Title

Coherent Digital Multimodal Instrument Design and the Evaluation of Crossmodal Correspondence

Permalink

https://escholarship.org/uc/item/3gb4h770

Author

Lee, Myungin

Publication Date

2023

Peer reviewed|Thesis/dissertation

University of California
Santa Barbara

Coherent Digital Multimodal Instrument Design and the Evaluation of Crossmodal Correspondence

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Media Arts and Technology

by

Myungin Lee

Committee in charge:

Professor JoAnn Kuchera-Morin, Chair
Professor Curtis Roads
Professor Misha Sra

September 2023

The Dissertation of Myungin Lee is approved.

Professor Curtis Roads

Professor Misha Sra

Professor JoAnn Kuchera-Morin, Committee Chair

August 2023

Coherent Digital Multimodal Instrument Design and the Evaluation of Crossmodal Correspondence

by

Myungin Lee

To every inspiration...

Acknowledgements

First and foremost, I would like to thank my family for their support, advice, patience, and prayers.

I thank Dr. JoAnn Kuchera-Morin for her constant support of my Ph.D. procedure and research at the intersection of art, science, and engineering. I thank Dr. Curtis for his thoughtful remarks and encouragement of my research. I thank Dr. Misha Sra for her invaluable feedback and sincere advice. I thank Dr. Peter Hinterdorfer and Dr. Yoojin Oh for providing scientific insights through collaborative work.

I would like to express my gratitude to the MAT community and the AlloSphere research group for overcoming the difficult time during the pandemic together. It would have not been possible to finish my Ph.D. study without colleagues who encouraged each other day and night together.

Lastly, I thank NSF for their support through the OAC in the Cyber-infrastructure for Interactive Computation and Display of Materials Datasets program.

Curriculum Vitæ
Myungin Lee

Education

2023	Ph.D. in Media Arts and Technology (Expected), University of California, Santa Barbara.
2017	M.S. in Electronics and Computer Engineering, Hanyang University.
2015	B.S. in Electronics and Computer Engineering, Hanyang University.

Research Experience

September 2018 - July 2023: Graduate Student Researcher under NSF Grant ”Elements: Cyber-infrastructure for Interactive Computation and Display of Materials Datasets” in the AlloSphere Research Facility, University of California, Santa Barbara, US. PI: JoAnn Kuchera-Morin.

June 2020 – August 2020: Nokia Bell Labs. Experiments in Art & Technology (E.A.T.), Murray Hill, New Jersey, US. Research Topic: Spatial-Acoustic parameter estimation research and implementation

March 2015 - February 2017: Graduate Student Researcher under National Research Foundation of Korea Grants in Acoustic, Speech Signal Processing and Machine Learning Lab., Hanyang University, Seoul, Korea. PI: Joon-Hyuk Chang

Publications

Myungin Lee, Sabina Hyoju Ahn, Yoojin Oh, JoAnn Kuchera-Morin, ”Parasitic Signals: Multimodal Sonata for Real-time Interactive Simulation of the SARS-CoV-2 Virus,” IEEE VISAP 2023, October, 2023.

Myungin Lee, “Entangled: A Multi-Modal, Multi-User Interactive Instrument in Virtual 3D Space Using the Smartphone for Gesture Control,” New Interfaces for Musical Expression (NIME’21), June, 2021.

Myungin Lee, “A Multi-User Interactive Instrument in the 3D Space Using the Gesture of Smartphones,” Korea Electro-Acoustic Music Society’s Annual Conference (KEAMSAC), October, 2019.

Myungin Lee, ”Deep neural network based music source conducting system,” International Computer Music Conference (ICMC), August 2018.

Myungin Lee, Joon-Hyuk Chang, ”Deep neural network based blind estimation of reverberation time based on multi-channel microphones,” Acta Acustica united with Acustica, 2018.

Myungin Lee, Joon-Hyuk Chang, “Blind Estimation of Reverberation Time on Multi-Channel Microphone using Deep Neural Network,” Master’s thesis, February., 2017.

Myungin Lee, Joon-Hyuk Chang, “Blind Estimation of Reverberation Time using Deep Neural Network,” IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC), Beijing, China, September, 2016.

Jeehye Lee, Myungin Lee, Joon-Hyuk Chang, “Ensemble of Jointly Trained Deep Neural Network-Based Acoustic Models for Reverberant Speech Recognition,” arXiv:1608.04983, 2016.

Myungin Lee, Joon-Hyuk Chang, ”A study of room acoustics estimation using neural network,” Korea Speech Communication and Signal Processing, The Acoustical Society of Korea, pp. 30, August, 2016.

Tae-jun Park, Bong-Ki Lee, Myungin Lee, Joon-Hyuk Chang, ”Integrated acoustic echo and background noise suppression based on data-driven method, Korea Speech Communication and Signal Processing, The Acoustical Society of Korea, Vol. 32, No. 1, pp. 145-146, August, 2015.

Songkyu Park, Jihwan Park, Myungin Lee, Joon-Hyuk Chang, ”A study of speech enhancement using microphone array structure,” Korea Speech Communication and Signal Processing, The Acoustical Society of Korea, Vol. 32, No. 1, pp. 153-155, August, 2015.

Myungin Lee, Jihwan Park, Songkyu Park, Joon-Hyuk Chang,, ”A study of crosstalk cancellation efficiency in reverberant environment,” Korea Speech Communication and Signal Processing, The Acoustical Society of Korea, Vol. 32, No. 1, pp. 167-169, August, 2015.

Performance and Exhibitions

(Scheduled) August 2024, Exhibition, “Sensorium: The Voice of the World Ocean,” PST 2024 Getty Biennale, Irvine, USA

(Scheduled) October 2023, Exhibition, “Parasitic Signals: Coexistence with the SARS-CoV-2 virus,” IEEE VISAP 2023, Melbourne, Australia

June 2023, Audiovisual concert (Tech Support), Premiere of ”Man in the Mangroves” by James Andy Moorer, AlloSphere, UCSB, Santa Barbara, USA

2019 – 2023, Regular Demonstration, AlloSphere Contents (Hydrogen-like atoms, Artificial Nature, Sensorium, Entangled, The Last Whisper, Musics of the Sphere, etc.), AlloSphere & AlloPortal, UCSB, Santa Barbara, USA

February 2023, Exhibition, “Coexistence with the SARS-CoV-2 virus,” Santa Barbara Center for Art, Science and Technology (SBCAST), Santa Barbara, USA

October 2022, Spatial audio concert (Tech Support), Premiere of ”Musics of the Sphere” by Dr. Robert Morris, AlloSphere, UCSB, Santa Barbara, USA

September 2022, Exhibition, “Coexistence with the SARS-CoV-2 virus,” Ars Electronica Festival 2022, Linz, Austria

May 2022, Audiovisual Concert (Direction), “AlloLib Audiovisual Concert,” SYMADES 2022, the California NanoSystems Institute, UCSB, Santa Barbara, USA

April 2022, Exhibition (Tech Support), “Last Whispers” by Lena Herzog, AlloSphere, UCSB, Santa Barbara, USA

June 2019, Art installation, “A Multi-User Interactive Instrument in the 3D Space Using the Gesture of Smartphones,” the MAT 2019 End of Year Show: MADE [at] UCSB, the California NanoSystems Institute, UCSB, Santa Barbara, USA

April 2019, CREATE Ensemble Performance, ”Ballet Mécanique (2019),” at Lotte Lehmann Concert Hall, UCSB, Santa Barbara, USA

August 2018, Art installation, “Deep neural network based music source conducting system,” International Computer Music Conference (ICMC), Daegu, Korea.

June 2018, Art installation, “Deep neural network based music source conducting system,” the MAT 2018 End of Year Show: Invisible Machine, the California NanoSystems Institute, UCSB, Santa Barbara, USA

June 2018, CREATE Ensemble Performance, ”Loading (2018),” at SBCAST, Santa Barbara, USA

May 2018, CREATE Ensemble Performance, ”Loading (2018),” at Lotte Lehmann Concert Hall, UCSB, Santa Barbara, USA

Patents

Multichannel Microphone-based Reverberation Time Estimation Method and Device which use Deep Neural Network Technical Field, US Patent: US10854218B2, 2017.

Multi-Channel Microphone based Reverberation Time Estimation using Deep Neural Network, Korea Patent: KR101871604B1, 2016.

Invited Talk

July 2023, Coherent Digital Multimodal Instrument Design and the Evaluation of Crossmodal Correspondence, Ewha Womens University, Seoul, Korea

July 2023, Coherent multi-modal instrument design in digital media, University of Maryland College Park, April, Maryland, USA

December 2022, ACM SIGGRAPH Digital Art Community SPARKS: New Media Architecture(s): A Speculative Vision of Change in the Arts, Design, & Sciences, Online.

January 2020, AR Interaction/Interface for CS291A Future User Interfaces. UCSB, Santa Barbara, USA

Teaching Experience

Fall 2022. Instructor, MUS109IA: Direct Digital Synthesis - Processing and Composition

Winter 2022. Instructor, MUS109IA: Direct Digital Synthesis - Processing and Composition

Winter 2021. Teaching Assistant, MAT240B: Digital Audio Programming: The Series (Instructor: Dr. Karl Yerkes)

Fall 2020. Teaching Assistant, MAT 240C: Digital Audio Programming: The Series (Instructor: Dr. Karl Yerkes)

Spring 2020. Teaching Assistant, MAT 276IA: Direct Digital Synthesis - Processing and Composition (Instructor: Prof. JoAnn Kuchera-Morin)

Spring 2019. Teaching Assistant, MAT 276IA: Direct Digital Synthesis - Processing and Composition (Instructor: Prof. JoAnn Kuchera-Morin)

Spring 2019. Teaching Assistant, MAT 240A: Digital Audio Programming: The Series (Instructor: Dr. Karl Yerkes)

Abstract

Coherent Digital Multimodal Instrument Design and the Evaluation of Crossmodal Correspondence

by

Myungin Lee

The rapid development of the current availability of advanced hardware and software is opening up new opportunities for digital creation every day. This situation provides great freedom for new artistic expressions with advanced audio, graphics, interface, and algorithms, including machine learning. However, while our nature is multimodal, these modalities in the digital domain are genuinely separate, and the computational platform allows innumerable varieties of linkages among them. For this reason, the holistic multimodal experience is highly dependent on the design and connection of different modalities. This dissertation explains the properties of coherent digital multimodal instruments and discusses their creative opportunity from the process of music composition and performance. The chapters introduce the related projects with their design process, demonstrating the role of crossmodal correspondence in scientific simulation and proposing a numerical method to evaluate the crossmodal correspondences using the correlation coefficient between the modalities. This dissertation aims to contribute to reorganizing the design process of multimodal instruments beyond the old and recent customs.

Contents

Chapter 1

Introduction

1.1 Motivation

We interact with the world as an integrated collection of sensations from multiple sensory modalities. Let us think about a simple example: finger snap. The activity utilizes the thumb, middle finger, and hand surface. With this given interface, lubing the thumb and middle finger with each other, we generate friction between the two fingers. This loads tension to the middle finger right before snapping. When we snap, this makes the middle finger hit the surface of the hand exceptionally fast. This impact generates a short impulse-like signal in the hand which vibrates throughout our hand. This physics generates sound, and the entire process can be observed visually. This activity can deliver some messages or generate rhythm, which sometimes requires training to master. Indeed, it is a unique instrument in our hands.

Throughout history, humanity has expanded these concepts from fingers to devices, developing countless types of musical instruments to express our voices to the world. While these instruments are all different, we can generalize the components of the instrument into four parts: interface, physics, sound, and visual. These modalities are genuinely connected, composing the multi-sensory experience. Likewise, modalities in our real life do not work independently but behave and are observed as dependent on other modalities. This dependency is inevitable in physical instruments.

However, such dependencies have become vague since electronic media have been available. New technologies allow us to observe what we could not see, hear, generate, manipulate, and compute before. This circumstance has opened up endless opportunities for novel artistic creations, allowing great freedom for instrument design. Such freedom is a huge opportunity but suggests a new challenge: While the dependencies between modalities are not determined, how do we represent these new observations multimodally? What opportunities can multimodal representation allow compared with monomodal representation? Is there a method to tell which multimodal design is more coherent?

Starting from these questions, this dissertation discusses the components of multimodal experience, introduces instruments as case studies, suggests a method to numerically analyze the multimodal relationship using multimodal signal processing, and defends their unique opportunities by introducing novel artistic creation and scientific observation.

1.2 Multimodal Instrument

Let us start by defining this dissertation's core concepts and terminologies: "*multimodal, instrument, coherence*." The thesis expands, questions, and connects different concepts from the definitions to articulate how to design "*coherent multimodal instruments*."

1.2.1 Multimodal

Joddy Murray describes multimodality as communication practices using visual, aural, spatial, textual, and linguistic resources to compose messages [1]. Even though our sensory modalities may appear to be working separately, we naturally try to find the connection between different perceptual modalities in a synergic way. When sensory inputs from different modalities are consistent or congruent, they enhance our perception and understanding of the world. This phenomenon is often referred to as "crossmodal integration" or "multisensory integration."

According to Alexa Bódog's research in 2012 [2]:

> "Babies learn to pick up environmental information in a synergic way, which means that their perceptual modalities work in a coordinated way as well as their sensory and motor abilities. For example, the coordination between vision and sound enables infants to prefer the face who is talking to them, or adults can read lips in an ambiguous or perceptually unstable environment."

Likewise, we naturally try to relate modalities and fill out the environmental information multimodally when we perceive ambiguous or perceptually incomplete information. To explain the interaction between different modalities, the next section describes the concept of crossmodal correspondence and coherence.

1.2.2 Crossmodal Correspondence & Coherence

In our multimodal experience, the signals sensed through one sensory modality can influence the processing of information received through another. Such a phenomenon is crossmodal interaction. To achieve a "natural" mapping of features, or dimensions, of experience across sensory modalities, we genuinely utilize crossmodal interaction to find crossmodal correspondence among different modalities [3]. This natural mapping requires systematic, consistent, logical connections bridging "coherent" multimodal experience. Here is an example of crossmodal correspondence, takete–maluma effect.

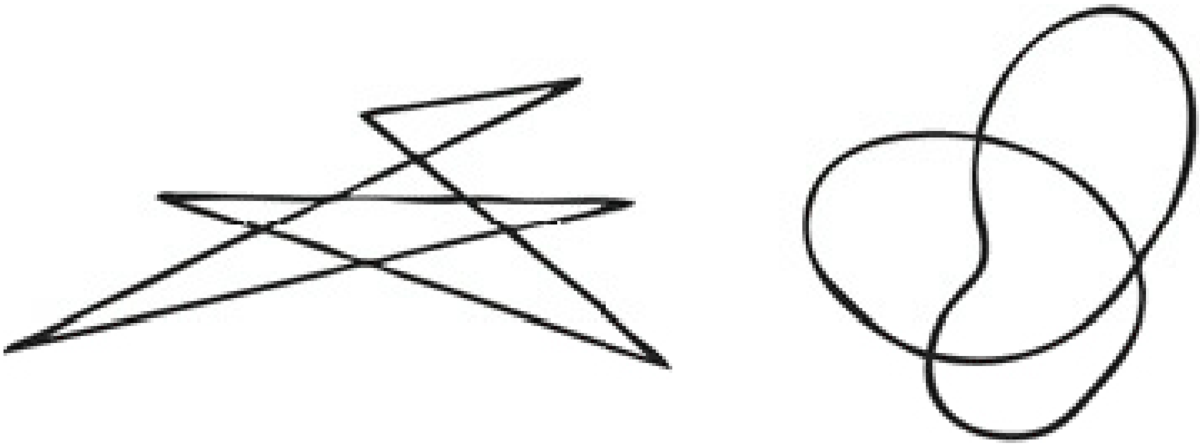


Figure 1.1: Two figures of takete–maluma effect

In this experiment by Wolfgang Köhler in 1929 [4], the participants match the word "takete" and "maluma" with the two figures above. About 97 % of participants relate the word "takete" better with the left shape, whereas the word "maluma" connects with the right shape. According to an additional experiment by Bremner in 2013 [5], participants who do not use a written language also exhibit this effect.

Let us extract a few features representing the characteristics of two figures and words to expand this experiment into numerical analysis. Figure 1.2 shows the features extracted: geometric observation, waveform from text-to-speech, audio spectrogram, and pitch estimation.

To start with the geometric observation, Takete has six vertexes with no curve, and Maluma has zero vertexes with curved edges. Based on the observation that the pronunciation of words affects the decision, a text-to-speech algorithm generates the sound of the words to represent the audio characteristic. The waveform, spectrogram, and

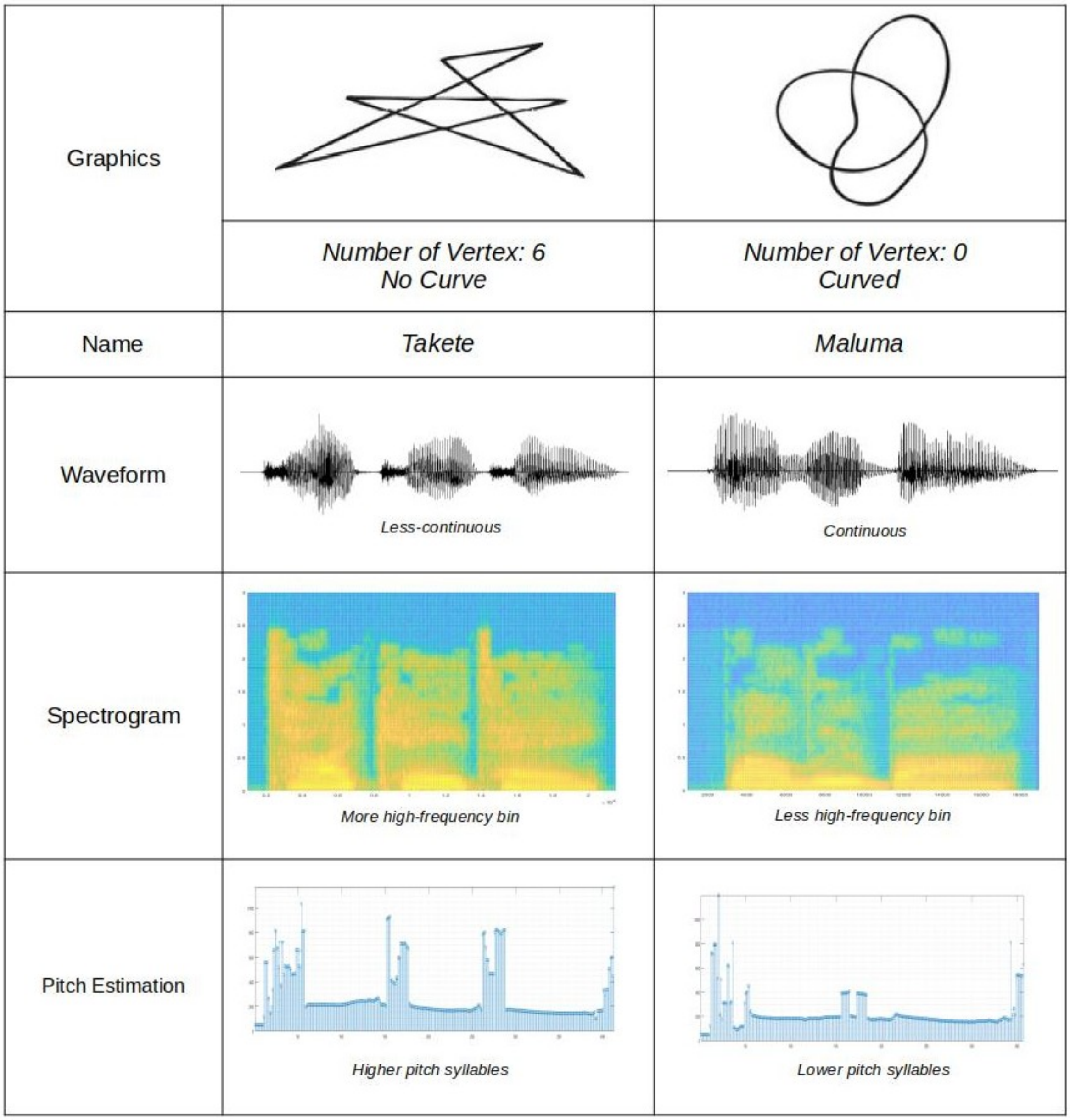


Figure 1.2: Analysis of Takete and Maluma in graphics and audio with features

pitch estimation are obtained from this audio using signal processing techniques written in MATLAB. Comparing the two waveforms, Takete is relatively less continuous than Maluma. To observe the spectrograms, Takete contains more high-frequency bins than Maluma, and the higher dimension feature, pitch estimation, shows a similar result. This observation indicates a strong positive correlation between graphical geometric features and audio frequency features. Likewise, further research can expand these feature selections, but at this stage, this proof of concept briefly shows the multimodal relationship between the graphics and audio of Takete and Maluma.

Explaining this phenomenon in neuroscience, the visual cortex has strong connections with other brain regions. In our brain, various types of information are fed into the visual cortex and integrated to form meaningful visual experiences. The output from the visual cortex is also sent to other brain areas, facilitating the connection and integration of diverse information, including audio and higher dimensional features. A similar process happens with the auditory cortex. Likewise, our brain synthesizes information from cross-modal stimuli through multisensory integration [6]. This observation is one of the fundamental notions of this thesis discussing how to design multimodal experiences in the digital domain.

1.2.3 Instrument

An instrument is a tool or implement, especially for delicate or scientific work. This definition can include various devices, including microscopes, calculators, musical acoustic instruments, and immersive environments such as the AlloSphere. Moreover, it is noticeable that the AlloSphere is called a one-of-a-kind immersive instrument. The AlloSphere is a three-story-sized, 26-projector, and 54.1-channel audio facility where researchers use multiple modalities to represent large and complex data, including immersive visualization, sonification, and interactivity.

