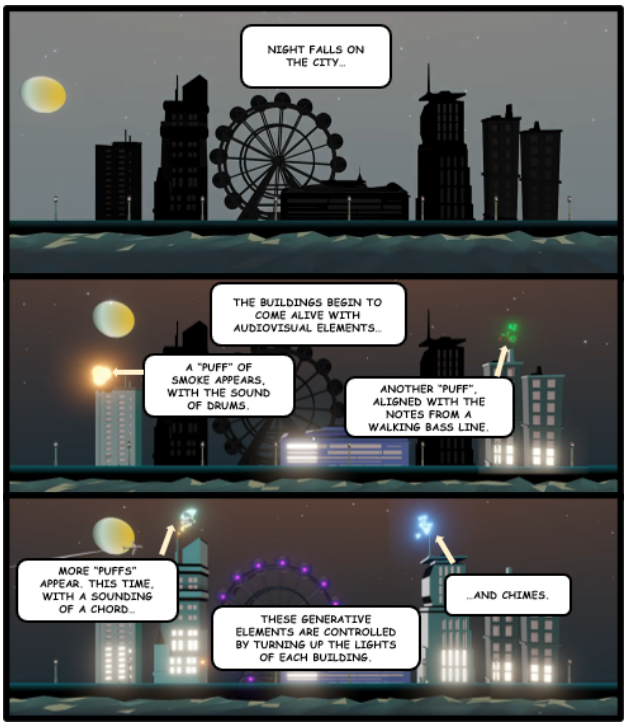
Ge Wang
CCRMA, Stanford University
ge@ccrma.stanford.edu

Figure 1: MIDI.CITI - a screenshot of the audiovisual experience in action.

ABSTRACT

MIDI.CITI is an interactive audiovisual musical sandbox that offers room for experiential narrative, playfulness, and expression. It contains a real-time algorithmically generated drum machine mapped onto a metaphorical cityscape environment. This paper unpacks the design of MIDI.CITI through the lenses of interaction, play, and the idea of designing tools "inside-out", i.e., designing from an intended aesthetic experience. We describe its design process and elucidate its underlying interactions and audio algorithms. We also introduce MIDI.CITI.VR, a translation of MIDI.CITI from desktop to a virtual reality medium. Lastly, we describe some underlying design principles that encompass prioritizing the experience, adopting an audio-first approach for tightly-coupled audiovisual correspondence, taking advantage of real-time generative audio, finding a balance between high-level and direct control, and aligning these elements to a fluid narrative. Through these discussions, we aim to provide "things to think with" for creating experience-oriented interactive audiovisual software.

