A Human-Agents Music Performance System in an Extended Reality Environment

Pedro Lucas
RITMO Centre for Interdisciplinary Studies in Rhythm, Time, and Motion
Department of Informatics
University of Oslo
Oslo, Norway
pedroplu@uio.no

Stefano Fasciani
Department of Musicology
University of Oslo
Oslo, Norway
stefano.fasciani@imv.uio.no

ABSTRACT

This paper proposes a human-machine interactive music system for live performances based on autonomous agents, implemented through immersive extended reality. The interaction between humans and agents is grounded in concepts related to Swarm Intelligence and Multi-Agent systems, which are reflected in a technological platform that involves a 3D physical-virtual solution. This approach requires visual, auditory, haptic, and proprioceptive modalities, making it necessary to integrate technologies capable of providing such a multimodal environment. The prototype of the proposed system is implemented by combining Motion Capture, Spatial Audio, and Mixed Reality technologies. The system is evaluated in terms of objective measurements and tested with users through music improvisation sessions. The results demonstrate that the system is used as intended with respect to multimodal interaction for musical agents. Furthermore, the results validate the novel design and integration of the required technologies presented in this paper.

Author Keywords

Interactive Music Systems, Multi-Agent Systems, Swarm Intelligence, Motion Capture, Spatial Audio, Mixed Reality, HoloLens

CCS Concepts

•**Applied computing** → **Sound and music computing; Performing arts;** •**Human-centered computing** → *Mixed / augmented reality;*

1. INTRODUCTION

Music performances with Digital Musical Instruments (DMI) expanded into a virtual 3D space can improve expressiveness and comprehensibility [10]. A human-machine music collaboration in a physical-virtual space can take advantage of 3D elements in terms of both sound and visuals.

This concept requires a multimodal interactive platform allowing the embodiment of sound sources as individual entities capable of interacting with a human performer. The development of such a platform presents challenging requirements with respect to the integration of *input modalities* and a proper *output generation*. It is essential that the output provides consistent feedback to users in terms of what they hear and see [6].

To take advantage of the 3D space and address the development challenges, we present an Interactive Music System (IMS) allowing human-machine music performances, in which the machine is depicted as *Musical Agents* represented as 3D objects. The performer interacts with the agents in real-time through a multimodal physical-virtual environment based on Extended Reality (XR) technologies. The contribution of this work is twofold: the *conceptual design* and *technical implementation* of a human-agents IMS in a 3D multimodal space.

The *conceptual design* is based on a *Multi-Agent* and *Swarm Intelligence* approach in a platform that integrates *Motion Capture*, *Spatial Audio*, and *Mixed Reality (MR)* technologies. These address limitations found in previous XR-related works such as: imprecision when interacting with virtual objects [15], inaccurate manipulation of sound parameters [21], and limited pre-trained AI models in music performances [7, 10].

The *technical implementation* reflects the design in a proof-of-concept prototype that includes hardware and software components of the above-mentioned technologies. Through this prototype, the proposed system is evaluated in terms of *effectiveness* and *efficiency* using objective measurements and tested with users by capturing data from live sessions and surveys.

The results between the measurements and user data are compared to assess the fluency of the experience. Moreover, captured data related to user attention and interaction during a performance allows analyzing the synergy between a user and the artificial agents. The objective is to validate the proposed way-of-making such a system as a human-agents platform for interactive music performances, contributing to designing new NIMEs based on XR.

The rest of the paper is organized as follows: Section 2 describes relevant works. Section 3 illustrates the system design. Section 4 specifies the implementation of the proof-of-concept prototype. The evaluation is detailed in Section 5, and finally, Section 6 presents conclusions and future work.

2. RELATED WORK

This work builds on the concept of *Musical Agents* as entities that automate creative tasks and interact with their environment [19]. These agents can be modeled as a *Multi-Agent* system, composed of simple units, that exhibit emerging intelligence to perform complex tasks [18].

NIME'23, 31 May–2 June, 2023, Mexico City, Mexico.

To enable interaction between humans and autonomous systems, Blackwell *et al.* developed the concept of a *Live Algorithm* in music [2]. Such systems can be expressed in a *PQf* architecture (P for analysis, Q for synthesis, and f for patterning) as used in the conceptual design proposed in this paper.

An additional field relevant to this work is *Music in Extended Realities* (*Musical XR*), which is rapidly developing, leveraging technologies such as Augmented Reality (AR), Augmented Virtuality (AV), Virtual Reality (VR), and Mixed Reality (MR) for applications that combine audio and music components with 3D representations of the physical world, necessitating proper spatialization techniques for environmental composition [20]. The integration of multiple modalities, such as visual, auditory, haptic, and proprioceptive modalities, can lead to immersive and engaging experiences in a musical context.

This paper explores the potential of Mixed Reality in combination with additional platforms, such as Optical Motion Capture (MOCAP) and Spatial Audio, to improve the perception of embodied experiences. A variety of works related to these technologies have been developed for sound and music applications. For instance, Costa [4] proposes a system that tracks head position to provide an accurate panning effect for binaural sounds, while The BoomRoom [13] allows users to manipulate sound in mid-air through real objects tracked in space. Spatial audio technologies have been adopted for virtual environments, where Grani et al. [8] found that audio-visual attractors can efficiently capture users' attention. Additionally, spatial audio can be used for environmental communication in human-robot interaction, as presented by Robinson [17]. These works can be improved by including feedback in the visual domain, as proposed in our integrated system.

In terms of MR, Hamilton et al. [9] describe environments that include networked music performances and spatial audio for the geolocalization of virtual spaces, where spatial coordinates are used to control music parameters. MR devices such as the Microsoft HoloLens[1] headset are used in music applications to support accurate augmented reality, such as the piano learning system from Das et al. [5] or to map virtual objects with sound properties for interaction, as in the work of Nakagawa et al. [14]. Furthermore, Riley [16] describes MR applications developed on HoloLens to explore music affordances between the physical and visual world, together with multi-track instrument mixing on inspirational landscapes. Moreover, recent NIME works make use of MR to explore the affordances of virtual objects and gestures mapped with sound and music properties [21], or introduce AI components as part of a music improvisation process [10]. Additionally, the analysis of frameworks and ergonomics for NIMEs based on MR offers important contributions to other Musical XR applications [22, 7].

This work integrates motion capture, spatial audio, and mixed reality technologies to implement a system based on a novel conceptual design for human-agent music performances. This system has not been attempted before according to the literature.

3. SYSTEM DESIGN

The proposed multimodal music performance platform is designed as a human-machine interactive system based on the principles and specifications listed below. The proposed design is implemented as a proof-of-concept prototype to validate and explore user interaction and experience.

[1]https://www.microsoft.com/en-us/hololens

3.1 Design Principles and Specifications

The proposed design is based on the following principles and specifications:

1. **Extending Traditional Interfaces:** The system enhances existing interfaces by providing a 3D performance experience that uses the ubiquitous *musical keyboard* as the physical interface.

2. **Multiple Sound Sources:** The system allows the creation of multiple sound sources, represented as 3D entities in space, which are instantiated using a multitrack looper operated by the physical interface. These sound sources are loops of musical material and are referred to as *Agents*.

3. **Real-Time Sound Synthesis:** The system generates sounds for each musical line in real-time, providing flexible timbre control that can be adjusted using the physical interface.