While it is likely to be focused on the immersive environment, the AlloSphere differentiates itself by creating technology enabling experts to use their experience and intuition to examine and interact with complex data to identify patterns, suggest theories, and test

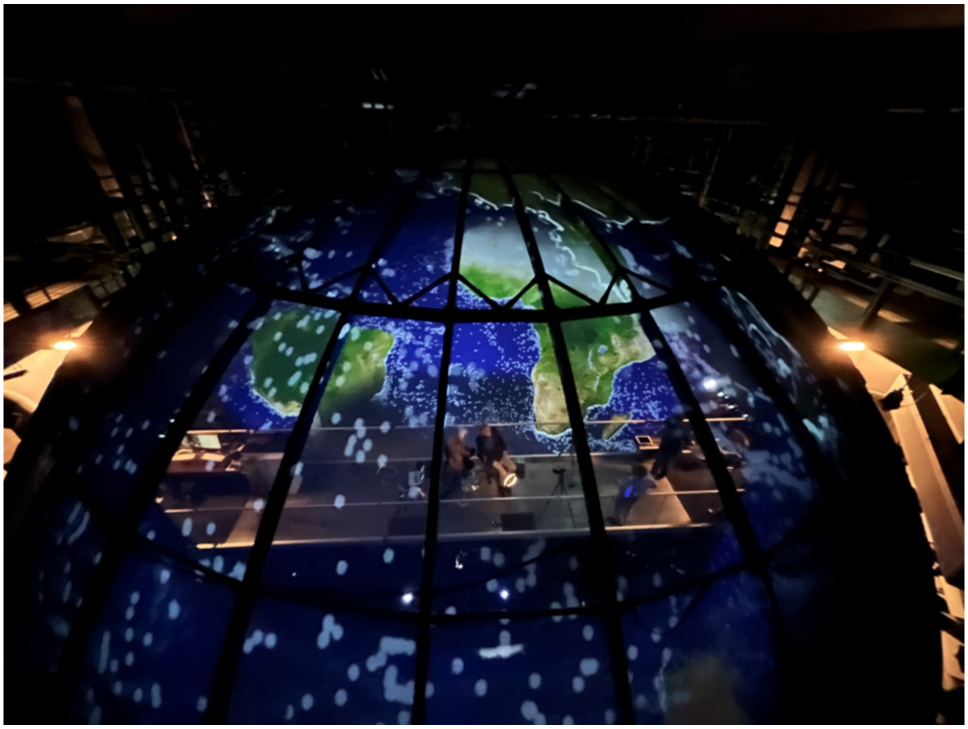


Figure 1.3: The AlloSphere

them in an integrated discovery loop. Furthermore, one of Dr. JoAnn Kuchera-Morin's interviews defines AlloSphere's research as follows:

> "A new and unique intersection of advanced science and art through the interactive visualization of complex systems using the creative process of music composition and performance."

Just like our brain relates different modalities and fills out the indefinite environmental information multimodally in real life, by involving auditory modality in complex interactive systems, the AlloSphere presents a coherent immersive link between the digital simulation domain. This approach allows the researchers to expand a compositional narrative. For example, a musical narrative is constructed by a succession of events in time and the interaction of simultaneous sounds. Likewise, the human brain constructs narratives from our experience by relating current perceptions to the past and anticipating the future. This process efficiently delivers temporal information within complex multimodal interactive systems, which aids specialists in accelerating their discovery, delivering pathological messages to the general audience, and allowing new creative opportunities for the artists. Therefore, the AlloSphere is simultaneously a microscope, calculator, immersive environment, and musical acoustic instrument.

1.2.4 Coherent Digital Multimodal Instrument

To abstract these concepts, a multimodal instrument in digital media is a tool or platform that allows users to create, share, or interact with digital content that combines multiple modes of communication. In this research, we constrain the components of the multimodal instruments as visual, audio, gesture, and rules that relate to those relationships. However, this concept can be expanded to include additional modalities.

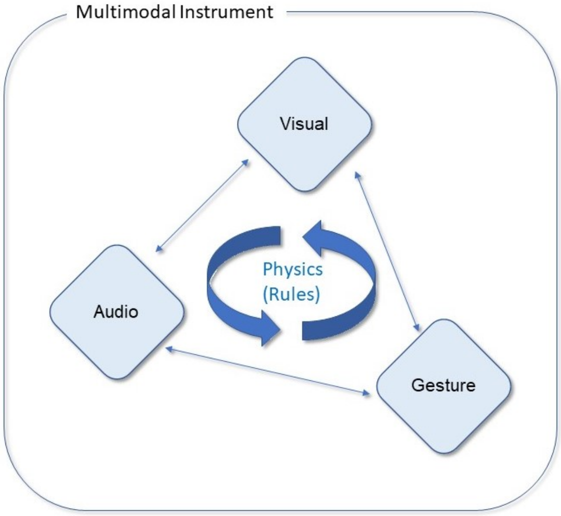


Figure 1.4: The multimodal instrument loop

Figure 1.4 shows an exemplary coherent multimodal instrument loop composed of modalities that can be related using physics or rules. These systematic, consistent, logical connections become the medium that enables coherent crossmodal interaction and builds crossmodal correspondence.

1.3 Problem Statement

The instrument's interface is the membrane of interaction between humans and technology. Especially designing an interface for the digital instrument requires a substantial effort to achieve coherent crossmodal interaction. While the characteristics of the acoustic instrument are determined by its physicality including structure, resonance, texture, and space, digital instruments with computational platforms are inherently non-physical. The digital instrument designer can separate sound production from the means. This circumstance gives excellent freedom to instrument design. At the same time, designing the interface to interact with the sound material that is now separated from the physicality is challenging. Nevertheless, our bodies and movements are the most expressive tool that humans can have.

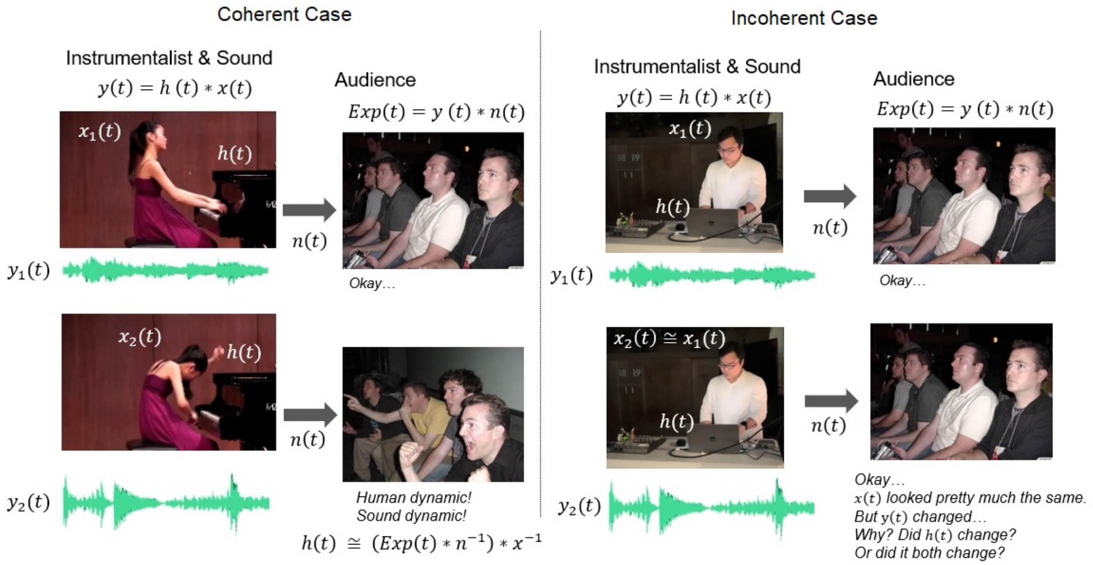


Figure 1.5: The multimodal instrument loop

To analyze the properties of coherent instruments, this research suggests establishing a model interpreting the music, instrumentalist, instrument, and audience's experience as a function. Figure 1.5 shows the exemplary comparison between coherent and incoherent cases where $h(t)$ is an instrument, $x(t)$ is the instrumentalist's gesture, $y(t)$ is the auditory output, $n(t)$ is the transfer function between the instrumentalist and the audience, and the $Exp(t)$ is the audience experience.

For the coherent case, $x_1(t)$ and $x_2(t)$ are noticeably different and the corresponding $y_1(t)$ and $y_2(t)$ changes drastically. Consequently, the audiences can presumably derive the functionality of $h(t)$. In contrast, for the incoherent case, $x_1(t)$ and $x_2(t)$ are similar whether $y_1(t)$ and $y_2(t)$ transforms.

This model derives two properties of coherent instruments: time invariance and perceptible interface. Time invariance means the input and output characteristics of a system do not change with time. The second property states the audience can identify and relate the gestural event to the sound. These properties are necessary for our brain to synthesize information from systematic, consistent, and logical cross-modal stimuli through multisensory integration.

These observations bring underlying questions:

> How do we decide which multimodal experience is coherent?
>
> What is the "natural" mapping of multimodal features?
>
> Is the user study the only method to evaluate the experience?
>
> Is there a way to numerically analyze the level of multimodal coherence?
>
> What opportunities do the multimodal experience allow compared with the monomodal experience?

Through the literature review and case studies, this thesis argues the essential aspects of multimodal instrument design and what opportunities it can present. Furthermore, this dissertation proposes a method to evaluate the crossmodal correspondence using correlation coefficients between modalities.

Chapter 2

Literature Review

2.1 History of Digital Instruments

While the history of digital instruments is vast and profound, this chapter mainly observes the evolution of their interface.

Research in digital audio began in the 1920s and 1930s with signal processing and the invention of analog-to-digital conversion. The first computer to play sound and music was the CSIR Mark 1, and Alan Turing's pioneering work in the late 1940s transformed the computer into a musical instrument. To generate the music, the engineer had to enter the programs bit by bit using a panel of hand-operated switches [7]. In 1957, the RCA Mark II sound synthesizer was invented as the first programmable electronic synthesizer and the flagship equipment at the Columbia-Princeton Electronic Music Center. The RCA Mark II used a binary sequencer as the interface, using a paper tape reader similar to a piano that would send instructions to the synthesizer automatically playing the sound [8].

In the late 1950s, modular synthesizers were developed, composed of separate modules for different functions. The modular synthesizers can be expanded by connecting patches using chords, enabling unlimited expandability. This design has been widely adopted, integrating keyboard synthesizers, sliders, knobs, and samplers until now [9]. This design has been widely adopted, integrating keyboard synthesizers, sliders, knobs, and samplers until now. However, due to their reliance on physical buttons and cables, modular synthesizers frequently had few features that could be operated remotely. While interest in modular synthesizers is still ongoing, a recent study by Beat Rossmy shows there was a controversial response from Eurorack-manufacturers that audiences do not realize the difference if the music is improvised or not [9].

In 1981, the MIDI (Musical Instrument Digital Interface) protocol was invented, and it became a method of connecting diverse electronic musical equipment [10]. Still, the genuine layout of MIDI is based on wired cables, so wires were typically included in the instrument design.

In 2001, New Interfaces for Musical Expression (NIME) conference started from the ACM Conference on Human Factors in Computing Systems (CHI). It became a crucial venue for advancing musical expression from the aspect of interface design, human-computer interaction, and computer music [11].

The invention of Open Sound Control (OSC) in 2002, a protocol for networking sound synthesizers, computers, and other multimedia devices, helped higher resolution and a richer parameter space to be transported across the wireless internet [12].

Recently, active research on machine learning has stimulated research on deep learning-based automatic compositions using brief interfaces such as keywords [13, 14].

Throughout the evolution of digital instruments, as computational capability has drastically upgraded, dependency on physicality has been reduced, automating sound production compared to acoustic instruments.

2.1.1 History of digital instrument with science

Scientific audio visualization inspired general audiences and experts from the early stage of digital instrumentation. In 1966, the featured artwork and computer graphics by Ken Knowlton and computer-generated music by Max Mathews from Bell Labs generated a stereoscopic computer animation showing the 3D simulated movement of the basilar membrane in the human ear. Especially even if real-time audio synthesis became available later in the video, the audience could find the audio was designed to deliver the corresponding coherence to the graphics and science. According to the conversation from Incredible Machine (1968) [15], we have evidence of the early digital art-science discovery:

> ”It allows people to see things which only mathematicians could see before.”
>
> ”It offers not only the means to quicken the pace of discovery but an ideal way of communicating what we may discover.”

2.2 Crossmodal Correspondence in Computer Music

This research finds examples of crossmodal correspondence in digital media from computer music. György Ligeti, one of the most important avant-garde composers in the latter half of the twentieth century, gave an unusually lucid explanation of his composing process through one interview [16]:

> “I always imagine music visually, in many different colors. The music is not growing. It is completely empty. The first sketches are always drawings.”

An exemplary case of comprehensive audiovisual multimodal experiences that shows crossmodal correspondence in his work is Artikulation.

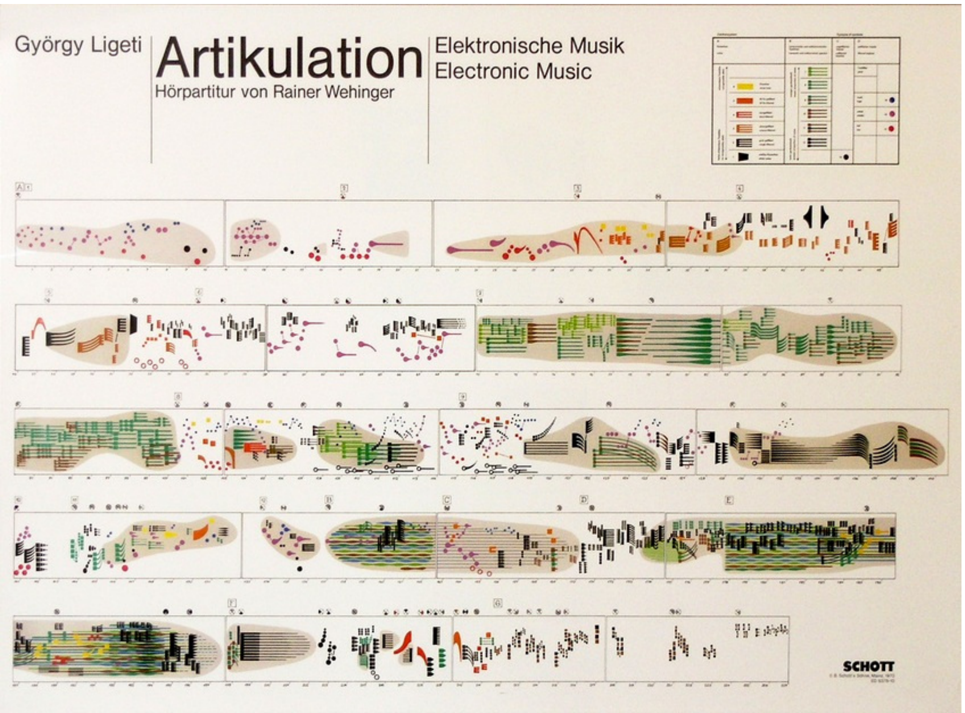


Figure 2.1: “Hörpartitur” or “score for listening” for the Artikulation, by Rainer Wehinger

In Artikulation, Ligeti said he wanted to build language-like or text-like structures using a kind of imaginary or pseudo-language with the help of electronic means. The piece’s narrative is linear and filled with questions - answers, high - low voices, polyglot speaking - interruptions, impulsive outbreaks - humor, and charring - whispering. Even if no explicit words exist, the listeners can infer conversation and expect and predict throughout the piece. Ligeti wrote this electronic piece in 1958 as a recording, and there was no score for musicians to visualize the music. In the 1970s, Rainer Wehinger created a visual explanation of the music relating and explaining all the colors and symbols with sonic elements [17]. Following the visualized score with the recording, we can simultaneously relate pitch, timbre, tempo, and overall structure in audio and visual through a congruent multimodal experience.

This chapter investigates the characteristics of multimodal experience in audio and visual to discover the property of crossmodal correspondence by analyzing Artikulation and its visual score. This piece's visualized score represents different sonorous effects with specific graphic symbols. Here I use spectral analysis to investigate the relation between the spectrum and the score's notations. Analyzing sound sources using the spectrum provides a visual approach to understanding the information that humans can miss in auditory analysis. The spectrum is obtained using a fast Fourier transform (FFT) with FFT size 4096, and the frequency on the vertical axis is scaled logarithmically. While the original piece is stereo, the two channels are equally mixed to focus the observation on the pitch.

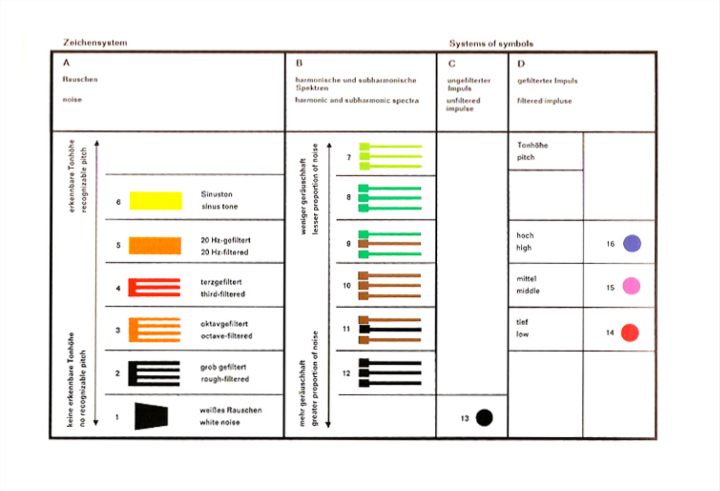


Figure 2.2: Sound materials used in György Ligeti's Artikulation

- Filtered-noise

Figure 2.1 describes the 16 notations used in the score. Four separate units compose the fundamental sonic elements: A. noise, B. harmonic and subharmonic spectra, C. unfiltered impulse, D. filtered impulse. The notes generated from noise have six types: Sine tone, 20 Hz-filtered noise, Third-filtered noise, Octave-filtered noise, Rough-filtered noise, and White noise. To observe the relation between the timbre and the color, we can find that as the pitch becomes more recognizable, the color becomes brighter and vice versa.

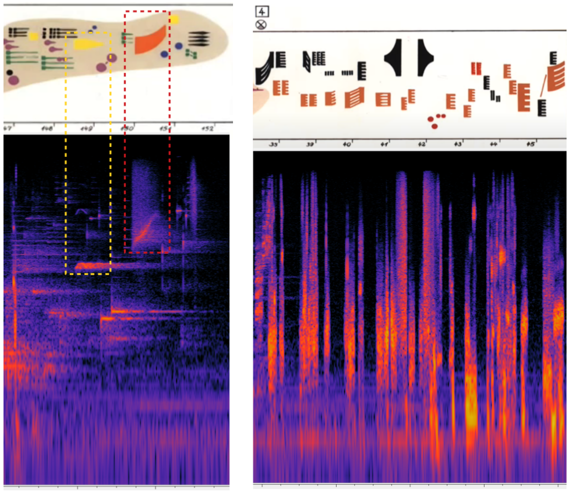


Figure 2.3: Example of a sine tone and 20 Hz-filtered noise (left) third-filtered, octave-filtered, rough-filtered, and white noise (right) in the piece

Figure 2.2 (left) shows a sine tone and 20 Hz-filtered noise notes at around 148 seconds of the piece. By observing the spectrum of the corresponding note, the sine tone (Yellow) shows a straight and concentrated tone with a particular frequency. To compare, the 20 Hz-filtered noise (Orange) shows a relatively blurred spectrum with elevating fundamental frequency. The outer mesh behind the notes refers to reverberation.

Looking at Figure 2.2 (right), from 38 seconds to 46 seconds, third-filtered, octave-filtered, rough-filtered, and white noise notes appear, and the corresponding spectrum shows the characteristic of each filtered noise. The shape of the notation recalls the filter's frequency response, and it helps the audience to match the pitch from the note. Also, the notation's direction and shape refer to the note's amplitude envelope. For example, when the filtered-noise notation is closed on the right side, it will start and diminish following the notation, and vice versa.

- Harmonic and subharmonic spectra

Sound materials made with harmonic and subharmonic spectra are the sequence of harmonic and subharmonic layers in the frequency domain. The combination of colorful lines visualizes the sonic material's noise proportion. As the line becomes darker, the material contains more noise. Figure 2.3 (left) shows brighter materials in the score correspond to clear lines in the spectrum, and darker materials match to noisier spectrum. Figure 2.3 (middle, right) shows variations of their usage using their width and elevation corresponding to their amplitude and frequency.

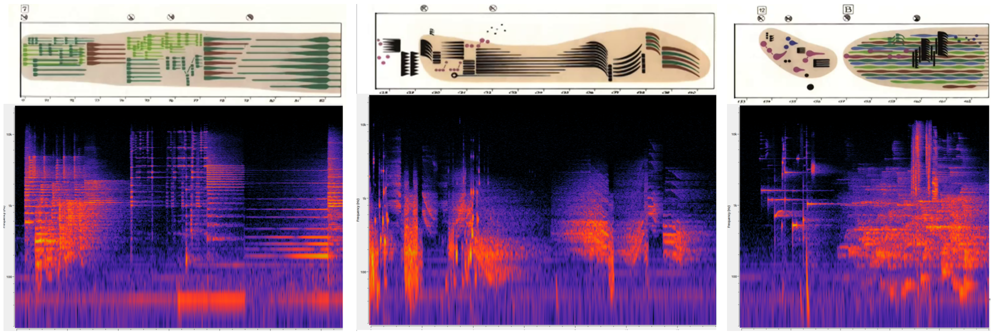


Figure 2.4: Variations of Harmonic and subharmonic spectra in the piece

- Unfiltered or Filtered impulse

The unfiltered impulse scored with a black dot is widely used over the piece. The idle impulse signal in the time domain is a uniform function in the frequency domain. Since the inactive impulse signal is impossible to implement in digital signal processing, theoretically, the material's characteristic is similar to a square wave in the time domain. This can be observed as a sync function in the frequency domain. Figure 2.4 (left) shows it matches to a short and widely spread spectrum with the diminishing pattern of sync function in high frequency. Unlike unfiltered impulse, filtered impulse shapes a sharper beamwidth with weaker sidelobes. Figure 2.4 (right) compares unfiltered and filtered

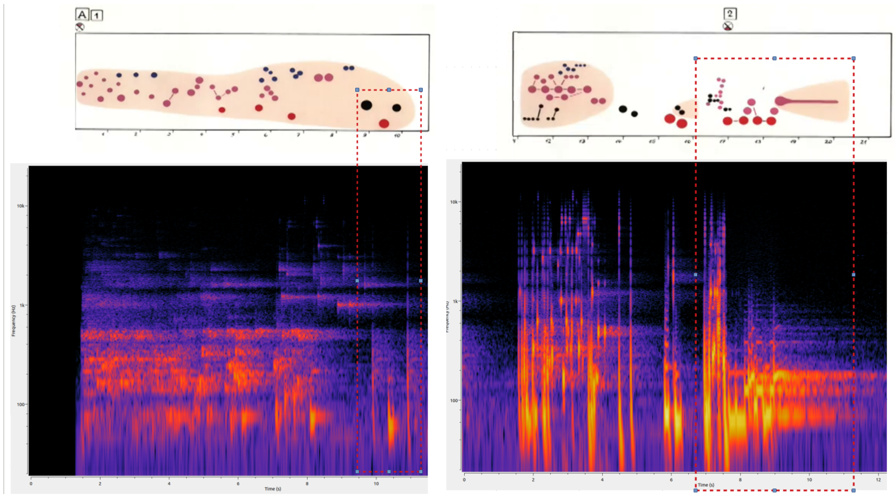


Figure 2.5: Unfiltered or filtered impulse in a phrase

impulse material in a phrase.

This investigation shows how different sonic elements are visualized, connoting their meanings through their notations. This observation presents an essential concept in real-time interactive instrument design with data, visualization, and sonification. When the relationship between data, physics, visualization, and audio is coherently designed within the instrument, the user can obtain comprehensive intuition from the media that could not be obtained from separate experiences. This thesis develops this notion for further multimodal instrument design. For example, Section 3.2 in Chapter 3 presents a similar approach to generating coherent audio from graphics and gravity-like physics. And Section 3.4 depicts research utilizing diverse sonic elements to be fused into the dynamic audiovisual narrative.