1. INTRODUCTION



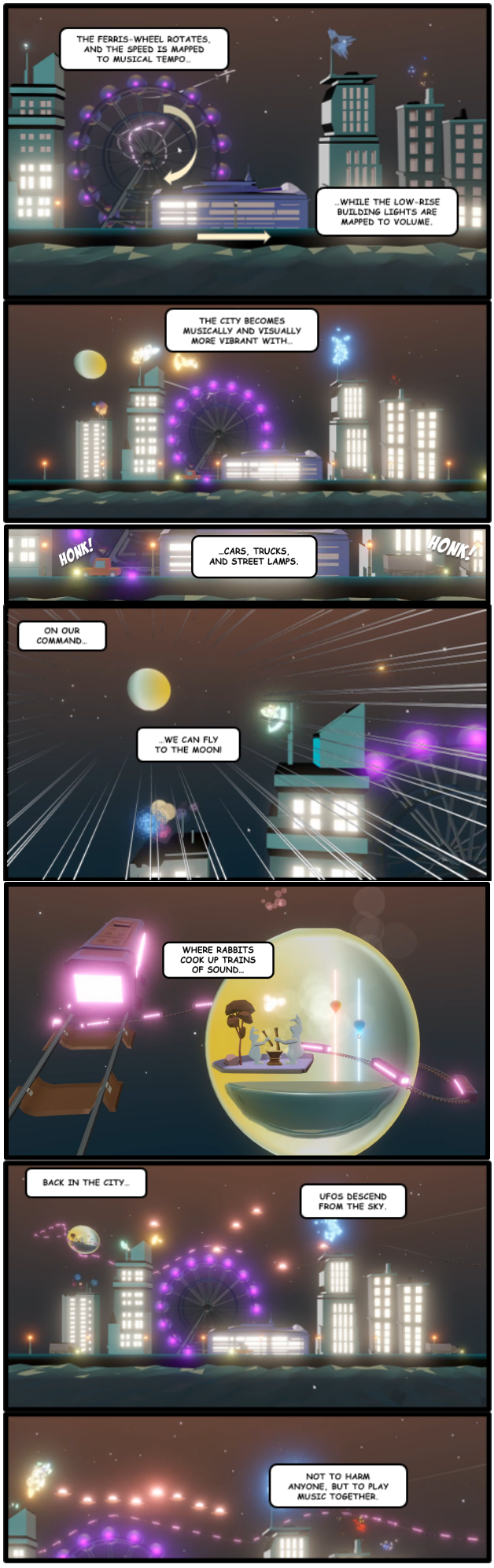


Figure 2. One possible narrative path of MIDI.CITI. The texts in the thought bubbles were added in this paper to describe the narrative. For a demonstation video of MIDI.CITI in action, see https://vimeo.com/794287322.

MIDI.CITI is an interactive musical sandbox, designed for users to have a playful and expressive experience as they experiment with musical ideas and visual metaphors embedded within a cityscape interface. Despite its name, the software has no affiliation with MIDI; it was simply chosen for its rhyme with "city". Instead, MIDI.CITI uses ChucK [1] for generative audio, and is interconnected with graphical and interactive elements using Chunity [2].

Programmable algorithmic processes in MIDI.CITI produce automated or semi-automated musical events, which are tightly coupled with the visuals. Using Chunity's strongly-timed mechanisms, audio generates at a much faster rate than graphics, while precisely governs the timing for the entire system. The synergy of the generative audio, precisely coupled graphics, and interactive elements curate an expressive sandbox that provides a unique combined experience.

MIDI.CITI adopts many of the lenses and design principles of Artful Design [3], aiming to build tools that attend to aesthetic considerations. This translates to critical questions in the design of MIDI.CITI: How might we effectively combine sound and visuals to create an expressive audiovisual experience? How can we meaningfully connect generative audio and interaction? How can we make the best use of these underlying mechanics and dynamics to foster playfulness?

In this paper, we unpack the design of MIDI.CITI, its underlying audio algorithms, and a few relevant design lenses. Next, we examine the re-design of MIDI.CITI in the VR medium and highlight the differences in these respective mediums. Lastly, we propose some design principles for creating experience-oriented audiovisual tools.

2. MIDI.CITI

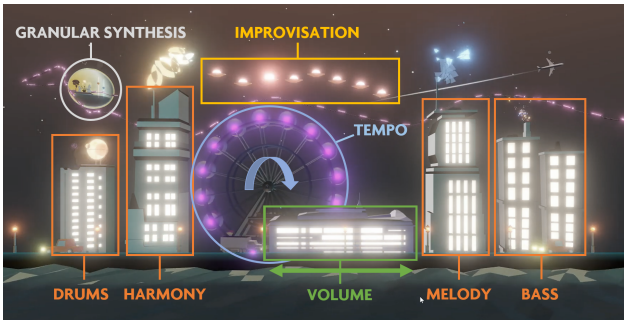


Figure 3. Audiovisual mapping of the city.

Each urban element in MIDI.CITI represents an instrument or an audio parameter such as tempo and volume. The overall produced music is a sequenced audio loop, algorithmically programmed in ChucK. This algorithm introduces pseudo-randomization of rhythms and pitches, ensuring sustained engagement across repeated iterations. In tandem with the musical development, the city becomes more vibrant with various visual correspondence, as well as the addition of cars, planes, trains, and even the Moon and UFOs. The rhythm, harmony, and meter are readily modifiable via the ChucK script, facilitating customization and experimentation within the musical framework.

