Master's Thesis

Multimodal Looper: Interactive Visual Music through Gestures

Pelin Kiliboz

Supervisor: Cumhur Erkut
Medialogy, Aalborg University
Department of Architecture, Design and Media Technology
September 2023

Abstract

This thesis presents the first prototype of a multimodal looper, an embodied interface that enables musical improvisation through connecting body movements to audiovisual forms. With the aim of building an intuitive system, the initial research focused on music cognition through an embodied paradigm, where sound perception is integrated with visual and kinesthetic modalities; as well as widely observed crossmodal correspondences. This led to the development of a live-looping system with multimodal objects: sound fragments that are gesturally activated and visually represented according to their dynamics, motion trajectories, pitch and texture. The design and implementation focused on the development of the various modules of the live-looping system. The evaluation aimed to determine if the system effectively established crossmodal correspondences for an integrated multisensory experience through split testing method, where the participants experimented with a variation of the system with randomized crossmodal mappings along with the original version. Although the results revealed no significant difference between the two conditions, as well as a perplexing user experience, the qualitative findings proved to be quite valuable for future iterations.

Contents

Chapter 1

Introduction

In the early 1940's, Pierre Schaeffer started experimenting with recorded sound loops on phonograph discs, and later magnetic tape, becoming a pioneer of sampling and paving the way of electronic and experimental music. This exploration led to the development of sonic objects, a basic musical unit determined by putting the perceived qualities of sound to the forefront, aiming to become an alternative view to traditional music notation. Schaeffer and his collegues viewed sound fragments according to their coherence, dynamic features and trajectories in time and space; features that are closely related to those of body motion. As our sensory inputs and our perceptions can be regarded as being heavily intertwined and mediated through our bodies, a lot of research has also been done in the field of multisensory perception, drawing connections between kinesthetic, auditory and visual modalities from an embodied cognition perspective.

This premise opens a lot of possibilities for innovative musical exploration and interactive experiences within multimedia. Using the body as an instrument and representing the movements through sounds and visuals can provide new and immersive ways to create and collaborate. For example, Godøy suggests a triangular method for the use of gestures within musical composition where gestures act as a binding element between sounds and multimodal gesture images, suggesting that visual representations can enhance the understanding of sound and gesture relationships [1]. Furthermore, a number of researchers have looked into general tendencies around relating sensory inputs, which seem to show distinct similarities across ages and cultures [2]. This would suggest the possibility of an an intuitive framework whereby gestures are correlated with sounds and visuals and multimodal objects can be used as a unit within musical composition.

In this paper, I introduce the multimodal looper, an interactive installation I started developing as a creative tool for making visual music through gestures. The primary motivation of this system is to offer a multisensory experience of musical improvisation by utilizing intuitive crossmodal mappings in order to create an easily understood language of music and a novel creative experience.

In order to achieve an intuitive method, it is imperative to understand the interconnection of the kinesthetic, auditory and visual modalities and the manner in which individuals tend to relate them. Therefore, a review of the related literature on this topic constitutes Chapter 1 of this thesis; whereby cross-modal mappings of sound and music within an embodied paradigm, including Schaeffer's sonic objects, are investigated. The research is concluded with state of the art designs, methods and technologies that can be used within interactive systems for making visual music through gestures. Based on the conducted research, a final problem formulation is achieved, and the design requirements are established for an initial prototype: a multimodal live looping system. Chapter 2 outlines the design methods and decisions taken for each module of the looper. This is followed by Chapter 3 where the technical implementation of each module, namely gesture analysis, the music and visualization,

as well as the combination of these modules within a live looping system are described. Chapter 4 presents the evaluation of the prototype which aimed to test the system's effectiveness in creating an integrated multimodal experience through split testing method, statistical analysis, participant backgrounds, interviews and behavioural observations. The results and future directions are discussed within Chapter 6. The thesis is concluded by the final Chapter 7.

1.1 Initial Problem Statement

Can a generative system for making visual music with gestures provide an intuitive way to collaborate on a multimodal level and enhance creative experience by using inherent connections between kinesthetic, auditory and visual sensations?

1.2 Related Research

1.2.1 Embodied Cognition

With its roots in Merlau-Ponty's phenomenology, embodied cognition is a theoretical framework which seeks to explain the cognitive processes, actions, and overall understanding of the environment by individuals [3]. Opposed to the traditional cognitivist approach, which views the mind as a computer working with an input-output method, embodied cognition model believes that the involvement of the human body and its interaction with the surrounding world play a vital part in cognitive processes. Instead of passively receiving incoming information, cognition is understood as an active process in which goal-directed activities and sensory feedback build a continuous and dynamic pattern of sensorimotor activation[4]. Also called perception-action cycle, this can be regarded as a feedback loop between the person (with their body as a mediator) and their environment. It is an especially useful framework for emerging fields within HCI (Human Computer Interaction) which seek to combine physical and digital processes in order to become a seamless part of the human experience or a means of enhancing it.

1.2.1.1 Embodied Music Cognition

The concept of embodied cognition has also been applied to music research, with embodied music cognition being conceptualized in 2007 by Marc Leman in his book *Embodied Music Cognition and Mediation Technology.* Leman argues that the body is a natural intermediary between musical experience and physical environment [5]. Regarding the fact that listening to music is a subjective experience, Leman aims to establish a scientific theory that retains subjectivity, suggesting a second-person description. This viewpoint differs from first-person perspectives in that it considers experience as expressed rather than examined. It emphasizes the embodiment of human musical interactions through bodily participation (corporeal intentionality) over subjective thinking processes interpretation (cerebral intentionality) [5]. This perspective is critical for musical systems based on embodied cognition whereby the interaction is mediated through bodily movements. Moreover, in his book, Leman argues that there is a need for a transparent mediation technology that directly connects musical involvement to sound energy, such as methods to access music gesturally [5]. The upcoming sections provide perspectives and classification methods that can facilitate the endeavour of accessing sound and music through gestural means.

1.2.2 Gestures in Sound and Music

Numerous scholarly works and theoretical frameworks have been developed to explore the connection between gestures and music. Edited by Rolf Inge Godøy and Marc Leman, *Musical Gestures* is a

collection of essays highlighting key concepts and frameworks that can be used to understand how gestures relate to music through the lens of embodied cognition, viewing bodily movement as both a response to our experiences and an active contribution to our understanding of the world [6]. Various concepts and categorization methods from the book that are relevant to this paper are hereby presented.

1.2.2.1 Intermodality of Gestures

Gestures can have a diverse range of purposes and may be intertwined with other sensations, as individuals perceive the world around them and actively contribute to their environment through their bodies. According to Jensenius et al., the study of gesture may be categorized into three overarching perspectives: gesture as a means of communication, gesture for the purpose of control, and gesture as a metaphorical expression [7]. Communication involves gestures as vehicles of meaning in social interaction, while control involves gestures as elements of systems, such as interactive technologies. The key difference between these two definitions is highlighted as control largely focusing on manipulative gestures involving physical contact, while communication mainly revolving around empty-handed gestures [7]. Alternatively, viewing gesture as metaphor is used when gestures function as conceptual representations that project physical movement, sound, or other forms of experience onto cultural subjects [7]. This viewpoint is especially interesting in the way that it focuses on the intermodality of gestures and human perception in general. In particular, it is argued by Middleton that the comprehension of musical sounds is structured by processual forms that have resemblance to physical motions [8]. Similarly, Hatten suggests that gesture can be viewed as a cognitive construct evoked from auditory stimuli, arguing that musical gestures involve manipulating energy over time, based on gestural competence, which is influenced by physical and social factors [9]. In essence, gestures can be stated to serve a variety of purposes, including communication, control, and metaphorical expression; highlighting their intricate role in human interaction and cognition, as they connect with various sensory modalities and cultural dimensions.

1.2.2.2 Functions of Musical Gestures

In view of the aforementioned perspectives on gestures, Jensenius and colleagues define musical gestures as "an action pattern that produces music, is encoded in music, or is made in response to music", underlining the interconnectivity of music and gestures [7]. Jensenius et al. further categorize musical gestures according to their functional aspects, into four categories [7]:

- Communicative gestures, which facilitate communication between musicians or between musicians and the audience.
- Sound-producing gestures, which refer to the bodily movements that are essential for the production of sound.
- Sound-facilitating gestures, that aim to prevent tiredness and pain when producing sound such as playing an instrument.
- Sound-accompanying gestures, which refer to various bodily movements that can be performed in conjunction with music. They are not essential for sound production but can to a degree be synchronized with musical occurrences. Examples include dancing, marching, gesticulating, nodding the head, and similar actions.

Moreover, in his essay "Gestural Affordances of Musical Sound" within *Musical Gestures*, Godøy expands further on the interrelation of gestures and musical sound, specifically focusing on the *affordances* associated with this relationship. Popularized by psychologist James J. Gibson, the term affordances refer to the perceived or possible actions that an object, environment, or system proposes to a user depending on its design and attributes [10]. In this context affordances relate to how

listeners react to musical sound; namely how they extract movement-inducing signals from musical sound, as well as how they make meaning of what they hear by conjuring up visual representations of sound-related movement [11]. Godøy explains that the study of gestural affordances of musical sound is a two-way process where sound causes images of movement, and previously learned images of sound-related movement are projected onto sound, forming the perception-action loop [11]. This would mean that music perception is embodied in the way that it is directly linked to bodily experience, and multimodal in the way that it is perceived with the assistance of visual/kinematic images and effort/dynamics sensations [11].

1.2.3 Sonic Objects

The foundations of viewing music from a gestural perspective can be traced back to Pierre Schaeffer, who emerged as a pioneering figure in the 1950s and 1960s. Schaeffer and his colleagues developed the concept of sonic objects through their work in electroacoustic composition for *musique concrète* [12]. They recorded sound loops on phonograph discs to facilitate sound splicing before the invention of tape recorders. While listening to the fragmented sounds, they discovered that the listener's focus shifted from the initial anecdotal significations of sound fragments to the intrinsic, subtle qualities of the sound [13]. This shift was called "écoute réduite" (reduced listening), and it paved the way for an elaborate theory of sonic objects, as described in Schaeffer's *Traité des objets musicaux* and associated books [14]. With the starting point being the question "What do we hear now?", Schaeffer and his colleagues' approach involved conceptualizing the characteristics of sound fragments as shapes that correspond to basic action categories such as sustained, impulsive, iterative, flat, curved, steep, and so on [1]. This inventive framework, based on the transformation of sound fragments into recognizable shapes, served as a conceptual link between auditory experiences and gestural representations.