Chapter 3

Case Studies

This chapter presents four different multimodal instrument design case studies. While the case studies handle diverse topics, including conducting (Section 3.1), gravity (Section 3.2), granular synthesis (Section 3.3), and SARS-CoV-2 virus (Section 3.4), each case can be represented as the proposed multimodal instrument model, Figure 1.4.

3.1 Music Source Conducting based on Gestural Control using Machine Learning

Conducting is defined as the art of directing the simultaneous performance of several players or singers using gestures [18]. Accordingly, even for the same piece of music, the conductor's directions cause a significant variation in the music.

Motivated by this concept, since the first attempt of Buxton [19], several studies involving the algorithmic analysis of conducting gestures have been made over the years. The methods adopt diverse approaches to track conducting gestures, including magnetic motion tracking [20], image processing [21, 22], and gyroscope [23]. However, conventional projects with analytic models are limited since they derive a specific command, such as tempo, from the sensors.

This case study proposes a musical-interactive system that enables the user to experience conducting activity with existing music sources with a machine learning approach. The user can control the music's tempo, amplitude, and frequency response using an accelerometer and gyroscope-based controller. The audience can have a multimodal experience by watching the user's gesture and pairing it with the manipulated audio. The conducting model can be extended with more complicated commands. The algorithm starts by extracting gyroscope and accelerometer data from smartphones. The ubiquitous smartphones with various sensors increase the accessibility of the system. At the same time, the rapid advancement of machine learning technology enabled the solution of many complex problems that analytical approaches could not solve. The proposed system is composed of various signal-processing techniques and machine-learning methods. The deep neural network (DNN), proposed by Hinton and Salakhutdinov [24] which is adopted as a state-of-the-art classification method in diverse pattern recognition tasks, is adopted to train and perform the classifications and regression of the conduction commands with the input features. Accordingly, classified commands are processed using sampling, modulation, amplification, and filtering of the music source in real time. The following chapters introduce the structure of the proposed system, the database generation, and the evaluation of the performance with discussion.

3.1.1 Music Source Conducting System

To achieve an intuitive music source conducting system, the rules of the commands are simply defined. The player generates the pattern by swinging a cellphone with the instructed shape, and the system classifies the pattern through the neural network. The shape of the swing is defined as the circles and 8-figure Mobius strips, while the detailed

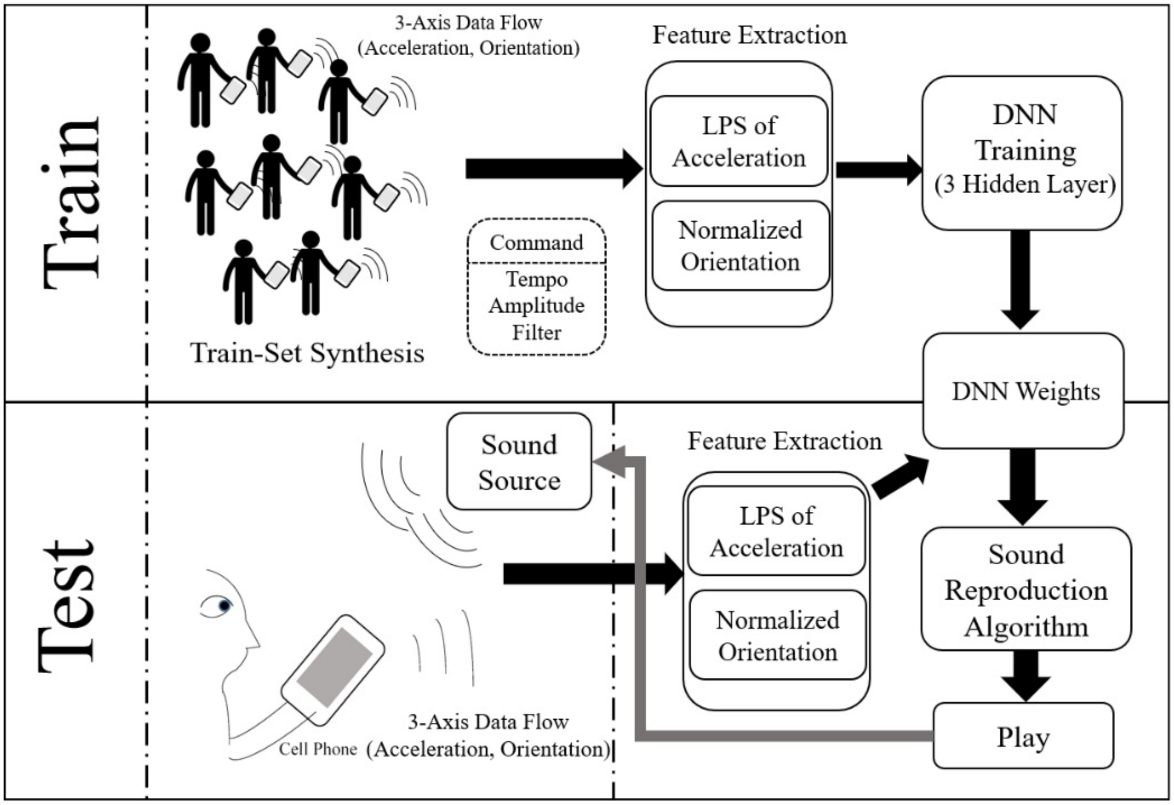


Figure 3.1: The overall structure of the proposed DNN-based music source conducting system.

commands differ according to the speed, width, and direction of the swing. Even though the shapes are not the only usable patterns, they are adopted since they can engage diverse beats which differ from various pieces of music.

The cycle of the swing determines the tempo. For example, when the cycle of the 8-figure Mobius strip and circle is short, the system interprets the command as a fast tempo. The width of the gesture determines the amplitude of the playback sound. When the width of the swinging is large, the playback volume is high. Meanwhile, the orientation of the controller has a different rule. When the player swings the controller in a circular shape with a head-up orientation, the pattern is interpreted as a treble boost. If the player swings the controller with the head-down orientation, the system interprets the pattern as a bass boost. Otherwise, when the player swings in an 8-figure Mobius strip with centered orientations, the frequency characteristic is maintained, and only the speed and amplitude of the pattern are controlled.

After the DNN model classifies the command, the system manipulates the sound source correspondingly. The system changes the playback rate of the sound, maintaining its pitch by modulating the frequency correspondingly. The volume of the sound simultaneously changes following the corresponding amplitude. When the command filters a specific frequency range, the sound source gradually boosts the corresponding frequency band.

Database Synthesis

In the DNN training process, the training set's database significantly determines the system's accuracy. For example, if DNN is trained with imbalanced datasets, the system is likely to be challenged with the overfitting issue. Therefore, I collected a database with

Real-time demonstration at the MAT 2018 End of Year Show: Invisible Machine, the California NanoSystems Institute, UCSB, Santa Barbara a diverse variance for each command. The database for the DNN training is obtained from the orientation and acceleration data from seven different people who were instructed on the system and the command. By diversifying the subjects, we expect to mitigate the bias of the experiment for a specific pattern caused by a certain height, width, or swinging habit.

The command is composed of three parts: tempo, amplitude, and frequency band boost. The database of tempo includes 40 to 80 beats per minute (bpm) for every five bpm. To obtain the quantitative dataset for the tempo, a corresponding metronome sound was provided while making each set of data. The subjects were requested to generate the amplitude command by swinging the controller approximately within the range of $0.1 - 0.8$ meters for five steps. To classify the frequency band boost, we requested the subjects to swing the controller with three different orientations: head down, centered, and head up. Each subject made 150 seconds of commands for each command following the instructions.

Feature Extraction

Converting the raw data into refined features is an efficient way to derive the DNN model. The proposed system extracts the log-power spectra (LPS) for each axis's acceleration data stream. For the 2019 version, the cellphone transmits each axis's acceleration value ten times per second through OSC. Currently, much higher stream rates are available to implement more rapid and sophisticated commands for further study.

We derived a short-time Fourier transform (STFT) of 32 samples added with 32 zeros using a 64 FFT size. This states that the range of the observation is equivalent to 3.2 seconds. The performance of the operating system determines the rate of commands. Throughout the process, the system refers to the data from the previous 3.2 seconds to determine the command. This prevents the abrupt transition of the command, which makes the system stable, but may result in delays on rapid commands. In addition, to utilize the controller's orientation for a bass-treble controller, the 3-axis orientation values are adopted as an input feature. After the feature extraction, every feature is normalized within the range of 0 to 1 to be employed as the input of the DNN. The normalization is made with empirical thresholds. These configurations are not optimal, and this practice became a chance to learn the importance of real-time signal processing using object-oriented programming languages for the interactive multimodal system.

DNN Training Setup

In the proposed method, the DNN plays a role in learning the complicated relationship between the input features and target commands through its multiple non-linear hidden layers. The structure of the DNN has three hidden layers, and the number of nodes for each hidden layer are 1500. A rectified linear unit (ReLU) is adopted for the activation function of the hidden layer. Also, pre-training is applied to avoid pool local optima or over-fitting. This procedure optimizes the training by making the model parameters trained with unsupervised greedy layer-wise learning before the fine-tuning process [25]. When training the DNN, each hidden layer was fine-tuned with 1000 epochs. The configuration of the DNN structure is chosen empirically to model the convoluted relations sufficiently while avoiding overfitting the training data.

Music Playback

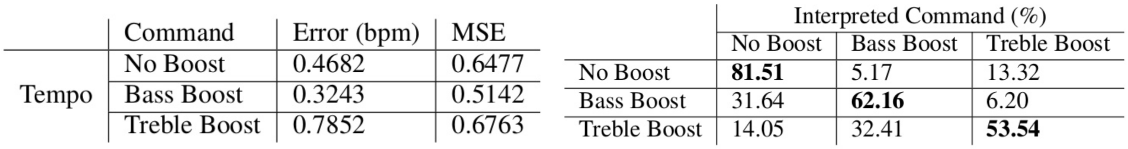
After deriving the conducting message through the DNN regression, the corresponding signal processing techniques should be applied simultaneously. Resampling, modulation, amplification, and filtering function the manipulation of tempo, amplitude, and frequency response. The system performs resampling and pitch restoration proportional to the soundtrack's original labeled bpm and the regression result.

While the training set's bpm is between 40 to 80, the regression applies to a wide range of music. When the bpm of the music is higher than the range, we conceptually combine the multiple beats in a beat and map the bpm within the proposed range. For example, if the target soundtrack initially has a bpm around 120, we reconfigure the bpm as 120 over two so the player can control the speed with the synchronized swing pattern. Likewise, by rescaling the bpm of the target audio, the system can control the tempo of a wide range of music. To control the playback volume, the swing within 0.1 – 0.8 meters is mapped to a volume parameter within 0.5 to 2.5. This volume parameter is linearly applied to the volume of the playback. The project also focuses on timbre manipulation using filters according to the controller's orientation. When the pattern is classified with the head-down command, a low pass filter is applied to the playback to emphasize the low-frequency range of the soundtrack. In contrast, if the pattern is recognized with the head-up command, the playback emphasizes the high-frequency range with the high pass filter. Integrating the overall methods results in an interactive conducting-like pattern recognition using DNN with audio playback.

3.1.2 Experiment and Result of Conducting Project

To evaluate the system's accuracy, three subjects who were not involved in the training procedure tested the performance. They also were instructed on the system and tested commands. The testing of the system is performed with Bolero by Maurice Ravel. The music selected, generally maintained a uniform bpm with repetitive patterns so subjects and listeners could compare the variation.

To obtain a quantitative analysis, the subjects were asked to follow the given metronome sound drawing the patterns. The left part of Table 1 shows the results of the performance evaluation. According to the evaluation result, the average error of the temporal command is less than one bpm for every amplitude and swing type. This shows that the proposed DNN structure can efficiently extract temporal information with the features.



	Command	Error (bpm)	MSE
Tempo	No Boost	0.4682	0.6477
	Bass Boost	0.3243	0.5142
	Treble Boost	0.7852	0.6763

	Interpreted Command (%)		
	No Boost	Bass Boost	Treble Boost
No Boost	**81.51**	5.17	13.32
Bass Boost	31.64	**62.16**	6.20
Treble Boost	14.05	32.41	**53.54**

Table 1. Results of the conducting-like gesture's performance evaluation

The right part of Table 1 shows the result of the classification of the timbre control. The highest values are marked in bold. The majority of commands are classified correctly, but there is room for their accuracy to be improved. After the experiment, the subjects and audiences agreed with the result that the amplitude of the music and the size of the swing have a strong positive correlation.

3.1.3 Discussion for the Conducting Project

This case study developed a demonstration of the DNN-based music source conducting system in the MATLAB platform. The program classifies the temporal, amplitude, and timbre information with a single DNN regression model. Signal processing techniques enable the implementation of the corresponding commands.

The interactive demonstration is enjoyable and presents a new musical experience to players with familiar devices, including a smartphone and computer. Even though classical orchestra music is adopted for our demonstration, the music applies to diverse genres. The proposed gestural recognition system can be expanded to computer-synthesized music, MIDI, or live-code performance.

This case study was premiered at the International Computer Music Conference in 2018 as a paper and interactive installation [26]. Still, the proposed system methodologically has many limitations. The recent machine learning approaches that efficiently observe dynamic temporal information will perform better for more reactive and accurate conducting gestures. Nevertheless, this study shows that the interaction model based on machine learning methods has huge potential which can be developed with numerous applications.

3.2 Entangled: A Multi-Modal, Multi-User Interactive Instrument in Virtual 3D Space Using the Smartphone for Gesture Control

This case study introduces a gravity-like physics-based multi-user interactive instrument, Entangled, for multimodal composition in VR environments. For the last decade, each component of this instrument including VR, gestural interface, smartphone control, and multi-user participation has been widely studied and used in NIME [27, 28, 29, 30, 31, 32, 33, 34, 35]. Especially, Atherton's study [31] discusses and suggests new design principles to design the audience and performers' relationships through the VR instrument. Entangled uses smartphones to connect our gestures to the virtual domain with a virtual embodiment using gravity-like physics. Accordingly, the performers and the audience can observe the immersive VR projection and performers' gestures simultaneously rather than using their smartphone screens. Utilizing the expressiveness of gestures and mitigating their artifact, the gestural commands are not directly mapped to certain sounds or graphics. This aspect is explained in detail in the Interface section.

Entangled is designed to achieve a novel experience that can be only obtained when different modalities are naturally entangled rather than simply merging different components. This entanglement between modalities is the crossmodal correspondence in this multimodal instrument design. Maintaining crossmodal correspondences in digital creation provides the most veridical estimate of environment or stimuli by integrating the estimates from different unisensory perceptual that refer to the same object while keeping those estimates separately [3]. For instance, Tsiros's study [36] states an adequate design of multimodal mappings with preceding perceptual knowledge can improve not only human-computer dialogue but also the analytical, creative, and pedagogical virtue of user interfaces.

From this perspective, the principles of multimodal design, including interface, physics, graphics, and sonification, are elaborated throughout the research explaining how and why each methodology establishes crossmodal correspondence to other modalities.

3.2.1 Design Criteria

This instrument is multimodal, and its aesthetic basis is based on computer music theory. Music is an interactive activity coordinated with structured acoustic motivation to capture the subtleties designated by musicians [37]. The structured acoustic stimulus formalizes sonic narrative by the interaction between simultaneous sounds and a succession of events in time [38]. Entangled expands the above concept to the audiovisual domain by building correspondence between different modalities.

Crossmodal Correspondence

When a human gesture is directly mapped to a sonification equation, there would be countless good or less-good ways to implement the mapping function. Considering the physical constraints of gestures and the compositional goal, finding satisfying mapping can be challenging.

By adding another dimension of physics or data, the instrument can have additional opportunities for a natural chain of causality between gestures and sonification. This crossmodal correspondence can aid and inspire the artist by opening an unique experience.

Proper articulation between modalities and physics, data, rules, signal processing, machine learning-based gesture recognition, or generative algorithm can establish crossmodal correspondences. For example, in the Reactable [39], the visualized algorithmic system interacts with the materials showing the interface by placing and spinning physical materials on the surface. The interface sonifies correspondingly using various synthesis methods, including an oscillator, sample player, and resonant filters. In this case, visualized algorithmic data becomes another modality that connects the device's trigger and sound material, opening new potentials for multimodal composition.

Likewise, multimodal approaches creating the mapping between gesture and sound in interactive music systems have been proposed [40, 41]. For instance, Françoise's method [40] employed both direct mapping and temporal mapping based on Hidden Markov Models (HMM). In this method, the classification of temporal command is obtained by observing the previous stream of gestural data. These approaches show that temporal interaction can trigger particular temporal events at specific times. This multimodal interface allows manipulating the combination of continuous-time parameters and discrete-time events synchronized to the input gesture data. Consequently, players can correlate and utilize multiple modalities with crossmodal correspondence.

The following sections elaborate on the instrument's concept and methodology. Every modality, including physics, sonification, graphics, and interface, is highly correlated from the idea to its implementation. Figure 3.2 shows the overall structure of the system.

Physics and Graphics

While we can establish any rule in the virtual domain, establishing causality based on real physics helps build intuition for the users. In this instrument, gravity-like force is the medium that combines every modality. Gravity explains the force that ties all the objects in the generated virtual space and grants acceleration and velocity to them. When the linear gravitation control is activated, and the player accelerates the smartphone, gravity occurs in the virtual domain where the smartphone points toward. The magnitude of gravity changes according to the magnitude of the acceleration. The gravitation simul-

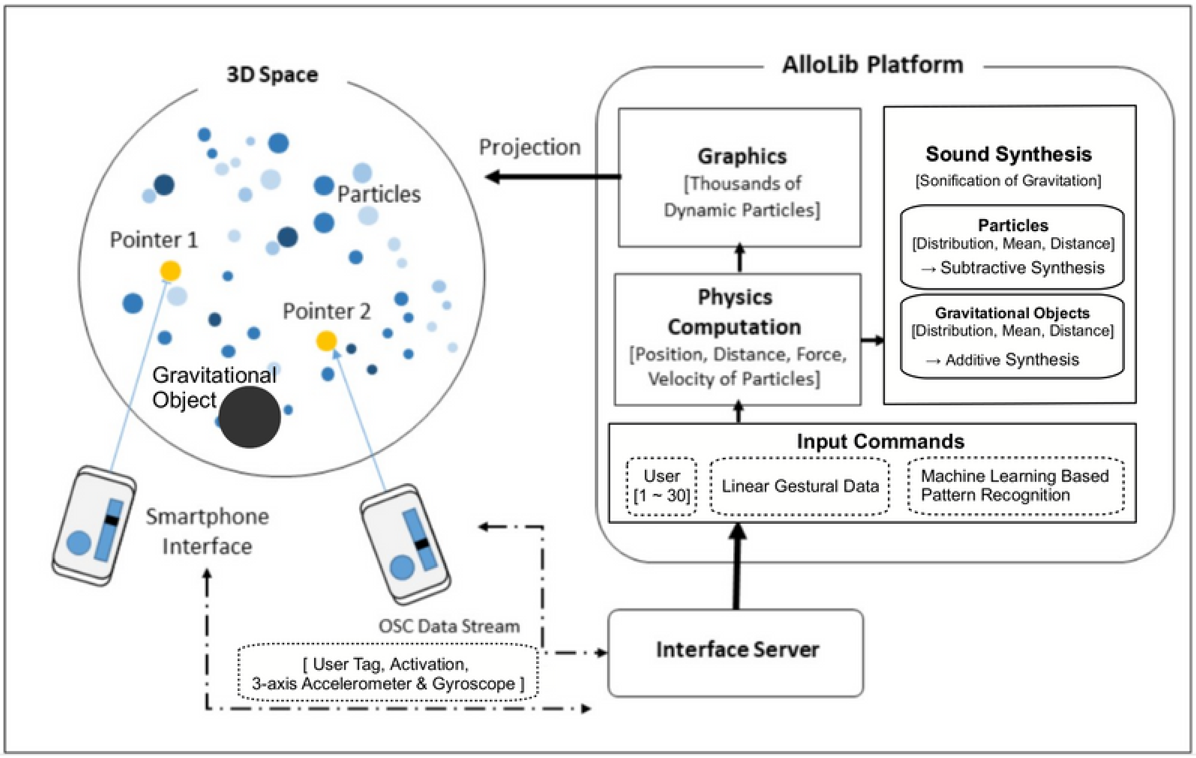


Figure 3.2: Overall structure of the instrument 'Entangled'

taneously affects thousands of particles in the virtual space, and the particles respond simultaneously.

The participants can also generate gravitation using specific gestural patterns using machine learning-based pattern recognition. By drawing an inward spiral using their smartphone, a black hole-like gravitational point pulls particles into the point. The gravitational point is visualized as a dark blinking sphere. In contrast, an anti-gravitational point occurs when an outward spiral is drawn pushing particles away from the center. This object is visualized as a white blinking sphere. These special objects blink and vibrate faster when it is close to lapse.

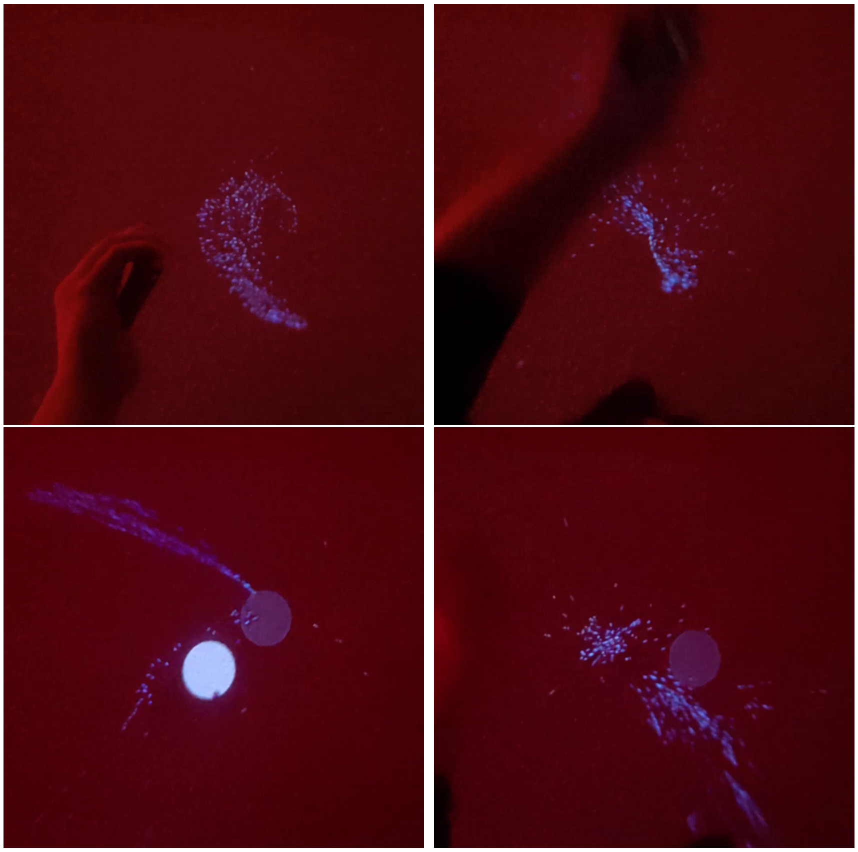


Figure 3.3: Exemplary demonstration of 'Entangled'

Interface

Gesturally controlled sonification methods have been actively proposed in NIME research [34, 35, 30]. Since the electronic medium allows the interface to be coupled or decoupled from the sound synthesis, it is crucial to determine how and why the interface uses the gestures. As described in Figure 1.5, mutualizing the gesture and interface identifies the properties of the instrument with three parts: input, mapping, and output. Since the human gesture is a smooth and continuous change of the limb to the body, the direct connection of the gesture to the interface may constrain the compositional process which sometimes has to be triggered and prepared. For instance, relying synthesis only on a direct mapping of the position of limps using a tracking system may limit the profound utility of human gestures.