2.1 The Four high-rise buildings

Users can control the rhythmic complexity of the four instruments—drums, harmony, melody, and bass—by raising or lowering the amount of window lights in corresponding high-rise buildings (orange in Figure 3). For instance, to intensify drum beats while minimizing melody lines, users can increase the number of drum building's window lights while decreasing that of the melody building's. Moreover, the overarching music, structured as an audio loop (e.g., 8 measures, each comprising 12 subdivided beats), maintains uniqueness in each iteration through a pseudo-randomized algorithm.

The adjustment of rhythmic complexity is facilitated through the density parameter. First, each instrument is assigned a "trigger array," where each element denotes one of three states: 0 ("do not play"), 1 ("maybe play"), and 2 ("play"), corresponding to the beat subdivision. Second, the density parameter ranges from 0.0 to 1.0, with unlit windows representing 0.0 density and all lit windows indicating 1.0 density. When the trigger array element is set to 1 ("maybe play"), the density parameter effectively serves as the probability of audio playback; the script employs a random number generator between 0.0 to 1.0. If the current density parameter exceeds the result, playback is allowed for that subdivided beat. Thus, while 0s and 2s establish a fundamental rhythmic structure for an instrument, 1s introduce interesting rhythmic nuances to the sequence.

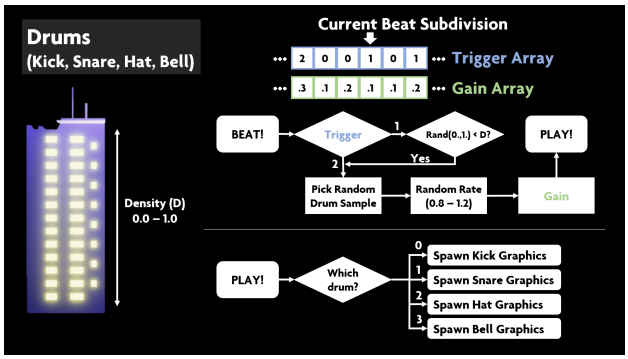


Figure 4. The algorithmic flow of sound playback for drums.

The drums consist of kick, snare, hat, and bell, each with a preassigned trigger array (Figure 4). Additionally, drum instruments feature a gain array, which dictates the volume level of playback for each beat subdivision, effectively differentiating between stronger and weaker beats. Upon receiving a playback signal, the algorithm initiates additional randomization procedures to maintain each beat fresh: it randomly selects an audio sample from a timbre pool and assigns a randomized playback rate ranging from 0.8 to 1.2.

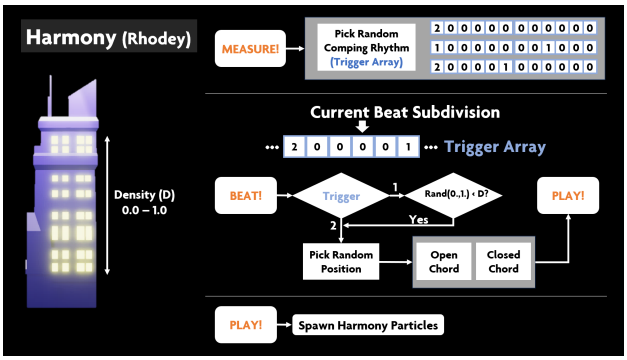


Figure 5. The algorithmic flow of sound playback for harmony.

The harmony consists of four Rhodey unit generators from ChucK, which establish a four-part chord structure (Figure 5). At the beginning of each measure, it selects one of three trigger array patterns for the measure to maintain rhythmic diversity. With each playback, it generates a chord in either open or closed position, ensuring the accompaniment remains engaging.