Schaeffer developed strategies to further define sonic objects. According to Schaeffer, any sound can be classified as a sonic object, given that it has a suitable duration (usually a few seconds) and can be perceived as a total and coherent entity [12]. What defines the beginning and end of a sound object, was then a matter of its perceived expression. Schaeffer proposed stress-articulation as a method for cutting sound objects at natural discontinuities in sounds, breaking up the continuum through distinct energetic events, resulting in short, clearly shaped objects [13].

Within Schaeffer's comprehensive framework, "typology" refers to the overall shape or envelope of a sound object, meaning how the sound changes over time considering its dynamic and timbre-related content. It can be seen as an initial classification of sound objects by considering their perceptually most prominent features [12]. "Morphology" focuses on the actions within these shapes, such as internal changes and fluctuations, and is meant to be a complimenting perspective, as there are interdependant features between the typological and morphological classifications [12]. For the purposes of this thesis, the focus will be on the typological categorization. Below are the three main sound object categories that can be mapped onto three types of bodily motion according to their dynamics and envelopes of duration.

- Sustained, meaning there is a relatively stable and continuous effort over time. Sustained objects generally maintain a constant amplitude and lack a clear attack and decay phase. Some examples may include continuous stretching or bowing.
- Impulsive, meaning there is a rapid onset or attack followed by a quick decay. Impulsive objects correspond to sharp or sudden events such as a door slam, a kick or a punch.
- Iterative, meaning a repetitive or cyclical pattern is exhibited. Examples may include scratching or brushing motions.

There are several other categories and sub-categories amongst Scaeffer's framework which focus on the internal features of sonic objects, as well as concepts that help define and utilize them in musical

compositions. Some examples of the latter include definitions such as singular objects, meaning one basic sound; or composite objects, meaning the fusion of rapid tone events into higher-level sonic objects [12]. However, in *Sonic Object Cognition*, Godøy states that classical sonic object research does not address long stretches of music, and that cognitive science research suggests that attention may fluctuate when listening to longer stretches of music [12]. This would mean that when used in a composition, sonic objects can be more difficult to define within the overall structure, as they are bound by a more integrated, subjective interpretation. Adressing this issue, Godøy suggests a triangular model involving gesture sensations, continuous sound and multimodal gesture images. These components interact with each other, with gesture sensations acting as the binding element between continuous sound, and multimodal gesture images [1].

In essence, Schaeffer and his colleagues' work paved the way to viewing sound fragments with their features described as trajectory shapes in time and space, with characteristics similar to gestures. Their emphasis on the dynamic changes and interactions among sounds encouraged a more holistic and multidimensional understanding of music, broadening the comprehension of musical expression beyond conventional melodic and rhythmic elements. The top-down classification approach developed by Schaeffer is a useful starting point in understanding sound and music within an embodied paradigm. Although there may be challenges with definition and representation of sonic objects within a larger composition, the potential for visualization through the integration of sonic and gestural features holds promise.

1.2.4 Intuitive Audio-visual and Audio-spatial Correspondences

The investigation of intuitive crossmodal correspondences starts with acknowledging the natural tendencies of the human brain in seeking to establish crossmodal correlations. As suggested by Bayesian integration theory, humans may integrate sensory inputs in a statistically optimal way by fusing sensory data with prior knowledge and weighing each one according to its relative reliability [15]. This suggests that prior perceptions play a significant role in the mappings between sensory modalities, referred to as crossmodal correspondences in the field of multisensory perception. Given the continuous integration of sensory information by the human brain, it is reasonable to suggest that crossmodal correspondances, whether innate or acquired via learning, constitute and integral aspect of human perception. Thus, our prior perceptions, such as the physical constraints of the world around us, aspects of our vision, and our bodily capabilities have an impact on our expectations and perceptions. This would mean that similarities can be found within individuals' mappings across modalites that may be attributed to the shared experience of being human. Therefore, examining crossmodal correspondences that are prominent across cultures and ages may facilitate the development of an intuitive framework for multisensory integration.

A number of researchers have examined the correspondences between audio-visual and audio-spatial stimuli, with several of these studies focused on the innateness of crossmodal correspondences by conducting their experiments with children, as well as adults across cultures. A well known phenomenon in the field of crossmodal perception which has been proven to be consistent across ages and cultures is the bouba/kiki effect. First proposed by German psychologist Wolfgang Köhler in the early twentieth century, it emphasizes the inherent association between sounds and their corresponding visual and auditory stimuli by proving that people consistently associate rounded, curvilinear shapes with the phonetically soft and melodious term "Bouba", while angular, jagged forms are associated with the phonetically sharp and abrupt designation "Kiki" [16].

One study which aimed to investigate the bouba/kiki effect across cultures and languages is the one by Bremner et. al, where they tested the effect with the semi-nomadic Himba tribe in Namibia, who do not use a written language [17]. The results showed that the majority of the participants (28 out of 34) associated the rounded shape with the name "Bouba" and the angular shape with the

name “Kiki” [17]. Moreover, Maurer et al.’ research on sound-shape correlations revealed that the Kiki/Bouba effect also emerges in toddlers as young as 2.5 years old, even before reading skills are developed [18]. These studies provide evidence to support the notion that there exists a fundamental connection between sound and shape perceptions in human beings that is beyond language.

In his doctoral thesis *Shape, Drawing and Gesture: Cross-modal Mappings of Sound and Music*, Mats B. Küssner explores the concept of shape in music cognition, focusing on the visual shapes of sound tracings created by the body as a mediator between the physical environment and the musical mind [2]. Küssner provides a connection between sounds, gestures and visual tracings while aiming to unite concepts such as Leman’s second-person descriptions and Godøy’s theories based on Pierre Schaeffer’s sonic objects [2].

As reviewed by Küssner, earlier research in correspondences between audio-visual and audio-spatial stimuli indicates that individuals tend to associate high pitch sounds with increased brightness, higher elevations in space and smaller objects [2]. Walker et al.’s findings indicating that 3 to 4 month old infants associate high pitch sounds with higher spatial positions and with spikier shapes highlight how these associations are either innate or developed at a very early age [19]. Eitan and Kohn further investigated children’s correspondances of musical parameters through their movement responses [20]. 106 participants aged 5 or 8 were instructed to move to musical excerpts “in an appropriate way”. Their movements were videotaped and analyzed by three referees, considering spatial directions, speed, and muscular energy [20]. The findings of their study suggest that distinct musical parameters impact distinct motion dimensions, with pitch affecting vertical motion, loudness affecting muscular energy, and tempo affecting speed and muscular energy [20].

Aiming to build on previous research and conceptualize music cognition as cognition by cross-modal sound shapes, Küssner focuses his thesis on real-time correlations of musical sound with the visual and kinesthetic domains as he conducts experiments both gestures and drawings. Küssner aimed to build on previous research by investigating whether musically trained individuals possess enhanced and transferable sensory-motor skills, by conducting his experiments with both musically trained and untrained participants. In a real-time drawing experiment, participants were asked to represent pure notes as well as musical excerpts by drawing on a Wacom tablet [2]. While proving that there were more varied interpretations within the musically untrained participants, Küssner’s findings demonstrate that most of the participants represented pitch onto a vertical axis, correlated loudness with size and represented time on the horizontal axis [2].

In terms of the gestural paradigm, Kussner focused his tests on representations of elapsed time, pitch, loudness and tempo, while also adding a three-dimensional visualization condition. In the first part of the experiment, participants were asked to gesture without visual feedback; while in the second part, their gestures were displayed real-time on a screen, through a black disk which had a 4 second decay time before it faded out. The disk’s size could be changed by moving the hand closer or farther from the screen, while movements on the x and y axis were directly represented. The results, which were obtained through motion capture with the use of Microsoft Kinect and Nintendo Wii, alongside video recordings, were analyzed through statistical methods such as ANOVA and Spearman’s rank correlation coefficient. The analysis consolidated previous research on gestural representation of pitch being associated with height, as all of the 64 participants showed such a correlation when it came to their hands movement. It also proved such correlations were more prevalent in musically trained participants. Furthermore, the results showed that elapsed time was strongly associated with the horizontal axis, while muscular energy was revealed to be associated with loudness, and speed of hand movement with increased tempo. Interestingly, the study found that visual feedback had a negative influence on pitch and height associations, as well as causing an increase in the speed of the hand movements. This could be due to a number of reasons, including the participants willingness to interact with the visualization, and proves that further research is required to study the varying effects of sound and gesture visualization.

1.2.5 Summary

Based on the research conducted in this chapter, it can be inferred that sound and music can be understood through an embodied paradigm, whereby gestural and sonorous features are integrated. This can be attributed to the manner in which music is perceived, which involves metaphors related to human perception, a combination of visual and kinesthetic sensations, as well as sensations related to effort and dynamics. With this perspective, sound fragments can be viewed and categorized in terms of their gestural affordances. Furthermore, individuals' perceptions of different sensory input often show similarities across ages and cultures, suggesting that there are intuitive qualities to audio-visual and audio-spatial correlations. Several notable examples include rounded and curved forms often being linked to softer and more melodic sounds, whereas jagged and spiky shapes usually being associated with sharper and more abrupt sounds. Other significant examples of crossmodal correspondences prominent in most individuals include the correlation of higher pitch sounds to brighter colors, as well as higher elevation in space; and louder sounds being associated with greater size a more muscular energy. It can be concluded that, if used distinctively enough, the aforementioned crossmodal correspondences can be viewed as intuitive and almost universal, allowing them to be an integral part of design and development of collaborative multimodal experiences.