4. **Physical Space:** The performer's body and the surrounding space are integral to the experience, requiring sufficient physical room for movement, exploration, and interaction with sound sources.

5. **Sound Source Spatialization:** The 3D immersive experience includes sound spatialization that does not interfere with body movements and does not require user-attached devices. A loudspeaker array for ambisonic playback is suitable in this case.

6. **Sound Source Spatial Visualization:** The audio position of a sound source is visually confirmed using a synchronized image representation of a simple colored shape in 3D space. MR technologies over a physical-virtual space generate this image.

7. **Human-Agent Interaction:** Users can move a sound source by representing it as a physical object and mapping its position to the target source. Motion tracking strategies enable this mapping. A sound source is considered a *Musical Agent*, allowing the user to enable/disable autonomous behavior via its 3D visual representation, which is possible through MR technologies.

8. **Agent Autonomy:** A sound source acting autonomously can move around the room, influenced by the user and other agents in the environment. As a musical agent, it can change the musical line based on the original material provided, attempting to keep the user's playing style. The user can override the material when the autonomous behavior is disabled.

3.2 Operation Description

The system is centered on *Agents*, which represent *sound sources* that can move autonomously in space. These agents are provided with *musical material* by a human performer, who is able to interact with them using a physical musical interface (keyboard) and a mixed reality (MR) headset. The system also includes Spatial Audio and Motion Capture (MOCAP) sub-systems. The physical-virtual interactive space is illustrated in Figure 1, where the performer wears the MR headset to visualize the agents as moving spheres.

The system architecture is presented in Figure 2, which shows how the different sub-systems are integrated through a *Core System* to capture input from and provide output to the user.

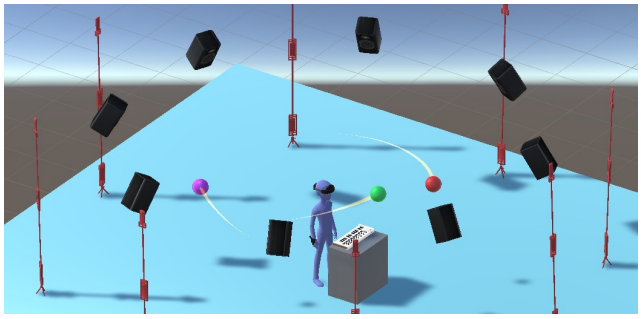

Figure 1: ***3D Representation of the Physical-Virtual System Layout.*** **A human performer interacts with sphere-like virtual entities (agents) wearing a MR headset and surrounded by MOCAP and Spatial Audio systems.**

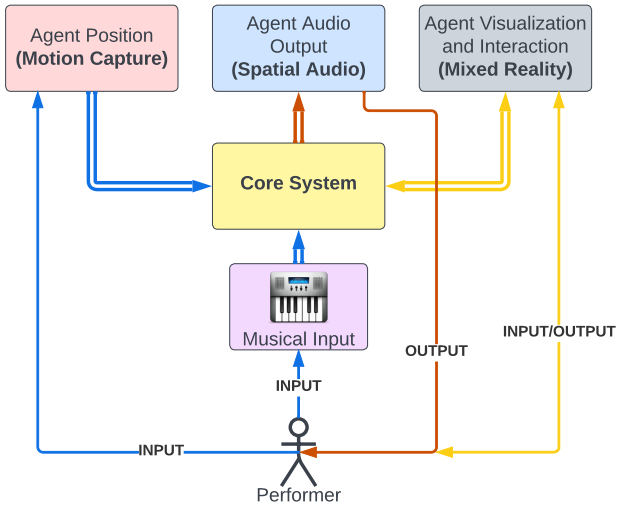

Figure 2: ***System Architecture.*** **The wider arrows represent inputs and outputs from the components to the system, and the thinner ones show inputs provided by the user and outputs received from the system.**

The performer creates a sound source by playing a musical line on the physical keyboard and modifying the sound properties using filters and effects using physical knobs. The *Core System* includes a looper that records and repeats this musical material, creating a sound source that can be heard and seen in space. This sound source is known as the *Musical Agent*, and it can be manually moved using a physical object known as *rigid body* (called *spatial positioner* in this paper), which is tracked by the MOCAP system. The *Spatial Audio system* maps the sound source position to a circular loudspeaker array, and the MR headset renders the agent as a colored sphere in the physical space.

The MR headset also recognizes simple hand gestures and the user's position, which enables agent interaction through an action for "tapping" the sphere-like agents from a distance. Once tapped, the agent is released and starts moving autonomously, changing the musical material (but keeping the sound properties) of the loop based on a machine learning algorithm that is fed in real-time while the user initially played the musical line. Upon releasing an agent, a new one is instantiated over the *spatial positioner*, allowing the user to initialize it with a new loop, and then release it again. This process can be repeated several times to generate a multi-track musical session, with each looping track associated with a sphere-like agent traveling around the 3D

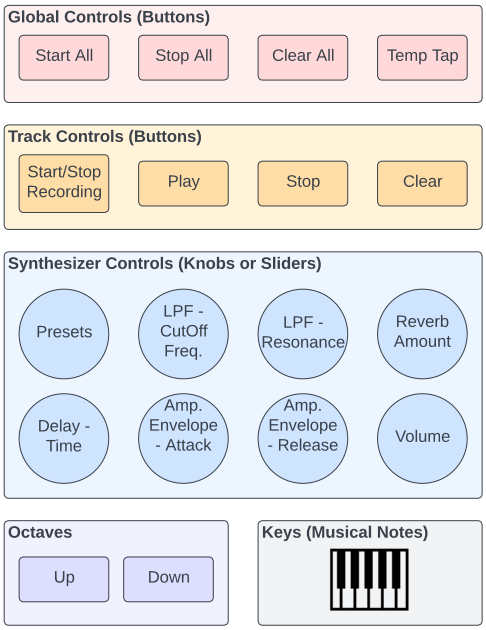

Figure 3: Abstract layout and mapping to synthesis parameters of the physical interface, which allows users to feed musical material into the system.

audio-visual space.

As the agents move freely in the performance area, the user can also move around the physical space. The user can catch released agents to modify the musical loop as well as the sound properties, and then release them once again in this human-machine music interaction. We describe the specific sub-components of this design below.

3.3 System Components and Integration

To provide the functionalities described above, the proposed design requires the integration of three key I/O technologies: *Motion Capture*, *Spatial Audio*, and *Mixed Reality*. These technologies are combined through a *Core System* which manages the data flow across sub-systems, as previously illustrated in Figure 2. The following sub-components are integrated into a proof-of-concept prototype.

3.3.1 Musical Input and Sound Synthesis

A physical interface is mapped on the control parameters of a multi-track looper and synthesizer, as shown in Figure 3. Presets, including pre-defined ones, allow users to store and recall the synthesizer's settings. The multi-track looper consists of individual instances working as illustrated in Figure 4, they record and play back polyphonic note messages synchronized with a metronome, which tempo can be changed in real-time using a "Tempo Tap" button on the physical interface. We use a digital sound synthesizer that operates through the architecture shown in Figure 5, with an instance created for each looper track. The outputs are routed to a *spatializer* module, which generates signals for the loudspeaker array based on the 3D spatial coordinates of the sphere-like agents.