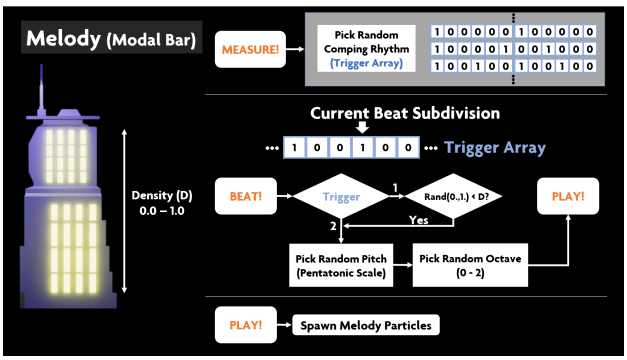


Figure 6. The algorithmic flow of sound playback for melody.

The melody is synthesized using the Modal Bar unit generator from ChucK. At the start of each measure, it selects one of five trigger array patterns (Figure 6). Unlike other instruments, the melody exhibits more random rhythmic patterns as its trigger array elements only consist of 0s ("do not play") and 1s ("maybe play"), with no 2s ("play"). During each playback, the algorithm selects a pitch from a pentatonic scale spanning across three octaves, adapting to the current harmonic context.

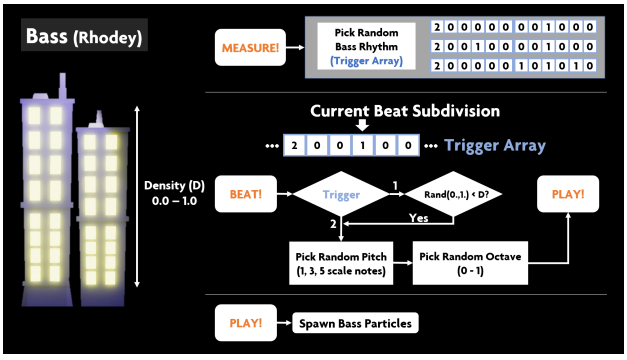


Figure 7. The algorithmic flow of sound playback for bass.

The bass is produced using the Rhodey unit generator from ChucK, employing three different trigger array patterns per measure (Figure 7). Unlike the melody, the bass introduces least amount of randomness to uphold the basic structure of the music. Specifically, the first beat of each measure is always played (i.e., trigger array element of 2), while its pitch is always the root of the current harmonic context. For subsequent subdivided beats, the pitch is chosen from scale degrees of 1, 3, or 5.

2.2 The Ferris wheel and the low-rise building

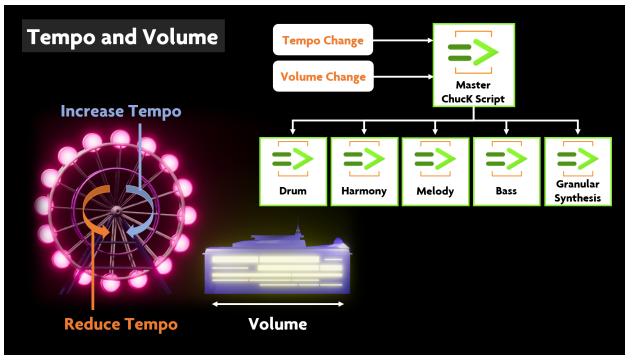


Figure 8. The algorithmic flow of adjusting tempo and volume.

Users can rotate the Ferris wheel using their mouse: clockwise to make it go faster and counter-clockwise to make it go slower. This speed of rotation corresponds to the tempo of the whole musical system by changing the beats per minute (bpm) parameter within the master ChucK script that sends global events to each generative instrument process (Figure 8). The tempo can be subtly influenced by rotating the mouse cursor to reinforce or dampen the rotation of the Ferris wheel.

In addition, the users can adjust the amount of window lights of the low-rise building, thereby adjusting the overall volume of the music. This adjustment is facilitated by the master ChucK script as well (Figure 8).

