VITRUVIAN WHOLE-BODY CREATIVE ACTION:

A PROPOSAL FOR THE STRUCK-STRING INTERACTION FRAMEWORK

SUTIRTHA CHAKRABORTY · DAMIÁN KELLER · JOSEPH TIMONEY[1]

ABSTRACT

The Vitruvian Creative-Action Metaphor is proposed as an approach to tackle whole-body interaction with piano-like sounds. We report on the design of a visual-tracking adaptive mechanism, implemented as a camera-based sys- tem deployable on browser technology. Our presentation documents the initial prototypes, available for usage by the ubimus community.

1 INTRODUCTION

Recent advances in ubimus research point to the emergence of an initiative targeting the development of infrastructure and the exploration of piano timbre without resorting to the acoustic-instrumental concepts tied to interacting with keyboards.

This tendency to equate musical interaction to the usage of acoustic instruments has not diminished, despite the significant changes im- plied by the adoption of mobile, embedded and network-based infrastructure. Given this state of affairs, various initiatives within our community are starting to

1. Ubiquitous Music Group - October 2024 - March 2025

unveil concepts and techniques that offer an alternative to the acoustic-instrumental perspective to deal with piano sounds: the struck-string interaction framework [Kramann and Keller, 2024], [Su et al., 2024].

Struck-string interaction involves a loose set of strategies for handling piano sounds, with a strong emphasis on the reduction of the cognitive load and the temporal investment typically demanded by the virtuosic mindset of the instrumental approach. We are interested in expanding the musical possibilities of the parametric control of piano-like sonic resources, encompassing both in-place and remote opportunities for shared musical experiences. Hence, our proposal involves revisiting well-established practices and also fostering an overhaul of what is understood as ”piano music”.

In this paper, we introduce the Vitruvian Creative-Action Metaphor within the context of struck-string interaction (henceforth ssi). The remaining sections encompass a presentation of the key characteristics of the Vitruvian Metaphor, a documentation of the computational components, highlighting the potentials and caveats of the current imple- mentation. We provide details of the system architecture, the adaptive components and their interplay.

2 RELATED WORK

Gesture- and body-driven mappings to musical sound have been investigated across various sensing modalities. Markerless, camera-based interfaces paired with interactive machine learning frameworks have enabled performers to map full-body motion to audio and visuals. For example, Schedel et al. demonstrated using the Wekinator toolkit to

translate Microsoft Kinect skeletal data into musical and visual outputs for the ensemble 000000Swan [Schedel et al., 2011]. Wearable IMU-based gloves, typified by the Mi. Mu Gloves designed by Imogen Heap's team, employ flex and orientation sensors to capture fine-grained hand and arm movements for expressive, low-latency control[2]. Non-contact ultrasonic systems have also been used for gesture recognition: Sang et al. proposed an ultrasonic active sensing array for hand-gesture classification, achieving a performance benchmark suitable for interactive applications [Sang et al., 2017].

These systems feature a mapping layer that transforms raw sensor data into sound synthesis or parametric control. Hunt and Wanderley introduced a two-layer model (a) separating sensor-specific feature extraction from synthesis parameter assignment (b) to decouple interface design from sound generation [Hunt and Wanderley, 2002]. Building on this, Verfaille et al. formulated a multi-level mapping framework for au- dio processing, distinguishing gestural control from adaptive, sound-driven modulation [Verfaille et al., 2006]. Our Struck-String Interaction (ssi) framework extends these approaches by adopting a Vitruvian Man–inspired creative-action metaphor, unifying discrete (inner-zone) and continuous (outer-zone) mappings in a browser-native, camera-based computational environment.

3 THE VITRUVIAN CONCEPT

Our starting point consists of stripping away the

2. https://mimugloves.com

requirements of piano-sound deployments to a bare minimum. This Ockham-Razor exercise prompts us to tackle interaction support as a three-component ecosystem: a human body that generates information by means of movements, a tracking component that translates these movements into data, and a rendering component that furnishes the sonic outcomes and informational feedback necessary to adjust the participants' behaviours to local demands and aesthetic goals. We emphasise that no genre prescriptions are adopted other than a focus on piano-related sounds (cf. [Chakraborty et al., 2022] and [Keller et al., 2023] for a similar approach ap- plied in banging interaction).