3.3.2 Spatial Audio

The *spatializer* processes the mono output from the synthesizer and renders a 3D sound environment using ambisonics. This environment is then reproduced on a circular array of

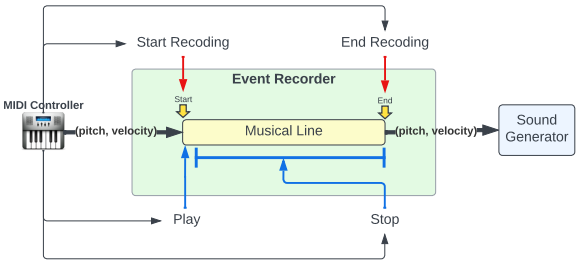

Figure 4: Looper operation diagram. It allows the user to record and reproduce note messages (pitch, velocity) for endless playback.

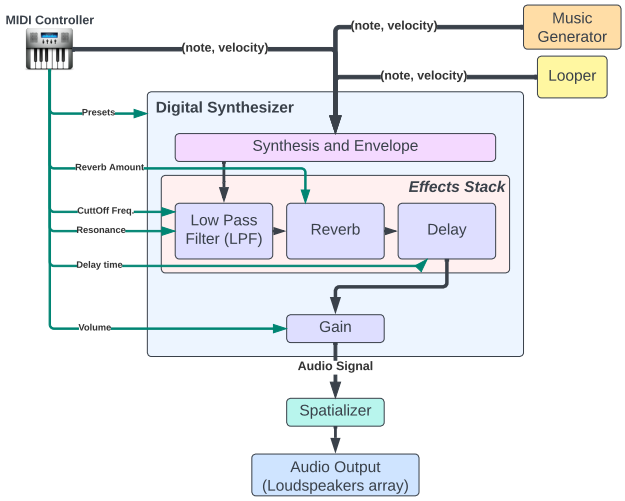

Figure 5: Sound generator and synthesis modules. A digital synthesizer produces the audio output for (note, velocity) messages received from several sources. The effect stack order differs across presets.

eight speakers, as shown in Figure 6. This approach eliminates the need for the performer to wear additional devices, such as headphones and head trackers required for binaural rendering. However, this choice may determine some loss of 3D audio rendering precision, which is compensated visually by the sphere-like agents. Additionally, although MR headsets such as the HoloLens provide a built-in solution for spatial audio, the sound quality and sense of distance are significantly lower than an infrastructure based on physical sources like a set of loudspeakers. Therefore, this spatial audio solution is chosen to enhance the user experience and improve the sound quality.

The ambisonic encoder receives not only the mono audio signals but also the control parameters, such as position coordinates corresponding to each sound source, whether from an *autonomous behavior* or a *manual movement* through the *spatial positioner*.

3.3.3 Motion Tracking

The *spatial positioner* is a physical object, specifically a *rigid body*, that can be precisely tracked using an *optical motion capture system*. This system is commonly used in spatial audio applications due to its accuracy and reliability [13, 8]. The use of this technology is essential to ensure that the target object remains as independent as possible from the user, enabling greater flexibility in object placement. As such, given this design context, it is not feasible

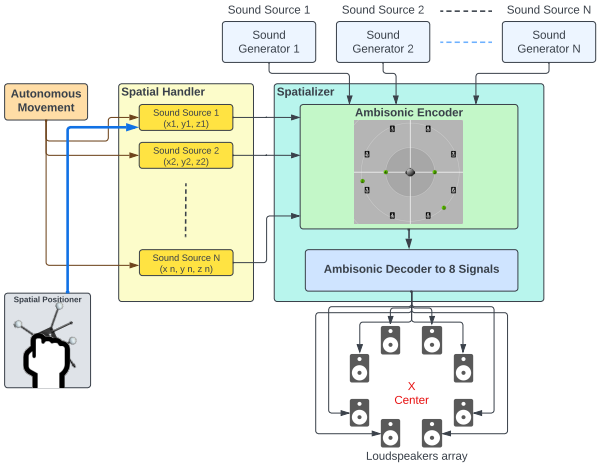

Figure 6: The spatializer receives all the *sound sources* (agents) and their corresponding control signals to arrange them in a 3D space through ambisonics. The agents can be manipulated either through the *spatial positioner* (one agent at a time) or through *autonomous movement*.

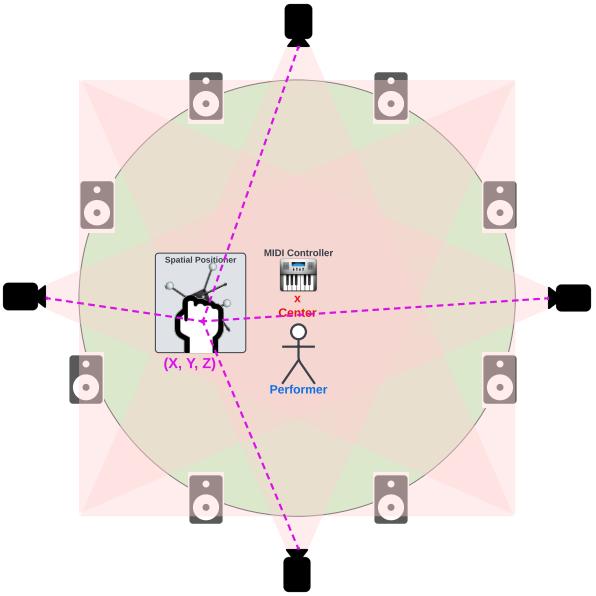

Figure 7: The *spatial positioner* being tracked in the performance area. Its position is estimated by a set of motion sensors spread throughout the room.

to develop a solution using current MR devices such as the HoloLens (version 1 used in this work), which has limited hand-tracking capabilities regarding specific gestures due to its limited Field Of View (FOV).

Figure 7 shows how this object is represented in the performance area as well as other elements that correspond to the system. This object allows the movement of a sound source in space and provides the coordinates to the Core System for audio and visual feedback.

3.3.4 Visualization

To enhance the immersive 3D embodiment of the musical agents, we reinforce the spatial auditory display with a coherent representation in the visual domain. This further confirms the presence of sound sources (agents) at a spe-

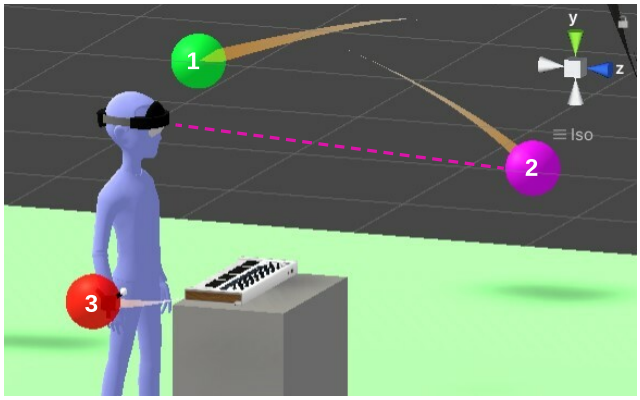

Figure 8: Agents visualization through a mixed reality headset. They are represented as colored spheres with numbers that identify their track number. In this image, attention is directed towards agent 2.