1.3 State of the Art

In a rapidly evolving landscape of artistic-technological convergence, innovative approaches to music creation through sensory integration and multimodal systems continue to emerge. The instances presented within this section exemplify systems that strive to utilize the capabilities of gesture and/or visualization within musical performance, ultimately offering a unified experience. Furthermore, methods and technologies that allow the production of generative and collaborative visual music are reviewed.

1.3.1 Gestural Interaction with Live Looping Systems

Live looping is a method which involves recording and playing back of audio loops, applied by numerous sensor driven real-time interactive musical systems, used for creating layered and evolving musical patterns. This can be a useful technique for gestural interfaces within real-time musical performance, since it obviates the need for explicit instructions through any other interface, fostering a seamless music generation process. Moreover, the technique of live looping enables the modification of musical elements and their parameters by using several methods of mapping gesture features to sound features, as well as to the controls of the looping system. This dynamic approach nurtures improvisation while maintaining a desired level of structure and control.

One innovative contribution to the realm of gesture-to-sound mappings within looping structures is MapLooper; an open source framework developed by Mathias Bredholt. Created with the aim of acting as an intermediary between gesture input and synthesis parameters, MapLooper employs a distributed mapping approach, meaning that mappings between gesture input and synthesis parameters, as well as synchronization with other devices, occur through direct communication through a distributed network. This is done through the use of libmapper, an open source software library for facilitating network connections to enable shared data access among interconnected peers [21]. This decentralized structure complements MapLooper's looping feature, allowing for the playback of previously recorded sensor data while being synchronized with other sequencers and loopers, ensuring a coherent musical output [21]. Rather than focusing on a single application, MapLooper aims to be an embedded platform, accommodating different hardware and software configurations in order to be a key element in enabling gestural mappings for musical performance [21].

The SoundGrasp system, created by Mitchell et al., is a multimodal looping system aiming to utilize distinct correspondences between gestural input and musical output [22]. SoundGrasp employs a gestural interface facilitated by gloves, which enables users to interact with the system using designated gestures that are metaphorically linked to certain actions. For instance, the act of grasping is associated with the recording of sound, while pinching is used to freeze a note [22]. The gesture recognition is performed through the use of an artificial neural network called multilayer perceptron, which has been extensively utilized for the non-linear control of audio and visual systems [22]. SoundGrasp features an audio control layer that enables users to transition between modes by using hand postures as gestures [22]. It employs two distinct categories of gestures: audio control gestures, which include operations such as recording, playing, stopping, restarting, reversing, and applying audio filters/effects; and state/mode control gestures, which facilitate switching between different audio control modes [22].

Another interactive system for gestural control and improvisation in live musical performances is called Gestate, presented by Mainsbridge et al. in their paper "Body as Instrument: Performing with Gestural Interfaces" [23]. They combine looping functions with mappings of communicative gestures, as well as performance with virtual instruments and open air gestures, integrating the gestures' temporal and spatial aspects. Also aiming to utilize intuitive connections between modalities, Mainsbridge et al. establish a connection between pitch and verticality, panning and horizontal movement, as well as volume and extention of the limbs [23]. The implementation makes use of a Kinect depth camera and Synapse motion tracking software for obtaining data such as position, speed, and the magnitude of the performers' gestures. This data allows gestural control over effects levels, a looper and virtual MIDI instruments within Max/MSP and Ableton Live [23].

In summation, it can be stated that live looping systems are an extensively used method that has shown its effectiveness within the domain of interactive multimedia, particularly in the context of musical performances that include gestural interaction. Several works have aimed to establish connections between gestures and music, with a multitude focusing on metaphors and intuitive connections between modalities. Such frameworks facilitate the design of multimodal systems for music creation.

1.3.2 Interactive Visual Music

Visual music is defined by Evans as visual imagery that exhibits a time-based architecture relative to music [24]. Similar to the aforementioned viewpoint of examining gestural-sonorous elements in relation to their temporal and spatial dynamics, visual music can be regarded as beeing driven by the resolution of tension [24]. Furthermore, incorporating movement input with auditory feedback as well as visual feedback can be stated to be a promising method for the intuitive understanding of a multimodal interface [25]. Today, there are many innovative installations and applications for making interactive visual music that utilize a variety of methods from visual programming software to deep learning algorithms. These methods aid in a generative design which is suitable for an engaging and collaborative experience. This section provides two illustrative instances of installations that exemplify such designs, as well as several methods that can be used for the creation of generative visual music.

1.3.2.1 Reactable

Developed since 2003 by a team of researchers led by Sergi Jordà, Günter Geiger, Martin Kaltenbrunner, and Marcos Alonso at the Pompeu Fabra University in Barcelona, the Reactable is an electronic musical instrument that provides a visual, as well as tangible interface for creating music collaboratively [25]. Its design consists of a round touch-sensitive translucent table which users can operate by manipulating acrylic pucks on its surface [25]. It is inspired by analogue modular synthesizers; with the pucks representing "the building blocks of electronic music" [25]. The tracking system uses reacTIVision, an open-source computer vision tracking software for efficient and reliable tracking of

fiducial markers and multiple touch points for finger-based interactions [25]. Once the pucks of the Reactable are placed on its surface, they become illuminated and start interacting with adjacent pucks based on their relative positions and proximity [25]. Visual feedback is provided through a short throw projector underneath the surface [25]. Today, the Reactable is sold as a product as well as a mobile application, but also offers its open-source models for wider utilization.

1.3.2.2 Anima

Anima is an interactive installation by Bureau Moeilijke Dingen which facilitates collaborative music creation between humans and AI systems[1]. Functioning as a sequencer, it deploys open-source AI algorithms such as Magenta Studio[2] and MusicAutoBot[3] to generate music which allows the system to contribute as an active musician during jam sessions. This two-way engagement allows users to adapt, build upon, and experiment with AI-generated sequences. The interaction is enhanced through LED lights, visual indicators, a virtual character guide, and adjustable AI parameters via rotation knobs. Users initiate sessions by placing genre-indicative disks on the "inspiration" module, prompting AI-generated beats and melodies in corresponding modules. The installation's architecture, governed by a central PC and custom hardware units, orchestrates AI tasks, user input/output, MIDI control through Ableton, and LED visualization. While, seamlessly combining rhythm and melody models, Anima also employs various modes of visualization, offering a multimodal and collaborative musical experience without the need of musical expertise or even a human collaborator.

1.3.2.3 Generative Visual Music

The utilization of generative techniques allows interactive systems to become a part of an ever-evolving feedback loop between the system and the users. These systems' developing dynamic are suitable to allow a collaboration akin to a musical jam session. Several generative and collaborative music systems such as MapLooper, Reactable or Anima employ models that utilize Ableton for MIDI control[4]. Amongst its versatile features for music production, Ableton also has capabilities such as its looper device that aids in live looping. In addition, it is possible to connect Ableton to visual programming software such as Max/MSP[5] or Touchdesigner[6], enabling interactive visualizations. Furthermore, several tools have recently been developed for using deep learning algorithms such as Facebook's MusicGen for music generation and Stability AI's Stable Diffusion for image synthesis and animation within Touchdesigner, which can further aid in a generative visual music experience [26][27].

1.4 Aims of the Thesis - Final Problem Formulation

The findings from the preceding chapter suggest that sound and music perception is interrelated with visual and kinesthetic modalities. In addition, audio-visual and audio-spatial correspondences are demonstrated to exist beyond age and language. This would allow the development of an intuitive system to make music multimodally, whereby sounds may be categorized into gesturally affording groups, and gestures can be mapped to sounds and visuals through widely observed crossmodal correlations. Furthermore, live looping is proven to be an efficient method for systems that employ real-time gestural interaction for musical improvisation.

All of this ultimately gives rise to the final problem formulation:

[1] www.moeilijkedingen.nl/cases/anima
[2] magenta.tensorflow.org/studio/
[3] github.com/bearpelican/musicautobot
[4] www.ableton.com
[5] cycling74.com/products/max
[6] derivative.ca

Is it possible to provide an integrated multimodal experience by taking gestures as input, classifying them based on their gestural-sonorous features and implementing intuitive correlations between audio-visual and audio-spatial modalities to generate multimodal objects? How can this be used in a live looping system for making visual music?

1.5 Requirements

Based on the final problem formulation and the acquired knowledge regarding intuitive crossmodal correspondances and interactive visual music systems, the software requirements for this project were established. In the first phase of system development, it was determined that the focus would be on developing the system and evaluating the user experience with a single user. Thus, for the requirements, the emphasis was placed on developing a testable prototype for making visual music through gestures, which highlights the aforementioned crossmodal correlations through the mappings between the gestures, sounds and visuals. Although still included in the design and development, some generative and collaborative aspects of the system were deferred for future consideration.

1.5.1 Functional Requirements

- The system should analyze gestures and classify them according to their gestural-sonorous features real-time.
- The model should map the users' gestures of sounds and visual forms that are contrasting in terms of their dynamic envelopes, shapes and textures in order to provide a clear distinction.
- In order to establish a musical structure with a repeating pattern, a live looping system in conjunction with a rhythmic component should be used for the sounds and visuals.
- The model should make use of common audio-visual and audio-spatial correspondences to provide an intuitive experience.

  - Softer, more melodic sounds should be matched with rounder, more fluid visual objects.

  - Sharper, more abrupt sounds should be mapped to spikier, more angular shapes.

  - Louder sounds should be correlated with greater size.

  - Sounds with higher pitch should correspond to smaller size and greater elevation in space.

  - Higher pitch sounds should also correspond to brighter hues in the visual objects.

- The usability of the system should be intuitive, requiring little instruction for users to operate effectively.

1.5.2 Non-functional Requirements

- The model should implement generative and audio reactive techniques for the production of the visual objects.
- The more subtle features of the sounds such as their texture or effects should be included in their visualization.
- Further mappings of the users' movement and interactions with their environment should be considered regarding the overall interaction.
- The design of the live looping system should consider collaboration.

Chapter 2

Design

Based on the findings of the conducted research and the requirements outlined in the previous chapter, several design choices were made to ensure an intuitive multimodal experience of making visual music through gestures.