In Entangled, as described in the physics and graphics section, the interface has two different types of commands: 1) gravity-like force using continuous gestures, and 2) generation of gravitational objects. These commands can be triggered using corresponding buttons. By separating linear and non-linear event buttons, the interface takes advantage of both the profound utility of human gesture and discrete control in higher dimensions at the same time.

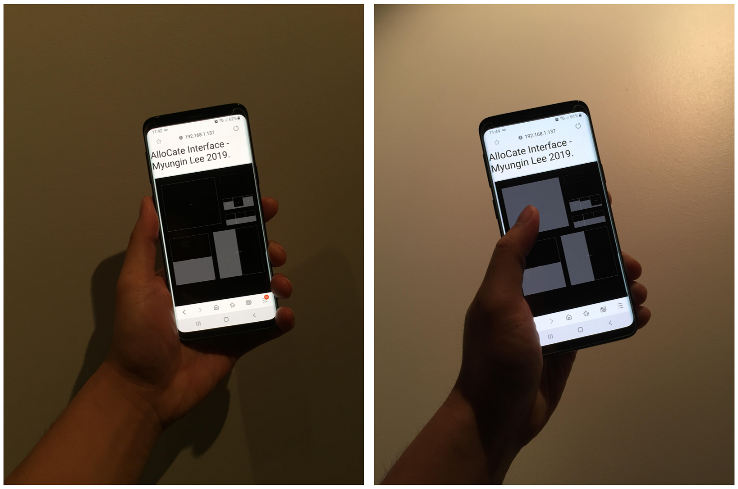

AlloCate Interface - Myungin Lee 2019.
AlloCate Interface - Myungin Lee 2019.

Figure 3.4: Interface.js based smartphone interface for 'Entangled'

Figure 3.4 shows the javascript-based interface [42] sending the gestural data and simple button commands through OSC to the computer program. When these buttons are off, the player's gestures do not affect the virtual world. Gravitation using continuous gestures occurs when the player triggers an activation button on the smartphone screen and the smartphone's acceleration is higher than a small threshold designed to mitigate the artifacts of the noisy sensor data.

On the other hand, gravitational objects, including gravity and anti-gravity points, require machine learning-based gestural pattern recognition classifying gestural patterns such as an inward spiral and an outward spiral. When the player triggers a pattern recognition button on the screen, the 6-dimensional stream of data (3-axis accelerometer and 3-axis gyroscope) is observed and classified.

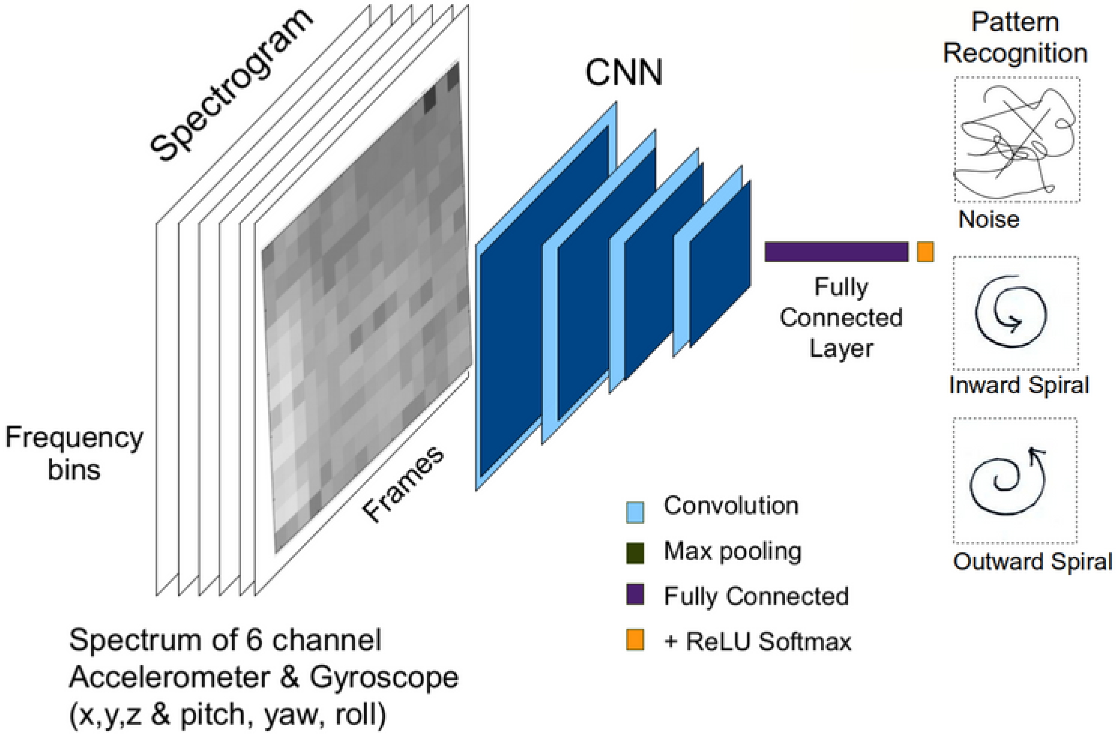


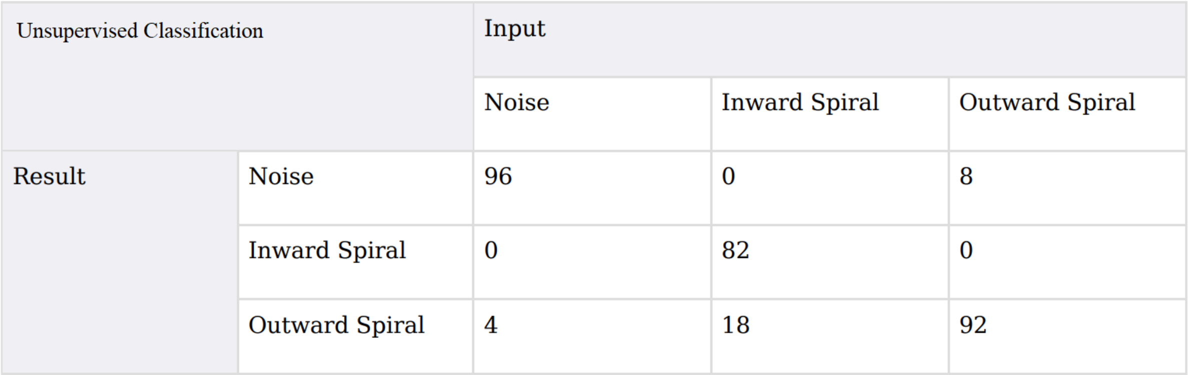
Figure 3.5: The structure of the adopted machine learning model

For the implementation, this project adopts a VGG-Style convolutional neural network for a multichannel spectrogram [43]. This model is known as a state-of-the-art image classification method and is widely used for various single or multi-channel data recognition [44, 45, 46]. While conventional study with Support Vector Machine (SVM) for 2D gesture exists [47], this project adopts the CNN-based method allowing additional expansion to more complicated patterns in 3D maintaining real-time executable complexity. Furthermore, this classification task differentiates from 2D character recognition because an inward spiral and an outward spiral can have the equivalent 2D footage if the temporal information is neglected.

In this algorithm, 256 samples of gesture data for every 6-channel data are marked as six different 17 * 17 spectrograms and adopted as the input of the CNN. The number of samples, 256, corresponds to about 5 seconds in time scale considering the employed sample rate of the gestural sensors. A window size of 32 and a 50% overlap derives certain spectrograms. Figure 3.5 shows the adopted multichannel convolutional neural network model. This implementation can be optimized with recent machine learning models which represent temporal information better.

To train three different gesture patterns (inward spiral/ outward spiral/ noise), I produced 250 incidences with various speeds, orientations, and sizes for each pattern.

The training set used 200 incidences and the remaining 50 incidences are used for the unsupervised test set. Table 2 shows the results of the unseen test. While the size of the dataset is not large, the test results show that the system effectively filters noise data and classifies the commands sensibly. Still, this requires further study for the advanced usage of additional commands. For the initial version, Pytorch C++ API was used for training, testing, and implementation. The current version is a work in progress to use generic Python-based Pytorch in interactive form.



Unsupervised Classification		Input		
		Noise	Inward Spiral	Outward Spiral
Result	Noise	96	0	8
	Inward Spiral	0	82	0
	Outward Spiral	4	18	92

Table 2. Classification result from the unseen testset

The proposed machine learning-based gestural pattern recognition increases the functionality of gestural commands to intervene in non-linear events in the linear timeline. Still, the specification of this experiment can be modified or optimized for different objects or resolutions. This investigation argues for the importance of experiment design considering the goal of the overall system rather than stating a specific performance of a certain model and dataset.

Sonification

Audio plays an essential role to represent the gravity of thousands of particles and their shape. This concept is motivated by Newton's third law stating every force in nature has an equal and opposite reaction. By sonifying the gravity, the player can experience a novel and immersive representation of physics which coordinates with the player's gestures and graphics.

The system models the shape of gravitation in the sonic spectrum using a stochastic model and sculpts the noise using subtractive synthesis and additive synthesis instead of sonifying the gravity of every single particle.

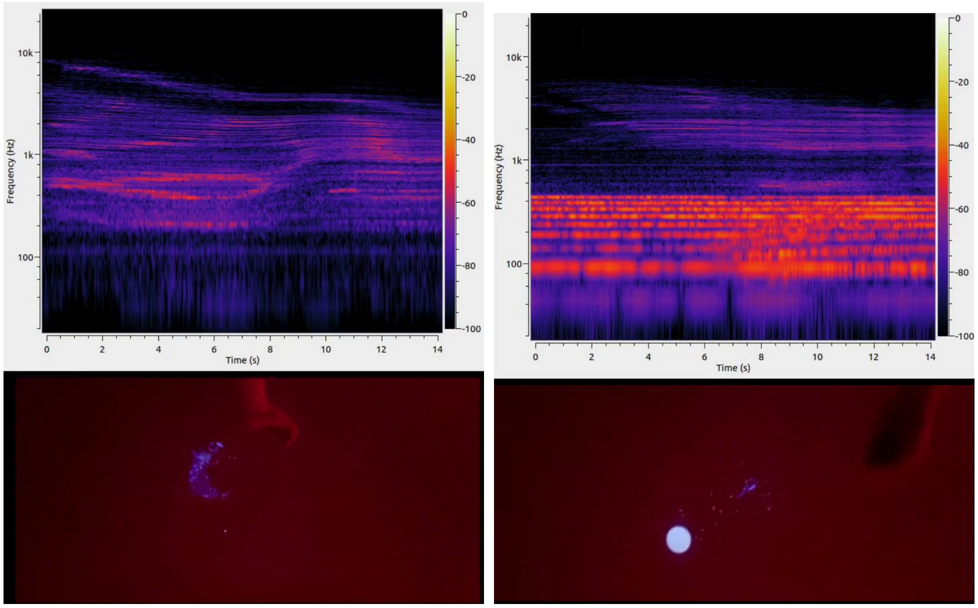


Figure 3.6: Audio spectrum matched with the gesture with the graphics

The average value of gravitation determines the cutoff frequency and the distribution value decides the bandwidth frequency of the subtractive synthesizer. The gesture triggers the synthesizer stochastically when the acceleration of the gesture is higher than the threshold, which is similar to the ringing bell.

Gravitational point and anti-gravitational point generate a harmony of bass frequency using additive synthesis based on the energy of gravity affecting the particles.

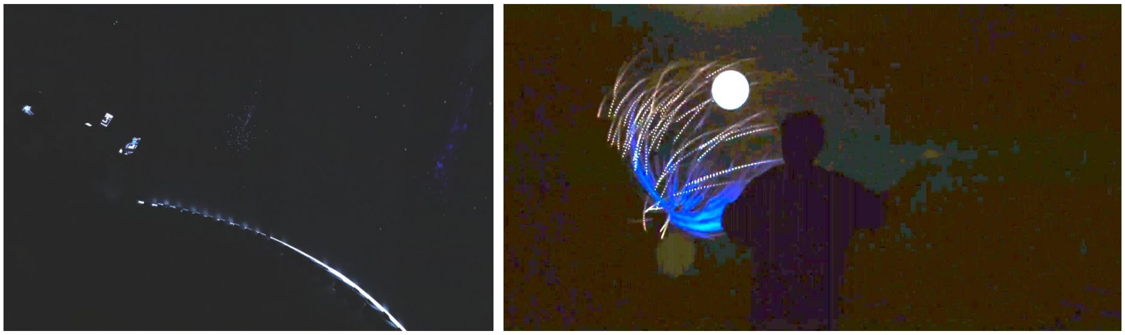


Figure 3.7: The AlloPortal, a prototype and exhibition space

Figure 3.6 shows the frequency spectrum of audio over time and the corresponding interactive audiovisual at the moment. To compare left and right in Figure 3.6, the stream of the spectrum is widely spread over a wide band when the particles are spread sparsely over the range. The right picture in Figure 3.6 visualizes the anti-gravitational point in the frequency domain which shows dominant power in the low-frequency band.

Even though the underlying role is straightforward, the player can create complexity or simplicity, intervals or morphologies, intuitionism or formalism, which can form a sonic narrative by combining these sonic elements.

Multi-User and Scale

This instrument allows a multi-user experience and has been tested by up to eight participants simultaneously while it can systematically allow more users. Using a web interface toolkit, Interface.js, the participants join the wireless network using their smartphone's browser without installing any application. As gravity is the medium of every modality, the multi-user experience generates more profound and dynamic physics in the system so that the players can explore ensemble-style composition with virtuosic potential, which could not be achieved from a single user's experience.

Entangled delivers the best immersive multimodal and multi-user experiences when played in the AlloSphere [48, 49]. This environment presents a seamless 3D surround vision for dozens of audiences at the same time in a shared virtual world using omnistereo imaging with their physical presence.

In this experience, the player's physical presence is especially important compared with conventional VR experiences since the player can observe the surroundings, including other players' gestures or facial expressions, to communicate, cue, and interact. Through this process, the participant can establish the crossmodal correspondence between other players' gestures and the virtual physics generating audiovisuals.

The fundamental platform of the instrument, AlloLib [50], allows its installation to be scaled from a laptop to a distributed system of any size with multiple displays. For example, the demonstration of the instrument for this study is made with multiple participants in the AlloPortal that accommodates a large frontal stereo projection screen. Figure 3.7 shows the AlloPortal prototype and exhibition space.

Likewise, the size of the performance is adjustable, and this implies the experience can not only be condensed but also be scaled even larger than the AlloSphere if the facility allows it. This allows the limitless opportunity for multi-user creation and further instrument designs.

Discussion for Entangled

This case study was premiered in NIME 2021 as a paper [51]. It elaborates on Entangled's design criteria, methodology, and implementation process. This research argues that a unique experience can be obtained when different modalities are naturally intertwined when the players and audience easily understand the relationship. Entangled has been demonstrated with different groups of participants over the years, including computer science students, media artists, and electronic musicians. After a brief explanation, the audience was extemporally invited to play Entangled with their smartphones. The participants and audience could intuitively understand how their gestures affect gravity in the virtual domain and relate it to the audiovisuals. This provoked the participants' curiosity to play Entangled proactively and dynamically with their gestures.

The instrument's name, Entangled, not only indicates the entangled states of gravity with particles but also connotes that various modalities are jointly established with crossmodal correspondence. The implementation on a single C++ platform, AlloLib, is suitable for building real-time interactive audiovisual applications. As a developer and the first player of Entangled, I have been writing exemplary phrases with narratives to make a multimodal piece [52].

For future works, I have been exploring more possibilities of multi-user experience and developing further sonic or visual elements for the next version of the instrument since machine learning-based gesture recognition allows more complicated commands to be involved. This case study demonstrates the practice of a scientific and artistic narrative allowing the general audience to gain insight and the scientists to be inspired by the nature of modern physics interacting with a multimodal experience.

3.3 AlloThresher: Multimodal Granular Synthesis

My case study of developing smartphone gesture-based instruments has continued over 6 years and has yielded diverse versions of instruments. Some of the instruments were involved in the CREATE laptop orchestra and performed regularly (Figure 3.8). Developing this idea over the years, the AlloThresher instrument was invented.

Figure 3.8: Performances of the CREATE laptop orchestra (2018, 2019)

AlloThresher is a multimodal instrument with audiovisual granular synthesis using the gestural interface. Granular synthesis is a sound synthesis method that creates intricate tones by combining and mixing simple micro-sonic elements called grains. Granular synthesis is often considered sound clouds, dynamic regions of sonic grains that move around the sonic landscape. Composers, including Iannis Xenakis, Barry Truax, and Curtis Roads, frequently utilized granular synthesis to shape large masses of sound with unique timbres [53]. Recently, EmissionControl2, an open-source standalone interactive real-time granular synthesizer with sound file granulation, was introduced [54]. Meanwhile, having many different parameters to control using the GUI is one of the challenges of using granular synthesis, especially in real-time.

From this observation, I have been developing a multimodal interactive granular synthesizer that utilizes two smartphones on each hand. The gestural interface interpreted from the smartphones' sensors enables the player to precisely and intuitively decide and play the multiple parameters of the granular synthesis in real time.

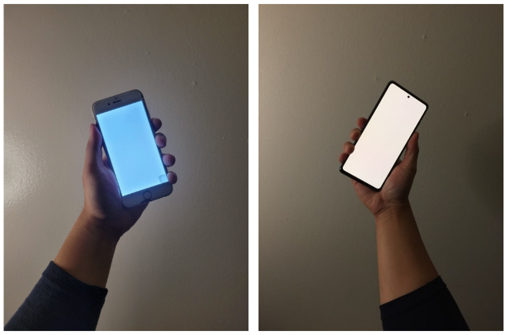


Figure 3.9: Two smartphones as controllers on each hand

Conceptually, the left hand decides the source to granulate, and the right hand grinds the grain. Technically, the left-hand gyroscope emanates the directional information of the gesture, which determines the sound source (y-axis), grain position (y-axis), and reverberation.

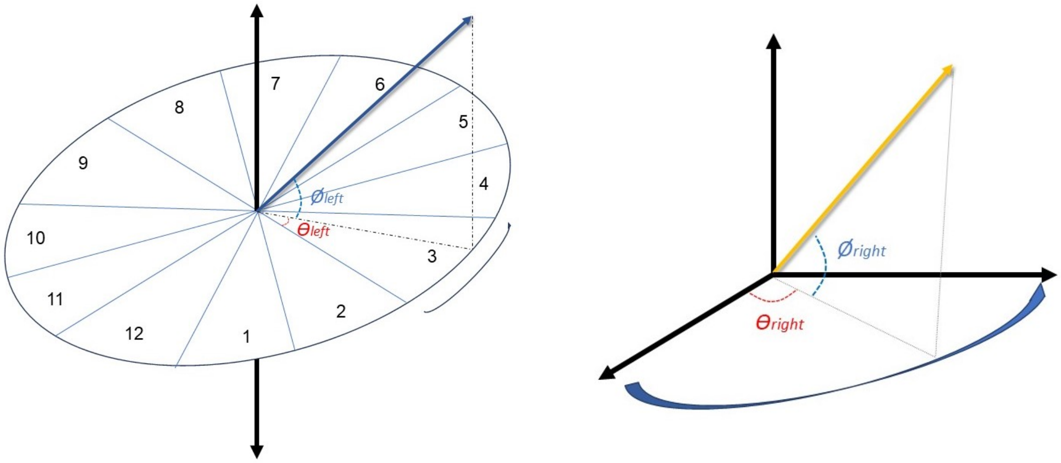


Figure 3.10: Parameter map of the left and right controllers

Figure 3.10 shows the exemplary parameters of the left and right controllers. The blue bold line indicates the left controller's vector, and the yellow bold line on the right points to the right controller's direction. In this case, the third sound source is selected from 12 sound sources. The grain position is $\theta/30$ between 0 to 1 and $\phi/90$ amount of reverberation between 0 to 1. The right controller behaves more dynamic role to actuate the grains. Its absolute acceleration determines the amplitude and duration of the synthesis, vertical direction ϕ_right decides the grain rate, and the horizontal direction θ_right playback rate from -1.5 to 1.5.

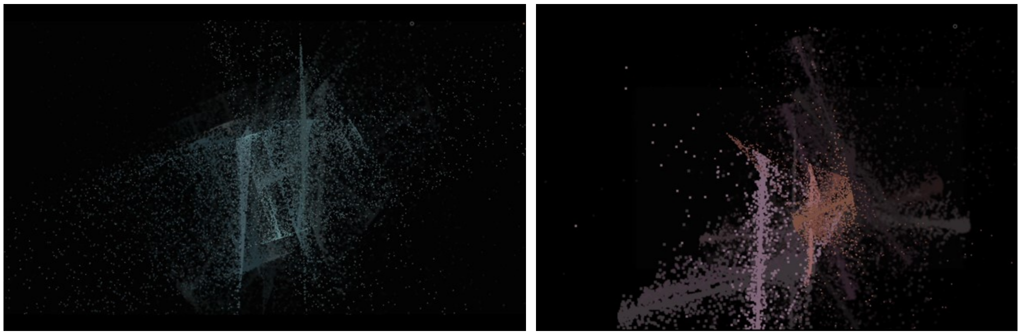


Figure 3.11: Screenshots of AlloThresher showing the dynamic visualization.

Figure 3.11 shows the visualization of the instrument. The corresponding visuals, inspired by the clouds of sound, are generated simultaneously for each granule based on the spectrogram of the sound that morphs and blends dynamically with the gesture. The color represents the granular source.

The genuine value of the instrument can be experienced when it is performed multimodally. Figure 3.12 shows the multimodal performance of the AlloThresher in the AlloPortal. The audiovisual demonstration is available online [55]. After the performance, the audience was invited to play the instrument too.

By breaking conventional interfaces like knobs and sliders, this seamless connection between modalities utilizes the profound advantage of the gestural interface. Moreover, the presence and gesture become part of the space and the performance so that the audience can simultaneously observe and cohesively connect the audio, visual, and interface. This instrument inspired many artists, and now I am collaborating with artists who want to compose with this instrument.