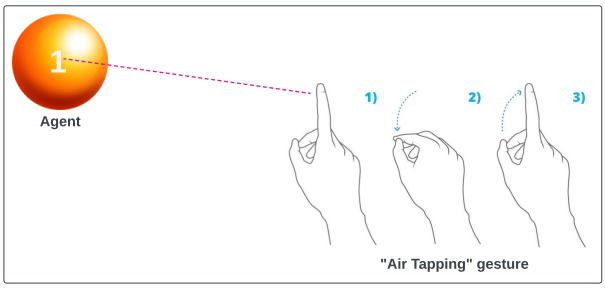

Figure 9: "Air Tapping" gesture for "catching" or "releasing" an agent-as-sphere object.

cific position in space, which changes over time (the agents dynamically move around the performer). We achieve this through a *mixed reality* (MR) system, as such, this solution connects sensory modalities and increases affordances [20].

Figure 8 depicts the agents as colored spheres around the performer, with numbers that identify their track number. We chose a spherical shape for the agents because it can unconsciously attract human attention, and curved objects have positive associations in human evolution [11].

Note that agent 3 is attached to the *spatial positioner* in the right hand, which means that this agent is "locked", while agents 1 and 2 move autonomously. Moreover, since the MR device is a headset, there are limitations in terms of FOV. Hence, attention mechanisms such as *directional indicators* (i.e., movement tails), *graphical feedback*, and *labeling* are important, as shown in the image.

3.3.5 Interaction

We aim to keep the interaction as simple as possible. We assume that performers are familiar with physical control interfaces, such as a piano-like keyboard, push buttons, and rotary knobs, which are used to operate the looper. Additionally, the *spatial positioner* provides a straightforward tangible medium to manually move an agent in space.

Another dimension of interaction involves how users "catch" and "release" agents. Users can perform a simple "air tapping" gesture[2] with either hand when the gaze is pointing to an agent-as-sphere object, as shown in Figure 9.

[2]https://learn.microsoft.com/en-us/hololens/hololens1-basic-usage#select-holograms-with-gaze-and-air-tap

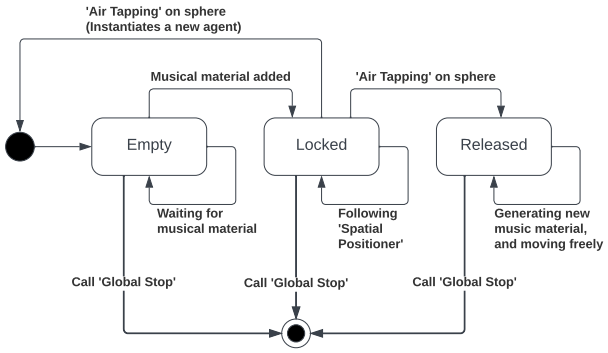

Figure 10: Finite State Machine (FSM) for agent behavior.

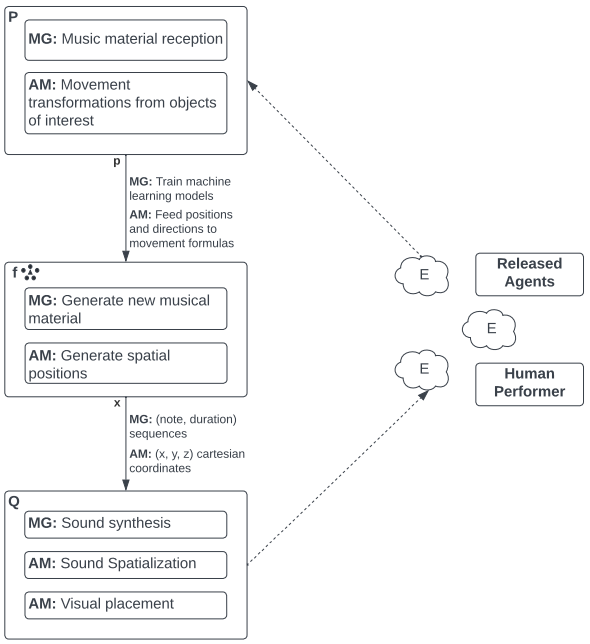

Figure 11: The system representation under the *PQf architecture* for computer music systems proposed by Blackwell [1]. MG stands for Music Generation, and AM for Autonomous Movement.

3.3.6 Autonomous Behavior: Agent Representation

An "agent" refers to an individual entity that can change musical material from the human performer and move freely in space when instructed. This behavior is defined in the *Finite State Machine (FSM)* shown in Figure 10.

An agent is associated with a sound source and a track in the musical session, making it part of a general synchronization commanded by a global metronome. It is part of a group consisting of other agents and the performer, as detailed below.

3.3.7 Autonomous Behavior: Swarm Representation

As a group, the agents act as an *artificial swarm* modeled under the *PQf architecture* illustrated in Figure 11. This architecture presents two algorithmic approaches for collective behavior: *Music Generation (MG)* and *Autonomous Movement (AM)*.

The performer and the released agents listen to each other in the environment (E). The analysis module (P) takes the

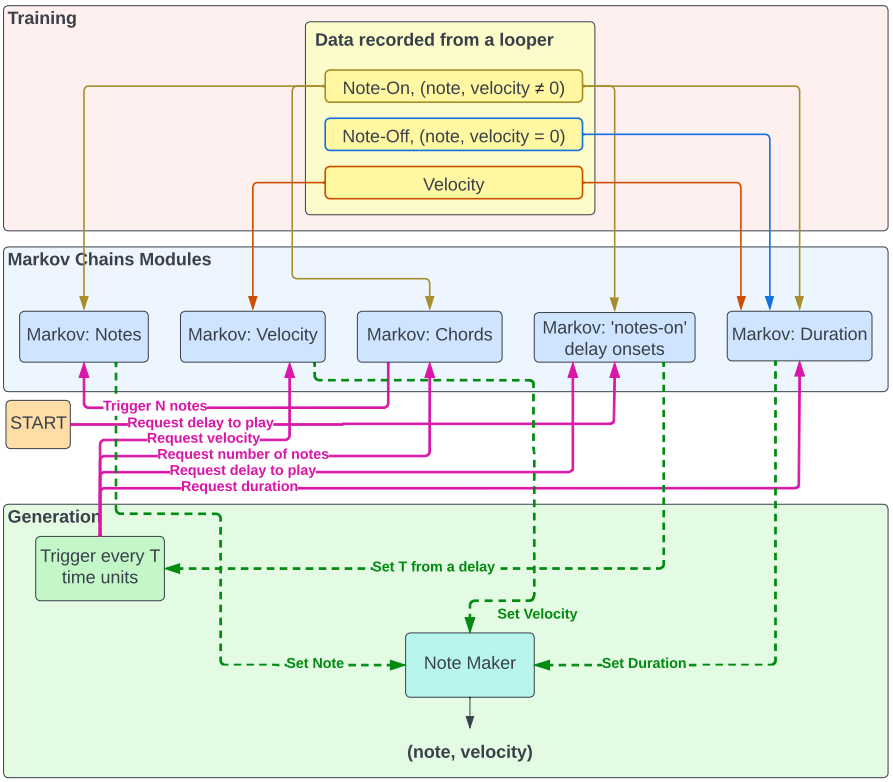

Figure 12: Markov Chains system designed by Samuel Pearce-Davies. It uses 5 Markov Chain modules to generate musical material with several properties, aiming to make it as "musical" as possible.

musical material from the performer to feed the MG algorithm, as well as the position data for AM. The real-time process takes the results from the (P) module to train MG models and evaluate AM formulas in the (f) module, in which new data is generated. This new material is synthesized by the (Q) module to produce spatialized sound and move the agents accordingly. It then returns to the environment (E) to start a new cycle.

