Portfolio of Original Compositions

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Humanities

2022

Hongshuo Fan
School of Arts, Languages and Cultures

Contents

Word Count: 17800

Portfolio of Musical Works

- **Handwriting · WuXing** (2018 - 2019) ca. 13’20
  For the hand-controlled gesture recognition system

- **Intangible Field** (2018 - 2019) ca. 07’00
  For three acoustic instrument and two actors

- **Audio Game - Music Force** (2018 - 2019) ca. 10’00
  For three acoustic instrument and custom design game system

- **Sound | Figuration** (2019 - 2020) ca. 12’20
  For piano and live multimedia

- **Without strings** (2019 - 2020) 24-hour cycle
  Real-time environmental data audio-visual installation

- **Metamorphosis** (2020 - 2021) ca. 21’20
  For one human performer and two artificial intelligence performers

List of USB content

- **Performance Video Recording**
  All files are in stereo, 1080P MP4 format.
  - *Handwriting · WuXing*
  - *Intangible Field*
  - *Audio Game - Music Force*
  - *Sound | Figuration*
  - *Without strings* (Installation concept demo)
  - *Without strings* (Time–Lapse recording)
  - *Metamorphosis*
  - *Conversation in the cloud*
- **Source code**
  Each project folder has included all the required dependencies and setup guidance.
  - *Handwriting · WuXing*
    - Performance notes, Technical specifications and Spatial diagram
    - Audio system (Max/MSP 8 Project)
    - Visual system (Max/MSP 6 Project)
    - Hand-controlled gesture recognition system（Max/MSP 8 Patches and Wekinator 2.1.0.4 project files）
  - *Intangible Field*
    - Performance notes, Technical specification and Spatial diagram
    - Audio system (Max/MSP 8 Project)
    - Visual system (Max/MSP 8 Project)
    - The body key-points tracking system (Max/MSP 8 Patches)
  - *Audio Game - Music Force*
    - Performance notes, score, Technical specification and Spatial diagram
    - Audio system (Max/MSP 8 Project)
    - Custom game system (Unreal Engine 4)
  - *Sound | Figuration*
    - Performance notes, score, Technical specification and Spatial diagram
    - Audio system (Max/MSP 8 Project)
    - Visual system (TouchDesigner099 project)

  - *Without strings*
    - Installation notes, Technical specification and Spatial diagram
    - Audio system (Max/MSP 8 Project)
    - Visual system (TouchDesigner099 project)
    - PerformerRNN-OSC (Application)
    - Thingy52 OSC source code
  - *Metamorphosis*
    - Performance notes, Technical specifications and Spatial diagram
    - Audio system (Max/MSP 8 Project)
    - Visual system (TouchDesigner099 project)
    - PerformerRNN-OSC (Application)
    - Joy-con receiver (Application and Source code)
  - PerformerRNN-OSC
    - Pretrained models
    - Source code

- **Software**
  Required software for the portfolio.
  - Max/MSP 8.22
  - Max/MSP 6.11
  - Wekinator 2.1.0.4
  - TouchDesigner 2021.15240

List of performances

- **Handwriting · WuXing**

– MANTIS Festival of Electroacoustic Music	Manchester, GBR
Performer: Hongshuo Fan	2019
– ECHOCHROMA XVII	Leeds, GBR
Performer: Hongshuo Fan	2019
– International Computer Music Conference	New York, US
Performer: Hongshuo Fan	2019
– Between Festival for Art, Science, Technology	Stockholm, SWE
Performer: Hongshuo Fan	2019
– The Giga-Hertz Award Ceremony	Karlsruhe, DE
Performer: Hongshuo Fan	2019

- **Intangible Field**

– University of Manchester Composition Workshop	Manchester, GBR
Performers: Vonnegut Collective and Animikii Theatre	2019

- **Audio Game - Music Force**

– University of Manchester Composition Workshop	Manchester, GBR
Performers: Distractfold Ensemble	2019

- **Sound | Figuration**

– MANTIS Festival of Electroacoustic Music	Manchester, GBR
Performer: Maria Palapanidou	2019
– International Society for Music Information Retrieval	Montréal, CAN
Performer: Maria Palapanidou	2020
– The New York City Electroacoustic Music Festival	New York, USA
Performer: Maria Palapanidou	2020

- **Without strings**

– Data Art for Climate Action Conference	Hong Kong, CN and Graz, AUT
	2022

- **Metamorphosis**

– MANTIS Festival of Electroacoustic Music	Manchester, GBR
Performer: Hongshuo Fan	2021

- New Interfaces for Musical Expression Conference Shanghai, CN
  Performer: Hongshuo Fan 2021
- RNCM PRiSM Future Music festival Manchester, GBR
  Performer: Hongshuo Fan 2021
- The Conference on AI Music Creativity Graz, AT
  Performer: Hongshuo Fan 2021
- International Computer Music Conference Santiago, CL
  Performer: Hongshuo Fan 2021
- AI and Music Festival Barcelona, ES
  Performer: Hongshuo Fan 2021
- International Computer Music Association Showcase 2022: Asia Online
  Performer: Hongshuo Fan 2022

List of figures

List of tables

Abstract

This portfolio of six compositions investigates how the IoT (the Internet of Things) era's technology may augment the current interactive multimedia performance systems and interactive music composition in terms of the novel forms of machine musicianship that can emerge from machine–machine communication and artificial intelligence technology.

The portfolio begins with *Handwriting · WuXing*, which integrates ML into the existing interactive multimedia performance system, followed by *Intangible Field* and *Audio Game - Music Force*, investigating the potential of composing interactive music that involves multiple performers by applying advanced computer vision and game engine technology. Next, *Sound | Figuration* explores the utilisation of deep learning to enhance the composition process and new visualisation technology, which provides the opportunity to realise complex compositional concepts with a creative hybrid. Then, regarding *Without strings*, not only is the implementation of the IoT concept realised through machine–machine communication without active human interference, but also the real-time application of the AI-aided composition tool. Finally, *Metamorphosis*, as a large-scale real-time interactive audiovisual composition that presents a compelling new combination of technologies and a unique aesthetic sensibility, which virtually draws upon Bianqing percussion, further extends the concept of interaction from human–machine to machine–machine and machine–human.

The thesis consists of a commentary on each piece, and supporting documentation and appendices. It includes details on each composition's initial inspiration, musical intention, compositional structure, and analysis of technology implementation.

Declaration

I hereby confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Acknowledgements

Firstly, my deepest gratitude goes to my main supervisor, Professor Ricardo Climent, for his constant motivation and guidance. Professor Climent has walked me through all the stages of my PhD journey. This portfolio could not have reached its present form without his support and inspiration.

I would like to acknowledge my second supervisor, Professor Richard Whalley, who has provided many suggestions and advice for my composition. In addition, I genuinely thank the professors and tutors at the University of Manchester: Professor David Berezan, Professor Camden Reeves and Dr Richard Allmendinger. They have advised and helped me a lot in my PhD studies. Also, I would like to express thanks to the performers who made these compositions possible.

Finally, I wish to sincerely thank my beloved parents and Lina Yan for their loving consideration and faith in me through all these years. Furthermore, I owe my heartfelt gratitude to all my colleagues at NOVARS Research Centre for their support, friendship, encouragement and positive spirit throughout my PhD studies.

Chapter 1

Introduction

1.1 Portfolio Contents

Technologies around us are becoming “smarter” and “closer” ; they are changing the world we live in and undoubtedly creating new avenues in music composition and performance practice involving them. Over the past few decades, there has been a dramatic increase in the research and composition of interactive multimedia due to technology innovation. On the one hand, advanced live audio-visual processing allows composers to realise their unique compositional intention, extend the existing language and grammar of music and enrich the musical expression. The works by Takayuki Rai [1]–[3], Hans Tutschku [4]–[6] and Wojciech Błażejczyk [7]–[9] are good illustrations of the augmented instruments approach that uses live signal processing to construct electronic sounds and interactivity to expand traditional music instruments. These works also enable the composer to rethink the compositional process as the designing process of multi-sensing experience, which could immerse the audience via live performance. On the other hand, many composers employ digital or data-driven music instruments [10] as the external musical prosthesis [11] in their creation and performance to capture musical gestures and afford new forms of sonic articulation and musical expression [12]. For example, the works by Jeffrey Stolet [13]–[15] and some projects [16]–[18] from IRCAM[1] show a great potential of human-computer interaction and co-creativity [19] in the compositional scope. Meanwhile, previous studies of AI (Artificial Intelligence) and music in the computational creativity [20] context demonstrate the possibility of applying AI to transform musical tasks, such as Google’s *Magenta* project [21], *SampleRNN* [22], *MuseNet* [23] and *Jukebox* [24] by OpenAI. These research outcomes furthermore shape new compositional ideas and enlighten numerous fantastic compositions, for instance, *Corpus Nil* [25] and *THIS IS FINE* [26]. Without a doubt, these motivate the composer to investigate the idea of human-machine collaboration in artistic practices.

This research is an artistic investigation exploring the evolution of music using interactive multimedia performance systems in the IoT (Internet of Things) [27] era. This music portfolio navigates the next generation of interactive multimedia performance systems combining cross-platform multidimensional communication and AI. The research outcomes include six interactive multimodal works exploring the evolution of a common interac-

[1] Institut de Recherche et Coordination Acoustique/Musique

tive media system. It ranges from a solo performer with acoustic instruments and electronics to various combinations of players, with live multimedia and interactive installations. The employed hardware also fluctuated as the software platform expanded and changed as the investigation progressed. The software often uses both Max/MSP Jitter programming language and a set of machine learning tools (especially deep learning). The system is a highly-integrated cross-platform multilanguage environment aiming to unveil never-before-seen forms of creative expression.

In the context of musical composition and performance, these emerging technologies, particularly connected to the human experience, provide unthinkable forms of musical expression and potential for new forms of audience engagement. This portfolio aims to comprehensively discuss and demonstrate how these technologies have shaped novel forms of machine musicianship [28] that were not possible before.

1.2 Research Enquiry

The following research questions inform the fundamental direction of the investigation:

- What new forms of machine musicianship can arise from machine-to-machine communication and artificial intelligence in composition and performance?
- How could novel forms of machine musicianship expand existing technology-based interactive multimedia systems?
- How can artificial intelligence and IoT influence existing compositional methodologies?
- Can the organic combination of media rediscover or reinforce musical paradigms?
- How can the design of an interactive performance cross-modal system be optimised to create a unique composer–performer–audience interaction?

In pursuit of answers to the research questions mentioned above, the research journey started with a piece for a single player using a basic interactive system as a point of departure, which interprets and exhibits changing behaviour in response to dynamic human musical input[29]. This was followed by a multiplayer interactive multimedia composition, expanding the system to machine responses and simultaneous user inputs. As the portfolio progressed, the source input and the starting point for musical interactions shifted from a human-centred system to a machine-centred environment, leaving the human performer to interpret and respond to the questions and answers created by the machine. This led to the investigation of machine-to-machine pure interaction, in the form of an interactive installation piece, which is exclusively based on cooperation with non-human neighbour "smart" components, but yet again, to reach common artistic goals as in previous models and to explore multidimensional interaction.

Chapter 2

Handwriting · WuXing

2.1 Description

Handwriting · WuXing (手书 · 五行) is an 8-channel multimedia interactive composition for live performance, using the hand-controlled gesture recognition system[1] of Chinese calligraphy strokes. The title contains two key components emerging from the initial inspiration for this composition: the Chinese calligraphy stroke-like dynamic hand gesture and the Chinese fivefold philosophy theory, WuXing 五行.

2.2 Inspiration and artistic goal

WuXing first appeared in the Eastern Zhou dynasty (770-256 BC) as a system that categorises the movement of the cosmos into five different agents: water, wood, fire, earth and metal [30]. After that, many traditional Chinese fields started to use the theory of WuXing to explain a wide range of phenomena, ranging from the nature of cosmic cycles to the interaction between internal human organs, and from the succession of political regimes to the composition of artistic works.

The creative goal of this piece is to combine the hand gestures with live audiovisual processing, in order to create an interactive immersive experience as an alternative perspective to WuXing. This is done by generating relationships between WuXing's five elements to provide a better understanding of its philosophy through sound and moving images, via the composer's input.

Based on the nature of the WuXing theory and its five different dynamic states, the whole composition was designed in five sections, which are strongly informed by the WuXing theory and its structural elements: water, wood, fire, earth and metal. Additionally, the theory of WuXing also provides the necessary routes to the starting material, using real-time software (Max/MSP and DIPS[2]) to generate sounds and images that mimic the aforementioned five elements. For example, the water phase starts with a raindrop-like audiovisual construction and the fire phase with the sound of firewood burning and flames.

[1] The Leap Motion Controller (an optical hand tracking module) for this version.

[2] Digital Image Processing with Sound, a set of plug-in objects that handle real-time digital image processing in Max/MSP programming environment.

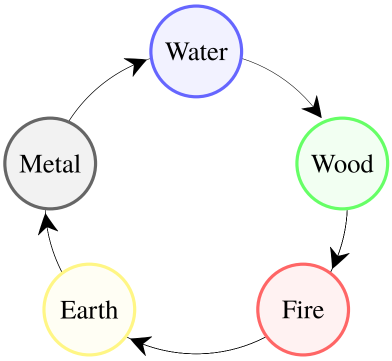


Figure 2.1: WuXing, five elements and internal cycle

Another key implementation of the WuXing theory is its internal cycle order, helping to generate various levels of interaction across nature (Figure 2.1). For example, wood can create fire, while water can put it out. This not only confirms the compositional structure of this work, but also, and more importantly, explains how basic material unfolds and evolves across time. This parallelism is found in Zhang Junfang's Seven Noties of Yunji, i.e. "Vitality can be divided into five elements, and the five elements belong into one spirit."[3]

Chinese calligraphy 书法, has been considered the quintessence of Chinese culture because it is an art that encompasses the Chinese language, history, philosophy, and aesthetics [31]. It strongly informs the design of the hand gesture recognition system for the piece, which is driven by how the Chinese writing brush is used and the way characters and strokes are structured. For instance, every character is formed by basic units called "strokes". A stroke is one continuous line of writing, drawn from beginning to end without any intentional break [31]. In the regular script of modern Chinese, there are eight major stroke types, namely, the dot (1), horizontal line (2), vertical line (3), down-left (4), down-right (5), up-right (6), hook (7), and turn (8) (Figure 2.2).

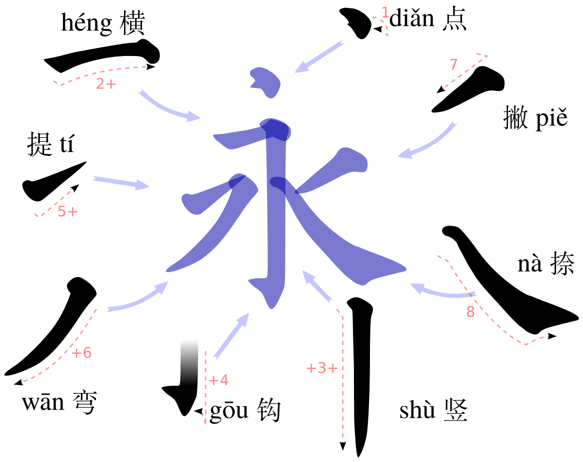


Figure 2.2: Eight Principles of Yong

My implementation of the Chinese calligraphy stroke gesture recognition system in this piece (Figure 2.3) combines machine learning and the Leap Motion controller. The machine algorithm applies a pre-trained model to recognise the eight major types of basic stroke-

[3] 《云笈七笺》北宋张君房辑："元气分而为五行，五行归于一气"，"Seven Notes of Yunji" Zhang Junfang Collection in the Northern Song Dynasty

like gestures in real-time. Machine learning also allows the system to recognise more complex gestures, similarly to the way in which different combinations of basic strokes are used to form different Chinese characters. The pattern recognition system analyses the performer's finger strokes whilst writing in the air, to identify any of the five elements.

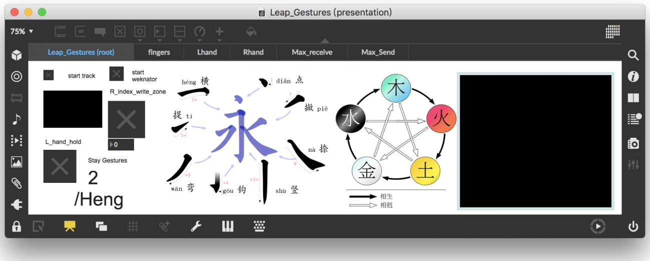


Figure 2.3: Chinese calligraphy stroke gesture recognition system interface

However, in *Handwriting · WuXing*, hand gestures not only control parameters or trigger sounds and visual events; they also constitute an integral part of the stage aesthetics and performance to communicate a higher-level compositional meaning to the audience.

2.3 Composition structure

Informed by the WuXing theory, the piece is structured into five sections and it is circa 13 minutes long (Figure 2.4). In terms of structural flow and direction, every two adjacent sections have a close relationship with each other, whereas all five sections combined provide additional meaning when the full creative cycle is completed.

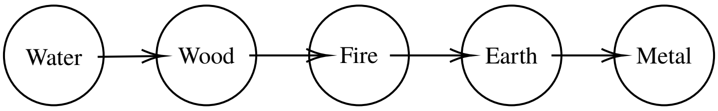


Figure 2.4: *Handwriting · WuXing* Composition Structure

Additional structural devices in the form of interlude sections were added to emphasise the central theme of the composition and to introduce new components.

2.3.1 Musical intentions

Each section starts with a synthesis sound mimicking a steady state of natural elements. For example, gentle raindrop sounds are introduced at the start of the water section at 0:50, and a breeze sound at the beginning of the wood section at 3:00. Audiovisual signal processing allows each element to shift from a steady to a dynamically intensive state. For example, in the latter part of the water section at 2:03, the raindrop gradually becomes denser in order to transform into a majestic current with greater spectral brightness. In the latter part of the wood section at 3:30, the breeze whirls and accelerates into a hurricane, which compositionally adds texture and gesture imperatives to the overall sound. Furthermore, signal processing facilitates the evolution from one element to another. Take the transition part from the wood section to the fire section at 4:35 as an example. The hurricane sound and spatial dynamics are compressed and eventually burst into flames. As a result, using these techniques of shifting visual artefacts to evolve sound, the overall music tension gradually increases across the following five sections.

This combination of machine learning applied to hand gesture recognition and audiovisual signal processing aims to explore new forms of musical expression of hand-controlled music. For instance, the key form of interaction is introduced at the very beginning of the piece, where the performer's hands directly map to control the sound frequency and related visual elements. Another example is revealed at minute 0:58. After the performer finishes a "Wan" gesture, it triggers a sonic event in which a sound object travels from the centre-front to the rear-right. In addition, dynamic gesture recognition enables the dynamic control of the composition progress, for instance, at minute 6:45, in the interlude section between fire and earth elements. The system holds all the current states until the performer finishes the Chinese character earth 土, which combines three gestures. Based on his interpolation of music, the performer can choose either to stay longer in the current section or to shift to the next section. Therefore, the infusion of machine learning in the system not only provides a novel way for the composer to use hand gestures to control musical tension and textural expression, but also allows the dynamic control of the music in order to progress during the live performance.

2.4 Interactive multimedia performance system

The system structure: According to Robert Row's classification of interactive systems [32], the interactive multimedia performance system of *Handwriting · WuXing* can be categorised into a score-driven system, which includes sequenced techniques and a player's paradigm. The overall structure of the system for this piece can be subdivided into three stages based on the three conceptual stages mentioned in chapter two, namely, "sensing", "processing", and "response", as shown in Figure 2.5).

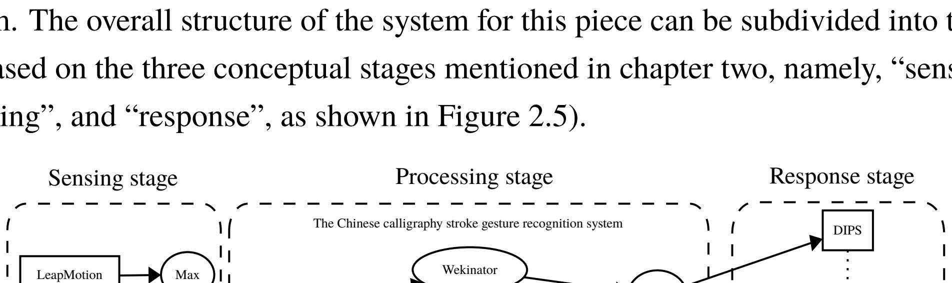


Figure 2.5: *Handwriting · WuXing* Interactive multimedia performance system structure

2.4.1 Sensing stage

In the sensing stage, the spatial raw data of the performer's hand is captured and passed on to the next stage. The Leap Motion controller, an optical hand tracking module, allows for spatial tracking. It reports comprehensive hand tracking data, such as the finger and palm positions, velocity, and acceleration. In this composition, the right index finger is used to mimic the behaviour of natural writing. This, alongside the number of recognised hands, is the primary information passed on to the next section.

2.4.2 Processing stage

The system receives raw data, which is firstly refined and then sent to the gesture recognition system to retrieve any of the eight stroke types. This is based on The Eight Principles of Yong[4] (Figure 2.2). As seen in Figure 2.6 below, the gesture recognition system combines a Max/MSP patch with a machine learning software called Wekinator[5].

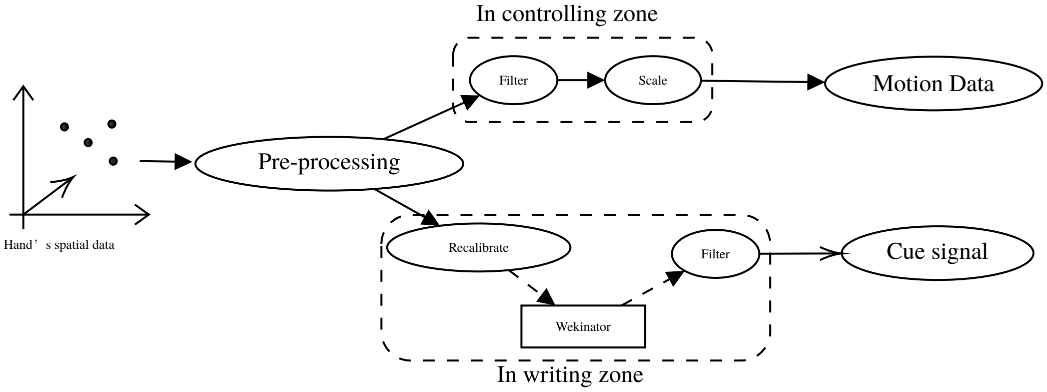


Figure 2.6: *Handwriting · WuXing* processing stage structure

1. In order to mimic Chinese calligraphy writing on paper, while maintaining a relatively natural and clear performance style, the performer uses 'air strokes' during the live performance.

2. Unlike calligraphy strokes on paper, where traces of writing remain, finger movements in the air disappear. However, it is sufficient for the system to encode them for classification purposes via the pre-trained model.

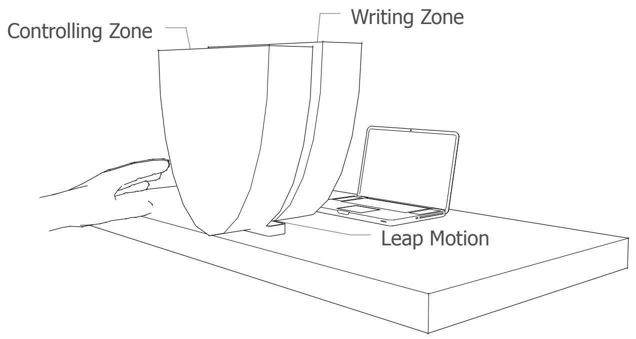
3. This recognition system uses a dynamic time-warping algorithm to recognise the gesture sequences. One of the essential conditions for this algorithm is to determine the beginning and end of the gestural sequence and it does so in two spatial zones (Figure 2.7):

   - The control zone is where the Leap Motion controller starts to sense the hands and fingers. In this space, the hands do not cross the controller's perpendicular centre plane. Here, the Max patch starts preprocessing all the hand's spatial data. For instance, the Max graph user interface part displays the finger position and directly scales and maps the data to generate refined data to pass on to the next stage. However, not all the data is sent to Wekinator.

[4] 永字八法; explain how to write eight common strokes in the regular script which are found all in one character

[5] Software for real-time, interactive machine learning

- The writing zone is the area where the right index finger passes the controller's perpendicular centre plane. Here, the Max patch re-calibrates and creates a starting point for the gesture sequence; then, it calculates the relative spatial location information and sends it to Wekinator via OSC—Open Sound Control [33]. Next, Wekinator is activated to measure the similarity between the input gesture sequence and the pre-trained gesture sequence model. Once operations in the writing zone are completed, the finger passes the perpendicular centre plane and returns to the controlling zone. At this moment, Max/MSP creates an ending point to wrap up the gesture sequence and to collect the recognition patterns extracted by Wekinator.



;

Figure 2.7: The control and writing zones of The Chinese calligraphy stroke gesture recognition system

The whole area for spatial interaction is similar to operating on imaginary paper. Finally, the stroke gesture recognition results are passed on to the response stage to trigger predetermined events. For instance, a couple of single stroke gestures that appear in a specific order could activate a special character recognition event.

2.4.3 Response stage

As indicated in Figure 2.8, the response stage contains two signal processing systems in response to the performer's real-time input. With the exception of some prestored data, most system parameters are modified in real-time by the performer as the piece progresses. For example, tracing the visual image is a real-time input process, but some parameters of the image are prestored in the system and cannot be modified. Another example of predetermined data is the fundamental frequency of the sound generated by the algorithm synthesizer subpatch in Max/MSP. However, this fixed value is modulated and mapped to the speed of the performer's hand in real-time.

- The audio signal process system comprises multiple sound-generating and signal-processing modules. They are connected via a multidimensional matrix to produce the

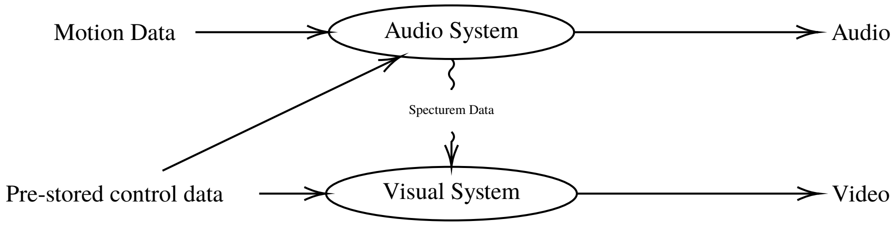


Figure 2.8: *Handwriting · WuXing* response stage structure

final output. A number of key sound modules implement sound algorithms, evoking natural sounds from Designing Sound [34]. However, the custom signal processing modules (Table 2.1) are custom designed by this author, based on the following rationale:

1. Custom modules can speed up the sonic exploration process since the user is also the creator. Their design is based upon the author's artistic aims and a deep understanding of their processing capabilities.
2. Their level of integration in the interactive system is high, as they were created for the same programming environment (Max/MSP).

Table 2.1: *Handwriting · WuXing* audio signal processing modules

Name	Function Description
Multilayer delay	Apply seven delay threads to the audio input, in which the delay time and level of each thread can be adjusted individually.
Resonator	Apply multiple classic comb filters with a feedback chain to the audio input to create a polyphonic resonance effect.
Stretch effect	Apply real-time loop recording to the audio input. Then, play back the recording through multiple layers with variable speeds.
Feedback with pitch shift	Multiple audio feedback with pitch shift during the processing loop.
Granular FM	Real-time audio granular synthesizer. The audio granular playback speed is based on a reference decimal (micro) pitch table.
Freeze reverb	A reverb effect taken from the implementation by Juhana Sadeharju.
Loop sampling	Apply loop recording and forward/backward playback to extend the input audio.
Vibrato	Apply frequency modulation to the delay thread of the audio input to create a vibrato effect.
Granulator	Apply multiple playbacks of the audio recording to create a granular cloud effect.
Granulation	Apply multiple fast random playback of the audio input clip to simulate the granular cloud effect.
Spectral delay	Combine the spectral filter and feedback to transform the input audio.

- The visual signal processing system is based on DIPS—Digital Image Processing with Sound, a library collection and a set of plug-in objects for Max, which handles real-time images [35].

Both systems follow a similar methodology and system structure:

1. According to the composition structure, each musical section interfaces with a relatively independent area of the designed systems. This includes data presets and control signals, which enable the navigation from section to section during the composition process and rehearsals.

2. All the signal processing modules are routed in a multidimensional signal matrix, as shown in Figure 2.9. In this way, every signal processing module can receive the signal from other modules, control the amount of received signal and send its output to other modules.

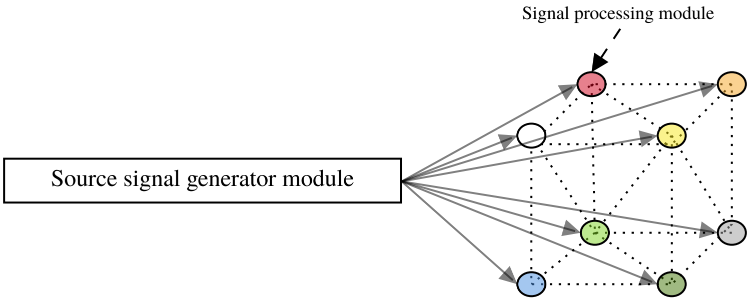


Figure 2.9: Multidimensional matrix structure

This modular system design and its multidimensional matrix structure can not only expand the avenues for sound transformation for each module but also maintain a coherent sonic character throughout. This allows raw materials to be transformed into complex new materials via the combination of various interventions in different modules, in line with WuXing's high-level concepts of modular interaction.

Chapter 3

Intangible Field

3.1 Description

This multimedia work is an 8-channel interactive composition for three music performers and two physical theatre actors. It was written for the Vonnegut Collective [36] and Animikii Theatre [37], as part of a series of composition workshops at the University of Manchester. The composition's title is a metaphor for a person's aura, which is influenced by others and also influences others. This work builds upon the previous exploration of musical performance interaction with an audiovisual system. It does so by incorporating a visual tracking system capable of detecting more than one rig/skeleton on stage with the aim of constructing multidimensional forms of interaction, for instance, exploring new relationships between the performers and how an extended interactive multimedia system can enhance those relationships.

3.2 Inspiration and artistic goal

The inspiration for this composition emerged from the workshop's theme: "breath". The focus was to explore a living creature's breath and how such a figure could create a "field" around it. The way the field is musically portrayed is by experimenting with concepts of attraction, repulsion and equilibrium, for instance, looking at how eager some fields are to get closer to another creature's field, and how excessive closeness can have a negative impact, eventually suffocating the interaction and becoming a symbiotic relationship. An example in nature of these mutual interactions exists between animals and plants, where the animal cannot live without the oxygen produced by the plant, and the plant also needs the carbon dioxide produced by the animal to survive. However, some level of equilibrium is needed because oxygen in a high content or an excess of carbon dioxide could be equally harmful. As a result, two fascinating ideas start to emerge:

1. The musical shape of acoustic instruments can be driven by the actor's movement and physical gestures. This piece attempts to establish a multidimensional interaction relationship (Figure 3.1) emerging from the choreography of actors and players on stage, in which their body movements influence one another's behaviour.