2.3 The Moon

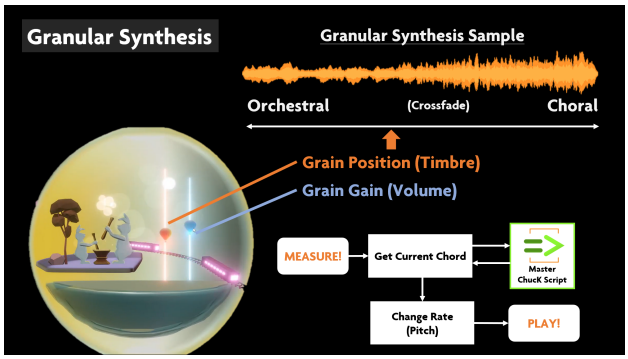


Figure 9. The algorithmic flow of the granular synthesis element.

Upon flying to the Moon, it is revealed that two rabbits occupy the Moon's dark side, inspired by the East Asian folklore of Moon rabbits. As they pound their mortals and pestle to the underlying rhythm, a granular synthesis train (i.e. grain train or "tram-ular" synthesis) departs from the Moon back to the city, a musical and symbolic link between the fantastical and the mundane ebb and flow of urban life.

The two air balloons on the Moon each control granular synthesis position and volume of the atmospheric soundscape, consisting of four voices in different pitch (Figure 9). The sample used for granular synthesis is a cross-faded interpolation between orchestral and choral sounds, offering a continuous timbre spectrum. The volume controls the gain of the projected soundscape.

2.4 The UFOs

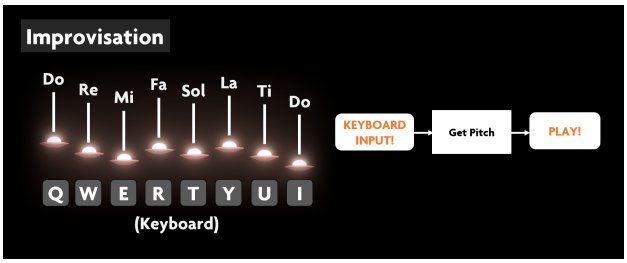


Figure 10. The algorithmic flow of the improvisational element.

Lastly, UFOs can be called to appear over the skyline, not to harm anyone, but to play music together. They are directly playable as a diatonic instrument, using the computer keyboard (Figure 10). This is the only non-algorithmic element in MIDI.CITI, intended for the users to directly improvise different melodies over the algorithmically generated musicscape.

3. DESIGN LENSES

The design of MIDI.CITI proceeded from three main lenses: interaction, play, and experience.

3.1 Interaction

The user's interaction sought a balance between automatic generation and human control. Principle 5.5 from artful design states, "Have your machine learning – and the human in the loop" [3]. MIDI.CITI uses algorithmically generative audio for its beats, chords, and melodies, but offers a number of simple, yet expressive, metaphorical "knobs" to the user. For example, increasing the density parameter of the building associated with melody introduces additional variations in pitch, rhythm, and dynamics. Similarly, rotating the Ferris Wheel accelerates its rotation speed in tandem with the musical tempo. Together, these simple interactions present many possible recombinations, where the user can create a narrative using the resulting sonic textures and densities. The system aims to provide a complex algorithmic generation with intentionally simple and combinable user control. This follows Cook's principle of "instant music, subtlety later" [4], where each interactive city element has one interaction that allows for instant change in one sonic-related parameter, yet combinations of elements provide room for subtle control over the holistic musical outcome.

3.2 Play

Callois states that play is a free, voluntary, uncertain, and unproductive activity, a source of joy and amusement [5]. MIDI.CITI embodies the paidia aspect of play, described as "more free-form, expressive, improvisational, recombination of behaviors and meanings" [3]. MIDI.CITI offers iterations of different narratives based on the user's choice by controlling various parameters. For example, the densities of different instruments can be maxed out with a fast tempo to achieve vibrancy or they can be minimized with a slower tempo to achieve serenity. The users of MIDI.CITI experience open-form aesthetics built on the dynamics of interaction and underlying mechanics of audio algorithms that drive the corresponding graphical output [6].