3.3.8 *Autonomous Behaviour: Music Generation Algorithm*

The algorithm used for music generation in the system is based on *Markov Chains*. This approach generates sequences based on transition probabilities and attempts to follow the performer's style using recorded music lines. Markov Chains have been extensively used in music generation and can potentially reach an adequate level of quality depending on the input. Additionally, this technique is recommended for design criteria that require a simple but efficient music generation tool [3]. In the case of real-time music generation, a computationally-efficient solution is essential.

This system integrates a custom version of the implementation proposed by Samuel Pearce-Davies[3] for a polyphonic music generator based on several instances of Markov Chains, as illustrated in Figure 12. It uses MIDI data based on notes, velocities, note groups, 'note-on' delay onsets, and duration. With this data, it is possible to generate melody, harmony, and rhythm as musical material. The original version was modified to include the global metronome for synchronization. Each agent incorporates an instance of this modified version to allow parallel music excerpts.

[3] `https://spearced.com/algorithmic-process-ai/`

3.3.9 *Autonomous Behaviour: Autonomous Movement Algorithm*

One of the motivations behind this work is the embodiment of sound sources as individual entities capable of interacting among themselves and with a human performer. To achieve this, a *"sense of agents' motion from the user"* and *"awareness from the agents regarding the user"* are necessary. Therefore, the system aims to enable agents to be spatially aware of the performer and other agents. According to that, the proposed algorithm controlling the movement considers three spatial sources to calculate the final position $\vec{P}_x$ for an agent x, assuming that the room center is the point $(0, 0, 0)$l, which are:

1. A base motion given by a circular path around the center, as point $\vec{P}_{x:base}$.
2. The position $\vec{P}_{x:swarm}$ such that all agents are equally spread around the performer position $\vec{P}_h$, providing agents' co-awareness.
3. The performer gaze direction given by $\hat{\mathbf{dir}}_h$.

For smooth movement, some position calculations need to be interpolated between points $\vec{a}$ and $\vec{b}$ during a time t. For this purpose, we use a *linear interpolation* model given by (1).

$$\vec{P}_{lerp(a,b)} = (1-t)\vec{a} + t\vec{b} : \vec{a}, \vec{b} \in \mathbb{R}^3; t \in \mathbb{R} \qquad (1)$$

For an agent x, the circular base movement for point $\vec{P}_{x:base}$ starts with spherical coordinates $(r_{x:init}, \theta_{x:init}, \phi_{x:init})$ such that r is the radius, θ is the azimuth angle, and ϕ is the elevation angle. This position is given when the user *releases*

agent x, i.e., the last position where it was attached to the *spatial positioner*. The circular movement keeps the same angles $r_{x:init}$ and $\phi_{x:init}$, while the azimuth angle changes according to (2) with a speed S_b. Here, Δt is the frame time of the calculation.

$$\theta_{x:base} \coloneqq \theta_{x:base} + S_b \Delta t : \theta_{x:base} = \theta_{x:init} \text{ when it starts}; \theta_{x:init}, \theta_{x:base} \in \mathbb{R} \quad (2)$$

Thus, the circular movement for point $\vec{P}_{x:base}$, in Cartesian coordinates, comes from the spherical coordinates ($r_{x:init}$, $\theta_{x:base}$, $\phi_{x:init}$).

The position $\vec{P}_{x:swarm}$ for the agent's co-awareness is based on the equation for calculating the center of a group of points. In this case, this center is the performer position $\vec{P}_h$. The position of an agent is $\vec{P}_i$ in a set of N individuals, including the target x. Hence, $\vec{P}_{x:swarm}$ is computed by (3).

$$\vec{P}_{x:swarm} = \vec{P}_h - \sum_{i=1, i \neq x}^{N} \vec{P}_i : \vec{P}_h, \vec{P}_i \in \mathbb{R}^3; i, x, N \in \mathbb{N}^+ \quad (3)$$

Since the gaze direction $\hat{\mathbf{dir}}_h$ is a vector of unit length, it will be used for the final calculation as if it were a point in space one unit away from the center. Thus, $\vec{P}_x$ can be obtained by (4).

$$\vec{P}_x = \alpha(\vec{P}_{x:base} + \vec{P}_{x:swarm} + \hat{\mathbf{dir}}_h) : \vec{P}_{x:base}, \vec{P}_{x:swarm}, \hat{\mathbf{dir}}_h \in \mathbb{R}^3; \alpha \in \mathbb{R} \quad (4)$$

The constant α allows the calculation of the average position, where this value is $1/2$ when the agent size is 1 (since there is no contribution from other agents) and $1/3$ when it is greater than 1.

To achieve agents' movement variability, we employ the following techniques:

- The direction of rotation in Equation (2) is inverted by multiplying S_b with -1 when the distance between $\vec{P}x$: swarm and an agent $\vec{P}i$ is less than m units (usually when agents are close and need to separate from each other).
- The gaze direction $\hat{\mathbf{dir}}_h$ is randomly inverted ($-\hat{\mathbf{dir}}_h$) every frame.
- $\vec{P}_{x:swarm}$ and $\hat{\mathbf{dir}}_h$ are interpolated in real-time using (1) with a small time value t_{lerp} for smooth movement.
- S_b and t_{lerp} are arbitrary values chosen by the designer to adjust the movement dynamics.

4. IMPLEMENTATION

A proof-of-concept was implemented in a 6 x 6 meter room using the following hardware and software resources.

- The *Core System* is programmed in *Max 8*[4].
- Communication between the *Core System*, the *Motion Capture*, and *Mixed Reality* systems is established using Open Sound Control (OSC) messages. *Spatial Audio* runs on the same computer as the *Core System* using Inter-process communication (IPC).
- The *Core System* runs on a 64-bit Windows 10 computer with an Intel Core i7-7700k 4.20 GHz processor and 16 GB RAM.
- The *spat*[5] library from IRCAM was used for spatialization. Ambisonic (aep2d panning) was used to render audio to an array of eight loudspeakers (Genelec 8030C) arranged in a circular configuration with a radius of 2 meters and a height of 2 meters.
- We use a *Midas M32 Digital Mixer*[6] as a USB soundcard through the Klark Teknik DN32-USB Expansion Module at 44.1 Hz and a buffer size of 1024 samples. The digital mixer routes the decoded ambisonic signals to the circular loudspeaker array.
- The physical controller interface is a standard keyboard, the *AKAI MPKmini II*[7], which includes buttons and rotary knobs for the looper and sound synthesis control.
- The motion capture system is an *OptiTrack*[8] running at 120 fps on an independent computer using the software *Motive*[9]. Communication with this computer is through a wired LAN using a router (150 Mbps TP-LINK TL-WR741ND) to share OSC messages.
- The mixed reality device is a *Microsoft HoloLens (Version 1)*[10] on which a custom application running at 60 fps was developed using the *Unity3D*[11] game engine. Communication is established using the same router for other computers but through a 2.4GHz Wireless LAN with WPA2 encryption.
- The digital synthesizers driven by the loopers are instances of the *Tunefish 4*[12] VST plugin.
- For the autonomous movement algorithm, we chose $S_b = 20$ degrees/s and $t_{lerp} = 0.1s$. The agents' position update process occurs every 30 ms.