2.1 Design Methods

For the design of the multimodal looper - a task involving three different sensory modalities - iterative incremental development model was chosen to be most suitable. This method assisted in dividing the project into smaller modules for each modality, prototyping them separately, and then integrating them gradually while revisiting and refining each module. Despite ultimately simplifying the design of the first model by reducing its generative and collaborative aspects, the design approach incorporated the potential and restrictions that were revealed through study in these domains.

2.2 Fitting Gestures to Sound

With the aim of creating an intuitive framework, the emphasis for the gesture to sound mappings was determined to be placed on the sound object categories first established by Pierre Schaeffer, and other widely observed audio-spatial correlations described in the related research section (1.2). Motion analysis through a depth camera and skeletal tracking was decided as a suitable method for the analysis and classification of gestures, as it can aid in obtaining wide variety of motion information including position and accelaration of various body parts; from which further calculations can be made. Although the potential for mapping gestures to MIDI control through integrating Ableton and Kinect within the visual programming tool Touchdesigner was explored, it was ultimately chosen to simplify the design of the music module. Therefore, it was decided that a collection of audio samples was to be used within Touchdesigner, a software that can handle a number of audio tasks with its channel operators. To determine which audio samples to use, the applied method was to listen to a variety of samples and identify those that had temporal, spatial, and dynamic features that allowed them to be distinctly perceived as iterative, impulsive or sustained; hence ensuring contrast and clarity within the crossmodal mappings.

2.3 Visualization

The visualization of the multimodal objects was also inspired by the aforementioned research on sound object classification and intuitive crossmodal correspondences. Various animation techniques

were designated to be applied in order to recreate the distinct features of each sound object such as their temporal expansion and dynamics. Textural aspects of the sounds were also chosen to be reflected in their visualization. Furthermore, as generative techniques for visual music allow for a stimulating experience through an ever evolving interaction, generative visualization techniques such as feedback loops, noise animation, and audio reactivity were determined to be applied in the production of the visual objects. These techniques may be stated to hold potential for future iterations of the prototype where the music is more generative, or collaboration is possible.

2.4 The Live Looper

Due to being a widely applied technique for generative music production that is also used in collaborative systems, a looping system was deemed suitable for the task of this project. A circular design that can be projected on the ground was selected for the looper, due to the shape being ideal for a collaborative experience such as a jamming circle. The circle was decided to be rotating clockwise with the multimodal objects positioned away from the center and divided in eight portions to represent the looper's time sequences. A turn taking model was decided to be applied, again with collaboration in mind. Consequently, it was determined to divide the looper into eight cycles of eight beats and allocate user input capability to every second cycle. Indicators were chosen to be added to the looper's design to clarify the input and loop cycles. A red triangular shape that rotates together with the looper was included in the design, with its appearance indicating the cycle being recorded. Circular indicators were also added to the bottom of the looper's rotating canvas to further illustrate the input cycles for a single user. The design of the looper is illustrated in figure 2.1.

2.5 Genre and Aesthetics

The prototype's genre and visual aesthetics were determined to be derived from the chosen audio samples, which evoked a futuristic and electronic ambiance. However, a conscious decision was made to set the tempo at 100 beats per minute (BPM) without incorporating bass elements. This choice was made to ensure that the beat remains inconspicuous and allows the multimodal objects to be the central focus. The visual objects, as well as the background were designed to compliment the space-like and electronic atmosphere.

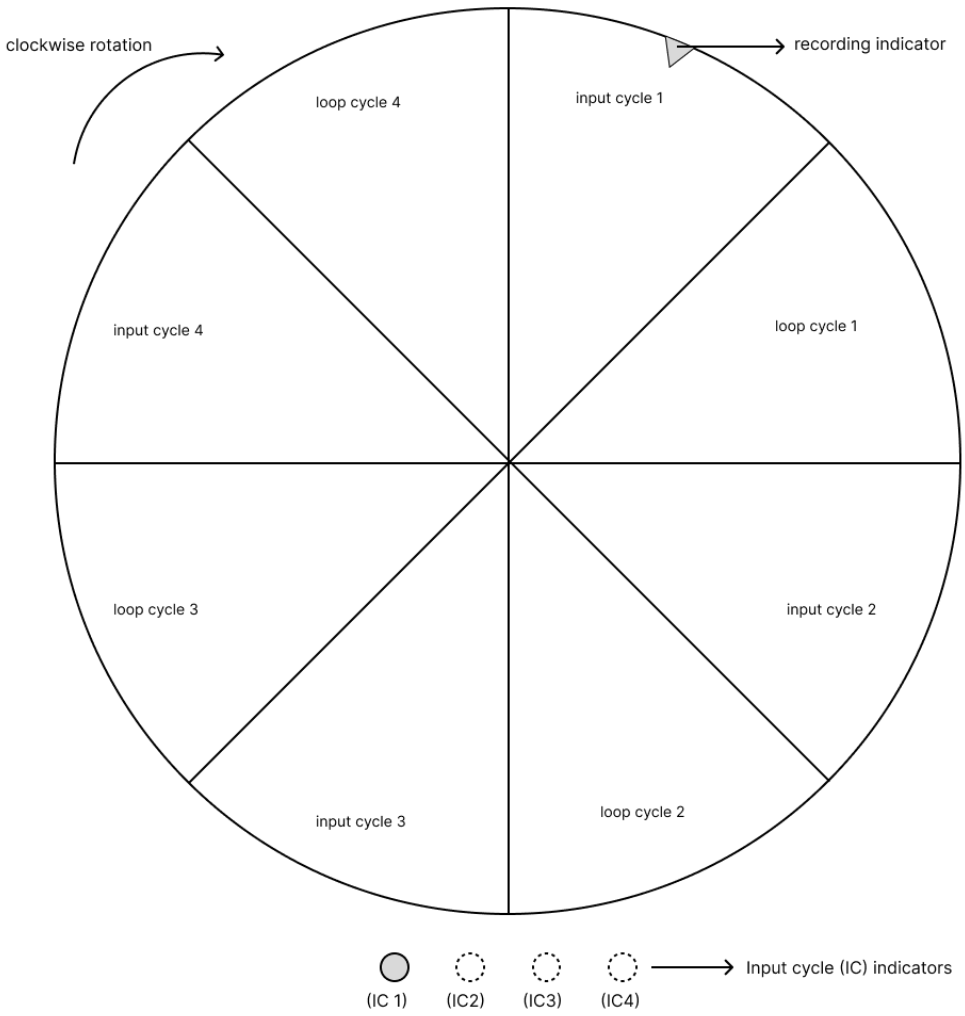

Figure 2.1: The design of the multimodal looper

Chapter 3

Implementation

3.1 Hardware and Software

3.1.1 Touchdesigner

Initially developed as a fork of Houdini, Touchdesigner is a powerful visual programming language and development environment that is widely utilized in the creation of multimodal interactive installations.[1] The prominence of Touchdesigner in the realm of interactive installations can be attributed to its ability to seamlessly integrate various data formats, such as sensor data, audio and visuals, as well as its ability to connect to a diverse range of hardware and software. Touchdesigner's power can also be attributed to its node-based visual programming model that allows for real-time modifications and integration of live data inputs, which can be used with generative techniques such as noise displacement and feedback loops. Furthermore, the use of Python scripting within Touchdesigner enhances the potential for implementing various algorithms and developing complex interactive systems.

3.1.2 Kinect

First released in 2010 as an Xbox accessory, Kinect is Microsoft's line of motion-sensing devices. Although the Kinect has since been discontinued, its low-cost depth information capabilities have rendered it useful beyond gaming; enabling applications such as 3D-simultaneous localization, object recognition, and human activity analysis[28]. Together with its software tool Kinect SDK, the Kinect detects and tracks the movements of specific skeletal joints with the help of depth data analysis and pattern recognition algorithms. Although skeletal tracking is also possible with the combination of a regular camera and convolutional neural networks such as PoseNet; the depth data obtained from infrared cameras such as the Kinect is especially convenient for motion analysis in situations such as installations with projector setups that require a dark environment[29]. For this reason, and its availability from Aalborg University, Kinect v1 (version one) was used for this project.

3.2 Gesture Analysis

In order to separate the user's gestures into three categories fitting the typology of sound objects defined by Pierre Schaeffer, a simple decision tree approach was used. Focusing on the movement of hands and feet, gestures are calculated to be sustained, impulsive, or iterative; which in turn trigger the associated AV(audiovisual) objects. This is achieved by getting the joint positions from the Kinect v1 sensor and using various channel operators in Touchdesigner. These operators give out signals of 0

[1] https://derivative.ca/

or 1 according to their set conditions and can be added together with "And" or "Or" options (amongst others), enabling complex decision trees. Furthermore, the use of space is considered vertically and horizontally in order to trigger different notes on the musical scale. For example, if the gesture is higher in vertical space, this plays the sample that has a higher pitch or involves notes that are higher on the musical scale.

Firstly, if there is a person identified by the sensor, this raises the volume of the beat. The space in front of the sensor is separated horizontally into three areas corresponding to three samples of continuous, atmospheric sounds that serve as a background to the more melodic sounds created by the gestures. Since standing can be considered a sustained movement, according to the person's horizontal position in front of the sensor, three variations of these sustained sounds that increase in pitch from left to right are triggered to play. If there is no person identified by the sensor, the overall volume goes down to zero, thus refreshing the output of the looper which will be described in detail in section 3.5.

Iterative, sustained and impulsive gestures are calculated and categorized simultaneously. Their modes of calculation are as follows:

- Iterative gestures are calculated by starting a timer at the beginning of every recording cycle and counting the number of times the hands have moved above one another during that time. If the number exceeds 4, this counts as an iterative gesture and the related sounds and visuals are played.
- Sustained gestures, on the other hand, are calculated with the use of Touchdesigner's slope operator which converts position data to speed data by calculating the rate of change of a channel's values over time. Thus, if the speed of the hand is between the threshold values of -0.005 and 0.005 on the x, y and z axes, this triggers different samples of sustained piano chords, increasing in scale according to the height of the hand's position at the time of the gesture.
- For the last category of impulsive gestures, the slope operator is used again together with the speed operator to calculate the acceleration of the hands and feet on the x and y axes. If the acceleration goes above the set threshold for one of these body parts, the AV objects categorized as impulsive are triggered to play.