3.3 Experience

The third lens is through the idea of designing "inside-out" from an intended experience. The idea expands from artful design's Principle 1.15, "Design not only from needs – but from the values behind them" and Principle 2.2, "Design inside-out" [3]. MIDI.CITI has a function of an audiovisual generative drum machine, but the design was executed with the highest priority on human values such as evoked emotion, playfulness, and expression. As a result, MIDI.CITI aims to follow artful design's principle 1.16, "Design is the radical synthesis of means and ends into a third type of a thing – both useful and beautiful" [3], where complex drum machine algorithms, audiovisual correspondence, and instant, yet subtle interactions are in service to the paramount value of experience.

MIDI.CITI encapsulates its audiovisual interactions within the overarching metaphor of a cityscape, thereby giving form to the experience. As one user reflected, it is "a very musical and expressive audiovisual experience that makes me think about the world we live in and how we interact with it." Another user remarked, "I am always left wanting to explore this space more deeply, to go to MIDI.CITI myself and take part in the musical lives of its inhabitants." Experience-oriented design, with metaphorically aligned audiovisual interactions, may evoke profound reflections and desire for deeper explorations.

4. MIDI.CITI IN VR

Figure 11. MIDI.CITI.VR - the city

MIDI.CITI was later translated from the desktop to a virtual reality (VR) medium, called MIDI.CITI.VR (Figure 11). While aiming to preserve the three main design lenses of interaction, play, and experience of its predecessor, MIDI.CITI.VR underwent strategic redesigns of its underlying mechanics to leverage VR's distinctive affordances, including immersion and heightened sense of presence.

First, the VR medium centers around the subject experience, enabling the users a free-form exploration of the virtual environment [7]. Therefore, all the elements in the city–ground surface, roads, buildings, railroad, trains, cars, and planes–were repositioned to surround the user 360°.

Second, audio spatialization is integral in VR by enhancing sense of presence while mitigating potential disorientation [8]. Each interactable audio source is localized based on user's position and head orientation by utilizing Unity's audio spatialization based on volume and angle, further enriching the immersive experience.

Third, the users in the VR medium can be susceptible to motion sickness from the inconsistency between the visual and vestibular information, especially when the medium implies motion [9]. During the design process, we noticed such discrepancies when flying toward the Moon at a high velocity. Consequently, we reimagined the Moon's location, positioning it atop the ocean's surface. Here, users embark on a journey to the Moon aboard an origami boat, characterized by a serene and leisurely pace (Figure 12).

Notably, although some audiovisual interactive elements and the overall experiences have been customized for VR, the foundational audio algorithms and pseudo-randomization techniques remain consistent. This may highlight the adaptability of MIDI.CITI's system across different audiovisual mediums.

Figure 12. MIDI.CITI.VR - the Moon

5. DESIGN PRINCIPLES

Atherton ruminated that design principles are helpful as they are "aphorisms intended to be consulted during the design process in moments of ambiguity" [10]. Here, we present six design principles embodied by MIDI.CITI, as "aphorisms" and "things to think about" in designing experience-oriented and music-driven audiovisual instruments and toys.

5.1 Don't forget the Experience

Prioritizing experience as the "North Star" of the design is central to value-based design approaches like artful design. This often translates to revisiting a number of questions throughout the design process. What is the end experience? How do we want a person to feel within and as a consequence of the experience? What kinds of emotions would we want to evoke? How does the system build

toward that experience? What unique qualities does the medium offer to shape the user's experience? How do we invite user participation for play and creative expression?

The metaphor of a "musical city" provided the framing for MIDI.CITI, where users engage in an experience that is musically creative, playful, whimsical, and expressive, yet uncomplicated and calm. This set of experiential "North Stars" guided the design choices on not only sonic, graphical, and interactive elements, but also the overall balance, narrative, flow, and presentation.