The use of this proof-of-concept is illustrated in several views in Figure 13. A video demonstration was recorded to show how the implemented system works from these perspectives.[13]

5. EVALUATION

We conducted experiments to validate and evaluate the proposed design and its implementation. This includes determining the maximum number of supported agents, collecting and analyzing system internal metrics (latency, jitter, and packet loss), recording and analyzing user-generated data during user studies, and collecting feedback and reflections from users using surveys.

The current system implementation can support up to 8 agents. Beyond this point, our implementation platform can no longer cope with the computation required to meet the audio real-time constraints. The system operates with an audio sampling rate of 44.1 kHz and a buffer size of 1024 samples. The internal metrics consider 1 to 8 agents, and the user evaluation uses a version where a performer can work with up to 8 agents simultaneously.

5.1 System Measurements

[4] `https://cycling74.com/products/max`

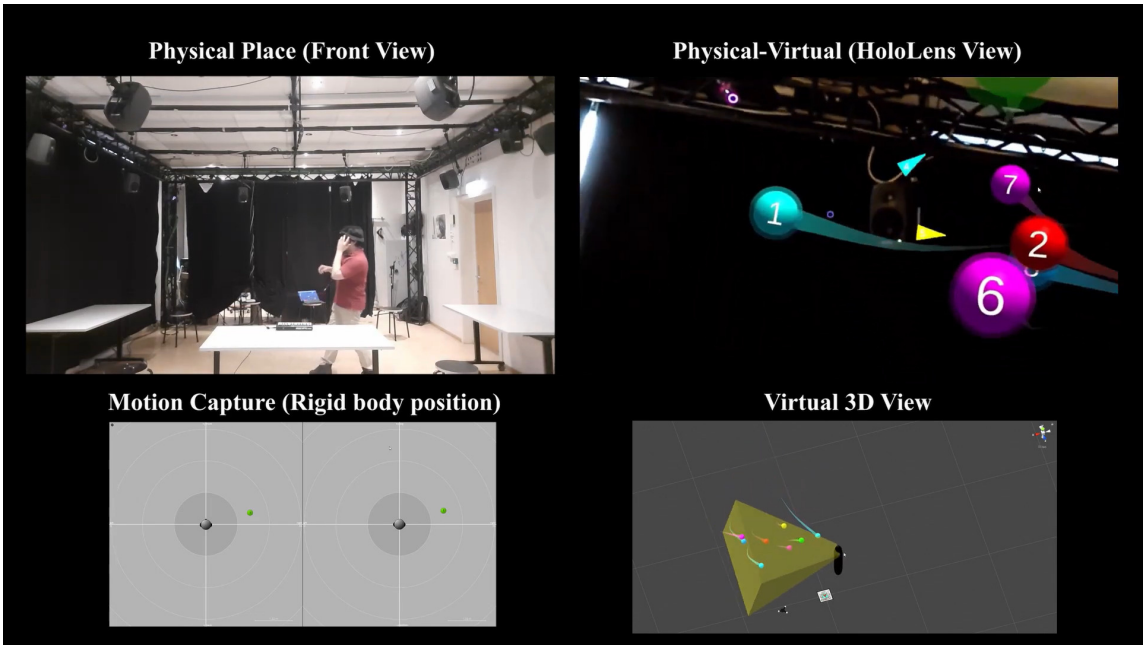

Figure 13: The prototype used in a live session. These four views allow an understanding of the operation of the system from several perspectives.

Table 1: Summary of the average system measurements with agents' group size from 1 to 8 active agents. Latency and jitter values are given in milliseconds and packet loss is in percentage.

Parameter	Agents' Group Size							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Physical keyboard to Sound Output Latency (ms)	40.56	38.43	36.20	39.62	37.33	38.93	35.83	39.07
Physical keyboard to Sound Output Jitter (ms)	6.00	6.45	6.78	8.04	6.81	6.73	7.83	7.05
Spatial Audio Placement Latency (ms) (Between the rigid body movement and the sound output panning from the loudspeaker array)	118.23	112.67	121.00	129.34	123.78	129.34	126.56	134.89
Sound to Visualization Latency (ms) (Between hearing a sound source - agent - from a point in the loudspeaker array and visualizing it through the HoloLens)	3.25	3.03	4.40	4.39	5.36	5.37	4.62	6.75
Sound to Visualization Jitter (ms)	8.62	9.81	12.30	12.81	12.89	15.16	11.66	13.44
Packet Loss Core-to-HoloLens (%)	19.08	20.80	20.33	20.71	21.72	24.14	23.56	23.89

We measure latency and jitter values in several stages, taking 30 readings for each case. Packet loss is measured from the *core system* to the *motion capture system* and the *mixed reality headset*. In both cases, repeated measurements are taken, increasing the number of agents from 1 to 8. The motion capture system uses a wired LAN connection, and packets are sent for one hour at 120 Hz without any loss. For the mixed reality headset, packets are sent over one minute at 33.33 Hz (30 ms period), and the test is repeated 30 times each time the size of the agent group is increased.

Table 1 shows the average values of the measurements. Latency values are high due to DSP for spatial audio calculations and sound generation, as well as graphics rendering for visualization. Additionally, packet loss is caused by wireless communication.

5.2 User Study

The system was evaluated with a user study focused on user-agents interaction involving seven participants. In each session, participants were allowed to improvise a musical piece with the system and manipulate up to 8 agents. The participants were experienced musicians with formal education and understanding of loopers, sound synthesis, music improvisation, spatial audio, and optical motion tracking, but no prior experience with the Microsoft HoloLens.

Each session was divided into three parts: an explanation of the system's features and functionalities together with the HoloLens standard tutorial, improvisation of a musical piece, and a survey about the experience. On average, participants spent 35 minutes using the system, with session times ranging from 24 to 43 minutes.

5.2.1 Captured Data

[5] https://forum.ircam.fr/projects/detail/spat/
[6] https://www.midasconsoles.com/product.html?modelCode=P0B3I
[7] https://www.akaipro.com/mpk-mini-mkii
[8] https://optitrack.com/
[9] https://optitrack.com/software/motive/
[10] https://www.microsoft.com/en-us/hololens
[11] https://unity.com/
[12] https://www.tunefish-synth.com/
[13] https://www.youtube.com/watch?v=6wm24BC5NLg

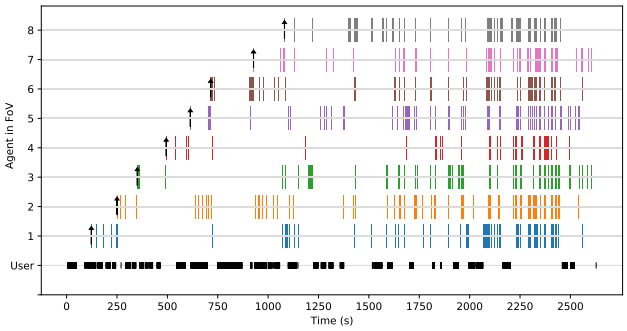

Figure 14: Periods when agents were in the HoloLens' FOV during the performance session for User 6. The first row 'User' represents the user activity on the physical keyboard during the session. The arrows represent the moment of the first appearance of an agent.