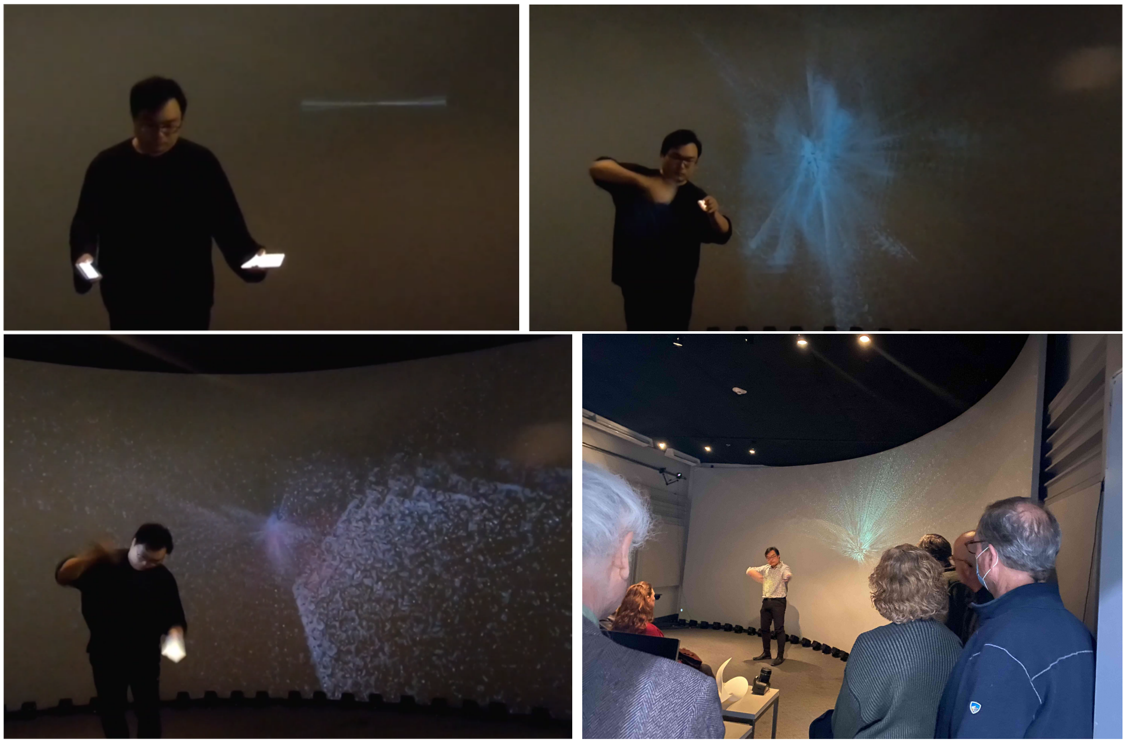


Figure 3.12: Multimodal performance of the AlloThresher

While some modern digital media arts tend to focus on the novelty of a specific technology in a single domain, this case study and instrument suggest there are unique and creative opportunities when the multimodal digital instruments are designed cohesively over the different modalities.

3.4 Multimodal Sonata for Real-time Interactive Simulation of the SARS-CoV-2 Virus

“Multimodal Sonata for Real-time Interactive Simulation of the SARS-CoV-2 Virus” is collaborative work with scientists and artists and myself as an instrument designer. The project focuses on SARS-CoV-2, known as coronaviruses presenting the interactive audiovisual experience in the virtual domain. Section 3.4.1 in Chapter 3 describes the fundamental biological discoveries resulting from this collaboration. From there, we designated the narrative of the simulation, elaborating on the interaction between the coronavirus and the human body. The following chapters deal with specific domains of the project, including simulation (Section 3.4.2), graphics (Section 3.4.3), audio (Section 3.4.4), interface (Section 3.4.5), and artistic representation (Section 3.4.6). These chapters provide a comprehensive examination of the case from a theoretical, technical, and artistic point of view.

3.4.1 Scientific Backgrounds

Over the years of the pandemic, the SARS-CoV-2 virus has been of significant interest in all scientific disciplines, especially research into how the coronavirus spikes can bind to a human cell component and how these findings can be utilized for therapeutic purposes [56, 57]. In 2020, biophysicists in the Department of Applied Experimental Biophysics at Johannes Kepler University showed the first short movie of the spike protein in action using a high-speed AFM. The behavior and dynamics of the spike protein were recorded under physiological conditions. In addition, the derivative technique of AFM successfully measures the interaction forces between the virus’s spike protein and the ACE2 receptors expressed on the cell membrane. Successful binding to ACE2 requires at least one of the three receptor-binding domains (RBDs) on a Spike trimer. The dynamic movement of the spike protein underlies the process of virus attachment to human cells during the early stages of viral infection. This scientific discovery describing the movement of the spike protein of SARS-CoV-2 in real-time has been introduced [58, 57].

3.4.2 Simulation of SARS-CoV-2 and Human Cell

This case study embraces the discoveries by bringing them into one programmable simulation. To achieve a comprehensive understanding of a complicated biosystem, we established a narrative based on the mechanism of SARS-CoV-2 propagation, and the narrative is developed into an interactive audiovisual. To achieve a coherent multimodal experience, the software is built on a single C++ platform using open-source libraries including *AlloLib, OpenGL, RtAudio, and OSC* [50, 59, 60, 12] with interface.js [42]. This method allows for access to the model establishment, data management, rendering,

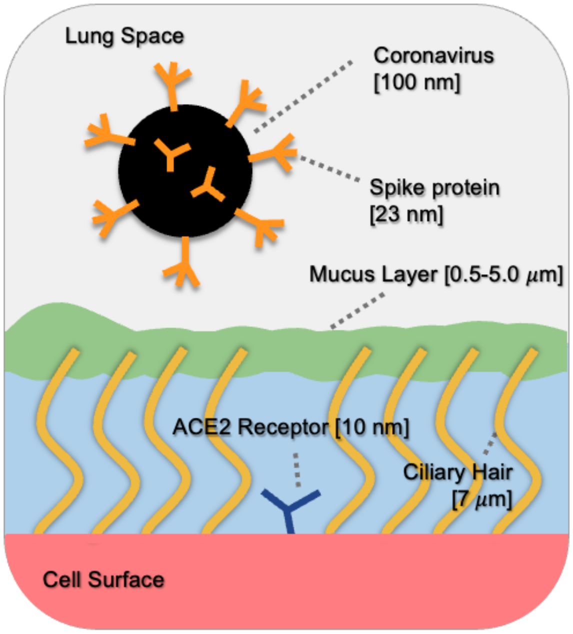


Figure 3.13: Schematic image of components in the simulation with the actual sizes.

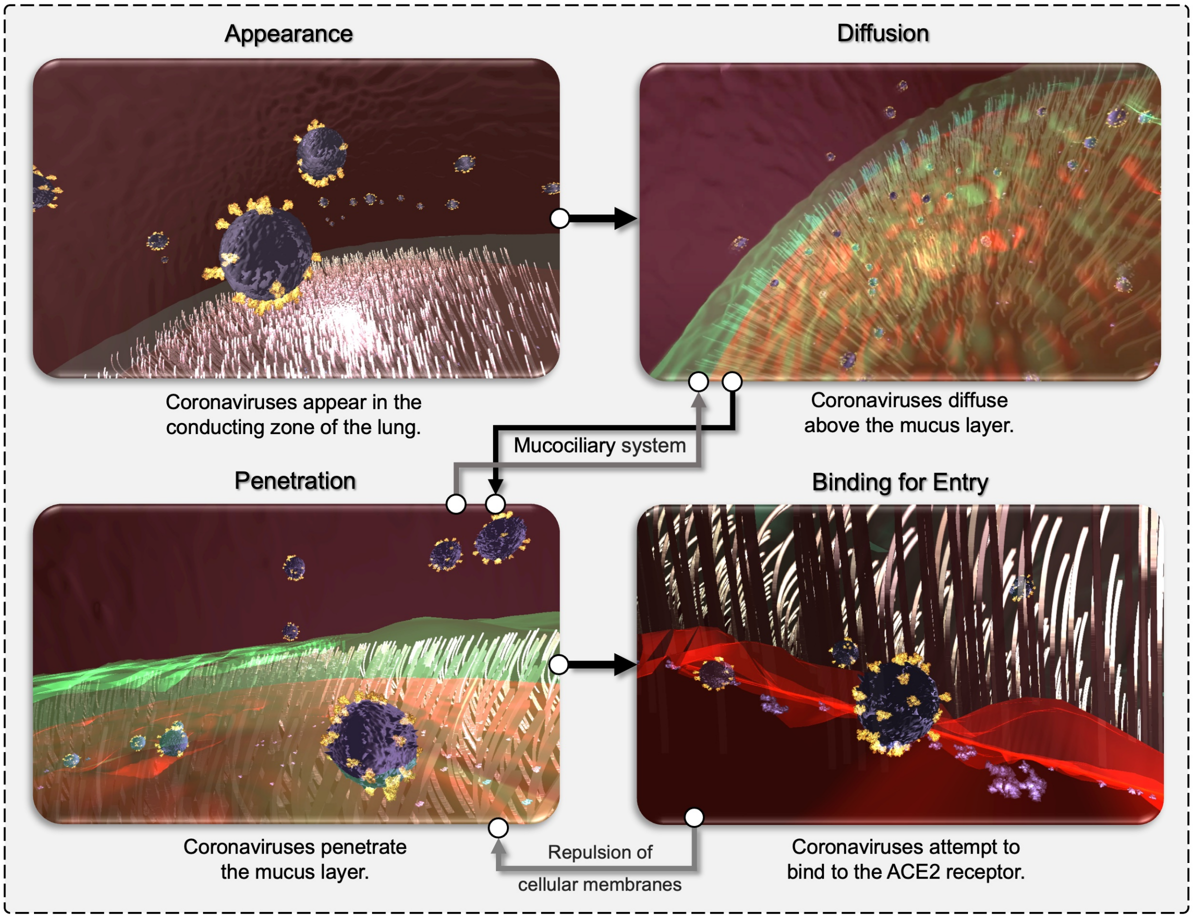

Appearance
Diffusion
Coronaviruses appear in the conducting zone of the lung.
Coronaviruses diffuse above the mucus layer.
Mucociliary system
Penetration
Binding for Entry
Coronaviruses penetrate the mucus layer.
Repulsion of cellular membranes
Coronaviruses attempt to bind to the ACE2 receptor.

Figure 3.14: Four stages of SARS-CoV-2 infection toward the lung along the airway: Appearance, Diffusion, Penetration, and Binding for Entry

and sonification. Consequently, we can interactively control myriad agents in the system with the composition of virtual and organic 3D audiovisual environments in real-time. Figure 3.13 shows a schematic illustration of the simulation's components, and Figure 3.14 depicts the infection mechanism. Figure 3.15 shows the process of achieving the graphics of the coronavirus spike from the AFM video. Figure 6 depicts the overall layout of the interactive multimodal design, and the following sections describe the methodology of simulation in detail mapped with the corresponding figures.

Narrative: Infection mechanism

This project presents the narrative based on the infection mechanism of virus within the lung along the airway. The narrative is composed with four stages (Figure 4): *Appearance, Diffusion, Penetration, and Binding for Entry.* These stages align with the steps of musical plot structure, such as the *Exposition, Development, Recapitulation, Climax.* This bond between the scientific mechanism and artistic representation strengthens the coherence of the overall portrayal.

Appearance (Exposition): The narrative begins with the appearance of the coronavirus as it follows the route of infection towards the lung along the airway.

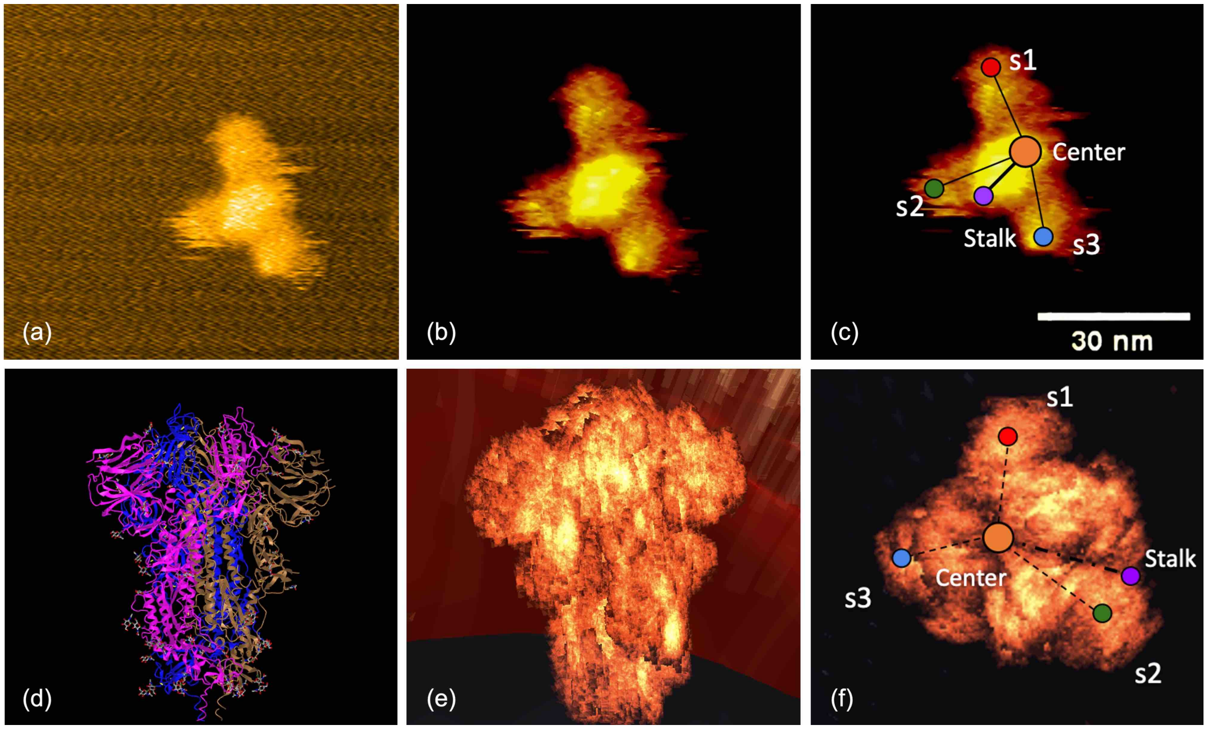


Figure 3.15: (a) Spike trimer of the SARS-CoV-2 virus observed with a high-speed atomic force microscope (AFM). (b) denoised frame using a deep neural network. (c) labeled frame. (d) 3D model of the SARS-CoV-2 virus spike trimmer. (e) rendering of the spike in the project. (f) rendered spike with the labels.

Diffusion (Development): A recent study by Johnson et al. [61] shows the behavior of viruses at the early stages of infection using imaging microscopy and 3D tracking. Inspired by these observations, a diffusion model simulates the movement of multiple agents with Brownian-like motions. In addition, a steering algorithm demonstrates the natural behavior of coronaviruses fluctuating the lung space and seeking cell-surface receptors. An exemplary trace of multiple coronaviruses is depicted in Figure 3.16(c). This algorithm is controlled by three parameters: *urging, grouping, and hunting.* The parameter *urging* determines how fast the viruses are accelerated in an random direction, *grouping* determines the tendency of the viruses to swarm, and the *hunting* parameter drags the viruses towards the ACE2 receptors. We decided on the presets of these parameters empirically, which change over the stages composing the flow of the entire narrative.

Penetration (Recapitulation): The ciliary hairs and mucus layer function as the primary defense mechanisms of the airway [62]. The slimy surface of the mucus layer greatly slows down the propagation of the virus and repels it from the ACE2 receptors. At this stage, some of the coronaviruses are removed from the mucus layer, recapitulating the penetration stage.

Binding for Entry (Climax): Coronavirus tries to bind with ACE2 receptors on the cell after it pierces the mucociliary defense mechanism. The spikes stretch and morph to bind to the receptor when the virus is close to an ACE2 receptor. The dynamics are programmed based on binding force data [58]. The virus has about a 30 % chance of binding to the receptor and entering human cells when the repulsion between the virus and the cellular membranes can be affected with sufficient energy [63, 64]. The narrative ends and loops back to the beginning when 90 % of coronaviruses bind to the receptors.

This process can be accelerated, decelerated, stressed, released, or initiated by the

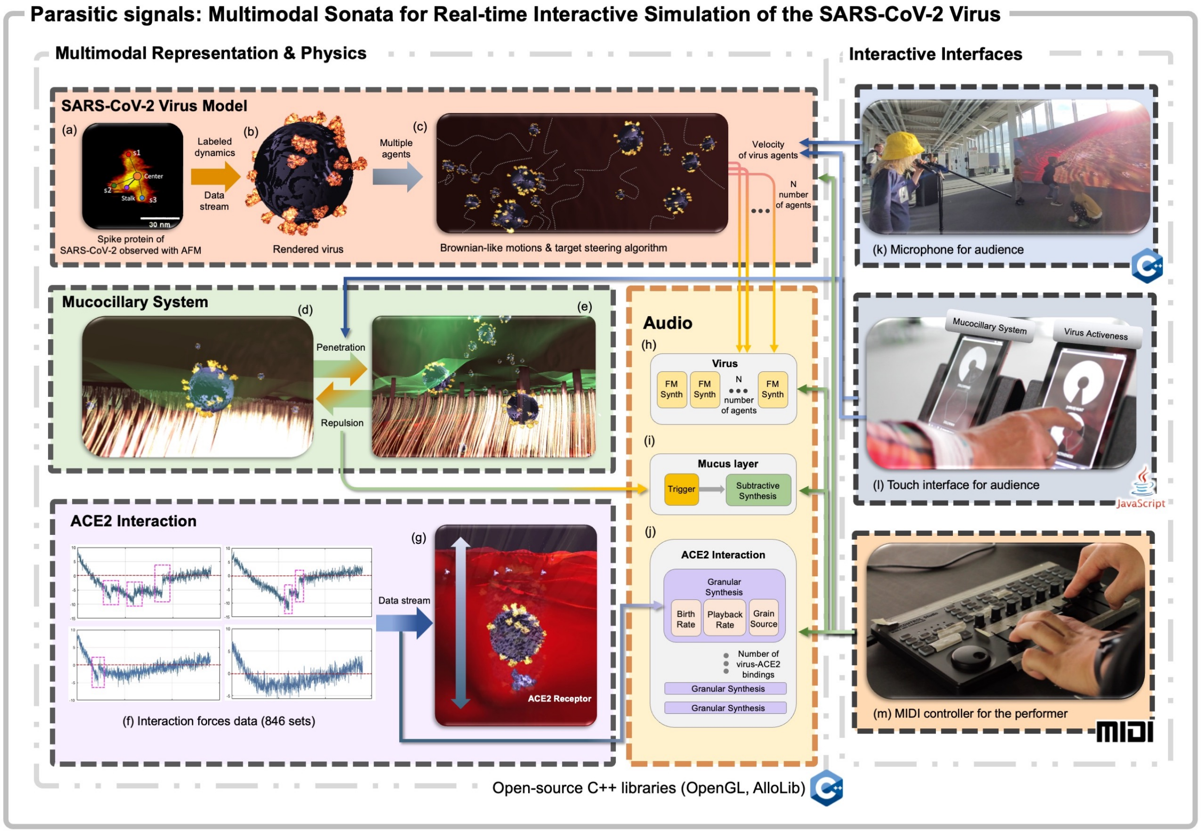


Figure 3.16: Overall structure of multimodal design. (a) spike protein of SARS-CoV-2 observed with AFM. (b) Rendered SARS-CoV-2. (c) an exemplary movement of coronaviruses. (d-e) the moment coronavirus penetrates the mucus layer. (f) interaction forces data observed with AFM. (g) coronavirus trying to bind with ACE2 receptor. (h) sound of coronavirus: frequency modulation. (i) sound of mucus layer: subtractive synthesis (j) sound of ACE2 receptor and coronavirus interaction: granular synthesis. (k) microphone for audience interaction. (l) touch interface for audience interaction, mucociliary system and virus activeness. (m) MIDI controller for the performer.

user's interaction in composing a musical structure. The interaction design is described in Chapter 3.4.5.

Probe: Atomic Force Microscopy (AFM)

We acquired the position of the stalk and each of the RBDs over time from AFM footage data. As shown in Figure 3.15(c), the spike center, three RBDs, and stalk were manually labeled initially. Using the ground truth length of the stalk and RBD, we derived the 3D position of three RBDs by regarding the length of the RBDs from the spike center and the angle using trigonometric functions [58, 57]. The 3D structure of the SARS-CoV-2 spike protein trimer, Figure 3.15(d), [65] was employed and we separted a branch from the trimer. Dynamic 3D spikes were reconstructed digitally, Figure 5(e-f), based on the real protein model, incorporating observed dynamics from temporal AFM data by rotating and scaling the individual branches.

The derivative technique of AFM can depict molecular scale dynamics and this technique enables the observation of mechanical interaction forces between the spike protein of the virus and ACE2 receptors expressed on the cell membrane. The initial interchange between the human protein ACE2 receptor and the coronavirus takes place at this moment. Figure 3.16(f) displays four cases of ACE2 interaction from hundreds of experiments. In the graph, the y-axis indicates the interaction force data and the x-axis represents the duration over 10 seconds. We can observe the impulsive spark, labeled as purple boxes in Figure 6(f), which tells the number of RBDs bonded with the receptor from the data. Likewise, the hundreds of experimental data offer diverse cases of human protein-coronavirus interaction. In this simulation, the coronavirus collides with the cell surface and ACE2 receptors based on randomly selected incidences from the hundreds of cases. This simulation provides a dynamic and profound scenery above the cell surface during the *Binding for Entry* stage.

Defense mechanism: Mucociliary system

Figures 3.16(d) and 3.16(e) illustrate the moment when the coronavirus pierces the mucus layer. When the coronavirus contacts the surface of the mucus layer, a slim wave occurs, causing oscillations in the layer. We implemented a simple 1D wave on the ciliary hairs that change its fluctuation based on the activity of the defense system. The functions and effects of the interactive parameters are detailed in Chapter 3.4.5.

Human protein

: Angiotensin Converting Enzyme 2 (ACE2)

ACE2 is a viral receptor that the coronavirus spike protein can find and combine. The virus enters human cells if the viral spike protein successfully binds to ACE2. The dynamics of ACE2 dimers are simulated with random angular spin that ranges from 0 to 50°, inspired by research on the flexibility of ACE2 in the context of SARS-CoV-2 infection [66]. The simulation elaborates on the interaction between ACE2 dimers and spikes of the coronavirus through AFM sonification. Chapter 3.4.4 provides detailed information on how this simulation involves the sonic domain into multimodal representation.

3.4.3 Graphics

The graphics of this project contain preprocessing and real-time rendering. Figure 3.15(a) displays a screenshot of the raw real-time AFM video, and Figure 3.15(b) shows a denoised frame using a deep neural network [67]. This process has enabled a more precise pictorial estimation of the coronavirus on the nano-scale.