2. The live generated audiovisual content could also be enacted as an additional layer, to enhance stage interaction. For example, when one actor moves toward a music performer, the interactive system responds to this action, and the distance between them will alter both the state of sound and visual elements generated by the system. Additionally, those elements could subtly guide the performer's improvisation and reveal the artistic and abstract concepts behind these movements and gestures.

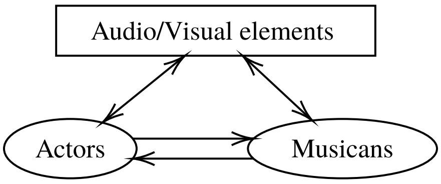


Figure 3.1: Multidimensional Interactive Relationship

To achieve the multidimensional interaction mentioned above, the composer must consider the following:

1. Limitations in the performance area: The camera tracks multiple people on stage within its field of view and estimates their poses [38]. However, if one actor or player is blocked by another or moves out of the camera's field of view, the system loses track of the performer.
2. Actors move more frequently than the instrument players and will require distinct settings and boundaries, which will be referenced by the musicians on stage.
3. Different performance notes and rules must be given to the two types of performers.

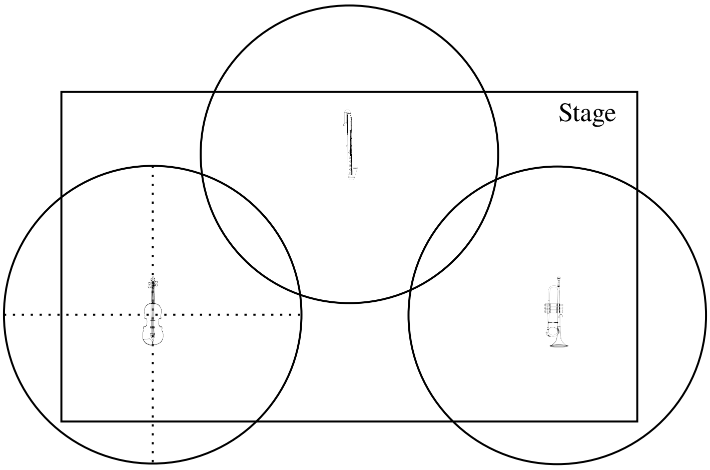


Figure 3.2: *Intangible Field* "Sound Control Field"

The above includes the initial position of each player and a number of basic rules relative to other performers. Alongside this, players receive musical directions in relation to pitch, register, instrumental techniques and musical directions.

All performers need to be aware of their respective stage boundaries and spatial outlines in relation to one another. For instance, the violinist needs to be on the far right of everyone else, the trumpet player on the far left and the bass clarinettist on the innermost side.

Additionally, each instrumentalist has a relatively independent "sound control field" (Figure 3.2), affecting the responses generated by the system, while being affected by any other stage performer who moves into this field. One of the essential rules for actors is that their performance area is relative to the location of the instrument players. For example, when the musicians are at their initial set position (Figure 3.3), actors can move freely around a trapezoidal area in the centre of the stage.

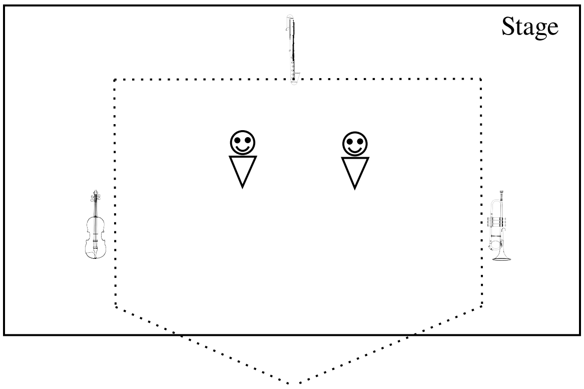


Figure 3.3: *Intangible Field* Initial Performance Area

Actors can affect the signal processing of instrumental sounds by crossing the boundaries of instrument players "sound control field". For instance, if one actor moves towards the violinist's "sound control field", the procedural sound of the violin will "freeze" and change based on the actor's movement, and the distance between the actor and the owner of the "sound control field", the violinist. Hence, these fundamental rules and indications effectively restrict the performance area and create a dramatic interaction between actors and musicians.

3.3 Composition structure

Intangible Field is circa 7 minutes in total, and it is structured into three themed movements, entitled: "seeking", "chasing", and "suffocation". Each movement contains different instructions for actors and musicians. The vast majority of instructions require the performer to improvise based on three elements:

1. The given musical materials.
2. The behaviour of the rest of the stage performers.
3. The nature of the audiovisual element.

Procedures: Before starting the piece, all performers must stand at the initial location on stage (Figure 3.4), and then wait for the interactive multimedia system initialisation and calibration. This includes the identification of performers according to location and the designation of boundaries.

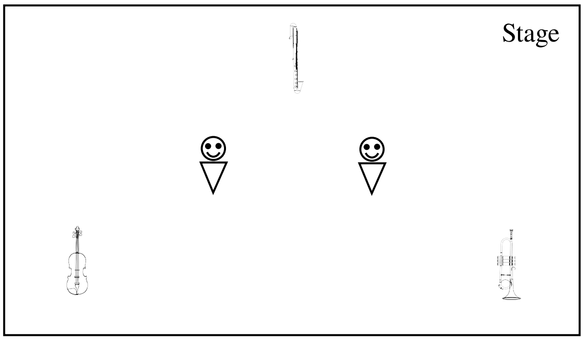


Figure 3.4: *Intangible Field* Initial Position

3.3.1 Musical intentions

For every movement, the musical tension gradually increases at a slow tempo, as the complexity of the texture builds up, to slowly scale down until a steady state is reached. The interactive relationships between audio, visual and movement evolve across time, ranging from one-dimensional to multidimensional relationships. They determine the nature of sonic transitions and their pace. In order to achieve the musical directions above, at the very start, musicians are instructed not to play while another musician is playing solo. They are also required to increase the frequency of the sound objects projected from their instrument. Similarly, actors can only move toward the musician who is playing and accelerate their movements along with the pace of the music.

Combining the technology and processes behind the system with the choreography of the stage performers aims to find musical expression at the intersection of dynamic physical motion and spectral/gestural content created by the players and signal processes. Take the start of the second movement 1:30 as an example. As the violinist plays and slowly moves around her control field, she musically projects the desired textural tension emerging from the technology implemented. Another textural variation is introduced at minute 4:17, where a chaotic gestural crunch is achieved via an increase in tempo and the juxtaposition of musical textures. On stage, this happens when actors interact with the musicians' auras and their boundaries. This compositional strategy provides the means for the composer to use the choreography on the stage as the main creative tool to control musical tension and textural expression.

3.4 Interactive multimedia performance system

Intangible Field's interactive multimedia performance system is realised in the Max/MSP graphical programming environment. In order to achieve a meaningful composition structure and multidimensional interaction in the piece, the system involves three stages, as Figure 3.5 indicates. Each section has prestored data for macro-control, which is triggered to set up the interactive environment. However, as the piece progresses, the system does not rely on any prerecorded music or image material. Instead, the system analyses the environment and applies meaningful transformations to the live input signal, both sounds and images, in order to extend the musical instrument and to enhance the levels of interaction between the stage performers. The system embraces a hybrid form, where at the macro-level, it acts as a score-driven program, and at the micro-level, becomes a performance-driven program, where transformations become the system's response.

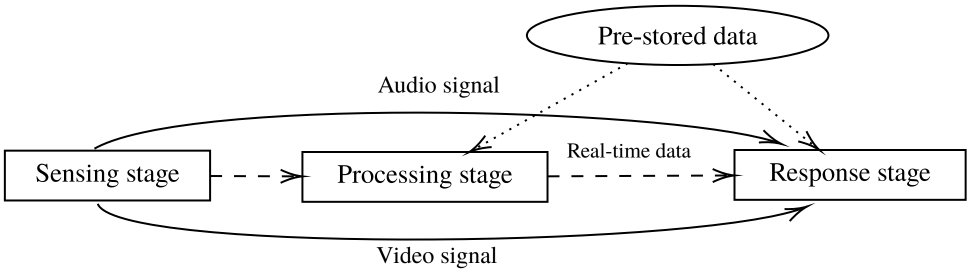


Figure 3.5: *Intangible Field* Interactive multimedia performance system structure

3.4.1 Sensing stage

In the sensing stage, as Figure 3.6 shows, the system collects three different types of raw data and passes them on to the next stage. These include:

1. Live audio signals, via individual wireless microphones for musicians.
2. The stage, using a live feed from a webcam. This serves two purposes: one is to apply transformations to the dynamic image to produce visual variants, and the second is to capture body movement and map the key points to the skeleton. The latter is achieved via the implementation of PostNet [39], a machine learning model, which allows the use of a single RGB image to estimate a human pose in real-time. The webcam reports the position data of 17 key points in 2D and the X–Y coordinates for every performer detected.

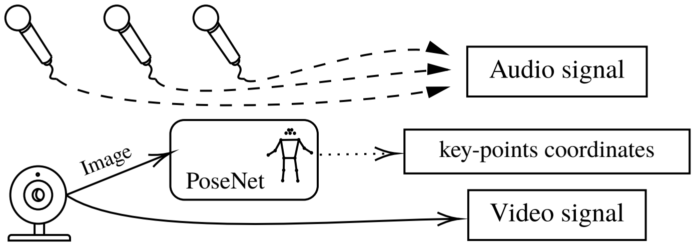


Figure 3.6: *Intangible Field* Sensing stage structure

3.4.2 Processing stage

The visual tracking system can only provide 2D coordinates on the vertical plane, which becomes problematic, as the primary interaction between performers occurs in the X and Z in horizontal planes. Therefore, this method requires two substages to calculate the depth coordinates, as Figure 3.7shows:

1. Firstly, in the preprocessing stage, the performer's key points, i.e., X and Y coordinates, are analysed to produce the Z coordinates . The human visual perspective can be understood as two points on a vertical flat surface, where the eye distance is only affected by the change in the X and Y coordinates. Based on the linear perspective principle, this method aims to estimate the Z coordinate in order to calculate the distance between performers as they move.
2. Secondly, in the postprocessing stage, the system reads and interprets each performer's coordinates and produces data in response, following the following steps:
   (a) It draws each performer's position on a bird-view map of the stage.
   (b) It identifies each performer's identity based on the system's settings in which the musicians are on the far-left, far-right and innermost areas on the map, and the actors are in the centre of the map.
   (c) It calculates the distance between each performer.

   Alongside the visual system above, the computer also analyses each instrumental sound and interprets several parametric data, including the pitch, velocity and frequency spectra. Finally, such data is sent to the next stage.

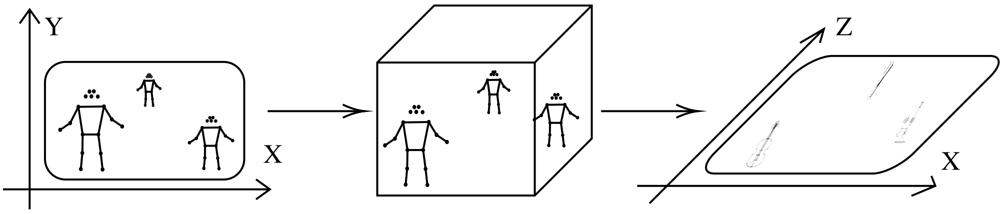


Figure 3.7: *Intangible Field* Processing stages substages

3.4.3 Response stage

The response stage embraces both audio and visual signal processes:

- The audio part: It comprises four different signal processing modules (Table 3.1) and is controlled by the "sound control field" mechanism. Three "sound control fields" are visualised in the form of three circled areas on the user interface's top-view map (Figure 3.8 right). At the centre of each circle area, there is an instrument player and its radius is eighty per cent of the distance between each player.

  Each circle is divided into four quadrants, as the left section of Figure 3.8 indicates. Each quadrant contains an opposite sector area, which represents one audio signal processing module. This provides plenty of flexibility for musical expression, as it maps the relative location to certain sound transformations, while setting the ground for aura interaction, as reflected in 1: 41. The centre of the sector area is located in the middle of the quadrant. As players move, the relative distance between the instrument player

Table 3.1: *Intangible Field* audio signal processing modules

Name	Function Description
Multilayer delay	Apply seven delay threads for the audio input, in which the delay time and level of each thread can be adjusted individually.
Feedback with pitch shift	Apply multiple audio feedbacks with pitch shift threads to the audio input during the loop.
Granular synthesizer	Real-time audio granular synthesizer. The audio granular playback speed is based on a reference decimal (micro) pitch table.
Loop sampling	Apply loop recording and backward playback to extend the input audio.

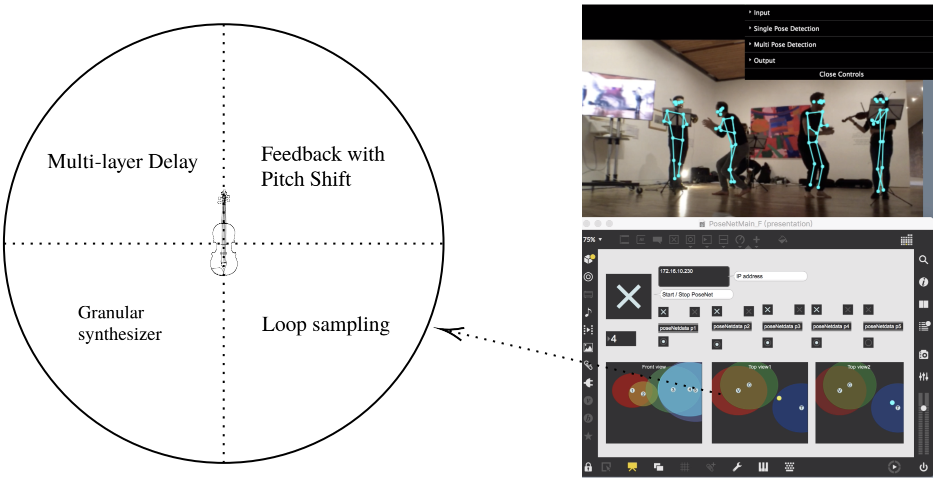


Figure 3.8: "Sound control field" quadrant diagram (left); tracking system graph user interface (right)

and the centre of the sector determines the module's output ratio. This ratio is essential for determining the timbral characteristics of the produced sound. For example, moving from a specific stage location to another would lead to changes in spectral occupation as pitch shifters become more active. Thus, with relatively slow action in the first movement, it directly reflects the interaction relationship between the performers and the system. Then, with the increase in the speed and frequency of their movements, it not only creates a more rich timbre to elevate the musical tensions, but also serves as positive feedback to encourage them to do so in the latter part. Equation 3.1 and Figure 3.9 indicate how a single module's output ratio is calculated, based on:

- The instrument player's position, $P(x, z)$.
- The actor's position, $A(x, z)$.
- Factor a: if the player is inside the quadrant, it is 1, or if not is 0.
- Factor b: if the actor is inside the field, it is 0.5, or if not, it is 0.
- The radius of the field, r.
- The centre position of the sector area, $S(x, z)$.

Similarly, the instrument player's position controls a number of spatial attributes in the music, including the panning allocation in the ring of the 8-channel loudspeaker system.

$$OutputRatio = 1 - \frac{a * Distance(P, S) + b * Distance(P, A)}{2r} \tag{3.1}$$

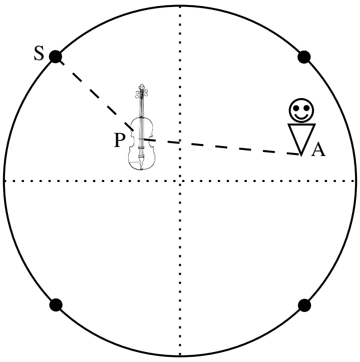


Figure 3.9: “Sound Control Field” module output ratio calculation diagram

- The video part: This signal process applies fast fluid dynamics simulation [40] to live and fixed visual materials in different sections of the piece. For example, the performer’s position in the first and second movements of the piece is linked to the visual element and is responsible for its several transformations, e.g., smoke-like real-time traces (Figure 3.10). This helps to visualise the complex nature of aura interactions resulting from performers’ exploration of the physical space.

  Another example of musical/visual interaction is found in the third movement of the piece, where a liquefied dynamic image serves as a metaphor for each person’s aura to further portray aura tension when performers start invading each other’s fields. Aura tension is mostly achieved by gradually reducing the distance between the musicians. It first ensures that all actors are inside all the fields; then, each musician is placed in a different sector, which activates different signal processing modules to produce highly complex timbre. When the distance between them is close to the limit, and the musicians are at the farthest point, the output ratio reaches the maximum, which pushes the music to the climax.

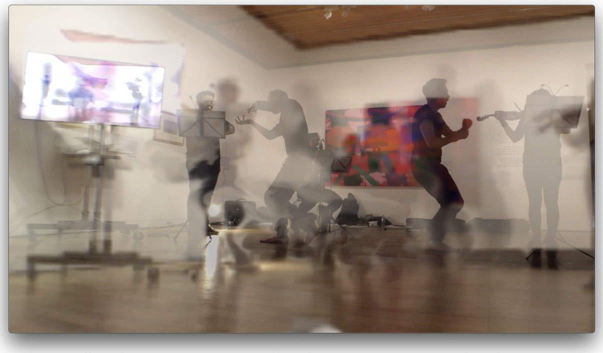


Figure 3.10: *Intangible Field* fluid dynamics simulation screenshot

Chapter 4

Audio Game - Music Force

4.1 Description

Audio Game - Music Force is an interactive composition for game audio with three musical instrument players. It aims to explore musical composition and performance as gamified experiences using the visual element (the game) as some sort of musical score to organise sound. The piece was written for the contemporary music ensemble Distracfold [41]. Technically speaking, the composer integrated Epic Games Unreal Engine [42] and Max/MSP (for sound). As a result, there is a high level of interaction between the players and the performance system, which is manifested in the audiovisual contract [43] of the piece. Furthermore, it brings a fascinating audiovisual music experience to the audience.

4.2 Inspiration and artistic goal

One of the main artistic goals of this piece was to apply the typical game-design aesthetics [44] as the compositional strategy to develop the audiovisual experience. The overall creative process is similar to designing a game but with three key differences:

1. Duration and structure: Unlike a traditional adventure or role-playing game, players have enough time and a certain level of autonomy to discover and explore the game space. Because the stage performance has a limited duration, a relatively fixed structure in the time domain and fixed game space are required.

2. The interaction method: This piece intends to use music to control the game character during the performance instead of the general game controller, similarly to voice-controlled games [45], such as *Bot Colony* [46]. However, the musicians can not learn this new method by trial and error like in regular gameplay. So, this piece required a novel interaction method that balances out the challenges in the control method with the need for musical expression.

3. The game mechanisms: Traditional games continually give goals to players, increase the difficulty of the challenge, measure the players' progress, and give rewards to motivate the players to continue playing. The composer aims to subtly apply the aforementioned classic game mechanisms as a compositional tool to navigate the audio-

visual experience. For example, in the first game level starting at 1:55, the musicians need to control the main character to avoid the incoming obstacles.

These three key differences allowed the composer to identify the most suitable game category: the side-scrolling or horizontal-scrolling game. A typical side-scrolling game is a game in which the player views from a side-view camera angle, and as the player-controlled character moves left or right, the screen scrolls with them; one of the most famous side-scrolling games is *Super Mario Bros* (1985). However, another side-scrolling shooter video game, *Gradius* (also known as Nemesis) [47], provides a more suitable model and inspiration for this piece. During the gameplay (Figure 4.1), the player needs to control the behaviours of a spacecraft that moves in a 2D space to collect items, and avoid obstacles and enemies, and control the gun system, in order to defeat the enemy.

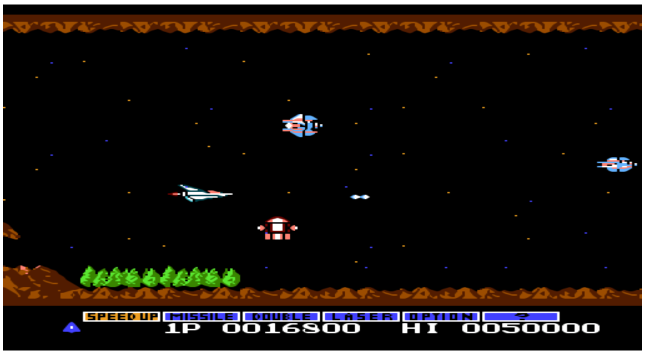


Figure 4.1: *Gradius* Main game view

These control mechanics allow each musician to control one part of the character's behaviour. However, they need to collaborate in order to progress further in the game and, therefore, the composition. For instance, players need to map a number of changes in musical parameters to the character's movement. For instance, pitch and velocity. The system converts these players' responses into data signals that control character-on-screen WASD movement (Figure 4.2) and her/his ability to shoot the enemies. For example, changes in the violin's pitch are mapped to the force that drives the spacecraft's vertical movement at minute 2:00.

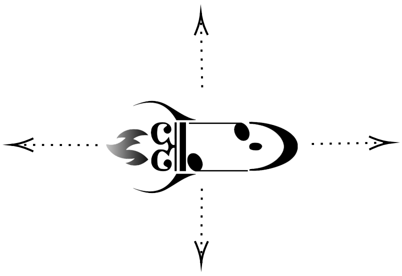


Figure 4.2: *Audio Game - Music Force* Player's character and moving direction

Additionally, the system produces audiovisual events as feedback emerging from players' actions during the game level, for example, from minute 4:20, as the second game level starts. Here, if the enemies are destroyed, the system will synthesise special sound effects to provide feedback on progress. As a result, signal processing is used as a guiding device, while creating more tension as a sense of progression between game levels. Take the third game level starting from minute 6:25 as an example; once the shape of the character shifts, the role of the transformation processes is to add external texture to the instrumental sound to reflect these changes.

Another essential reference for this composition is Prof. Ricardo Climent's game-audio composition, *Duel of Strings: for Violin (non-virtual) vs. Virtual Strings*. It explores issues of representation in interactive media composition incorporating virtual reality through an onstage musical battle between two performers as they navigate a virtual world [48]. More importantly, his compositional methodology centres on the game-engine interactive system, which uses the custom system to reconstruct the deployed sounds, phrases and musical ideas through playing the game. This approach allowed the composer to design the system as a dynamic score that can reconstruct the composed music. For example, in the first cutscene section at the beginning of the piece, the system applies a slight reverb to the instruments' sound to recreate a sense of vast space; then, it gradually adds the feedback with pitch shift to the sound to add a sound layer that enriches the harmonicity. Furthermore, it provides a model to optimise the performers' musical expression on stage via improvisation in response to real and surreal cues. Take the second game-level section that starts at minute 4:16 as an example. During this section, the sonified enemy's behaviour creates a collision with the instrument sound to augment the musical tension. It also becomes a cue to lead the musicians' improvisation.

4.3 Composition structure

Audio Game - Music Force is circa 10 minutes in total. It has eight sections, namely, four semi-interactive cutscenes [49] and four fully interactive game levels, as Figure 4.3 shows. The composition structure follows a basic storyline in which musicians control the "music spacecraft" to defeat the enemy and to retrieve the "music core" together. To achieve their mission, musicians need to pass four game levels, including challenges with increasing difficulty, namely, the first level, passing the danger zones to get the "weapon"; the second level, collecting energy to upgrade the "music spacecraft"; the third level, breaking the "enemy defences zone"; and the fourth level, defeating the final boss.

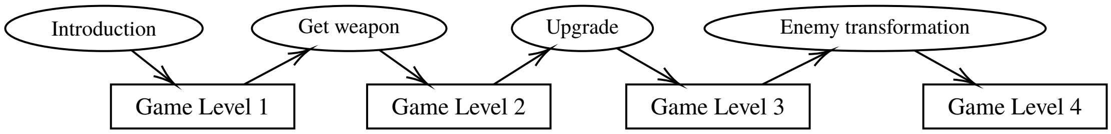


Figure 4.3: *Audio Game - Music Force* Composition structure figure

To provide the musicians with a certain level of autonomy and control, the score uses three staff lines for musical notation instead of the standard five staff lines. These lines do not indicate specific pitches but a defined register, low/medium/high, which is relative to the range of the instrument and technique. The middle line always indicates the pitch of the first note being played after each interval. For example, as Figure 4.4 indicates, if the violin reads the score, the first note on the middle line could be C5; the second note on the top line could be G6; and the third note on the bottom line could be G3.

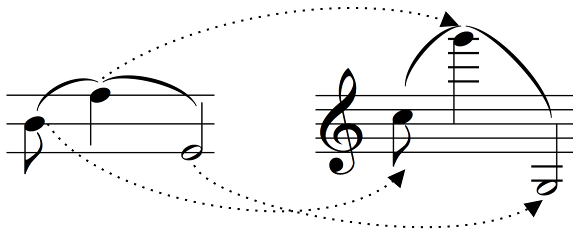


Figure 4.4: Three staff lines notation example

The cutscenes are interlude sections between each game level to narrate the storyline. They contain a fixed animation sequence and various signal processing techniques, allowing the musicians to pin the different game-level narratives. Take the first cutscene section at the very beginning of the piece as an example. It not only establishes the fundamental atmosphere of the work, but also introduces a number of in-game relationships and the overall storyline.

4.4 Interactive multimedia performance system

In *Audio Game - Music Force*, the Max part processes and analyses the audio input in real-time, converting the music information into a control signal to navigate the visual part in the Unreal Engine, using the OSC (Open Sound Control) protocol. The Unreal Engine processes all the logical events to generate the visual element, and then returns the signal back to Max via OSC. In doing so, the Max part applies sound transformations and produces sonic events in real-time for the video-game part in a constant feedback system. As a distinctive exploration of interactivity methods, the piece explores the use of predetermined score-driven cutscenes and performance-driven game levels.

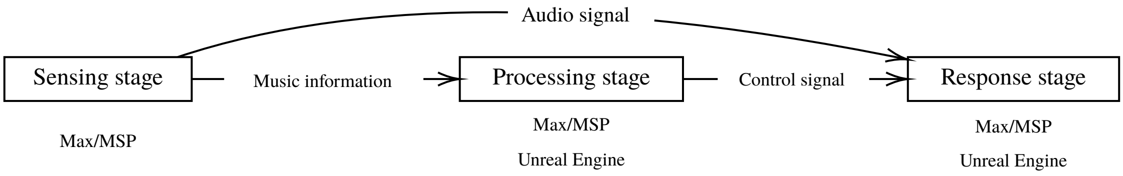


Figure 4.5: *Audio Game - Music Force* Interactive multimedia performance system structure

4.4.1 Sensing stage

As Figure 4.6 shows, the system uses three microphones to capture each instrument sound in the sensing stage and analyse their variation state in Max/MSP patch. The YIN algorithm-based monophonic fundamental pitch estimation module [50] analyses the filtered input fundamental frequency and energy data in real-time. It also prepares the data to be passed on to the following stages.

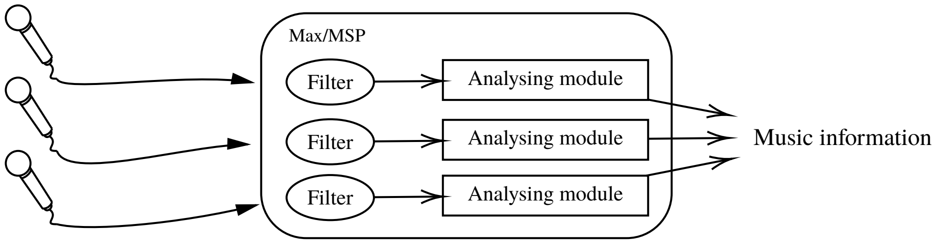


Figure 4.6: *Audio Game - Music Force* Sensing Stage Structure

4.4.2 Processing stage

The processing stage applies two steps to achieve the music control mechanism of the game character. Figure 4.7 indicates one example of this:

1. Firstly, it converts the continuous music information into nonlinear trigger signals: Because the game character is designed to be operated by three musicians simultaneously, the game character is intended to respond to five types of nonlinear control signals, namely, moving up and down, moving left and right, and shooting. Hence, to convert the data from linear to nonlinear, the system requires the interruption of the fundamental frequency process chain using the changes in audio energy, as it crosses a specific threshold. However, if the incoming signal does not change in a short period, this indicates that the input audio energy is stable. Thus, the system does not produce any control signal since the frequency data has been intercepted and cannot move on to the following step. In contrast, if the level changes in a short amount of time, it means that the input audio energy is changing. Therefore, the system passes the incoming frequency data to the next step.

2. Secondly, it determines the direction of the frequency variations to generate corresponding control data based on its tendency, as follows:

   (a) The system computes the difference between the new input and the previous values to obtain the direction of its transition.

   (b) Each musician's frequency signal is mapped to the corresponding control signal according to their role assigned in the composition.

3. Finally, within the engine, data collection sent back to Max includes game state data, such as the character's respawning, position and number of enemies destroyed. Max and Unreal reciprocally prepare key performance data, which they pass on to each other before the piece moves to the next stage.

This mechanism enables a new form of a dialogue between the musicians and the engines that contributes to the creation of novel texture, timbre, and harmonicity. This encourages the musicians to utilise different musical gestures to respond to the in-game changes. For example, at minute 4:26, the musicians use a less dynamic musical gesture to control the character since there are only two enemies to defeat. As a result, the sound's texture and

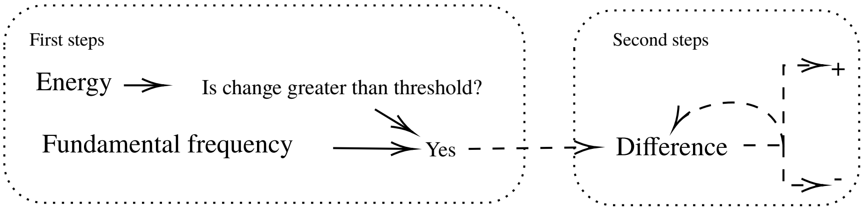


Figure 4.7: *Audio Game - Music Force* Processing Stage

harmonicity are relatively uncomplicated. However, from minute 6:45, the musicians increase the rhythmic density and pitch variation to combat the increasing number of enemies.

4.4.3 Response stage

The response stage has two models: the cutscene and the game-level models.

- During the cutscenes (Figure 4.8), the system plays back a predetermined animation sequence in Unreal for Max to trigger some prestored parameters, which are mapped into the signal processing modules (Table 4.1). These modules apply transformations to the audio signals to extend the musical instrument and produce an extra layer of sound texture.This highly synchronized audiovisual contract seeks musical expression and a sense of narrative at the intersection of the two media. For example, events such as the upgrade of the shape of the spacecraft at 6:00 hold new (audio-to-visual) mapping settings. These are used to trigger distinctive musical devices, revealing more colourful textures and timbres across the instruments: the violin's spiccato sound is used to create a granular sound layer that matches the particle visual effects; the clarinet ʻs sound expands its harmonic textures via the pitch shift and feedback.

Table 4.1: *Audio Game - Music Force* audio signal processing modules

Name	Function Description
Multilayer delay	Apply seven delay threads for the audio input, in which the delay time and level of each thread can be adjusted individually.
Feedback with pitch shift	Apply multiple audio feedbacks with pitch shift threads to the audio input during the loop.
Granular synthesizer	Real-time audio granular synthesizer. The audio granular playback speed is based on a reference decimal (micro) pitch table.