Moreover, MIDI.CITI was not crafted from a top-down pre-determined blueprint, but through a series of improvisatory design choices that reinforces the intended experiential outcomes. In other words, the design may be initially inspired by its "North Star": a starter idea along with a set of experiential priorities. After a few design iterations, however, the design-in-progress itself becomes the source of inspiration, leading to specific design questions and concrete choices that ultimately becomes the substance of the experience: What if the buildings represented the four instruments in the drum machine? What if parameters were controlled by window lights? What if a Ferris wheel controlled the tempo? What if we could fly to the Moon? etc. All the while, the "North Star" remains as a set of experiential goals to evaluate each design choice.

5.2 Achieve precise audiovisual correspondence (using audio timing).

A precise correspondence between the audio and the visual promotes functional understanding of the system and enhances the overall aesthetic experience.

In MIDI.CITI, every sound we hear is visualized in the scene with precise audiovisual synchronization (and we can think of the reverse as being true: each visual event is sonified). Whether they are particle effects coming from the rooftops or the animations of rabbits on the Moon, we take the advantage of strongly-timed audio provided by ChucK and use it to precisely drive both audio and graphics. In particular, MIDI.CITI is able to generate audio in ChucK while sending relevant parameter values and event triggers to Unity, enabling a visual element to occur in the nearest frame after a sound has been played. In this way, the audio retains its continuous generation while precisely synchronized with the graphics. Note this synchronization may fail if such a relationship is instead driven by graphics. A typical video frame rate of 30 fps to 60 fps may not be fast enough to control a continuous stream of generative audio and its varying nature may result in asynchrony, jitter, or other undesirable artifacts.

5.3 Take advantage of generative audio

Generative audio has many advantages over using pre-recorded music in audiovisual interactive software. Various low-level control of audio synthesis parameters including filters and effects, as well as high-level musical parameters such as tempo, dynamics, texture, and rhythmic complexity can be manipulated interactively and expressively.

In MIDI.CITI, by introducing pseudo-randomization, the looped sequences of music rarely gets repeated in the same manner, yet it preserves similar aesthetics by keeping a prescribed parametric boundary. Also, granular synthesis is used to enrich the overall sonic atmosphere, allowing the user to resynthesize a sound stream that interpolates between orchestral and choral timbres.

5.4 Combine high-level and direct control

An important consideration in generative music is finding an appropriate balance between automated/algorithmic processes and meaningful user input. This supports artful design's Principle 7.11A, "That which can be automated should be.", and 7.11B, "That which cannot be meaningfully automated should not be". For example, the user has high-level influence over rhythmic variations, musical density, and timbre, but this does not preclude the design from also having direct low-level musical input, such as pitch and note onset.

In MIDI.CITI, the UFOs are mapped in a diatonic scale of the music's overall key and allow for one-to-one action-to-sound interactions. Combined with the semi-automated music environment, this direct-input element enables moments of play, improvisation, and expression.

5.5 Surprise! Break the established expectation

As users spend a good amount of time with the design, the mechanics and dynamics of the system get understood and expectations begin to form. After establishing the ostensible core element, an effective way of making the design more interesting is to break such expectations by introducing a novel element that "no one asked for". This can result in a reframing and broadening of the aesthetic experience.

For example, in MIDI.CITI, the Moon in the sky remains a visual ornament as users interact with different city elements in the scene. As an unexpected feature, the user can fly toward it, and experience fantastical elements such as Moon rabbits creating grains of sound, resulting in the appearance of the sky-trains that flow back into the city. In addition, UFOs appear over the city, adding another layer of interaction to the audiovisual experience. Each additional event set up new mechanics and dynamics, leading to novel ways to aesthetically experience MIDI.CITI.

Breaking the established expectation within the system can be considered a form of defamiliarization—a technique that presents a familiar object or situation in an unfamiliar manner to protract the perceptive process and invite fresh context [11]. Such break in expectation sustains the experience, calls attention to its own aesthetics, and invites users to see and hear familiar elements anew.