The three gestural object categories trigger brighter and higher pitched AV objects relative to their height on the y axis. Through the use of logic operators, one category of AV objects is programmed to only be activated by the gestures if there are no other categories of AV objects being activated at that time. This was done in order to minimize false triggers due to the rather basic nature of the gesture calculations.

3.3 The Music Module

For the first prototype of the multimodal looper, a rather simple approach was taken in terms of music production. From a free pack of 300 artisan FM sounds from various Yamaha DX synthesizers, which were obtained from the Dutch composer Legowelt, four groups of three to five samples were chosen in accordance with Schaeffer's typology[2]. The beat for the looper was obtained for free from the website looperman.com and chosen to be a percussive beat without bass, at 100 BPM, to be as unobtrusive as possible and let the attention stay with the melodic sounds that are created by the user's gestures[3]. While also seeking to establish an ambient electronic style, the choice of the foreground samples was aimed at highlighting the diversity of their designated categories.

[2] legowelt.org

[3] www.looperman.com

The continuous sounds to accompany the score as long as a player is identified were determined to be three synth pads; meaning sustained tones or chords meant to be used as harmonic background material. The pads are thirteen seconds long samples of grainy ambient sounds with varying pitches. Moreover, five samples of piano chords were chosen to correspond to sustained objects due to their soft and continuous nature. Contrastingly, four samples named "Techno Blip", which are striking electronic sounds less than one second long with a fast attack and decay, were utilized in correlation with impulsive objects. Furthermore, iterative objects were represented by three samples labelled "Pure Tones Announcement", which contain three notes played relatively rapidly after one another and have a joint decay phase similar to an echo. These samples were imported into Touchdesigner with the audio file in operator which allows references from other operators for cueing and volume, as can be seen in figure 3.1. The samples were then added together with the math operator in order to connect to the same audio input for the looper.

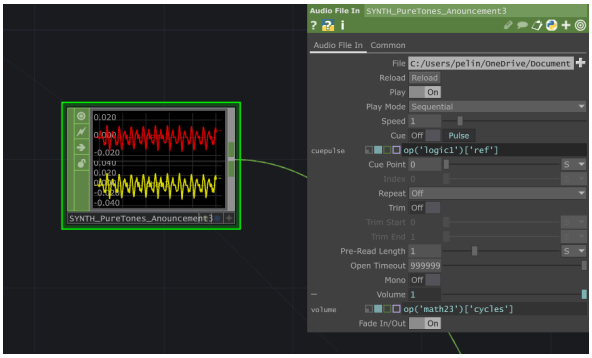

Figure 3.1: The audio file in operator with cue and volume references

3.4 Multimodal Object Visualization

The visuals for the multimodal objects were created in Touchdesigner with the aim of reflecting the properties of their corresponding categories of sustained, iterative, and impulsive. While keeping the theory of intuitive crossmodal correspondences in mind, reduced listening technique was also used by the author to bring up mental images of the sound, which inspired the ideation phase for the visuals. Generative techniques such as feedback loops allowed the production of different variations for each audio sample. These variations were rendered as videos to then be played in accordance with the user's gestures. Instancing was another technique used to duplicate and render multiple instances of a 3D object or geometry and animate them according to the audio. The visuals for each sample are composited with the blending mode set to screen mode, which inverts top and bottom layers' colors and multiplies them; creating a brighter, more light-like appearance when the multimodal objects overlap.

3.4.1 Sustained Objects

To accompany the synth pads meant as harmonic background elements mentioned in section 3.3, visuals were developed through feedback loops and noise displacement techniques to create a grainy, sweeping animation aimed to fit sustained objects (see figure 3.2). Decreasing size, as well as increasingly brighter and warmer hues of colors were utilized in correspondence to the increasing pitch of the three samples.

The sustained piano chords were represented by multicolor "blobs" with a liquid or paint-like texture. This effect was produced by instancing a circle on three axes and animated through displacing it

with a ramp TOP (Texture Operator). With the animations length fit to match the length of each of the chord samples, animating the circular ramp TOP's phase and position provided an effect that resembled sound waves expanding through space and finally shrinking to disappear. The expanding animation is demonstrated in figure 3.3. Colored with a lookup TOP and a vertical ramp of gradients, the multitonal colors of the animation were aimed to reflect the notes of the chords; minor chords were associated with darker midtones, and major chords were more colorful and increasingly brighter relative to their musical scale.

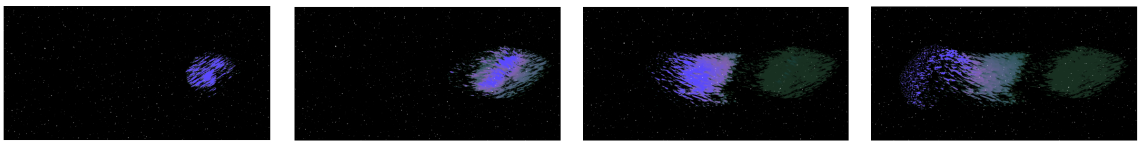

Figure 3.2: Sweeping animation for the sustained object corresponding to the synth pad

Figure 3.3: Expanding animation frames from a sustained object representing a piano chord sample

3.4.2 Iterative Objects

For the visualization of iterative objects, audio reactive animation techniques were applied together with instancing to represent the consecutive musical notes. To make the visual objects audio reactive, audio analysis techniques were used in conjunction with the samples in order to turn the audio signal into a fit channel for instancing the visuals. After resampling is used to reduce the sample rate of the audio, left and right channels are combined by obtaining their average to create a mono channel. The audio spectrum CHOP (Channel Operator) is utilized in order to boost the higher frequencies for a more detailed visualization. After the modifications made to the audio input, the audio channel is converted through a CHOP to TOP operator which converts numerical data into pixel format. The network then connects to different noise TOP's for position, rotation and color instancing. This data is used to instance circle geometry, resulting in a generative audio reactive visualization, frames of which can be seen in figure 3.4. Furthermore, the echo in the decay phase of the audio was visualized through the use of the edge TOP, which finds and highlights the edges in an image, within a feedback loop to create the visual effect of ripples.

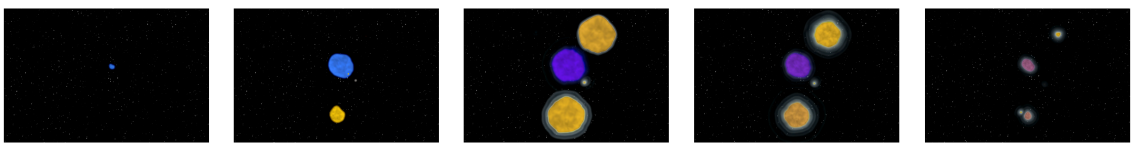

Figure 3.4: Iterative object animation frames

3.4.3 Impulsive Objects

In the case of impulsive objects, the same techniques of audio reactivity and instancing used for the iterative objects were applied, only this time with a tube instanced instead of a circle. This was done with the aim of matching the impulsiveness and higher pitch of the samples with faster and spikier objects. Different variations were rendered, increasing in brightness and decreasing in size according to the sample's pitch rise. A glowing effect was also added to match the corresponding audio sample's digital nature. Example animation frames can be seen in figure 3.5.

Figure 3.5: Impulsive object animation frames

3.4.4 The background

A starry look was achieved for the background by connecting an animated noise to a function TOP which performs mathematical operations; in this case, the RGB values of the noise were raised to the power of 15. Due to the rotating design of the looper, which will be explained in the next section, the background choice was aimed to be dark and not too distracting to the eye. The background was also intended to compliment the space-like quality of the music and visuals.

3.5 The Live Looper

For the first iteration, it was decided to test the prototype with a single user, although the features of the looper were designed for a collaborative experience. Therefore, a circular design was applied; motivated by the idea of a jamming circle with the looper projected in the middle. The looper runs clockwise, completing one full rotation every eight cycles of eight beats. The system also involves certain additional functions for the overall interaction, such as fading out the multimodal objects with the absence of a user detection, as well as fading out the beat if the user's head is lowered below their core.

The looper is mainly controlled by a timer CHOP, python scripting and logic operators, which further steer the input and output operators for the audio and pixel channels. The timer controlling the looper is set to repeatedly run for eight cycles of 4.85 seconds, which is the length of 8 beats of the 100 BPM audio sample used for the rhythm. Inputs from the user is taken every second cycle, visualized by four circles outside the rotating looper, alongside a triangle on the side, rotating together with the input multimodal objects (see figure 3.6). The four circles in the bottom, going from lighter to darker blue, represent each input cycle, and were added to indicate each specific cycle for a single user. The looper is also visually divided in eight portions representing each cycle. The visual inputs are added at the location of the active cycle, also indicated by the rotating triangle, and positioned according to their pitch with higher pitch sounds positioned higher on their relative y axes.

Furthermore, in order to fit the input to the beat, a beat CHOP is programmed to match the BPM of the audio. If there is a user input during an input cycle, this activates a trigger CHOP, which conditioned together with the beat CHOP, delays the recording of the input until there is a beat. Inputs are taken through four audio input operators and four video input operators, respectively for

each of the four input cycles. After delaying the cycle change by one frame per second, the outputs are reloaded for each input cycle and are played for the remaining seven cycles until a new input is taken. Figure 3.7 illustrates the main processes of the multimodal looper.