A further constraint of our proposed design entails enhancing the sustainabil- ity of infrastructure. As proposed by other ubimus frameworks, we employ browser- based technologies that are compatible with stationary, embedded or mobile devices [Lazzarini et al., 2014] [Lazzarini et al., 2020] [Yi and Letz, 2020].

Drawing inspiration from Da Vinci's (1487) painting, The Vitruvian Man - which epitomizes human-body proportions through geometric ratios - the proposed creative- action metaphor defines spatial interaction zones to determine the type and character of musical outcomes. The Vitruvian Metaphor allows participants to perform struck-string sonic models by means of simple body movements. It employs a webcam to enable MediaPipe-based pose detection and it incorporates MIDI protocols for parametric control and audio synthesis. Key features include the following items.

- **Pose Tracking:** Synchronous detection of body, hands, and face landmarks.

- **Movement Tracking and Gesture Detection:** Algorithms to compute motion-speed, direction and positional changes.

- **Vitruvian Visuals:** A dynamically scaled square with inner and outer circular zones aimed at whole-body interaction.

- **MIDI Integration:** Configuration of MIDI channels and parametric control of piano-like audio-synthesis models.

- **Visual Feedback:** On-screen graphics that display geometric containers, labels, and colour-coded visual anchors to complement proprioceptive cues.

4 SYSTEM ARCHITECTURE

The current implementation is structured as a real-time processing pipeline that converts video input into musical output, furnishing visual feedback of body actions. The primary components are as follows.

4.1 OVERVIEW

Figure 1 depicts the high-level data flow.

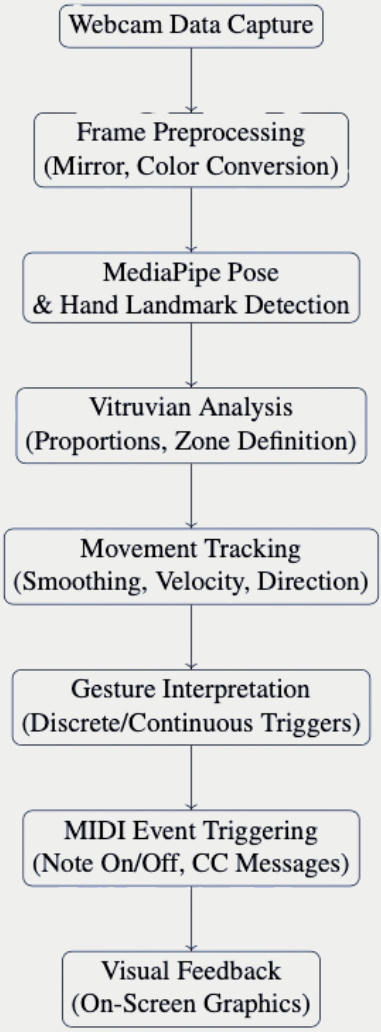

FIGURE 1: VITRUVIAN PROTOTYPE WORKFLOW: FROM VIDEO CAPTURE TO MIDI OUTPUT AND VISUAL FEEDBACK

4.2 DETAILED COMPONENTS

- **Webcam Capture:** Utilizes OpenCV to obtain a live video feed.

- **Frame Preprocessing:** Applies mirroring and colour space conversions.

- **Pose Detection:** Leverages MediaPipe's holistic model to extract 33 body tokens, 21 per hand tokens, and 468 face tokens.

- **Vitruvian Analysis:** Computes a dynamic Vitruvian container based on the user's height (distance from head to feet) and overlays a square with two concentric circles defining interaction zones.

- **Movement Tracking:** Uses a MovementTracker class (employing a circular buffer via Python's deque) to calculate speed, direction, and trigger conditions.

- **Gesture Interpretation:** Differentiates between continuous (e.g., sustained index finger position) and discrete (e.g., foot or wrist movement) gestures.