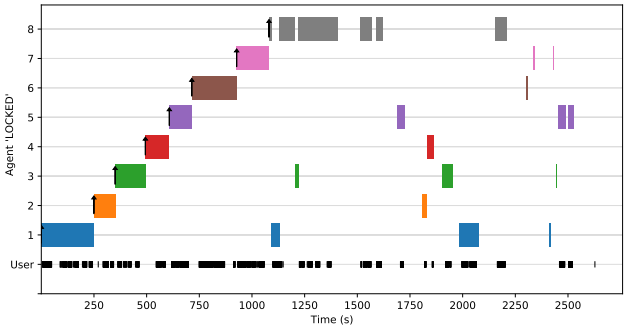

Figure 15: Periods when agents were "LOCKED" during the performance session for User 6. The first row 'User' represents the user activity on the physical keyboard during the session. The arrows represent the moment of the first appearance of an agent.

A Core System's recording module captured anonymous user data related to human-agent interaction. We present the results for one user (User 6) to illustrate how the data is processed and analyzed for an individual.

To assess users' attention to the agents, we checked whether they observed the agents. We estimated this by determining the number of agents that appeared in the FOV of the HoloLens. Figure 14 shows when each of the eight agents was visible in User 6's FOV during the session, along with the user's physical keyboard activity.

Aside from manipulating the physical keyboard, the user could *lock* or *release* agents using the air tapping gesture. Figure 15 shows the time when agents were "LOCKED". The rest of the time, agents were "RELEASED", meaning they were behaving autonomously. In Figure 14 we can estimate when the user observed the agents during these moments.

Figure 16 shows heatmaps for every agent in User 6's session. The heatmaps illustrate different movement patterns that last for a considerable amount of time, as seen in the hot spots, which can contribute to movement predictability.

5.2.2 *Survey*

After the session, the seven participants completed a survey with 38 questions organized into five groups. Results are discussed qualitatively due to the small sample size.

1. Latency and Jitter: Participants perceived minimal delay related to *Physical keyboard to sound output latency* (44.46 ms) except for User 2, who accelerated the tempo

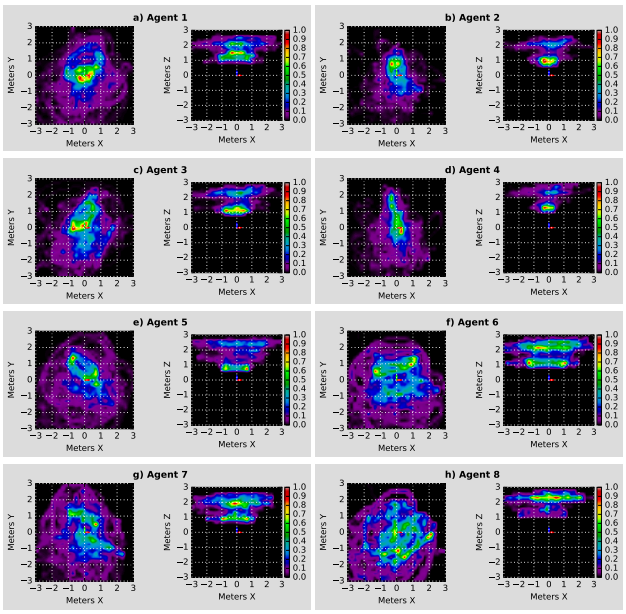

Figure 16: Heatmaps for the movement of every agent during the performance session for *User 6*. It shows the locations where they spent most of their time, as well as trajectories. The view is from top (x, y) and from back (x, z) per each agent.

in some parts of the performance. Scores for *Spatial Audio Placement latency* (134 ms) and *Sound to Visualization latency* (6.747 ms with a jitter of 13.44 ms) were around 8 out of 10 for alignment perception. Given the amount of latency, this relatively good perception can be attributed to the human limitation in identifying directional sounds [12]. Participants had difficulty localizing sounds but found the HoloLens helpful to support this task.

2. Music Improvisation: Participants felt familiar with the physical keyboard for the looper operation, but not in complete control of the human-machine music interaction.

3. Autonomous Movement: Five participants preferred a slower speed for agents in some situations and noticed quasi-predictable patterns and behavior similar to a school of fish.

4. MR Experience: Three users found the HoloLens uncomfortable initially but became immersed in the experience afterwards. The FOV did not significantly restrict their ability to identify agents.

5. Overall User Experience: Participants rated aesthetics as 6 out of 10, with varying opinions on the number of agents. Investigating the motivations behind selecting a specific number of agents is a potential future work. Ease-of-use and enjoyment were also rated (Figure 17 and Figure 18). Participants found the system enjoyable and regarded it as a non-conventional medium for music composition.

5.2.3 *Observations and Reflections*

Despite latency limitations, users were able to improvise a musical piece for a relatively long period, which indicates that fluency was not significantly affected. On some occasions, all users felt that the agents changed their musical intention, which was undesirable for 5 of them, but for the other 2, it was an opportunity to follow the agents' music composition and perform accordingly. Most of the time, users contemplated the agents and focused their attention on the machine's performance, showing curiosity in the de-

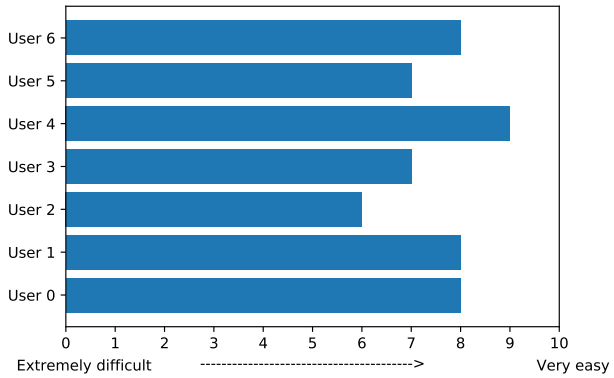

Figure 17: How easy was to use the whole system?

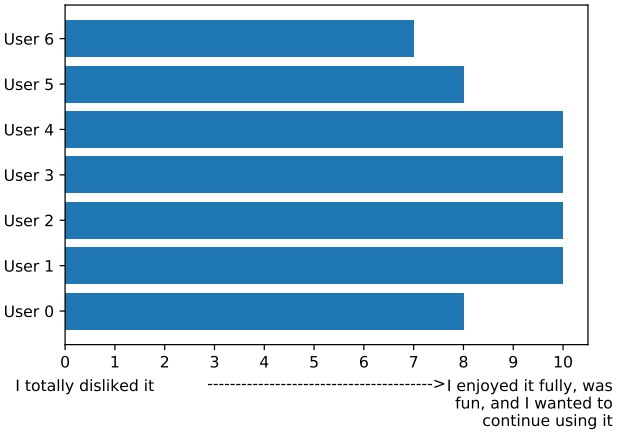

Figure 18: How much did you enjoy the performance?

velopment of the musical piece. Moreover, all users agreed that the machine was able to replicate their style to some extent. In general, users rated the system as *easy-to-use*, and the experience as *highly enjoyable*, which can be related to the relatively long time they spent in the free improvisation session.

Table 2 provides reflections from further observations, comments, and captured data regarding the design principles established earlier.