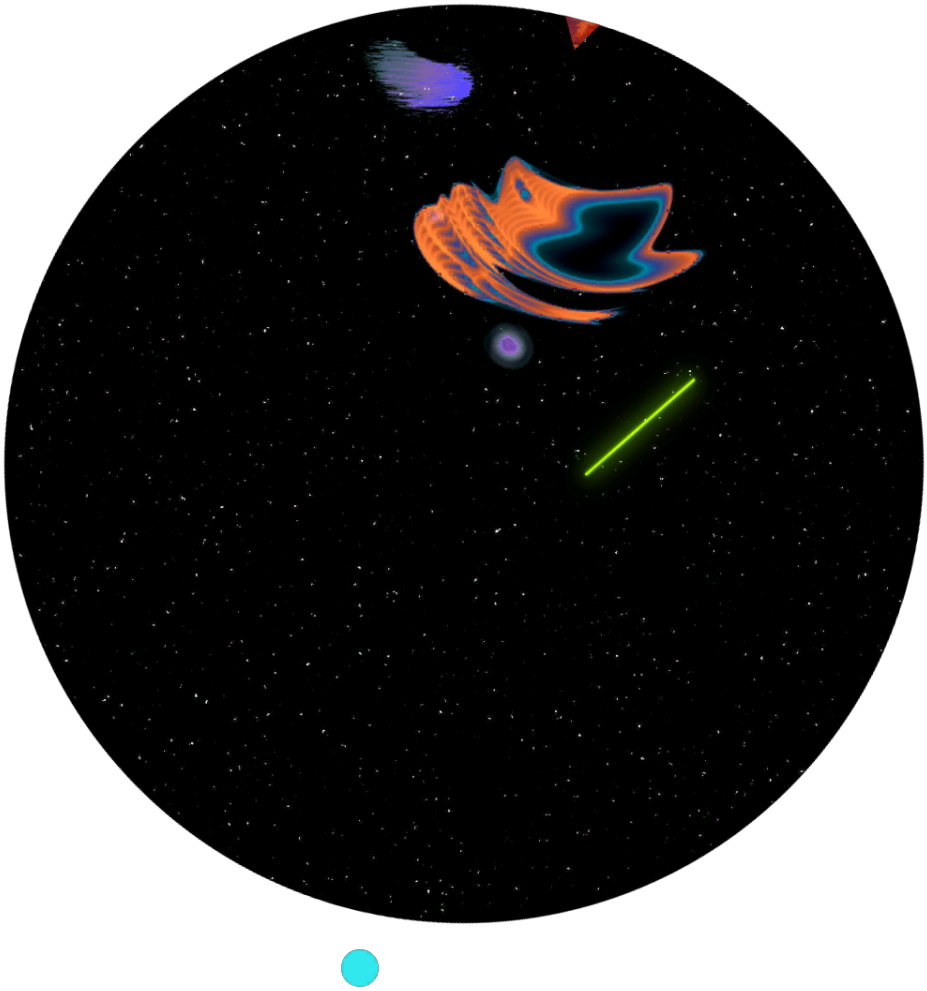

Figure 3.6: The looper visualization during an input cycle

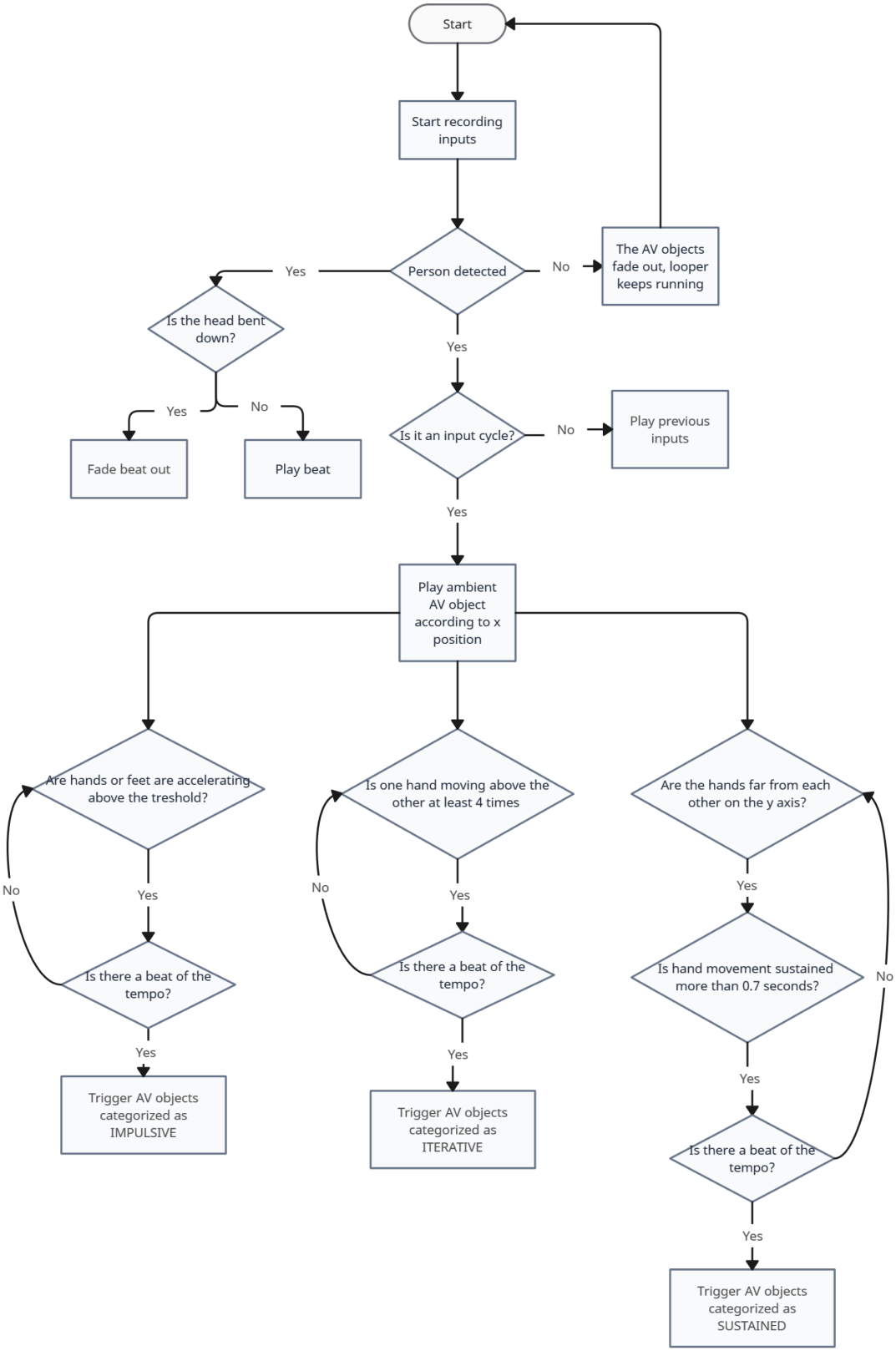

Figure 3.7: The processes of the multimodal looper

Chapter 4

Evaluation

4.1 Evaluation Methods

For the evaluation, the focus was set on determining whether the system was successful in creating crossmodal correspondences to provide an integrated multisensory experience. A hypothesis was formed in line with Bayesian integration theory's suggestion (explained in section 1.2.4), that the strength of crossmodal couplings depend on how much the sensory input correlates with our sensory system's prior knowledge. It was hypothesised that, if presented with two versions of the prototype, one with stronger correlations of gesture to sound, as well as sound to visuals; and one with more random correlations, the users would perceive a weaker coupling between their modalities in the randomized variation. This would mean that the prototype was effective in creating crossmodal mappings in line with their prior perceptions.

Thus, in terms of the experimental setup, A/B testing (or split testing) method was applied. A variation of the system was created where the aimed coupling strength between the different modalities was lessened by using the same sounds and visuals, but randomizing the mappings of gesture to sound, as well as sound to shape and colors of the visuals. As an example, in the randomized version, gestures that were classified as impulsive now produced iterative sounds. Furthermore, smooth and fluid visuals, which were meant for sustained sounds of piano chords, now represented fast and high-pitched impulsive sounds and so forth. Because the same visuals were shuffled and used, the speed of the videos were then adjusted to match their new sounds.

Two trials were designed where participants were presented with the version of the prototype with stronger correspondences and the version with weaker correspondences between the modalities. This was done in randomly varying order for each participant in order to reduce the effect of their learning curve increasing in their second attempt. After each trial, the participants answered ten identical questions in the form of a survey with 5 scale Likert questions, with responses ranging from "Strongly Agree" to "Strongly Disagree". These questions all inquired about their perceived correlations between their gestures, the sounds and the visuals, in varying forms and combinations. The results of the survey were analyzed by calculating and comparing the average perceived correlation ratings for each coupling condition. Moreover, statistical tests were applied to evaluate whether there are significant differences in perceived correlations between the two coupling conditions.

Aside from the quantitative evaluation, several methods were applied with the aims of gaining qualitative insight about the users and their experience. In the beginning of testing, the participants answered two questions about their previous involvement in music or multimedia design, to which they could reply with "yes", "no", or "other" with an explanation. To assess the participants' musical involvement, they were asked if they played an instrument, performed or composed music in the past

six years. Moreover, they were asked if they were in any way affiliated with sound or multimedia design. This was done with the aim of understanding if any affiliation with music or media influenced the participants' performance, due to musical training having been demonstrated to increase the use of common crossmodal correlations within individuals in a multimodal interaction[2]. Furthermore, as the final part of the testing, the participants were asked to type their answers to some open-ended questions about the non-randomized version of the prototype, such as asking them to describe the experience, what they think could be improved, or if they could see this being used in a collaborative way. To finalize, the participants were invited to vocally express their remaining thoughts, with the aim of initiating a dialogue and eliminating the potential presence of undisclosed sentiments. Additionally, the participants were observed by the researcher during the experiment and notes were taken about their behaviours, as well as their interactions with the system and their environment for further assessment.

4.2 Procedure

The evaluation process took place over the course of two weeks in the Augmented Cognition Lab of Aalborg University Copenhagen. A big portion of the 9 participants were recruited based on convenience sampling. Due to the fact that the prototype was to be tested with a single user, it was decided that projecting the visualizations on the wall instead of the ground would be sufficient for the testing. Thus, the set-up was the Kinect, a portable computer running the software, and a short-throw projector with speakers; projecting on a wall in a space that was surrounded with dark curtains to block the light and minimize distractions. Behind the curtained area, there was a computer for the participants to answer questions after each of their turns.