- **MIDI Communication:** Sends MIDI messages through a virtual port using the mido library, mapping gestures to different channels.

- **Visual Feedback:** Renders real-time graphics – including the Vitruvian container, zones, labels, and colour panels – through the video feed.

5 POSE DETECTION WITH MEDIAPIPE

Pose detection is a core functionality. MediaPipe's Holistic

model is configured with a minimum detection confidence of 0.6 and a tracking confidence of 0.6 to robustly extract various landmarks. A key feature of the model is to keep a consistent tracking behaviour, regardless of the distance between the device and the subject. Thus, within the camera's field of vision, whole-body motion is supported.

- **Body Landmarks:** Nose, shoulders, hips, wrists, and feet determine the overall pose and Vitruvian scaling.

- **Hand Landmarks:** The right-hand index fingertip is crucial for fine gesture con- trol.

- **Face Landmarks:** Although not directly used for MIDI mapping, they enhance tracking reliability.

Figure 2 (a placeholder diagram) illustrates the key landmark groups.

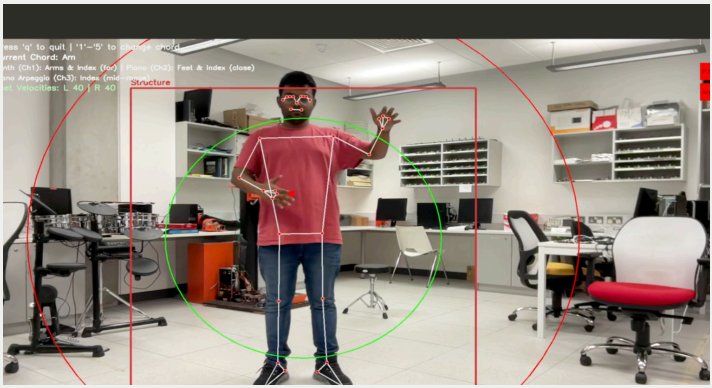

FIGURE 2: SYNCHRONOUS MOTION CAPTURE AND INTERACTION INTERFACE FROM THE VITRUVIAN CREATIVE-ACTION METAPHOR. THE PERFORMER IS TRACKED USING SKELETAL KEYPOINTS, WITH DISTINCT ZONES (HIGHLIGHTED IN RED AND GREEN) MAPPED TO VARIOUS PIANO-LIKE SOUND MODELS. THIS SYSTEM ENABLES EXPRESSIVE FULL-BODY MUSICAL INTERACTION THROUGH SPATIAL BODY-MOTION RECOGNITION

6 MOVEMENT TRACKING AND GESTURE DETECTION

The system relies on accurately tracking movements and interpreting gestures. A dedicated *MovementTracker* class processes sequential landmark data to compute speed of motion and direction.

6.1 FEET AND WRIST TRACKING

Feet: Movements are tracked using a 15 position buffer.The average y-coordinate difference (speed of movement) is mapped to MIDI velocity. Direction changes (e.g., upward vs. downward movement) trigger piano onset events on MIDI Channel 2.

6.2 INDEX FINGER TRACKING

The right-hand index finger provides nuanced control based on both its position and movement:

- Crossing Detection: A function computes the side of a point relative to a line (from the right shoulder to the right hip) using the following cross product.

$$side=(Bx-Ax)(Py-Ay)-(By-Ay)(Px-Ax)$$

A sign change indicates that the index finger has crossed a defined boundary.

- **Distance Zones:** The distance from the body centre is used to define three zones.

Far Zone **(outside of the outer circle):** Activates a continuous synth-pad note (Channel 1).

Close Zone **(inside the inner circle):** Triggers a piano event (Channel 2).

Intermediate Zone **(between the circles):** Initiates a piano arpeggio (Channel 3).

Figure 3 outlines the movement-tracking algorithm.

7 VITRUVIAN VISUALS AND PROPORTIONAL ANALYSIS

Inspired by Da Vinci's Vitruvian Man, the system dynamically computes a Vitruvian Size based on the distance between the subject's head and feet. This measurement scales a square - the Vitruvian container – drawn on the screen. Within this square, two concentric circles are overlaid.