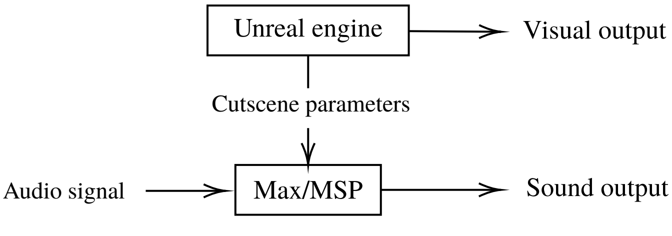


Figure 4.8: *Audio Game - Music Force* Response stage cutscene model

- As Figure 4.9 indicates, Unreal applies a moving force to the game character based on musical energy from the instrument, to naturally handle the game character's movement. This basically means that the mapping between musical parameters and game control is the key compositional method to create unique modes of interaction and musical deployment in the game engine that audiences perceive as natural artefacts in musical performance. The reappropriation of game-engine technology for compositional purposes provides endless opportunities for the composer to map music and moving images through the mediation of instrumental performers on a stage. For instance, at minute 8:30, as the player character sustains damage, Unreal shakes the screen, and the Max part triggers a noise synthesis method to generate matched sound worlds. Another example of a musical-to-visual device occurs at minute 8:52, when the enemies are destroyed. As a result, both parts produce a synchronised visual and sonic explosion.

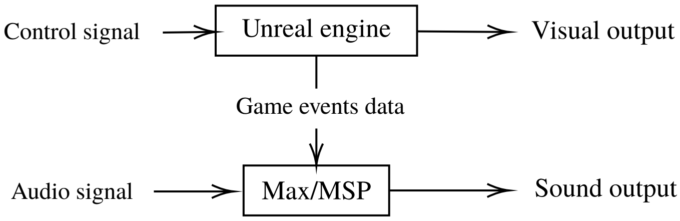


Figure 4.9: *Audio Game - Music Force* Response stage game level model

Chapter 5

Sound | Figuration

5.1 Description

Sound | *Figuration (声|形)* is a live interactive composition for piano, multimedia and machine learning. In this piece, machine learning technology serves as a compositional tool, which utilises deep learning [51] to explore the aural world and the sonority of the piano. The composition makes use of TouchDesigner software for the visual part and Max/MSP for the sound part, as the basis for live transformations carried out by the pianist during the performance. As a result, the piece provides the listener with an aural experience aiming to unveil processes of musical emergence, growth and distillation.

5.2 Inspiration and artistic goal

An important part of this portfolio explores the next generation of machine musicianship via novel IoT technologies and machine learning [52], in particular, neural networks and deep learning, as increasingly important methods to unveil new forms of creative expression rarely seen before. For instance, *Handwriting · WuXing* combines a dynamic time-warping algorithm to recognise and classify a series of finger gestures via machine learning. Similarly, *Intangible Field* uses deep learning pre-trained models to track multiple human bodies in real-time. These supervised learning technologies [53] provide the possibility of widening musical ideas and levels of expression. However, they are not directly controlled during the music composition process, making the creative flow rather unusual. Nevertheless, the main artistic goal in *Sound* | *Figuration* is to intentionally utilise unsupervised learning technologies to inspire creativity during the composition process by enhancing creative methods via the incorporation of non-human decisions.

The first step toward integrating AI into the compositional process of this work was to construct an aid tool that could generate music information based on the pre-trained model and specific rules. With the assistance of this tool, the composer obtained a number of source materials from the piano, which helped to compose the whole score for the piece. An essential inspiration for this piece is the art manifesto by composer Takayuki Rai, who defined his music as 'four-dimensional sculptures', being himself a sculptor in a five-dimensional world [54]. The concept of music as a four-dimensional sculpture provides a clear direction for visualisation in the piece. As a result, the music acts as the driving force, shaping the transformation of dots into lines, lines into faces, and faces into bodies. It also visually represents the transformation of low-dimensional objects into high-dimensional ones, while becoming a metaphor for the projection of moving images of high-dimensional objects in a low-dimensional space.

5.3 Performer RNN, AI-aided compositional tool

Another example of AI for this piece includes supervised learning, where neural networks determine optimal solutions to tasks from a labelled dataset. For example, it uses a classical "dog vs cat" classifier dataset [55], where all the image files are labelled. In this example, the neural network attempts to learn the features of each category in order to distinguish new input images. However, unlike supervised learning, in typical unsupervised deep learning methods, the training dataset fed into the neural network is unlabelled. Therefore, the neural network needs to learn the relationship between each data value to be able to predict the next one. For instance, by feeding a neural network with a full year of weather conditions, it will try to find out any temperature change pattern, in order to forecast future temperatures. Therefore, unsupervised learning is usually suitable for dealing with time-series-based questions, like weather forecasts, signal processing, and natural language processing. However, it is also potentially disruptive when incorporated as part of the composition methodology.

Many successful studies, such as OpenAI's Jukebox [24] and DeepMind's WaveNet [56], focus on applying deep neural networks to generate and transform sound sample by sample. This research raises awareness about the great creative potential behind applying deep learning algorithms to music and sound. It also highlights the massive amount of computational power, time and size of the dataset required to train a model that can generate unique and meaningful musical results [57]. Therefore, it is quite challenging to utilise this technology in a live interactive multimedia performance with the off-the-shelf available computer hardware. In this piece, the composer decided to apply a more practical approach to the real-time generation of music using this type of algorithm. Therefore, the idea was to process and produce symbolic music information, such as pitch, velocity and time parameters, instead of computationally expensive raw audio data, and then apply that information to drive signal processing modules and for resynthesis. This approach significantly reduced the amount of computational power and time required for training and evaluating the data but also provided the means for the composer to reuse the trained model in various compositional contexts.

The first version of this AI compositional assistant tool, Performer RNN, integrates the Magenta's Performance recurrent neural network structure [58] in the Python programming environment and the Bach library [59] in the Max/MSP via OSC (OpenSoundControl protocol). In this system, the Max part serves as the main tool interface (Figure 5.1), and the core of the tool is a pre-trained long short-term memory [60]-based recurrent neural net-

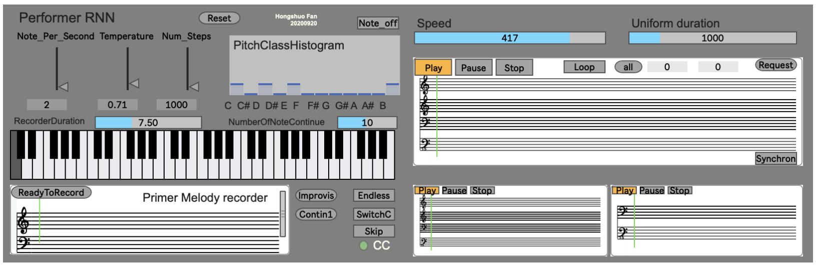


Figure 5.1: *Performer RNN* user interface

work model. The latter generates music information based on received rules; then, it sends it back to Max. To achieve a higher efficiency, more sensitive results and a greater macro-control of the trained neural networks, the system has a similar three-stage structure to the interactive music systems explored before, as Figure 5.2 shows.

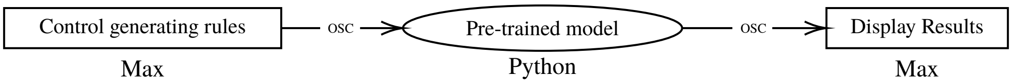


Figure 5.2: *Performer RNN* System structure

1. The system collects and encodes the generated rules in the first stage, then passes the data on to the next stage. The Max graphic user interface allows the user to change the rules generated for the different musical parameters, such as rhythmic density, temperature[1], principal melody, pitch-class range and the number of steps (Figure 5.1 left side). In musical terms, these generated rules provide a probabilistic framework for musical parameters and sound devices to fluctuate. For instance, a higher temperature means a higher possibility of getting an unexpected musical response. As Figure 5.3 indicates, after the user requests that the system generates a new result, the system reorganises the data rules into a single list. Then, it sends it to the next stage via OSC.

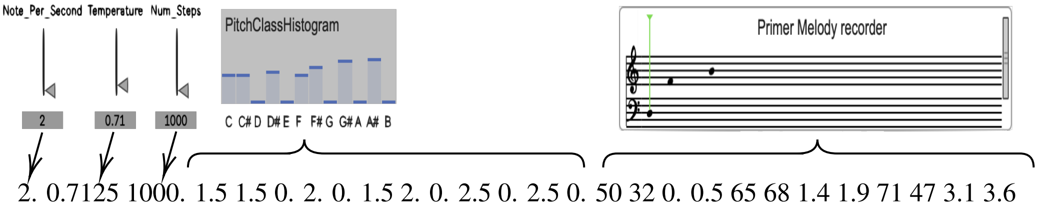


Figure 5.3: *Performer RNN* Generating rules example

2. The second stage makes use of the Python programming environment. After this block receives the data rules, it triggers the generated sequence, in three steps:

    (a) The system changes the generated option based on the received data rules.

    (b) The activation of the on-hold model, as Figure 5.4 shows, applies a sliding-window algorithm to produce multiple results based on the given number of steps.

    (c) Once the generative process is finished, the model goes back to the on-hold state, and the system loops through the results to send each note back to the Max part via OSC.

[1] Temperature is a hyperparameter involved in logits that affect the final probabilities from softmax.

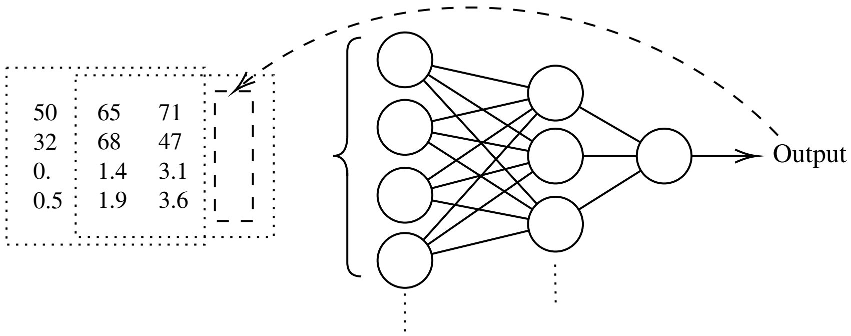


Figure 5.4: *Performer RNN* Generating process

3. The third stage returns to the Max block. In order to display the results in proportional notation, the system filters some repeated notes and recalculates the received results within the Bach notation library. Then, the notes are sent to the Bach interface object to articulate music notation. As a result, the user can directly read the score in the interface (Figure 5.1 right side). When required, the system can also play back the result and output standard MIDI notes based on a set speed.

5.4 Composition structure

Sound | *Figuration* lasts about 12 minutes in total; as Figure 5.5 shows, it can be divided into five sound sections and three visual scenes based on a basic motive and sound material. Each section of the piano score is based on the source material generated using specific rules via the AI-aided compositional tool, while various live signal processing modules extend the live piano performance capabilities. Live sound also drives the visual algorithm to generate and evolve the visual elements as the piece progresses.

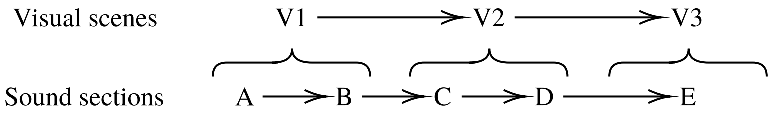


Figure 5.5: *Sound* | *Figuration* Composition Structure

5.4.1 Musical intentions

This work aims to enhance the range of compositional approaches through the construction of human–AI hybrid creative practices. An AI-aided tool can enrich the possibilities of developing musical materials and allow the composer to handle musical tension created as a result using higher-level hyperparameters. The first approach is the alteration of rhythmic densities. For example, the first line of the score is the outcome that the composer provides, consisting of a main motive and pitch classes. It uses the AI tool to extend the idea using lower rhythmic density and temperature. In contrast, the second line is developed using the same rules but with a higher rhythmic density. As a result, the musical material gradually increases the rhythmic density to create changes in musical tension. Another approach is to adjust the temperature to control the tension of the musical material. Take section B on the third page as an example. It starts with the material with a relatively high rhythmic density and lower temperature but without a clear motive. Then, it gradually develops into a higher temperature material and, therefore, has a higher chance of returning unpredictable musical responses. This creates a fluid movement, starting with a fast arpeggio constrained within a small pitch range. Then, the range expands and appears as non-harmonic tones. The final result forms a rich rhythmic combination between the different voice elements, which enhances the timbre and increases the tension of the harmonic materials.

This composition applies an intelligent system to enhance musical expression in order to extend the realm of acoustic instruments. For example, the system applies different signal processing methods to specific frequency ranges at 0:50, resulting in a multilayer timbral continuum. Another example can be found at 4:23, which applies different spatial movements to the more robust notes being detected, in order to reinforce the pianist's performance and to enrich the rhythmic element.

From the visual perspective, the real-time graphics aim to expand the artistic endeavours of this composition and helps the listener to navigate the musical narrative. Take the first sound section as an example. The visual part starts from a state of nihility; then, it uses a generative algorithm and sound as the driving force to gradually construct a complex structure. It provides a concrete image to illustrate the relationship found in the interactivity across media. Moreover, as a mirror of the piece's sound world, the visual content aims to gradually reveal its musical identity and form.

5.5 Interactive multimedia performance system

The interactive multimedia performance system in *Sound | Figuration* is a combination of two real-time subsystems communicating via OSC and Syphon:

The system recalls prestored event collections to match against a set of live piano sounds during the performance, since it does not expect the playback of any prerecorded music or video fragments. It not only applies transformations to live audio signals to produce variants as the response, but it also records piano fragments as the source material to feed the algorithm and produce a complete musical output. Therefore, the system is capable of recreating audiovisual environments that are slightly different to alternative live interactions emerging from each performance. In summary, the interactive multimedia performance system for *Sound | Figuration* is a score-driven program and instrument paradigm system that utilises a combination of transformative and generative algorithms as the response methods to generate meaningful musical outcomes.

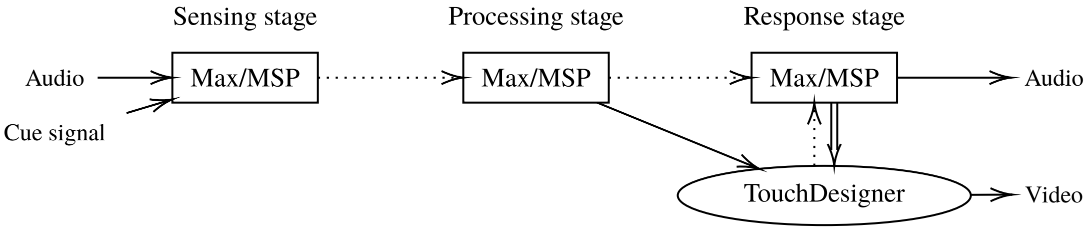


Figure 5.6: *Sound* | *Figuration* Interactive multimedia performance system structure

5.5.1 Sensing stage

During the sensing stage, the system collects two types of data, as Figure 5.7 shows:

1. The live piano sound, which is captured via a microphone(s).
2. The cue signal, for the system to recall prestored event collections.

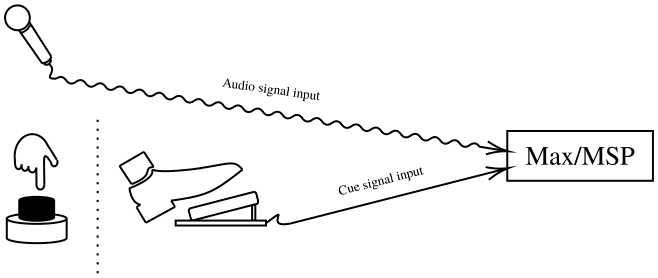


Figure 5.7: *Sound* | *Figuration* Sensing stage signal flow

In order to facilitate performance flexibility, the cue signal can be triggered via two different methods:

1. In the first method, the second person, i.e., a technical assistant sitting by the pianist, can trigger the cue signal. The benefit of this method is that the pianist can focus on performing the instrument, while the assistant deals with the score matching system in order to trigger the cue signal. The technician is also responsible for observing the system's statistics and for adjusting them during the live performance, aiming to produce the best possible result. This method of cooperation is comparable to a musical duet, requiring the pianist and the assistant engineer to rehearse multiple times to find the ideal timing.
2. The second method allows the pianist to trigger the cue signal without any technical assistance, using a MIDI pedal while performing. It provides the pianist with complete freedom for improvisation. This method holds the current interactive environment and only progresses to the next cue section once the performer has pressed the MIDI pedal, which may affect the musical flow. Thus, the solo method requires the pianist to be familiar with a number of responses and timing of events, in order to cue the system at appropriate moments.

5.5.2 Processing stage

As Figure 5.8 indicates, the system interprets the incoming raw audio signal to produce specific music information parameters and spectrum data and to recall prestored events during the processing stage. After this, data is prepared and passed on to the next stage.

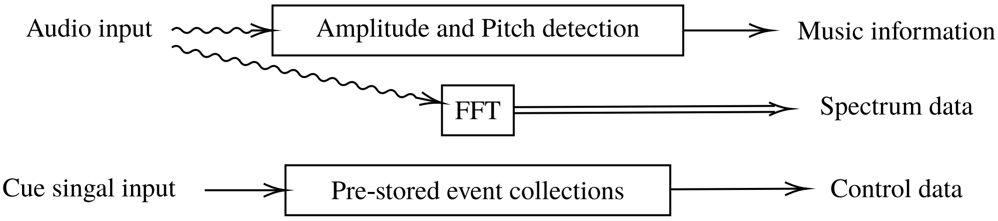


Figure 5.8: *Sound* | *Figuration* Processing stage signal flow

1. In order to obtain relevant real-time audio information for each input note, including the pitch, attack, velocity and interval relationship, the system filters the noise band of the input signal and subsequently checks the velocity peak using a fixed timing interval. Next, it judges whether a new attack has appeared by comparing the change in the direction of the peak velocity. Analysing the pitch at the moment of attack helps the system to obtain more accurate results. Lastly, it identifies intervallic relationships via calculating the absolute value of the relative pitch change.

2. The system processes the input signal using Fast Fourier Transform [61] with a fixed window size, which produces the required spectrum data every 33 milliseconds. Each frame of the spectrum data contains 2048 frequency bins of the data amplitude. Then, it stores the audio spectrum in Max's jit.matrix object and shares this large-scale datablock with the next stage via the Syphon interface.

3. The cue signal sequentially triggers Max's Qlist object to read the prestored collection of messages inside each text file. The messages include all the necessary control data for the response stage, including signal processing control parameters. During the next stage, the trigger data and the control parameters for the visualisation are sent to TouchDesinger via OSC.

5.5.3 Response stage

As Figure 5.9 shows, two subsystems produce corresponding visual and music outputs once the signal flow reaches the response stage.

The audio subsystem realised in Max/MSP contains the signal processing modules (Table 5.1), the audio multidimensional matrix for routing and the 8-channel ambisonic system. The whole signal process can be divided into three steps, as Figure 5.10 indicates:

1. The audio subsystem adjusts the matrix to change the input signal's ratio based on the control messages from the previous stage. For example, the comb filter input could be a mix of the live audio input and the output of other modules.

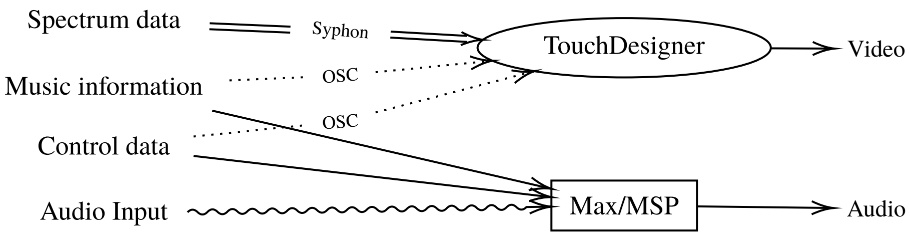


Figure 5.9: *Sound* | *Figuration* Response stage signal flow

2. Each module applies the transformation to the incoming signal according to control parameters received in advance, in order to produce variations in the input signal.

3. The multichannel ambisonic system integrates IRCAM's Spat library [62]. Each output module can be placed in any position in the virtual space and then be mapped onto the speaker configuration on stage. Finally, because of the recursive structure of the multiple module's input and output, the system is provided with infinite alternative results, making each performance even more unique.

Table 5.1: *Sound* | *Figuration* audio signal processing modules

Name	Function Description
Vibrato	Apply frequency modulation to the delay thread of the audio input to create a vibrato effect.
Comb filter	Apply a polyphonic comb filter effect to the input audio signal.
Freeze	Audio capture and reconstruction through real-time FFT spectrum analysis.
Granular FM	Real-time audio granular synthesizer. The audio granular playback speed is based on a reference decimal (micro) pitch table.
Freeze reverb	A reverb effect taken from the implementation by Juhana Sadeharju.
Loop sampling	Apply loop recording and forward/backward playback to extend the input audio.
Granulator	Apply multiple playbacks of the audio recording to create a granular cloud effect.
Spectral delay	Combine the spectral filter and feedback to transform the input audio.
Stretch effect	Apply real-time loop recording to the audio input. Then, play back the recording through multiple layers with variable speeds.

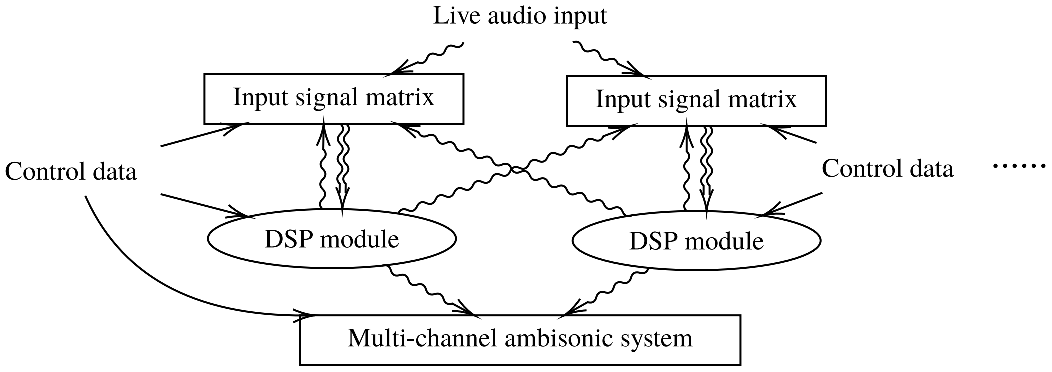


Figure 5.10: *Sound* | *Figuration* Audio subsystem structure and signal flow

The visual subsystem driven by TouchDesigner applies a different system structure to the audio subsystem, and it is controlled by messages sent from Max/MSP via OSC. Again, this holds a similar three-step processing chain structure, as seen in Figure 5.11:

1. Each visual scene has an individual processing container based on the incoming signal generated by the visual material. For instance, the $V1$ scene receives parametric information via OSC to handle the unfolding of geometrical node networks; the $V3$ scene receives the audio spectrum via Syphon in order to drive the internal motion of the point cloud.
2. It is possible to apply different image augmentation algorithms to the source image to produce highly varied visually stunning results, for example, a blurring effect at the edge of the image or the adjustment of the brightness, gamma and contrast to match each output image.
3. Adjusting the transparency of each container output produces the final composited image, achieving a natural transition between each visual scene and the final output image on the screen.

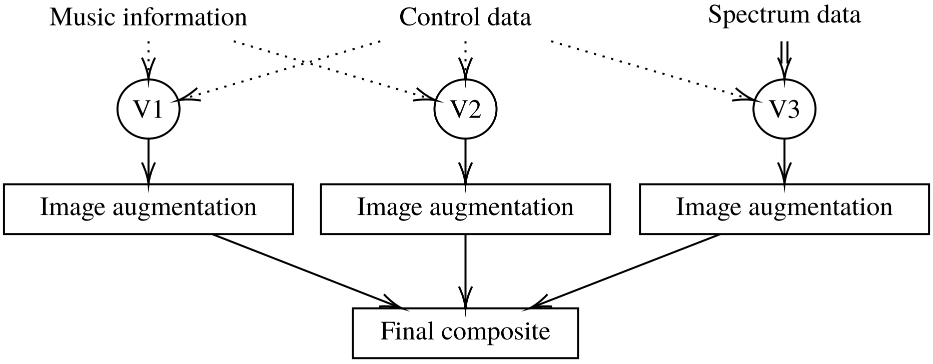


Figure 5.11: *Sound* | *Figuration* Video subsystem structure and signal flow

Chapter 6

Without strings

6.1 Description

Without strings (无弦) is an interactive audiovisual installation driven by real-time environmental data. It is inspired by traditional Chinese ink-wash painting and Guqin music. It emphasises the "hidden" and "blank" visual features from the ink-wash painting, and applies Guqin improvisation by feeling the surrounding space. At the core of the visual element is a virtual tree that grows by absorbing environmental data. The sound aspect is represented by a physical model of the Guqin traditional musical instrument in China, which is driven by a pre-trained AI model in real-time. The Guqin synthesizer is a polyphonic physical model based on the IRCAM's Modalys library [63]. The combination of audiovisual content creates an artistic reflection of how machines can also "sense the world", while providing an immersive experience to the audience.

6.2 High-level questions: Can machines feel? Artistic goals, format and the Guqin

How can machines sense the world? Can a machine musically express such a "feeling"? What forms of artistic practice can emerge from a novel interactive multimedia performance system exploring machine musicianship without human involvement? By exploring artificial intelligence in music, the composer provides a pathway to the listener in order to advance on the above research questions. Adopting an audiovisual installation format for this work was considered a potentially suitable art form and context, since it often does not need human interaction at all. The Guqin: As for the use of this musical instrument, also known as Qin (Figure 6.1[1]) by Chinese Scholars, it is considered to be the first amongst the Four Arts[2]. The Guqin is one of the most respected musical instruments in ancient Chinese culture [64], through which many musicians and repertoire express their emotions about the environment. Metaphorically, the instrument is used to recreate the artistic aural surroundings of the composer's feelings. Another indication of its important cultural heritage is *Flowing Water* [3]. Such recording by Guan Pinghu aims to represent a vivid image of the

[1] Muziekinstrumentenmuseum, Europe - CC BY-NC-SA. `https://www.europeana.eu/en/item/09102/_RMAH_109010_NL`

[2] 四艺, Qin (琴, Guqin), Qi (棋, Chess of Go), Shu (书, Calligraphy), Hua (画, Painting)

[3] 流水, a Guqin piece composed around 220 BC by Yu Boya was included in the Voyager Golden Record and sent into outer space in 1977.

water flowing through mountains and woods. Traditional Guqin music is typically transcribed in Wenzipu (word tablature) [65]. The notation only records pitch and fingering without indicating the note value, tempo or rhythm. Therefore, playing the Guqin involves substantial improvisation with the given materials. The improvisation part is based on the musician's understanding of the piece, and it aims to reflect the musician's deeper levels of musicianship. From the compositional perspective, the unique tone structure of the Guqin and its particular musical scale was an ideal source material and a form of inspiration to reveal the complexity of the system design and influences behind the piece.

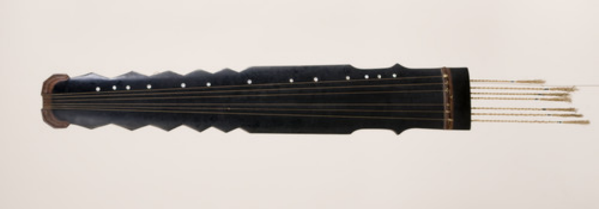


Figure 6.1: Kin, Qin, Guqin, Ch'in

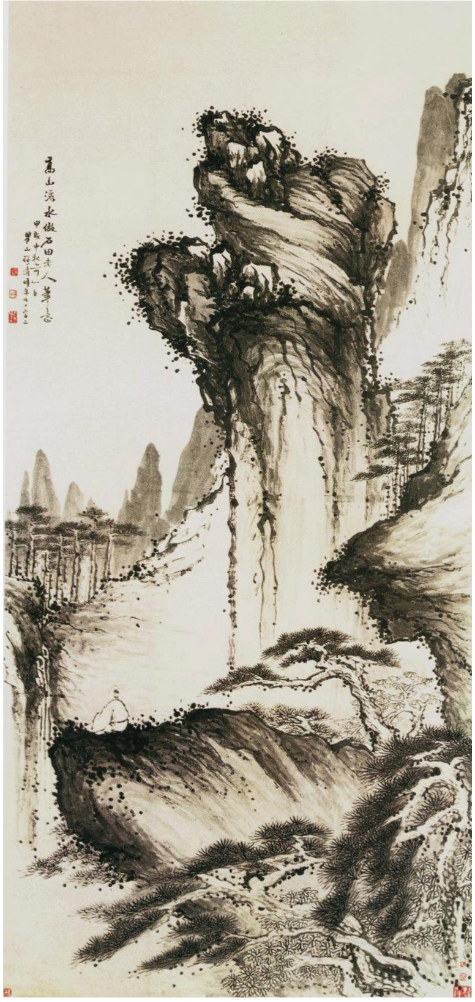


Figure 6.2: "High Mountains and Flowing Water", By Mei Qing, Qing Dynasty, Beijing Palace Museum, China

Another strong source of inspiration comes from another one of the Four Arts, Hua, traditional Chinese ink-wash paintings [66], and it is used to drive the design of the visual element. In particular, it was informed by the Shan Shui style of landscape painting, as seen in Chinese ink-wash art [67]. Take Figure 6.2 as an example; this Shan Shui style of landscape painting is typically monochrome, and it uses only black shades to depict very loose recreations of mountain and water landscapes. However, its unique style also informs back and forth the sculpting of sound and visuals; for instance, a mountain's silhouette can be affected by the shape of sound and the overall colour of the image. Trees are often leading elements in Shan Shui style painting, and as a result, a virtual tree is placed in the staged scenery. This virtual tree serves as a visual reflection of the system sensing the environment, which provides the trajectory to the environmental data. The remaining white space between the mountains in the distant scenery and the trees in close view conveys the perceived "spirit" of the flowing water and the air [68].

In summary, this installation utilises AI to drive the Guqin performance by sensing the environmental data, and it does so, not only to create a combination of traditional Chinese art and cutting-edge technology, but also to provide new routes for the expression of the system's musicianship.

6.3 Installation structure: Data mapping

Without strings is a real-time audiovisual and evolving installation, as it can complete an entire growth cycle in 24 hours. The sound element holds two distinctive layers:

1. The leading sound source material: The Guqin sound is produced from the physical model synthesiser, which is driven by an AI algorithm affected by the environmental data in real-time. It does so via a real-time implementation of Performer RNN (Performer RNN, AI-aided compositional tool), including two features:

   (a) Performer RNN applies a pre-trained model, which is trained in 61368 steps with the Guqin dataset [69], in order to generate realistic Guqin music information. The dataset contains 71 Guqin pieces and 408 phrases, with 9860 measures in total. The generated results achieve about 0.98 in accuracy and around 1.0 in perplexity[4].

   (b) The environmental data affect the generating rules of Performer RNN, including the note density and temperature. For example, the higher the temperature readings, the greater the rhythmic density and complexity.

2. The background sound source material: Ambient sound is produced by combining several sound processing modules. It applies several signal processes to the source material, and it is directly controlled by fluctuations in the environmental data. For example, humidity and air pressure affect the module's feedback rate by controlling the reverb and the attenuation time. It aims to directly reflect environmental changes and enhance the sense of space to create an immersive experience for the audience.