Firstly, the participants answered two questions about their previous involvement in music or multimedia. Then, they were told that, what they were about to try was a multimodal looper, which allows one to create music and visuals through their gestures. They were then invited to take their shoes off, explore the space and their bodies fully and "see what comes out". The only detail that the participants were told about the system was that it was recording their gestures when the circles and the triangle appeared. This was done in order to understand how intuitive the system was to use with little instruction, as well as to observe the participants' inclinations towards different ways of interaction. After experimenting with the prototype for two minutes, while they were also observed by the researcher, the participants answered ten survey questions about their perceived crossmodal couplings.

The participants then tried the prototype a second time for two minutes, although this time it was the other variation. Subsequently, they answered the same ten survey questions, with the addition of the open-ended questions in the end. They were instructed to only consider the non-randomized version of the prototype when answering the open-ended questions. Finally, the participants were invited to verbally share any insights or feelings regarding the experience, and testing was complete.

Chapter 5

Results

5.1 Survey Results

The survey results from the two testing conditions of strong and weak cross-modal correlations were analyzed and compared. This was done by inverting the scores for the questions where a higher score implied a weaker perceived correlation between modalities and obtaining the medians of the answers for both the strong and weak coupling conditions. Although the prototype with the strong aimed correlation had overall higher scores, it was only by a small margin, as can be inferred from the median comparisons for the two testing conditions in figure 5.1 The results were then compared by performing Wilcoxon Signed-Rank Test, a statistical test which is suitable for comparing rankings between two sets of data with a small sample size[30]. The result showed that there was no significant difference between the average scores of the answers in either coupling condition.

It may be further remarked that, when asked about their perceived correlation between their movements and the created sounds and visuals, the survey answers were very varied amongst the participants, with a homogeneous distribution across the scale. Moreover, the seventh and eighth questions asking if the participants sensed discrepancies between the speed of their gestures and the speed of the visuals, or between the timings of their gestures and the timings of the sounds; most participants' answers were in the middle point of the scale, signalling uncertainty. Additionally, the pre-testing survey questions revealed that most of the participants were either affiliated with sound or multimedia design or had played an instrument, performed or composed music in the past six years.

5.2 Open-ended questions

When asked to describe the experience, several participants indicated that it was confusing and difficult to understand which gestures triggered which outputs. When questioned about if they felt the system was intuitive to use, one participant remarked that the fact that the sounds and visuals appearing only in one corner of the screen felt counter intuitive. Two participants described it as straightforward and easy to understand, although the consensus was that it required experimentation and getting used to. Another participant remarked that it was difficult to know when the system was recording their gestures because the indicator changed position. Most of the participants thought that the prototype had potential for being used in a collaborative way similar to a jam session, with one participant remarking each corner of the looper could be allocated to different people. They also mostly expressed positive feelings towards a version where the sounds and visuals evolve with AI. Finally, when asked about possible improvements, one thought that was shared among most of the participants was that the AV objects could benefit from better timing, as well as a more distinguished

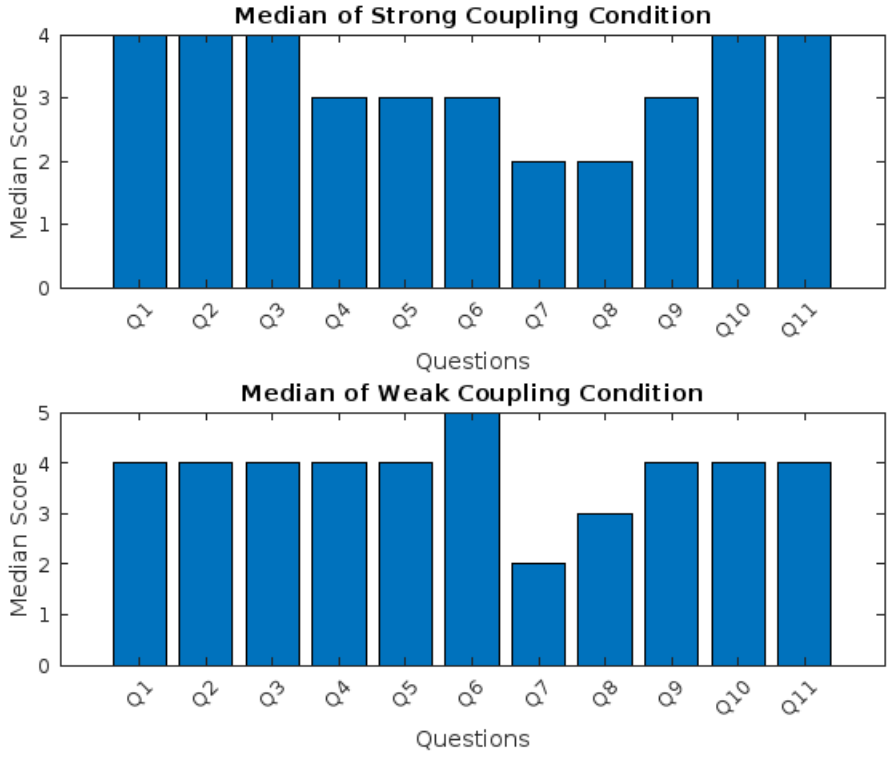

Figure 5.1: Median comparison for both conditions

placement instead of being on top of each other. Another shared argument for further improvement was clearer and better placed visual indicators. One compelling argument for enhancement made by a participant was to be able to modify the rotating canvas like a DJ (disc jockey).

5.3 Interviews and Observations

In the post-task interviews most participants shared their feelings and remarked once again what they thought could be improved. Many of the remarks signalled a perplexing user experience. It was mentioned by one participant that the indicators were confusing, and they did not know when to move and when not to move. Some participants mentioned that they felt like the system was less responsive in their try with the weaker coupling condition, although responsiveness had not changed. One participant had the opposite feeling; they felt that the version with the weaker coupling worked better in terms of responsiveness. One participant remarked that they sensed latency in the outputs of the system. It was mentioned by two participants that they understood the system better in their second try.

Furthermore, a general confusion was observed by the researcher during the testing. Although the participants were instructed that their gestures were only being recorded when they saw the indicators, most participants tended to move during the entire experience. The subjects exhibited obvious uncertainty about the mappings between gestures and sounds. The mapping of the gestures categorized as impulsive were observed to be more evident to the participants. Moreover, it was noted that the participants generally did not consider when to make the sound but were more interested in

getting some feedback from the system during the entirety of their interaction. This often resulted in the overloading of the musical composition.

Some further observations on the participants reactions towards the gestural interaction, for which they had very little instruction, are as follows: Only one participant intentionally tried to make sustained gestures such as holding their hand up in a fixed position. Only two of the participants used their legs to make kicking motions. None of the participants tried lowering their heads below their core. Another observation that is noteworthy was the fact that some of the participants did not explore the space horizontally. While some of the participants tried to make big gestures with the use of the entirety of their bodies, others tended to stay in one area and only try arm and hand gestures. Some participants moved more in dance-like movements. Overall, it was observed that the participants exhibited an interest towards exploring different modes of interaction in order to comprehend the system, and there was certainly an element of perplexity evident.

Chapter 6

Discussion

6.1 Achievements and limitations

As shown by the results, there was not a significant difference in the participants' perceived crossmodal couplings between the two testing conditions. Moreover, a significant amount of confusion regarding the interaction was observed by the researcher and remarked by the participants. This might imply that the prototype was not very effective in establishing coherent and intuitive crossmodal mappings, or that there were further flaws within the design and implementation of the system which contributed to a perplexing user experience and impeded the perception of the mappings. Based on an analysis of the qualitative findings, together with the data obtained from interviews and observations, several issues emerge as potential shortcomings of the system in question.

Firstly, as it was observed that there was widespread confusion in terms of the input and loop cycles, it is possible to say that the system was not successful in creating an intuitive user experience. The visualizations that were meant to indicate the input cycles, although also explained before the testing sessions, were ignored by most of the participants, as they showed a willingness to constantly move and try to provide input. This could have been due to being tested as a single user for a system that incorporates turn taking, or a willingness to explore a process that was unclear for which they had received minimal instruction. The projection of the looper on the wall instead of the ground could have also contributed to the confusion regarding the visual indicators, as well as the loopers rotating design which is more suitable for collaboration.

Another important limitation of the prototype also noted by most of the participants was the placement of the AV objects on the rotating looper. Several issues including insufficient mapping between the temporal structure of the music and the visuals, as well as disparities between the looper's design and the evaluation procedure may have contributed to this. Firstly, due to being projected on the wall, visual objects intended to be positioned at a greater distance to the user only appeared to be in closer proximity to the center. This may have had a big impact on the mapping of high pitch to further elevation in space not being perceived at all. Moreover, musical structure and its correct visualization in terms of timings was certainly a shortcoming of the prototype, as noted by a number of participants. For example, as it was noted in post-task conversations, the AV objects should be placed in the looper according to when they are played within their cycles, instead of being layered on top of each other.

Furthermore, it is crucial to note that the analysis and classification of the gestures was not sufficient for an intuitive system. As a lot of participants demonstrated a wide range of movements, it can be stated that, if the three sound object categories are to be utilized alone, further instruction should have been given about the gesture categories. If aiming for a more intuitive interaction that requires

no instruction, higher-level movement qualities should have been included, as well better analysis of movement patterns. Additionally, another limitation was the separation between the gestural categories, as it was observed that certain gesture categories were simultaneously triggered. This may have possibly caused a false perception of responsiveness and added to the confusion.

Finally, it is worth mentioning that it may not have been possible to accurately test the participants' perceived crossmodal correlations. This could be partly due to the aforementioned design flaws and technical limitations of the system, which may have shifted the participants' focus towards understanding the interaction. A second reason could be that, although split testing method was applied, more than one variable was changed, which may have blurred the results. Focusing on each crossmodal coupling separately may be a more advantageous future evaluation method. Lastly, correlations such as the association of loudness with greater size, and faster tempo with more muscular energy, were not included in the first iteration. These mappings, and other overall gestural interactions with the system could benefit from further research and evaluation in terms of their use within the looper's design.