An instanced rendering method was utilized using OpenGL Shading Language (GLSL) to depict rapid and dense scientific models with convoluted details. Instanced rendering is a computer graphics technique that optimizes the drawing of multiple objects by reusing a single instance, such as a triangle [59, 50]. For instance, the 3D model of a spike, Figure 3.15(e), is rendered using the structure of the SARS-CoV-2 spike protein trimer, Figure 5(d) [65], by filling out 125,632 vertexes with translucent instances. This approach stuffs the skeleton of the 3D model, presenting imagery that closely resembles the original AFM image. To perform efficient real-time rendering of considerable coronaviruses, each with 35-40 spikes, the number of vertices can be sampled.

Furthermore, real-time frame feedback is used to illustrate the nano-scale scenery in the human body. This involves duplicating the frame buffer into a texture and adding it to the following frame. This real-time postprocessing technique ensures a smooth transition of the scenery.

3.4.4 Audio

Audio in this project is a vital modality that represents the events coherently in the scenery. The agencies including coronaviruses, the mucus layer, and ACE2 receptors, behave as individual sources for various digital synthesizers. The yellow box in Figure 3.16 depicts how each element synthesizes the audio. Regarding the dozens to hundreds of agents involve in real-time simulation, innumerable sonic events happen simultaneously in the scenery, which requires an all-around strategy during the design procedure.

In order to designate a coherent multimodal structure, we paired the narrative and the infection mechanism with a sonic structure based on a music-compositional approach. Figure 3.17 depicts an exemplary mono-channel auditory structure using a spectrogram with labels that explain the interconnected stages. This musical structure spawned by diverse sonic elements goes back and forth but eventually flows to the end following the probability model of the events described in the narrative chapter.

Each coronavirus emits violin-like audio at its 3D position, synthesized using frequency modulation (Figure 3.16(h)). Their speed and density determine their frequency and modulation parameters. The audio evokes a sonata-like structure. In the initial stage, the *Exposition*, dozens of coronaviruses gradually appear, forming an ambient soundscape. In the *Development*, the coronaviruses begin to diffuse and develop a chaotic 3D orchestral-like structure. As the diffusion progresses, the coronaviruses spike the mucus layer causing visible fluctuations on the mucous surface emitting low-pitch wave-like sounds created using subtractive synthesis (Figure 3.16(i)). These wave-like sounds re-

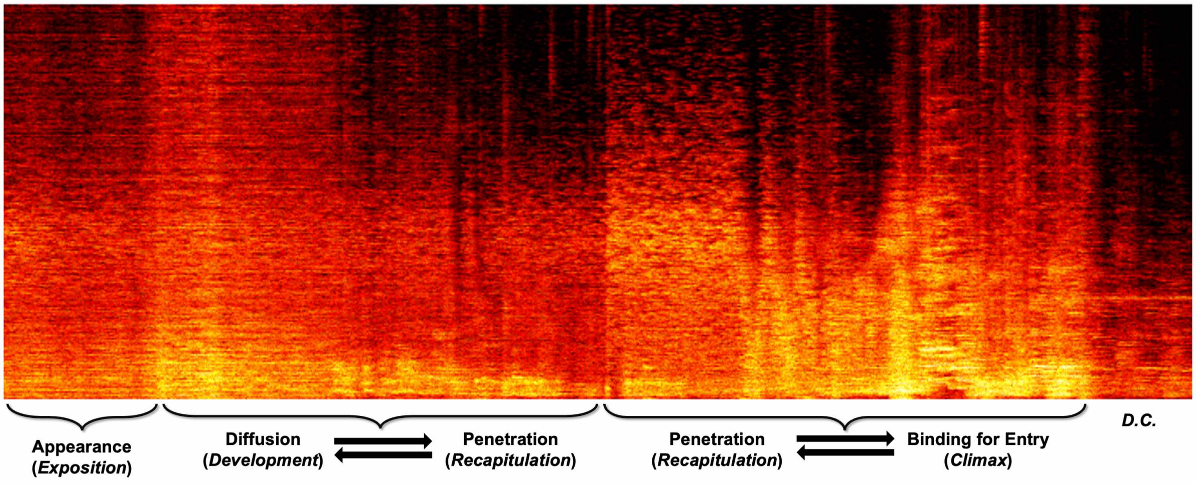


Figure 3.17: Linear-scale spectrogram of an exemplary narrative over 4 minutes

lieve the tension generated by the chaotic virus sounds. These two opposing sonic objects compose the buildup and release of tension, shaping and recapitulating the overall structure.

When the coronavirus interacts with the ACE2 receptor, the AFM data is employed to parameterize the duration, pitch, and birth rate of the granular synthesis (Figure 3.16(j)). We effectively emphasize the singular points of the AFM data (Figure 3.16(f)) by applying the data to multiple parameters of the granular synthesis. This approach represents the intense interaction between the coronavirus and the ACE2 receptor while also comprising symbolic messages through its granular sources, such as coughing, whispering, crying, and laughing sounds. These intricate tones intensify the tension directing to the *Climax* as multiple coronaviruses bind to ACE2 receptors. Finally, the narrative returns to the birth with a *Da Capo*, reprising the succession. This real-time sonic landscape not only audifies the data or theory but also comprises a musical structure that aligns with the occasions in the simulation. As a result, the audio is filled with dynamic and unusual patterns intertwined with the simulation, visualization, and narrative.

3.4.5 Interface

The interface invites the audience to participate in the real-time audiovisual narrative. A voice-based interface and touch-based tablets allow users to intuitively probe the dynamic biophysics of the human cells and coronavirus. Figures 3.16(k) and (l) show the installation spaces and demonstration of the interaction.

Participants can collaboratively build the audiovisual narrative by manipulating the defense mechanism (the activeness of the mucus layer & mucociliary clearance) and the activeness of the virus (diffusion of viruses & seeking ACE2). This interactive setup enables the audience to actively participate in the narrative unfolding.

In addition to the touch interfaces, a microphone captures the voices and noise within the installation space. This symbolizes human activity and the characteristics of coronavirus transmission through the respiratory tract. As the microphone input increases, the mucociliary system responds by showing oscillations in color and height. Simultaneously, the virus becomes vibrant, and its spike structure changes morphologically. These transformations raise the likelihood of the virus infiltrating the mucociliary system and binding to the ACE2 receptor.

The design also offers explicit control over the parameters of the simulation through a MIDI controller (Figure 3.16(m)). The controller furnishes various options for audiovisual manipulation, such as adjusting time, speed, camera angles, and corresponding audio effects. This functionality extends the options for audiovisual composition, allowing the performer to improvise or plan the narrative, similar to playing a musical instrument.

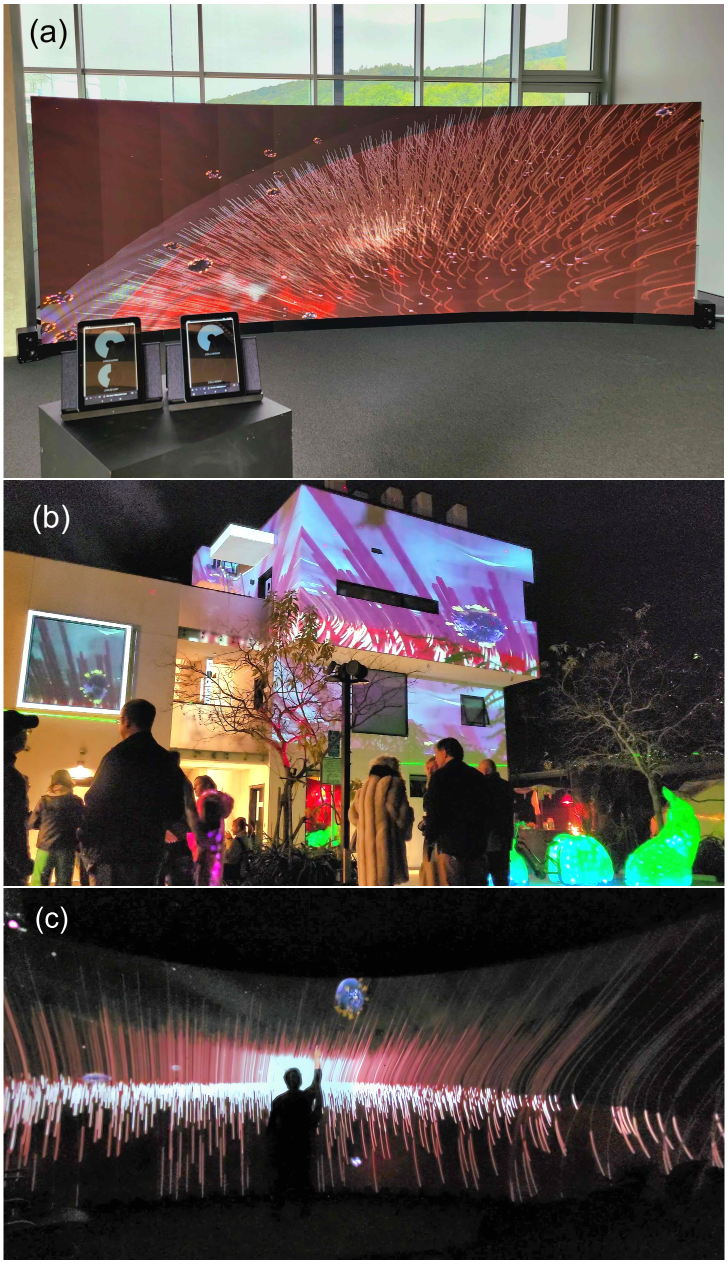


Figure 3.18: The installations are made in various environments and diverse audiences are invited for the interaction. (a) Ars Electronica Festival 2022. (b) Santa Barbara Center for Art, Science and Technology (SBCAST) 2023. (c) AlloPortal (3D) in the California NanoSystems Institute 2022-2023.

3.4.6 Artistic representation

Performance

Even if we have organized the overall narrative, the audience's active participation and the performer's control influence the progression. As if a waterfall maintains its continuous flow yet never pours down the same water, numerous multimodal narratives can be composed for each incidence involving different audiences and performers. The structure transforms within the programmed mechanism through the ensemble of participants and performers, sometimes quickly, sometimes gradually. This aspect also provides creative opportunities for individuals or groups to improvise or develop their virtuosity, allowing them to deliver special live performances with a distinct creative vision. The online audiovisual documentation and demonstration are available in [68].

Space Design

This project has been shown in different forms in immersive environments. The customized open-source C++ components, including *AlloLib* [50], allow miscellaneous audio and graphic configurations that change with the installation space and environment.

In the Ars Electronica Festival 2022 [69], we used a 6-meter-wide and 2-meter-high curved LED wall, accompanied by a 4.1 channel sound system (Figure 3.18(a)). This immersive multimodal experience offered interactive features allowing visitors to engage with complex biophysical data through multiple senses with tactile sensations through a 16-inch sub-woofer.

Furthermore, this project was showcased as a building projection on the complex surfaces of the Santa Barbara Center for Art, Science, and Technology (SBCAST) in 2023, engaging a broader audience in the city and providing an immersive experience (Figure 3.18(b)).

Currently, the program is archived and accessible at the *AlloPortal* in the California NanoSystems Institute at the University of California Santa Barbara, presenting 3D visuals and spatial audio (Figure 8(c)) [70]. This space welcomes diverse audiences regularly.

3.4.7 Discussion of SARS-CoV-2 Virus Project

Throughout the case study and the project's evolution, we received assorted types of feedback from multifarious audiences, including artists, engineers, scientists, physicians, and the general public. Artists expressed curiosity about the aesthetics of audiovisual representation, and engineers asked about the implementation methods and interaction design. Physicians and scientists recognized the project's value in raising concepts through interactive multimodal means that could not be observed. Their expertise deepened their attention, and they shared invaluable insights and perspectives. This collaboration brought multifaceted dialogs to the immersive audiovisual simulation, enriching the project.

We acquired practical feedback during the development phase on the scale of particles, dynamics, and mechanisms of the coronavirus. By comprising advice and insights, we improved the accuracy and authenticity of the simulation, providing a precise representation of the scientific notions.

This aspect encouraged precise conversations among scientists and accentuated distinct perspectives on the imaging and interpretation of the coronavirus and its mechanisms. The multimodal representation of science functioned as a catalyst for discussion among a manifold audience, fostering a deeper understanding and gratitude for scientific discovery.

Chapter 4

Evaluation of Crossmodal Correspondence

This thesis started by stating the crucial role of coherent digital multimodal instruments and introduced the case studies. Meanwhile, projects motivated by artistic goals sometimes struggle to determine how to measure the research's significance quantitatively. This chapter investigates a methodology to explain the crossmodal correspondence in multimodal instruments quantitatively. This approach is motivated by the observation described in Chapter 1. Section 1.2.2, quantizing the multimodal properties of Takete-Maluma. Through the experiment, we could find a positive correlation between the graphical features and audio features. Expanding on this concept, this research suggests a method to extract monomodal features from multimodal media and compute the correlation coefficients between the features.

4.1 Corpus of Interactive Audiovisual from the Audience's Perspective

The proposed method investigates the experience of the multimodal instrument from the audience's point of view and derives correlations between different modalities from each video in which an instrumentalist plays an instrument.

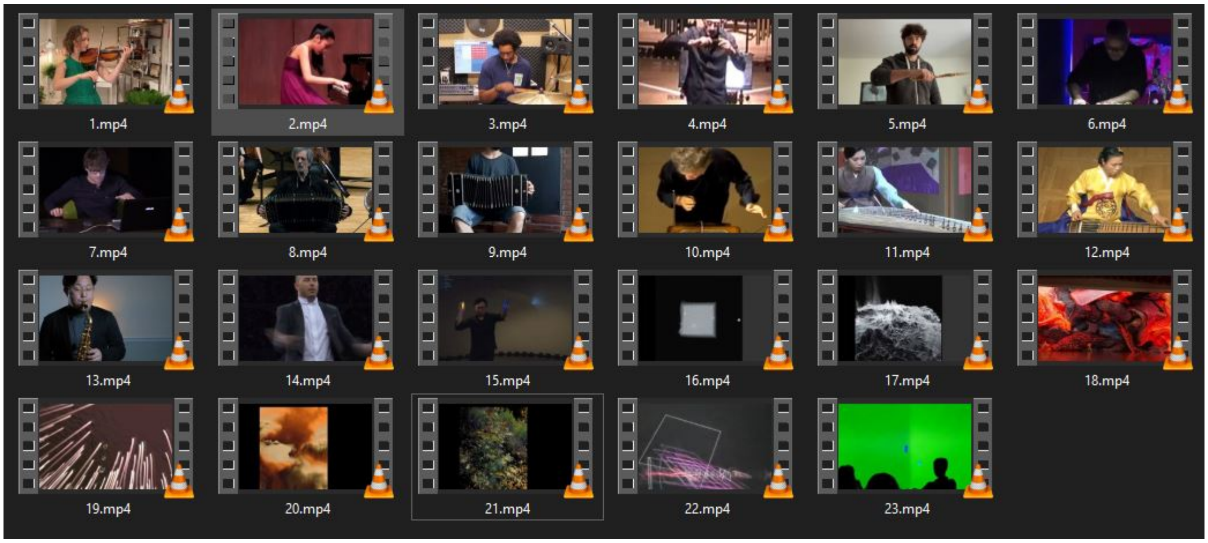


Figure 4.1: Corpus of interactive audiovisual from audience's point of view: 1. Violin [71], 2. Piano [72], 3. Drum [73], 4. Mephisto [74], 5. Airsticks [75], 6. Modular synthesis [76], 7. Conundrum [77], 8. Bandoneón(1) [78], 9. Bandoneón(2) [79], 10. Theremin [80], 11. Gayageum [81], 12. Korean lyre [82], 13. Saxophone [83], 14. Conductor [84], 15. AlloThesher [85], 16. Multidimensional pulse [86], 17. Time loop [87], 18. Living Paintings by Refik Anadol [88], 19. SARS-CoV-2 Virus Project [89], 20. Luminism by Ben Heim [90], 21. GARDEN by Ben Heim [91], 22. MAT276IA [92], 23. MUS109IA [93]

To provide controlled and vast observation, this research establishes a new database composed of videos of an instrumentalist playing an instrument solely from a similar distance and observed from a static point of view. The videos with at least 35 seconds are collected to obtain consistent observation. The database is collected from various sources, including YouTube and Instagram. However, it is challenging to discover the clips satisfying the given conditions. Figure 4.1 shows the current inclusion of the database comprised of acoustic instruments, NIME, and audiovisual without an instrumentalist.

4.2 Feature Selection

As discussed in Chapter 1. Section 1.2.4, this method hypothesizes three sensory modalities in the multimodal instrument: visual, audio, and gesture. The proposed method investigates the experience of the multimodal instrument from the audience's point of view. It derives correlations between different modalities from each video in which an instrumentalist plays an instrument. This method is expected to derive the data-centric and signal-processing-wise quantitative observation of multimodal information from the audience's perspective. The selected features are designated to represent each modality. The code to obtain the feature is shared and will be updated through GitHub [94].

4.2.1 Visual

For the visual, the histogram difference between the two frames is selected. This inter-frame feature detects changes in the weighted color histogram of two consequent images. It is widely used for video segmentation tasks and codec to represent variations of graphical information over frames [95, 96, 97]. This feature can detect changes in the graphics of digital instruments and the physical dynamics of acoustic instruments and instrumentalists.

4.2.2 Audio

Audio envelope and pitch are obtained to describe amplitude and frequency information in short. The envelope is acquired by retrieving the amplitude of the waveform, and the pitch information over time is obtained by estimating the fundamental frequency of the audio signal using the pitch estimate filter. [98].

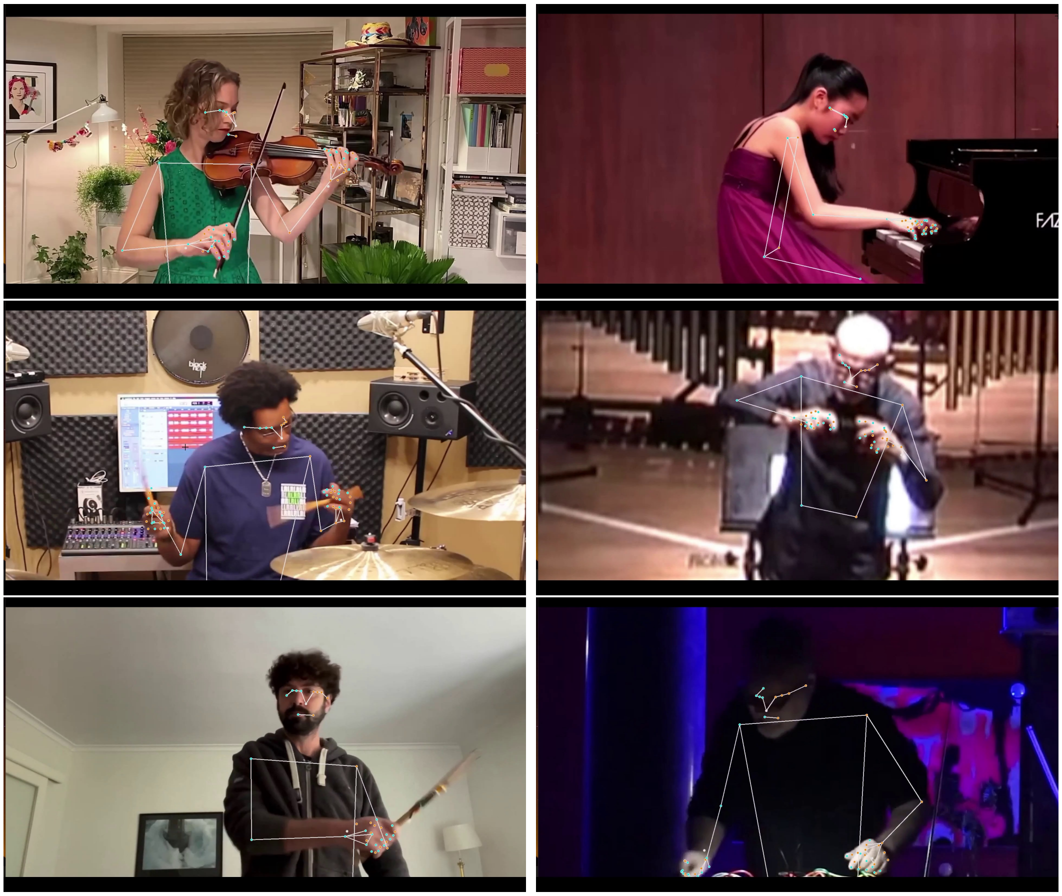


Figure 4.2: Examples of successful gestural detection of the instrumentalists from videos

4.2.3 Gesture

To retrieve human gesture information from video, I use machine learning-based gesture recognition to label major joints of the motion [99]. While this algorithm is widely used to obtain real-time gesture recognition using a camera, the gestural information from video can be labeled by feeding video input using OpenCV. Figure 4.2 shows the successful gestural detection of the instrumentalists from the video. In this case, the framerate is fixed at 30 fps but can be increased drastically with GPU acceleration. From the position of the joints, the difference between each frame is calculated and accumulated to characterize the activeness of the gesture.