5.6 One aesthetic theme to rule them all

In MIDI.CITI, the governing theme is a night cityscape, and it unifies the abiding elements of audio, graphics, and interaction. The musical elements incorporate instruments, rhythm, and chords related to lo-fi hip-hop. The visual elements use bloomed lighting, neon colors, and low poly graphics. The interactive elements consist of changing the lights of the city or turning the Ferris wheel around. Achieving this "thematic alignment" facilitates users to understand the design system without external guidance: for example, more lights, more sound, and more visual effects.

In addition, every element in the scene revolves around this central theme. Even ornamental elements such as cars

(and car honks), trucks, street lamps, and airplanes are introduced one by one as the city becomes sonically and visually more complex with user interactions. These strengthen the governing theme, adding to the fluidity of the overall aesthetic experience.

6. CONCLUSION

As for future work, mechanics for customizing and even programming the underlying musical context of MIDI.CITI directly within the software are considered. Currently, the ChucK backend contains information on probabilities of beats and pseudo-randomized algorithms of pitches. If the user has access to editing these components within the software, it would create another layer for creativity and expression. Moreover, a "free-edit" mode of a city would be an interesting implementation, where the user places different kinds of buildings (or even their own 3D models) and maps them to different audio parameters to customize their own version of a musical city.

In conclusion, we designed an experience-oriented musical cityscape and unpacked its design through lenses of interaction, play, and experience. We outlined the audio algorithms and pseudo-randomization techniques that aim to promote engagement in repeated iterations. We proposed design principles for using a generative audio-driven system, and interconnecting graphics, interactions, and music with the aim of creating a sense of narrative and experiential flow. We presented a case study of translating the design between desktop and VR and highlighted the adaptability of MIDI.CITI's algorithmic system. Overall, if we had a "call to action" in writing this paper, it would be to promote designing more aesthetics-driven audiovisual experiences, as tools for human creativity and expression.

• For demonstration video: https://vimeo.com/794287322

Acknowledgments

Thanks to all students in Music 220B and members of the CCRMA VR Lab at Stanford University for feedback and reflections during the design process.

7. REFERENCES

[1] G. Wang, P. R. Cook, and S. Salazar, "Chuck: A strongly timed computer music language," *Computer Music Journal*, vol. 39, no. 4, pp. 10–29, 2015.

[2] J. Atherton and G. Wang, "Chunity: Integrated Audiovisual Programming in Unity," in *NIME*, Conference Proceedings, pp. 102–107.

[3] G. Wang, *Artful Design: Technology in Search of the Sublime, A MusiComic Manifesto*. Stanford University Press, 2018.

[4] P. Cook, "2001: Principles for designing computer music controllers," *A NIME Reader: Fifteen years of new interfaces for musical expression*, pp. 1–13, 2017.

[5] R. Caillois, *Man, play, and games*. University of Illinois press, 2001.

[6] R. Hunicke, M. LeBlanc, and R. Zubek, "MDA: A formal approach to game design and game research," in *Proceedings of the AAAI Workshop on Challenges in Game AI*, vol. 4. San Jose, CA, Conference Proceedings, p. 1722.

[7] J. Lanier, *Dawn of the new everything: Encounters with reality and virtual reality*. Henry Holt and Company, 2017.

[8] J. Atherton and G. Wang, "Doing vs. Being: A philosophy of design for artful VR," *Journal of New Music Research*, vol. 49, no. 1, pp. 35–59, 2020.

[9] N. Tanaka and H. Takagi, "Virtual reality environment design of managing both presence and virtual reality sickness," *Journal of physiological anthropology and applied human science*, vol. 23, no. 6, pp. 313–317, 2004.

[10] J. Atherton, "Artful Design for Positive Design: A Case Study in VR," 2020.

[11] L. Crawford, "Viktor Shklovskij: Diffrance in Defamiliarization," *Comparative Literature*, pp. 209–219, 1984.