6.2 Future Directions

In future iterations of the multimodal looper, an imperative focal point should be improving the analysis of gestures. Building upon the idea of the gestural-sonorous categories, future considerations should include better integration and separation of these categories within a continuous stream of movement. This can be done through the use of interactive machine learning software such as Rebecca Fiebrink's Wekinator, or deep learning methods for pattern recognition and classification[31]. For example, Jiuqiang Tang uses as a sliding window approach with multiple dynamic time warping (DTW) instances for continuous gesture recognition, enabling adaptable starting points. The system preprocesses the data it is analyzing by considering where the gesture or movement began and compares it with DTW. If similarity meets a threshold established by the training data, potential matches are identified and optimal result is determined through the K-Nearest-Neighbor algorithm[32]. Furthermore, for a more intuitive gestural interaction in a real-time setting, the analysis of higher-level movement qualities, and expressive qualities of movement through frameworks such as Laban Movement Analysis, could help expand the model's capacity to meet the level of variety and expression shown by the participants.

As mentioned in the previous section, musical structure should be considered further and visualized better in the future iterations. One way to achieve a better visualization would be to map time on an inner circle rotating clockwise and placing the AV objects according to when they are played within their loop cycle. It can also be helpful to visualize the timings within the musical structure to make the system more comprehensible. Interaction with the musical structure should also be considered further for a system involving more controls such as changing the tempo or manipulating the loops. One framework that could be beneficial to employ for this purpose could be the distributed mapping approach of the previously described (1.3.1) MapLooper [21]. Evolving towards a more adaptable musical structure and adding further generative functionality could also be done through the integration of music specific software such as Ableton, or machine improvisation software such as MASOM (Musical Agent based on Self-Organizing Maps), which can learn musical structure during live performances through self-organizing maps [33].

Moreover, the interaction with the multimodal objects within the musical structure can benefit from further consideration. As it becomes a bit too repetitive, instead of playing each input 7 times, there could be a way to take the objects out of the composition or change their timing, placement, shape and effects within the loops. This could help build further crossmodal mappings within the interaction. For example, changes in placement can affect the pitch, and changing the visual texture

can influence the timbre of the sounds. Rather than the gestures only creating the objects, this would allow the users to have more control and compositing possibilities. Whether this will enhance the improvisational experience would then be a matter of investigation.

Finally, for the circular design of the looper and the turn-taking system to make sense, the next iteration should allow multiple users and/or an optional turn for deep learning algorithms for music generation such as MusicGen or MusicLM to continue the melodies created by the users [26][34]. The visualizations could also be enhanced through other mediums than projection that would allow them to be three-dimensional, such as AR (Augmented Reality). Alternatively, although still posing difficulty for real-time animation, open-source image generation algorithms such as Stable Diffusion can be used to generate ever evolving forms and shapes with their parameters determined by the gestures [27]. Incorporation of these methods could help create a collaborative system where, each multimodal object can be manipulated, as well as interpolated to evolve together with inputs from other users; with each collaborative experience resulting in a unique composition.

Chapter 7

Conclusion

The aim of my thesis was to investigate the potential of an interactive system for creating music through multiple sensory modalities. Aiming to build a framework for an intuitive system, I based my research on crossmodal relationships of sound and music within an embodied paradigm, where music cognition is interwoven with kinesthetic and visual modalities. This led to the development of multimodal objects; sound fragments that can be triggered gesturally and represented visually according to their gestural-sonorous features within a live looping system. The user evaluation revealed that the system did not exhibit a significant ability to construct intuitive cross-modal mappings, or to provide an integrated multisensory experience. Rather, a considerable amount of confusion concerning the interaction was experienced by the participants. The qualitative findings revealed key limitations of the prototype that may have caused this perplexity; namely, insufficient analysis and integration of gestures within a live performance setting, deficient visualization of musical structure in terms of the placement of the objects, and a disparity between the multiple user aspects of the design and the single user evaluation procedure. These revelations led to valuable insights concerning further refinement and integration of each of the prototype's modules, as well as inspiration for added functionalities for future iterations.

In conclusion, while the system may not have achieved great success, this does not mean that building intuitive multimodal experiences for musical improvisation is an unattainable goal. In fact, with further functionality and refinement, the use of gestural-sonorous features combined with common cross-modal associations remains as an exciting opportunity to be explored further within the context of interactive visual music systems, such as the multimodal looper.

Chapter 8

Bibliography

[1] Rolf Godøy. Gestural-sonorous objects: Embodied extensions of schaeffer's conceptual apparatus. *Organised Sound*, 11:149 – 157, 08 2006.

[2] Mats Küssner. *Shape, drawing and gesture: Cross-modal mappings of sound and music*. PhD thesis, King's College London, UK, 09 2014.

[3] Jelle van Dijk, Remko van der Lugt, and Caroline Hummels. Beyond distributed representation: Embodied cognition design supporting socio-sensorimotor couplings. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction*, TEI '14, page 181–188, New York, NY, USA, 2014. Association for Computing Machinery.

[4] Jelle Van Dijk. Designing for embodied being-in-the-world: A critical analysis of the concept of embodiment in the design of hybrids. *Multimodal Technologies and Interaction*, 2(1), 2018.

[5] Marc Leman. *Embodied Music Cognition and Mediation Technology*. The MIT Press, 2007.

[6] R.I. Godøy and M. Leman. *Musical Gestures: Sound, Movement, and Meaning*. Taylor & Francis, 2009.

[7] Alexander Jensenius, Marcelo Wanderley, Rolf Godøy, and Marc Leman. Musical gestures: concepts and methods in research. *Musical Gestures: Sound, Movement, and Meaning*, 01 2009.

[8] Richard Middleton. Popular music analysis and musicology: Bridging the gap. *Popular Music*, 12(2):177–190, 1993.

[9] R.S. Hatten. *Interpreting Musical Gestures, Topics, and Tropes: Mozart, Beethoven, Schubert*. Musical Meaning and Interpretation. Indiana University Press, 2004.

[10] J.J. Gibson. *The Ecological Approach to Visual Perception*. Resources for ecological psychology. Lawrence Erlbaum Associates, 1986.

[11] Rolf Godøy. Gesture affordances of musical sound. *Musical Gestures: Sound, Movement, and Meaning*, 01 2010.

[12] Rolf Inge Godøy. Sonic object cognition. In Ralf Bader, editor, *Springer Handbook of Systematic Musicology*. Springer, Berlin, Heidelberg, 2018.

[13] B. Kane. *Sound Unseen: Acousmatic Sound in Theory and Practice*. Oxford University Press, 2014.

[14] P. Schaeffer. *Traité des objets musicaux*. Pierres vives. Editions du Seuil, 2016.

[15] Charles Spence. Crossmodal correspondences: A tutorial review. *Attention, perception & psychophysics*, 73:971–95, 05 2011.

[16] Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, Gary Lupyan, Grace E. Oh, Jing Paul, Caterina Petrone, Rachid Ridouane, Sabine Reiter, Nathalie Schümchen, Ádám Szalontai, Özlem Ünal-Logacev, Jochen Zeller, Marcus Perlman, and Bodo Winter. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377, 2021.

[17] Andrew Bremner, Serge Caparos, Jules Davidoff, Jan Fockert, Karina Linnell, and Charles Spence. "Bouba" and "Kiki" in Namibia? A remote culture make similar shape-sound matches, but different shape-taste matches to Westerners. *Cognition*, 126, 10 2012.

[18] Daphne Maurer, Thanujeni Pathman, and Catherine Mondloch. The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental science*, 9:316–22, 05 2006.

[19] Peter Walker, J. Bremner, Uschi Mason, Jo Spring, Karen Mattock, Alan Slater, and Scott Johnson. Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological science*, 21:21–5, 01 2010.

[20] Dafna Kohn and Zohar Eitan. Musical parameters and children's movement responses. In Päivi-Sisko Eerola, editor, *ESCOM 2009 : Abstracts I& programme*, pages 233–241. Dep. of Music, Univ. of Jyväskylä, 2009.

[21] Christian Frisson, Mathias Bredholt, Joseph W. Malloch, and Marcelo M. Wanderley. Maplooper: Live-looping of distributed gesture-to-sound mappings. In *New Interfaces for Musical Expression*, 2021.

[22] Thomas Mitchell and Imogen Heap. SoundGrasp : A Gestural Interface for the Performance of Live Music. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 465–468. Zenodo, June 2011.

[23] Mary Mainsbridge and Kirsty A. Beilharz. Body as instrument: Performing with gestural interfaces. In *New Interfaces for Musical Expression*, 2014.

[24] Brian Evans. Foundations of a visual music. *Computer Music Journal*, 29(4):11–24, 2005.

[25] Sergi Jordà. The reactable: tangible and tabletop music performance. *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, 2010.

[26] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2023.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[28] Ziyun Cai, Jungong Han, Li Liu, and Ling Shao. RGB-D datasets using Microsoft Kinect or similar sensors: a survey. *Multimedia Tools and Applications*, 76, 02 2017.

[29] Wenxia Bao, Zhongyu Ma, Dong Liang, Xianjun Yang, and Tao Niu. Pose ResNet: A 3D human pose estimation network model. *2023 2nd International Conference on Big Data, Information and Computer Network (BDICN)*, pages 264–267, 2023.

[30] Mohammed Usman. Power efficiency of Sign Test and Wilcoxon Signed Rank Test relative to T-Test. *Mathematical theory and modeling*, 5:53–59, 2015.

[31] Rebecca Fiebrink and Perry Cook. The wekinator: A system for real-time, interactive machine learning in music. *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)*, 01 2010.

[32] Jiuqiang Tang and Roger B. Dannenberg. Extracting commands from gestures: Gesture spotting and recognition for real-time music performance. In *Computer Music Modeling and Retrieval*, 2013.

[33] Kıvanç Tatar, Philippe Pasquier, and Remy Siu. Revive: An audio-visual performance with musical and visual ai agents. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[34] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

List of Figures