[4] Perplexity is a metric to evaluate the model, and is used to measure the probability distribution. [70]

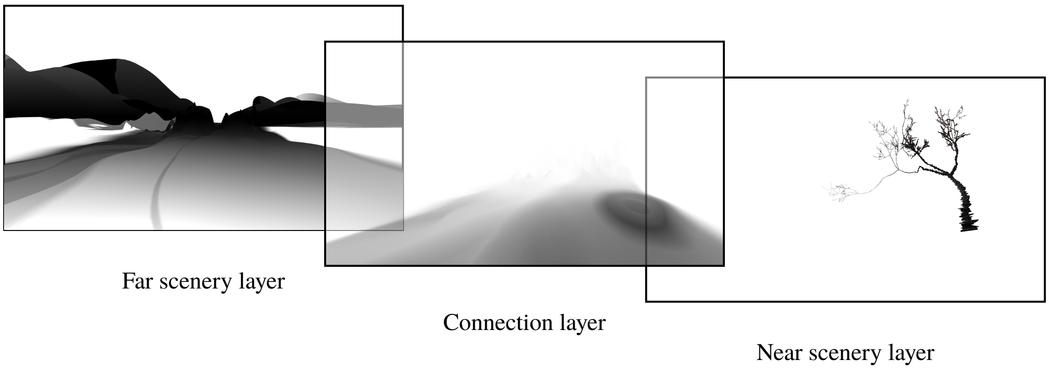


Figure 6.3: *Without strings* three layers of the visualisation

As Figure 6.3 shows, the moving image is constructed of three layers:

1. The near scenery layer: This is where the main object tree is placed. The tree is a 3D model generated via Equation 6.1, which implements the L-system in TouchDesigner [71][72]. The L-system is a parallel rewriting system and a type of formal grammar used to describe the behaviour of plant cells and to model their growth processes [73].

$$\begin{aligned} context_ignore &: F + - \\ premise &: A \\ A &= ";T \sim (c)F[+(d) - (d)A]/[-(d - 20)A]F/A \\ F &= \sim (-b)!F//F \end{aligned} \tag{6.1}$$

   In order to achieve real-time growth, the system binds the L-system's rules to real-world elapsed time. The virtual tree completes an entire growth cycle in 24 hours, as Figure 6.4 indicates. The system maps the real-time environmental data to variables b, c and d in Equation 6.1 to reflect real-world changes in the environment.

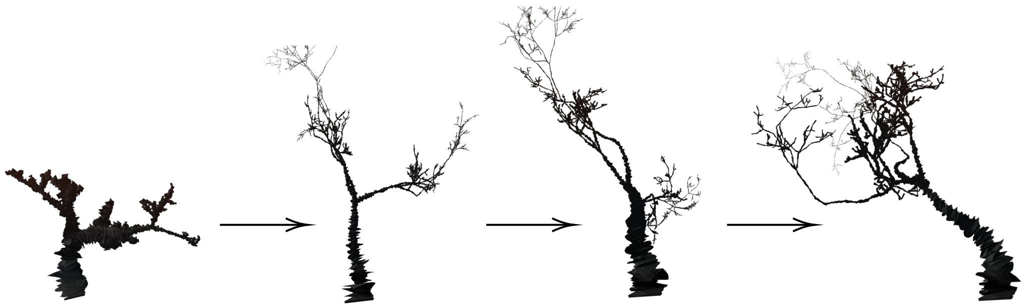


Figure 6.4: *Without strings* different phases of the virtual tree

2. The far scenery layer: This is the main sound-to-visual representation. Here, the system collects spectrum data from the audio output through real-time spectral analysis. Then, the system applies it to the height data to reshape the 3D geometry grid, in order to produce a mountain silhouette-like shape. While this happens, the system applies its dynamical texture to the 3D geometry, using texturised spectral data in a feedback loop. As a result, the sound gradually shapes the mountain's silhouette and

builds a scenery of the water flowing down from the top. Simultaneously, the elapsed real-world time is mapped onto fluctuations in the visual texture to add a sense of time passing.

3. The connecting layer: This is represented by the "water", and it is placed in the middle of the virtual space to establish a dynamic connection between distant and close scenery layers. The connection layer utilises a geometry grid, which has a dynamic texture that is a rippled image associated with the spectral data of the Guqin sound. Such rippled image emerges from a circle geometry, which evolves according to real-time spectral data extracted from the Guqin-synthesised sound. As a result, the centre of the rippled image overlaps with the virtual tree to reach the distant scenery layer.

As Figure 6.5 shows, the blend of all visual layers provides a mountain and water scene and a window to display an imaginary sound world, which strongly informs the music.

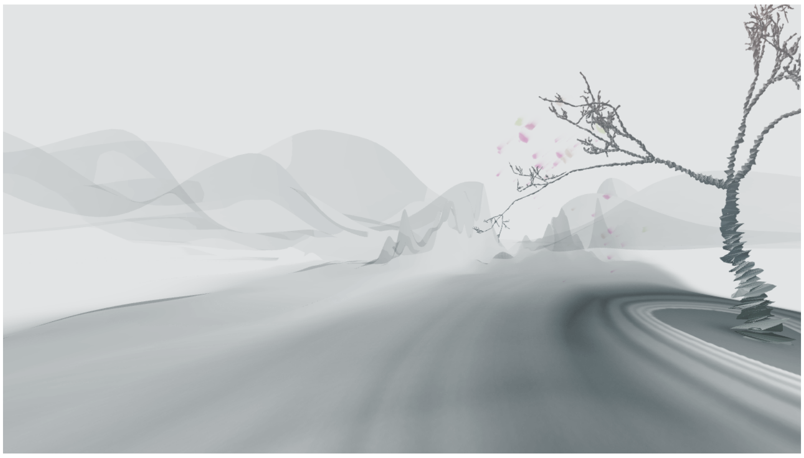


Figure 6.5: *Without strings* Visualisation example shot

6.4 Interactive multimedia performance system

As Figure 6.6 indicates, the *Without strings* interactive multimedia performance system does not require active human intervention during the audiovisual generative process. The system automatically collects real-time environmental data and applies it to generate responses accordingly. Such a system can be categorised as a performance-driven program using generative response methods and an instrumental paradigm [74].

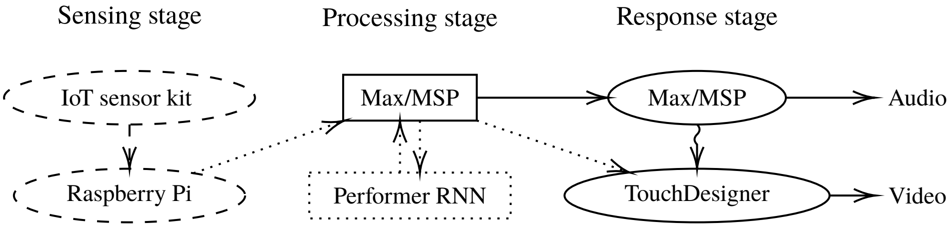


Figure 6.6: *Without strings* Interactive multimedia performance system structure

6.4.1 Sensing stage and music mapping

During the sensing stage, the system collects the real-time environmental data via the IoT sensor kit, the Nordic Thingy:52 [75], which is a compact prototyping platform with motion- and environment-related data acquisition. The built-in environmental monitoring sensors can capture temperature, humidity, air pressure, CO2 levels, TVOCs (Total Volatile Organic Compounds), light and colour intensity. From the compositional viewpoint, a key aspect of this work is the mapping between musical devices (phrasing, textures, and micro-structures) and the dataset, especially using a pre-trained AI model of the musical instrument. But, it also uses a layering system to project aural representations of the environment in spatial form (close and distant space and in between).

The Thingy:52 is connected to the host device via Bluetooth 5 standard [76], which has a relatively limited connection distance range of up to 10 meters long. To overcome this difficulty, the installation uses a combination of the Raspberry Pi [77] and the Thingy:52 in order to extend the distance range, since the Raspberry Pi has both Bluetooth and Ethernet connection capabilities.

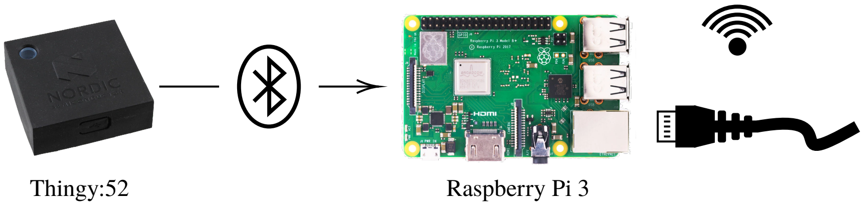


Figure 6.7: *Without strings* Sensing stage signal flow

To achieve the above-mentioned method, as Figure 6.7 indicates, the sensing stage involves three concatenated steps:

1. The Nordic Thingy:52 sensor kit collects data.
2. The Raspberry Pi reads the sensor values over Bluetooth.
3. The Raspberry Pi converts the sensor reading values to the OSC (OpenSoundControl) protocol-based message, then transfers it via Ethernet connection to the host system.

As a result, the Thingy:52 and Raspberry Pi can be installed outside the venue to collect the outdoor environment-related data, and the host system can access this data via wireless or cable connection.

6.4.2 Processing stage

As the first step of the processing stage, the collected environmental data is displayed on the graphical user interface (Figure 6.8) developed in the Max/MSP programming environment.

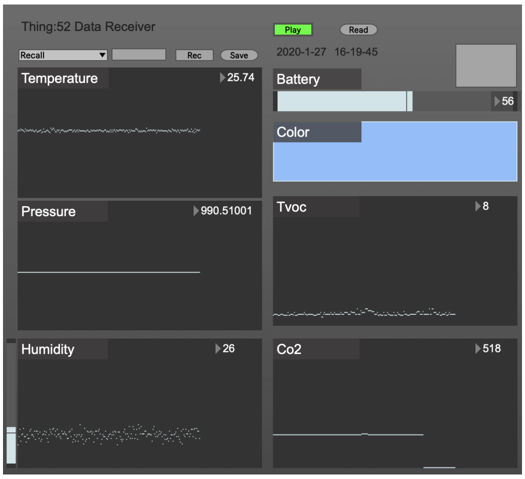


Figure 6.8: Thingy:52 Max/MSP GUI

As Figure 6.9 indicates, to generate meaningful musical responses from the pre-trained model, the data is required to convert and re-scale to an optimal range. For instance, humidity, air pressure, and CO2 levels must be mapped to control the correct range of musical pitches.

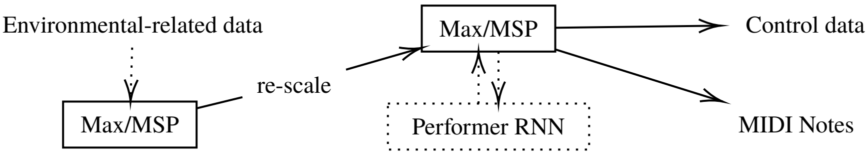


Figure 6.9: *Without strings* Processing stage signal flow

The system uses two buffers to achieve parallel music retrieval from the AI and playback with a few hundred milliseconds of precision. For example, to generate ten seconds of responses with around two notes per second takes about two seconds of computation. The system can control the playback speed independently, with both buffers switching their states based on each other's performance. Figure 6.10 shows how the process can be divided into four steps to gain a better understanding:

1. The first buffer starts playing.
2. The second buffer requires the model to generate new information and store it inside, remaining on standby for playback.
3. Once the first buffer has finished playback, the second buffer starts.
4. At the same time, the first buffer clears the allocated memory and requires the model to generate new information to store it inside whilst waiting to be called for playback.

In the final step of the processing stage, the music playback is converted into standard MIDI notes, and the re-scaled environmental data is ready to be passed on to the next stage.

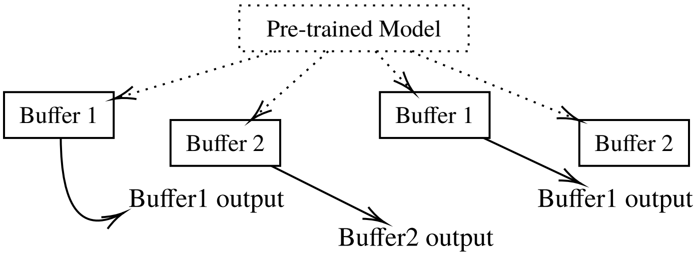


Figure 6.10: Performer RNN the implementation of linear sequence reaction

6.4.3 Response stage

As Figure 6.11 shows, the response stage consists of two parts: the audio part (realised in Max) and the visual part (in TouchDesigner).

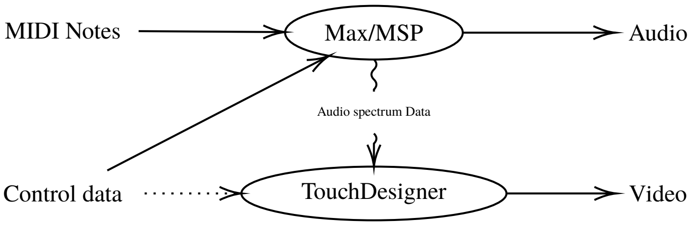


Figure 6.11: *Without strings* Response stage signal flow

Figure 6.12 shows how the audio pipeline works:

1. Once the Guqin synthesizer receives the notes from the previous stage, it produces the source sound materials informed by the data. The traditional Guqin uses silkworm silk strings to produce its unique sound. To simulate this musical colour, the physical model applies similar material properties (Table 6.1) to five mono-string engines. It also utilises different lengths and filters for each mono-string engine to simulate their aural nuances. As a result, the model can produce sound material close to the original Guqin plucking sound in real-time, providing the composer with a powerful tool to embrace the mapping of the dataset.

Table 6.1: Material properties of silkworm silk string

Length	**Density**	**Young**	**Poisson**
$1.12 - 1.18m$	$1300 - 1380kg/m^3$	$5.0e9 - 6.0e9N/m^2$	0.4

This part takes the sound material from the physical model synthesizer and implements transformations to produce musical variations involving spectral content, sound gestures, dynamics and timbral changes in the original instrument. It extends the Guqin sound material to create an immersed sonic experience, sculpting space into it. To do so, it makes use of control parameters for further signal processing (Table 6.2), which is driven by the environmental dataset, evoking the aforementioned aspects of spatial distance (from close/intimate to peripheral and back). This is achieved by adjusting each audio signal to a specific spatial position, which can be resolved into different speaker configurations. In this case, all the audio signals are sent to a multichannel

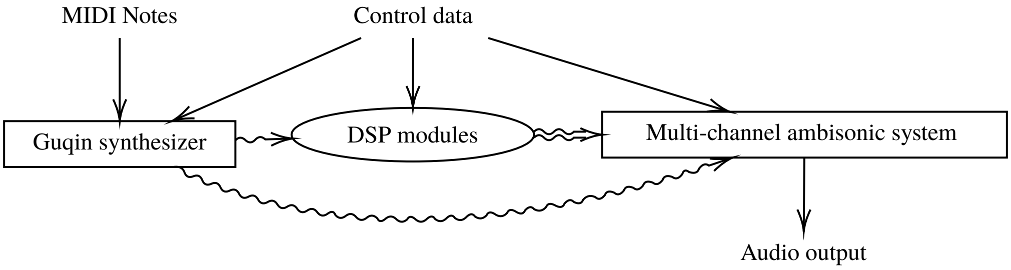


Figure 6.12: *Without strings* response stage audio part signal flow

ambisonic system, which integrates the Spat library [62]; this sends the final output to the multichannel system, which is adaptable to the needs of the gallery or installation space.

Table 6.2: *Without strings* audio signal processing modules

Name	Function Description
Freeze	Audio capture and reconstruction through real-time FFT spectrum analysis.
Stutter granular playback	Real-time audio granular synthesizer. The audio granular playback speed is based on a reference decimal (micro) pitch list.
Freeze reverb	A reverb effect taken from the implementation by Juhana Sadeharju.
Loop sampling	Apply loop recording and forward/backward playback to extend the input audio.

As for the visual part seen in Figure 6.13, it takes two steps to build the final dynamic image. Firstly, based on the received data to produce the source image, the visual system receives the spectral data via Syphon [78] and the environmental data via OSC, both within Max. Then, the spectrum data is mapped onto the height data of the 3D geometry, in order to reshape the silhouette-like shape of the mountain. It applies the environment-related data to the L-system variables to affect the tree growth. Secondly, the system applies visual augmentation to the source image to generate the desired results at the intersections of music and visuals. For instance, the colour of the original image is replaced by a poly-chrome image based on a dark–azure–light ramp texture. This is not only to match the Shan Shui style colour via a colour lookup function, but also to obtain specific sonorities connected to this change. Another example is to apply very slight noise to the bright part of the image to increase the sense of paper texture, which again, is reflected in the audio part.

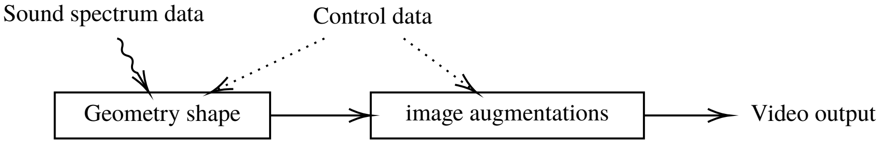


Figure 6.13: *Without strings* Response stage visual part signal flow

Chapter 7

Metamorphosis

7.1 Description

Metamorphosis (蜕变) is a real-time interactive audiovisual composition for one human performer and two AI (artificial intelligence) performers. Performers make use of a virtual ancient Chinese percussion instrument called the Bianqing (磬) to learn and imitate one another's musical expressions. The composition presents scenarios of both confrontation and cooperation to explore alternative performance situations often seen on stage in the interaction between human performers. The musical direction is parallel to the shape and sound of the instrument, which gradually evolves as the piece progresses and among the AI performers themselves, which also morph throughout. From ancient to modern and from concrete to abstract, the piece creates an immersive experience, as a metaphor for exploring the complex co-evolution of humans and machines.

7.2 Inspiration and artistic goal

In previous works, such as *Handwriting · WuXing* (Chapter 2) and *Intangible Field* (Chapter 3), the infusion of machine learning into the composition not only provides the possibility to realise complex compositional concepts, but also demonstrates the potential of novel machine musicianship. The first AI-aided composition tool used in the portfolio, Performer RNN, provided room for hybrid creativity to facilitate the composition process in *Sound | Figuration* (Chapter 5). Then, in *Without strings* (Chapter 6), Performer RNN was further developed to work independently in real time, unveiling a number of unique and unexpected AI musicianship features. In this piece, the new goal was to integrate this state of the art technology into one interactive multimedia composition involving a higher complexity of human–machine relationships on stage. To achieve this goal, a strong source of inspiration comes from George E. Lew's computer music composition, *Voyager* [79]. It not only gives the direction of this piece's fundamental paradigm, which is multiple improvisers' performance in real-time, but its software-based music system has influenced the composer to investigate the model-based agent as the primary structure for the AI performer's system.

As Figure 7.1 indicates, a general model-based reflex agent [52] is used to sense the environmental state to rationally drive the actuators based on the internal AI model and the given condition-action rules. This basically allows the AI performer to make sense of what is heard and to compose convincing responses based on a pre-trained model and certain conditional rules. This provides the means to drive the instrument in order to perform expressive musical materials [80]. As Figure 7.1 shows, it has a similar three-stage system structure to previously discussed interactive music systems. The sensing stage collects the current state; the processing stage chooses the action; and the response stage executes it.

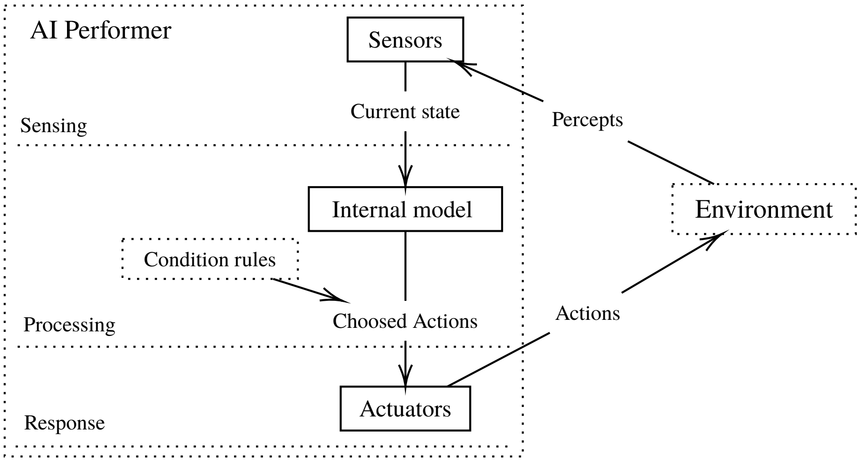


Figure 7.1: AI Performer conceptual structure[81]

While exploring the application of suitable emerging technologies for this new work, two interesting concepts emerged, which later became a primary inspiration for the composition:

- What would it be like to design an AI performer that does not have human-like limitations? For instance, an AI performer that could directly listen and understand datasets and drive a virtual music instrument as a result. An AI performer whose sensors and actuators do not restrict the potential for interaction.
- How difficult would it be to challenge current AI models that can respond to data and sound in real-time? Where are the existing technological and creative barriers? By the nature of the model-based agent and the limitation of currently available technology, the AI performer's internal model is a pre-trained long short-term memory-based recurrent neural network [58]. So, the two problems to tackle are:

  1. The network's over-reliance on the training dataset leads to a lack of generalisation, which gradually discards the sense of long-term structure in the composition [82].
  2. Each generated process and step takes a few milliseconds, and on occasions when the information is vast, it does not work as fast as expected in real-time question–answer musical scenarios you may find when two human performers play or improvise together.

In an attempt to address the above musical challenges, the composer adopted three compositional solutions to address these challenges:

1. With regard to human interaction and prestored events to handle the larger-scale structure in the composition, the composer's strategy was to apply a similar control mechanism for the structure employed in *Audio Game - Music Force* (Chapter 4). In this piece, the system switches between fully interactive improvised sections and semi-interactive cutscene sections. These two combine different modes of human and AI interaction, which can be further developed to create structural meaning in large-scale works.

2. With regard to the limitations of speed in the responses, as Figure 7.2 indicates, the method aims to think outside the box, and to supply the necessary time for the generation process using musical means, rather than the technological enforcement of a problem that is quasi-impossible to solve. This is explored in the selected type and number of performers: a trio including one human performer and two AIs on stage. For instance, the human performer starts the piece with a musical motive as the question, and then, the AI performer develops it. After that, the third performer provides a different answer based on the second performer's response. Then, the first performer is asked to answer the musical variation from the second AI, and so on. As a result, this triple musical dialogue not only provides enough time for each block of data to be processed and deployed musically, but also demonstrates that, human to machine, machine to machine and machine to human interaction can create a feedback loop of multidimensional relationships on stage.

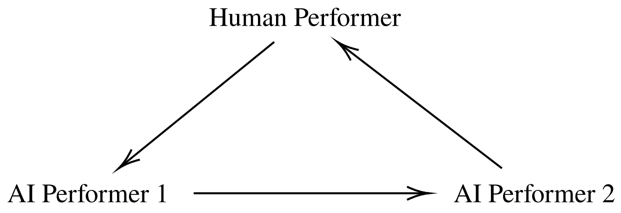


Figure 7.2: The human performer and AI performers trio relationship

3. Finally, to combine the two above items, common human–AI instrumental references are found. In order to achieve coherence between both human and AI performers and musical expression, the traditional Chinese percussion instrument Bianqing (编磬 Figure 7.3) [83] was chosen as the essential reference for the virtual musical instrument. The pitch range of Bianqing is up to three octaves, its bass sound is thick and the treble tone projects itself incredibly well. Its timbre is enchanting and relatively achievable via physical modelling. The percussion striking action can be simulated and detected via a motion sensor, such as a Nintendo Wii Joy-Con controller. The model is trained on a much larger piano performance dataset [84]. It generates MIDI-like music information, which can be passed on to the Bianqing for musical expression.

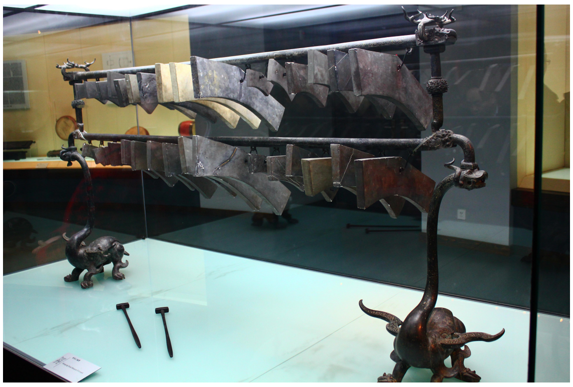


Figure 7.3: Bianqing from Marquis Yi's tomb, Hubei Provincial Museum

7.3 Composition structure

Metamorphosis lasts about 21 minutes; as Figure 7.4 shows, it applies a non-peer structure for the visual and sound elements. It consists of five music sections and three visual scenes. Each aural section has its own compositional aims, or creative themes, which are used to determine the interaction models between the musical performers. For example, in the first section, a learning method starts and each performer tries to mimic one another. In this part, the compositional strategy is to design a musical path to control the evolution of musical tension, while instruments work separately. However, in the second section, starting at 2:30, the theme/aim is cooperation. In this cooperative mode, the interaction model changes from a solo to a choral performance. Musically, the objective is to enrich the rhythm and harmonicity environment, and to build up a conceptual foundation for more complex relationships that appear as the large-scale structure unfolds. Digital signal processing plays an essential role in supporting musical expression across different creative themes, especially when it comes to transforming the timbre and texture of sonic materials emerging from the physical model. In the third section, starting at 8:29, the processing of aural materials strengthens the sense of dissonance by gradually alienating the timbre of the instruments. As a result, the musical conflict between the performers and their sound world becomes exacerbated.

From the visual stand point, the composition's layout also changes dynamically in response to the developed theme, aiming to enhance the musical expression of each section. For example, in the first section, a multiangle image is presented, in order to introduce every performer, while it enlarges the visual area of performers in action with the aim of guiding the listener's attention. Another example is at 12:06, which marks the beginning of the last section. The visual coats are transformed into a single-image layout, aiming to work as a metaphor of the musical fusion emerging from each performer's action.

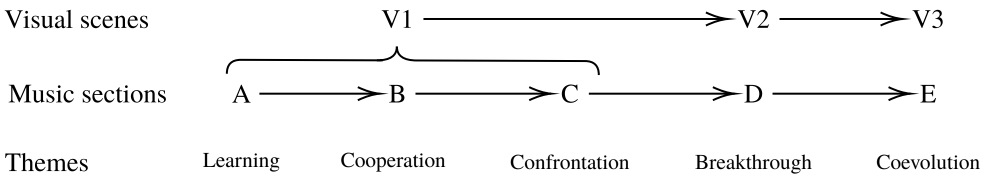


Figure 7.4: *Metamorphosis* Composition Structure

7.4 Interactive multimedia performance system

As Figure 7.5 indicates, the interactive multimedia performance system in *Metamorphosis* integrates a modified version of Performer RNN to achieve human and AI real-time interaction. It not only relies on the retrieval of any prestored musical device during the performance, but it is based on the performers' real-time input signal, which generates musical responses. However, as the composition progresses, the system relies on some predetermined event collections. For example, at 8:26, the human performer triggers a number of event changes in order to match the nature of music arriving at the system's input. Similarly, the system analyses the human performer's body gestures to guide an elaborated output exceeding the standard mode of response seen in previous models. For instance, at 13:10, the autonomy of the agents' cluster increases, and the human performer's motion gesture leads them to aggregate or disperse. Moreover, the system takes incoming sonic material and produces transformations and variations of it. A clearer example of this is at 16:37, where a sound is triggered by a striking body gesture of the human performer, producing a variation of the Bianqing sound, which is rearranged by the system to modify its texture by adding a granular system effect. Therefore, *Metamorphosis*'s interactive multimedia performance system is not only a hybrid system in terms of its dimensionality (score and performance-driven dimensions), but also in the way it deals with the instrument's and player's paradigms, and in the way, it applies transformative and responsive methods to a given input.

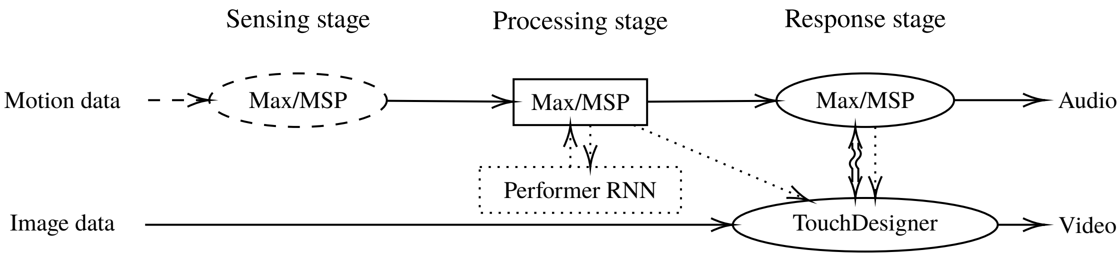


Figure 7.5: *Metamorphosis* Interactive multimedia performance system structure

7.4.1 Sensing stage

In order to design the simulation of the percussion instrument and its performance style, the system expects an off-the-shelf human–computer interface (HCI) solution to capture the motion of the human hands in real-time.

Among the solutions explored, the leap motion controller reported the best comprehensive hand tracking data, as previously seen in *Handwriting · WuXing* (Chapter 2). This, in combination with the PostNet machine learning model, allowed the use of a single RGB image to detect real-time human pose estimation using key points, as demonstrated in *Intangible Field* (Chapter 3). However, such a solution proved to have limited recognition space and requires specific lighting conditions on stage, which is often out of the control of the composers when performing at festivals, in order to provide an accurate image suitable for accurate tracking. After further investigation, the Nintendo Joy-Con controller[1] (Figure 7.6) was the preferred choice as the HCI solution. Among its advantages, it has two separate units for both hands and a similar holding method to drumsticks, which was perfect for the Bianqing. For example, the gesture at 0:08 enables the performer to mimic the striking sound gesture. It also allows the performer to use two hands to simultaneously control different elements of the system and instrument. For example, at minute 8:38, the performer has individual control of both the density and the spatial location of the sound, using different hands for each task. An added technical advantage is the fact that the Joy-Con has an accelerator, a gyroscope, buttons, and other sensors to track hand motion in real-time, and it transmits data wirelessly via Bluetooth [76]. This device enables dynamic gestures and provides a relatively large performance space to accommodate the capture of dynamic data from moving objects on stage. Finally, the interface can also control the visual element, as seen in 17:13 after the striking hand gesture, where the performer modifies the audiovisual element with their hand (which holds the interface).

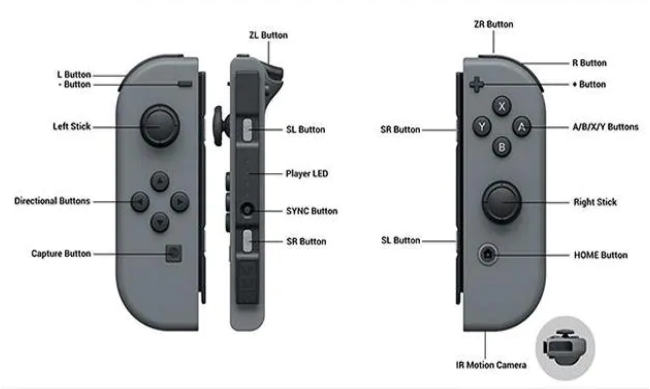


Figure 7.6: Joy-Con: Nintendo Switch video game console controller

Despite all the aforementioned advantages, the Joy-Con controller only operates in a simple human interface device (HID) mode, where the motion sensors are not enabled, although it pushes for updates every time a button is pressed. Further investigation led to setting an input report mode from the standard to full mode, in order to enable motion sensor technology. Technically, the user needs to send the sub-command $0x03$ with argument $x30$ to the Joy-Con after pairing with the computer [85], in order to add additional data with the potential for added nuances in terms of musical expression.