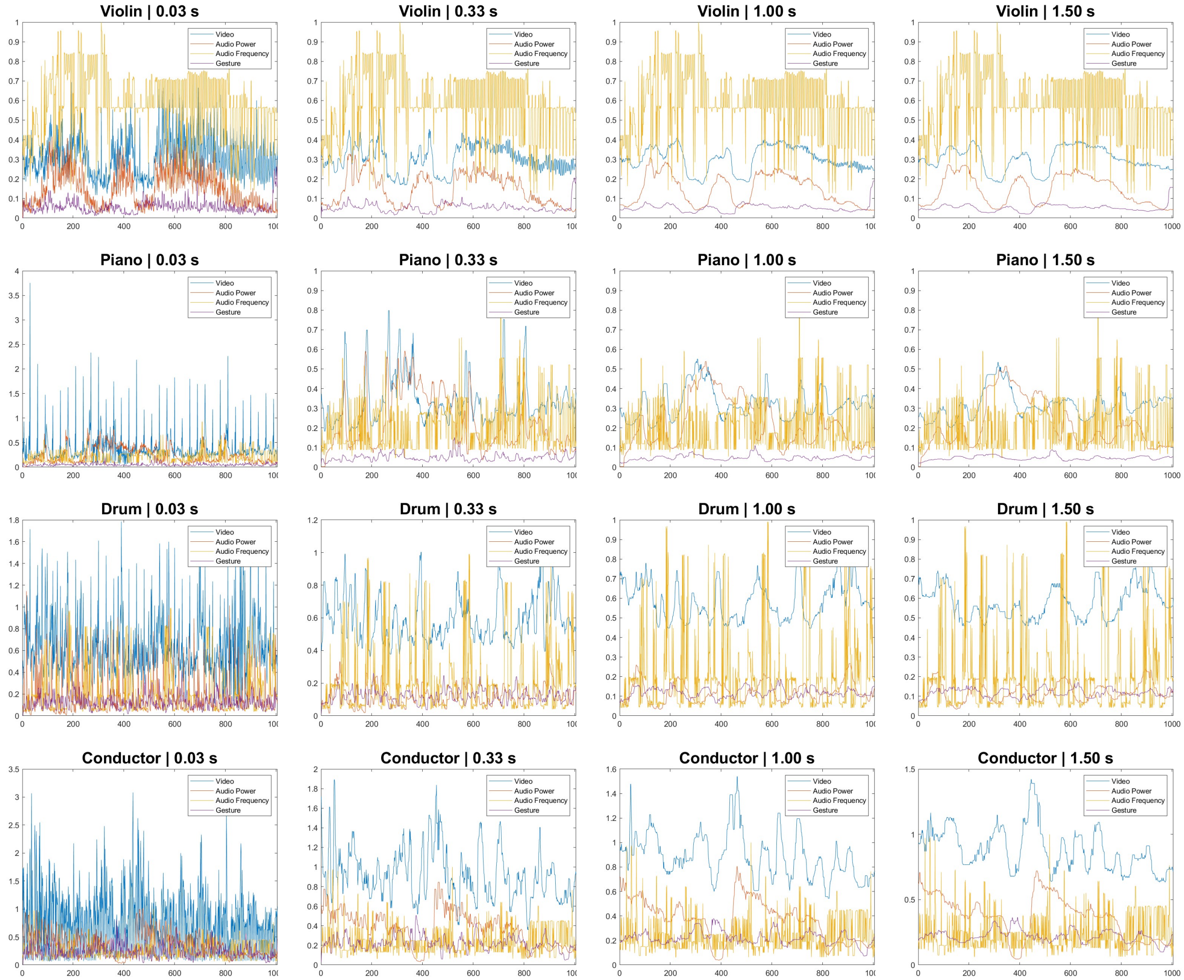


Figure 4.3: Multimodal features of four exemplary acoustic instruments with various observation times using smoothing (0.03-1.50s)

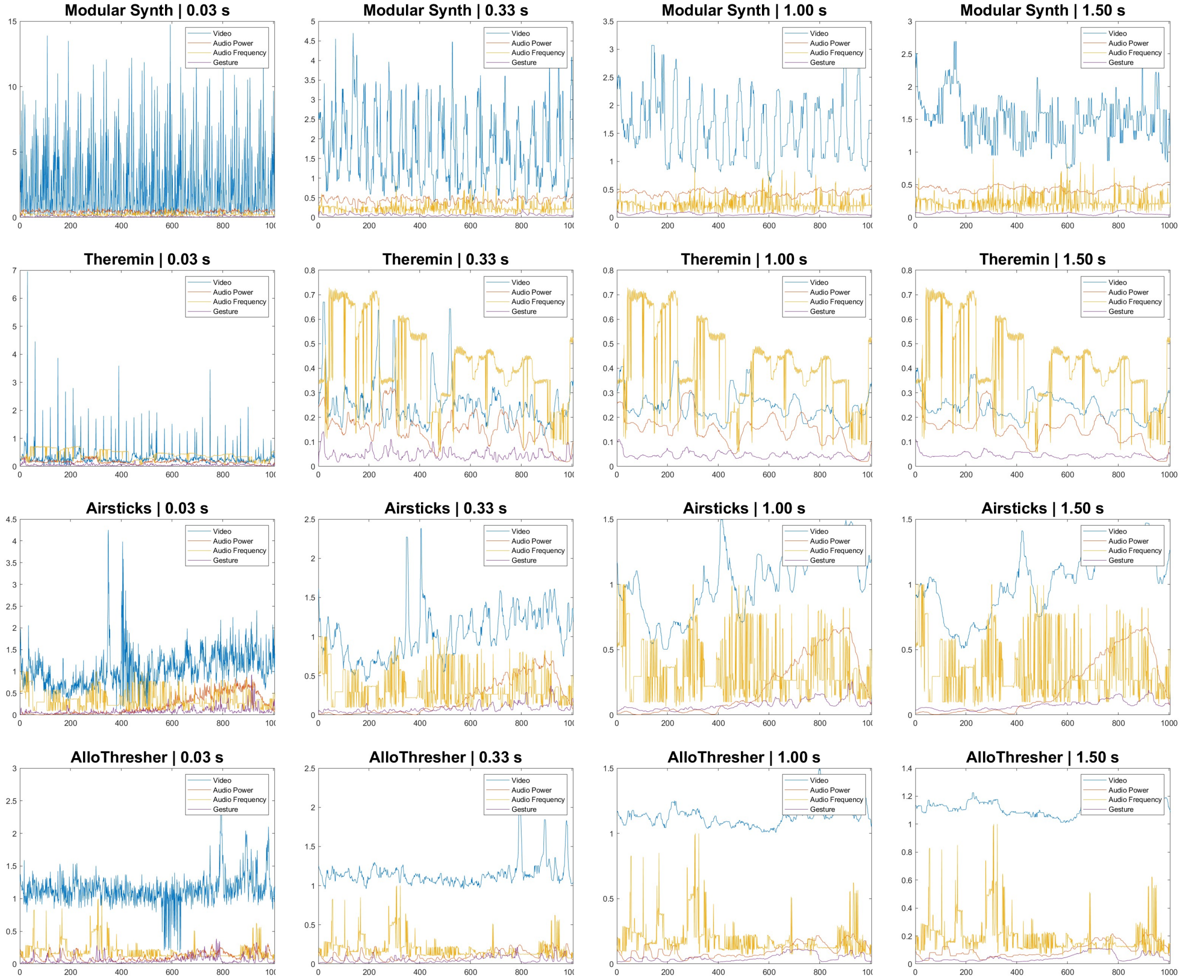


Figure 4.4: Multimodal features of four exemplary NIME instruments with various observation times using smoothing (0.03-1.50s)

4.3 Analysis of the Features

4.3.1 Audio, Visual, Gesture

The features from different domains are temporally synchronized with the same number of samples using proper sampling. Combining and observing these features at the same time gives an interesting insight into this dataset. Figure 4.3 shows the features extracted from acoustic instruments, and Figure 4.4 depicts the features from NIME instruments. Blue lines indicate the visual inter-frame feature, red lines show the audio power, yellow lines depict audio frequency, and purple lines present the gestural feature. While these features are derived from different modal data, some instruments show apparent correlations between features. The correlation can be observed more clearly when the parameters are smoothed over time. This mitigates the noisy data artifacts and increases the observation's temporal window. A diverse range of observations is also meaningful, considering some instruments are triggered or cued by another modality.

While Figures 4.3 and 4.4 give a glimpse of the data, this method derives a numerical analysis by computing the correlation coefficient between modal features. The correlation coefficient is a numerical barometer of correlation indicating a statistical connection between two variables. While there are types of correlation coefficients, the Pearson correlation coefficient evaluates linear connections, and Spearman's correlation evaluates not only linear but also monotonic relationships by calculating the Pearson correlation between rank values of the two variables [100].

From four features, including video, audio(amplitude, pitch), and gesture, we can derive $\binom{4}{2}$ combination of correlation coefficients from the features. The six combinations of correlation coefficients are defined as follows:

V-A = (Video frame —Audio amplitude)
V-F = (Video frame —Audio frequency)
V-G = (Video frame —Gestural activeness)
A-F = (Audio amplitude —Audio frequency)
A-G = (Audio amplitude —Gestural activeness)
F-G = (Audio frequency —Gestural activeness)

Each combination depicted the correlation between two different modal features. For example, if one instrument's audio amplifies when gestural features increase, the A-G

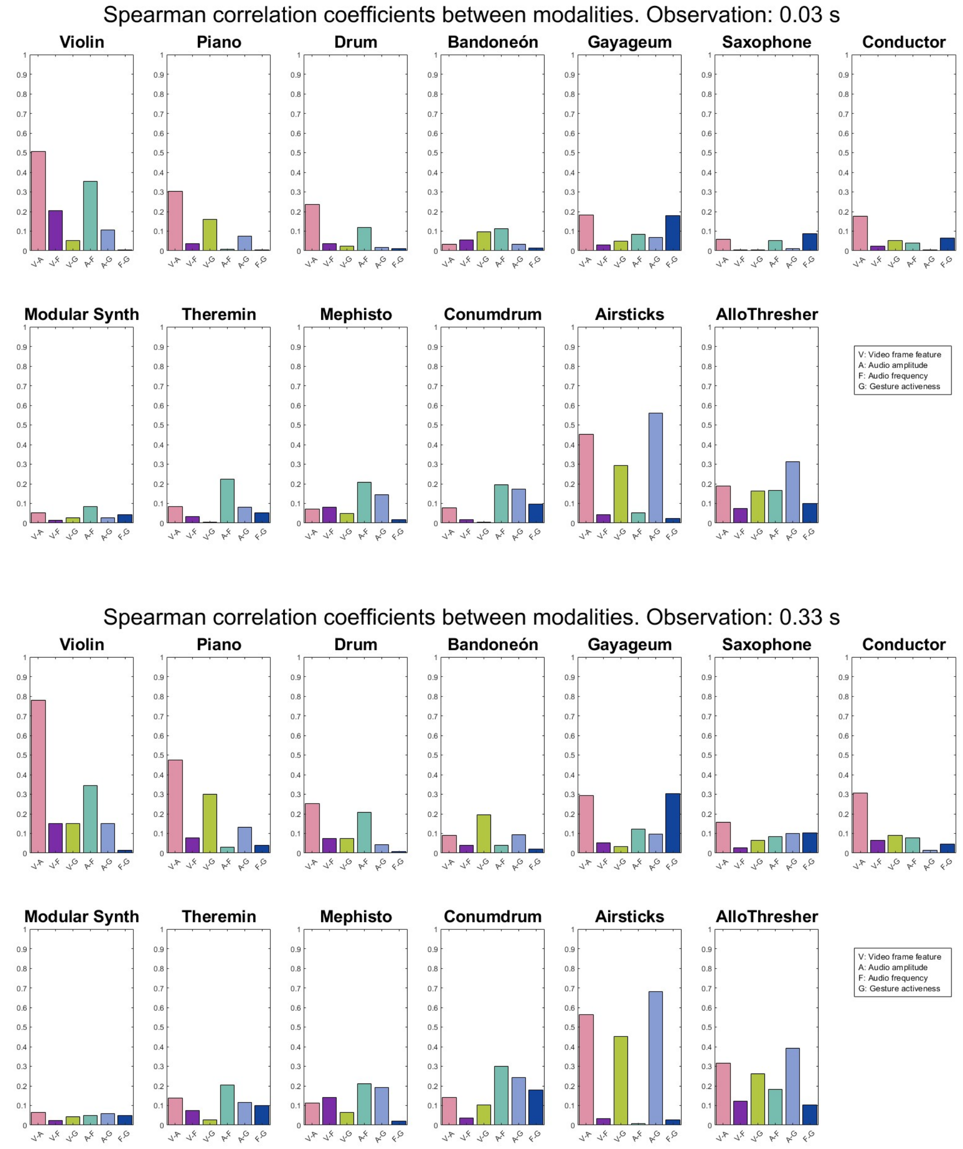


Figure 4.5: Spearman correlation coefficients between modalities with short observation time. V: video inter-frame, A: audio (amplitude), F: audio (pitch), G: Gesture

correlation coefficient will be close to 1. This can be endorsed by observing Figures 4.5 and 4.6, where Figure 4.5 shows shortly observed cases and Figure 4.6 shows more extended observations of correlation. The correlation coefficients tend to be higher when features are related over a long period. NIME instruments that actively use gestures to

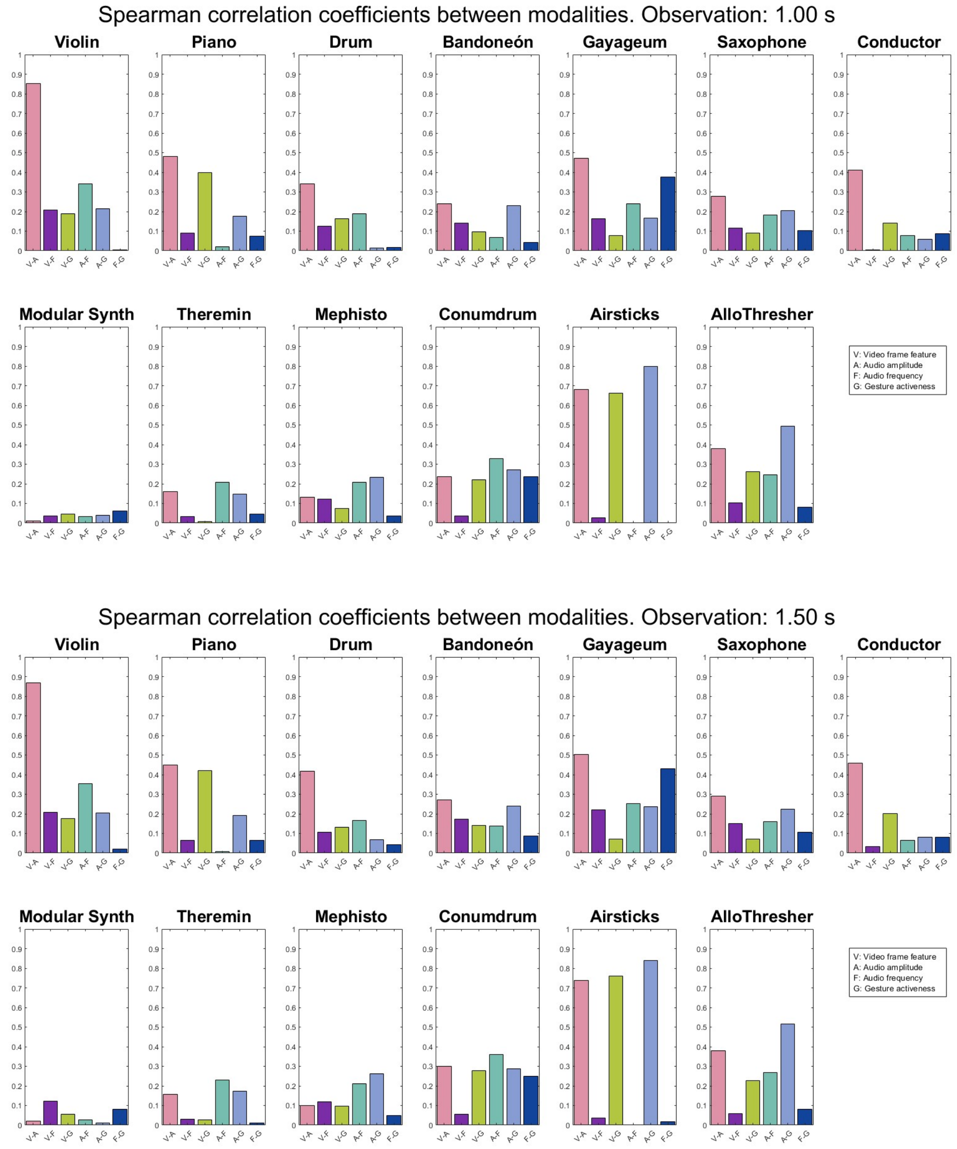


Figure 4.6: Spearman correlation coefficients between modalities with long observation time. V: video inter-frame, A: audio (amplitude), F: audio (pitch), G: Gesture

actuate audio (Airsticks, AlloThresher) show higher A-G values than other instruments.

Besides, professional instrumentalists playing violin or piano dynamically move their bodies and instruments, and correspondingly, high correlations between the video frame feature and audio amplitude (V-A) can be observed. In the case of the gayageum, a traditional Korean instrument, the instrument features extensive vibrato using direct hand fluctuation. Such a characteristic can be observed from the gayageum's high F-G correlation coefficient value. On the other hand, it is noticeable that modular synthesis shows low correlation coefficients in every possible combination. As discussed in Chapter 1. Section 1.3, coherent multimodal experience requires consistent, systematic, and logical cross-modal stimuli through multisensory integration. If the multimodal characteristics are inconsistent, the audience cannot logically connect the modalities and fail to understand the system.

Likewise, this combination of numerical observation grants a notional understanding of the instrument from the audience's point of view.

4.3.2 Audiovisual

This method can also evaluate the crossmodal correspondence of audiovisuals without a gestural interface. In this case, gestural features are neglected and three combinations of Spearman correlation coefficients can be derived:

V-A = (Video frame —Audio amplitude)
V-F = (Video frame —Audio frequency)
A-F = (Audio amplitude —Audio frequency)

The Figures 4.7 and 4.8. show the corresponding Spearman correlation coefficients of audiovisuals. While fewer combinations of features are available, the correlations between visual and audio amplitude are dominantly high in most cases. The projects with high V-A tend to have rapid audio-reactive graphics [86, 87, 90]. However, the observation states that each audiovisual shows different crossmodal characteristics. For example, a project with generative graphics with ambient audio shows low correlation coefficients [91].

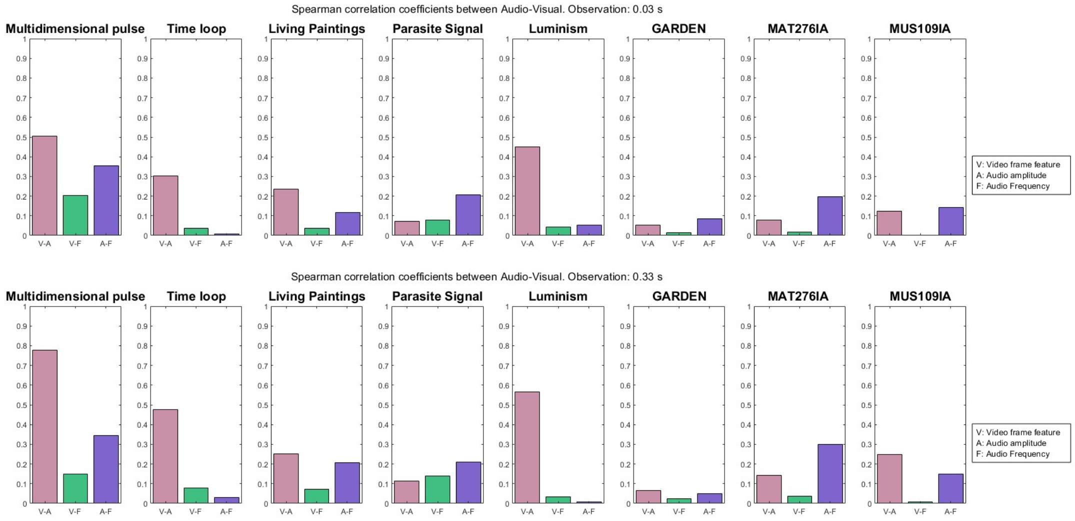


Figure 4.7: Spearman correlation coefficients between audio-visual modalities with short observation time. V: video inter-frame, A: audio (amplitude), F: audio (pitch)

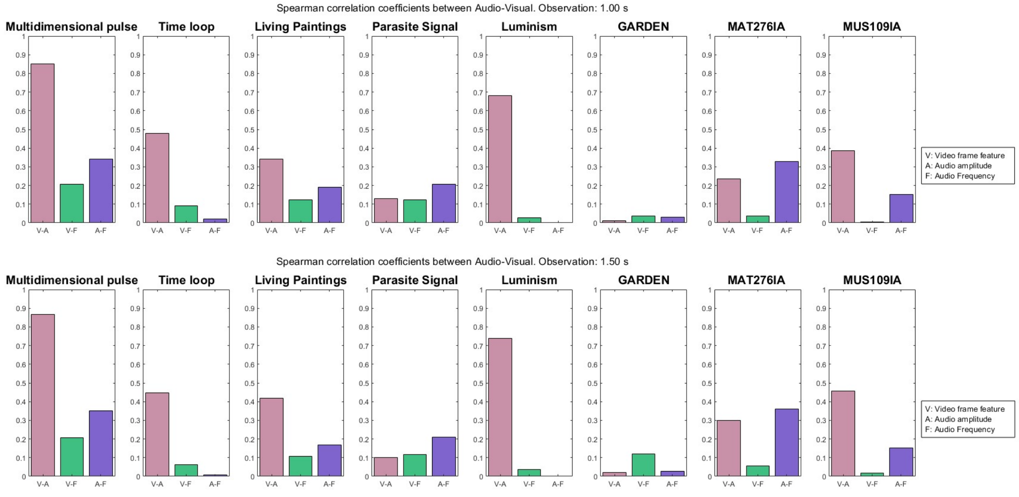


Figure 4.8: Spearman correlation coefficients between audio-visual modalities with long observation time. V: video inter-frame, A: audio (amplitude), F: audio (pitch)

4.3.3 Discussion on the Evaluation Method

The presented evaluation method offers a numerical way to analyze the characteristics of the multimodal instrument from the audience's point of view. Meanwhile, additional statistical measures such as mean and variance of the correlation coefficients are derived in this research since the purpose of the evaluation is not to line up the diverse instruments but observe the attributes of the tool from multimodal aspects.

For instance, in the case of the SARS-CoV-2 virus project, Parasite Signal, the correlation coefficients do not show a strong pattern compared with audio-reactive visualizations. This brings the question of whether the higher correlation coefficient always states a better instrument design. Indeed, the apparent connection between modalities builds strong multisensory integration. To rephrase the audiovisual design of the project, while individual agents have strong audiovisual coherence, compounds of agents with a probability model compose the audiovisual landscape simulating the complex biosystem. Humans can distinguish complex sound sources and extract higher dimensional information from stimuli. Likewise, the SARS-CoV-2 virus project involves biophysics as the rule that connects the different modalities, and it is designed to guide the audience to comprehend complex biophysics through presented audiovisuals interactively.

From this observation, we can derive the advantage and limitations of the proposed evaluation method. The proposed analysis provides a way to evaluate monotonic relationships between multiple modalities. The pattern in the set of correlations represents the crossmodal characteristics of the instrument. From the audience's perspective, the low correlation between other modalities suggests crossmodal correspondence is uncertain, and the performance information is hard to be perceived thoroughly.

On the other hand, the instrument with complex multimodal relationships has moderately distributed correlation coefficients rather than a single high correlation coefficient.

Compared to a singular bimodal connection, since our brain synthesizes information from cross-modal stimuli through multisensory integration, expanding modality can stimulate new enlightenment from coherent multimodal experience. For instance, we could drastically expand the possible combination of correlations between modalities by considering the gestural feature as an input. While the physicality and profound expressiveness of our gestures tend to have been diluted due to the automation of programmable audio-visuals, by reconnecting our body and expressiveness with multimodal instruments, we can expect to leverage novel creative opportunities and audience experience at the same time.

Chapter 5

Future Research Directions

The research on coherent multimodal instruments will be continued by developing, collaborating, presenting, and sharing new systems with new physics and data. Continuing a career as a researcher and educator, I will continue sharing, teaching, and learning new perspectives, expanding the multimodal instrument design criteria.

As a practical commitment, my research will build up the complexity of systems, even climate studies. Sensorium: The Voice of the World Ocean is one of the core projects I will focus on. Climate change is a complex system dealing with multiple interweaved dimensions that should be studied coherently to affect a positive change. Sensorium is a work of art and science motivated by Newton Harrison, categorizing diverse ocean data and allowing researchers to analyze extensive amounts of information interactively, accelerating the discovery [101, 102].