- **Inner Circle:** Defines a zone for discrete, percussive triggers.
- **Outer Circle:** Establishes the boundary for continuous sound control.

Figure 4 illustrates the geometric layout.

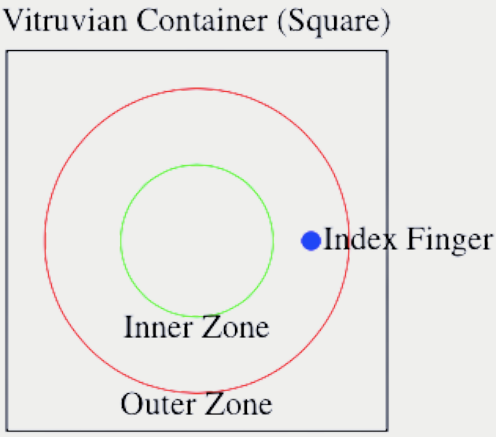

FIGURE 3: VITRUVIAN VISUALS: A SQUARE AND INNER AND OUTER CIRCULAR ZONES DEFINING GESTURE-BASED CONTROL AREAS

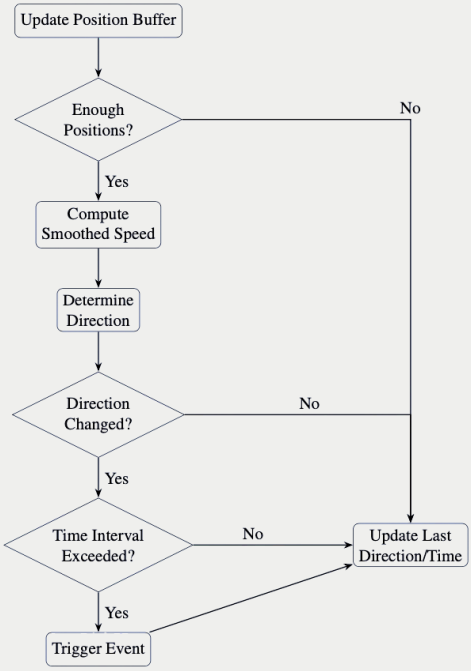

FIGURE 4: MOVEMENT TRACKER ALGORITHM: PROCESSING POSITION DATA TO TRIGGER MIDI EVENTS

8 MIDI COMMUNICATION AND SONIC RENDERING

The system converts body gestures into musical output by means of MIDI events. MIDI messages are implemented using the Python mido library, routed over a virtual MIDI port.

8.1 MIDI CHANNEL MAPPING

MIDI channels are assigned to various sound types:

• **Channel 1 (mido channel 0):** Synth or warm pad sounds. Triggered by arm/wrist movements and continuous note mode (when the index finger is far).

• **Channel 2 (mido channel 1):** Piano events. Activated by feet movements or when the index finger crosses into the inner zone.

• **Channel 3 (mido channel 2):** Piano arpeggios. Initiated when the index finger is positioned between the inner and outer circles.

Table 1 summarizes the parametric MIDI-based mapping.

MIDI Channel	Instrument	Trigger Source & Description
1 (Ch 0)	Synth / Warm Pads	Activated by arm/wrist movements and continuous note triggers when the index finger is far.
2 (Ch 1)	Piano Strum	Triggered by foot movements and discrete index finger crossings within the inner zone.
3 (Ch 2)	Piano Arpeggio	Initiated by the index finger in the intermediate zone (between circles).

TABLE 1: MIDI CHANNEL MAPPING AND ACTIONS

8.2 MIDI WORKFLOW DIAGRAM

The following diagram (Figure 5) illustrates the mapping from gesture detection to MIDI event dispatch.