As Figure 7.7 indicates, during the sensing stage, the Joy-Con pushes the current state at $60Hz$ to the computer via Bluetooth to unlock further features.

[1] The primary game controllers for the Nintendo Switch video game console, which was released worldwide in most regions on March 3, 2017

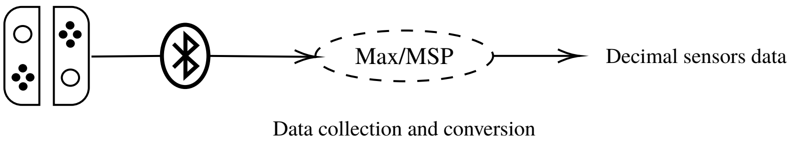


Figure 7.7: *Metamorphosis* Sensing stage data flow

On the visual side of things, in the last visual scene, Touchdesigner enables the webcam in order to capture the landmarks from the performer's body. This not only allows the immersion of the human performer into the virtual space, but it also becomes an essential strategy for mixing virtual and real motion interaction between performers. For example, at minute 18:17, the agents move closer to one another to embrace the projection of the human performer in the virtual world, which provides a unified visual and sound experience.

7.4.2 Processing stage

When the raw data from the sensors arrives at the processing stage, it operates both in pre- and post-processing modes. These two steps allow for the generation of derivative data, which is then passed on to the next stage.

As shown in Figure 7.8, the raw data is first filtered and re-scaled into a stable data stream. This helps communication with the controllers' pitch, yaw and roll data, which is re-scaled from its original range to -1 to 1. As a result, it provides a smooth change in data by limiting the amount of data passing by at a certain speed and by adding an interval value within the given time constraint.

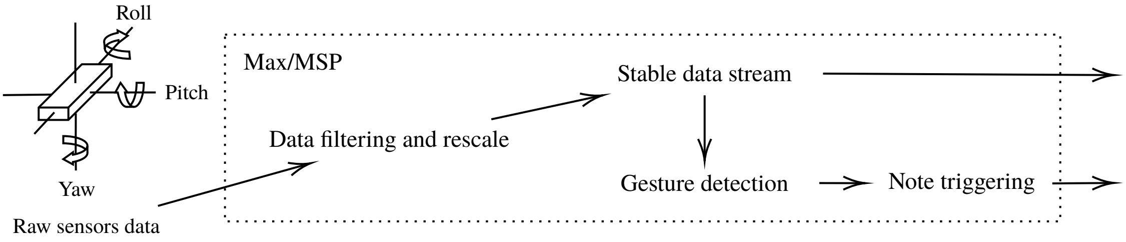


Figure 7.8: *Metamorphosis* Processing stage pre-processing step

The following step consists of applying the stable data to drive the virtual instrument. Here, the interaction mechanism uses the left-hand controller to strike a virtual stone-chime sound, while the right-hand controller simulates the natural playing gesture (strike) of the Bianqing. As indicated in Figure 7.9, the interaction process has two steps:

1. The system maps the left controller's yaw value to a one-dimensional slider to enable the controller's horizontal position to select the desired pitch.
2. During a short period, the system measures the changes in the right controller's pitch to obtain the difference value per unit of time. Then, it compares it with a threshold value to determine whether it is a rapid gesture or not.

This mechanism also provides room for additional gestural nuances, such as non-sudden gestures, to subtly enhance the control of musical parameters in sound, and even provide control over slow scrolling and panning gestures, as heard at 8:40.

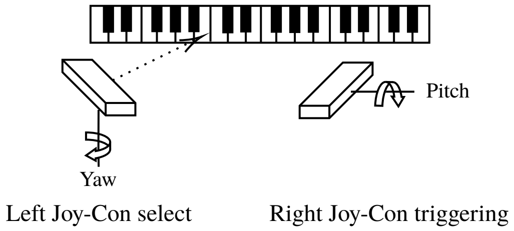


Figure 7.9: *Metamorphosis* “striking note” interactive mechanism

Once a known gesture is spotted by the system, it packs the data as a standard MIDI note [86]. Then, it passes the note on to the postprocessing block in order to create responses that can trigger both sound and visual materials.

The postprocessing step is where the human performer interacts with the AI performer. The latter consists of three primary modes of interaction with the human performer (Figure 7.10), which are also associated with the three skills emerging from a deserving machine with musical abilities: listening, composition, and performing, respectively.

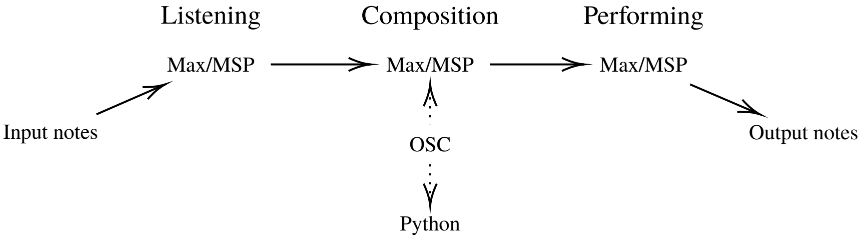


Figure 7.10: *Metamorphosis* “AI Performer” three primary modes

- The listening mode aims to record and analyse the incoming data information. The AI performer stores any incoming note in the main melody storage space once this mode is activated. Then, it outputs the analysed result, which includes the note density and pitch distribution [87], once the listening mode is deactivated.
- The composition mode connects Max with a Python script, and it applies three sequential steps, as shown in Figure 7.11:

  1. Once the composition mode is activated, the AI performer collects all the current conditions and rules, such as note density, pitch distribution, duration, number of steps and randomness temperature in the Max patch. Then, the system pushes all the generated information onto the Python script via the OSC protocol.
  2. Once the Python script receives the data, a generative loop is activated. Then, the pre-trained Recurrent Neural Network model produces the required music information based on the received data while looping. After that, and at the ending of the loop, such information is sent to the Max part via OSC.

3. Once the Max part receives all the resulting data, the system ties up the dataset by integrating the repeat notes that are too near and filtering notes that are out of range. Then, the system stores and displays the final result score, ready for the performing mode.

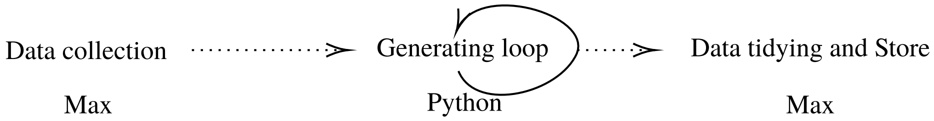


Figure 7.11: *Metamorphosis* "AI Performer" composition mode three steps

- The performing mode is where the system plays back the final result data. It holds the ability to change the playback speed dynamically during the performance, which is a method used by the composer to control rhythmic tension. For example, at the beginning of the piece, the playback speed is relatively slow, but it starts to increase from minute 05:40 onward using this method.

Finally, the resulting playback information is packaged into a standard MIDI note format and passed on to the response stage.

7.4.3 Response stage

As shown in Figure 7.12, the response stage consists of two parts, which are necessary to generate the final audiovisual response. All the predetermined instructions are stored in the Max part, waiting to be sequentially called back to re-create the interaction environment in a meaningful way as the composition progresses. Additionally, both Max and TouchDesigner communicate small-scale data via the OSC protocol, including notes and control data. Then, it exchanges a large array of data via the Syphon protocol [78], including the audio spectra and agent cluster representing the position data.

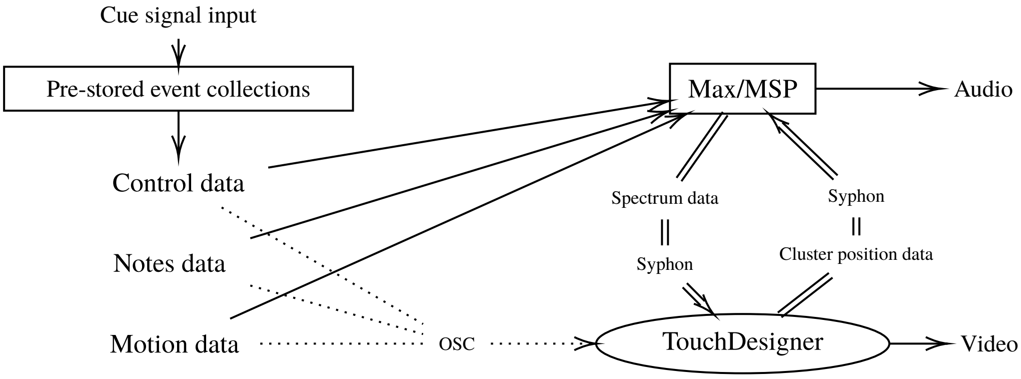


Figure 7.12: *Metamorphosis* Response stage signal flow

The audio part utilises three steps to produce the final audio response, as illustrated in Figure 7.13:

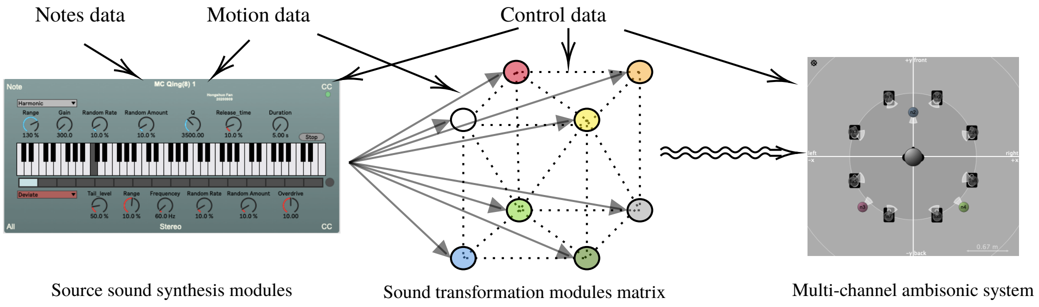


Figure 7.13: *Metamorphosis* Response stage sound part three process steps

1. The first step is designed to produce the sound source material using three sound synthesis modules, which listen to the incoming notes from the previous stage. These sound synthesis modules are based on a Karplus Strong algorithm [88], combined with Max, and produce multiple audio channel objects [89] as the basis to model the Bianqing sound. Parallel to this, the system maps the controller's motion data to specific modulation parameters determined by the composer. As a result, it obtains a continuous sound response from the subtle gestures and physical nuances. For example, at minute 0:16, the performer subtly changes the sound duration of the phrase with the motion of the hand.

2. The second step aims to extend the source materials by producing the sonic variations from them. The system calls back the stored parameters of the audio signal processing modules (Table 7.1) to control the timbre of these variations. Then, it sends the signal to the audio matrix using a different mixing ratio to control the generation of related materials. For example, at the beginning of the piece, no sound source is sent to the matrix. Instead, the signal goes directly to the output. However, from minute 12:39, the sound source is exclusively sent to the matrix, avoiding rerouting it to the direct output. The latter allows for the new materials to evolve away from the source, creating some unexpected variations. In parallel to this process, the role of the dataset from the visual part is to manipulate the timbral characteristics of each new sound directly. For instance, at minute 13:17, the cluster distribution determines the difference in the centre frequency of a resonant bandpass filter to apply subtle variations to different source materials owning particular spectral content.

3. In the third step, the resulting audio signals from previous processes collide and are sent to the multichannel ambisonic system. This uses a predetermined mixing ratio, which integrates the Spat library [62] to assign a spatial position to the final sonic materials in real-time.

Table 7.1: *Metamorphosis* audio signal processing modules

Name	Function Description
Freeze	Audio capture and reconstruction through real-time FFT spectrum analysis.
Granular FM	Real-time audio granular synthesizer. The audio granular playback speed is based on a reference decimal (micro) pitch table.
Freeze reverb	A reverb effect taken from the implementation by Juhana Sadeharju.
Audio hub	Four-channel spectrum bandpass filter. Each channel's pass frequency range can be altered in real time.
Granulator	Apply multiple playbacks of the audio recording to create a granular cloud effect.
MC resonant	32-channel resonant bandpass filter
Stretch effect	Apply real-time loop recording to the audio input. Then, play back the recording through multiple layers with variable speeds.
Loop sampling	Apply loop recording and forward/backward playback to extend the input audio.

The visual element consists of three virtual scenes and applies two steps to produce the final visual response, as shown in Figure 7.14. Similarly to the sound part, Max prestored data is sent to TouchDesigner to produce a specific distinct visual environment, distinctive from each other, as the piece progresses.

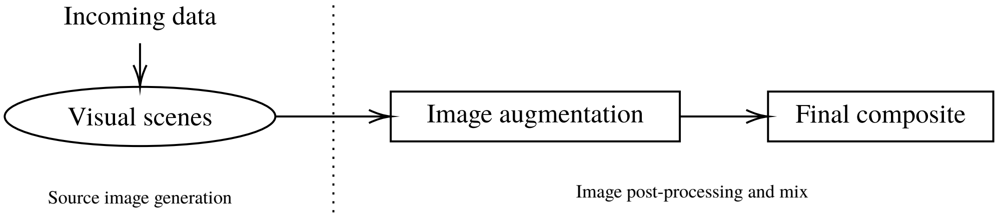


Figure 7.14: *Metamorphosis* Response stage visual part two process steps

1. As Figure 7.15 indicates, the first visual scene aims to introduce a framework to host the primary form of interaction. The notation materials from the audio stage are sent to the visual realm to trigger the corresponding movement of the virtual chimes made of stone. The aim is to visually match both elements and make the audiovisual contract effective [43]. This first scene applies three different camera images to display the virtual space from different angles. The size of each image is bound to individual instrument dynamics. For example, at 0:28, when the trio musically interacts, it

enhances the visual tension across their correspondent images, driven by contrasting sonic energies competing for space in the piece. Furthermore, there is also a feedback loop, where the sound directly affects some visual elements to enhance the organic aspect and continuity of the audiovisual continuum. For example, at minute 6:55, the spectrum data of the sound variation moves the ground position of the virtual instrument up and down.

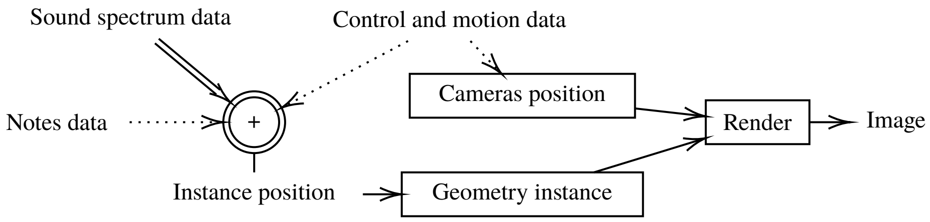


Figure 7.15: *Metamorphosis* first visual scene structure

2. The second visual scene aims to manifest creative breakthrough and to provide a contrast with previous visual environments. Starting at minute 10:40, the human performer's gesture of striking the air is sent to the visual element to trigger the light spot on a dark background. Then, the light spot gradually grows to match the sonic intensity before leading to the next scene.

3. The third visual scene showcases multidimensional interaction between performers and the gradual merge of the audiovisual element. As indicated in Figure 7.16, it consists of three steps to produce the final output:

    (a) The first step produces the initial placement of the visual elements following two parallel processes:

    - Using cluster point positions based on a Boids algorithm [90]. This applies the incoming control data as the aggregation force and the sound spectrum data as the distribution force, which affects the Boids' behaviour.
    - Using human landmark point positions, which are achieved via implementing Open-CV library with a live webcam [91], specifically, background subtraction [92] and Good Feature tracking [93], in order to capture the behaviour of the human performer in real-time.

    (b) The second step calculates the distance between the position of the human landmark and the cluster points.

    (c) The third step utilises all data generated in previous steps to produce the final visual environment.

    - The system applies position data from the human landmarks to generate points in the virtual space. In the meantime, it calculates the distance between each landmark point; if the result is below a given threshold, it generates a line to connect the point with the neighbour.

- It utilises the position of the cluster point to drive the ribbon's head point. The ribbon is a line with 100 points that employs a feedback material method to pass the position from the head to the tail point.
- Finally, the connecting lines are based on the distance data between the body landmark and the cluster, which are only enabled within a given range.

To conclude, the role of the additional visual effects is to enhance musical expression and create a strong sense of musical interaction with the moving image. For example, at minute 16:37, the performer's striking gesture is visually represented as a rippling effect on the ground.

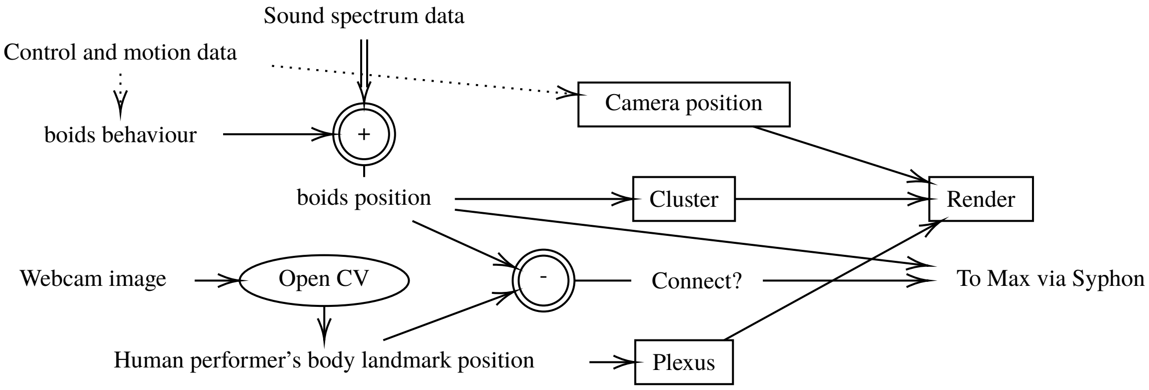


Figure 7.16: *Metamorphosis* third visual scene structure

Chapter 8

Conclusion

The novel paradigm “Internet of things” (IoT) refers to the concept of the pervasive presence of a variety of “things” or “objects”, which can interact with one another and cooperate with their neighbouring “smart” components to reach common goals [27]. This portfolio set out to explore how the IoT era’s technology may expand existing interactive multimedia performance systems and interactive music composition, especially in terms of new forms of machine musicianship, which can evolve from human-to-machine interaction to machine-to-machine communication using artificial intelligence technology and conceptual frameworks.

Although each composition is different in many respects, this portfolio has proposed two core creative devices：

1. The first one is the design of a Hybrid Interactive System gradually departing from human interaction. As the portfolio progressed, the system evolved (Figure 8.1) to explore the intersections between the IMS and the IoT system, in order to address new possibilities for musical expression. This system evolution is built upon existing interactive music design (usually deployed in three stages) by incorporating additional creative value into existing frameworks.

    - Adding further modularity: Each stage was subdivided into independent interactive subsystems, where each sub-module was relatively independent of the rest, although aligned with the main system targets. This not only supported more complex compositional structures and methods, but was also proven to be a creative catalyst for the compositional process.
    - Extending connectivity: Each stage connected to the next stage in the signal flow, taking the prior stage’s output as its current input. When multiple interactions occurred, it allowed for different stage sub-systems to form an extended framework for sub-modular cooperation, in order to reach common goals.

2. The second core aspect was to develop additional hybrid composition methodologies for the above: developed in parallel to the Hybrid Interactive System, it was presented in three evolving musical directions:

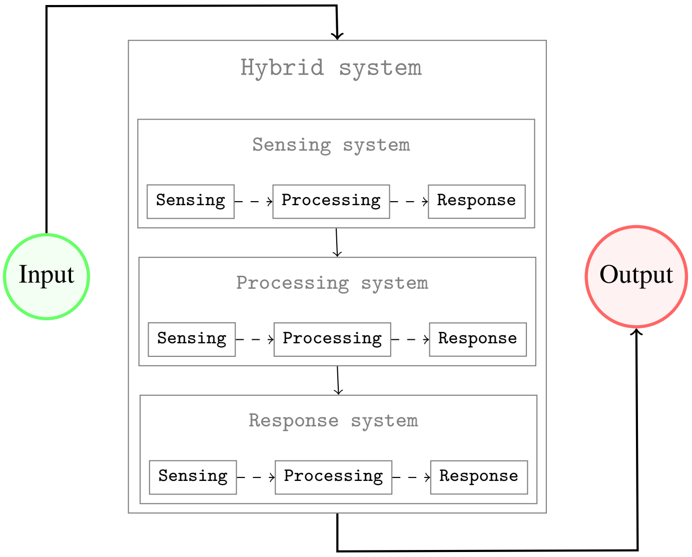


Figure 8.1: Structure and signal flow of the Hybrid system

- Multiparadigm: The composition paradigm shifted from single- to multiparadigm combinations, combining score- and performance-driven features. For example, when switching between paradigms, the composer could better shape the balance between fixed notated music and improvised material.
- Structure: The compositional structure evolved to navigate different relationships in the audiovisual contract, from strong to freer interlinks, to enhance and optimise the integration of media and the musical narrative.
- Interactivity: The interaction between different media evolved from static and one-dimensional to dynamic and multidimensional. For instance, a one-way interaction from human to machine shifted to a much more complex flux including human to machine, machine to machine and machine to human interaction. As a result, the composer could design the frequency and nature of these variations by switching the initial role of interaction between them in order to control the degree, and musical character of information in the signal flow [94].

In summary, this hybrid interactive system design provided further directions for combining musical thinking (composition and performance) and state-of-the-art emerging technology in AI. It allowed for the enhancement of existing HCI systems and enabled new forms of musical expression, with AI not only as a companion but elevated to the rank of a true performer. The involvement of deep learning in the creative process led to the deployment of advanced levels of musicianship from the AI and from human to AI, which validated the prior sentence. The unique personality of the AI and behaviour helped the composer to reveal a profound relationship between compositional elements across the dataset, the musical material and the performers' interactions.

8.1 *Conversation in the cloud*

In relation to compositional directions after this portfolio, *Conversation in the cloud*[1] is a good example of how the work in this portfolio can be extended to explore further musical practices across humans and AI to create music and media. This new piece is a live multimedia composition for one human musician and one AI musician, and shares some similar features with the work in this portfolio. For example, it applies similar sub-themes in the *Metamorphosis*, such as conversation, confrontation and cooperation, to explore performance circumstances frequently visited on stage in the interaction between human performers; it uses live audio-visual elements as a tool to control and enhance musical tension and textural expression like in *Intangible Field*. Also, this piece aims to converge compositional practices with additional emerging technologies in AI. For instance, it includes performer RNN (Section 5.3); the real-time, simultaneous perception of human pose, face landmarks and hand tracking [95] to provide an additional embodiment of the machine's intelligentsia. However, the human musician in *Conversation in the cloud* is a human bass clarinet player without sensors and intentionally uses the instrumental sound as the primary media to interact with the AI musician. Thus, the AI musician in this piece applies the existing system structure in the *Metamorphosis* but adjusts its machine musicianship to achieve reasonable interaction with the audio signal input. For instance, adding an analysing step in the sensing stage converts the raw audio signal to symbolic musical data. As a result, the AI musician can precisely capture the human musician's music and also provide a way to involve the sounds produced by contemporary instrumental playing techniques in the composition. For example, in the second section, starting at 4:04, the human musician is able to use various sounds produced by the bass clarinet to interact with the AI musician. It not only enhances the musical expression but also pushes the boundaries of their musicianship.

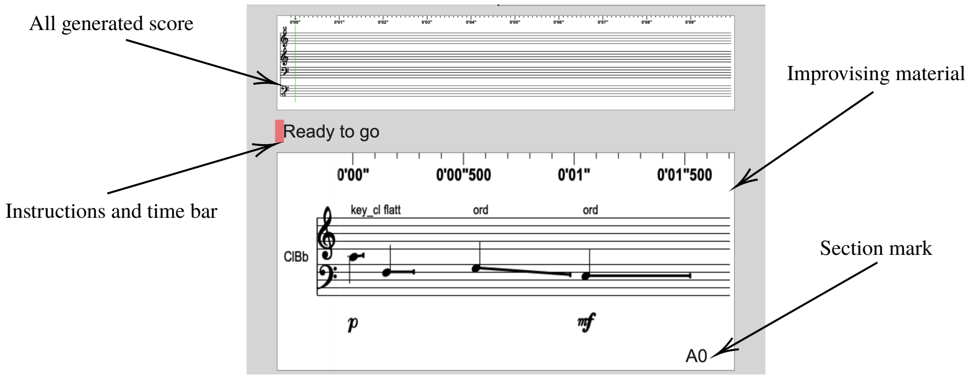


Figure 8.2: *Conversation in the cloud* real-time score interface

Besides, a web browser-based real-time score interface system [96] is applied to help the human musician view some essential information (Figure 8.2) during the performance, such as the live-generated score, additional instruction and section marks. It helps the human musician to follow the progress during the performance. More importantly, with its real-time score display, the human musician can not just be aware of musical content by listening but also visually capture the context and structure to organise various responses. Take

[1] Commissioned by the SWR experimental studio, the performance video recording is included in the USB content.

the third section, starting at 7:03, as an example. The AI musician generates musical material about two minutes in duration and visualises it on the interface. Therefore, the human musician can observe and consider all the generated music and build a relatively long-term musical tension flow through their improvisation.

8.2 Future research

The portfolio's research leads toward two interesting strands, which the composer would be willing to investigate:

- Deepening the language and grammar of AI for Music: AI could be used to design immersive experiences, that further explore the constructions found at the intersections of traditional culture and cutting-edge technology. Moreover, AI and ML can be rethought as yet another compositional tool, for example, using deep learning technology to enhance music and sonic exploration via transforming sound or as an alternative method for sound syntheses, such as RAVE [97].
- Exploring the cloud as a home for an expanded interactive multimedia performance system: With more advanced technology involved in the interactive multimedia composition process, including live performance, the need for computing power is becoming critical. Local computing power using currently accessible hardware has limitations because of the cost, mobility, and difficulty of the live performance setup. Therefore, shifting the entire (or partial) live interactive multimedia performance systems to a cloud server in order to address calculation-hungry tasks, such as 3D rendering, music information retrieval, and deep neural network inference is becoming a promising reality with the novel generation of network technology in mapping [98]. In one idealised scenario, the sensing stage signals could be sent to the cloud server during a live performance, for all the sound processing and image rendering to be accomplished remotely and fed back to the concert stage, including audio and visuals, within reasonable latency.

Alongside the exploration of new compositional practices for an evolving AI system, this portfolio of musical compositions also constructed a new path for future trial-and-error experimentation on new hybrid methodologies, but also in terms of the deconstruction of questions in relation to human–AI interaction, which in the past were partially unsolved. This may include identifying the harmonious state between automation and creativity when utilising AI technology for musical creation and determining how to reach it. These questions will be at the core of future investigation with the aim of contributing to a compositional field, which is constantly shaping the emergence of new technologies and the way humans interact with them.

References

[1] Takayuki Rai, *Discrete Transfer*, for piano and computer, 2012.

[2] ——, *Facade*, for guitar and computer, 2003.

[3] ——, *Impulse*, for percussion and computer, 1997.

[4] Hans Tutschku, *SprachSchlag*, for percussion and live-electronics, 2000.

[5] ——, *under*, for flute, oboe, clarinet, two violins, viola, violoncello, percussion, piano and electronics, 2013.

[6] ——, *virtual bodies*, for piano and live-electronics, 2017.

[7] Wojciech Błażejczyk, *#NetworkMusic*, for voice, ensemble and electronics [Baritone, 2 vn, vl, vc, fl, cl, b cl, sax, acc, e gt / b gt, perc], 2017.

[8] ——, *LoPassHiCut*, for double bass and live electronics, 2013.

[9] ——, *Trash Music*, for voice, objectophones and acoustic instruments [vc, acc, e gt, b cl], 2014.

[10] Simon Hutchinson, *Data-Driven Instruments*, 2020. [Online]. Available: `https://simonhutchinson.com/2020/08/03/data-driven-instruments/`.

[11] J. Impett, "Situating the invention in interactive music," *Organised Sound*, vol. 5, no. 1, pp. 27–34, 2000, ISSN: 14698153. DOI: `10.1017/S1355771800001059`.

[12] Atau Tanaka, "Sensor-Based Musical Instruments and Interactive Music," in *The Oxford Handbook of Computer Music*, Oxford University Press, Sep. 2012, pp. 233–257, ISBN: 9780199940233. DOI: `10.1093/oxfordhb/9780199792030.001.0001`.

[13] Jeffrey Stolet, *BetweenTheWords*, for Wacom tablet & Kyma, 2016.

[14] ——, *Lariat Rituals*, a real-time performance composition for Kyma, Max, and Gametrak controller, 2012.

[15] ——, *Tokyo Lick*, for two infrared MIDI controllers, 2007.

[16] B. Caramiaux, N. Montecchio, A. Tanaka, and F. Bevilacqua, "Adaptive gesture recognition with variation estimation for interactive systems," *ACM Transactions on Interactive Intelligent Systems*, vol. 4, no. 4, 2015, ISSN: 21606463. DOI: `10.1145/2643204`.

[17] J. Li, Z. L. Wang, H. Zhao, R. Gravina, G. Fortino, Y. Jiang, and K. Tang, "Fluid gesture interaction design: applications of continuous recognition for the design of modern gestural interfaces," in *BodyNets International Conference on Body Area Networks*, 2017.

[18] F. Bevilacqua, F. Baschet, and S. Lemouton, "The Augmented String Quartet: Experiments and Gesture Following," *Journal of New Music Research*, vol. 41, no. 1, pp. 103–119, Mar. 2012, ISSN: 09298215. DOI: `10.1080/09298215.2011.647823`.

[19] F. Pinel, L. R. Varshney, and D. Bhattacharjya, *Computational Creativity Research: Towards Creative Machines*. 2015, vol. 7.

[20] J. McCormack and M. D'Inverno, *Computers and Creativity*, J. McCormack and M. d'Inverno, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 9783642317279, pp. 39–60, ISBN: 978-3-642-31726-2. DOI: `10.1007/978-3-642-31727-9`. [Online]. Available: `http://link.springer.com/10.1007/978-3-642-31727-9`.

[21] G. Kamhi, A. Novakovsky, A. Tiemeyer, and A. Wolffberg, "MAGENTA," 2009. DOI: `10.1145/1629911.1630080`.

[22] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.

[23] A. Pal*, S. Saha, and Anita, "Musenet : Music Generation using Abstractive and Generative Methods," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 6, pp. 784–788, Apr. 2020, ISSN: 22783075. DOI: `10.35940/ijitee.F3580.049620`. [Online]. Available: `https://www.ijitee.org/portfolio-item/F3580049620/`.

[24] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A Generative Model for Music," Apr. 2020. [Online]. Available: `http://arxiv.org/abs/2005.00341`.