Similarly, the demand for the artistic representation of scientific data is emerging, and collaborative opportunities with artists, scientists, and engineers are growing. In these projects, I will play a core role in developing the system's comprehensive multimodal representation of scientific information into artistic languages.

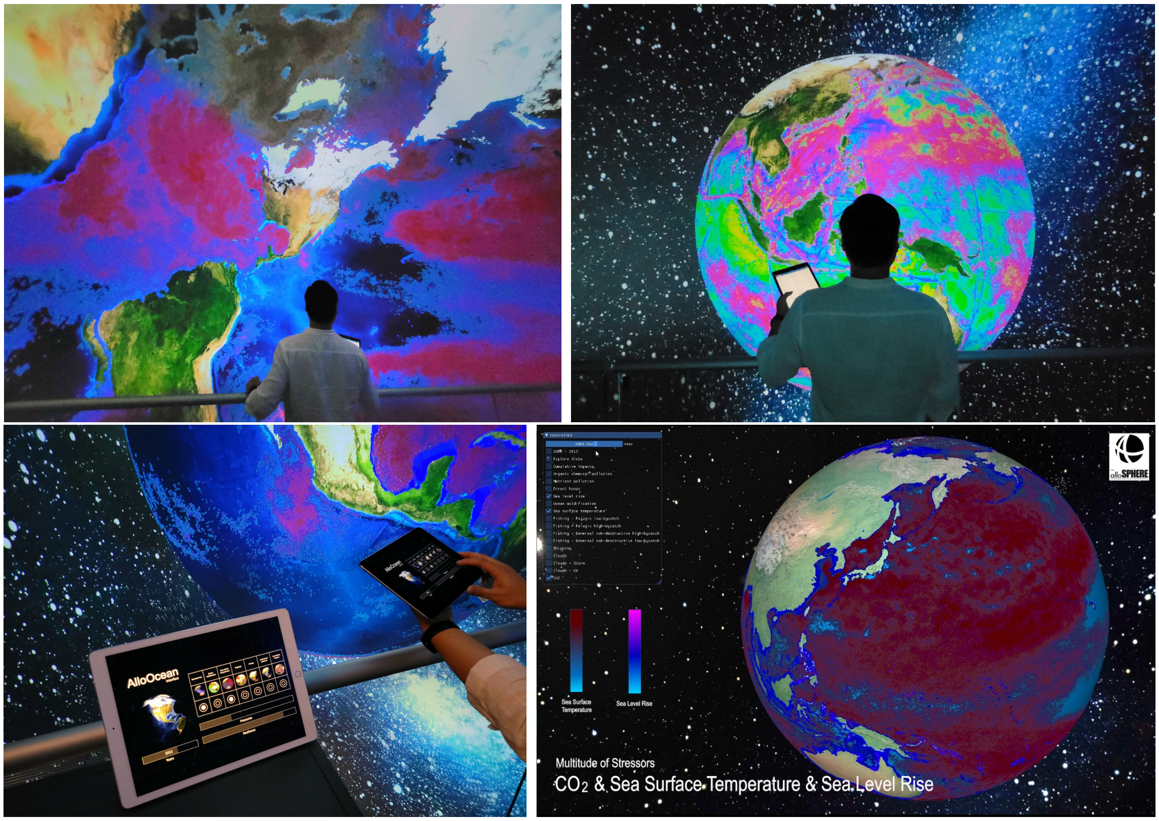


Figure 5.1: Sensorium in the AlloSphere

Chapter 6

Conclusion

This dissertation explored how to design coherent digital multimodal instruments and argues the opportunities that can provide. The introduction and literature review analyzed the properties and components of coherent instrumental experience. Empirical four multimodal instrument design cases elaborated different attributes and practices of the coherent experience. Especially the role of gestures and physicality have been highlighted to relate complex multimodal correlations.

By computing the correlation coefficients between multimodal features, a quantitative method to evaluate the crossmodal correspondence was proposed. To comprehensively understand diverse instruments, the constrained database from acoustic instruments to new digital instruments has been gathered, and the multimodal feature extraction code has been shared with open access. This research states that such measurements are not proposed to assess which instrument is better but to understand the multimodal characteristics of the instrument from the audience's perspective and expand to further instrument design criteria. By designing coherent multimodal instruments with complex rules, such as scientific theory or data, the multimodal experience can allow new observations, aesthetics, and pedagogical methods for scientists, artists, and general audiences in our society. From this aspect, we can find the importance of shared group experience. For us to understand complex systems, sometimes one person will not be able to know the whole thing. To make a breaking-through discovery, a group of people with various expertise work together to understand the complex system. Through this process, we become each representation of modality, which can be greater than the sum of the parts. Therefore, this dissertation and further research are trying to bring science into an artistic domain that could not be achieved only through papers and charts. To quote Albert Einstein's words,

> ”We do art when we communicate through forms whose connections are not accessible to the conscious mind yet we intuitively recognize them as something meaningful.”

Likewise, this series of research states that digital art created from coherent multimodal instruments can be a language that brings a group of people together with various expertise to create, share, and understand a novel experience.

Bibliography

[1] J. Murray, *Composing multimodality*, *Multimodal composition: A critical sourcebook* (2013) 325–350.

[2] A. Bódog, *Multimodality and spontaneity in human-computer interactions analogies of ontogeny*, *IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (2012).

[3] C. Spence, *Crossmodal correspondences: A tutorial review*, *Attention Perception Psychophysics* (2011).

[4] W. Köhler, *Gestalt psychology.* Liveright, New York, NY, 1929.

[5] A. Bremner, S. Caparos, J. Davidoff, J. de Fockert, K. Linnell, and C. Spence, *"bouba" and "kiki" in namibia? a remote culture make similar shape-sound matches, but different shape-taste matches to westerners*, *Cognition* **126** (2013), no. 2.

[6] B. E. Stein and T. R. Stanford, *Multisensory integration: current issues from the perspective of the single neuron*, *Nature Reviews Neuroscience* **9** (2008).

[7] J. Copeland and J. Long, *Alan turing: How his universal machine became a musical instrument*, *IEEE Spectrum* (2017).

[8] *RCA Mark I and Mark II Synthesizers*, 2015. [Online] Available: https://ethw.org/RCA_Mark_I_and_Mark_II_Synthesizers.

[9] B. Rossmy and A. Wiethoff, *The modular backward evolution – why to use outdated technologies*, *Proceedings of the International Conference on New Interfaces for Musical Expression* (2019).

[10] J. Rothstein, *MIDI:A Comprehensive Introduction.* 1992.

[11] A. R. Jensenius and M. J. Lyons, *A NIME Reader: Fifteen Years of New Interfaces for Musical Expression.* Springer, 1992.

[12] M. Wright, A. Freed, and A. Momeni, *2003: Opensound control: State of the art 2003*, *A NIME Reader. Current Research in Systematic Musicology* (2017).

[13] A. Roberts, C. Hawthorne, and I. Simon, *Magenta.js: A javascript api for augmenting creativity with deep learning*, *Joint Workshop on Machine Learning for Music (ICML)* (2018).

[14] E. R. Miranda, *Handbook of Artificial Intelligence for Music*. Springer, 2021.

[15] M. Noll, *The beginnings of computer art in the united states: A memoir*, *Leonardo* **27** (1994), no. 1.

[16] *György Ligeti: 'I always imagine music visually, in many different colours'*, 1974. [Online] Available: https://www.theguardian.com/music/2016/feb/02/from-the-classical-archive-gyorgy-ligeti-interveiw-1974.

[17] R. Wehinger, *Artikulation : Electronic Music / Aural Score by Rainer Wehinger*. Schott, 1970.

[18] M. Noll, *The beginnings of computer art in the united states: A memoir*, *Leonardo* **27** (1994), no. 1.

[19] W. Buxton, W. Reeves, G. Fedorkow, K. C. Smith, and R. Baecker, *A microprocessor-based conducting system*, *Computer Music Journal* **4** (1980), no. 1.

[20] T. Ilmonen and T. Takala, *Conductor following with artificial neural networks*, *International Computer Music Conference* (1999).

[21] P. Kolesnik and M. Wanderley, *Recognition, analysis and performance with expressive conducting gestures*, *International Computer Music Conference* (2004).

[22] E. Lee, I. Grull, H. Kiel, and J. Borchers, *conga: A framework for adaptive conducting gesture analysis*, *Computer Music Journal* (2006).

[23] A. Hofer, A. Hadjakos, and M. Mühlhäuser, *Gyroscope-based conducting gesture recognition*, *International Computer Music Conference* (2009).

[24] G. Hinton and R. Salakhutdinov, *Reducing the dimensionality of data with neural networks*, *Science* **313** (2006), no. 5786.

[25] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, *Greedy layer-wise training of deep networks*, *Neural Information and Processing Systems* (2006).

[26] M. Lee, *Deep neural network based music source conducting system*, *International Computer Music Conference (ICMC)* (2018).

[27] G. Essl, G. Wang, and M. Rohs, *Developments and challenges turning mobile phones into generic music performance platforms*, *International Mobile Music Workshop* (2008).

[28] S. W. Lee and J. Freeman, *echobo: A mobile music instrument designed for audience to play*, *NIME'13* (2013).

[29] E. Egozy and E. Y. Lee, *12: Mobile phone-based audience participation in a chamber music performance*, *NIME'18* (2018).

[30] K. Nishida, A. Yuguchi, K. Jo, P. Modler, and M. Noisternig, *Border: A live performance based on web ar and a gesture-controlled virtual instrument*, *NIME'19* (2019).

[31] J. Atherton and G. Wang, *Curating perspectives: Incorporating virtual reality into laptop orchestra performance*, *NIME'20* (2020).

[32] A. Çamcı, M. Vilaplana, and R. Wang, *Exploring the affordances of vr for musical interaction design with vimes*, *NIME'20* (2020).

[33] A. Xambo and G. Roma, *Performing audiences: Composition strategies for network music using mobile phones*, *NIME'20* (2020).

[34] T. Brizolara, S. Gibet, and C. Larboulette, *Elemental: a gesturally controlled system to perform meteorological sounds*, *NIME'20* (2020).

[35] J. Leonard and A. Giomi, *Towards an interactive model-based sonification of hand gesture for dance performance*, *NIME'20* (2020).

[36] A. Tsiros, *The parallels between the study of crossmodal correspondence and the design of cross-sensory mappings*, *In Proceedings of the Conference on Electronic Visualisation and the Arts* (2017).

[37] E. Large, *Resonating to musical rhythm: theory and experiment*, *The psychology of time* (2008).

[38] E. Large, *Microsound.* The MIT Press, 2008.

[39] S. Jordà, *On stage: the reactable and other musical tangibles go real*, *International Journal of Arts and Technology* (2008).

[40] J. Françoise, N. Schnell, and F. Bevilacqua, *A multimodal probabilistic model for gesture based control of sound synthesis*, *ACM Multimedia* (2016).

[41] C. Traube, P. Depalle, and M. Wanderley, *Indirect acquisition of instrumental gesture based on signal, physical and perceptual information*, *NIME'03* (2003).

[42] C. Roberts, G. Wakefield, and M. Wright, *2013: The web browser as synthesizer and interface*, *A NIME Reader* **3** (2017) 433–450.

[43] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, *arXiv:1409.1556* (2014).

[44] C. Tong, J. Li, and F. Zhu, *A convolutional neural network based method for event classification in event-driven multi-sensor network*, *Computers Electrical Engineering* (2017).

[45] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, M. R. C. Jansen, Aren, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, *Cnn architectures for large-scale audio classification*, *ICASSP* (2017).

[46] N. Yalta, S. Watanabe, T. Hori, K. Nakadai, and T. Ogata, *Cnn-based multichannel end-to-end speech recognition for everyday home environments*, *EUSIPCO* (2019).

[47] C. Tong, J. Li, and F. Zhu, *Mgra: Motion gesture recognition via accelerometer*, *Sensors* (2016).

[48] J. Kuchera-Morin, M. Wright, G. Wakefield, C. Roberts, D. Adderton, B. Sajadi, T. Höllerer, and A. Majumder, *Immersive full-surround multi-user system design*, *Computers and Graphics* **40** (2014) 10–21.

[49] J. Kuchera-Morin, L. Putnam, L. Peliti, D. Adderton, A. Cabrera, K. Kim, G. A. Rincon, J. Tilbian, H. Wolfe, T. Wood, and K. Youn, *Probably/possibly?: An immersive interactive visual/sonic quantum composition and synthesizer*, *MM '17: Proceedings of the 25th ACM international conference on Multimedia* (2017).

[50] *AlloLib: A cross-platform suite of C++ components for building interactive multimedia tools and applications*, 2023. [Online] Available: https://github.com/AlloSphere-Research-Group/allolib.

[51] M. Lee, *Entangled: A multi-modal, multi-user interactive instrument in virtual 3d space using the smartphone for gesture control*, *NIME'21* (2021).

[52] *Entangled: Demonstraion*, 2021. [Online] Available: https://www.myunginlee.com/entangled.

[53] C. Roads, *Composing electronic music: a new aesthetic.* Oxford University Press, 2015.

[54] *EmissionControl2*, 2023. [Online] Available: https://github.com/EmissionControl2/EmissionControl2.

[55] *AlloThresher*, 2023. [Online] Available: https://www.myunginlee.com/allothresher.

[56] D. Hoffmann, S. Mereiter, Y. J. Oh, V. Monteil, E. Elder, R. Zhu, D. Canena, L. Hain, E. Laurent, C. Grünwald-Gruber, M. Klausberger, G. Jonsson, M. J. Kellner, M. Novatchkova, M. Ticevic, A. Chabloz, G. Wirnsberger,

A. Hagelkruys, F. Altmann, L. Mach, J. Stadlmann, C. Oostenbrink, A. Mirazimi, P. Hinterdorfer, and J. M. Penninger, *Identification of lectin receptors for conserved SARS-CoV-2 glycosylation sites*, *The EMBO Journal* **40** (2021), no. 19 e108375.

[57] J. F.-W. Chan, Y. J. Oh, S. Yuan, H. Chu, M.-L. Yeung, D. Canena, C. C.-S. Chan, V. K.-M. Poon, C. C.-Y. Chan, A. J. Zhang, J.-P. Cai, Z.-W. Ye, L. Wen, T. T.-T. Yuen, K. K.-H. Chik, H. Shuai, Y. Wang, Y. Hou, C. Luo, W.-M. Chan, Z. Qin, K.-Y. Sit, W.-K. Au, M. Legendre, R. Zhu, L. Hain, H. Seferovic, R. Tampé, K. K.-W. To, K.-H. Chan, D. G. Thomas, M. Klausberger, C. Xu, J. J. Moon, J. Stadlmann, J. M. Penninger, C. Oostenbrink, P. Hinterdorfer, K.-Y. Yuen, and D. M. Markovitz, *A molecularly engineered, broad-spectrum anti-coronavirus lectin inhibits SARS-CoV-2 and MERS-CoV infection in vivo*, *Cell Reports Medicine* **3** (2022), no. 10 100774.

[58] R. Zhu, D. Canena, M. Sikora, M. Klausberger, H. Seferovic, A. R. Mehdipour, L. Hain, E. Laurent, V. Monteil, G. Wirnsberger, R. Wieneke, R. Tampé, N. F. Kienzl, L. Mach, A. Mirazimi, Y. J. Oh, J. M. Penninger, G. Hummer, and P. Hinterdorfer, *Force-tuned avidity of spike variant-ACE2 interactions viewed on the single-molecule level*, *Nature Communications* **13** (2022) 7926.

[59] D. Shreiner, G. Sellers, J. Kessenich, and B. Licea-Kane, *OpenGL Programming Guide: The Official Guide to Learning OpenGL, Version 4.3.* Addison-Wesley, Boston, MA, USA, version 4.3 ed., 2013.

[60] G. P. Scavone., "Rtaudio version 5.2.0." [Online] Available: http://www.music.mcgill.ca/ gary/rtaudio/.

[61] C. Johnson, J. Exell, Y. Lin, J. Aguilar, and K. D. Welsher, *Capturing the start point of the virus–cell interaction with high-speed 3D single-virus tracking*, *Nat Methods* **19** (2022) 1642–1652.

[62] X. Bustamante-Marin and L. E. Ostrowski, *Cilia and mucociliary clearance*, *Cold Spring Harb Perspect Biol.* **9** (2017), no. 4 1147–1150.

[63] J. Shang, Y. Wan, C. Luo, G. Ye, Q. Geng, A. Auerbach, and F. Li, *Cell entry mechanisms of SARS-CoV-2*, *Proceedings of the National Academy of Sciences (PNAS)* **117** (2020), no. 21 11727–11734.

[64] C. Jackson, M. Farzan, B. Chen, and H. Choe, *Mechanisms of SARS-CoV-2 entry into cells*, *Nature Reviews Molecular Cell Biology* **23** (2022) 3–20.

[65] X. Xiong, K. Qu, K. A. Ciazynska, M. Hosmillo, A. P. Carter, S. Ebrahimi, Z. Ke, S. H. W. Scheres, L. Bergamaschi, G. L. Grice, Y. Zhang, J. A. Nathan, S. Baker, L. C. James, H. E. Baxendale, I. Goodfellow, R. Doffinger, and J. A. G.

Briggs, *A thermostable, closed SARS-CoV-2 spike protein trimer*, *Nature Structural Molecular Biology* **27** (2020), no. 10 934–941.

[66] E. P. Barros, L. Casalino, Z. Gaieb, A. C. Dommer, Y. Wang, L. Fallon, L. Raguette, K. Belfon, C. Simmerling, and R. E. Amaro1, *The flexibility of ACE2 in the context of SARS-CoV-2 infection*, *Biophysical Journal* **120** (2021), no. 6 1072–1084.

[67] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, *Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising*, *IEEE Transactions on Image Processing* **26** (2017), no. 7 3142–3155.

[68] M. Lee, *Coexistence with the SARS-CoV-2 virus*, 2023. [Online] Available: https://www.myunginlee.com/covid.

[69] Y. Oh, S. H. Ahn, and M. Lee, *Coexistence with the SARS-CoV-2 virus*, 2022. [Online] Available: https://ars.electronica.art/planetb/en/coexistence-sars-cov-2/.

[70] J. Kuchera-Morin, A. Cabrera, K. H. Kim, G. Rincon, and T. Wood, *MYRIOI*, *ACM SIGGRAPH 2020 Art Gallery* (2020) 468–469.

[71] *Hilary Hahn - J. S. Bach: Partita No. 3 for Solo Violin, BWV 1006* , 2022. [Online] Available: https://www.youtube.com/watch?v=Pr_gK9fzwSo.

[72] *Tiffany Poon plays Beethoven Moonlight Sonata*, 2010. [Online] Available: https://www.youtube.com/watch?v=Ey4n8vqlX_o.

[73] *5 Slick Easy Fills For 'Just OK' Drummers - Practice Aid Video!*, 2019. [Online] Available: https://www.youtube.com/watch?v=TDj6GFJrg08.

[74] *Mephisto for smartphone solo by Xavier Garcia (Performed by Jean Geoffroy)*, 2016. [Online] Available: https://www.youtube.com/watch?v=IkS5cjcCXpI.

[75] S. Trolland, A. Ilsar, C. Frame, J. McCormack, and E. Wilson, *Airsticks 2.0 instrument design for expressive gestural interaction*, *NIME'22* (2022).

[76] *Recursion 1, moers festival. richard scott, anita damas*, 2020. [Online] Available: https://www.youtube.com/watch?v=-_TLHN6xAc4.

[77] *Conundrum - new stage of alexandrinsky theater(andrey bundin)*, 2020. [Online] Available: https://www.youtube.com/watch?v=aVgc1wgA7Bw.

[78] *BACH / Prelude in C Major / bandoneón PASSARELLA*, 2016. [Online] Available: https://www.youtube.com/watch?v=rf1oILpoqp4.

[79] *Desde el alma - solo de bandoneon*, 2017. [Online] Available: https://www.youtube.com/watch?v=l2VFEt7l68Y.

[80] *THEREMIN - Over The Rainbow*, 2009. [Online] Available: https://youtu.be/K6KbEnGnymk.

[81] *25 String Gayageum Solo*, 2019. [Online] Available: https://youtu.be/zHIjGZIPykM.

[82] *Korean lyre (geomungo) Solo*, 2015. [Online] Available: https://www.youtube.com/watch?v=u-FbVlWYnK0.

[83] *Classical Saxophone Solo Performance- Astor Piazzolla Tango Etude No.3 by Wonki Lee*, 2021. [Online] Available: https://youtu.be/EwpvsUrgTxg.

[84] *Waldbühne 2016 – Czech Night (Conductor Camera - Clip 1)*, 2016. [Online] Available: https://youtu.be/G90Erh1hMuI.

[85] *AlloThresher*, 2023. [Online] Available: https://www.myunginlee.com/allothresher.

[86] *Multidimensional pulse*, 2023. [Online] Available: https://www.instagram.com/p/CpYAFLtNmOo/.

[87] *Time loop)*, 2023. [Online] Available: https://www.instagram.com/p/CpSw6f7tnp7/.

[88] *Refik Anadol "Living Paintings" 4k*, 2023. [Online] Available: https://www.instagram.com/p/CpSw6f7tnp7/.

[89] *Coexistence with the SARS-CoV-2 virus*, 2023. [Online] Available: https://www.myunginlee.com/covid.

[90] *Luminism - Ben Heim*, 2022. [Online] Available: https://www.instagram.com/p/CcjPxDApJOT/?img_index=1.

[91] *GARDEN - Ben Heim*, 2023. [Online] Available: https://www.instagram.com/p/CvsY-NXJTKq/?img_index=1.

[92] *MAT276IA*, 2022. [Online] Available: https://www.youtube.com/watch?v=fWPrWyy943Y.

[93] *Alien march - brandon nadell*, 2022. [Online] Available: https://www.youtube.com/watch?v=HGcq9THVkTI.

[94] *Multimodal Correlation Research*, 2023. [Online] Available: https://github.com/MyunginLee/xcorrelation_research.

[95] A. Bovik, *The Essential Guide to Video Processing*. Academic Press, 2009.

[96] N. Ejaz, T. B. Tariq, and S. W. Baik, *Adaptive key frame extraction for video summarization using an aggregation mechanism*, *Journal of Visual Communication and Image Representation* **23** (2012), no. 7.

[97] P.-N. Zhao, R.-K. Wang, and Z.-M. Lu, *Inter-frame passive-blind forgery detection for video shot based on similarity analysis*, *Multimedia Tools and Applications* **77** (2018).

[98] *pitch*, 2023. [Online] Available: https://www.mathworks.com/help/audio/ref/pitch.html.

[99] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, *Mediapipe: A framework for building perception pipelines*, *arXiv:1906.08172* (2019).

[100] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing. Springer Topics in Signal Processing.* Springer, 2009.

[101] *Sensorium: The voice of the world ocean*, 2023. [Online] Available: https://allosphere.ucsb.edu/research/sensorium/.

[102] N. Sazevich, *Sensorium, a new project of uc santa cruz's center for the study of the force majeure, to be an unprecedented experience of the world ocean*, *UCSC News Center* (2023).