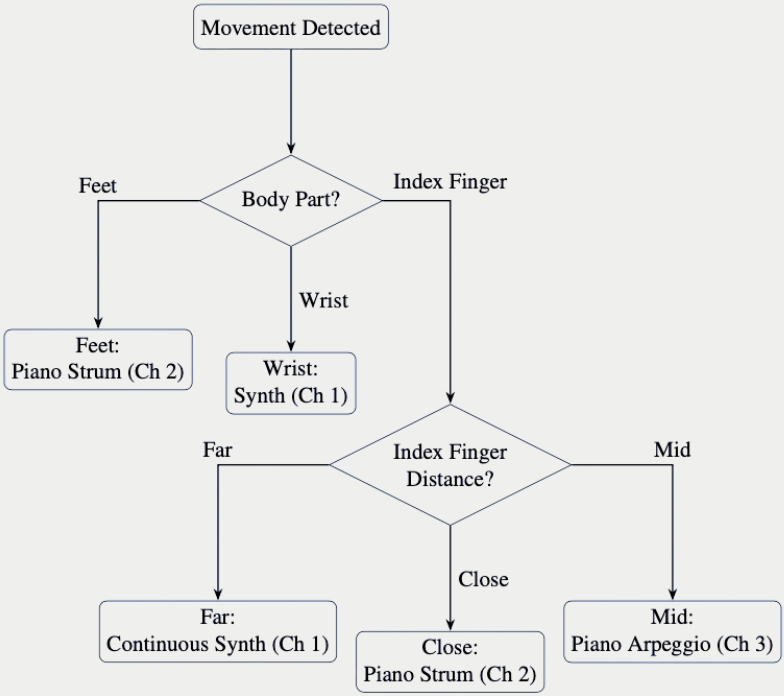

FIGURE 5: MIDI TRIGGER MAPPING: FROM BODY PART DETECTION TO SOUND GENERATION ON DESIGNATED CHANNELS

8.3 MOVEMENT–TO–SOUND MAPPING RATIONALE

Each zone's behaviour of prototype v.1 draws on common practice in gesture-based de- sign.

- **Inner Zone (0–30% radial range):** discrete index-finger crossings trigger piano onsets, leveraging the fine motor affordances of finger taps.

- **Intermediate Zone (30% - 60%):** index-finger position modulates arpeggios, preserving pitch continuity while allowing control of melodic contour.

- **OuterZone (60% - 100%):** sustained arm/wrist motion controls synth-pad timbre and volume, aligning continuous spatial gestures with sustained events.

We frame these choices as hypotheses grounded in gesture-based common practice, to be validated in future empirical studies.

8.4 MIDI MODULE

The MIDI module tackles the basic functionality for the generation of piano-like events.

1. **Initialization:** A virtual MIDI port is opened (e.g., using mido.open output).

2. **Program Changes:** Instrument settings are configured via program change messages (e.g., Program 90 for synths, Program 1 for piano).

3. **Onset Triggering:** On detecting a gesture, a note on message (with velocity proportional to movement speed) is sent, often accompanied by control change (CC) messages (e.g., reverb on CC 91 or modulation on CC 1).

4. **Note Off:** After a predetermined duration for discrete events, a note off mes- sage is dispatched.

9 VISUAL FEEDBACK

Visual feedback reinforces the interaction by overlaying dynamic graphics onto the video feed.

- **Vitruvian Container:** A square scaled to the user's height, representing ideal proportions.
- **Concentric Zones:** Two circles (inner and outer) demarcate regions for discrete and continuous triggers.
- **Labels and Colour Panels:** On-screen text displays pitch-aggregates names, movement velocities, and MIDI pitches, while a colour panel maps MIDI octaves to specific colours (e.g., blue for lower octaves, red for higher octaves).

10 VITRUVIAN CONCEPTS AND THEIR IMPLEMENTATION

The system's aesthetic and functional design is inspired by Leonardo da Vinci's The Vitruvian Man.

- **Proportional Scaling:** The distance from the head to the feet is used to dynamically size the Vitruvian container (square), ensuring the visual feedback remains proportional to the user.
- **Zone Allocation:** Two concentric circles within the square define distinct interaction areas:

    Inner Zone: Reserved for discrete, percussive triggers (e.g., piano strums or onsets).

Outer Zone: Engages continuous sound control (e.g., sustained synth-pad sounds).

- **Symbolism:** The geometric forms not only serve a functional role but also suggest embodiment of musical actions, highlighting their relationship to human anatomy.