[25] B. Caramiaux and M. Donnarumma, "Artificial Intelligence in Music and Performance: A Subjective Art-Research Inquiry," in *Handbook of Artificial Intelligence for Music*, 2021. DOI: `10.1007/978-3-030-72116-9{\_}4`.

[26] Sam Salem, *THIS IS FINE*, for solo cello, performative electronics, tape and video projection., 2021.

[27] D.Giusto, A.Iera, G.Morabito, and L.Atzori(Eds.), *The Internet of Things*. Springer, 2010, ISBN: 978-1-4419-1673-0.

[28] R. Rowe, *Machine Musicianship*. The MIT Press, 2004, pp. 1–2, ISBN: 0-262-18206-8.

[29] ——, *Interactive Music Systems*. The MIT Press, 1994, pp. 1–2, ISBN: 0-262-18149-5.

[30] X. Wang, "Wuxing : An Investigation Into the Interpretations of Traditional Chinese Cosmology in Contemporary China," *The Asia Pacific Journal of Anthropology*, vol. 20, no. 2, pp. 129–146, Mar. 2019, ISSN: 1444-2213. DOI: 10.1080/14442213.2019.1572783. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/14442213.2019.1572783.

[31] W. Li, *Chinese Writing and Calligraphy*. University of Hawaii Press, May 2010, ISBN: 9780824860691. DOI: 10.1515/9780824860691. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/9780824860691/html.

[32] Robert Rowe, *Interactive Music System*. 1994, p. 6, ISBN: 0-262-18149-5.

[33] M. Wright, "Open Sound Control: An enabling technology for musical networking," *Organised Sound*, vol. 10, no. 3, 2005, ISSN: 14698153. DOI: 10.1017/S1355771805000932.

[34] Andy Farnell, *Designing Sound*. MIT Press, 2010.

[35] DIPS Development Group, *Digital Image Processing with Sound*, 2013. [Online]. Available: https://dips.kcm-sd.ac.jp/.

[36] Gary Farr, *Vonnegut Collective*, 2014. [Online]. Available: https://vonnegutcollective.co.uk/.

[37] H. M. Adam Davies, *Animikii Theatre*, 2014. [Online]. Available: www.animikiitheatre.com.

[38] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, 2017. DOI: 10.1109/CVPR.2017.395.

[39] G. Papandreou, T. Zhu, L. C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11218 LNCS, 2018. DOI: 10.1007/978-3-030-01264-9{_}17.

[40] Mark J. Harris, "Fast Fluid Dynamics Simulation on the GPU," in *GPU Gems*, Randima Fernando (Series Editor), Ed., Addison Wesley; Har/Cdr edition, 2004, ch. Chapter 38, ISBN: 978-0321228321. [Online]. Available: https://developer.nvidia.com/gpugems/gpugems/part-vi-beyond-triangles/chapter-38-fast-fluid-dynamics-simulation-gpu.

[41] *Distracfold*, 2014. [Online]. Available: http://www.distractfold.co.uk.

[42] S. Pv, "Introduction to Unreal Engine 4," in *Beginning Unreal Engine 4 Blueprints Visual Scripting*, 2021. DOI: 10.1007/978-1-4842-6396-9{_}1.

[43] M. Chion, *Audio-Vision: Sound on Screen*. Columbia University Press, Dec. 2019. DOI: 10.7312/chio18588.

[44] F. J. Gallego-Durán, C. J. Villagrá-Arnedo, R. Satorre-Cuerda, P. Compañ-Rosique, R. Molina-Carmona, and F. Llorens-Largo, "A guide for game-design-based gamification," *Informatics*, vol. 6, no. 4, 2019, ISSN: 22279709. DOI: 10.3390/informatics60400

[45] D. Strzałko, "Voice Controlled Games – The approach and challenges of implementing speech recognition and voice control in games," in *Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems*, vol. 26, 2021. DOI: 10.15439/2021f143.

[46] E. Joseph, "Bot Colony-a Video Game Featuring Intelligent Language-Based Interaction with the Characters," Tech. Rep.

[47] Clare Edgeley, "Arcade Action: Nemesis," *Computer and Video Games No. 48*, p. 96, Sep. 1985. [Online]. Available: https://solvalou.com/arcade/reviews/219/595.

[48] Ricardo Climent, *Duel of Strings: for Violin (non-virtual) vs. Virtual Strings*, Manchester, Mar. 2019. [Online]. Available: https://vimeo.com/gameaudio%20https://www.research.manchester.ac.uk/portal/en/publications/duel-of-strings(696cb9ee-8bea-43b1-b4d4-a20e0c9c47c9).html.

[49] H. Hancock, "Better Game Design Through Cutscenes," *Gamasutra*, 2002.

[50] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, 2002, ISSN: 0001-4966. DOI: 10.1121/1.1458024.

[51] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, ISSN: 14228890. DOI: 10.1007/s12525-021-00475-2.

[52] S. J. (J. Russell, *Artificial intelligence : a modern approach*, eng, Third edition / c..., P. Norvig and E. Davis, Eds., ser. Prentice Hall series in artificial intelligence. Harlow: Pearson, 2016, ISBN: 9781292153964.

[53] A. Géron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow - Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition (2019)," in *Journal of Physics A: Mathematical and Theoretical*, 8, vol. 44, 2019, pp. 7–9.

[54] Takayuki Rai, *Takayuki Rai*, 2014. [Online]. Available: `http://www.t-rai.net/`.

[55] J. Elson, J. R. Douceur, J. Howell, and J. Saul, "Asirra: A CAPTCHA that exploits interest-aligned manual image categorization," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2007. DOI: `10.1145/1315245.1315291`.

[56] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," Sep. 2016. [Online]. Available: `http://arxiv.org/abs/1609.03499`.

[57] P. Verma and C. Chafe, "A Generative Model for Raw Audio Using Transformer Architectures," Jun. 2021. [Online]. Available: `http://arxiv.org/abs/2106.16036`.

[58] Ian Simon and Sageev Oore, "Performance RNN: Generating Music with Expressive Timing and Dynamics.," *Magenta Blog*, 2017.

[59] A. Agostini and D. Ghisi, "Bach: An environment for computer-aided composition in max," in *ICMC 2012: Non-Cochlear Sound - Proceedings of the International Computer Music Conference 2012*, 2012.

[60] S. Hochreiter and J. Schmidhuber, "Long Short Term Memory. Neural Computation," *Neural Computation*, vol. 9, no. 8, 1997, ISSN: 21695717.

[61] G. Plonka, D. Potts, G. Steidl, and M. Tasche, "Fast fourier transforms," in *Applied and Numerical Harmonic Analysis*, 2018. DOI: `10.1007/978-3-030-04306-3{\_}5`.

[62] T. Carpentier, "A new implementation of spat in Max," in *Proceedings of the 15th Sound and Music Computing Conference: Sonic Crossings, SMC 2018*, 2018.

[63] Richard Dudas and Robert Piéchaud, *Modalys Documentation*, 2021. [Online]. Available: `https://support.ircam.fr/docs/Modalys/current/index.html`.

[64] L. Tan and M. Lu, " "I Wish to Be Wordless" : Philosophizing through the Chinese Guqin," eng, *Philosophy of music education review*, vol. 26, no. 2, pp. 139–154, 2018, ISSN: 1063-5734. DOI: 10.2979/philmusieducrevi.26.2.03.

[65] C. Yingshi, "Ancient chinese music notation," eng, *Anuario musical*, vol. 44, pp. 239–239, 1989, ISSN: 0211-3538.

[66] Sharron Gu, *A Cultural History of the Chinese Language*. 2011, pp. 99–100, ISBN: 978-0-7864-8827-8.

[67] M. J. Powers and K. R. Tsiang, *A companion to Chinese art*, eng, M. J. Powers and K. R. Tsiang, Eds., ser. Wiley Blackwell Companions to Art History. West Sussex, England: Wiley Blackwell, 2016, pp. 177–178, ISBN: 9781118885208.

[68] Z. Fa, *The history and spirit of Chinese art. Volume 2, From the Song to the Qing dynasty*, eng. Honolulu: Silkroad Press, 2016, pp. 177–178, ISBN: 9781623201289.

[69] Yusong Wu and Shengchen Li, "Guqin Dataset: A symbolic music dataset of Chinese Guqin collection," Beijing University of Posts and Telecommunications, Beijing, Tech. Rep., 2019. [Online]. Available: https://github.com/lukewys/Guqin-Dataset.

[70] S. Chen, D. Beeferman, and R. Rosenfeld, "EVALUATION METRICS FOR LANGUAGE MODELS," Tech. Rep.

[71] P. Prusinkiewicz, *The algorithmic beauty of plants*, eng, P. Prusinkiewicz, Ed., ser. The Virtual Laboratory. New York, [New York: Springer, 1990, ISBN: 9781461384762.

[72] Derivative, *LSystem SOP*, Jul. 2021. [Online]. Available: https://docs.derivative.ca/index.php?title=LSystem_SOP.

[73] P. Prusinkiewicz, "A look at the visual modeling of plants using L-systems," *Agronomie*, vol. 19, no. 3-4, 1999, ISSN: 02495627. DOI: 10.1051/agro:19990303.

[74] Robert Rowe, *Interactive Music System*. 1994, pp. 1–10, ISBN: 0-262-18149-5.

[75] Nordic, *Nordic Thingy:52*, 2019. [Online]. Available: https://www.nordicsemi.com/Products/Development-hardware/Nordic-Thingy-52.

[76] M. Collotta, G. Pau, T. Talty, and O. K. Tonguz, "Bluetooth 5: A Concrete Step Forward toward the IoT," *IEEE Communications Magazine*, vol. 56, no. 7, 2018, ISSN: 15581896. DOI: 10.1109/MCOM.2018.1700053.

[77] filipeflop, "Raspberry Pi 3 Model B - Raspberry Pi," *Raspberry Pi 3 Model B*, 2016.

[78] T. Butterworth and A. Marini, *Syphon*, 2019. [Online]. Available: https://github.com/Syphon/Syphon-Framework.

[79] G. E. Lewis, “ Too Many Notes: Computers, Complexity and Culture in Voyager,” *Leonardo Music Journal*, vol. 10, 2000, ISSN: 0961-1215. DOI: 10.1162/096112100570585.

[80] R. Rowe, *Machine Musicianship*. The MIT Press, 2004, pp. 1–2, ISBN: 9780262256896. DOI: 10.7551/mitpress/4361.001.0001. [Online]. Available: https://direct.mit.edu/books/book/1822/machine-musicianship.

[81] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach (4th Edition)*. 2021, pp. 36–60.

[82] G. Van Houdt, C. Mosquera, and G. Nápoles, “A review on the long short-term memory model,” *Artificial Intelligence Review*, vol. 53, no. 8, 2020, ISSN: 15737462. DOI: 10.1007/s10462-020-09838-1.

[83] “磬,” chi, 中文自修, no. 3, pp. 64–64, 2016, ISSN: 1000-7245.

[84] *Yamaha e-Piano Competition dataset*, 2018. [Online]. Available: https://www.piano-e-competition.com/.

[85] dekuNukem, *Nintendo Switch Reverse Engineering*. [Online]. Available: https://github.com/dekuNukem/Nintendo_Switch_Reverse_Engineering.

[86] J. Chadabe, “The MIDI World,” in *Electric Sound: The Past and Promise of Electronic Music*, 1, 1997, pp. 185–212.

[87] A. Agostini and D. Ghisi, “A Max Library for Musical Notation and Computer-Aided Composition,” *Computer Music Journal*, vol. 39, pp. 11–27, 2015. DOI: 10.1162/COMJ. [Online]. Available: http://direct.mit.edu/comj/article-pdf/39/2/11/1856136/comj_a_00296.pdf.

[88] K. Karplus and A. Strong, “DIGITAL SYNTHESIS OF PLUCKED-STRING AND DRUM TIMBRES.,” *Computer Music Journal*, vol. 7, no. 2, 1983, ISSN: 01489267. DOI: 10.2307/3680062.

[89] Cycling74, *Multiple channels of audio*. [Online]. Available: https://docs.cycling74.com/max8/vignettes/mc_topic.

[90] C. W. Reynolds, “FLOCKS, HERDS, AND SCHOOLS: A DISTRIBUTED BEHAVIORAL MODEL.,” *Computer Graphics (ACM)*, vol. 21, no. 4, 1987, ISSN: 00978930. DOI: 10.1145/37402.37406.

[91] J. Howse, *OpenCV Computer Vision with Python*. 2013.

[92] A. Vacavant, T. Chateau, A. Wilhelm, and L. Lequièvre, “A benchmark dataset for outdoor foreground/background extraction,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics), vol. 7728 LNCS, 2013. DOI: 10.1007/978-3-642-37410-4{_}25.

[93] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1994. DOI: 10.1109/cvpr.1994.323794.

[94] Karlheinz Stockhausen, "STRUCTURE AND EXPERIENTIAL TIME," *Die Reihe musical journal*, vol. 2, pp. 64–75, 1958.

[95] Ivan Grishchenko and Valentin Bazarevsky, *MediaPipe Holistic — Simultaneous Face, Hand and Pose Prediction, on Device*, Oct. 2020.

[96] S. Tarakajian, D. Zicarelli, and J. Clayton, "Mira: Liveness in iPad Controllers for Max/MSP," *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2013.

[97] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," Nov. 2021. [Online]. Available: http://arxiv.org/abs/2111.05011.

[98] X. You, C. X. Wang, J. Huang, X. Gao, Z. Zhang, M. Wang, Y. Huang, C. Zhang, Y. Jiang, J. Wang, M. Zhu, B. Sheng, D. Wang, Z. Pan, P. Zhu, Y. Yang, Z. Liu, P. Zhang, X. Tao, S. Li, Z. Chen, X. Ma, I. Chih-Lin, S. Han, K. Li, C. Pan, Z. Zheng, L. Hanzo, X. S. Shen, Y. J. Guo, Z. Ding, H. Haas, W. Tong, P. Zhu, G. Yang, J. Wang, E. G. Larsson, H. Q. Ngo, W. Hong, H. Wang, D. Hou, J. Chen, Z. Chen, Z. Hao, G. Y. Li, R. Tafazolli, Y. Gao, H. V. Poor, G. P. Fettweis, and Y. C. Liang, *Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts*, 2021. DOI: 10.1007/s11432-020-2955-6.

[99] R. Rowe, *Machine Musicianship*. The MIT Press, 2004, p. 317, ISBN: 0-262-18206-8.

[100] ——, *Interactive Music Systems*. The MIT Press, 1994, p. 9, ISBN: 0-262-18149-5.

[101] Eng Tat Khoo, R. L. Peiris, and M. Rauterberg, "3D Guqin: Digital Playground to Explore Music that Embodies Chinese Culture and Philosophy," eng, in *2011 Second International Conference on Culture and Computing*, IEEE, 2011, pp. 145–146, ISBN: 9781457715938. DOI: 10.1109/Culture-Computing.2011.41.

[102] L. Henbing and M. Leman, "A Gesture-based Typology of Sliding-tones in Guqin Music," eng, *Journal of new music research*, vol. 36, no. 2, pp. 61–82, 2007, ISSN: 0929-8215. DOI: 10.1080/09298210701755073.

[103] C. Vernallis and N. Cook, "Analysing Musical Multimedia," *American Music*, vol. 19, no. 4, p. 480, 2001, ISSN: 07344392. DOI: 10.2307/3052426. [Online]. Available: https://www.jstor.org/stable/3052426?origin=crossref.

[104] R. Rada, *Artificial intelligence. E. Rich, (McGraw-Hill, New York, 1983); 411 pages, $30.95*, 1986. DOI: 10.1016/0004-3702(86)90034-2.

[105] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach (2nd Edition).* 2002, pp. 48–50.

[106] S. Russel and P. Norvig, *Artificial intelligence—a modern approach 3rd Edition.* 2012. DOI: 10.1017/S0269888900007724.

[107] Ricardo Climent, *B is for Bird*, Jun. 2017. [Online]. Available: https://vimeo.com/229351488.

[108] W. Jiang, Z. Wang, J. S. Jin, Y. Han, and M. Sun, "DCT – CNN-based classification method for the Gongbi and Xieyi techniques of Chinese ink-wash paintings," eng, *Neurocomputing (Amsterdam)*, vol. 330, pp. 280–286, 2019, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2018.11.003.

[109] P. D. Bharathi, V. Ananthanarayanan, and P. Bagavathi Sivakumar, "Fog Computing-Based Environmental Monitoring Using Nordic Thingy: 52 and Raspberry Pi," in *Smart Innovation, Systems and Technologies*, vol. 141, 2020. DOI: 10.1007/978-981-13-8406-6{_}27.

[110] Y. Gong, *Guqin Yanzoufa*, 2nd ed. Shanghai: Shanghai Educational Press, 1999, ISBN: 7-5320-6621-5.

[111] W. Ertel, *Introduction to Artificial Intelligence (Undergraduate Topics in Computer Science).* 2017.

[112] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A Generative Model for Music," Apr. 2020. [Online]. Available: http://arxiv.org/abs/2005.00341.

[113] M. Wright, R. Dannenberg, S. Pope, X. Rodet, X. Serra, and D. Wessel, "Panel: Standards from the computer music community," in *Proceedings of the 2004 International Computer Music Conference, Miami, FL*, 2004.

[114] Jan Erik Solem, "Programming Computer Vision with Python," *Programming Computer Vision with Python*, 2012, ISSN: 1098-6596.

[115] Michel Chion, "Sound An Acoulogical Treatise,"

[116] J. Françon, "The algorithmic beauty of plants," *Plant Science*, vol. 122, no. 1, 1997, ISSN: 01689452. DOI: 10.1016/s0168-9452(96)04526-8.

[117] E. O. Brigham and R. E. Morrow, “The fast Fourier transform,” *IEEE Spectrum*, vol. 4, no. 12, 1967, ISSN: 00189235. DOI: 10.1109/MSPEC.1967.5217220.

[118] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010, ISSN: 13891286. DOI: 10.1016/j.comnet.2010.05.010.

[119] M. Wang, “The Meaning of Xing ’Greek Passage’ and Moral Transformation in Wuxing,” *Frontiers of Philosophy in China*, vol. 11, no. 2, 2016, ISSN: 1673355X. DOI: 10.3868/s030-005-016-0017-3.

[120] J. Mccarthy, “WHAT IS ARTIFICIAL INTELLIGENCE?” Tech. Rep., 2007. [Online]. Available: http://www-formal.stanford.edu/jmc/.

[121] 润. 李, 懿. 侯, 若. 何, and 春. 初, “ “以琴载德” ——古琴艺术的道德内涵探讨,” 教学方法创新与实践, vol. 4, no. 9, 2021, ISSN: 2661-4367. DOI: 10.26549/jxffcxysj.v4i9.7294.

[122] 韩晓莉, “中国民族器乐发展控究,” chi, 音乐创作, no. 3, pp. 144–145, 2013, ISSN: 0513-2436.

[123] 杨帆, “古琴声学特性与音响表现关系解析,” chi, *Ren min yin yue*, no. 1, pp. 85–87, 2015, ISSN: 0447-6573.

[124] 潘蕾雅, 梁锦钰, 姚欢夏, and 李昂, “浅论古琴制作工艺,” chi, 艺术评鉴, no. 9, pp. 180–181, 2019, ISSN: 1008-3359.

Appendix A

Compositions' narrative explanation

A.1 *Handwriting · WuXing*

A.1.1 The water section

The whole composition starts from the water section since "Water is the driving force of all nature" (Leonardo da Vinci). However, one short intro section is presented before the start of the water section. There are two important reasons to add this intro section:

1. To introduce the primary interaction mechanism to the audience, which uses finger and hand gestures to trigger and control the sound and image changes.
2. Reveal the conformance and complementation [103] relationships between the sound, image, and performer's gestures.

During the intro section, the performer starts with the "hook" gesture because it is the first stroke of the Chinese character: water (水). To keep the audience focused on the gesture, the intro's visual part has only white traces on a black background as a metaphor for ink on paper. The sound part uses procedural sound to mimic the friction between the pen and paper; it is also the primary sound material to reflect the movement of the hand. After the performer has finished all five strokes (Figure A.1), the white traces will form the Chinese water character and indicate the start of the water section.

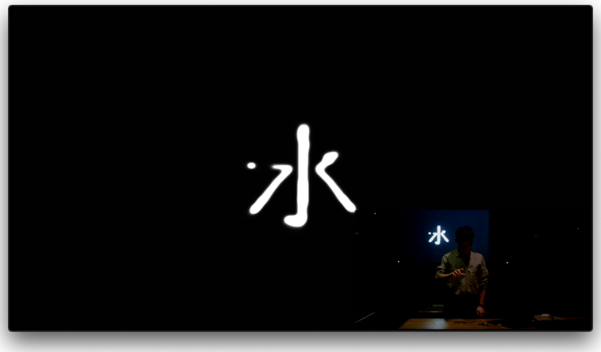


Figure A.1: *Handwriting · WuXing* the water section beginning screenshot

The water section starts with a real-time procedural raindrop-like sound and image. At this moment, all the sound is only output from the front speakers, which means a sound field is only created directly in front of the audience. After one horizontal line stroke gesture has been made, the raindrop-like sound diffuses into all eight channels, in which the sound field from the front side expands into the whole area. This spatial sound change creates a window through which the audience can immerse all their senses in the composition, and serves as psychological preparation for the later section, which has more complex spatial sound change.

After this starting point, both sound and visual elements start to increase in density. Next, the raindrops begin to gather, and the frequency of sound begins to rise, gradually forming a new shape for both parts. During this subsection, sound and visual shape development is a metaphor for different water elements' states and the shift between those different states; for instance, ice is in a solid state, formed by highly polymerized liquid water. Additionally, the three gestures in this section are chosen from the Chinese character: big (大), which is one horizontal line, one down to the left and one down to the right. Then, after those three-stroke gestures have finished, the process moves on to the visual scene in the following subsection. Finally, when the image of a large amount of water has formed (Figure A.2), it indicates that water cycles have been established; it also indicates the entrance into the wood section's preparation stage.

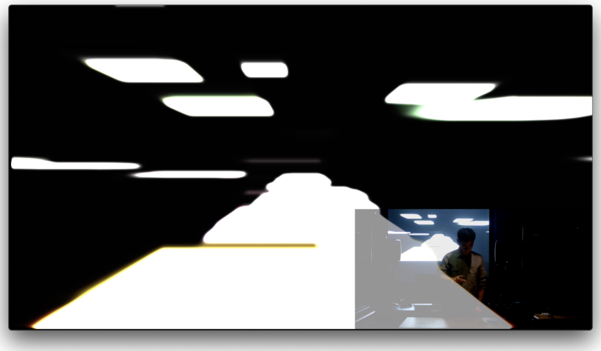


Figure A.2: *Handwriting* · *WuXing* the water section ending screenshot

A.1.2 The wood section

In the wood section's preparation stage, similarly to in the intro section, the performer needs to finish four-stroke gestures of the Chinese character wood 木, in order to trigger the start of the wood section. When the gesture condition has been fulfilled, the Chinese character wood will form in the visual part, as shown in Figure A.3. Next, the previous section's sound will fade out, and the primary sound material for the wood section will fade in, which is the procedural sound that mimics the sound of the wind blowing over trees.

At the beginning of the wood section, the sound and visual parts have returned to a relatively simplistic scene, in which only the wind sound and image of traces left by moving

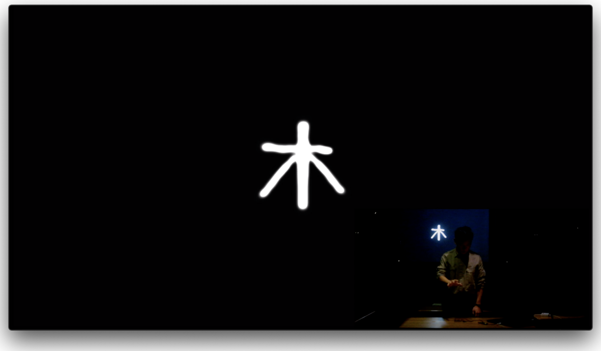


Figure A.3: *Handwriting* · *WuXing* the wood section beginning screenshot

objects remain (Figure A.4). After that, however, the sound part starts to introduce relatively complex spatial movements, for example, starting with the clockwise movements around the whole listening area, which then gather in the centre. The movement of the objects in the visual part also represents the projection mapping to the spatial sound movements, providing another abstract expression layer of those movements.

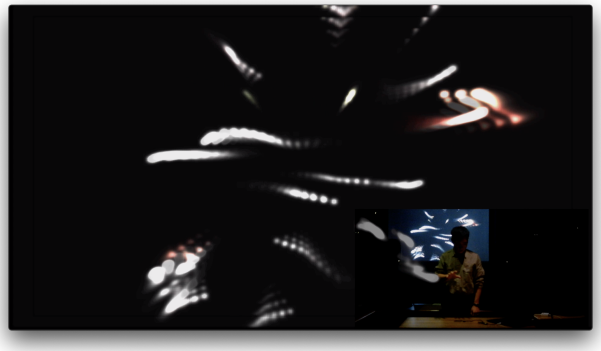


Figure A.4: *Handwriting* · *WuXing* the wood section moving objects screenshot

Next, with several stroke gestures being made, which are chosen from the Chinese character wind (风), the composition will gradually increase the complexity of the wind sound's spatial movements and the amount and speed of the visual objects. Finally, when the wind takes over the entire sound field and visuals (Figure A.5), it suggests that the composition approaches the end of the wood section and is ready to enter the next section's preparation stage, the fire section.

A.1.3 The fire section

Similar to the previous section's preparation stage, the performer needs to finish four-stroke gestures of the Chinese character fire (火). During this preparation stage, with the gradual and continuous completion of the gesture, the sound and visual parts also indicate the accruement of energy; this procedure also builds up an expectation among the audience of the next higher peak. When the gestures have been completed, it will trigger three short events, one after another:

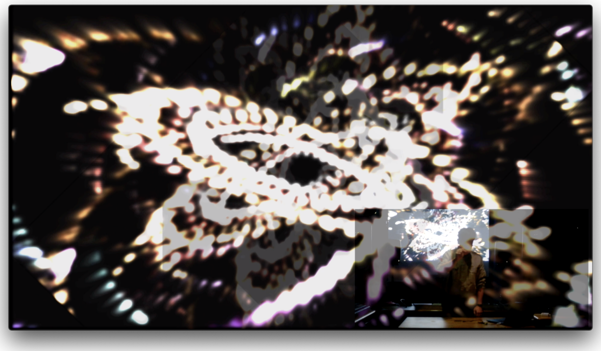


Figure A.5: *Handwriting · WuXing* the wood section ending screenshot

1. The sound and visual elements are quickly gathered to the central point, then sound vanishes; this creates a sense of a highly concentrated vacuum state of the objects.
2. The highly concentrated visual objects in the centre of the image explode alongside the sound of an explosion. These two highly contrasting statuses, silence and explosion, create more than one peak moment in this composition and represent the shifting process between the two highly different phases in WuXing.
3. The Chinese character fire (Figure A.6) will form in the visual part and reveal the fire section's fundamental sound material at this moment.

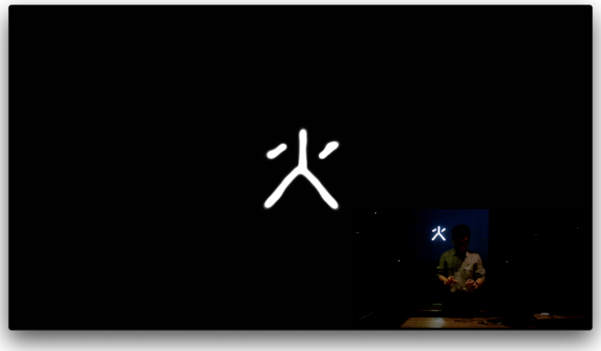


Figure A.6: *Handwriting · WuXing* the fire section beginning screenshot

The fire section is different from the previous sections, in that the purpose of the stroke gestures is not only to trigger the subsections, but each subsection also has a predetermined duration, which will automatically move on to the following subsection when the current subsection has finished. These automatically processed subsections have two main purposes in the fire section. On the one hand, the fire section aims to be more dynamic than before; thus, using fast hand movement data directly generates parameters to handle the sound and visual elements. On the other hand, avoiding slow stroke gestures breaks the sense of continuity during these fast hand movement sections. Nevertheless, the fast-moving hand gesture is a variant of the stroke gestures from the continuous writing of the Chinese character fire. Furthermore, the performer can fully concentrate on the performance without the need for additional gestures and attention to trigger the following subsections.

Moreover, during this section, the sound and visual elements start from concrete object status, the sound of firewood burning and the flame's shape, then gradually reform into more abstract sounds and images, as shown in Figure A.7. This change symbolises the shift between the fire section and the earth section.

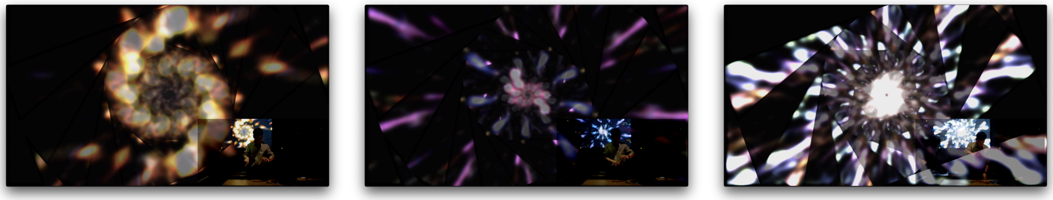


Figure A.7: *Handwriting* · *WuXing* the fire section visual process screenshot

A.1.4 The earth section

After reaching the fire section's peak point, the system will enter the earth section's preparation stage, waiting until the performer makes the three-stroke gestures from the Chinese character earth(土) (Figure A.8).

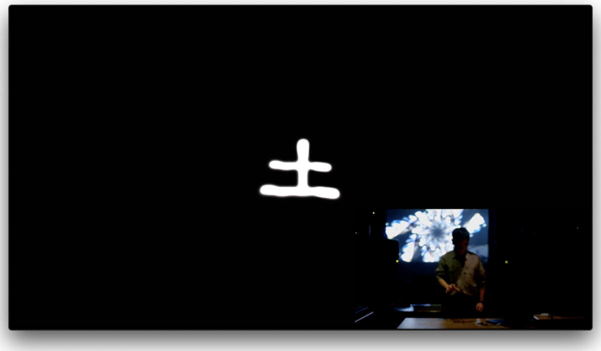


Figure A.8: *Handwriting* · *WuXing* the earth section beginning screenshot

At the beginning of the earth section, one of the most noticeable differences is the unblemished objects that the visual elements start to reveal, which are cubes. At this moment, all of the postvisual efforts are removed, and only the image of multiple cubes rolling remains (Figure A.9), accompanied by a sound that simulates the sounds of rolling stones. Thus, this combination builds up a rustic and concrete scene in the earth section. With the gradually circling and rising gestures, the relationship between gesture and sound becomes more apparent, as well as the relationship with the visual elements' movement. Within this exploration process, when the gesture reaches the highest point, the whole scene naturally transitions to the next section, the metal section.