6. CONCLUSIONS

In this paper, we presented the design, implementation, and evaluation of a human-machine interactive music system based on autonomous agents in mixed reality. The system integrates concepts related to *Musical Agents*, *Live Algorithms*, *Swarm Intelligence*, and *Extended Reality*, and combines *Motion Capture*, *Spatial Audio*, and *Mixed Reality* technologies to provide a multi-modal interactive experience.

The system was evaluated through a user study involving 7 participants who performed free music improvisation with the system. The study revealed limitations in terms of latency, jitter, and package loss, which were identified during system measurements before the user study. However, despite these limitations, the study validated the conceptual design and demonstrated the feasibility of the integrated solution for human-machine music performances and the interaction possibilities for the users.

Future work includes system optimization through more efficient software components and higher-performance hardware. It also includes exploring alternative multi-agent algorithms and conducting a larger study with a representative sample of participants to perform a quantitative analysis. Additionally, further human aspects will be explored from a musicological perspective.

7. ACKNOWLEDGMENTS

We would like to thank Egidijus Pelanis, a researcher from the *Norway University Hospital – Rikshospitale*, for providing a HoloLens headset and supporting this work with insights and suggestions.

8. ETHICAL STANDARDS

This study followed all ethical and data protection guidelines from the University of Oslo (UiO). Master's students from the Department of Musicology at UiO, including international programs, participated in the study, and their identities were anonymized in the collected data. Written consent was obtained to publish any musical content (only audio) they produced during the sessions. The Microsoft HoloLens device was provided by the *Norway University Hospital – Rikshospitale*, with no potential conflicts of interest with this work.

9. REFERENCES

[1] T. Blackwell. Swarming and Music. In *Evolutionary Computer Music*, pages 194–217. Springer London, London, 2007.

[2] T. Blackwell, O. Bown, and M. Young. Live Algorithms: Towards Autonomous Computer Improvisers. In *Computers and Creativity*, volume 9783642317, pages 147–174. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[3] M. Catak, S. AlRasheedi, N. AlAli, G. AlQallaf, M. AlMeri, and B. Ali. Artificial Intelligence Composer. In *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pages 608–613. IEEE, 9 2021.

[4] M. Costagliola. Multi-user shared augmented audio spaces using motion capture systems, 2018.

[5] S. Das, S. Glickman, F. Y. Hsiao, and B. Lee. Music Everywhere - Augmented Reality Piano Improvisation Learning System. *Proceedings of the 207 International Conference on New Interfaces for Musical Expression*, pages 511 – 512, 2017.

[6] A. D'Ulizia. Exploring Multimodal Input Fusion Strategies. In *Multimodal Human Computer Interaction and Pervasive Services*, pages 34–57. IGI Global, 2009.

[7] M. Graf and M. Barthet. Mixed Reality Musical Interface: Exploring Ergonomics and Adaptive Hand Pose Recognition for Gestural Control. In *NIME 2022*, pages 1–26. PubPub, 2022.

[8] F. Grani, D. Overholt, C. Erkut, S. Gelineck, G. Triantafyllidis, R. Nordahl, and S. Serafin. Spatial Sound and Multimodal Interaction in Immersive Environments. In *Proceedings of the Audio Mostly 2015 on Interaction With Sound - AM '15*, volume 07-09-Octo, pages 1–5, New York, New York, USA, 2015. ACM Press.

[9] R. Hamilton, J.-P. Caceres, C. Nanou, and C. Platz. Multi-modal musical environments for mixed-reality performance. *Journal on Multimodal User Interfaces*, 4(3-4):147–156, 12 2011.

[10] R. Ishino and N. Tokui. MIAMI: A Mixed Reality

Table 2: Reflections on every design principle addressed in the system.

Design Principle	Value
1. Extension of traditional interfaces	A piano-like controller is a suitable device for providing musical material in an extended 3D experience due to its familiarity. Latency issues must be addressed in the system if they increase as a result of intensive resource usage.
2. Multiple musical sound sources	The multi-track looper is an appropriate means for multiple sources and an excellent complement to the piano-like controller.
3. Sound synthesis	As a complement to the previous two principles, the sound synthesis approach provides flexibility to custom compositions in terms of timbre.
4. Physical space	A 6 x 6 meter room provides a suitable space for the performance. Tables, chairs, and other supporting objects can be used to place the *spatial positioner* and explore the agents' movements.
5. Sound source spatialization	A loudspeaker array with ambisonic playback allows the performer to move fluidly around the room. Limitations in terms of latency and spatial resolution are not significant issues due to human constraints.
6. Sound source spatial visualization	A 3D representation of sound sources compensates for spatial sound limitations and adds precision to agent identification. A MR headset is suitable for this purpose, but ergonomic issues need to be solved for long performances.
7. Human-agent interaction	The spatial positioner, as a physical object, facilitates the manual movement of a sound source and can be left in several places around the room as the performer carries out other actions. The MR agent selection gesture provides a straightforward way for agent interaction.
8. Agent autonomy	Agents' movement and music generation is a spectacle that attracts user attention. Moreover, performers try to modify the composition through agent interaction to observe and hear how the result changes. However, users demand more manual control of agents in some situations.

Interface for AI-based Music Improvisation. In *NIME 2022*. PubPub, 2022.

[11] M. Lima. *The Book of Circles*. Princeton Architectural Press, 1st edition, 2017.

[12] A. W. Mills. On the Minimum Audible Angle. *Journal of the Acoustical Society of America*, 30(4):237–246, 1958.

[13] J. Müller, M. Geier, C. Dicke, and S. Spors. The BoomRoom: Mid-air Direct Interaction with Virtual Sound Sources. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 247–256, New York, NY, USA, 4 2014. ACM.

[14] R. Nakagawa, R. Komatsubara, T. Ota, and H. Ohmura. Air Maestros: : A Multi-User Audiovisual Experience Using MR. In *Proceedings of the Symposium on Spatial User Interaction*, number 18, pages 168–168, New York, NY, USA, 10 2018. ACM.

[15] A. Riddershom Bargum, O. Ingi Kristjánsson, P. Babó, R. Eske Waage Nielsen, S. Rostami Mosen, and S. Serafin. Spatial Audio Mixing in Virtual Reality. In *Proceedings of the 19th Sound and Music Computing Conference*, pages 100–106, Saint-Étienne, 2022.

[16] I. T. Riley. *Touching Light: A Framework for the Facilitation of Music-Making in Mixed Reality*. PhD thesis, West Virginia University Libraries, 1 2021.

[17] F. A. Robinson. Audio Cells: A Spatial Audio Prototyping Environment for Human-Robot Interaction. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 955–960, New York, NY, USA, 2 2020. ACM.

[18] Y. Tan and Z.-y. Zheng. Research Advance in Swarm Robotics. *Defence Technology*, 9(1):18–39, 3 2013.

[19] K. Tatar and P. Pasquier. Musical agents: A typology and state of the art towards Musical Metacreation. *Journal of New Music Research*, 48(1):56–105, 1 2019.

[20] L. Turchet, R. Hamilton, and A. Camci. Music in Extended Realities. *IEEE Access*, 9:15810–15832, 2021.

[21] Y. Wang and C. Martin. Cubing Sound: Designing a NIME for Head-mounted Augmented Reality. In *NIME 2022*, pages 1–17. PubPub, 2022.

[22] K. C. Zellerbach and C. Roberts. A Framework for the Design and Analysis of Mixed Reality Musical Instruments. In *NIME 2022*, pages 1–21. PubPub, 2022.