Figure 6 illustrates the integration of the visual Vitruvian components.

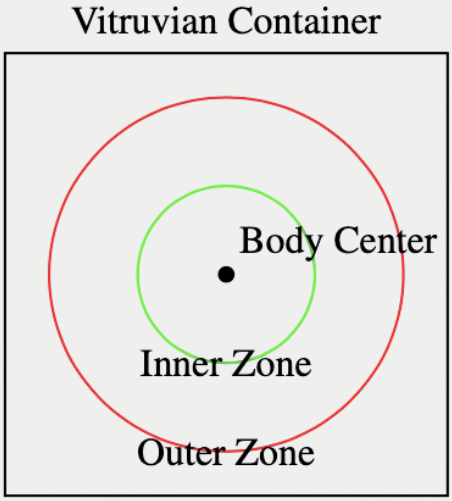

FIGURE 6: VITRUVIAN FRAMEWORK: DYNAMIC SCALING AND ZONE ALLOCATION INSPIRED BY MODELLED HUMAN PROPORTIONS

10.1 CAMERA-BASED TRACKING: RATIONALE AND LIMITATIONS

We chose a browser-native, webcam-only approach for maximum accessibility. These are some of the trade-offs.

- **Latency:** Typical per-frame delay of 30-50 ms at 30 fps (mitigated by downsam- pling to 15 fps for tracking loops).

- **Orientation Constraint:** Performers must face the camera; we mitigate this via on-screen alignment guides and wide-angle lens correction.

- **No Wearables:** Unlike IMU gloves, our system requires zero setup after opening a web-based interface.

11 ALGORITHMS AND IMPLEMENTATION DETAILS

Several key algorithms enable the conversion of body movements into musical outputs.

11.1 MOVEMENTTRACKER ALGORITHM

The *MovementTracker* class employs a circular buffer (using *Python's deque*) to store recent positions and computes:

- **Speed:** Calculated as the average Euclidean distance between successive positions:

$$\text{speed} = \frac{1}{n-1} \sum_{i=1}^{n-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$$

- **Direction:** Inferred by comparing the first and last positions in the buffer.

- **Trigger Conditions:** A note is triggered only if a significant change in direction occurs and a minimal time interval (e.g., 0.25 seconds) has passed.

11.2 DISTANCE AND CROSSING CALCULATIONS

- **Distance from Center:** The Euclidean distance of the index finger from the body center is computed as:

$$\text{dist} = \sqrt{(x_f - x_c)^2 + (y_f - y_c)^2}$$

- **Point-Line Crossing:** Using the cross product method (as shown earlier), the system determines when the index finger crosses a defined boundary.

11.3 PITCH SELECTION AND SCALE MAPPING

Musical events are selected based on user-defined scales and chords. Table 2 lists example scales employed by the current prototype.

Chord	Scale Notes
Am	[57, 60, 62, 64, 65, 67, 69, 72]
C	[60, 62, 64, 65, 67, 69, 71, 72]
G	[55, 57, 59, 60, 62, 64, 66, 67]
F	[53, 55, 57, 58, 60, 62, 64, 65]
Em	[52, 55, 57, 59, 60, 62, 64, 67]

TABLE 2: EXAMPLE MUSICAL SCALES FOR PITCH SELECTIONS

12 PRELIMINARY DISCUSSIONS: CAVEATS AND POSSIBILITIES

Musical events are selected based on user-defined scales and chords. Table 2 lists example scales employed by the current prototype. We report on the design of the Vitruvian Creative-Action Metaphor - proposing a prototype based on computer vision tracking, synchronous pose tracking and MIDI connectivity to turn human gestures into whole-body musical experiences. Inspired by Leonardo da Vinci's The Vitruvian Man, human-body proportions and movements are captured through adaptive techniques based on a combination of MediaPipe's holistic landmark detection, custom movement-tracking and a MIDI-based communication.

In this report we described the prototype's architecture, algorithmic approaches and the ubimus strategies required to handle piano-like sounds. We have shared our tool within a research focus-group. A working prototype can be accessed online. Our aim with this presentation is to gather feedback from the ubimus community on the proposed concepts, the design choices and the targeted types of musical experiences.