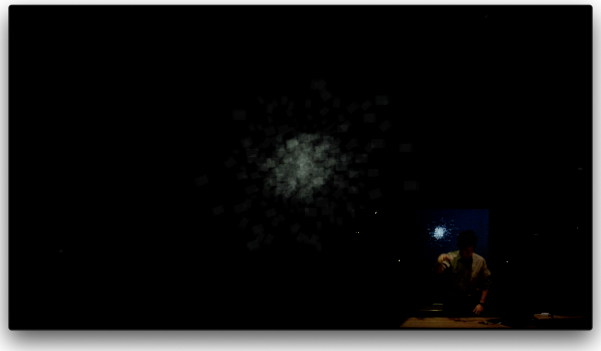


Figure A.9: *Handwriting · WuXing* the earth section “rolling cube” screenshot

A.1.5 The metal section

Unlike the previous sections, because of continuity, the metal section does not start with any preparation stage, but with the continuation of the more profound exploration of sound and visual elements. On the visual side, the previous randomly rolling cubes collectively transform into a tunnel with a metallic lustre (Figure A.10). The sound part also transforms from the previous granular state to a more continuous shape. Then, the view gradually passes through the tunnel with seven hand gestures, in which the hand enters and exits the recognition area.

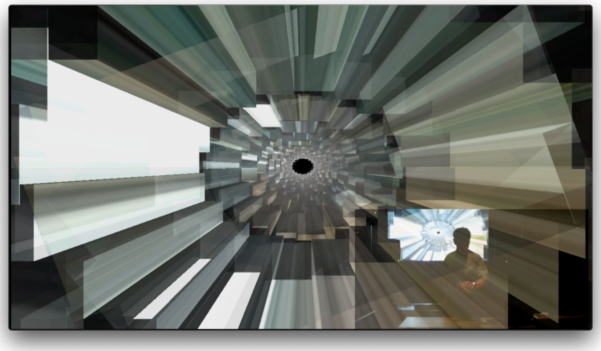


Figure A.10: *Handwriting · WuXing* the metal section beginning screenshot

After this exploration process is finished, the camera moves to a relatively distant position to reveal the whole picture, as shown in Figure A.11. Next, all the cubes move together to reform a multilayer circle shape, which is a metaphor for the relationship between the internal and overall circles of the WuXing theory.

Then, with the stroke gestures from the Chinese character metal(金) finished, the composition progresses to the next scene. In the visual part, the camera starts to move closer, and all the cubes start to transform from a multilayer circle into a moving spiral, and finally into a large grid formation (Figure A.12). The sound part also transforms into a more hollow and expansive ambient sound in the meantime.

The completion of the cubes’ shape transformation represents the start of the final subsec-

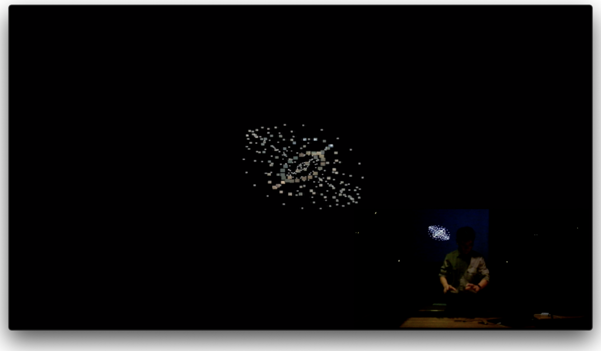


Figure A.11: *Handwriting* · *WuXing* the metal section distant view screenshot

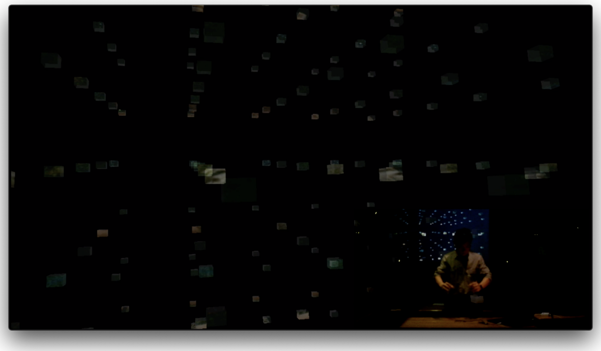


Figure A.12: *Handwriting* · *WuXing* the metal section “grid formation” screenshot

tion. In the final subsection, a new two-handed gesture is introduced, which serves three different purposes:

1. The distance between the hands can manipulate the camera’s position, therefore realising the control of shifting between different perspectives.
2. The distance between the hands can drive the sound source position as well as achieve conformance with the visual part.
3. When the distance between the hands is minimal, for example, the palms touching, it could be treated as a trigger event to progress the composition.

Additionally, no more new sound or visual elements will be added during this final subsection. Nevertheless, all those elements will be revealed via different perspectives. The camera view, from a distant to an extremely near position, from a detail to the whole shape, the break down and reorganisation of the cube formation, the different perspectives and formations of the cubes, and the synchronised sound transformation together illustrate the ultimate theme. Finally, all those cubes form a flat grid (Figure A.13), which reveals the secret that the texture on the cubes is the fragmented projection of a real-time captured image of reality; it also serves as the exit point of this composition’s world, thus achieving the conceptual closed loop between the real and virtual worlds.

Figure A.13: *Handwriting · WuXing* the metal section ending screenshot

A.2 *Intangible field*

A.2.1 First movement

The first section of the first movement has a duration of approximately a minute and aims to help the performers and audience establish the basic concept of this multidimensional interaction. Because musicians are the initiators of the interaction in this movement, all the instrument players are given the same musical material (Figure A.14) to improvise. Moreover, the instructions given to all musicians are also the same: stop playing if any actor moves too close to you. Correspondingly, all the actors try to freely and slowly move towards the instrument player. Hence, when one instrument player begins to play, the actor moves toward them. Nevertheless, when the actor is near to the player, that player will stop; then, the actor will seek and move towards another player who is improvising. Moreover, these interlocking interactions also leave space between real and virtual elements. In the procedural sound part, the “sound control field” will be triggered by any actor who enters the field, which enhances that interaction. As for the visual part, only the real-time capture of images of the moving actors is liquefied to smoke-like trace images. These procedural visual elements serve as a direct approach to presenting a real-time human–machine interaction.

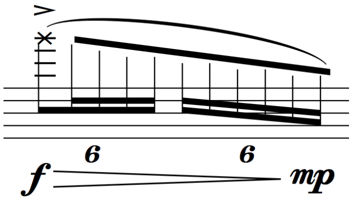


Figure A.14: *Intangible Field* A1 Music Material

Next, the actors will naturally lead the performance to the second section, which will continue for approximately a minute. Both the actors and players will continue to follow the same instruction as in the first section, but the actors gradually increase their moving speed. This transition also encourages the instrument player to increase their tempo to match the actor's movement. Moreover, the instrument players are given three new musical materials (Figure A.15), shorter variants of the first movement's material. After the actor's moving speed and the music tempo reach their peaks, the actors and instrument players experience breathlessness. Then, as the actors step back to their initial position and lie down, the player plays a fixed phrase that consists of a long note ranging from fast vibrato to normal. Finally, the trace images gradually dissolve to black, and the procedure sound slowly diffuses, indicating the end of the first movement.

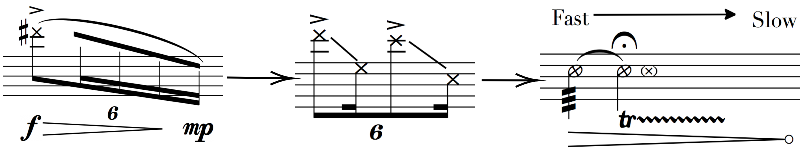


Figure A.15: *Intangible Field* A2 Music Material

A.2.2 Second movement

The second movement starts with a relatively short section of a solo violin performance. The violinist is given the music material (Figure A.16) with a fixed technique, dynamic and tempo change, but can choose any pitch. Meanwhile, the violinist is required to move around the initial position clockwise. This short section aims to smoothly connect the first and second movements and help the audience understand the interactive relationship between the performer's movement and the generated sound. Additionally, the system tracking the violinist's position will be synchronised to a smoke-like trace generator's position in the visual part to demonstrate this interactive relationship further. On the actors' side, they need to stand up gradually, and then look in the direction of the sound source, but do not move their position. These actions help them prepare for the next section and draw everyone's attention to the instrument player.

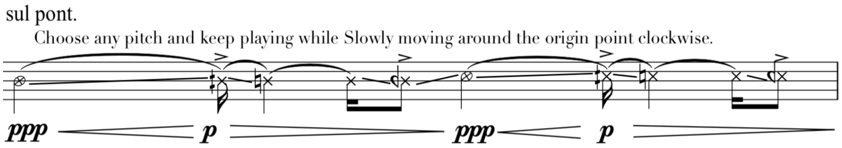

sul pont.
Choose any pitch and keep playing while Slowly moving around the origin point clockwise.
ppp
p
ppp
p

Figure A.16: *Intangible Field* B1 Music Material

After the violinist starts to repeat the music, other instrument players join the performance with new musical material (Figure A.17), which is the signal for the actors to begin their next performance, in which they move towards any sound source, but with caution and whilst maintaining a distance and slowly wandering around it.

Next, the third section of the second movement starts immediately after the trumpet player finishes his last musical material. The instruction for musicians in this section is the opposite of that in the first section of the first movement, which is that instrument players should only play when an actor is near to and interacting with them. This shift also means the initiator of the interaction switches to the actors. Additionally, the musicians are given three

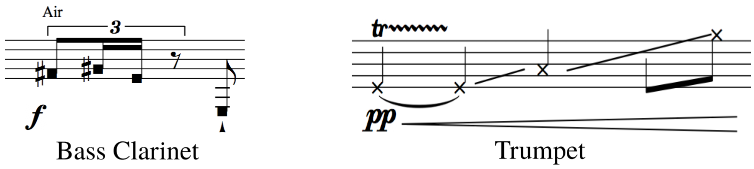


Figure A.17: *Intangible Field* B2 Music Material

different musical materials (Figure A.18). Nevertheless, they need to choose the musical materials based on the actors' actions. For example, they should play the initial material when an actor starts interacting with them or use fade-out material when an actor moves away. However, the actors' performance instruction is similar to the first movement in that they are required to move near to the musicians and interact with them, as if caring for a "plant" whilst it grows and then absorbing the "energy" from it. Then, when the actor feels they have enough "energy", they can move on to another musician and repeat this performance.

The procedural elements in this section are the visualisation and sonification of the abstract concept and the interactions. In the sound part, when the two performers are inside the same "sound control field", the transformed sound will become more evident as they move closer to each other. In the visual part, each musician's position synchronises to a smoke machine but with different colours, and the amount of smoke generated is linked to their music, for example, the volume and pitch variation. Additionally, each actor synchronises to an obstacle that interferes with the diffusion of the smoke. This visualisation aims to vividly represent the abstract concept and the interaction of these virtual objects to project an image of reality.

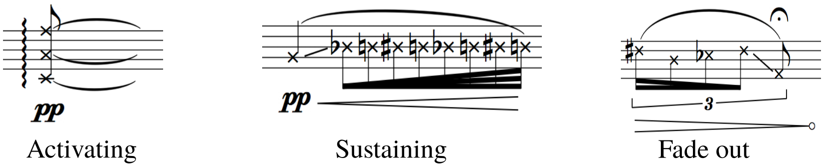


Figure A.18: *Intangible Field* B3 Music Material

After "the energy floating all over the space", the musician gradually increases their tempo and reduces the resting time between each musical material, and the "smoke" fills the area. Meanwhile, the actors can freely improvise during this time, moving around and enjoying the energy; this also indicates the end of the second movement.

A.2.3 Third movement

The previous section lasts approximately one minute and gradually enters the first section of the third movement, which has two significant changes. On the one hand, all the instrument players start random performances and move around to create chaos. On the other hand, the actors need to catch the musicians who are causing chaos and lead them back to a normal state. More specifically, when an actor catches a musician and tries to fix their performance, the musician will stop moving and play the previous musical materials. However, if the actor moves away, the musician will start moving around and causing chaos again.

After around 30 seconds, the musicians will start to increase their tempo and volume to create more chaos. With the increase in tempo, the actors need to move faster, creating an image of "smoke" of different colours mixed together by their movement. Next, when the actors cannot follow the music, they need to return to the centre position. Furthermore, all the musicians will suddenly stop playing and slowly move back to the initial position. Meanwhile, the picture full of chaotic "smoke" also fades into a liquefied real-world image, and the procedural sound fades away.

The following section starts after everyone moves back to the initial position. All the performers then follow the new instruction. The instrument players are asked not to play simultaneously and try to move close to an actor slowly, then use music to scare away the actor. Furthermore, the three musical materials (Figure A.19) that can be chosen are relatively short, with robust and sudden accent attacks. Correspondingly, the actors are provided with two instructions. Firstly, they should display anxious behaviour when someone is near to them. Secondly, they should run to another field if someone scares them. Thus, these new instructions also switch the role of the interaction initiator back to the instrument player side.

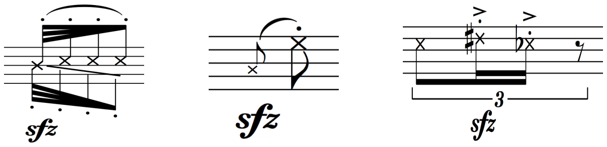


Figure A.19: *Intangible Field* C3 Music Material

The previous section eventually allows the actors to move between different areas quickly. However, the following instruction given to the instrument players is that every two players form a pair and use new music materials (Figure A.20) to build a "sound wall" (Figure A.21) to block the actors' movements. For example, if an actor tries to pass between the violin and the trumpet player, two players should play together to force him to change direction. The actor will then try to escape. However, if actors encounter a "wall", they will first try to break it but will be bounced back and discover it is unbreakable. Additionally, the procedural sound generated by the system' s deformation and stretching of the instrument's sound also strengthens the concept of the "sound wall". All the movements and actions of the performers will also cause the liquefied real-life image to become muddier, to express an increasingly depressing atmosphere.

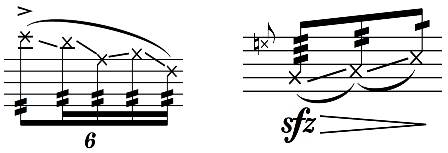


Figure A.20: *Intangible Field* C4 Music Material

Finally, all instrument players slowly move towards the centre, and the three "sound walls" form a triangular cage that compresses the actors' activity space. Additionally, the closer they are to one another, the faster and louder they will play. In contrast, the actors still try to escape but have nowhere to run, struggle, and are finally trapped in the centre and feel suffocated when the activity space is compressed to its limit. Eventually, when the suffocating atmosphere fills the room, everyone will suddenly stop at the same time, ending the whole piece.

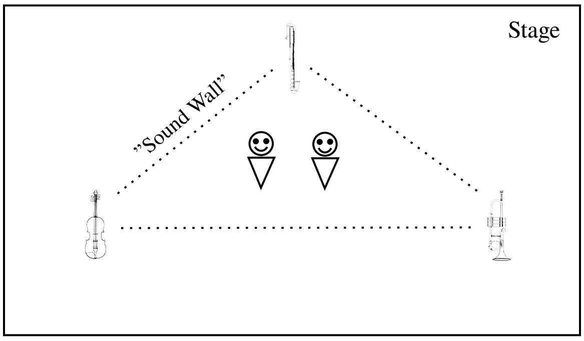


Figure A.21: "Sound Wall" diagram

A.3 *Audio Game - Music Force*

A.3.1 Cutscene 1

The introductory cutscene section is the beginning section of the whole composition. This cutscene section is a fixed-sequence animation lasting approximately 1 minute and 30 seconds, which is used to introduce the role of the game and lead the musicians and audience into the game space. Since the music score for the cutscene section is based on the time marker rather than regular rhythm and bar markers, the musicians can follow the sequence animation to perform together.

The first scene of the animation sequence starts from the outer space scene, and the camera gradually moves close to the "musical core". Then, a black portal emerges unexpectedly, and an energy body rushes out, followed by three sequential steps:

1. The "energy body" wanders around the "musical core" and uses the black energy to take it away.
2. The "energy body" leaves this space through the black portal.
3. Suddenly, a small musical energy body escapes from the black portal and becomes a "musical spacecraft". To chase the "musical core", the "musical spacecraft", which speeds up and enters the portal without hesitation.

Finally, the camera follows the "musical spacecraft" entering the portal, and the light fades to black, demonstrating the end of the first cutscene.

A.3.2 Game level 1

After the introductory cutscene section, the camera view switches to the side-view camera angle, and the first game level section starts. Because the first game level section aims to help the musicians to become familiar with the control mechanism and introduce the interactive relationship to the audience, the first challenge is to control the "musical spacecraft" movement to avoid "dangerous areas" and reach the destination. Only two musicians in charge of the spacecraft's movement are given the musical material (Figure A.22) to overcome this challenge. The musicians can use these musical materials for improvisation, and the system recognises the pitch alteration direction to control the movement of the spacecraft. For instance, if the musician who controls the up and down movement chooses to play an ascending scale, the "musical spacecraft" will move upward and vice versa.

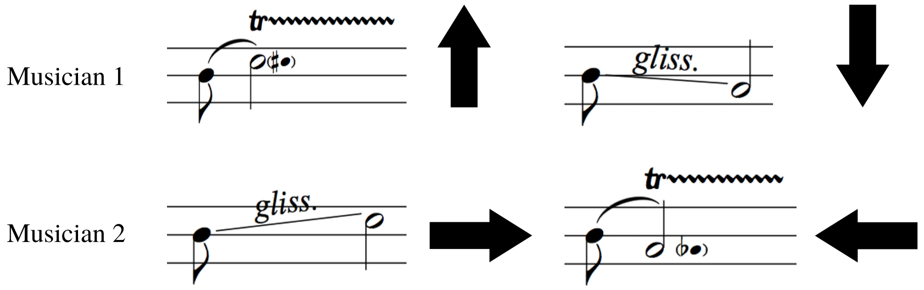


Figure A.22: *Audio Game - Music Force* Movement music material

During the first game level (Figure A.23), a staff-like background is continuously generated at the left side of the view and moves to the left at a constant speed, then disappears after moving out of view. Moreover, the "musical spacecraft" initial position is at the right side of the view. Thus, the combination creates an illusion that the "musical spacecraft" is constantly moving to the right side. In addition, two kinds of "dangerous areas", one fire area and one electric grid area, randomly appear on the right side out of view and move towards the left side; then, the musicians control the up, down, left and right movements of the "music spacecraft" to avoid them. However, when the "musical spacecraft" cannot avoid these "dangerous areas" and collides with them, it causes damage to the "musical spacecraft". The system produces corresponding sound as sonic feedback. In the meantime, as visual feedback, the staff-like background will collapse, and the player's view will falter.

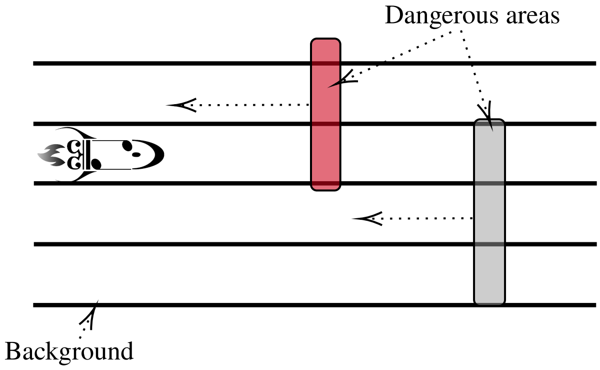


Figure A.23: *Audio Game - Music Force* game level 1 view figure

A.3.3 Cutscene 2

After the "musical spacecraft" passes through dangerous areas to reach the destination, the performance reaches the second cutscene section. The second cutscene section is approximately 20 seconds, introducing the attack ability to the "musical spacecraft". During the second cutscene, the "musical spacecraft" meets and absorbs another "musical energy body", which is a metaphor for acquiring a new ability. Then, the character moves on to the next game level.

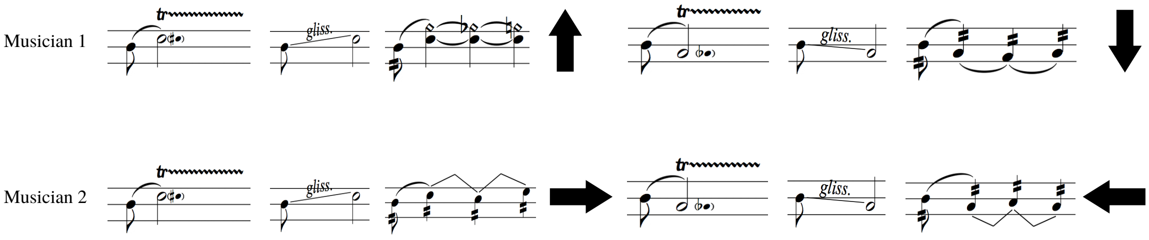


Figure A.24: *Audio Game - Music Force* game level 2 "Movement" music material

A.3.4 Game level 2

The second game level has the same background scene as the first game level. However, the challenge is entirely different, in that the players need to form two groups to control the "musical spacecraft" to destroy moving enemies and collect musical note energy:

1. The two musicians who control the spacecraft movement are given similar musical materials (Figure A.24) to in the first game level to extend the music expressiveness.
2. The third musician also performs, improvising based on two new musical materials (Figure A.25) to control the "shooting".

Since the system can recognise a more significant pitch interval change to trigger the "shooting" event, the "shooting" materials start from a long soft note and jump up and down to a short, strong note. This musical material is also in line with the concept of aiming and shooting in the game.

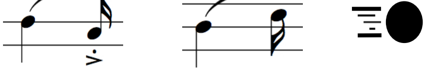


Figure A.25: *Audio Game - Music Force* game level 2 "Shooting" music material

Moreover, as Figure A.26 shows, the enemies randomly appear from the left side of the view and move around to avoid the player's attack. In contrast, the "musical spacecraft" also needs to keep moving to avoid the enemies and shoot at the right times to destroy them. When an enemy has been destroyed, it will spawn musical note energy in that position. In addition, the system generates a sonic response to the spawn, movement and destruction of the enemy behaviour. Additionally, the “musical spacecraft” must move close to the musical note energy to collect it. When the musical note energy has been collected, it will follow the spacecraft around as the visual feedback. The second game level is finished when enough musical note energy has been collected, and the third cutscene section starts.

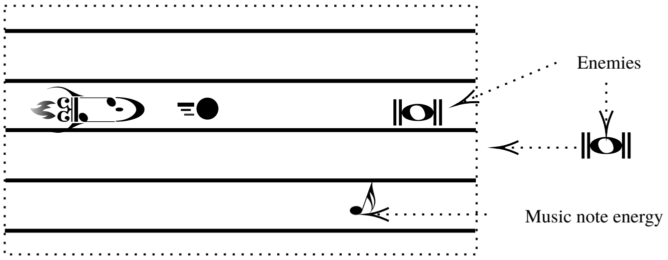


Figure A.26: *Audio Game - Music Force* game level 2 view figure

A.3.5 Cutscene 3

The third cutscene section is approximately 20 seconds; it aims to promote the development of the game story and connect to the next game level. In the cutscene, the collected musical energy fuses with the spacecraft. Then, the appearance of the “musical spacecraft” upgrades from a simple to a more spectacular ship (Figure A.27).

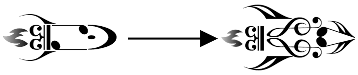


Figure A.27: *Audio Game - Music Force* “Music spacecraft” upgrade

A.3.6 Game level 3

As Figure A.28 displays, the third game level is similar to the second game level in terms of the scene and challenge, but with one significant difference: introducing a new enemy to increase game difficulty. The new type of enemy has a different appearance composed of three musical score elements, one of which will be destroyed when hit once by a player, but it needs to be hit by a player three times to be annihilated. Furthermore, it can fight back, shooting a “red sharp” sign as a bullet to damage players’ “musical spacecraft”. Hence, the musicians need to perform more actively to avoid the enemies and bullets and destroy the enemies to gain musical note energy. Furthermore, as the number of enemies increases, the difficulty of the game and the intensity of the performance also increase.

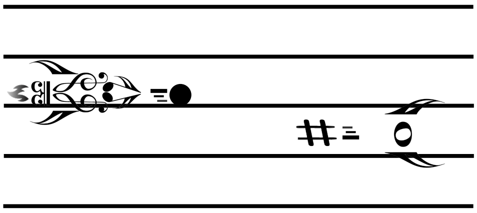


Figure A.28: *Audio Game - Music Force* game level 3 view figure

A.3.7 Cutscene 4

After the third game level, the last cutscene will introduce the final enemy, which is the energy body from the first cutscene utilising the “musical core” it took to transform into another “dark musical spacecraft” (Figure A.29). At this stage, the entire game model has been revealed, and the musicians and audience have mentally prepared to meet the final challenge.

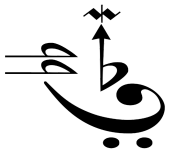


Figure A.29: *Audio Game - Music Force* “dark music spacecraft”

A.3.8 Game level 4

The final game level aims to defeat the “dark musical spacecraft” to retrieve the “musical core”. The difficulty of this level is the highest of all the game levels regarding two aspects:

1. The “dark musical spacecraft” has three phases. Therefore, the players need to hit it ten times for each phase to destroy it.
2. The enemy’s AI will continue to increase according to its phase, that is, the movement frequency to avoid the player’s attack, and the attack frequency will increase phase by phase.

Additionally, the appearance of the “dark musical spacecraft” will change to indicate its current phase instead of the regular graph widget in the view. Moreover, the system will sonify all the actions of the “dark musical spacecraft”, including movement, shooting, being hit and being destroyed. As a result, the musicians’ performances and the procedure sound become more intense as the battle becomes more intense. Eventually, with the explosion of the “dark musical spacecraft”, the “musical core” is released; the “musical spacecraft” takes it and moves towards deep space, thus ending the performance.

A.4 *Sound | Figuration*

A.4.1 Visual scene V1

The first visual scene, $V1$, includes the first two sound sections, A and B, and lasts a total of 5 minutes and 30 seconds. The first visual scene applies a basic algorithm in which each piano note sound will trigger the visual system to generate a node in the virtual space. More importantly, the newly generated node’s position is determined by the previous position and follows three rules:

1. The new node can only be placed in six positions perpendicular to the previous node: above, below, left, right, front and rear.
2. The new node's chosen direction should not be at the same axis as the previous time.
3. The distance between the new node and the previous one is determined by the interval relations between the current piano note and the previous piano note; this means the distance between nodes is more significant if the two notes' interval relations are farther apart.

For example, as Figure A.30 shows, if the previous node is placed at the right of the one before the previous one, the new node can only choose the other four directions on the Y and Z axes.

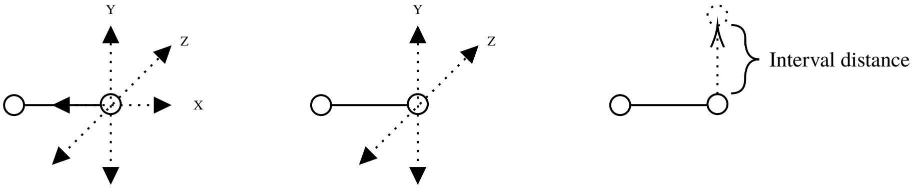


Figure A.30: Visual scene $V1$ generating rules example

Additionally, the new node will move from the previous to the new position and generate a line to connect each node. As a result, the visual part starts from nothingness then gradually constructs complex node geometrical networks that visualise the music development in an abstract manner.

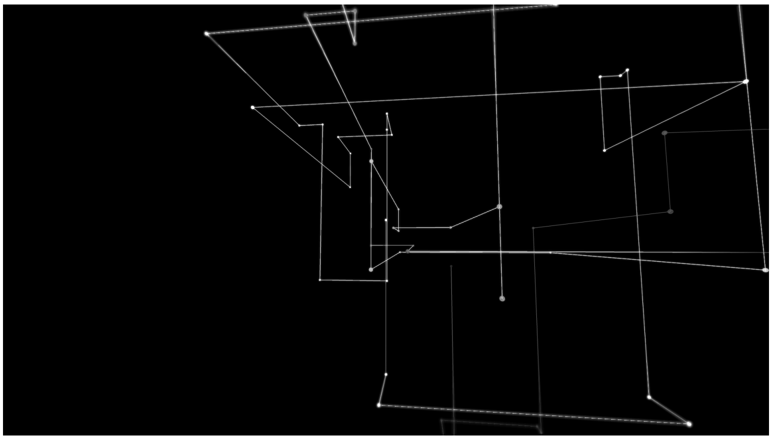


Figure A.31: Visual scene $V1$ Screenshot

A.4.2 Sound sections A and B

As for the sound part, the A and B sections' fundamental materials share the same pitch class in the generating rules, but with different note density and temperature. To begin with, the A section lasts about 3 minutes and 40 seconds, and aims to gradually lead the audience to the compositional space. Therefore, the primary musical material-generating rules are set to one note per second and 0.8 for the temperature. The resulting material has a relatively slow rhythm, and the interval relationship is comparatively close and coherent. Moreover, with the combination of the spectrum filter and the reverb module, the mid–high- and low-frequency bands of the piano sound are over-amplified and extended to create a boundless sense of space.

After a transition section of about 5 seconds, the B section starts with a much denser rhythm and high pitch repetition rate material generated in the Performer RNN with eight notes per second and 0.5 for the temperature. The B section lasts about 1 minute and 50 seconds, increasing the overall tension through fast-flowing music and switching views in the space. Additionally, the electron part gradually detaches from the piano sound in timing and harmony to build up the multilayer sound.

A.4.3 Visual scene V2

Once the view zooms in on the central node, the second visual scene, $V2$, appears over the $V1$ section. Because the $V2$ starts with one straight line, which coincides with the last scene of the $V1$, the transition of the two visual scenes is barely perceivable. The $V2$ scene contains the C and D sound sections and lasts 3 minutes and 25 seconds. The new generating algorithm for the $V2$ scene utilises the piano sound to generate multiple curvilinear forms composed of points. The algorithm takes the piano sound amplitude and note interval distance as the increment factor mixed with the noise factor to produce the particle position using the sine function. Moreover, the continuously developed curvilinear forms (Figure A.32) create a noticeable contrast to the rectangular networks in the $V1$ scene.

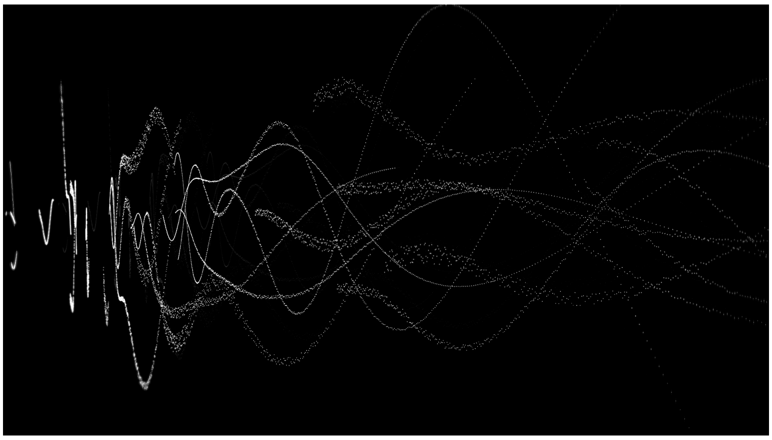


Figure A.32: Visual scene $V2$ First part screenshot

A.4.4 Sound sections C and D

During the visual transition from $V1$ to $V2$, section C also starts and serves as a connection section in the musical part. The C section lasts about one minute, and the music shifts from a smooth to wavy movement. Furthermore, the generating rules are set to four notes per second and one for the temperature to obtain the piano material with a wavy motion.