During our discussions toward the early deployments of the Vitruvian prototype, two approaches to implementation emerged: visual tracking and accelerometer-based sensing. Both techniques feature advantages and limitations within the context of struck-string interaction. Arguably, visual tracking might present some caveats when dealing with finger movements. This is an area of active research and our demonstrations will feature problems and solutions.

Another aspect to consider is the increased cognitive demands of multiple simultaneous body movements while tackling complex musical events, such as pitch-aggregates that feature multiple dynamic levels in sequence-based interactions. We address strategies to tackle these issues.

The Vitruvian Metaphor is proposed as an approach to struck-string interaction that furnishes a simplification of means for parametric control which is deployable on browser-based technology. The implemented prototype supports interaction with piano- like sounds through visual tracking of body movements. Alternative approaches can complement our proposal. We look forward to suggestions and insights from the ubimus community to expand and refine the vitruvian design.

13 FUTURE WORK

Future work involves quantifying the cognitive load of multi-limb gestures using assess- ment tools such as the NASA-TLX and dual-task techniques. We will also tackle artistic applications, collecting performance recordings, participants feedback, and expert analyses, to gauge its creative caveats and potentials. By profiling the required skill levels, we will develop adaptive gesture-to-sound techniques tailored for novices and experts. We will also incorporate frameworks from ubimus theory to organize and justify our movement→sound assignments and interaction-design decisions. Finally, to support multiple approaches to temporalities, we will optimize the tracking algorithms and adjust the WebRTC settings to drive end-to-end latency below 20 ms.

REFERENCES

Chakraborty, S., Yaseen, A., Timoney, J., Lazzarini, V., and Keller, D. (2022). Adaptive touchless whole-body interaction for casual ubiquitous musical activities. In Proceedings of the International Computer Music Conference (ICMC2022), pages 132–138. Limerick, Ireland: University of Limerick.

Hunt, A. and Wanderley, M. M. (2002). The importance of parameter mapping in electronic instrument design. Organised Sound, 7(2):97–108.

Kramann, G. and Keller, D. (2024). Struck-string interaction: Exploring piano timbre beyond the piano. In Proceedings of the Ubiquitous Music Symposium (UbiMus 2024). Macau, China: Ubiquitous Music Group.

Lazzarini, V., Costello, E., Yi, S., and ffitch, J. (2014). Development tools for ubiquitous music on the world wide web. In Ubiquitous Music, Computational Music Science, pages 111–128. Heidelberg Berlin: Springer International Publishing.

Lazzarini, V., Keller, D., Otero, N., and Turchet, L. (2020). Ubiquitous Music Ecologies. London: Taylor & Francis (Routledge).

Sang, Y., Shi, L., and Liu, Y. (2017). Micro hand gesture recognition system using ultrasonic active sensing. arXiv preprint arXiv:1712.00216.

Schedel, M., Perry, P., and Fiebrink, R. (2011). Wekinating 000000swan: Using machine learning to create and control complex artistic systems. In Proceedings of the International Conference on New Interfaces for Musical Expression (NIME), pages 453–456, Oslo, Norway.

Su, R., Timoney, J., and Keller, D. (2024). Just intonation and inharmonic- ity in struck-string interaction. In Proceedings of the Ubiquitous Music Symposium (UbiMus 2024). Macau, China: Ubiquitous Music Group.

Verfaille, V., Wanderley, M. M., and Depalle, P. (2006). Mapping strategies for gestural and adaptive control of digital audio effects. Journal of New Music Research, 35(1):71–93.

Verfaille, V., Wanderley, M. M., and Depalle, P. (2006). Mapping strategies for gestural and adaptive control of digital audio effects. Journal of New Music Research, 35(1):71–93.

Yi, S. and Letz, S. (2020). The browser as a platform for ubiquitous music. In Lazzarini, V., Keller, D., Otero, N., and Turchet, L., editors, Ubiquitous Music Ecologies, pages 185–206. London: Routledge.