Hence, the source material generated from the compositional tool has a low pitch repetition rate and rapid flow between a wide range of piano notes. Additionally, the granular sound is produced by the live audio processing module from the live piano sound, which also enhances the wavy movement and the consistency of the visual. Once the composition reaches the final part of the C section, the sound expands both harmonically and spatially due to the audio processing modules. In the meantime, the multiple curvilinear forms fission again and compose a complex symmetric swirling image (Figure A.33), indicating the start of the D section.

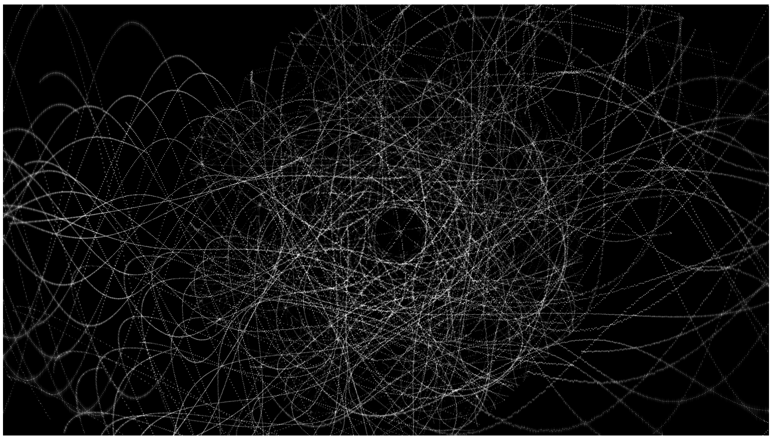


Figure A.33: Visual scene $V2$ Symmetric swirling image screenshot

The duration of the D section is about two minutes. In the sound part, the piano material increases the pitch range to form the long wavy lines of motion. Meanwhile, the over-dense and high-pitch granular transformation module exaggerates the high-frequency piano sounds, creating a symmetric sense of the sound. In the visual part, the camera view from a far distant panorama angle slowly moves closer to the centre. Thus, the image changes from the overall swirling image to details of the flowing particle. Then, after reaching the final part of the D section (Bar 70), the piano and the electric sound fuse together, extending the entire image to the audience's listening space. Finally, the sound part simultaneously flows and evolves with the view back to the distance, which reveals the entire swirling form's contraction and expansion process; then, the system will hold its current status and wait for the pianist to be ready to start the next section together.

A.4.5 Sound section E and Visual scene V3

The last section, section E, lasts about three minutes and can be divided into two parts:

1. The beginning part of the E section (Bar 72) serves as the transition and introduction, in which the arpeggio chords unveil the pitch class used to generate the instrumental material. To be more specific, once the pianist starts to play the first chord of section E, the system will follow this action chord by chord to reduce the density of the granular sound, then slowly reduce the frequency of the chosen part. At the same time, these chord sounds will be recorded by the system and kept as the source sound ma-

terial for later parts. In the visual part, to match the sound changes, the system will increase the degree of dispersion of the points that make up the curve. Hence, the shape of the swirling gradually becomes blurry. Additionally, the camera view gradually moves away from the swirling image to display the full picture of the transformation process.

2. The remaining part of section E serves as the reappearance and intensification of the theme and emotion of the whole work in three aspects:

    (a) The material for the piano part has similar properties to section A's material, which has a lower note density and relatively smooth pitch movement.

    (b) It fills in the sonic space with the combination of the chosen sound from two dimensions: one is the emptiness of harmony; the other is the vacuum in the rhythm, which are both created by the single piece of slow piano material.

    (c) Point cloud transformation is used to create a complementary visualisation of the music.

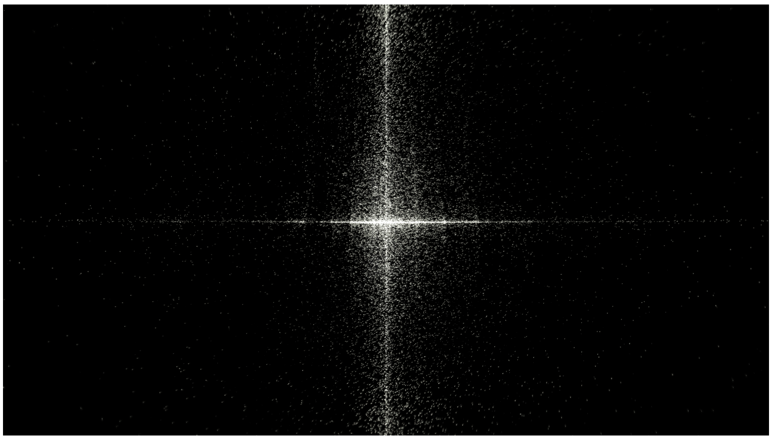


Figure A.34: Visual scene $V3$ point cloud screenshot

In order to achieve the goals mentioned above, the system utilises the recorded material from the beginning part of section E as the source material to drive the granular synthesiser module to produce polyphonic support sound layers from Bar 74. Correspondingly, the harmony structure changes as the piano material develops. In the meantime, the combination of the spectrum delay and the loop sampler module extracts and amplifies the high-frequency band of the live input piano sound. Hence, it not only increases the density and variation in the rhythm, but also extends the piano's timbre. However, the visualisation detaches from the music via the explosion of the point cloud, which switches to the $V3$ scene. Instead, it forms a nebula-like shape (Figure A.34) via the interior motion of the point cloud to complement the insufficient motion in the music. Finally, as the music's motion slowly comes to a stop, the camera view gradually moves extremely far away, till the nebula-like shape becomes a star among the galaxy (Figure A.35); then, the sound and image fade out, bringing the whole composition to a close.

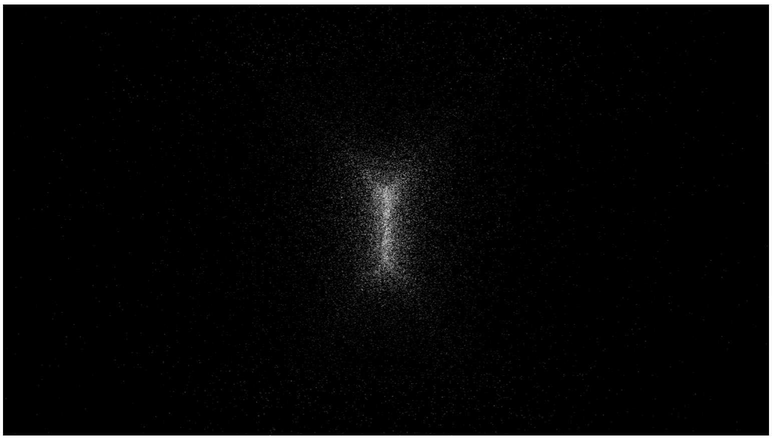


Figure A.35: Visual scene $V3$ Ending screenshot

A.5 *Metamorphosis*

A.5.1 Sound section A

The central theme for the A section is learning, which aims to establish the essential concept of the interaction relationship between the human performer and AI performers. The A section consists of two parts and lasts about four minutes; both the human and AI performers only utilise a striking action, which is the standard way of playing the Bianqing. During the first part, three performers successively perform a solo improvisation (Figure A.36) based on the previous performer's music with their own virtual musical instrument. More specifically, the solo performance mode has three steps:

1. The human performer initiates the performance and plays the virtual Bianqing, providing the first motive.
2. Then, the first AI performer follows the performance, developing the primary motive.
3. After the first AI performer is finished, the second AI performer will continue, further extending the motive based on the first AI performer's music.

Next, the attention goes back to the human performer, who starts a new round and provides a new motive based on the second AI performer's performance. After the solo performance mode loop has been performed twice, the performance enters the second part of the A section.

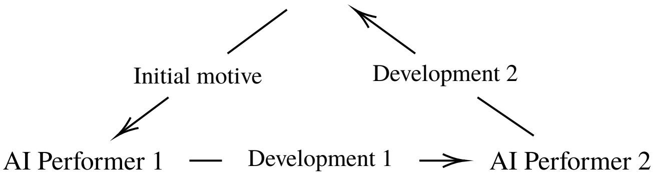


Figure A.36: *Metamorphosis* section A solo improvisation loop

The second part keeps the sequential performance mould, but extends the performance from solo to duo mode (Figure A.37), further unveiling the AI performers' musicianship. The performance of each segment is performed by two different players at the same time, one leading and one supporting, and likewise with three steps:

1. The performance starts with the human leading, and the second AI performer supports it.
2. After that, the first AI performer leads the performance, and the second AI performer supports it.
3. Then, the second AI performer takes the leading position, and the human performer supports it.

Moreover, by continuously increasing the temperature and note density in the AI performer's generating rules, the resulting music gradually varies from the original motif, which increases the variation in the music.

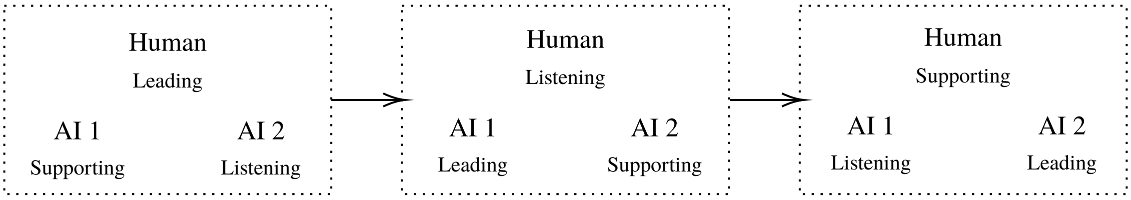


Figure A.37: *Metamorphosis* section A duo improvisation loop

As Figure A.38 shows, the first visual scene consists of three views of a single image: the three different camera angles camera of the same virtual space.

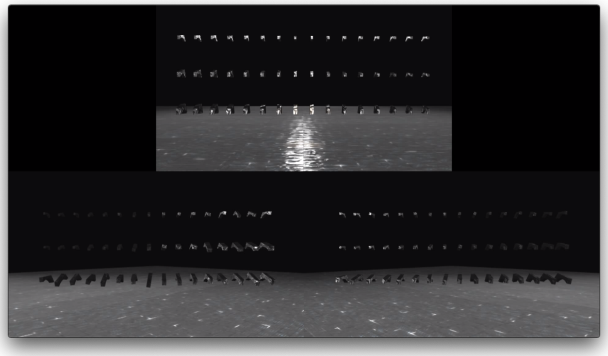


Figure A.38: *Metamorphosis* Initial screenshot

Three virtual Bianqings are placed at the three corners of an equilateral triangle shape with the origin as the centre point in the virtual space, as Figure A.39 indicates. Furthermore, the Bianqings face the centre point initially, ensuring that each camera can only catch the right opposite instrument.

Furthermore, as Figure A.38 indicates, these cameras images' final layout mimics a typical video conference screen layout: one is on the top, and the other two are on the bottom left

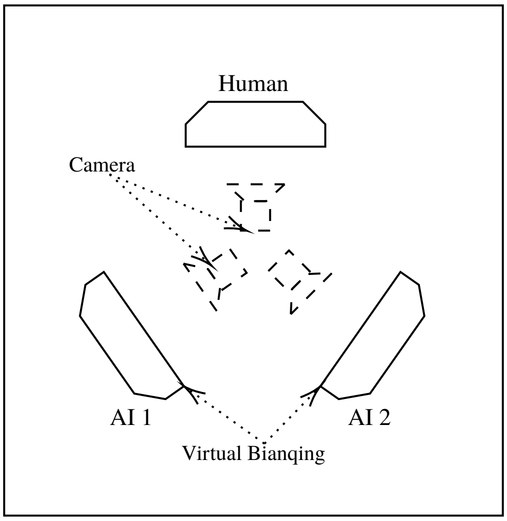


Figure A.39: *Metamorphosis* section A visual spatial diagram

and right. More importantly, each image size is changed based on each performer's amplitude level, which guides the audience's attention. For instance, when the first AI performer plays, they have a higher amplitude than other performers. As a result (Figure A.40), its image is scaled in, while the other two images are correspondingly scaled out, dynamically adjusting the layout. Additionally, when the virtual chimes are struck, they emit a particle, which is pulled by an attractor, then gradually form a particle cluster. During the A section, the cluster of particles moves to the front of the Bianqing being played, visually representing the leading performer's role exchange. Finally, once the performers have completed the duo performance loop twice and the cluster of particles are back on the human side, the piece progresses to sound section B.

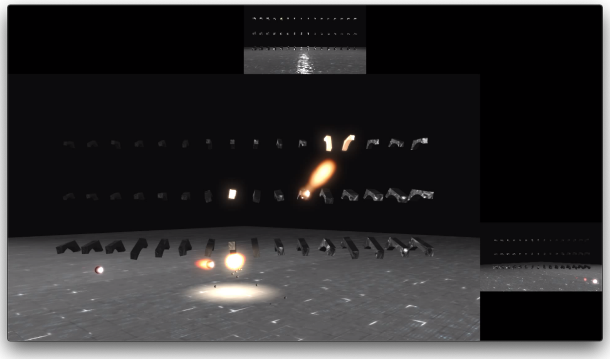


Figure A.40: *Metamorphosis* Section A camera image zoom in

A.5.2 Sound section B

The B section duration is about four minutes, in which the performance mode gradually changes from cooperation to confrontation. Firstly, the performance switches to the trio mode from the duo mode. Two AI performers lead the performance first and progressively increase the tempo to create more tension between the performers. Then, the human performer cannot follow the performance by simply striking but introducing new gesture control into the interactive mode to achieve the extended control of the instrument and sound.

In the meantime, several digital sound processing modules start to extend the sound of the basic instrument and generate more layers to correspond to the changes happening in the visual content.

To visualise the interactive relationship developments that happen during section B, as Figure A.41 shows, four primary changes apply to the virtual space, as follows:

1. The image size transformation stops at the beginning of the B section, then reverts back to the same size to provide a stable view of each camera angle.
2. To symbolise the timbre change, the virtual instruments will alter the formation if they are struck.
3. The formation of the human performer's instrument will transform into three concentric rings. Their rotation angle is bound to the gyroscope data of the controller, which shows the extended control mechanism in real time.
4. The position of all the cameras increasingly moves away from the instrument towards the centre, which not only allows a broader view to observe the deformed Bianqing, but also gradually reveals the virtual space environment.

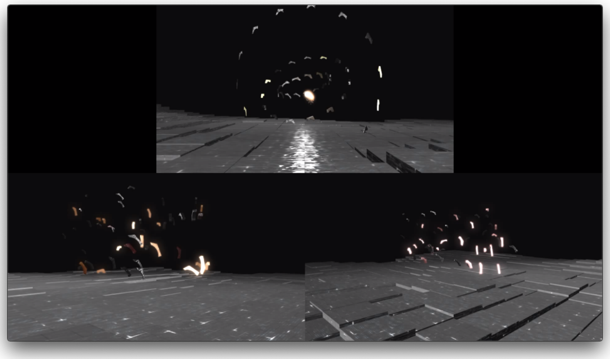


Figure A.41: *Metamorphosis* Section B screenshot

After AI performers gradually detach the control of the human performer, the solid ground in the virtual space starts to disintegrate into floating blocks, indicating the disintegration of the old relationship. Next, with the one strong striking action of the human performer, as Figure A.42 exhibits, the discrete ground's block spreads around to form a surrounding ring, which symbolises the starting point of the sound section C.

A.5.3 Sound section C

At the beginning of section C, the overall sound texture continues on from the previous section, but with two main differences:

1. The spatial panning control starts to link to the human performer's gesture control, and the rapidly changed gesture will add a sense of motion and chasing.

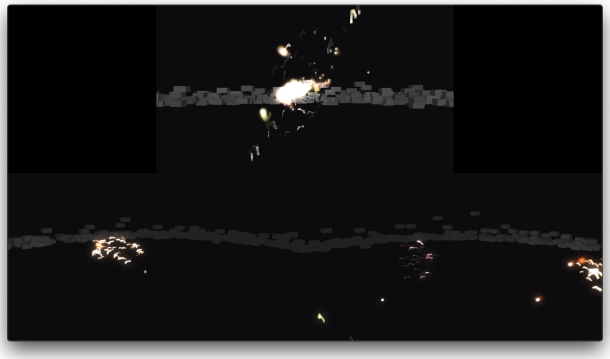


Figure A.42: *Metamorphosis* Section C beginning screenshot

2. The frequency structure of the extended sound layer is connected to the human performer's controller pointing direction, which further extends the interaction mode.

Furthermore, the position of each virtual instrument moves in accordance with the projection of the spatial motion of each sound source. However, two cameras for the AI performers are moved outside to the surrounding ring, and their lenses are still locked to each instrument, following the virtual instrument movement, which displays different angles of the virtual space and confrontation mode. After the surrounding ring separates and forms a cylinder, the camera moves back inside the cylinder to show the details of the internal motion. Next, the surrounding blocks separate again and form a sphere. With the gradual shrinking of the sphere, all the high-frequency sounds are slowly filtered as the consistency of the visual changes. Finally, the blocks block all the light (Figure A.43), and only a low roar remains; all the images are plunged into darkness, which concludes section C.

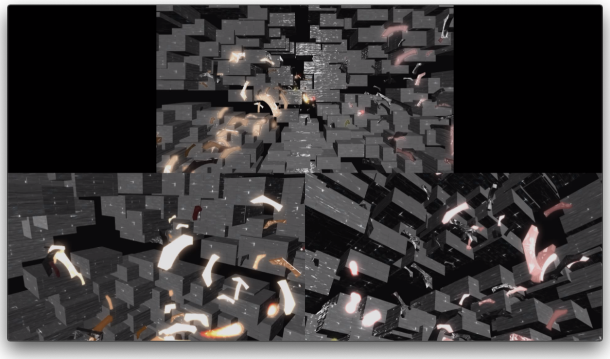


Figure A.43: *Metamorphosis* Section C Ending screenshot

A.5.4 Sound section D

The following section is section D, which takes about two minutes to complete a scene that presents a metaphor of breaking out of a shell and changes the image layout from three camera views to one, which serves as the bridge between the concrete and abstract worlds. As this is a transition section, section D starts from the end of the previous section's audio-visual scenes, which, via the combination of a dark image and low roaring sound, build a depressive environment. Next, with the heavy strikes of the human performer, accompanied by a crisp knocking sound, a crack is formed in the image to let a ray of light into the dark virtual space (Figure A.44). Then, after multiple strikes, more cracks are formed, and the low roar sound starts to fade out. Eventually, when the cracks merge, the virtual space gradually lights up, and the new synthesis sound fades in, together symbolising the rebirth of the performers; the image then turns to white, signifying that the transition section is finished.

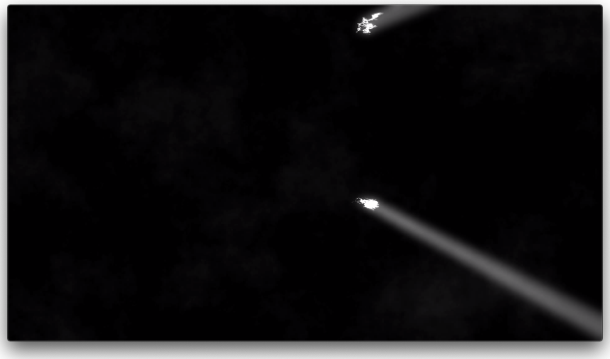


Figure A.44: *Metamorphosis* Section D “Crack” screenshot

A.5.5 Sound section E

The E section is about eight minutes in duration and consists of three parts that present the co-evolution between humans and AI through different interaction relationships and camera view angles. It starts with traceless cross-fades into section E, and sequentially presents the novel components and their relationships with the current virtual space via a distant top view. On the visual side, as Figure A.45 shows, the first significant contrast between the previous virtual space and the current space is that the background turns pure white.

Additionally, the form of all AI performers is abstract to the cluster of artificially intelligent agents, which are made up of lines. The AI agents’ basic motion rules are the modified implementation of the Boids algorithm [90]. In addition to applying three principal rules: separation, alignment and cohesion, the human performer also leads the AI agents’ motions by pointing in any direction, which adds another force to the algorithm to affect the agent’s action.

The AI performers begin to drive the spectrum-based synthesizer’s centre frequency structure on the sonic side. More precisely, the AI agents’ cluster motion also affects the spectrum structure. For instance, if the separation force between agents is greater than the cohesion force, the distance between them will become more prominent and form a massive cluster; then, the output sound will consist of richer frequencies. In contrast, if the cohesion force between agents is greater than the separation force, the distance between them will become smaller, creating a single line; then, in the end, the output sound will consist of fewer different frequencies. As a result, each synthesizer thread produces a subtle sound consistent with each AI agent’s action. Additionally, the position of each agent is projected

Figure A.45: *Metamorphosis* Section E beginning screenshot

onto the spatial ambisonic system, which alters the spatial position of the output sound accordingly. Accordingly, as all agents together build a vibrant cluster form (Figure A.46), the sound of each agent mixes to produce a dynamic sonification of the cluster. Moreover, the human performer combines gestures with minimalist audiovisual content to reveal the new interaction relationship.

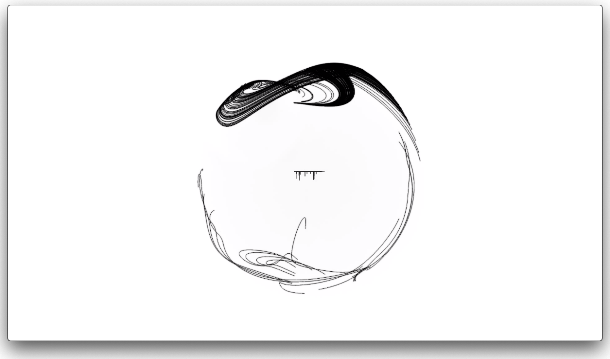


Figure A.46: *Metamorphosis* Section E “Cluster” screenshot

Next, an additional interaction mechanism between the human and AI performers begins to unveil after the camera gradually moves down. The new interaction mechanism is that the human body feature point detection, which aims to realise the human performer’s projection in the virtual space, can touch the intelligent agents. More specifically, the human performer’s image captures and extracts the body feature points first, and projects the feature points onto the centre of the virtual space. Lastly, when the distance between any feature point and AI agent is smaller than a certain threshold, it will trigger two response events:

1. As Figure A.47 shows, a line to connect the feature point and the AI agent is generated as the visual response.
2. The granular sound synthesizer module is activated and applied to the output sound to produce a variant sound as the sonic response.

Consequently, this new interaction mechanism signifies the closer interconnection between virtuality and reality.

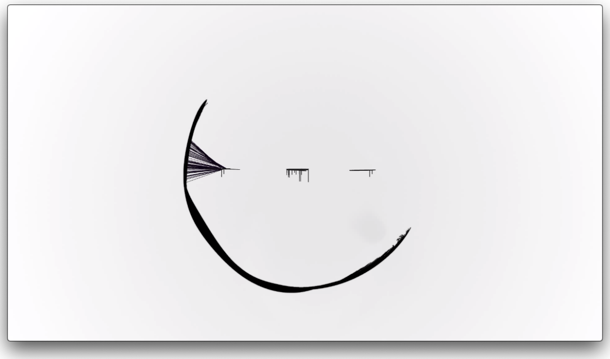


Figure A.47: *Metamorphosis* Section E “Connect” screenshot

Meanwhile, to enhance the sense of depth in the virtual space, when the AI agents make contact with the ground, it will create a ripple in the contact position. Furthermore, it will trigger the modified Bianqing synthesizer to produce another sonic response, which thus also enriches the sound. Additionally, this internal interaction between virtual objects and space foreshadows the external interaction between human performers and virtual space in the next section.

Next, when approaching the end of the introduction part, the camera’s position shifts from the top to the front (Figure A.48), which not only exhibits the three-dimensional virtual world but also fully reveals the human performer’s projection form, which is also composed of lines.

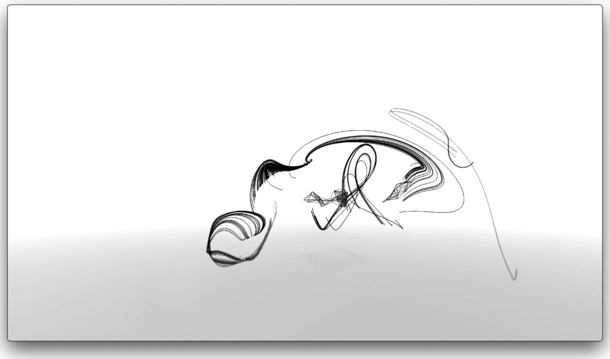


Figure A.48: *Metamorphosis* Section E front view

After the introduction part, the development part aims to present the process of the AI agent cluster’s transformation and the expansion of the interaction relationship from various angles. At the beginning of the development part, the cluster suddenly spreads out to form a sparse umbrella shape surrounding the human performer’s shape, as Figure A.49 illustrates. This step also drives the synthesizer to expand its spectrum range to produce a sound that contains an enriched frequency and opens up a more expansive sound space.

Then, each agent’s action starts gradually shift from smooth to unstable. Correspondingly, the camera starts to lock on the cluster’s centre and moves from the front to the rear. This camera movement guides the audience’s attention, focusing on the change that happens to the AI agents and creating camera moving space for the following part. After the camera

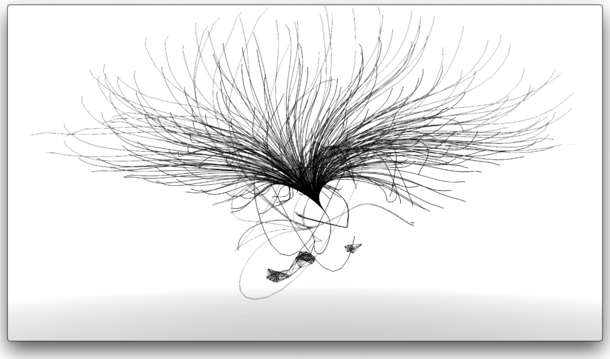


Figure A.49: *Metamorphosis* Section E "cluster" shape transformation

arrives at the rear position and unlocks it, the human performer will start to apply the new gesture to sequentially progress the composition. The new gesture is an intense downward striking action instead of the previous section's forward striking action. Furthermore, the downward striking action produces an audiovisual response similar to the response events when the agent makes contact with the ground. Accordingly, this human–machine gesture and its response logically continue the interaction relationship between objects in the virtual space, symbolising the switch in the learning relationships between the machine and human.

Following the first downward striking action (Figure A.50), the cluster swiftly gathers above the human figure. Subsequently, after the second striking action, the overall cluster's movement mimics the human hand movement. Next, after the third striking action, the camera moves back to the front, and the cluster's internal motion shifts to a chaotic state. As the camera arrives at the front position, the human performer will carry out the fourth striking action, and the cluster becomes further compressed. Then, with the fifth striking action, the cluster will gradually descend, entangle, and eventually merge with the human body figure and move on to the final part.

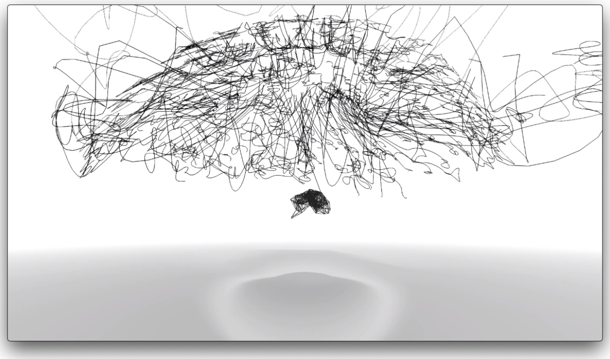


Figure A.50: *Metamorphosis* Section E first "Striking"

For continuity, the last part utilises the same downward striking gesture to progress the composition and aims to present the re-established collaborative relationship between AI and humans. At the start of the last part, the interval between the sixth and seventh striking actions is relatively short, in contrast to the previous part. After these two quick striking actions, the cluster quickly pulls away from the human body figure and starts to circle the centre stably, as Figure A.51 shows. In the meantime, because the sound synthesizer's spectrum structure is linked to each agent's stats, the timbre of the sound changes as the cluster motion changes, which gradually shifts from a chaotic to a harmonious state. More importantly, the base frequency of the synthesizer starts to bind with the human performer's gesture and the spatial position of the output sound. The human performer then starts to carry out an arm gesture that draws a circle in the air, which has the same motion feature as the agent's movement. As a result, by producing a sound that reflects the combination of the human performer's gesture and the AI agent's motion, the system dynamically sonifies the harmonious state of the human and AI performers.

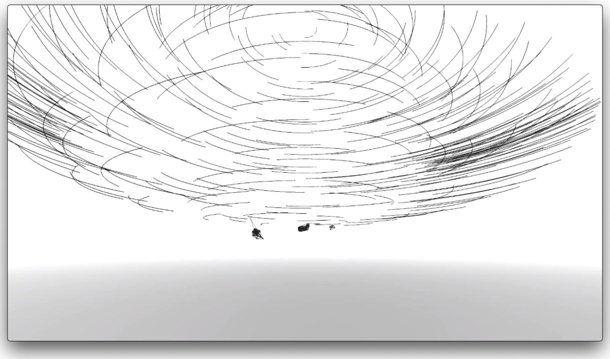


Figure A.51: *Metamorphosis* Section E last part beginning screenshot

Next, with the eighth striking action, each agent's speed increases, leading to their trail beginning to extend and the final form of an apparent orbit. Furthermore, the frequency of the human performer's gesture movement starts to increase in accordance with the AI agent's changes. After the speed approaches its peak, the camera moves closer, and the human performer starts to touch the cluster again. Again, this touching gesture triggers the system to produce an additional layer of sound and signifies the deeper interaction between the real and virtual space. Finally, following the ninth striking action, the cluster immediately unites into a single waving line at the head position of the human figure (Figure A.52). In the meantime, the sound converges into a single-frequency bin. As the piece ends, the camera gradually moves away; then, both the sound and image fade out eventually.

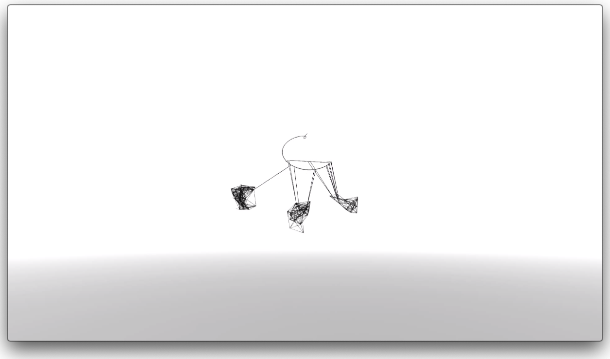


Figure A.52: *Metamorphosis* Section E ending screenshot

Appendix B

Supporting Material

1. **Conversation in the cloud** ca. 14’48
   For one human musician and one AI musician

   - Program Notes: *Conversation in the cloud* is a live multimedia composition by one human musician and one AI musician. The conversation between the two musicians starts at the intersection of reality and virtuality via music. Then, both musicians will drive their limits through improvisation founded on each other’s music during the live performance. Finally, the two worlds slowly merge as the conversation deepens. The AI musician is a comprehensive system that involves multiple machine learning techniques to enhance its machine musicianship, such as DNN and human body pose estimation. Thus, the combination of live multimedia and performances from the two musicians unveils a multidimensional music conversation.
   - Premiere on November 27, 2021 at the ZKM Karlsruhe Media Center, Germany.